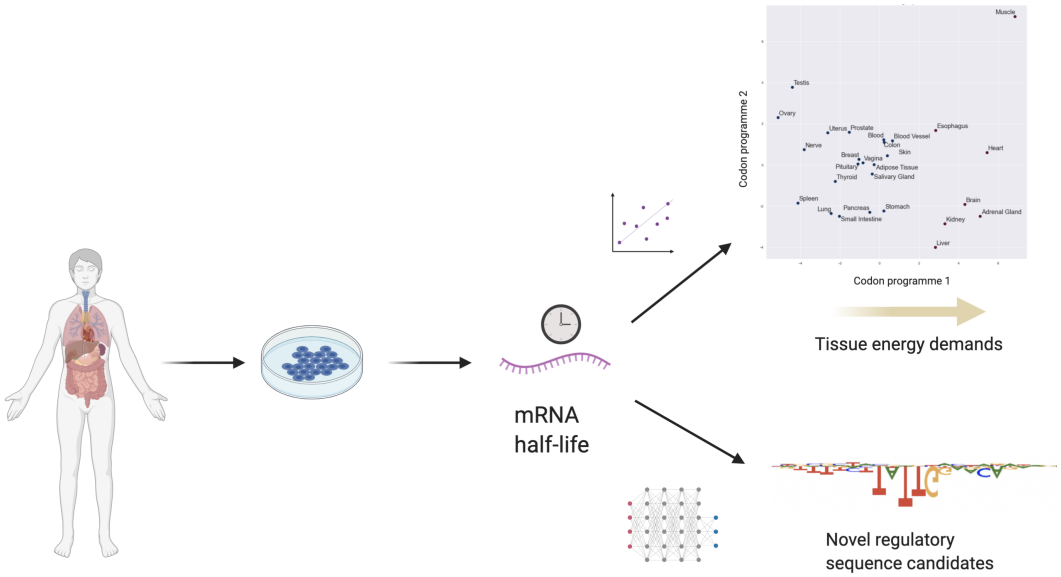
TÉCNICO LISBOA



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

Pedro Miguel Tomaz da Silva

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. Maria Margarida Campos da Silveira
Prof. Julien Gagneur

Examination Committee

Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Prof. Maria Margarida Campos da Silveira
Member of the Committee: Prof. Tiago Paulo Gonçalves Fernandes

January 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Acknowledgments

I would like to thank professor Julien Gagneur for all invaluable mentoring and the opportunity of doing exciting research in his laboratory. I thank Vangelis Theodorakis for his crucial contribution and feedback to the work I present. Furthermore, I would like to thank Florian Hölzlwimmer for the excellent support, tutoring and engaging discussions.

I am grateful to professor Margarida Silveira for all mentoring during my master's year at Instituto Superior Técnico and for her input as supervisor of this thesis.

Finally, I would like to thank my family for their foundational support during my academic journey.

Abstract

In humans, the DNA encodes a program written in a 4-character language using a 3-billion-long-text, which defines the behaviour 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, from which the work I present is now a part of.

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 has to be copied from its storing location and delivered to its production site, where it is used as a template to produce proteins. The molecule responsible for carrying the code is called messenger RNA (mRNA).

Curiously, the program encoded in this message also determines the amount of time it is allowed to run until it finally stops - when the message gets destroyed. The longer the message is available, the more proteins will be produced from it. Hence, the amount of proteins produced from each gene can be regulated by the message contained in the mRNA.

This thesis addresses two overlapping questions: how the variable length 4-character-language sequence contained in the mRNA molecule determines the time window it stays functional - measured by its half-life; what patterns can be found between multiple mRNAs' sequence determinants of half-life in one cell line and throughout populations of cells from different human tissues.

In order to tackle these questions, the approach followed repeatedly in this thesis comprises two steps. Firstly, mRNA half-life is modelled through a prediction task using features derived from its sequence as input. Secondly, information is extracted from the resulting models using different interpretation methods.

Following this approach, firstly I developed a linear regression model explaining the quantitative influence of the most well known sequence features influencing half-life in a human cell line.

Secondly, I created a deep convolutional neural network for mRNA half-life prediction and applied interpretation tools to discover the quantitative influence of each sequence position in the prediction output, revealing possible new regulatory sequence elements.

Thirdly, I developed a multi-task neural network model for the prediction of mRNA half-life variation between cells from different human tissues which can be integrated in models predicting tissue-specific

mRNA levels.

Fourthly, linear modelling of mRNA half-life variation between cells from different human tissues and its subsequent interpretation through principal component analysis, points to a previously unknown connection between a tissue's cell specific mRNA translation effects on half-life and its energy consumption.

Overall, the models created 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. Furthermore, the uncovered biological pathways add to our understanding of cell biology and can further be exploited for both clinical and knowledge purposes.

Keywords

mRNA half-life; modeling; Deep Convolutional Neural Networks, model interpretation; regulatory sequence elements; tissue-specific mRNA half-life variations

Resumo

Nos seres humanos, o DNA codifica um programa escrito numa língua de 4 caracteres, usando um texto com mais de 3 mil milhões de letras. Este texto define o comportamento e a função de cada célula no organismo.

Entender de que forma o código deste texto (genótipo) produz o seu resultado (fenótipo) é um tópico que inclui décadas de investigação, do qual a tese que apresento agora faz parte.

Porções deste código denominadas genes contêm instruções para a produção de proteínas. A regulação da quantidade de proteínas numa célula num dado momento é de importância vital para o seu funcionamento correto.

O código utilizado para a produção de proteínas tem de ser copiado a partir do local onde está armazenado e subsequentemente transportado até ao local de produção de proteínas, onde é usado como modelo para as fabricar. A molécula responsável pelo transporte deste código denomina-se RNA mensageiro ou mRNA.

Curiosamente, o programa codificado nesta mensagem também determina o intervalo de tempo em que este está ativo, ou seja, até à destruição desta mensagem. Quanto mais tempo a mensagem está ativa, mais proteínas são produzidas a partir dela. Deste modo, a quantidade de proteínas produzidas a partir de um gene pode ser regulada a partir da mensagem contida no mRNA.

Esta tese aborda duas questões relacionadas entre si. Uma consiste em investigar de que forma a sequência escrita a partir de uma língua de 4 caracteres contida no mRNA determina o intervalo de tempo em que esta molécula permanece funcional - medido pelo tempo de semivida. Outra questão consiste em investigar que padrões existem entre as várias porções do código que determinam o tempo de semivida do mRNA numa linhagem celular humana e entre várias populações de células provenientes de tecidos humanos.

Com o objetivo de responder a estas questões, foi seguida uma abordagem repetida ao longo da tese que consiste em dois passos. No primeiro, o tempo de semivida do mRNA é modelado a partir de uma tarefa preditiva usando componentes obtidos da sua sequência. Num segundo passo, é extraída informação a partir dos modelos obtidos usando variados métodos de interpretação.

Seguindo esta abordagem, em primeiro lugar desenvolvi um modelo de regressão linear que explica quantitativamente a influência de conhecidos componentes da sequência do mRNA que influenciam o

seu tempo de semivida numa linhagem celular humana.

Em segundo lugar, criei uma rede neuronal convolucional que infere o tempo de semivida do mRNA a partir da sua sequência e apliquei ferramentas de interpretação para quantizar a influência de cada posição da sequência na inferência do modelo, revelando possíveis novos códigos regulatórios do tempo de semivida do mRNA.

Em terceiro lugar, desenvolvi um modelo baseado em redes neuronais que infere as variações do tempo de semivida do mRNA entre differentes células de múltiplos tecidos humanos. Tal modelo mostra potencial para ser integrado em modelos de inferência dos níveis de mRNA entre tecidos humanos.

Por último, a análise a um modelo de regressão linear desenvolvido, que infere a variação do tempo de semivida do mRNA entre tecidos humanos e a sua subsequente interpretação a partir de uma análise de componentes principais, sugere uma associação entre os efeitos da tradução do mRNA no seu tempo de semivida e o consumo de energia em cada tecido.

Os modelos criados a partir deste trabalho podem ser integrados em avaliações ao impacto de mutações genéticas no código de indivíduos, o que por sua vez auxilia ao diagnóstico de doenças e ao desenvolvimento de medicamentos e tratamentos. Álem disso, os mecanismos biológicos sugeridos por esta investigação ampliam o nosso conhecimento em relação à biologia celular e podem ser explorados no futuro para propósitos clínicos e científicos.

Palavras Chave

Tempo de semivida do mRNA; Redes neuronais convolucionais; interpretação de modelos; componentes regulatórios da sequência do mRNA; variações do tempo de semivida do mRNA entre múltiplos tecidos humanos

Contents

List of Figures

List of Tables

Acronyms

A adenine

ATP adenosine triphosphate

CNN convolutional neural network

C cytosine

DeepLIFT Deep Learning Important FeaTures

DNA deoxyribonucleic acid

DNN deep neural network

G guanine

GSEA gene set enrichment analysis

GTP guanosine-5'-triphosphate

mRNA messenger RNA

MT-DNN multi-task deep neural network

PC principal component

PCA principal component analysis

RNA ribonucleic acid

RNA-seq RNA sequencing

T thymine

TF-MoDISco Transcription Factor Motif Discovery from Importance Scores

UTR untranslated region

TPM transcripts per million

tRNA transfer RNA

U uracil

uAUG upstream AUG

uORF upstream opening reading frame

1

Introduction

Contents

1.1 Genome

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, Cytosine, Guanine and Thymine. 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.

1.2 The central dogma of molecular biology

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. 1.1). 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).

1.3 Regulation of cellular protein levels

The need for specific levels or amounts of different proteins by the cell varies between cell type (ex. pancreas cells, muscle cells, tumor cells, etc) or cell conditions (ex. proliferation, lack of oxygen, lack of nutrients, etc). The reason for this variation stems from the fact that both different cell types and cell conditions involve their own set of cellular processes which are supported by specific proteins in specific amounts. Not having the protein amount which allow for each cell's specific set of processes will either

impair or unable proper cell functioning. On the other hand, having a higher amount of specific proteins than needed for normal cellular functioning can be toxic and resourcefully expensive to the cell [5].

1.4 mRNA degradation

We have established that the regulation of the protein levels is key for cellular functioning. One mechanism which allows this regulation is mRNA degradation.

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 from it.

An example for the regulation of protein levels from mRNA degradation is when the cell faces osmotic stress (ex. when concentration of salt surrounding the cell is high). In this situation, mRNAs which translate proteins involved in stress responses have high degradation rates. These allows for the rapid removal of those mRNAs after stress [6]. Generally, mRNAs encoding proteins involved in stress responses, which must rapidly react to environmental signals, degrade faster [7].

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

$$\text{mRNA abundance}(t) = \text{mRNA abundance}_{\text{steady state}} * e^{-\text{degradation rate} \times t}. \tag{1.1}$$

where, t stands for the time interval from the last instant with steady state mRNA abundance to the current instant [8]. 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:

$$\text{mRNA half-life} = \frac{ln(2)}{\text{degradation rate}} \tag{1.2}$$

1.5 The mRNA half-life program

The program encoded in each mRNA, defined by its sequence, not only encodes the design of a protein but also the amount of time the program is allowed to run until it finally stops - how fast the mRNA degrades, or in other terms, how long its half-life is. In fact, depending on the sequence encoded by the mRNA, its half-life can range from minutes to days [9].

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 [7] a model for mRNA half-life prediction on yeast - the 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.

1.6 Thesis contribution and scope

This thesis focuses on understanding and predicting the influence of the mRNA sequence on its half-life across human cells. It addresses the questions: How much of half-life variabilty in human cells can we predict from sequence? What are the roles of the main mRNA known sequence features such as codons and UTRs? How do this features' influence vary between human cells from different tissues? Are there novel mRNA sequence motifs?

Its contribution can be summarized in the following points:

- Development of a linear regression model explaining the quantitative influence of the most well known sequence features impacting half-life in a specific human cell line.
- Creation of a deep convolutional neural network for mRNA half-life prediction and application of the interpretation tools DeepLIFT and TF-MoDISco to evaluate the quantitative influence of each sequence position in the prediction output, which revealed possible new regulatory sequence elements.
- Development of a multi-task neural network model for the prediction of mRNA half-life variation between cells from different human tissues.
- Development of a new model-interpretation-based metric which characterizes the effect of the mRNA sequence translation in tissue specific variations of its half-life.
- Uncovering of 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 [10].

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.

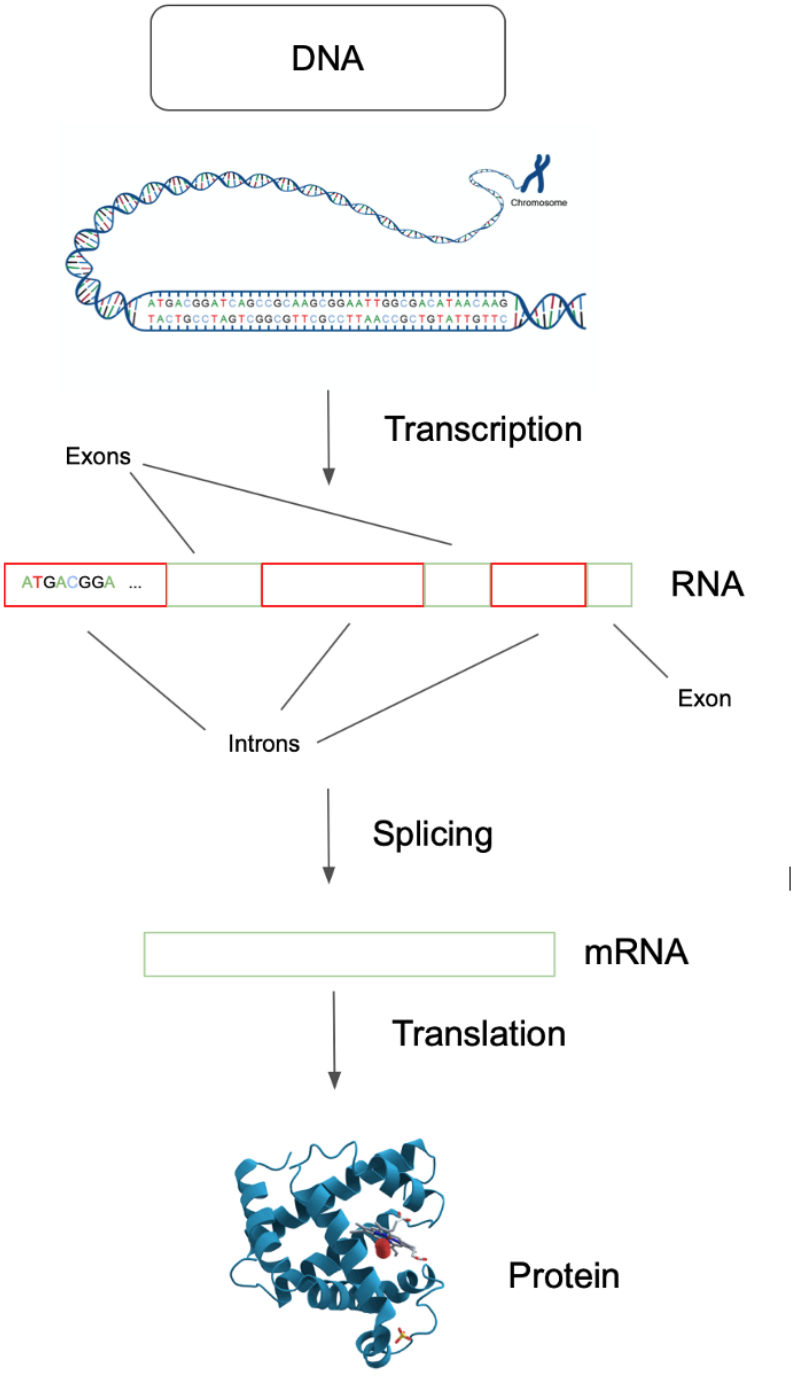


Figure 1.1: The steps from a gene to a protein. The DNA molecule is stored in the nucleus of the cell on chromosomes. A portion of this DNA molecule - gene - is transcribed into RNA, processed in a process called splicing and subsequently used as a template to form multiple equal proteins.

2

Background

Contents

2.1 Biology

2.1.1 mRNA structure

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 in the RNA along this thesis.

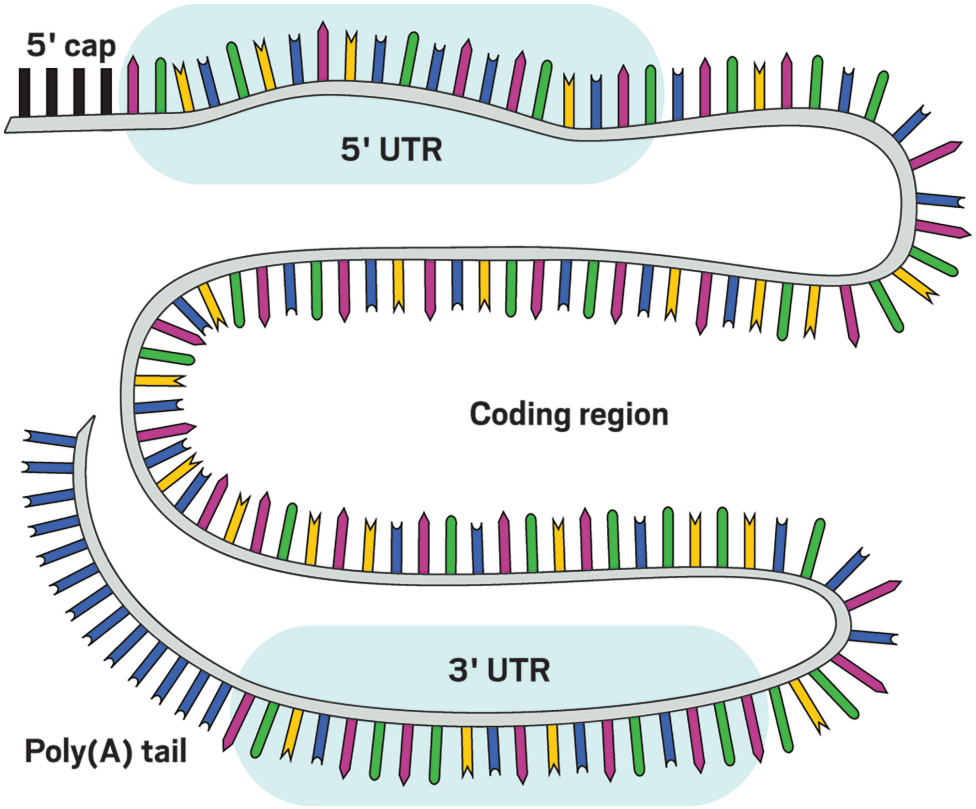


Figure 2.1: Schematic illustration of an mRNA molecule and its structure. Each blue, red, green and yellow positions in the figure represent the bases A, U, G and C respectively. [1]

2.1.1.A Coding Sequence

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 $4^3 = 64$. The only codons not encoding an amino-acid are the stop ones. 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.

2.1.1.B 5'UTR

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.

2.1.1.C 3'UTR

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). This region is involved in regulating multiple mRNA fates such as its degradation and localization in the cell's cytoplasm and can also interfere with the translation process [11]. Two main characteristics of this region allow for this regulation: its three-dimensional structure and its binding affinity with different proteins or RNAs. Some proteins or small non coding RNAs (microRNAs) bind to specific sequence portions of the mRNA called motifs. This RNA binding molecules directly influence the fate of the mRNA and the translation process. The 3'UTR can encode diverse motifs suitable for regulation of the mRNA in different cell conditions [11].

2.1.2 mRNA translation

The mRNA translation is the process of the assembly of a protein through the instructions in its sequence, in particular, the coding sequence of the mRNA. This process occurs in the cytoplasm of the cell.

2.1.2.A tRNAs

tRNAs or transfer RNAs are a specific kind of RNA molecules found in the cell's cytoplasm, which contain a sequence triplet and sometimes an amino-acid. The sequence triplet they contain is complementary to one specific mRNA codon. Therefore, this triplet is defined as anti-codon. The function of tRNAs is,

during translation, to transfer an amino-acid mapping to one specific codon. In this way tRNA molecules are the units of the dictionary which maps codons to amino-acids.

2.1.2.B Ribosome

The ribosome is a molecular machine composed of RNA molecules and proteins, responsible for the protein synthesis of every living cell. It's function is to construct a protein by linking its composing amino-acids in the order defined by the codons of the mRNA.

A ribosome can be seperated into two subunits 40s and 60s. The process of ribosome assembly starts with the binding of the 40s subunit to the 5' end of the mRNA. Then this subunit moves along the sequence in the 5'UTR searching for a suitable translation initiation site - a sequence generally composed of the start codon and a specific combination of bases surrounding it [12]. In certain mRNAs, this sequence is well studied and corresponds to the Kozak sequence. Subsequently, the 60s subunit binds and the ribosome is assembled allowing translation to begin.

The ribosome has three biding sites for the accommodation of tRNAs with a different role in translation. The aminoacyl site (A), the peptidyl site (P) and the exit site (E).

2.1.2.C Translation process

Once assembled in the begining of the coding sequence, the ribosome firstly decodes the start codon. The following steps of translation are then common to every codon.

First, the tRNA containing the anti-codon complementary to the next codon enters the **A** site of the ribosome (Fig. 2.2 tRNA with AAG anticodon). Secondly, the tRNA in the **P** site (Fig. 2.2 tRNA with ACC anticodon) leaves its amino-acid on the growing chain of amino-acids (poly-peptide) and occupies the **E** site. At the same time, the tRNA previously on the **A** site occupies the **P** site. Lastly the tRNA on the E site exits the ribosome and leaves room for the tRNA on the **P** site to enter it.

This cycle continues in the 5' to 3' direction, by decoding each subsequent adjacent triplet for one specific amino-acid and connecting each amino-acid to the previous one in the P site, forming a chain. Such cycle allows a protein to be iteratively formed. In the end, the ribosome dissociates itself from the mRNA when it reaches the stop codon. At this stage, the full chain of amino-acids is complete and the protein is therefore formed.

2.1.3 mRNA degradation

Even though mRNA degradation can be triggered by multiple events, the degradation itself generally happens in a similar way across all mRNAs. In this section we will focus on explaining the mRNA degradation process and its most frequent triggers. Such can be categorized in two main groups: triggers

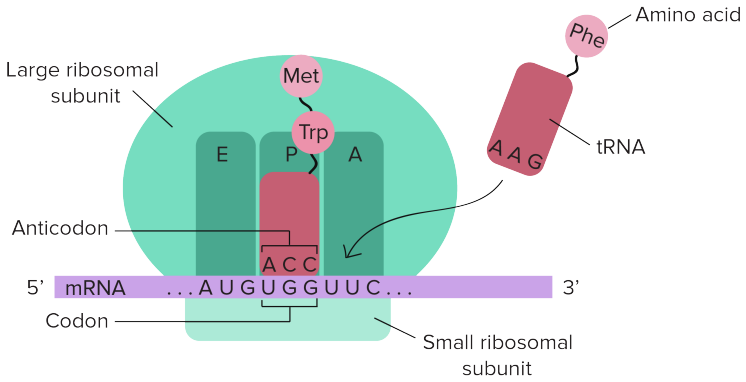


Figure 2.2: Schematic illustration of mRNA translation. [2]

from translational events and triggers from binding of functional molecules like miRNAs (micro RNAs) or proteins mainly to untranslated regions (3'UTR or 5'UTR).

2.1.3.A How mRNA degrades

The most common mRNA degradation pathway starts with the shortening of the poly-A tail, a process termed deadenylation. Afterwards either the 5'cap of the mRNA is removed, allowing the mRNA to be degraded by the molecular complex XRN1 exoribonuclease in the 5' to 3' direction, or the unprotected 3' end is exposed to the molecular complex called exosome and degraded in the 3' to 5' direction [13].

2.1.3.B miRNA and protein binding triggers

mRNA degradation can be regulated by miRNA or proteins, which bind to the mRNA at specific sequence regions (sequence motifs), or structural regions (structural motifs) [14], [15].

miRNAs or micro RNAs are RNA molecules encoding a sequence of about 20 bases long [15]. Like RNA binding proteins, they have known roles in regulating mRNA translation and degradation.

These binding regulatory elements can control mRNA degradation by repressing or promoting the assembly of molecular complexes which degrade the mRNA. One example is the Puf3 protein in yeast, which promotes the binding of molecular complexes that degrade the poly-A tail [16].

2.1.3.C Surveillance mechanisms

The surveillance mechanisms are pathways which are triggered in order to quickly degrade an mRNA which has errors in its sequence that would produce defective proteins and therefore harm the cellular environment and waste resources. There are two main surveillance pathways: Nonsense-mediated decay and Nonstop-mediated decay.

Nonsense-mediated decay (NMD) is a pathway responsible for signaling an mRNA for degradation if it has a stop codon positioned in the coding sequence before its end and ”in frame” with the start codon. For a codon to be ”in frame” it is meant that the codon is part of the set of adjacent non-overlapping triplets starting from the coding sequence.

