An Atlas of the Human uORFome and its Regulation across Tissues

Håkon Tjeldnes, author

Eivind Valen, supervisor

University of Bergen
Department of Informatics

Abstract

Upstream open reading frames (uORFs) are in frame start and stop codons starting in the 5' leader of mRNAs. They have been found to regulate gene expression, primarily through translational inhibition by hindering ribosomes from reaching the protein coding ORF. Initial estimates concluded that almost half of the genes in the human genome contain uORFs, and studies have shown that uORF mediated mis-regulation can lead to health issues and disease. While some efforts have been made towards annotating uORFs, a comprehensive annotation of uORFs across the transcriptome and its regulation across tissues is lacking.

This thesis presents methods for large scale detection of uORFs based on experimental and sequence-based data, and presents an atlas of the human uORFs and their use and regulation across >1000 samples in 122 tissues. uORFs need to be translated to act as regulators and from an initial population of > 2 million candidates, my method identifies 21,766 uORFs as actively translated. Collectively, they show a strong bias towards a ATG/CTG start codon and disfavour codons known to disfavour translation, indicating that my method produces predictions of high accuracy.

Acknowledgments

I want to express my gratitude to my supervisor Eivind Valen for his support and making my ideas clear. I want to thank my advisor Gunnar Schulze for answering all my strange questions. I also want to thank Kornel Labun for teaching me the ways of package development. I want to thank Adam Giess for all his good ideas and Katarzyna “Kasia” Chyżyńska for helping me to prepare the data. The rest of the Valen group also deserves my gratitude for discussing with me the biological aspects of uORFs. I want to thank Martin Morgan from the Bioconductor team, for explaining me the inner workings of the R programming language. Lastly, I want to thank my girlfriend Marianne Espelid, for her love and support.

For all who might read this, sincerely Håkon Tjeldnes.

Contents:

Introduction

The central dogma of molecular biology describes the flow of genetic information between three main states; DNA, RNA and protein. The DNA sequences contain the genetic information which is copied into RNA molecules in a process called transcription. These RNA molecules then serve as templates for the assembly of amino acids into proteins - in a process called translation. This thesis will focus on the latter of these steps, translation, and how specific regulators in mRNAs called uORFs can regulate this process. I will show how these uORFs can be identified, catalogued and how to predict differential usage in human tissues. The results will be validated by correlating to known biological features and compared with a small set of experimentally verified uORFs. The resulting pipeline can be used to predict uORFs in other species or on additional human tissues in the future.

I will first give a general introduction to the biology related to and of uORFs, followed by the methods and results constituting the work of this thesis.

From RNA to Protein

Ribonucleic acid (RNA) is the product of DNA transcription and, like DNA, consists of nucleic acids, which are: A, G, C and U (corresponding to T in DNA). One of RNAs important functions is to serve as copies of genomic regions that can then be translated into proteins. RNA can also act as active regulatory molecules and as building blocks in structures like ribosomes [1].

The first of these functions, as templates for translation, is accomplished through messenger RNAs (mRNA), which contain protein encoding regions. These regions can be read in a process known as translation to convert an mRNA nucleotide sequence into its corresponding protein sequence. Proteins consist of amino acids, of which several hundred naturally occur, but only 20 are in the genetic code [2]. The genetic code describes which RNA bases encode which amino acids. These codes are written in triplets of nucleotides, called codons, each of which specifies a single amino acid. E.g. the triplet AGA codes for the amino acid Serine. Some amino acids are encoded by multiple codons making the genetic

code partially redundant. So for instance, in addition to AGA, Serine is also encoded by AGG.

The CDS is typically the longest stretch of an mRNA that starts with a start codon (AUG, CUG and others), and ends with a stop codon (UUG, UAG and UAA) in the same 3 nucleotide frame. In general, regardless of whether these stretches encode proteins, they are known as open reading frames (ORFs). Since codons consist of triplets, the length of these ORFs can only be multiples of 3.