NMD can be triggered when, for example, a start codon is recognized before the beginning of the coding sequence, in the 5’UTR. If that happens, then, in case there is a stop codon in the coding sequence or 5’UTR in frame with that start codon and no surveillance mechanism, a defective protein could be produced.

The other common surveillance mechanism, nonstop-mediated decay, surveils if an mRNA has no stop codon, preventing the ribosome from translating 3’UTR sequence.

2.1.4 Codon usage

Each mRNA is composed of a different coding sequence and therefore has its own usage pattern of codons during translation. The usage of a different codon composition is known to be associated with increased or decreased mRNA half-life. For the yeast organism, a linear model for mRNA half-life prediction from sequence explained most of the mRNA half-life variability through the mRNA’s specific codon composition, approximately 55% out of 59% explained variance obtained from a linear model with all sequence features as covariates [7].

The effect of a particular codon composition in mRNA half-life is thought to happen through the influencing of the dynamics of translating ribosomes [17]. In particular, the accepted hypothesis is that the translation elongation rate - the number of codons decoded by the ribosome per unit of time - is influenced by the codon composition of the mRNA. Subsequently, a low translation elongation causes an mRNA to degrade faster by inducing the recruitment of degradation molecular machinery.

Among other factors, the decoding time of a codon depends on two: the availability/concentration of cognate tRNAs and the codon-tRNA-ribosome interaction. In this context, the metric codon optimality was introduced in the literature as a quantity which defines how optimal a codon is, based on its cognate tRNA concentration, demand and the particular codon-tRNA binding dynamics [18]. Codon optimality is capable of capturing a codon’s decoding rate. Optimal codons - which have a high codon optimality score - should be associated with higher decoding rates compared to low non optimal codons - which should have a low optimality score. mRNAs composed of a higher amount of non optimal codons compared to optimal ones should translate slower and therefore have a lower half-life.

In fact, replacing an mRNA coding sequence with different but synonymous codons while maintaining their number, can increase or decrease half-life tenfold [19].

Recently, a causal relationship between a codon’s decoding time and mRNA half-life was found. If a codon takes a sufficiently high amount of time to decode due to a lack of an available cognate tRNA,

then the tRNA in the E site of the ribosome exits and leaves it with just one tRNA in the P site. As a consequence, the ribosome acquires a specific structural conformation which ends up recruiting the Ccr4-Not complex [20]. This complex is responsible for the deadenylation of the poly-A tail of the mRNA, which makes it vulnerable to degradation by the XRN1 exoribonuclease or the exosome.

2.1.5 Codon distribution in the genome

2.1.5.A Codon bias

Codon bias is a term for the uneven presence of synonymous codons in the genome. If the composition of synonymous codons for each mRNA was uniformly and randomly distributed, then we would expect a close to uniform distribution of the quantities of codons encoding the same amino-acid. It turns out that for every organism there is a specific codon bias - a specific uneven representation of some synonymous codons over others in the genome [21].

As presented previously, the codon content of an mRNA is a major factor for its degradation/half-life and translation elongation rate. So one may hypothesize that the codon bias can be a way of regulating protein levels by decreasing/increasing the mRNA levels and decreasing/increasing the translation elongation rate. By choosing specific synonymous codons, the overall protein will have the same amino-acid sequence, while its number can be regulated [21].

Although the significant effect of synonymous codons on mRNA half-life is acknowledged in the literature, an explanation for the codon bias in the genome remains a topic of debate with two main hypothesis proposed.

One argues that the uneven presence of synonymous codons in the genome is a product of natural selection to increase the fitness of an organism [22]. By enabling gene expression regulation through the ability of synonymous codons the organism would accumulate an advantage.

The other hypothesis argues that even though synonymous codons have a significant influence on mRNA half-life, there was no natural selection for specific synonymous codons. It argues that codon bias is the product of biases not related to translation of mRNAs, but to mRNA-independent factors such as the structural organization and replication of the genomic code [23]. Following this hypothesis, it was proposed that the codon bias could be the product of a genome-wide inherit bias towards guanine and cytosine bases, measured by GC content - the ratio of guanine or cytosine bases in the genome. High ratios of guanine or cytosine confer more stability to the DNA molecule.

2.1.5.B Functional-related genes

For a cellular function to be performed there is often a need for a unique set of proteins to be involved and interact. These proteins are termed functional-related and the genes that encode them are defined

as functional-related genes.

Functional-related genes often have mRNAs with similar codon usage patterns and similar half-lives [19]. This allows protein expression levels to be coordinated in a compound way. One example of this is the clock genes, which are responsible for the control of circadian rhythms in numerous life forms, including cyanobacteria and Neurospora crassa. In both organisms, the clock genes are lacking in optimal codons, and their non-optimality is key for their ability to drive the transient response which characterizes circadian rhythms [24] [25].

2.1.5.C Different cell states and conditions

We have established that the codon content of functional-related sets of genes can drive their coordinated expression. However, cells are not static entities, they differentiate into tissues, they proliferate and they face different environmental conditions. All of these cell states and conditions require the coordinated expression of the same groups of genes in different amounts. As codon content is already defined in the sequences of the genes and is therefore immutable to the cell state or condition, then one could consider that the effect of codon content is only a baseline regulatory mechanism, not influencing the specific needs of cells during their life-time.

Despite codon bias being a fixed property, the condition or cell-state-specific environment is able to tune the impact of a certain codon in mRNA half-life or the translation rate.

In fact, a study performed on mouse embryonic fibroblast cells, revealed that mRNAs whose levels increase during cell proliferation are enriched in rare codons - codons with an overall low presence in the coding sequence of expressed mRNAs - and undergo a higher translation boost than transcripts with common codons [26]. Other study linked changes in the concentration of specific tRNAs to the increase of expression of specific mRNAs during cell proliferation [27].

Although it has been pointed out that the regulation of specific tRNAs can drive mRNA translation, it is still not clear the magnitude of the impact of this regulation neither how and in what specific conditions or cell types it happens. Furthermore, factors other than tRNA regulation can possibly drive the condition or cell-state specific effect of codon usage. These factors remain, to my knowledge, largely unexplored and unknown.

2.1.6 High throughput sequencing methods

In the context of DNA and RNA, sequencing refers to the process of determining the sequence of nucleic acids or bases that compose a DNA or RNA molecule.

High throughtput sequencing encompasses a class of sequencing methods capable of sequencing high amounts of RNA or DNA molecules in parallel.

2.1.6.A RNA-seq

RNA-seq (RNA sequencing) is a high throughtput sequencing technique used to quantify and sequence the transcriptome - meaning all RNAs present in a single cell or population of cells.

In a first step, the RNA molecules are fragmented into smaller portions, therefore containing smaller sequence intervals. Secondly, these fragmented RNAs are copied into cDNA molecules, which are more stable than RNA fragments. Lastly these fragments are sequenced into reads which are then mapped to the position in the genome they came from.

2.1.7 Exonic and intronic reads as a proxy for half-life variations

The reads obtained through RNA-seq can map either to intronic or exonic sequences. As intronic reads correspond to RNAs which were not processed (spliced), it can be said that these reads map to RNAs still in the nucleus, in the specific case of the RNA encodes for a protein, these RNAs are called pre-mRNAs. Exonic reads can both belong to pre-mRNAs or mRNAs.

In [28], it was shown that under some assumptions such as steady state mRNA levels and spliced introns short life-time:

$$\Delta \log_2(m) = \Delta \log_2(\beta) - \Delta \log_2(\alpha) \tag{2.1}$$

where m is the amount of an mRNA, β is the transcription rate and α the degradation rate. Δ stands for variation between two different conditions such as two different tissues or cell states.

Furthermore, it was shown that $\Delta \log_2(\beta)$ can be approximated by Δintron and $\Delta \log_2(m)$ by Δexon, where Δintron and Δexon refer to variations in exon and intron amounts between conditions. Lastly, as degradation rate, α, is inversely proportional to half-life, it follows that:

$$\Delta \log_2\left(\mathrm{t}_{1/2}\right) = \Delta\text{exon} - \Delta\text{intron} \tag{2.2}$$

where, $\Delta \log_2(\mathrm{t}_{1/2})$, stands for the difference in log2 half-life between two different conditions. The mRNA's amount of exons and introns can be approximated by the RNA-seq exonic and intronic read's amount respectively.

2.2 Modeling and Analysis

2.2.1 Supervised Learning

Supervised learning corresponds to the task of creating a parametrized function which maps an input to an output from a set of known input-output pairs. The set of input-output pairs is also called a dataset,

and each pair can be defined as a tuple (X_i, Y_i), where X_i is an input and Y_i is an output. Both X_i and Y_i are tensors which can be of any shape, usually with real numbers.

The parametrized function mapping inputs to outputs can be defined as $f(X_i; \mathbf{w}) = Y_i$, where **w** is the function's parameters which are in the form of a tensor that can be of any shape. The parameters of this function are obtained from a process called learning. This process corresponds to a search on the function parameters' space which aims to select the parameters which provide the best X_i to Y_i mapping of a subset of the dataset termed test set using a different subset of the dataset termed training set. The quality of the mapping is defined based on a selected evaluation metric, which compares the target output Y_i with the one provided by the function $\hat{Y}$ - prediction output.

This evaluation metric is in the form of a loss function $L(Y, \hat{Y})$, which measures how close the prediction variable $\hat{Y}$ and the target variable Y are. The smaller the loss function output the better.

The aim of supervised learning is to minimize an objective function J on the training set

$$J(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i^{\text{train}}, f\left(\mathbf{x}_i^{\text{train}}; \mathbf{w}\right)\right) \tag{2.3}$$

such that the learned parametrized function performs best on the test set, where the performance is measured by an evaluation function.

2.2.2 Linear regression

Linear regression is in the class of supervised learning methods. Here the parametrized function is defined as:

$$f(\mathbf{x}; \mathbf{w}) = w_1 x_1 + \ldots + w_D x_D \tag{2.4}$$

The loss function is the squared error: $L(y, f(\mathbf{x}; \mathbf{w})) = (y - f(\mathbf{x}; \mathbf{w}))^2$. We define the input vector of a sample i as $\mathbf{x}_i^T$ and it's target y_i, where $\mathbf{x}_i^T \in R^D$, $y_i \in R$ and T is the transpose operator.

If we define X as the matrix containing all training set samples:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1D} \\ \vdots & \ddots & \vdots \\ x_{N1} & \ldots & x_{ND} \end{pmatrix} = (\mathbf{x}_1, \ldots, \mathbf{x}_D) \in R^{N \times D} \tag{2.5}$$

then, if there exists an inverse $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$, the parameters of the learned function $\mathbf{w} \in R^D$ can be analytically obtained by:

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y} \tag{2.6}$$

2.2.3 Regularization

As discussed before, the aim of the learned parametrized function in supervised learning is to have the best possible performance on the test set, meaning that the aim is to have a function which can best generalize to previously unseen sets of data. The objective function J can be minimized in a way that the learned function performs excellently in the training set and poorly in the test set - if this happens it is said that the function overfited the training set. This means that, instead of learning general rules which direct the prediction of the target variable, the learned function has memorized input to output mappings.

Regularization is a way of dealing with overfitting and increasing the generalization power of the learned function. One way regularization can be implemented is by introducing "penalization" terms in the loss function, which direct the optimization to produce weights with lower magnitude.

In linear regression, when the penalization term introduced is $\alpha|\mathbf{w}|_2$, where $|\mathbf{w}|_2$ is the L2 norm of the parameters **w** and $\alpha > 0$, the loss function changes to the following:

$$L(y, f(\mathbf{x}; \mathbf{w}), \mathbf{w}) = (y - f(\mathbf{x}; \mathbf{w}))^2 + \alpha|\mathbf{w}|_2^2 \tag{2.7}$$

A linear regression subject to this type of regularization is also called Ridge regression. The coefficient α determines the strength of the regularization. This type of regression forces the parameters **w** to be closer to 0. The objective function $J(\mathbf{w})$ is still convex and therefore easy to minimize. The parameters **w** can be obtained by:

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.8}$$

where T denotes the matrix transpose operation.

2.2.4 Hyperparameter optimization

Parameters of the supervised learning predictive function, the optimization procedure or loss function like the regularization parameter α from section 2.2.3, that are not optimized during the optimization of the objective function on the training set (model training stage), are termed hyperparameters. Hyperparameters are optimized to provide the best evaluation metric over a different set, usually termed validation set, which is independent from the test set.

One common approach to optimize hyperparameters is selecting a bounded hyperparameter space and sample randomly from it until the desired evaluation metric on the validation set is achieved. Such approach is also termed random search [29].

Another approach, Bayesian hyperparameter optimization, rather than searching randomly over the hyperparameter space, relies on the construction of a probabilistic model of the value of the objec-

tive function conditioned on the hyperparameters. The selection of the hyperparameters to search is informed by this probabilistic model [30].

2.2.5 Deep Learning

Deep learning is a class of machine learning algorithms characterized by progressively applying multiple transformations to the input data in order to extract higher level features.

In the context of deep learning in supervised learning, the parametrized function mapping inputs to outputs $f(X_i; \mathbf{w}) = Y_i$ (see section 2.2.1) is composed of n functions $f_1, f_2, ..., f_n$ also called layers, such that $f(X_i; \mathbf{w}) = f_n(f_{n-1}(...(f_1(X_i; \mathbf{w}_1))...; \mathbf{w}_{n-1}), \mathbf{w}_n)$, where $\mathbf{w}_1, ...; \mathbf{w}_n$ comprise the parameters $\mathbf{w}$.

2.2.5.A Fully connected deep neural network

Fully connected deep neural networks are a class of deep learning models composed of fully connected layers. A fully connected layer can be defined as a function $f(X; \mathbf{w})$ such that:

$$f(X; \mathbf{w}) = \sigma(\mathbf{w}X^T) \tag{2.9}$$

where $\mathbf{w} \in R^{H \times D}$, $X \in R^{N \times D}$ is a matrix containing N training set samples and σ is a non-linear activation function. ReLU (rectified linear unit) is a frequently used activation function, defined as $\sigma(x) = \max(0, x)$.

2.2.5.B Convolutional Neural Networks

Convolutional neural networks (CNNs) are a class of deep learning models which rely on the convolution operation of parametrized learnable filters on the input data. Usually, such data is either 2 or 3-dimensional (ex. images), or 1-dimensional like in RNA or DNA sequences. In this section we will focus on convolutional neural networks applied to RNA or DNA sequences [31].

The mRNA molecule can be represented as a sequence of length L, where each position is one of A, C, G or T bases. The most common way of representing a DNA-like sequence is through the method of one-hot encoding. One-hot encoding the mRNA sequence means representing it as a matrix of dimension $4 \times L$, where each column represents a sequence position and each row an mRNA base. Each column contains a value of 1 for one of the rows, corresponding to the base present on that position, and all remaining rows contain a 0 (Fig. 2.3).

A convolutional neural network is composed of convolutional layers. A convolutional layer contains n filters with specific dimensions. In the context of one-hot encoded mRNA sequences, the first layer comprises n filters of dimension $4 \times l$ where l is the filter's length and each filter, i, is parametrized by its own set of weights, w_i. The filters convolve with a set of l contiguous positions in the mRNA

sequence, producing a scalar value, also called activation. The convolution can be on every position of the sequence, which produces $L - l + 1$ scalar values or activation, or spaced by a number of elements s, termed stride. Often, the sequence is concatenated to its end $l - 1$ elements containing zeros so that the number of values produced by the convolution is equal to the length of its input. Such concatenation is also termed zero-padding.

The vector of produced scalars for one filter is termed channel. Every value of each channel is then transformed through a non-linear function (typically ReLU). The set of n channels correspond to the output of the convolutional layer. A second convolutional layer would input the set of n channels from the first layer and perform convolutions with a new set of parametrized filters of dimension $n \times h$, where h is the filter's size, which can be different from the first layer's filter size l. The number of convolutional layers is arbitrary, taking into account that each new convolutional layer can be applied on the output of the previous one.

In the context of convolutional neural networks, a common operation performed over each convolutional layer channel is pooling. Pooling consists of aggregating the values of contiguous activations in the channels by usually computing their mean or maximum value (max pooling) along the position dimension. When the aggregation is made on the whole channel this operation is called global pooling.

Once the desired number of convolutional layers and pooling operations have been performed, the output of the final layer/operation can be flattened to a d-dimensional vector and serve as the input of a fully connected layer which returns the predictions of the convolutional neural network model.

2.2.5.C Multi-task Neural Networks

Multi-task neural networks are a class of deep learning models which aim to predict multiple different tasks simultaneously given some input data and using neural-network-based models (ex. fully connected deep neural networks, CNNs, ...). Consider the example of predicting variations in mRNA half-life on multiple human tissues (prediction tasks). The mRNA features that can eventually be useful for predicting one task (tissue) can also be for a different one. In this way, a multi-task neural network allows for the leveraging of possible interrelationships between tasks to increase the predictive power of the final model.

2.2.6 Deep Learning Model interpretation techniques

Deep learning models, are frequently criticized for being difficult to interpret, which has proven to be an obstacle for its adoption in some fields.

Although their interpretation may not be as straightforward as other supervised learning methods, it is still possible to interpret it to find out information on the model's inference.

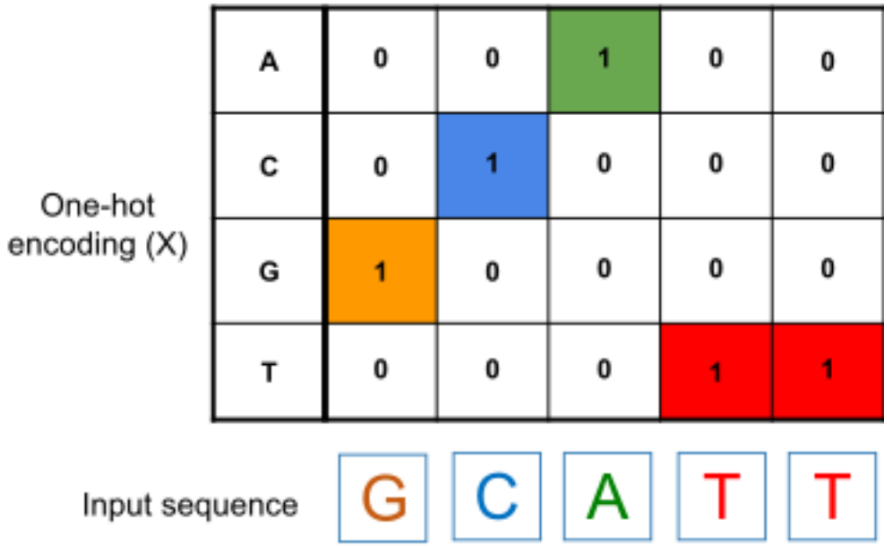


Figure 2.3: One hot encoding illustration of an mRNA sequence of length 5.

2.2.6.A Convolutional filters

When the deep learning model is a CNN, the convolutional filters contain information on what pattern is being matched or extracted from the convoluted sequence. Therefore, the weights of each CNN filter can be used to inspect what sequence or image patterns the model finds relevant to predict the target variable.

The main limitation of this approach is failing to capture the distributed nature of the representations learned by the CNN. By distributed nature, it is meant the fact that the search for a specific pattern on the sequence or image may be the result of a compound search distributed by multiple filters. If the pattern search is distributed, looking at an individual filter's weights would not be insightful.

2.2.6.B Backpropagation-based approaches

Backpropagation-based approaches are a class of interpretation methods which can overcome the distributed learned representations limitation of the convolutional filters.

In this class of methods, an importance score is back-propagated from each output neuron through the models' layers until each input sequence position.

One method belonging to backpropagation-based approaches is using the gradient of the output with respect to each input. In [32], the authors used this approach in order to compute a "saliency map" of an image in the context of image classification tasks, by computing the gradient of the classification prediction output with respect to each pixel.

Although this approach is able to capture the contribution of distributed representations, it fails in

mainly two aspects: neuron saturation and discontinuous gradients.

Figure 2.4 illustrates the neuron saturation problem. For the output y, its gradient with respect to i_1 and i_2 is 0 if $y = 1$, or if $i_1 + i_2 \geq 1$. Therefore using the gradient in this case would not flag either i_1 or i_2 as relevant to the output y.

Figure 2.5 illustrates the discontinuous gradient problem. In case the gradient has a discontinuity at some point of its domain, in this example at $x = 10$, then an infinitesimal change near its discontinuity will produce a strikingly different value leading to numerically unstable and misleading results.

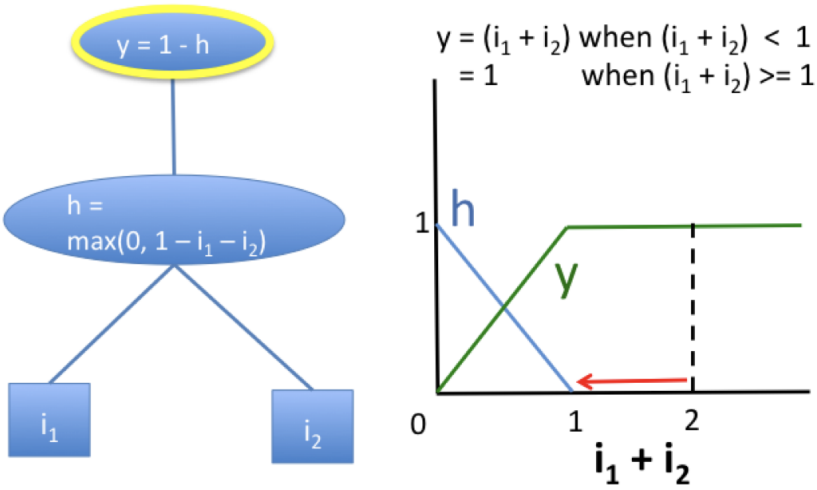


Figure 2.4: Example of a neuron saturation problem. Figure taken from [3].

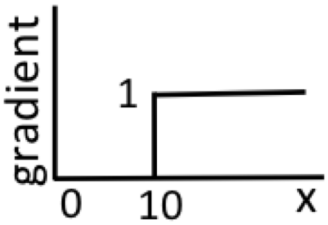


Figure 2.5: Example of a discontinuous gradient problem. Figure taken from [3].

2.2.6.C DeepLIFT

DeepLIFT [3] was one of the algorithms belonging to the class of backpropagation based approaches introduced to solve the neuron saturation and discontinuous gradient problems. It relies on explaining a difference from reference prediction, Δt, as the sum of contribution scores, $C_{\Delta x_i, \Delta t}$, for each difference from reference input $\Delta x_i = x_i - x_{\text{reference}\,i}$.

$$\Delta t = f(\mathbf{x}) - f(\mathbf{x}_{\text{reference}}) = \sum_i C_{\Delta x_i, \Delta t} \tag{2.10}$$

DeepLIFT defines the contribution scores, $C_{\Delta x_i, \Delta t}$, for a range of neural network layers. For example, for the fully connected layer, $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, the contribution scores are defined as:

$$C_{\Delta x_i, t} = w_i \left(x_i - r_i\right) \tag{2.11}$$

The contribution scores of a layer composed of other layer, $f^2(f^1(x)))$ can be computed using the chain rule. If we define a multiplier $m_{\Delta x, \Delta t}$ as:

$$m_{\Delta x, \Delta t} = \frac{C_{\Delta x, \Delta t}}{\Delta x} \tag{2.12}$$

Then, the multiplier $m_{\Delta x_i \Delta t}$ is given by:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i, \Delta y_j} m_{\Delta y_j, \Delta t} \tag{2.13}$$

where x_i is one neuron of the input layer, $y_1, y_2, ..., y_j$ are the outputs of the first layer and t a single target output.

The fact that the contribution scores can be computed by backpropagating the multipliers adds to the efficiency of this algorithm.

The problem of defining an input reference is domain-specific. When dealing with mRNA sequences as input, the reference input can be a randomly shuffled mRNA sequence.

2.2.6.D TF-MoDISco

A common focus when studying mRNA is to find specific sites in the sequence where regulatory elements such as proteins or miRNAs bind, or structural elements reside. This sites are usually defined as a set of contiguous bases, each one having a stronger or weaker contribution for the binding of the regulatory element. The particular combination of bases with their heterogeneous contributions forming the binding site is termed (sequence) motif. A specific motif is usually found on multiple mRNA sequences.

After having each mRNA sequence's importance scores calculated by a gradient-based approach or DeepLIFT, inside each sequence, several close-to-contiguous portions will have a distinct importance score pattern - usually either a high negative or positive contribution score. These close-to-contiguous portions can be considered possible motif candidates. If a motif candidate is spotted in several mRNA sequences' with similar high magnitude contribution scores, then we can assume that candidate was captured by the model as an important sequence element predictor, indicating that the captured motif

candidate can be a sequence with biological relevance.

Given the importance scores of thousands of sequences (see Fig. 2.6), finding candidate motifs shared across several mRNAs is a hard problem to be done manually. Furthermore, the motif candidate found in one sequence can have a similar but not exactly equal set of contribution scores to other sequences but still represent the same motif.

For this reason TF-MOdisco was developed. TF-MOdisco is a tool to extract sequence patterns, or motifs, from a set of sequence's importance scores [33]. This importance scores can be calculated using, for example, a gradient-based approach or DeepLIFT.

Overall, the TF-MoDISco algorithm can be described in 3 steps. In the first step, a sliding window with size n and stride 1 is passed through each sequence's contribution scores, and all contributions inside that window are summed. Afterwards, the windows with highest sum are selected and filtered using a null distribution computed as the distribution of the sliding window sums of randomly shuffled sequences. These selected windows are also termed seqlets. Seqlets are then flanked by k amount of bases on both sides and further filtered in a way that any pair of seqlets doesn't overlap more than 50% (when such happens the seqlet with highest contribution sum is selected).

In the second step, a similarity metric between all pairs of seqlets is calculated. For each pair, this metric is computed by trying out different alignments between 2 seqlets and choosing the most similar one. Then, the seqlets are clustered using the computed pair-wise similarity metrics. The seqlets within a cluster are then aligned and aggregated into a new seqlet which covers all the positions inside the aligned ones and the positions outside the aligned but inside the clustered seqlets (Fig. 2.7). The final result is a set of candidate motifs.

In the final step, the motifs follow a postprocessing procedure. Motifs belonging to a cluster with less than l seqlets are merged into motifs with clusters with more seqlets. Furthermore, motifs apparently consisting of 2 separate ones are splitted and similar motifs are merged.

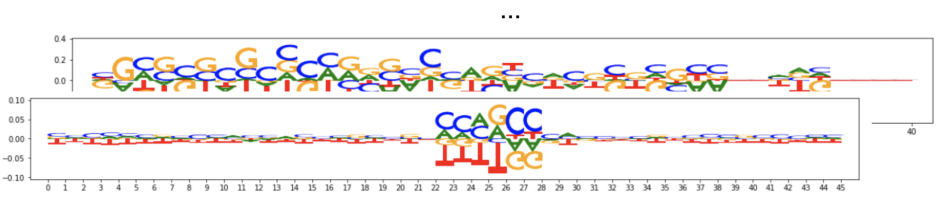


Figure 2.6: Figure illustrating the set of input contribution scores to the TF-MoDISco algorithm computed using DeepLIFT. Each position of the sequence has a score for each possible base (A, T, G or C).

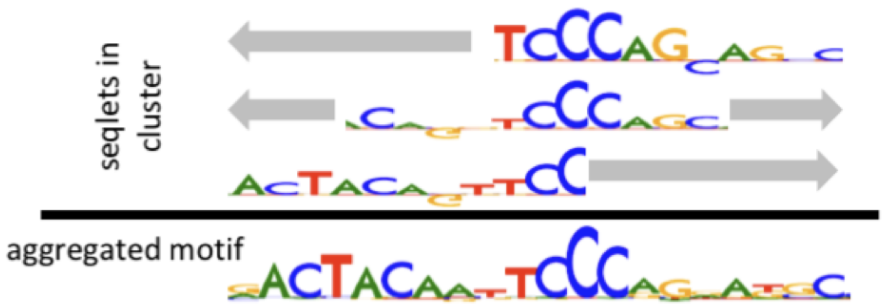


Figure 2.7: Seqlet alignment and aggregation.

2.2.7 Evaluation metrics and statistical tests

2.2.7.A Explained variance

Explained variance is defined as:

$$\text{explained variance } (y, \hat{y}) = 1 - \frac{\text{Var}\{y - \hat{y}\}}{\text{Var}\{y\}} \tag{2.14}$$

.

It measures the proportion of the variance of a given dataset, y, captured by a mathematical model whose output is $\hat{y}$.

2.2.7.B Pearson correlation coefficient

The Pearson correlation coefficient measures the linear correlation between 2 random variables as a number between -1 and 1. A value of 1 or -1 indicates a complete positive or negative linear correlation respectively. 0 indicates no linear correlation.

For 2 random variables, X and Y the pearson correlation coefficient $\rho_{X,Y}$ is calculated as:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2.15}$$

where, $\text{cov}(X, Y)$ is the covariance between X and Y and σ_X and σ_Y are the standard deviation of X and Y respectively.

2.2.7.C Spearman's correlation coefficient

The Spearman correlation coefficient measures the extent to each two variables X and Y have a monotonic relationship. It evaluates the strength and direction of this relationship and is defined as the Pearson correlation coefficient between the ranks of X and Y. The Spearman correlation coefficient, r, can be computed by:

$$r = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}\left(\mathrm{rg}_X, \mathrm{rg}_Y\right)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}} \tag{2.16}$$

where, rg_X and rg_Y are the ranks of X and Y respectively, $\mathrm{cov}\left(\mathrm{rg}_X, \mathrm{rg}_Y\right)$ is the covariance between rg_X and rg_Y and σ_{rg_X} and σ_{rg_Y} are the standard deviation of rg_X and rg_Y respectively.

2.2.7.D Wilcoxon rank sum test

The Wilcoxon rank sum test is a non parametric test aimed at evaluating significant differences between observations from two populations X and Y. It assumes that all observations are independent of each other and all responses are ordinal.

Under the null hypothesis, H_0:

$$P(X > Y) = P(Y > X) \tag{2.17}$$

To performe this test, all observed values x_i, y_i are ranked and a metric U_x is defined as:

$$U_x = R_x - \frac{n_x(n_x + 1)}{2} \tag{2.18}$$

where R_x is the sum of the ranks of all x_i and n_x is the number of observations of X. Analogously to U_x, a U_y metric can also be defined in the same way. Then, the Mann-Whitney U statistic is calculated, defined as:

$$U = min(U_x, U_y) \tag{2.19}$$

In a Wilcoxon rank sum test, it has to be determined whether the observed U statistic supports the null hypothesis. It can be shown that, under the null hypothesis:

$$E(U) = \frac{n_x n_y}{2} \tag{2.20}$$

and:

$$Var(U) = \frac{(n_x n_y)(n_x + n_y + 1)}{12} \tag{2.21}$$

For a large amount of observations, the distribution of the Mann-Whitney U statistic under the null hypothesis can be approximated by a Gaussian distribution.

2.2.8 Principal Component Analysis

Principal component analysis (PCA) is a method frequently used to project a high-dimensional set of points to a lower dimensional space, while preserving the highest possible amount of variance. It is often used as a way of visualizing a dataset or selecting lower dimensional uncorrelated (orthogonal)

features without loosing too much information [34].

In order to find the desired low dimensional space, the PCA method starts by defining a new set of n orthogonal axes or principal components - the same number of axes as the starting n-dimensional space. Such axes are constructed in a way that the highest amount of variance in the dataset is captured by one axis (first principal component), the second highest amount of variance in the dataset captured by other axis (second principal component), and so on, until the axis explaining the least amount of variance from all axis (last principal component).

Finding the principal components is equivalent to finding the eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$, where X is a matrix containing the set of n-dimensional points (one for each row). The eigenvalues of the covariance matrix provide the explained variance of the dataset by each corresponding principal component. The lower k-dimensional space chosen to project the n-dimensional points is selected by picking the k first principal components.

3

Materials and Methods

Contents

3.1 Modeling mRNA in a human cell-line

3.1.1 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 [35].

3.1.2 Distribution of half-life measurements

The log10 transformed half-life measurements follow a distribution approximately symmetric to the median (Fig. 3.1). 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 and its mRNA corresponds to the gene PPDPF - pancreatic progenitor cell differentiation and proliferation factor. The minimum half-life is 3 minutes and 45 seconds and belongs to an mRNA belonging to the gene SLC22A13, which encodes a protein involved in the transport of small molecules.

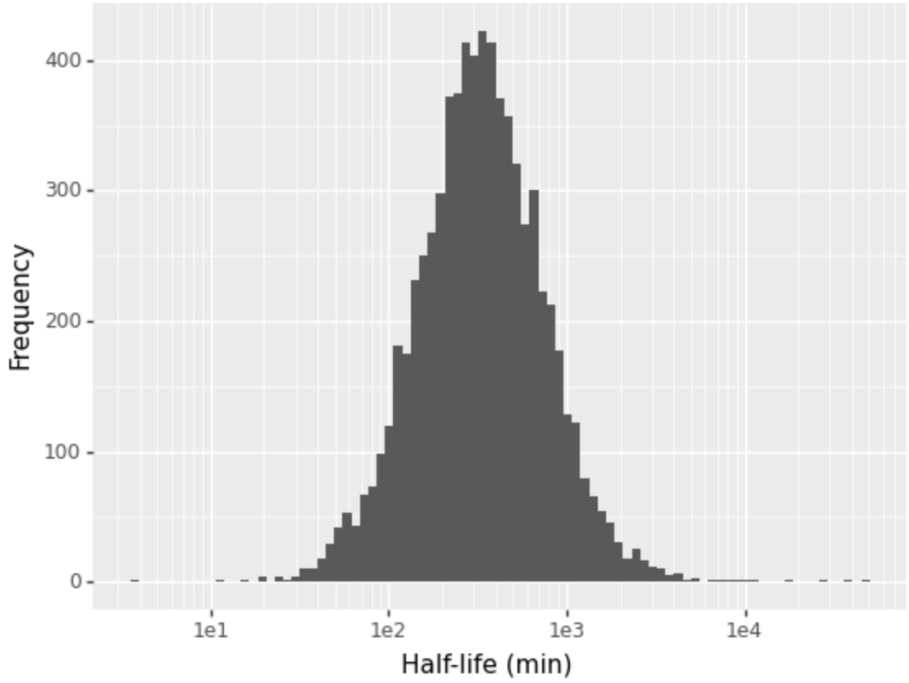


Figure 3.1: Histogram of half-life measurements. The scale of the half-life axis was log10 transformed.

3.1.3 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 (which is equal to the length of the coding sequence divided by 3).

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.

3.1.3.A 5'UTR length distribution

The length distribution of the 5'UTR has a median of 158 bases and a standard deviation of 219 bases. 75 % of the length values are bellow 270 bases and 95% below 562 bases. The minimum length is 1 base (ex. transcript ANAPC7-201) and the maximum 3693 bases (transcript INO80-210).

3.1.3.B 3'UTR length distribution

The length distribution of the 3'UTR has a median of 645 bases and a standard deviation of 1594 bases. 75 % of the length values are bellow 1607 bases and 95% below 4178 bases. There are mRNAs with no 3'UTR (length=0, ex. transcript POM121C-206) and with 3'UTR with 1 base only (ex. transcript DPF3-210). The maximum length is 25999 bases (transcript ENST00000369851.5 from gene GNAI3).

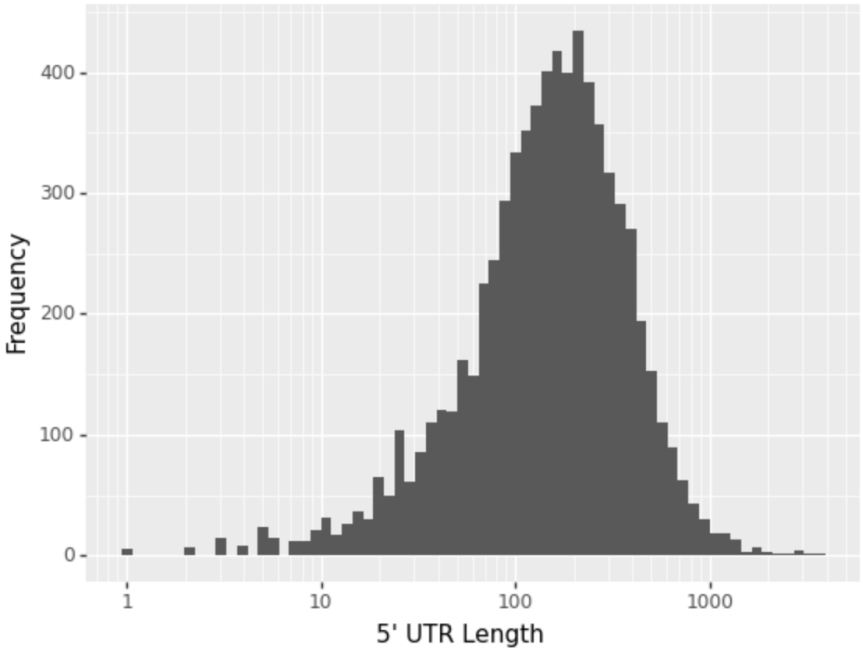


Figure 3.2: Histogram of the length of the 5'UTR of every mRNA. The scale of the length axis was log10 transformed.

3.1.3.C Coding sequence (CDS) length distribution

The length distribution of the coding sequence has a median of 1158 bases and a standard deviation of 1502 bases. 75 % of the length values are bellow 1971 bases and 95% below 4164 bases. The minimum length is 9 bases (transcript PEX2-203) and the maximum 17670 bases (transcript AHNAK-202).

3.1.4 Assessing statistical significance of correlations

The p-value of the Pearson correlation coefficient between 2 variables x and y, was calculated under the assumption that x and y are drawn from independent normal distributions and is a two-sided p-value.

The p-value of the Spearman correlation coefficient is a two-sided p-value. The SciPy python package documentation gives further details on the p-value computation.

Both Pearson and Spearman correlation and their p-values were computed using the SciPy package.

3.1.5 Modeling

3.1.5.A 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

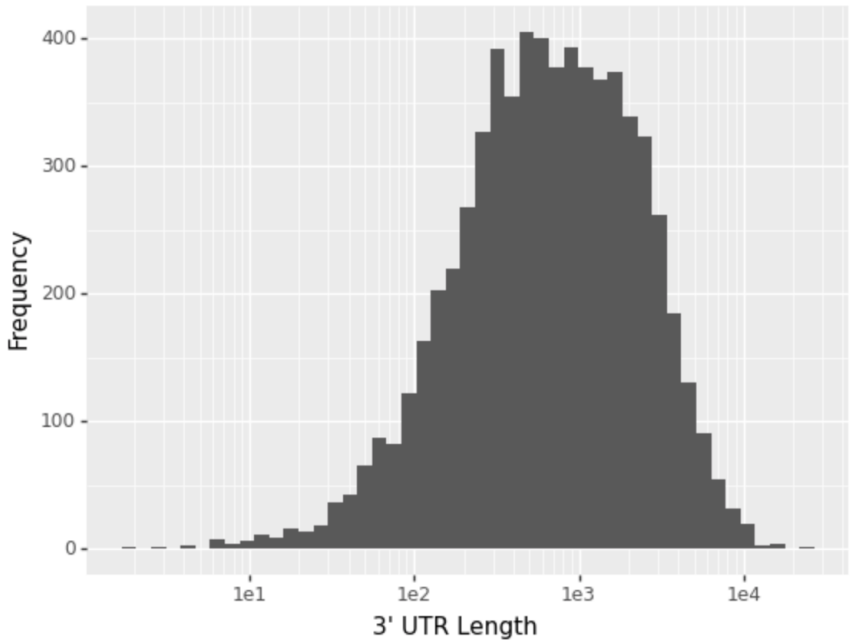


Figure 3.3: Histogram of the length of the 3'UTR of every mRNA. The scale of the length axis is log10 transformed.

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.

3.1.6 Convolutional neural network models

Two convolutional neural networks were trained, both for predicting log2(half-life). One network was trained using as input one-hot encoded sequence from the 5'UTR and the other was trained using as input one-hot encoded sequence from the 3'UTR as input.

The data was split into 3 sets - train, validation and test. After hyperparameter optimization, the models 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) were 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). The total amount of mRNAs for the 5'UTR and 3'UTR datasets was 7135 and 6601 respectively.

Both models where optimized with the mean squared error as loss function and evaluator of the performance of the validation set. Each convolutional neural network was implemented with the python package keras and optimized using the Adam optimizer with a learning rate of 1e-4.

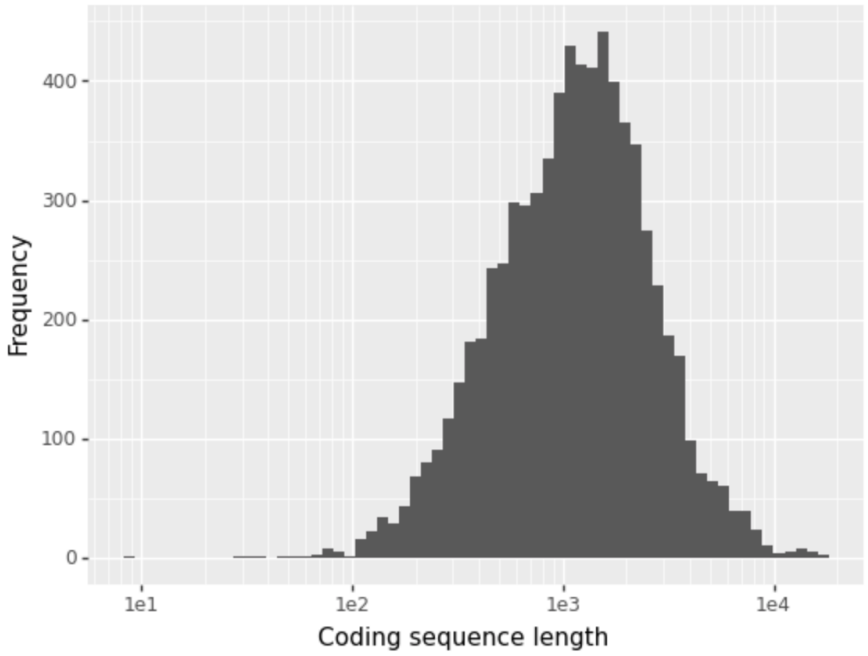


Figure 3.4: Histogram of the length of the coding sequence of every mRNA. The scale of the length axis is log 10 transformed.

3.1.6.A 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, 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

3.1.7 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.

3.1.8 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
```

3.2 Modelling the variation of mRNA half-life across human tissues

3.2.1 Data source and brief description

The used dataset comes from the Genotype-Tissue Expression (GTEx) project version 7. It comprises 11688 RNA-seq samples from 714 individuals on 53 different tissue types. These samples were collected after the death of the individual.

3.2.2 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 [28], 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:

$$\log_2(\text{exon}) - \log_2(\text{intron}) = \log_2\left(\frac{\text{exon}}{\text{intron}}\right) \tag{3.1}$$

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 a rearranging of equation 2.2:

$$\Delta\log_2(\text{half-life}) = \log_2\left(\frac{\text{exon}}{\text{intron}}\right)_{s_1} - \log_2\left(\frac{\text{exon}}{\text{intron}}\right)_{s_2} \tag{3.2}$$

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.

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

3.2.3 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).

3.2.4 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 3.1.3.

3.2.5 mRNA half-life variations distribution

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 (see 3.2.2). The tissue-specific distribution of mRNA half-life variation is shown in figure 3.5. There are on average 9798 values per tissue and on total 18551 mRNAs. Of note that the number of mRNAs is bigger than the average mRNA values per tissue because some mRNAs are only expressed in some tissues.

3.2.6 Multi-task DNN Model

A multi-task deep neural network model based on a fully connected deep neural network architecture 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 found to be relevant for half-life prediction in section 4.1, 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 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.

3.2.6.A Hyperparameter optimization

The hyperparameters batch size and model architecture parameters 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 model architecture parameters comprised the number of dense layers and the size of all dense layers.

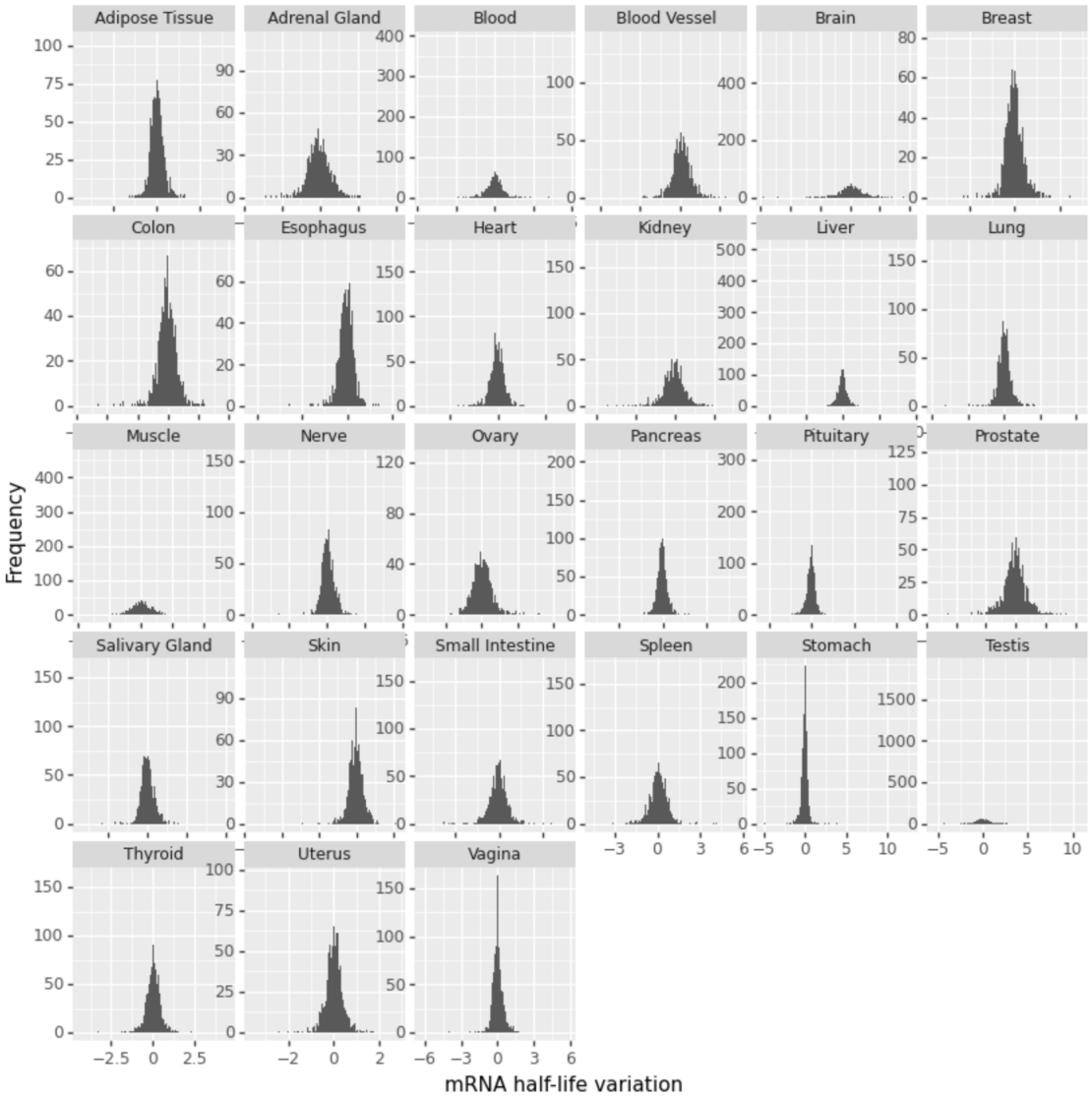


Figure 3.5: Per tissue histogram of mRNA half-life variations. The y axis scale is specific to each tissue.

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). 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.

3.3 A tissue-specific codon effect program

3.3.1 Data source

The used dataset corresponds to the mRNA half-life variations from section 3.2.

3.3.2 Linear regression model

The linear model used was a Ridge regression with regularization strength $\alpha = 0.01$. Its implementation followed the same characteristics as 3.1.5.A.

3.3.3 Tissue-specific gene's transcripts amounts

The gene's transcript amounts were picked as the sum of the TPMs of all transcripts belonging to the gene. Such TPM values are available at the GTEx website (GTEx version 7).

3.3.4 Galactose and glucose samples

The RNA-seq samples of an individual's cell's cultured in glucose and galactose mediums were provided by Holger Prokisch's Lab.

3.3.5 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) is a widely used algorithm to detect if an a priori known set of genes shows statistically significant changes, or enrichment, between two conditions [4]. In particular, when using a ranked list of genes, it determines whether a priori known sets of genes show statistically significant enrichment at either end of the ranking. A statistically significant enrichment for a set of genes indicates that the biological pathway composed of that set correlates with the ranking of the supplied list of genes.

This algorithm was ran using pyGSEA python package on a list of 10000 genes.

4

Results

Contents

4.1 Modeling mRNA in a human cell-line

4.1.1 Analysis and visualization of associations between mRNA half-life and sequence features

4.1.1.A uAUG

As discussed in section 2.1.3.C, uAUGs can have a negative impact on half-life. The boxplot in 4.1 indicates that having more uAUGs is associated with lower half-lives. In fact, having one or more uAUGs is associated with having, on median, an half-life 19% lower (Wilcoxon ranksum test with p-value = 1.36e-25). Having 4 or more uAUGs associates with having, on median, an half-life 27% lower (Wilcoxon ranksum test with p-value = 1.16e-20).

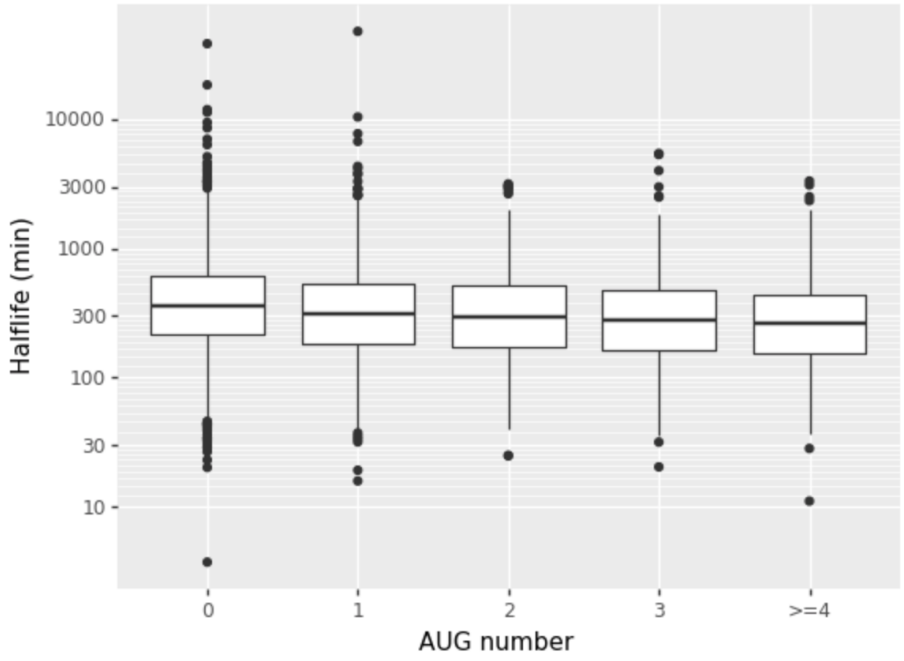


Figure 4.1: Boxplot depicting the distribution of half-life for mRNAs with 0,1,2,3 and 4 or more uAUGs.

When comparing the half-life distributions sorted by the number of uAUGs in frame with the coding sequence (Fig. 4.2) to the previous approach (Fig. 4.1) the half-life distribution corresponding to 2 uAUGs has higher half-lives on median (median fold change from the 0 uAUGs distribution 0.94 vs 0.82; Wilcoxon ranksum test p-value 4.40e-2 vs 2.20e-08).

Having 3 or more uAUGs in frame with the coding sequence is associated with having on median 33% lower half-lives (Wilcoxon ranksum test P value = 5.78e-07).

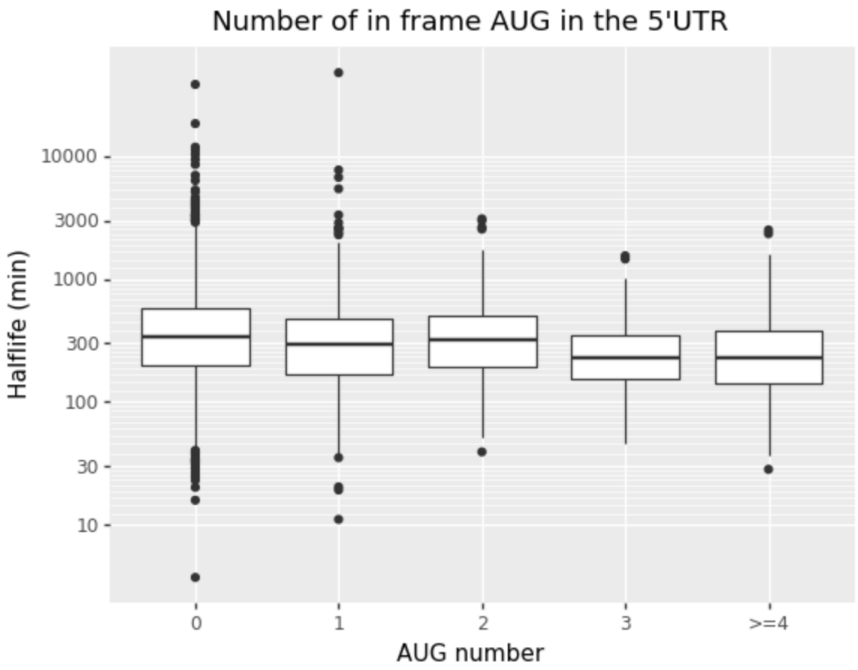


Figure 4.2: Boxplot depicting the distribution of half-life for mRNAs with 0,1,2,3 and 4 or more in frame uAUGs.

4.1.1.B uORF

In section 2.1.3.C, it is mentioned that mRNAs with start codons in the 5'UTR (uAUG) in frame with a stop codon in the 5'UTR can promote degradation through the NMD pathway, therefore lowering the half-lives of these mRNAs.

The median half-life of the mRNAs with one uORF is 13% lower (Wilcoxon ranksum test p-value = 7.03e-07). Furthermore, mRNAs with 3 or more uORFs have on median 23% lower half-lives (Wilcoxon ranksum test p-value = 1.02e-27).

4.1.1.C Kozak sequence

The Kozak sequence 2.1.2.B is known to facilitate translation initiation and, when present in the mRNA sequence, is comprised of the first 4 bases of the coding sequence plus 6 to 9 5'UTR bases immediately before the start codon AUG. Here the strongest 10 and 8 bases of the Kozak sequence are used and compared: GCC(A or G)CCAUGG or (A or G)CCAUGG.

The Kozak sequence with the strongest 8 bases (A or G)CCAUGG was found in 876 mRNAs. These mRNAs have a median half-life 9.8% higher than the remaining ones (Wilcoxon ranksum test p-value=3.2e-4). The Kozak sequence with the strongest 10 bases was found in 72 mRNAs. Despite having a median half-life 8.0% higher than the remaining mRNAs, its effect was not found to be significant (Wilcoxon ranksum test p-value=0.18).

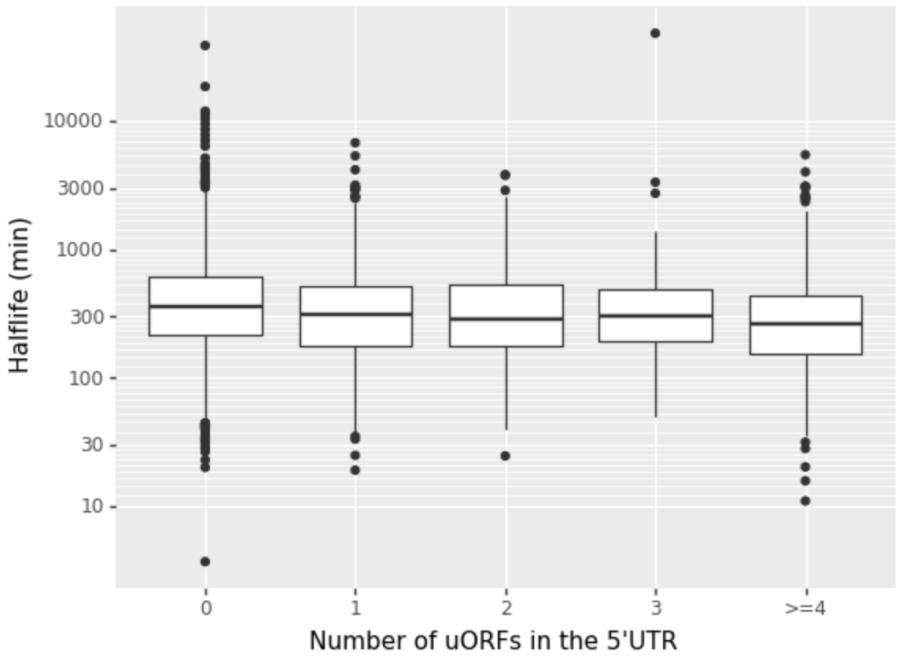


Figure 4.3: Boxplot depicting the distribution of half-life for mRNAs with 0,1,2,3 and 4 or more uORFs.

4.1.1.D PUM proteins binding motifs

The PUM1 and PUM2 proteins discussed in 2 have binding motifs UGUANAUA, where N stands for any base. These proteins bind to the mRNA and cause degradation. Having one instance of one of these motifs is associated with having a median half-life 19% lower compared with having 0 instances (Wilcoxon ranksum test P value = 8.09e-15). Furthermore, having 3 or more instances of these motifs associates with a median half-life 28% lower compared to 0 instances (Wilcoxon ranksum test P value = 1.06e-04).

4.1.1.E Codon content

The plot in figure 4.6 represents in the y axis the pearson correlation coefficient between the codon frequency (or ratio) 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 3 associates with half-life differently, negatively or positively, depending on the specific codon.

Figure 4.7 shows how CSC varies between codons encoding the same amino-acid. It indicates that the association of codons with half-life varies between synonymous ones.

The median frequency of a codon in all mRNAs was calculated as a measure of the median presence of a codon in the set of expressed mRNAs. Figure 4.8 depicts the relation between the CSC of a codon and its median frequency. The variables' relation point to a non-existent correlation (Pearson correlation

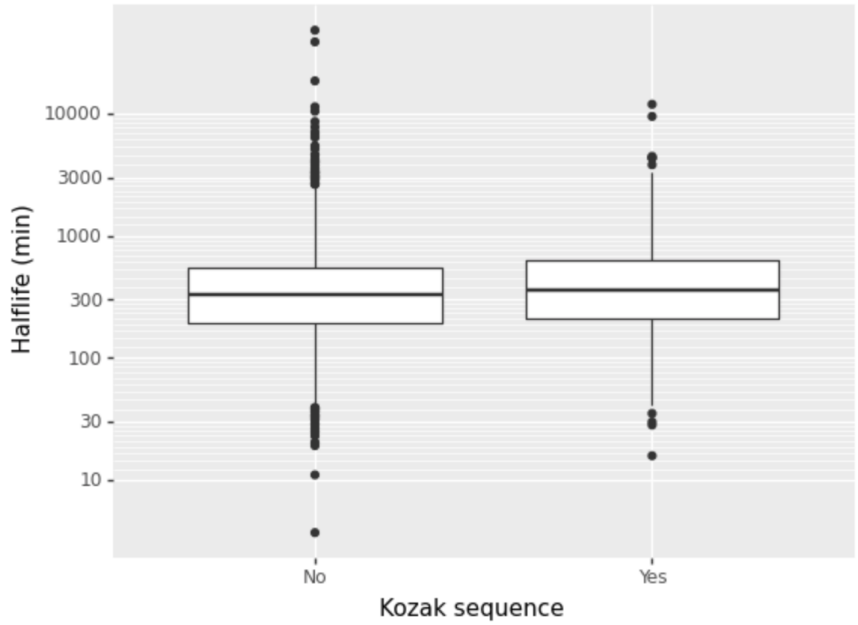


Figure 4.4: Boxplot depicting the distribution of half-life for mRNAs with and without the first of the 8 strongest bases of the Kozak sequence - (A or G)CCAUGG.

coefficient = 0.155; p-value = 0.232).

4.1.2 Modeling results

4.1.2.A 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 4.1 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 (see section 5).

Others like the stop codons TGA and TAG have a negative explained variance individually which

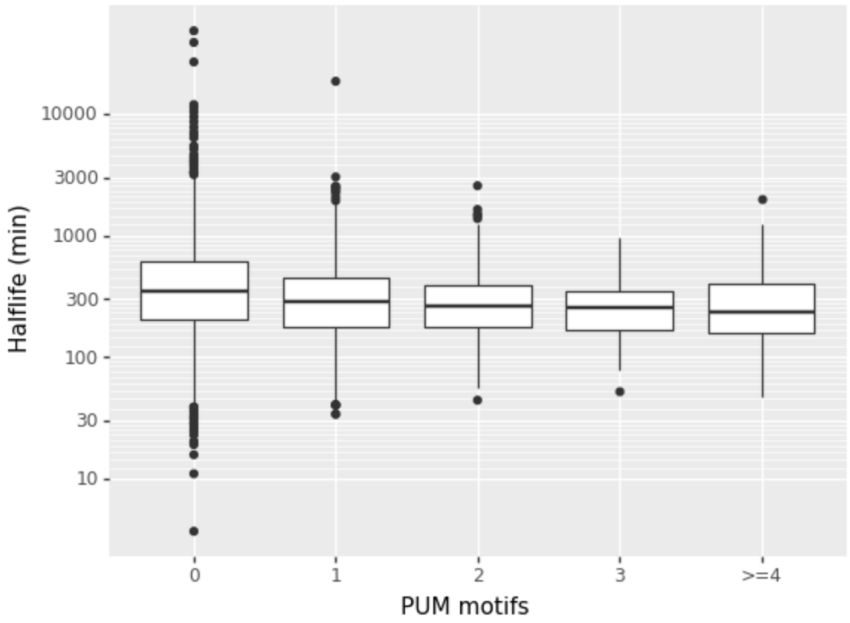


Figure 4.5: Boxplot depicting the distribution of half-life for mRNAs with 0,1,2,3 and 4 or more PUM protein binding motifs in the 3'UTR.

indicates a possibly non-existent relevant contribution to the predicted log(half-life).

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

Table 4.1: Individual and drop explained variance score for each feature in the ridge regression.

4.1.2.B Convolutional Neural Networks

The training and validation set performance during the training process for the 5'UTR CNN can be seen on figure 4.9. 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

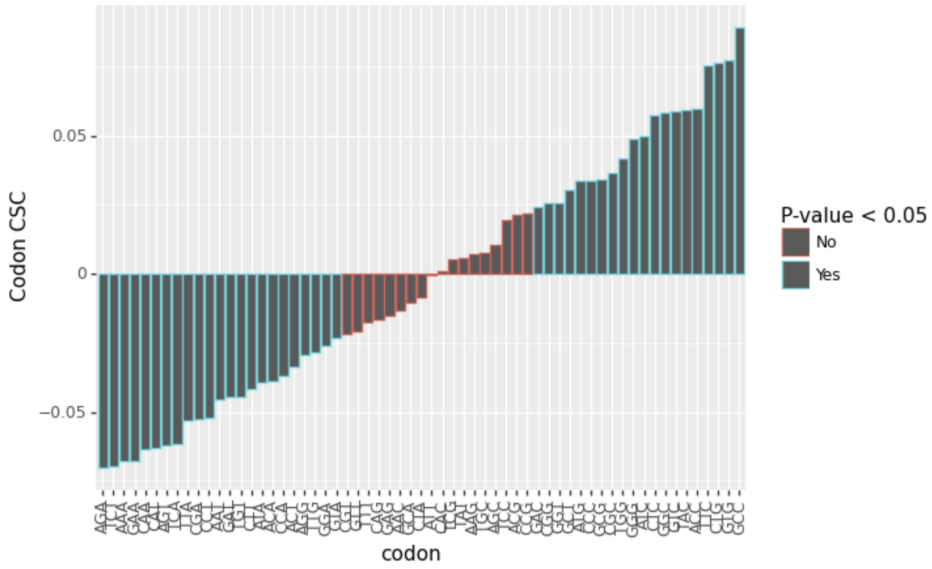


Figure 4.6: CSC per codon.

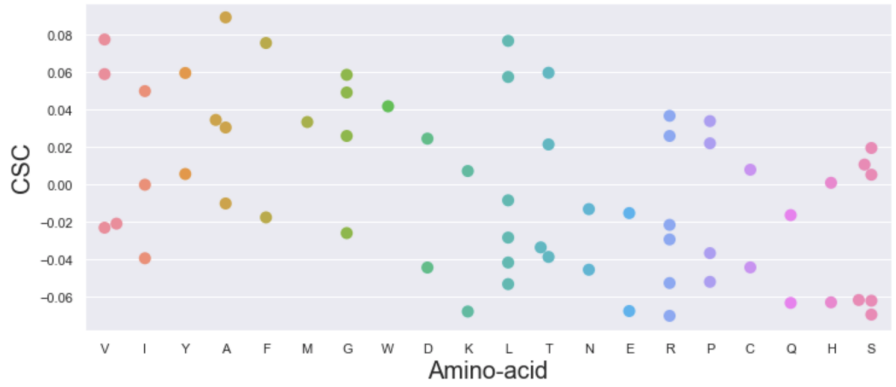


Figure 4.7: CSC grouped by amino-acid.

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).

The training and validation set performance during the training process for the 3'UTR CNN can be seen on figure 4.10. 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).

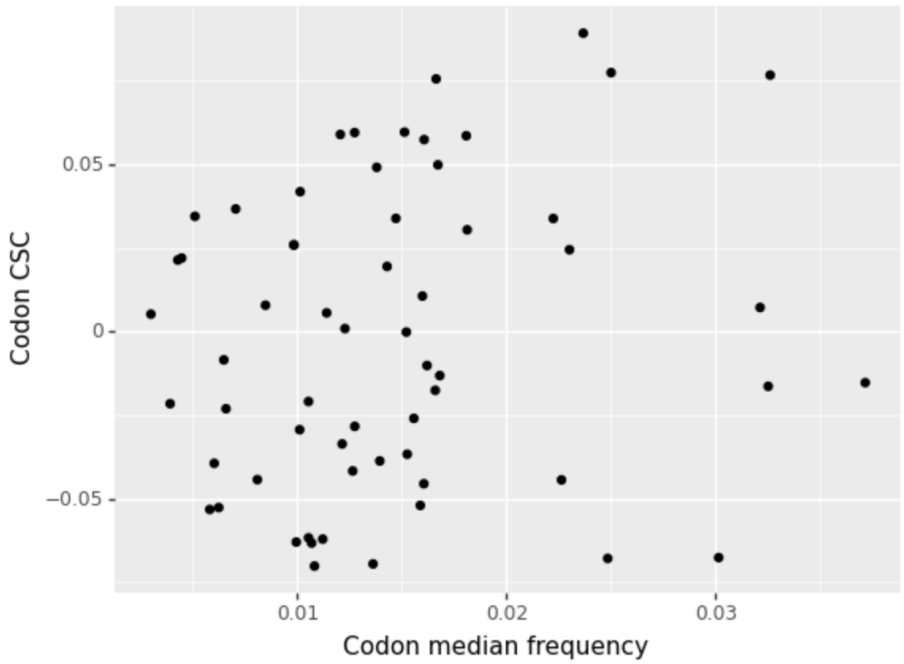


Figure 4.8: Codon stability coefficient (CSC) vs Codon median frequency.

4.1.3 Model interpretation

4.1.3.A 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 4.11 shows the contribution scores for the 3'UTR of an mRNA (PUSL1-201). It's possible to see several contiguous sequence regions with high positive or negative contribution scores.

4.1.3.B TF-MoDISco

The TF-MoDISco algorithm was applied for each set of 3'UTR and 5'UTR DeepLIFT contribution scores. The 4 motifs with the most amount of seqlets are represented in figure 4.12 for the 3'UTR and figure 5'UTR 4.13.

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

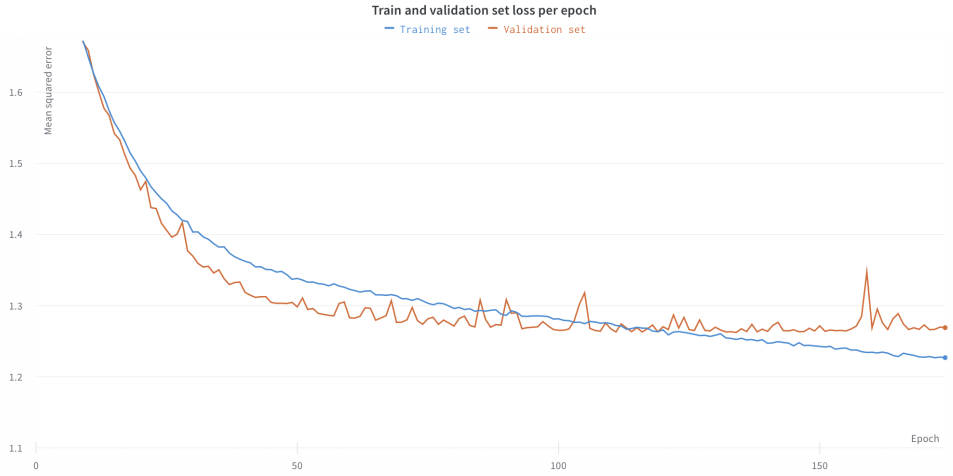


Figure 4.9: Training and validation set mean squared error per epoch on the 5'UTR CNN model.

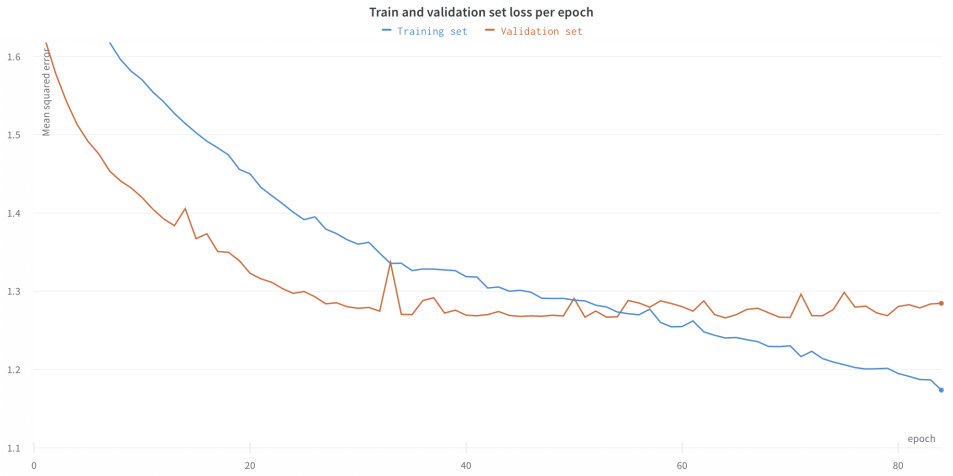


Figure 4.10: Training and validation set mean squared error per epoch on the 3'UTR CNN model.

sequence or not [3]. 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.

In order to evaluate further 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 4.14 represents such a comparison, for the motif with the most amount of seqlets on the 3'UTR (AGNCTCA). Notice that, as depicted in the 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

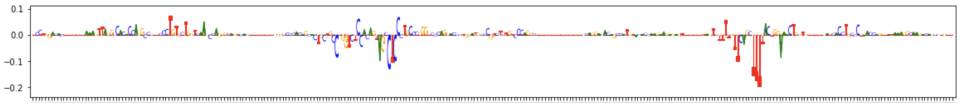


Figure 4.11: DeepLIFT contribution scores for the mRNA 3'UTR belonging to the PUSL1-201 transcipt. The y axis represents the contribution score and the x axis each position of the 3'UTR. The height of the letters reveals the magnitude of the contribution and the orientation (facing left or down) indicates the sign of the contribution.

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).

In figure 4.15 we can see the same motif distribution comparison as previously, although 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 (1.37e-3).

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) (Fig. 4.16).

By looking at what position these motifs are found in their respective UTRs some patterns were found. Figures 4.20 and 4.19 show the distribution of the motifs TATTG and AAAA position relative to the total length of the 3'UTR and 5'UTR. 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.

4.2 Modeling tissue-specific mRNA half-life variations

4.2.1 Multi-task DNN

The training and validation set performance during the training process can be seen on figure 4.21. The best validation performance (mean squared error = 1.26) was achieved on epoch 134. The training

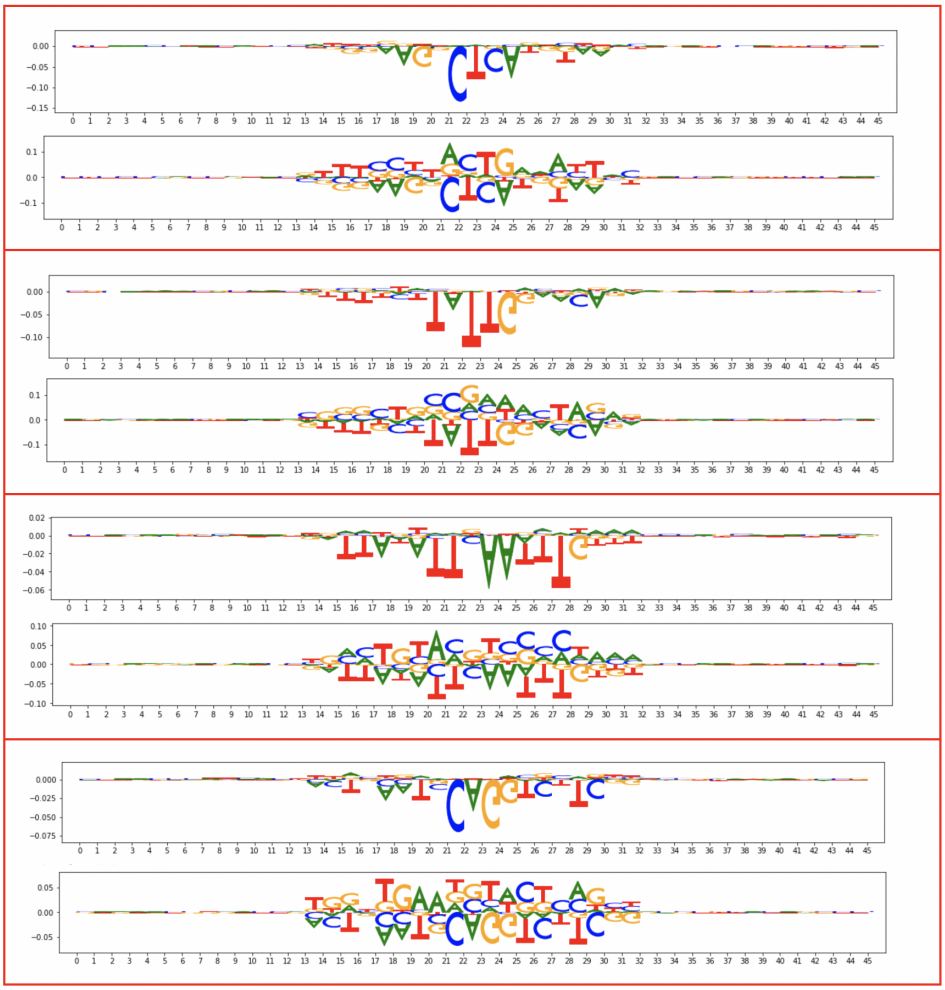


Figure 4.12: 3'UTR motifs corresponding to the most amount of seqlets ordered from highest on the top to lowest on the bottom. Each red box contains one motif, where the top sequence shows the motif with the "real" contribution scores for each seqlet and the the bottom sequence shows the motif with the "hypothetical" contribution scores. A letter facing up indicates a positive contribution to half-life a letter facing down indicates a negative contribution. Number of seqlets per motif: 662, 635, 553, 516.

process was stopped after reaching 150 epochs with no improvement on the validation set performance.

The explained variance and Pearson correlation coefficient for each tissue is shown in figures 4.22 and 4.23.

4.3 A tissue-specific codon effect program

4.3.1 Codon content as a predictor of tissue-specific mRNA half-life variations

Given the mRNA half-life variations derived from the Exonic/Intronic ratios (also used on 4.2), we set out to explore how the specific content of codons in an mRNA influences its half-life variation on each

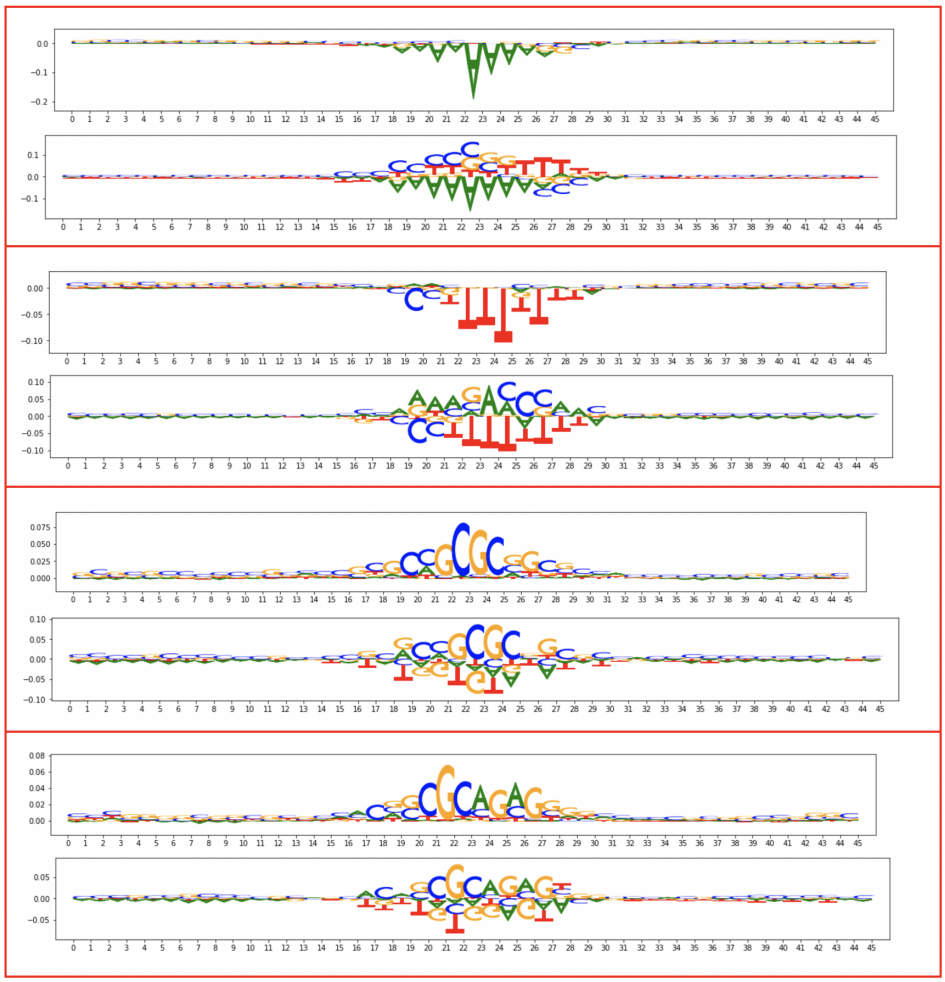


Figure 4.13: 5'UTR motifs corresponding to the most amount of seqlets ordered from highest on the top to lowest on the bottom. Number of seqlets per motif: 1361, 1057, 337, 274.

tissue.

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, and as target output the mRNA half-life variation for 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:

$$\Delta\text{log(mRNA half-life)}_{t_i} = \beta^{t_i}_{\text{AAA}} f_{\text{AAA}} + \beta^{t_i}_{\text{AAC}} f_{\text{AAC}} + ... + \beta^{t_i}_{\text{TTT}} f_{\text{TTT}} + \beta^{t_i}_0 \tag{4.1}$$

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

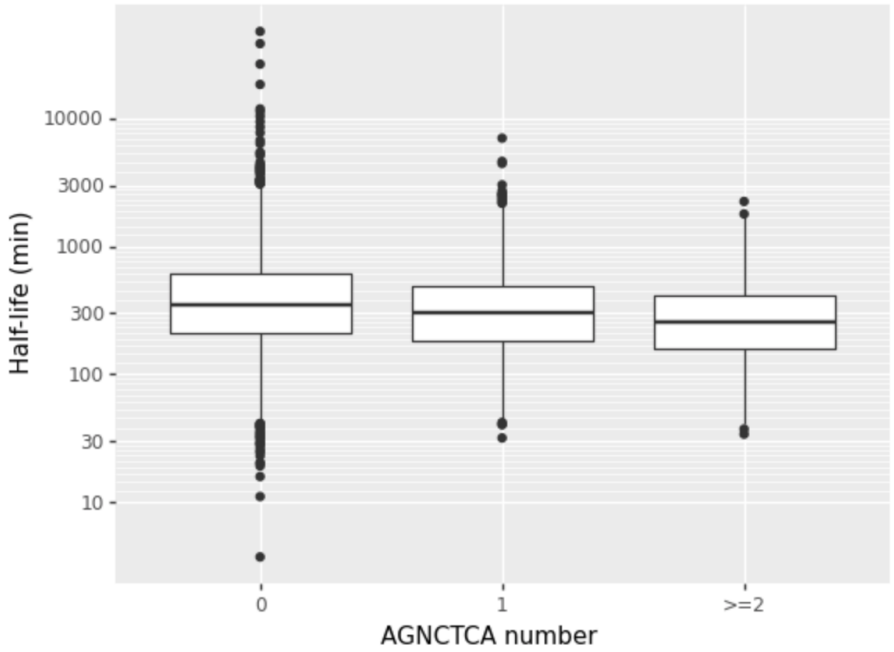


Figure 4.14: Boxplot depicting the distribution of half-life for mRNAs with 0,1 or 2 or more AGNCTCA motifs in the 3'UTR.

on the predictor for tissue t_i, and f_{codon_k} is the frequency of codon k in the mRNA as defined in 3.1.3.

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 the 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. Figures 4.24 and 4.23 give further detail on the evaluation of each tissue model's prediction on the test set.

4.3.1.A Model interpretation - tissue-specific codon effect on ΔmRNA half-life

The linear model described in 4.1 can be used to further inspect the effect of each codon in ΔmRNA half-life, as captured by the model. As

$$\frac{\partial \Delta \log(\text{mRNA half-life})_{t_i}}{\partial f_{\text{codon}_k}} = \beta^{t_i}_{\text{codon}_k} \tag{4.2}$$

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 ΔmRNA 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 ΔmRNA half-life$_{t_i}$ and the magnitude of $\beta^{t_i}_{\text{codon}_k}$ indicates the strength of this effect.

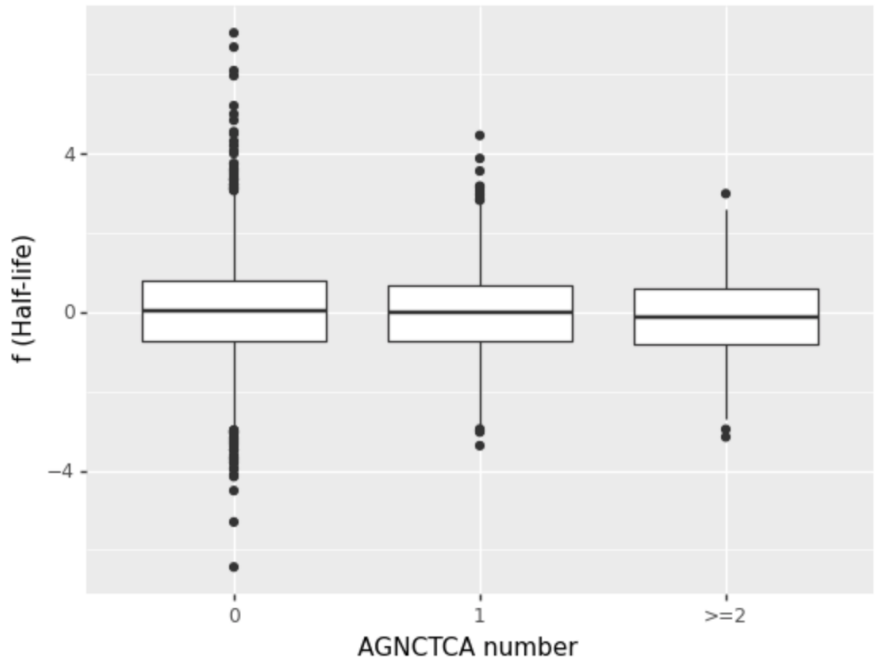


Figure 4.15: Boxplot depicting the distribution of the corrected half-life effect $f(\text{Half-life})$ for mRNAs with 0,1 or 2 or more AGNCTCA motifs in the 3'UTR.

4.3.1.B Visualizing $\beta^{t_i}_{\text{codon}_k}$

The tissue-specific codon effects on Δlog(mRNA half-life) as measured by $\beta^{t_i}_{\text{codon}_k}$ can be visualized in the form of the clustered heatmap of figure 4.25.

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.

4.3.1.C Principal component analysis of β_{codon_k}

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 4.26 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

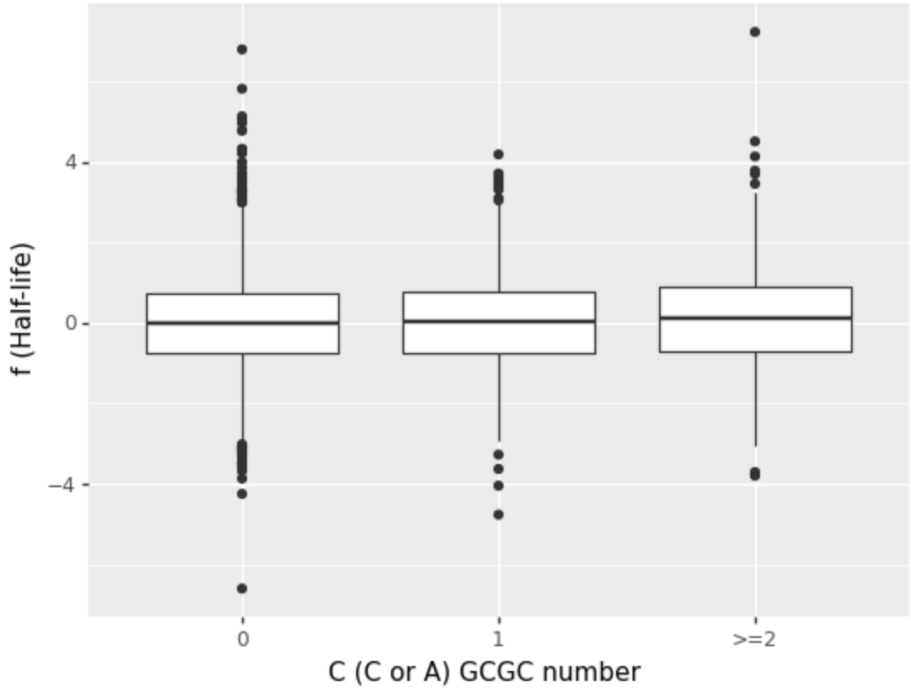


Figure 4.16: Boxplot depicting the distribution of the corrected half-life effect $f(\text{Half-life})$ for mRNAs with 0,1 or 2 or more ”C (C or A) GCGC” motifs in the 5’UTR.

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

4.3.1.D Tissue-specific codon signature relationship with transcript amounts

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 (see 3). 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. Figure 4.27 illustrates in a graph the relationship between those 2 vectors, where the specific gene is MT-CO1.

All 38793 genes’s TPMs were then ordered by their Spearman correlation with the codon signature.

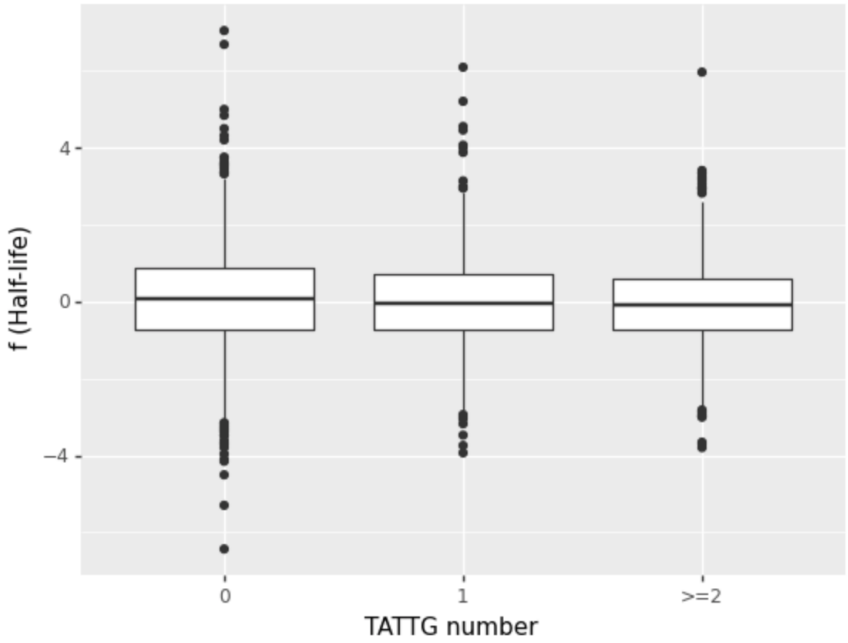


Figure 4.17: Boxplot depicting the distribution of the corrected half-life effect $f(\text{Half-life})$ for mRNAs with 0,1 or 2 or more TATTG motifs in the 3'UTR.

Table 4.2 illustrates the 15 top correlating genes. Surprisingly, all genes in the table are related to mitochondrial pathways and more than half are encoded by the mitochondrial DNA.

We then performed a gene set enrichment analysis using the ordering obtained previously. Figure 4.28 illustrates one of the top enriched pathways. The high enrichment for respiration-related pathways like oxidative phosphorylation and krebs cycle (A.10) highlight the strong relationship between the mitochondrial genes and codon signatures seen previously.

4.3.2 Tissue-specific codon signatures across human individuals

Having analyzed the tissue-specific codon signatures on the average human individual (3), 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 in 4.3.1.A, $\beta^{t_i}_{\text{codon}_k}$ was obtained for each individual. 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 4.29. 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. Such can also be observed by looking at the overall patterns of β_{codon_k} across samples in figure 4.30.

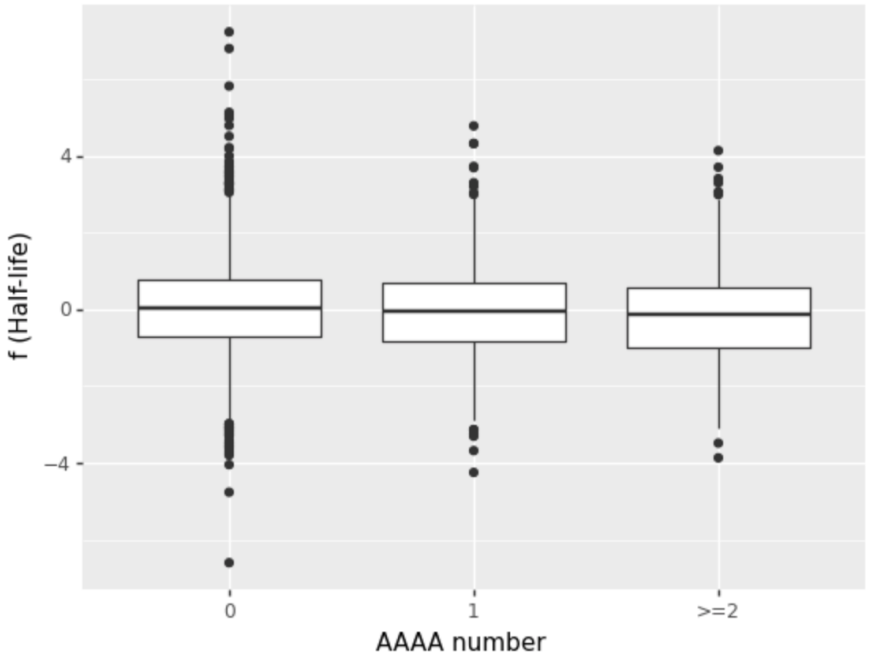


Figure 4.18: Boxplot depicting the distribution of the corrected half-life effect $f(\text{Half-life})$ for mRNAs with 0,1 or 2 or more AAAA motifs in the 5'UTR.

Previously we have shown the consistency of the codon signature metric across same tissue types of different individuals. We now set out to explore how this metric relates to specific individual traits such as age and sex, and specific sample acquisition characteristics like ischemic time and RNA integrity number (RIN).

4.3.2.A Age

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 (appendix A.5). The strongest negative Spearman correlation of value -0.35, was found for blood vessel tissue (Figure 4.31).

4.3.2.B Sex

Overall, the individual's sex didn't show any correlation with codon signature (figure 4.32). 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 (figure A.6).

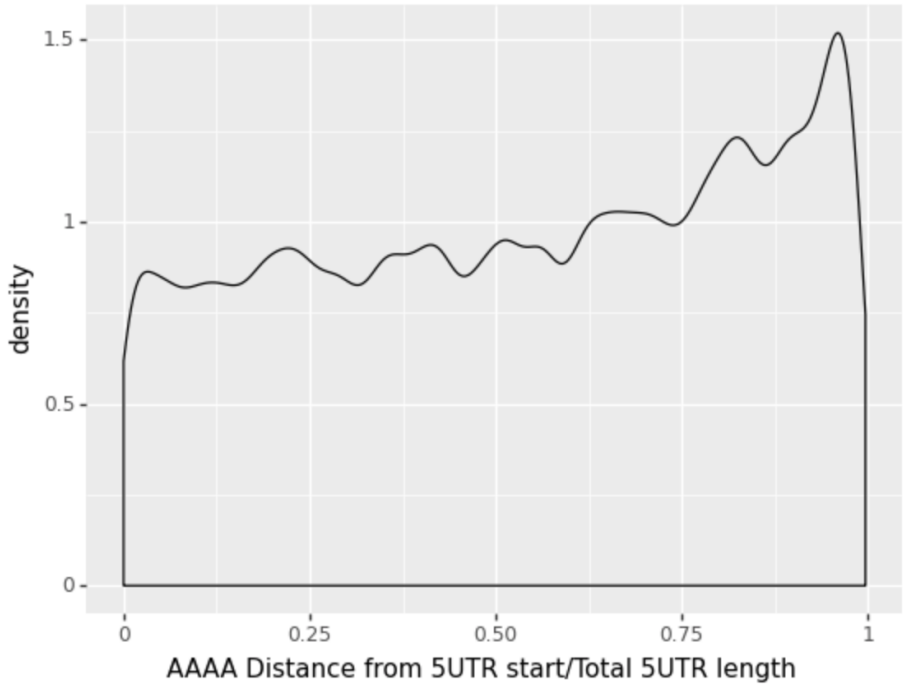


Figure 4.19: Density plot showing the distribution of the relative position of AAAA on the 5'UTR of mRNAs. The relative position is computed as the quotient between the distance from the beginning of the 5'UTR (5' end) and the total 5'UTR length.

4.3.2.C Ischemic time

Ischemic time 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 (figure A.7), showing a high negative correlation on heart and lung tissue (Spearman correlation = -0.69 for heart and -0.64 for lung; see figure 4.33).

4.3.2.D RNA integrity number (RIN)

Overall a RIN's sample correlates positively with its codon signature (Spearman correlation coefficient = 0.34, p-value = 5.20e-226) with correlation varying across tissues (figure A.8)). The highest correlation was obtained for ovary and heart tissue (Spearman correlation coefficient = 0.61 for heart and 0.58 for ovary; see figure 4.34)

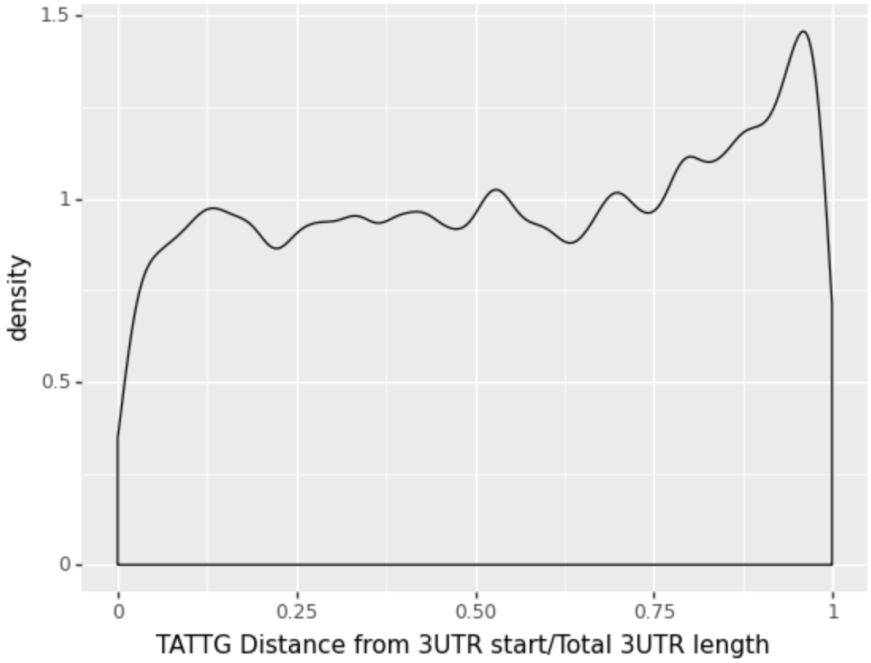


Figure 4.20: Density plot showing the distribution of the relative position of TATTG on the 3'UTR of mRNAs. The relative position is computed as the quotient between the distance from the beginning of the 3'UTR (after the stop codon) and the total 3'UTR length.

4.3.3 Codon effects on mRNA half-life variation between glucose and galactose cell cultures

In 4.3.1.D we have seen a connection between codon signatures and mitochondrial-pathway-related transcript abundance. Compared to glucose mediums, galactose is known to enhance mitochondrial metabolism [36]. Therefore, studying the effect of codon content in mRNA half-life differences between these two mediums can possibly add more insight into the codon signature differences we see across tissues.

Following the same procedure as described earlier, a linear regression model was fitted to predict the log difference between an mRNA's half-life for cells grown in a glucose medium relative to cells grown in a galactose medium, using the codon content as input. The model performance on the test set was 0.18 (Pearson correlation coefficient between predicted and ground truth values).

The resulting $\beta_{\text{codon}_k\,\text{Gal/Glu}}$ were then extracted from the model and are shown in figure 4.35.

4.3.4 Comparing codon effects on half-life and its variations

Having obtained the effect of a codon in the mRNA half-life variation between these two conditions as measured by $\beta_{\text{codon}_k\,\text{Gal/Glu}}$, we set out to analyze its relationship with the tissue specific codon effects obtained on section 4.3.1.A.

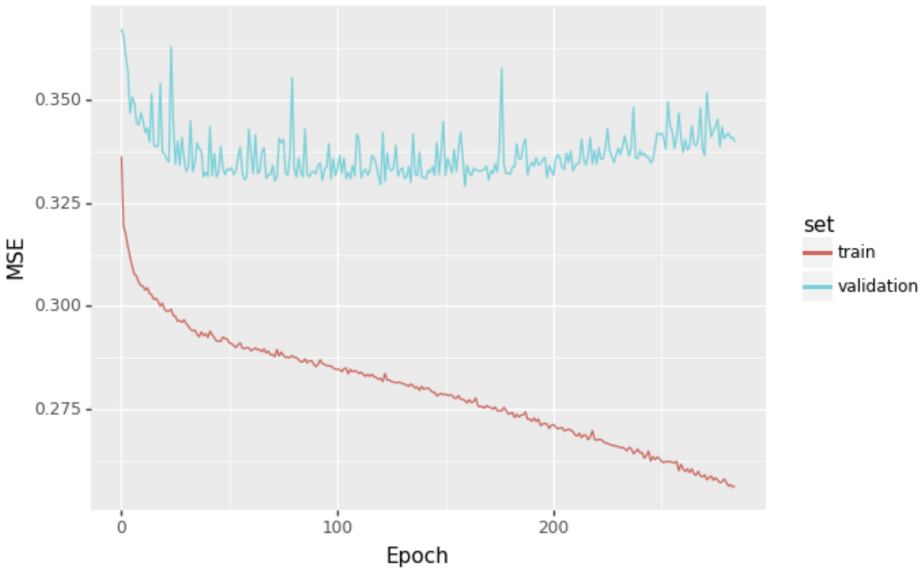


Figure 4.21: Training and validation set mean squared error per epoch for the multi-task DNN model.

To that end we computed the projections of each codon axis into the first principal component (the codon signature axis). Each codon projection gives us a measure of the sign and magnitude of that codon on the codon signature axis, suggesting its contribution to the overall trend captured by the codon signature axis.

These codon projection values were then tested for correlation with $\beta_{\text{codon}_k\,\text{Gal/Glu}}$. The correlation was 0.464 (Pearson correlation with p-value = 1.62e-4). Figure 4.36 illustrates the relationship between these two variables.

Furthermore, we extracted the regression coefficients corresponding to codons from the linear model fitted to predict half-life on a human cell line (section 4.1). Following the same reasoning as in 4.3.1.A, these coefficients, now termed as $\beta_{\text{codon}_k\,\text{half-life}}$, can be used to interpret the influence of a frequency of a codon in half-life.

We then tested the correlation between $\beta_{\text{codon}_k\,\text{half-life}}$ and both $\beta_{\text{codon}_k\,\text{Gal/Glu}}$ and the codon projection values. $\beta_{\text{codon}_k\,\text{half-life}}$ and $\beta_{\text{codon}_k\,\text{Gal/Glu}}$ show a correlation of 0.443 (Pearson correlation with p-value=3.45e-4, figure 4.38). The codon projection values and $\beta_{\text{codon}_k\,\text{half-life}}$ show a correlation of 0.532 (Pearson correlation with p-value=1.02e-5, figure 4.37).

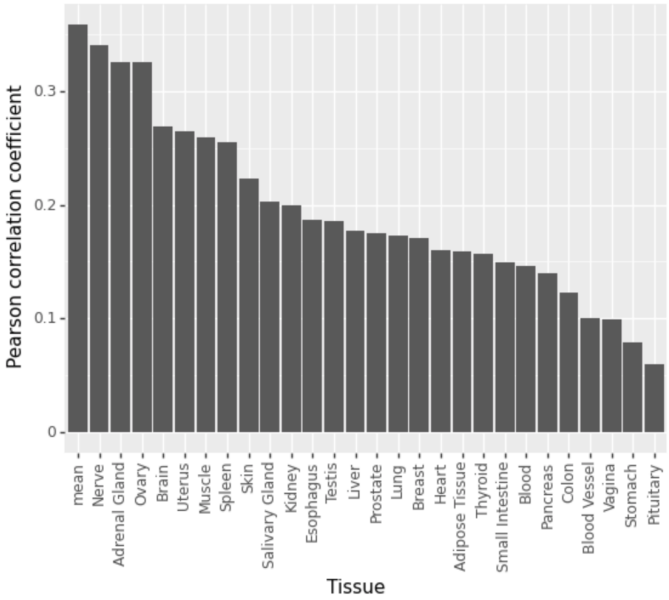


Figure 4.22: Pearson correlation coefficient per tissue and for the mean value (mean Exonic/Intronic ratio).

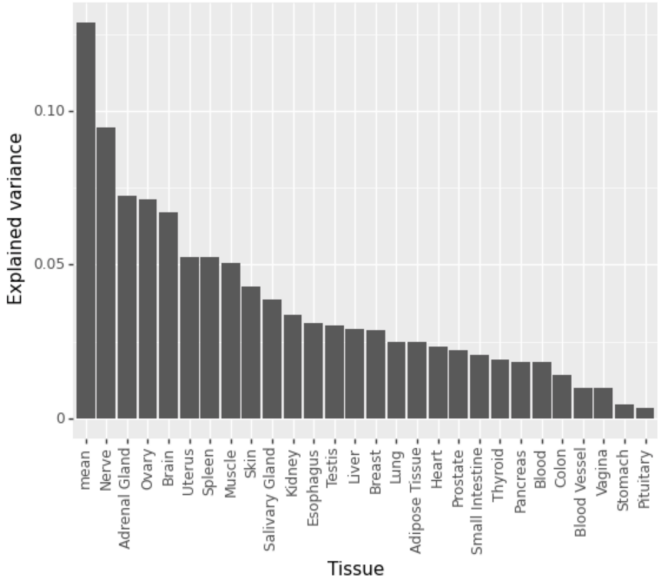


Figure 4.23: Explained variance per tissue and for the mean value (mean Exonic/Intronic ratio).

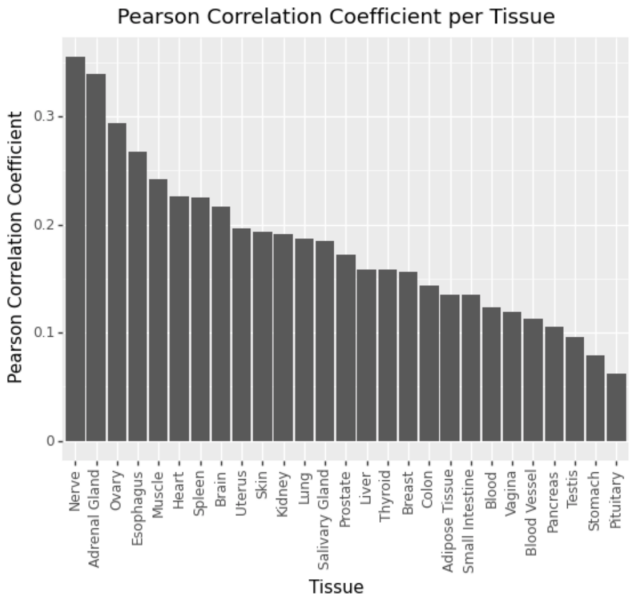


Figure 4.24: Pearson correlation coefficient for each tissue's linear regression model.

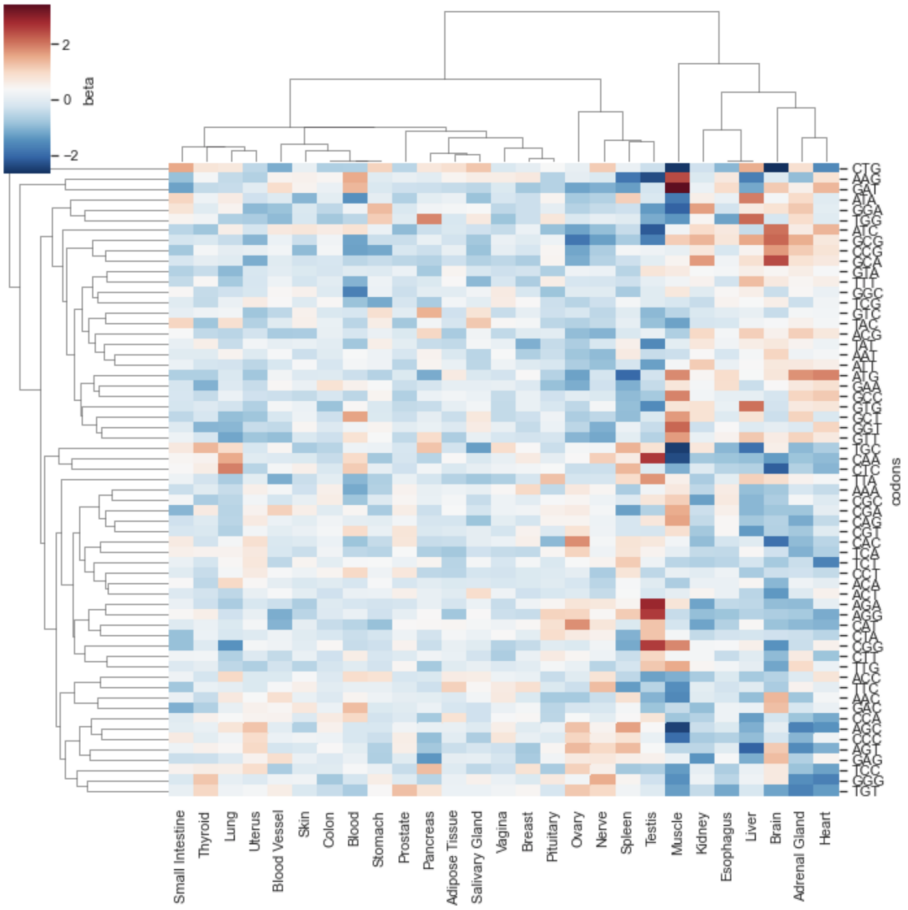


Figure 4.25: Clustered heatmap depicting the relationship of β across different tissues and codons.

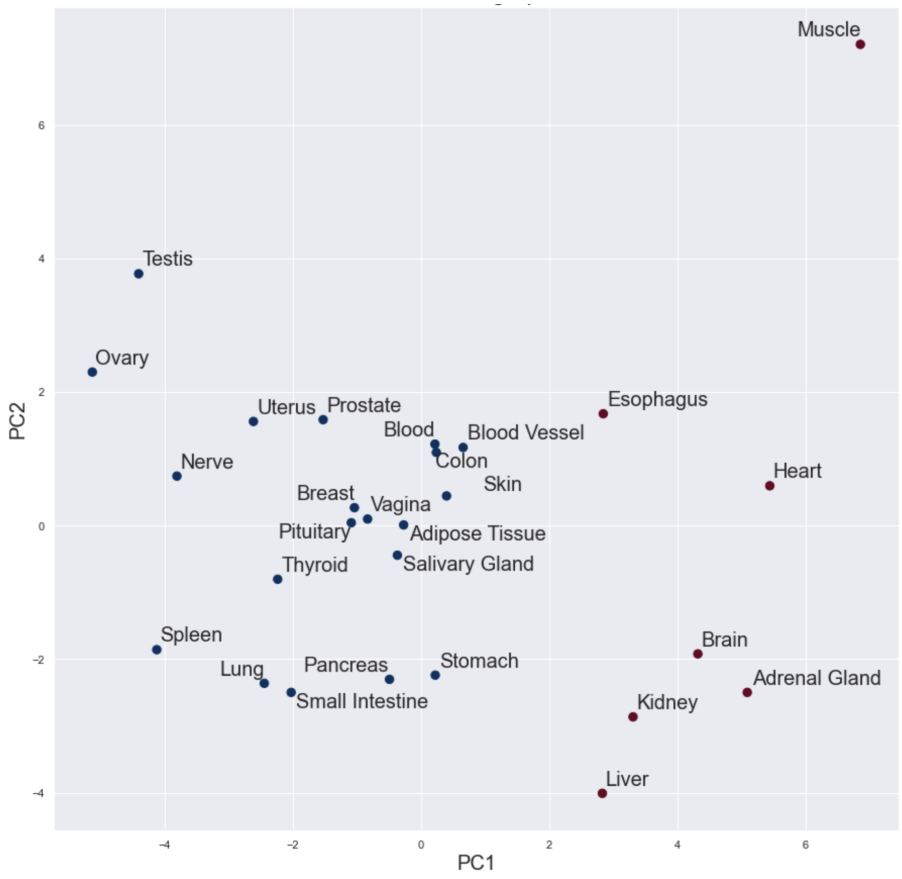


Figure 4.26: Tissues projected into the PC1 and PC2 components. The colors represent 2 clusters obtained by applying k-means clustering on the tissue's coordinates in the n-dimensional space.

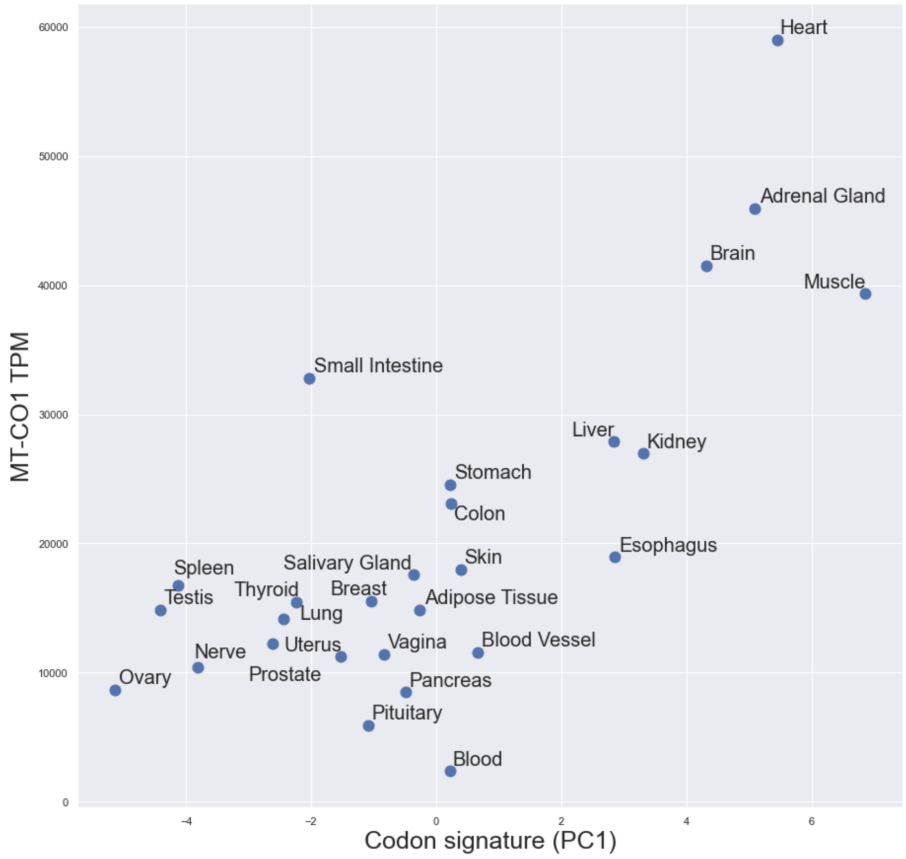


Figure 4.27: Tissue-specific MT-CO1 transcripts per million (TPM) vs tissue's codon signature (PC1).

Gene name	DNA from:	Type	Description	Spearman correlation coefficient
MT-CO1P53	Mitochondria	Pseudogene	MT-CO1 pseudogene 53	0.712
AC006064.4	Nucleus	Long non-coding RNA	Antisense To GAPDH	0.686
VDAC1	Nucleus	Protein coding	Voltage-Dependent Anion-Selective Channel Protein 1	0.679
MTCO1P40	Mitochondia	Pseudogene	MT-CO1 pseudogene 40	0.676
MTCO1P2	Mitochondia	Pseudogene	MT-CO1 pseudogene 2	0.668
COX5B	Nucleus	Protein coding	Cytochrome C Oxidase Subunit 5B	0.657
MT-ND5	Mitochondia	Protein coding	Mitochondrially Encoded NADH: Ubiquinone Oxidoreductase Core Subunit 5	0.651
MTCO1P12	Mitochondia	Pseudogene	MT-CO1 pseudogene 12	0.642
MTCO1	Mitochondia	Protein coding	Mitochondrially Encoded Cytochrome C Oxidase I	0.637
MDH1	Nucleus	Protein coding	Malate Dehydrogenase 1	0.627
MT-CYB	Mitochondia	Protein coding	Mitochondrially Encoded Cytochrome B	0.625
VDAC1P2	Nucleus	Pseudogene	Voltage Dependent Anion Channel 1 Pseudogene 2	0.621
UQCRQ	Nucleus	Protein coding	Ubiquinol-Cytochrome C Reductase Complex III Subunit VII	0.617
MTCO2P12	Mitochondia	Pseudogene	MT-CO2 Pseudogene 12	0.606
ATP5MC3	Nucleus	Protein coding	ATP Synthase Membrane Subunit C Locus 3	0.605

Table 4.2: Table containing the first top 15 correlating genes. The columns contain the location of the gene's DNA (mitochondria or nucleus), the gene type and a brief description of the gene's known characteristics.

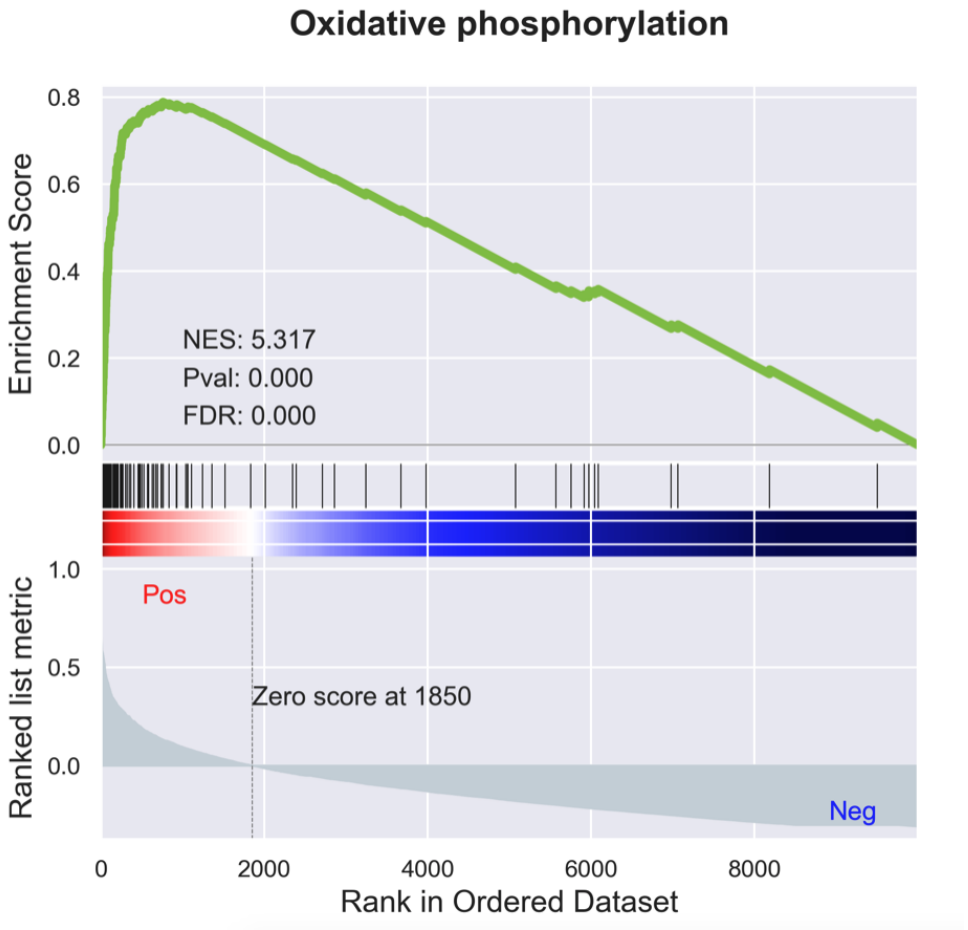


Figure 4.28: Gene set enrichment analysis results for the oxidative phosphorylation pathway. The black stripes correspond to the position in the ranking of the genes composing the gene set of the oxidative phosphorylation pathway. The ranked list metric axis is the spearman correlation between transcript abundance and codon signature discussed previously. The shape of the curve on the enrichment score axis indicates that the gene set involved in oxidative phosphorylation shows an enrichment for the first end of the ranking. NES stands for normalized enrichment score, Pval stands for p-value and FDR stands for false discovery rate. For more details on the GSEA algorithm see [4].

Figure 4.29: 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.

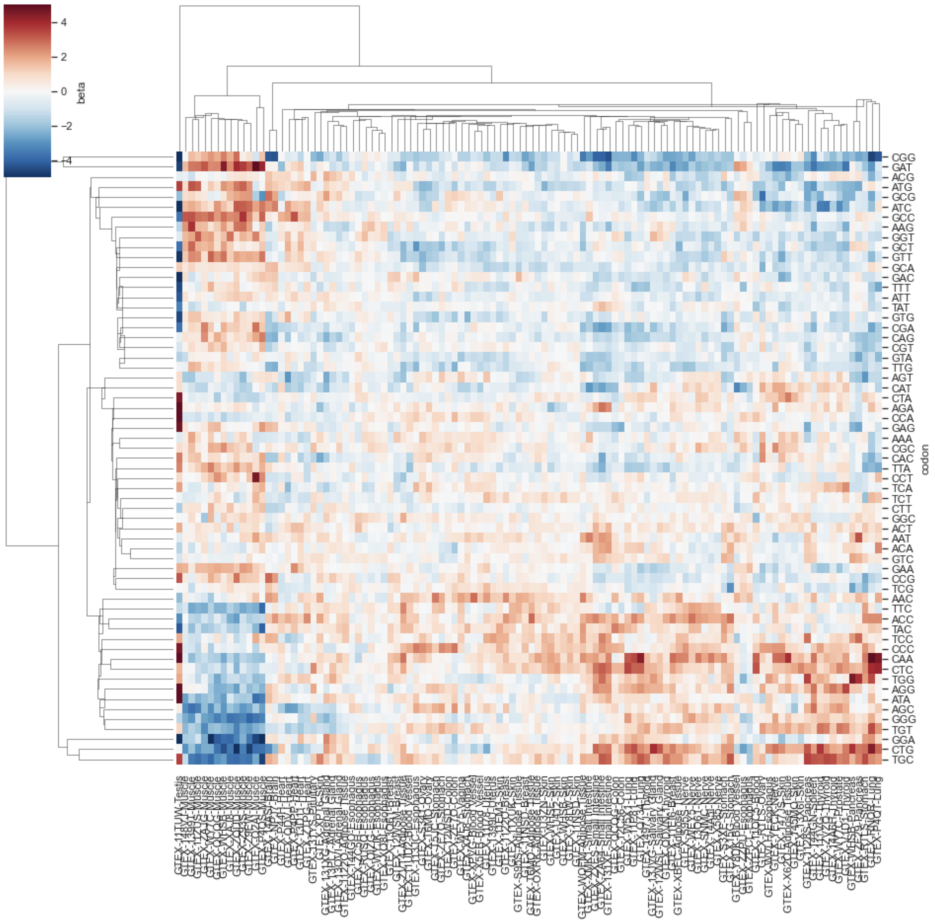


Figure 4.30: Clustered heatmap containing β_{codon_k} for 110 randomly chosen samples.

Figure 4.31: Boxplot showing the codon signature distribution per age group on blood vessel tissue samples. Spearman correlation coefficient=-0.35, p-value=2.17e-13.

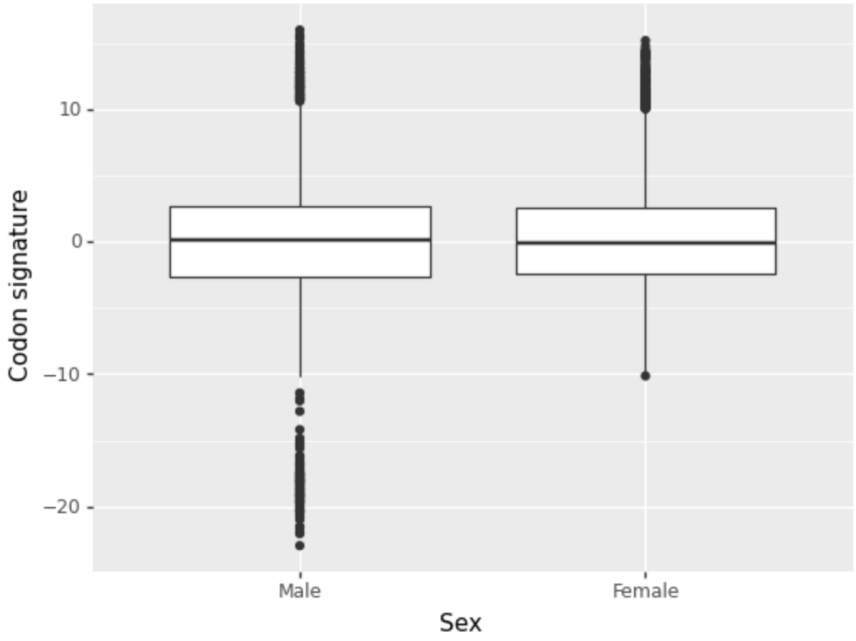


Figure 4.32: Boxplot showing the codon signature distribution per sex.

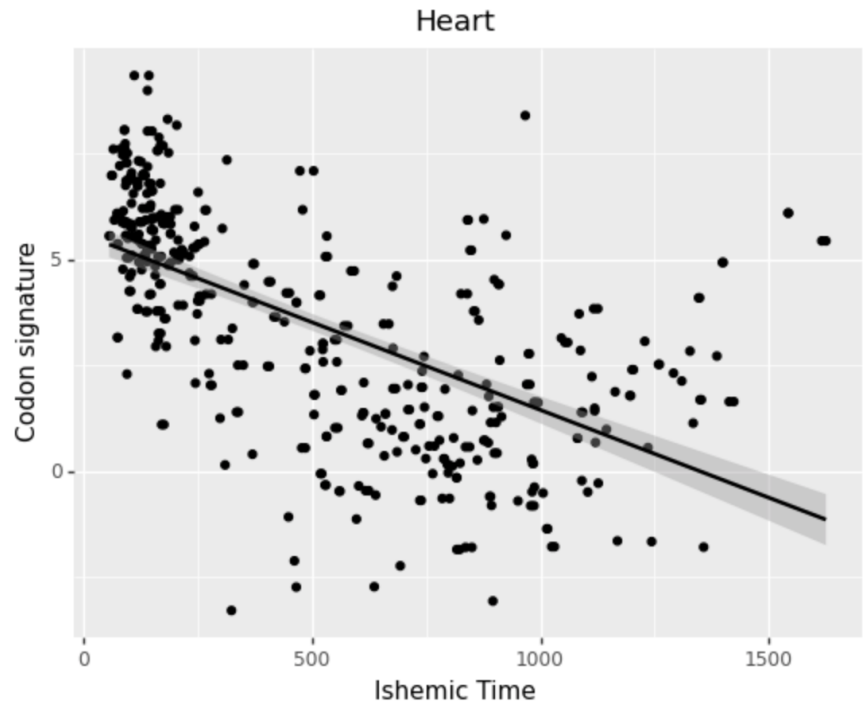


Figure 4.33: Ischemic time vs codon signature in heart samples.

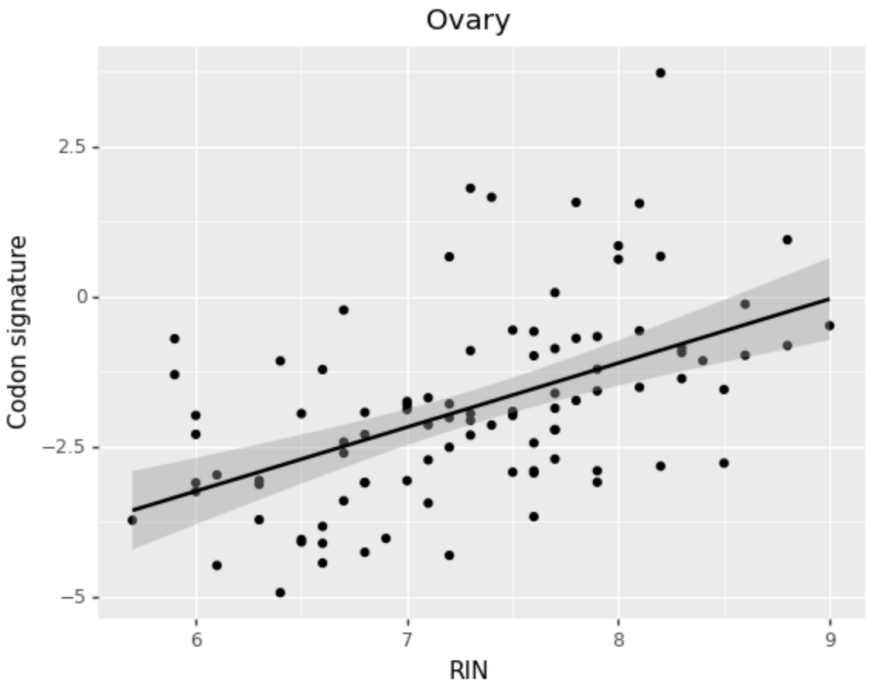


Figure 4.34: RIN and codon signature in ovary samples.

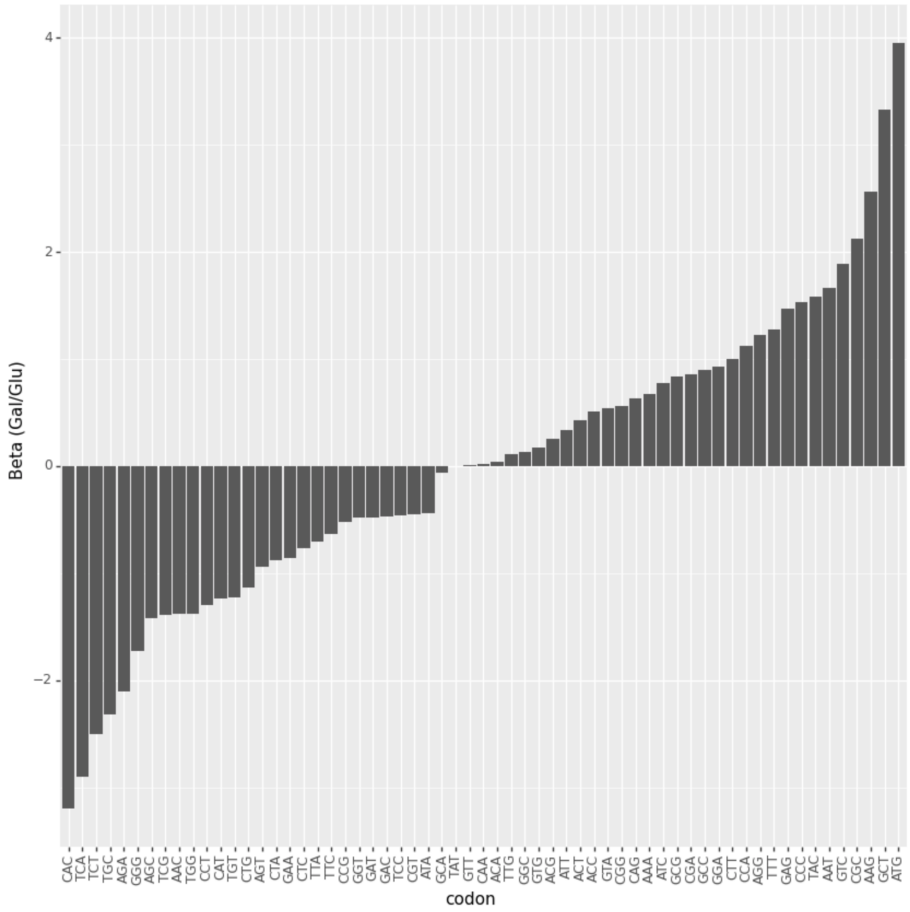


Figure 4.35: $\beta_{\text{Gal/Glu}}$ for each codon.

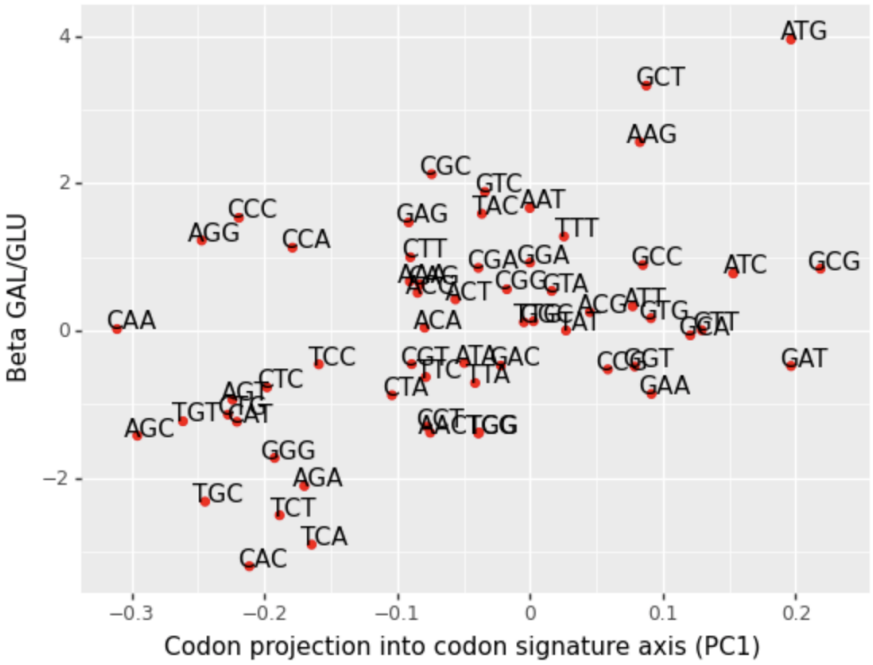


Figure 4.36: Scatter plot between $\beta_{\text{codon}_k\,\text{Gal/Glu}}$ and the codons projections into the codon signature axis.

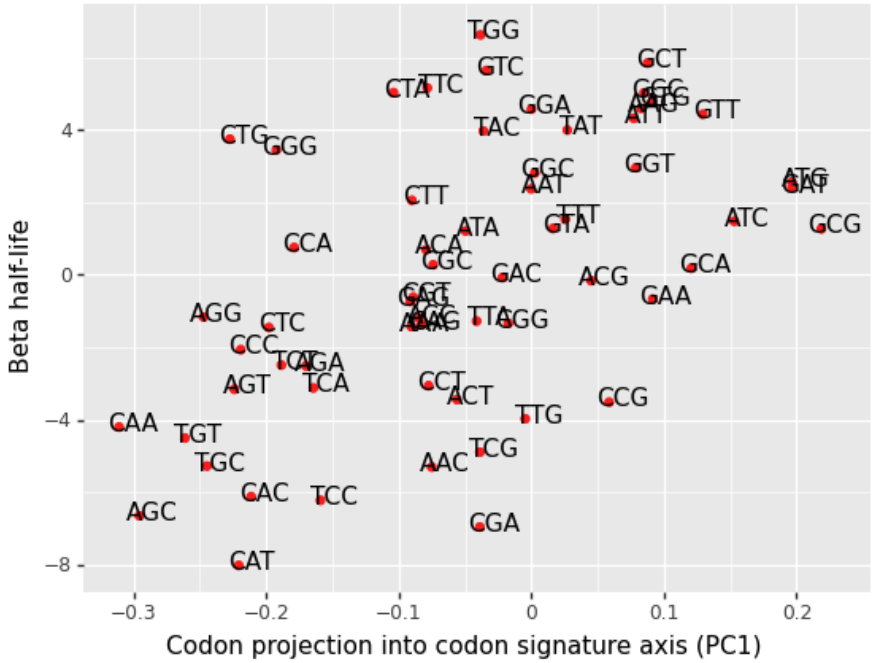


Figure 4.37: Scatter plot between $\beta_{\text{codon}_k\,\text{half-life}}$ and the codons projections into the codon signature axis.

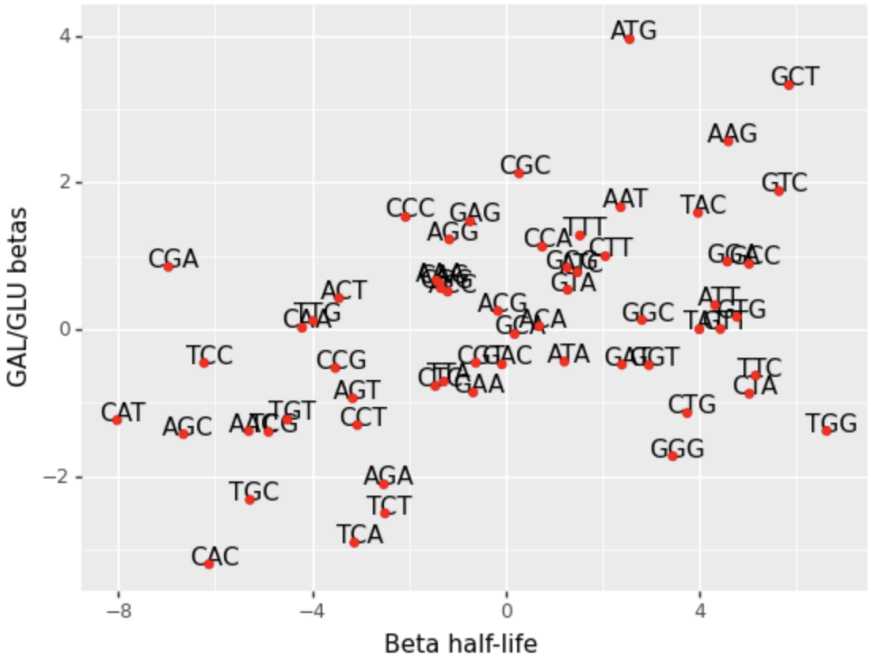


Figure 4.38: Scatter plot between $\beta_{\text{codon}_k\,\text{Gal/Glu}}$ and $\beta_{\text{codon}_k\,\text{half-life}}$.

5

Discussion

Contents

5.1 Modeling mRNA in a human cell-line

In line with the literature [7], start codons or opening reading frames in the 5'UTR (uAUGs and uORFs respectively) were found to be associated with lower mRNA half-lives. Having uAUGs in frame with the coding sequence is associated with lower half-lives than having an opening reading frame in the 5'UTR, indicating that alternative in-frame translation starting sites may have a more destabilizing impact than translated uORFs.

The main 8 bases of the Kozak sequence were found to be associated with higher half-lives. Furthermore, in line with previous research [16], mRNAs with PUM protein binding motifs were associated with lower half-lives.

Although all these mRNA features were found to significantly associate with half-life, their contribution to the final linear regression model was minor for PUM motifs and insignificant for uORF, uAUG and Kozak sequence. Such low contributions can indicate that the effect of this features on half-life is minute or that such effect is captured by other features, such as the length of the UTR. Indeed, the longer the UTR the more probable it is to have a certain motif and generally the lower the half-life of an mRNA. Although this could explain the poor predictive power of uORFs, uAUGs or PUM binding motifs, it cannot explain the poor performance of the Kozak sequence, as it is fixed to the same positions in all mRNAs and those positions are found in practically all. Having used a specific combination of bases to characterize the Kozak sequence may have been part of the explanation for its low predictive power, as this sequence is not defined only as a combination of immutable bases but as a combination of bases with different contributions for each position of the sequence.

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.

As discussed in 2, the effect of each codon in an mRNA's half-life is thought to be caused mainly because of its decoding rate. Through this view, codons with higher CSCs should be associated with higher decoding rates. In turn, the decoding rate depends, between other factors, on the available amount of ready-to-translate tRNAs of a certain codon and its demand. A rough measure of its demand is obtained by retrieving the median frequency of each codon on all expressed mRNAs. As measured by this approach, the ready-to-translate tRNA demand of each codon did not correlate with its effect on stability. Such can be explained by the quality of metric used for ready-to-translate tRNA demand, as the amount of expressed mRNAs was not considered, meaning all expressed mRNAs were considered equally important for codon demand. Other explanation is that the decoding rate also depends on the supply of ready-to-translate tRNAs, which was not captured by the previous metric. To sum up, these results can suggest that the frequency of a codon in the genome's expressed mRNAs doesn't capture

its specific impact on half-life or perhaps its decoding rate.

Quantitative modeling of mRNA half-life from sequence on human cells has not been made on previous research. However, similar to what was found in the yeast organism on [7], codons explain the bulk of the mRNA variability as predicted by a linear model. This result indicates that mRNA translation may be a strong determinant of its half-life.

After codon content, the 3'UTR length and the GC content of the 5'UTR were the features explaining the second and third highest amount of variability. The GC content of the 5'UTR had already been documented to be connected with mRNA secondary structures in the 5'UTR which in turn influence translation initiation [37] and possibly mRNA half-life. Furthermore, a higher 3'UTR length increases the probability of having a higher amount of regulatory motifs and possibly changes the overall structure of the mRNA.

As discussed previously, the task of predicting mRNA half-life in human cells from sequence was not, to our knowledge, topic of previous research. In that way, it is not possible to compare the developed convolutional models on the 5'UTR and 3'UTR to any baseline model outside of this work. Nevertheless, the development of such models was mainly intended as a tool to find new regulatory motifs through the use of model interpretation techniques.

Both 5'UTR and 3'UTR models 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 3'UTR model performed slightly better than the 5'UTR one. Although no clear explanation for this fact was found, one could hypothetically attribute it to the fact that 3'UTR is a longer sequence on average, which possibly has more regulatory elements associating with mRNA half-life.

The fact that the validation loss during training was lower than the training loss for several initial epochs could indicate that the model didn't have enough predicting capability, in the form of the model architecture complexity or number of parameters. Such reason doesn't seem plausible as multiple networks with much higher number of parameters were optimized and lead to a similar validation loss behavior. Perhaps a more convincing reason comes from the possibility of the validation set distribution being more predictable than the training set, with, for example less outliers. Furthermore, the training epoch loss function is calculated as the average of all loss function values for each batch during training. If the loss presents high variations given the batch and due to optimization during training, then such can direct the overall epoch training loss into a higher value. In contrast, the epoch validation loss is calculated as the average of all validation batches' loss, for a model with immutable parameters in the end of the epoch. Although the validation loss is lower than the training loss during the first dozens of

epochs, the lowest validation loss is achieved when the training loss is lower.

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.

In the end, more candidate motifs obtained from TF-MoDISco should be investigated and, instead of using a fixed candidate motif sequence with the highest contributing bases, the specific scores of each base should be considered and scanned in the UTR sequence, producing a more accurate description of the presence of the candidate motif. Furthermore the development of a tool could be considered in order to automate the process of analyzing the candidate motif's properties that endorse the possibility of it being biologically significant, such as the motif's position distribution, the overall mRNA half-life association or its conservation across multiple species.

The model architecture and mRNA sequence representation also offer room for improvement. Transformer based architectures were found to be successful in producing sequences' representations or embeddings from sentences to proteins [38]. Such representations can be learned in an unsupervised way, which given the limited amount of half-life data, offers the prospect of increasing the predictive power of models using such representations.

Convolutional layers are capable of capturing sequence motifs through the distributed use of filters. However, the long-range relationship between motifs far apart in the sequence can be a constraint. The self attention layer conveys the possibility of capturing dependencies between motifs, regardless of their position in the sequence. Therefore, one possible model architecture could take use of the convolutional layers to detect motifs and subsequently self attention layers to evaluate relationships between them.

5.2 Modeling tissue-specific mRNA half-life variations

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, although the fact that the explained variance score is positive for all tissues indicates that the

model is a better predictor of tissue-specific mRNA half-life variations than the mean value. Most tissues had better performances than the corresponding linear models fitted only on codon content from section 4.3, although overall the increase in performance was small. This fact points to codon content as the major driver of the prediction of mRNA half-life variations between tissues. Furthermore, it indicates that the linear regression models are solid predictors of mRNA half-life variation with codon content alone despite not being able to capture eventual interrelationships between tissues.

Modeling tissue-specific mRNA half-life variations from sequence has never been done on previous research. The developed multi-task model sets a first baseline for these prediction tasks. 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 predicting tissue-specific mRNA levels.

5.3 A tissue-specific codon effect program

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, further validated through gene set enrichment analysis and the positive correlation between $\beta_{\text{codon}_k\,\text{Gal/Glu}}$ and codon's projection into codon signature axis. Such connection opens up new research directions and questions.

As presented in 2, 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.

Besides capturing translation, the codon signature metric can actually be encoding biological processes which are driving the changes in codon effects across tissues. The uncovered connection between codon signatures and mitochondria gene expression sheds light into such biological processes.

We now propose that changes in the amount and state of the resources available for translation differently influence the effect each codon has on decoding rate and mRNA half-life. Such resources consist of the energy molecules GTP/ATP and elongation factor proteins, all needed for translation elongation.

Mitochondria produce energy in the cell by making the energy molecule ATP, using oxygen and biomolecules such as derivatives of glucose. ATP is converted in the cell to GTP by molecules termed kinases. As measured by the amount of transcripts of mitochondrial genes and reported in [39], 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. The same can be said about GTP production and concentration, as it is in part dependent on ATP.

Elongation is the most energy demanding step of translation [40]. For the decoding of one codon two GTP molecules and two ATP molecules (hydrolysis of ATP to AMP) are required. ATP is used on the charging of an amino-acid to the tRNA. One GTP molecule is involved in tRNA delivery and recognition and the other involved in the translocation of the ribosome. As discussed in 2, the decoding time of a codon is limited, between other factors, by the amount of corresponding tRNAs. We argue that the GTP/ATP availability can be another 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.

Similar to the interpretation proposed for $\beta^{t_i}_{\text{codon}_k}$, the codon effects on half-life measured by $\beta_{\text{codon}_k \text{ half-life}}$ can convey information on the influence codon k has on its decoding rate/time. Comparatively to $\beta^{t_i}_{\text{codon}_k}$, $\beta_{\text{codon}_k \text{ half-life}}$ was obtained by predicting mRNA half-life and not mRNA half-life variation (Δlog(mRNA half-life)). This means that $\beta_{\text{codon}_k \text{ half-life}}$ captures the effect of each codon on mRNA half-life, for mRNAs under the same condition - the same cell-line. Because the condition is the same, we should expect both GTP/ATP availability, and the quantity and state of elongation factor proteins to be constant between translated mRNAs. This leaves us with tRNA concentration as the main translation-related factor capable of explaining the differences of $\beta_{\text{codon}_k \text{ half-life}}$ between codons.

The way GTP/ATP availability can influence each codon differently is still not clear, however our results point in one possible direction. The fact that $\beta_{\text{codon}_k \text{ half-life}}$ is positively correlated with the codon's projection into the codon signature axes, can indicate that the decoding is more determined by tRNA availability in higher energy production tissues. Hypothetically, in these tissues, GTP and ATP would be highly available, which would possibly change their effect as translation rate determiners.

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 [41]. The older the individual gets, the lesser the quality of the mitochondria and therefore its energy production capability. Indeed, the codon signature decreases with age.

The RNA integrity number can convey the overall degradation state of mRNAs in an RNA-seq sample [42]. As previously mentioned, the translation rate of an mRNA is associated with its half-life. In particular, higher translation rates associate with higher half-lives, or decreased mRNA degradation. The observed positive association between RNA integrity number and codon signature may indicate that the translation of mRNAs happens at higher rates for tissues with higher energy production. In those tissues, the more available ATP and GTP molecules would increase the overall codon decoding time, and tRNA availability could be a higher translation rate limiter factor.

In 2.1.5, we discussed how the distribution of codons in the genome can be a product of natural selection and facilitate the expression of certain groups of genes relative to others by enhancing their mRNA's half-life/stability and translation rate. Furthermore, we discussed how and why the effect of this codon distribution can vary between cell condition/state.

From the research presented on this thesis, a new exciting possibility arises. The possibility that codon effects on mRNA stability and possibly translation rate observed across tissues are driven in part by energy production. This suggests that, depending on the cell's energy production, a specific set of codons will have an enhanced positive/negative impact on mRNA stability 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 results raise the 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.

Bibliography

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

[2] "Stages of translation," https://www.khanacademy.org/science/biology/gene-expression-central-dogma/translation-polypeptides/a/the-stages-of-translation, accessed: 2020-11-09.

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

[4] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[5] B. Bolognesi and B. Lehner, "Protein overexpression: reaching the limit," *eLife*, vol. 7, p. e39804, 2018.

[6] C. Miller, B. Schwalb, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken *et al.*, "Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast," *Molecular systems biology*, vol. 7, no. 1, p. 458, 2011.

[7] 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.

[8] 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.

[9] A. Lugowski, B. Nicholson, and O. S. Rissland, "Determining mRNA half-lives on a transcriptome-wide scale," *Methods*, vol. 137, pp. 90–98, 2018.

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

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

[12] A. G. Hinnebusch, "The scanning mechanism of eukaryotic translation initiation," *Annual review of biochemistry*, vol. 83, pp. 779–812, 2014.

[13] N. L. Garneau, J. Wilusz, and C. J. Wilusz, "The highways and byways of mRNA decay," *Nature reviews Molecular cell biology*, vol. 8, no. 2, pp. 113–126, 2007.

[14] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *FEBS letters*, vol. 582, no. 14, pp. 1977–1986, 2008.

[15] D. P. Bartel, "Metazoan micrornas," *Cell*, vol. 173, no. 1, pp. 20–51, 2018.

[16] W. Olivas and R. Parker, "The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast," *The EMBO journal*, vol. 19, no. 23, pp. 6602–6611, 2000.

[17] G. Hanson and J. Coller, "Codon optimality, bias and usage in translation and mRNA decay," *Nature reviews Molecular cell biology*, vol. 19, no. 1, pp. 20–30, 2018.

[18] S. Pechmann and J. Frydman, "Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding," *Nature structural & molecular biology*, vol. 20, no. 2, p. 237, 2013.

[19] 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.

[20] R. Buschauer, Y. Matsuo, T. Sugiyama, Y.-H. Chen, N. Alhusaini, T. Sweet, K. Ikeuchi, J. Cheng, Y. Matsuki, R. Nobuta *et al.*, "The Ccr4-Not complex monitors the translating ribosome for codon optimality," *Science*, vol. 368, no. 6488, 2020.

[21] T. E. Quax, N. J. Claassens, D. Söll, and J. van der Oost, "Codon bias as a means to fine-tune gene expression," *Molecular cell*, vol. 59, no. 2, pp. 149–161, 2015.

[22] R. S. Dhindsa, B. R. Copeland, A. M. Mustoe, and D. B. Goldstein, "Natural selection shapes codon usage in the human genome," *The American Journal of Human Genetics*, 2020.

[23] J. Li, J. Zhou, Y. Wu, S. Yang, and D. Tian, "GC-content of synonymous codons profoundly influences amino acid usage," *G3: Genes, Genomes, Genetics*, vol. 5, no. 10, pp. 2027–2036, 2015.

[24] Y. Xu, P. Ma, P. Shah, A. Rokas, Y. Liu, and C. H. Johnson, "Non-optimal codon usage is a mechanism to achieve circadian clock conditionality," *Nature*, vol. 495, no. 7439, pp. 116–120, 2013.

[25] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. S. Sachs, and Y. Liu, "Non-optimal codon usage affects expression, structure and function of clock protein FRQ," *Nature*, vol. 495, no. 7439, pp. 111–115, 2013.

[26] J. C. Guimaraes, N. Mittal, A. Gnann, D. Jedlinski, A. Riba, K. Buczak, A. Schmidt, and M. Zavolan, "A rare codon-based translational program of cell proliferation," *Genome biology*, vol. 21, no. 1, pp. 1–20, 2020.

[27] H. Gingold, D. Tehler, N. R. Christoffersen, M. M. Nielsen, F. Asmar, S. M. Kooistra, N. S. Christophersen, L. L. Christensen, M. Borre, K. D. Sørensen *et al.*, "A dual program for translation regulation in cellular proliferation and differentiation," *Cell*, vol. 158, no. 6, pp. 1281–1292, 2014.

[28] 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.

[29] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.

[30] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, pp. 2951–2959, 2012.

[31] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.

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

[33] 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.

[34] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[35] 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.

[36] L. D. Marroquin, J. Hynes, J. A. Dykens, J. D. Jamieson, and Y. Will, "Circumventing the Crabtree effect: replacing media glucose with galactose increases susceptibility of HepG2 cells to mitochondrial toxicants," *Toxicological Sciences*, vol. 97, no. 2, pp. 539–547, 2007.

[37] K. Leppek, R. Das, and M. Barna, "Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them," *Nature reviews Molecular cell biology*, vol. 19, no. 3, p. 158, 2018.

[38] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, "Bertology meets biology: Interpreting attention in protein language models," *arXiv preprint arXiv:2006.15222*, 2020.

[39] 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.

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

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

[42] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, "The RIN: an RNA integrity number for assigning integrity values to RNA measurements," *BMC molecular biology*, vol. 7, no. 1, pp. 1–14, 2006.

A

Supplemental figures

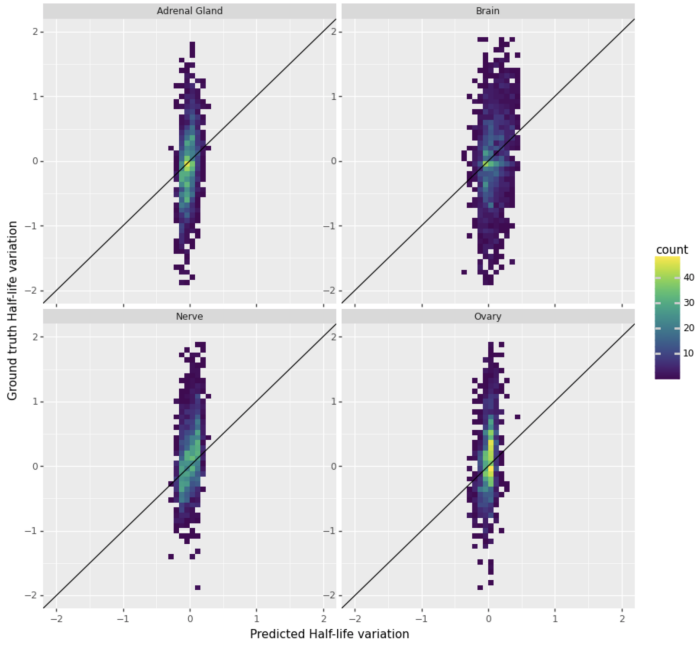


Figure A.1: Ground truth vs Prediction for the Multi-task DNN model - top 4 tissues.

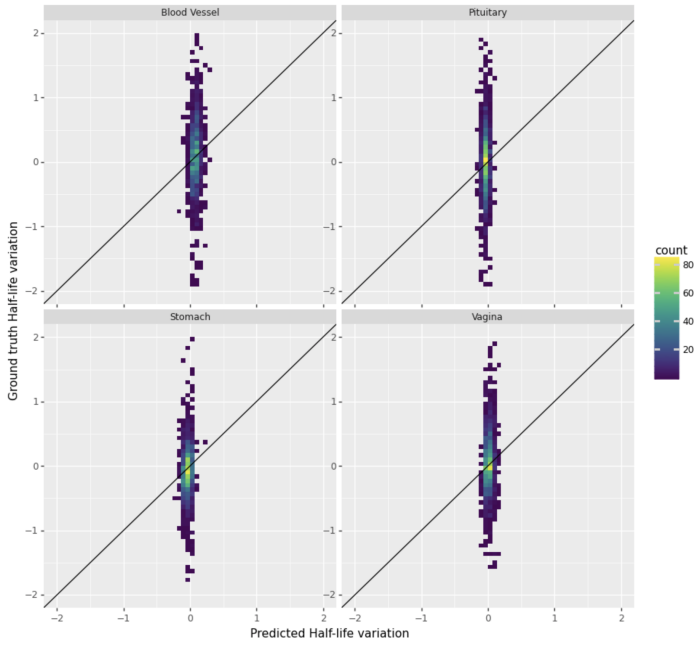


Figure A.2: Ground truth vs Prediction for the Multi-task DNN model - worst 4 tissues.

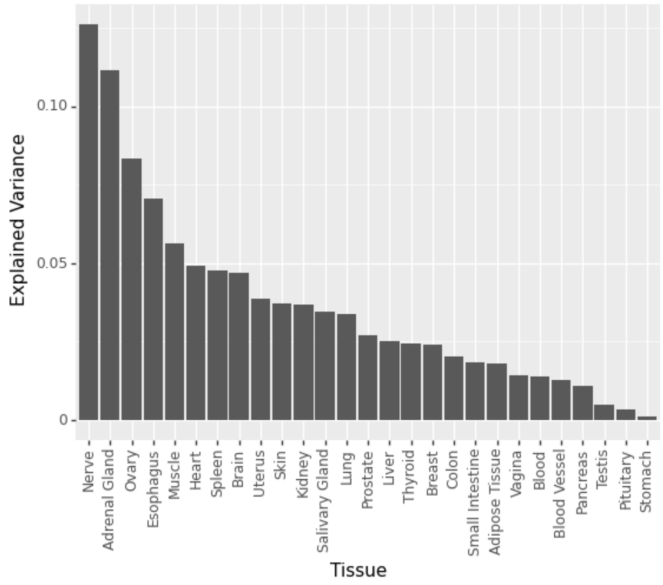


Figure A.3: Explained variance for each tissue's linear regression model.

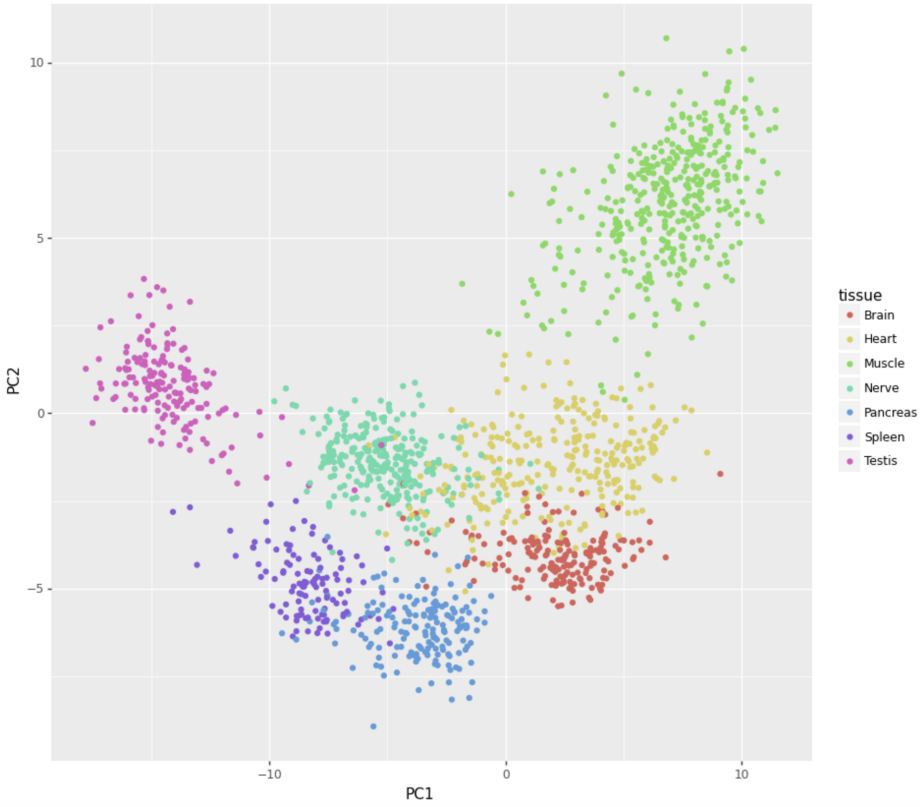


Figure A.4: Individual samples in the PC space as obtained by the analysis of the average human individual tissue (4.3.1.A).

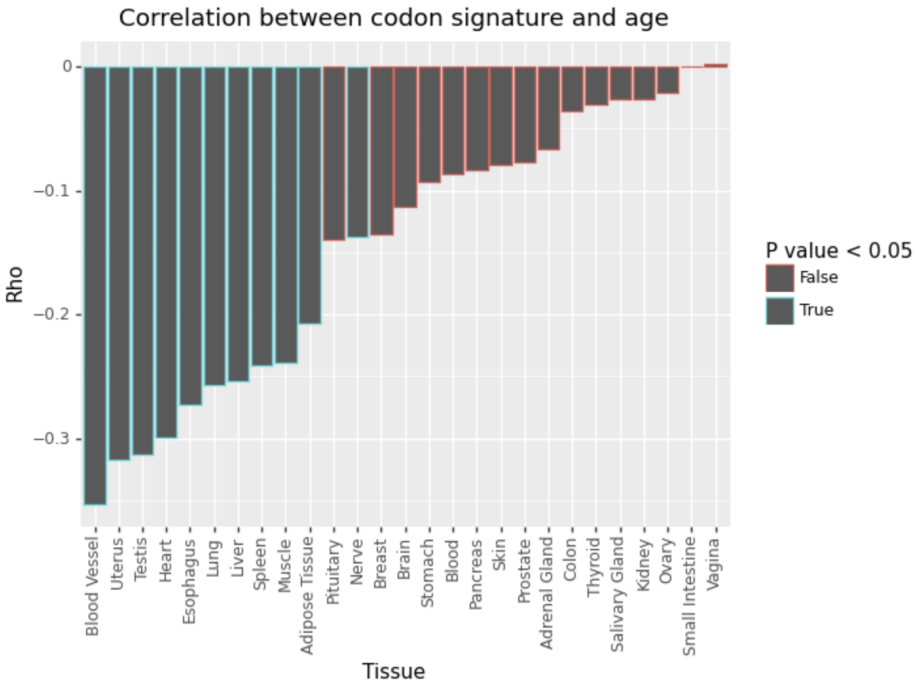


Figure A.5: Spearman correlation coefficient between age and codon signature per group of samples belonging to one tissue.

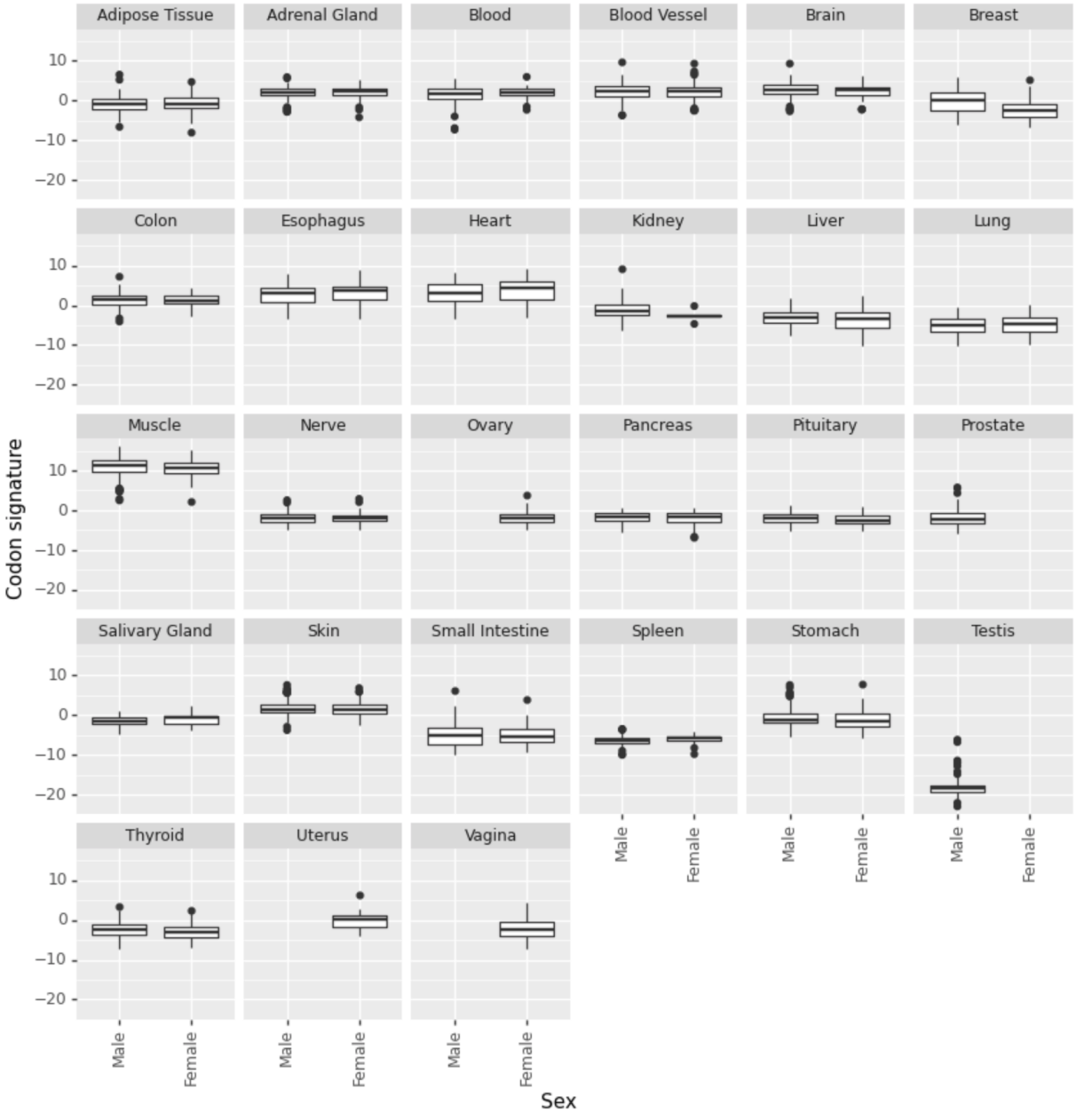


Figure A.6: Boxplot showing the distribution of codon signatures per tissue for male and female samples.

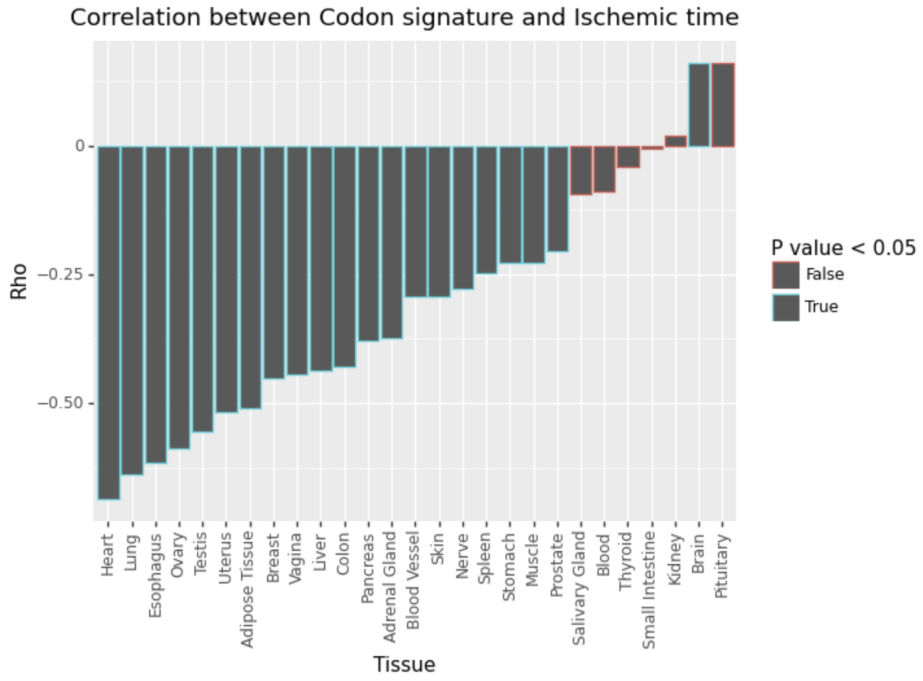


Figure A.7: Spearman correlation coefficient between ischemic time and codon signature per group of samples belonging to one tissue.

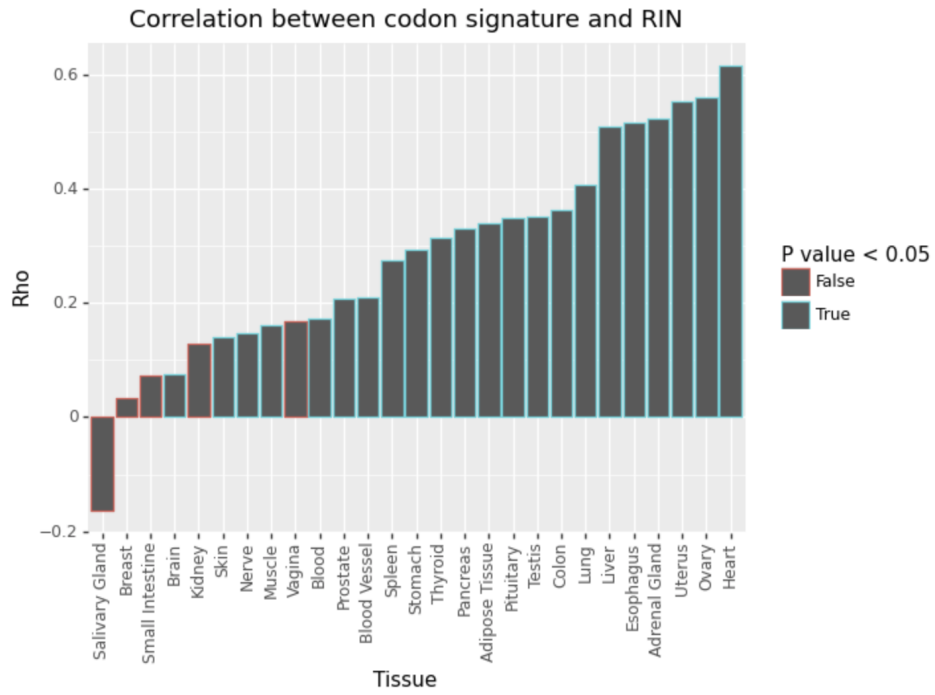


Figure A.8: Spearman correlation coefficient between RIN and codon signature per group of samples belonging to one tissue.

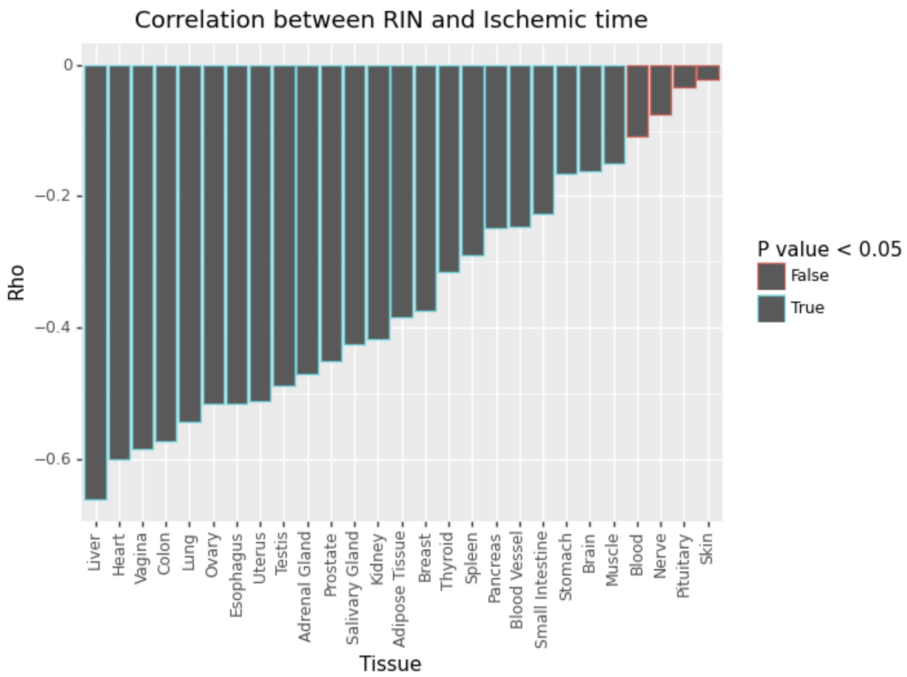


Figure A.9: Spearman correlation coefficient between RIN and Ischemic time per group of samples belonging to one tissue.

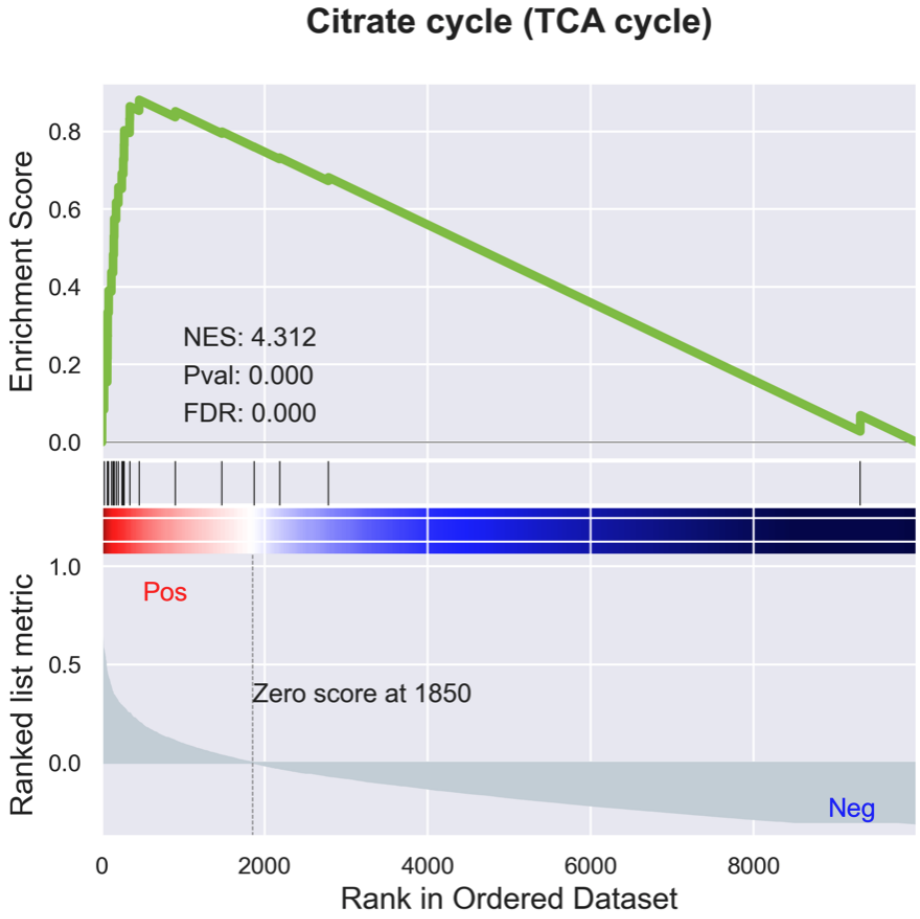


Figure A.10: Gene set enrichment analysis results for the TCA (Krebs cycle) pathway. The black stripes correspond to the position in the ranking of the genes composing the gene set of the TCA cycle pathway. The ranked list metric axis is the spearman correlation between transcript abundance and codon signature discussed previously. The shape of the curve on the enrichment score axis indicates that the gene set involved in the TCA cycle pathway shows an enrichment for the first end of the ranking. NES stands for normalized enrichment score, Pval stands for p-value and FDR stands for false discovery rate. For more details on the GSEA algorithm see [4].