The canonical model of translation states that mRNAs consists of three main parts, from start to end; the 5' leader sequence, (also called 5' untranslated region or 5' UTR), the coding sequence (CDS) and the 3' trailer (also called 3' UTR). The ribosome initiates translation by first binding to the 5' cap at the start of the 5' leader of the mRNA, it then moves downstream, scanning until it finds a special codon, known as a start codon, which is the start of the CDS region. The genomic location of the start codon is called the translation initiation sites (TIS), since this is where translation is initiated. These TIS' are often surrounded by a favourable sequence called the Kozak sequence which helps to facilitate the initiation [3]. Upon recognition of a start codon, the ribosome will assemble into an elongation complex and start translation. During translation the ribosome will produce the CDS-encoded peptide. This will continue until a special stop codon is reached [4], whereupon the ribosome is released. The end of the mRNA, the 3' trailer, is typically not reached by ribosomes. It contains many regulatory signals and ends in a stretch of nucleobase adenine (A) - referred to as the poly-A tail - which protects the mRNA from degradation and is used in regulation [5]

Since the CDS encodes the primary translation product in the mRNA, the other parts of the mRNA have been called untranslated regions. However, 5' leader sequences can also code for peptide products [6], and these regions are the topic of my thesis.

Upstream open reading frames

While traversing the 5' leader, the ribosome may encounter sequences that closely resemble translation initiation sites of CDS'. Such sites are said to be upstream open reading frames (uORFs), since they begin upstream of the main ORF, see figure 1 for examples. Recent experimental techniques estimate that 20-50% of all transcripts contain uORFs [7].

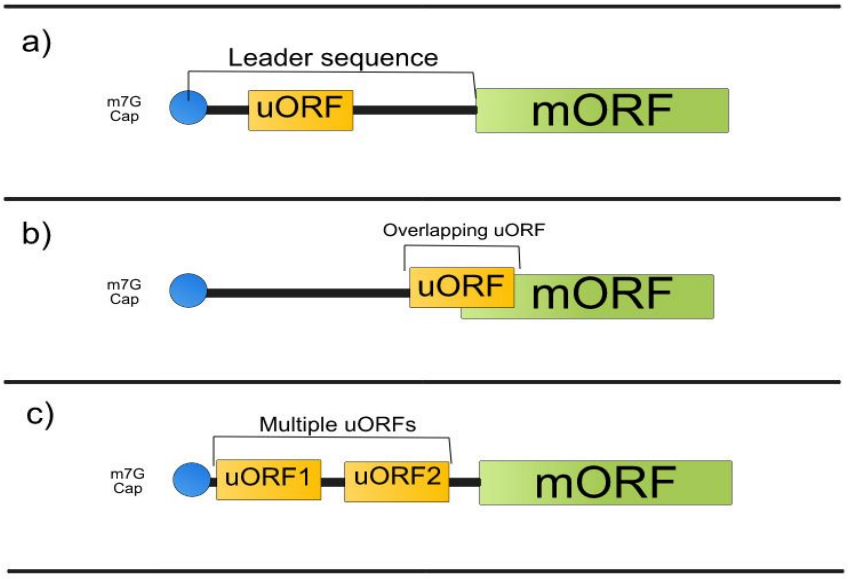

Figure 1: a) *Overview of a mRNA containing a uORF. b) Example of uORF overlapping the mORF (CDS). c) Example of mRNA with 2 uORFs. The m7G cap (in blue) on the 5' end protects the mRNA from degradation and is involved in regulation.*

Certain uORFs are thought to have an effect as a regulator for the CDS. These uORFs are called functional uORFs. While there exist examples of uORFs that produce small proteins or increase translation of the downstream mORF, the main mechanism of most functional uORFs, is to repress expression of the CDS through evicting ribosomes that would otherwise translate the CDS. This occurs when a scanning ribosome recognizes the uORF start codon as a site of translation initiation, leading to translation of the uORF. Since the common mode of action is to release ribosomes after translation, once the ribosome reaches the uORF stop codon it will often disassemble and be removed from the mRNA. In this way the protein product of the CDS will not be created. The uORF may act as a on/off switch for the CDS translation.

The ribosome is sometimes not disassembled, but regains its scanning capabilities and therefore continues scanning past the uORF. Whether the main CDS can be initiated after translating a uORF is determined by multiple factors. If the distance between the stop site of the uORF and the CDS is too short, this can repress the CDS expression. The ribosome will be unable to re-initiate on the CDS start, because the distance would not be enough to regain its ability to initiate translation. Studies have shown that changing this distance between the uORF and the CDS affects translation rate of the CDS[8]. The distance needed to re-initiate may depend on other factors like concentrations of regulatory molecules in the cell. This can control the required reinitiation time and distance, for translation to start again

[9]. The size of the uORFs also varies. As an example, in the ATF4 gene a uORF of only 12 bases (a start codon, two codons and then a stop codon) regulates the main ORF [10].

There are also other, less common, mechanisms for uORF functionality. If a uORFs is translated, it can lead to nonsense mediated decay (NMD), a process where the mRNA is degraded [11]. The NMD pathway uses the fact the mRNA in eukaryotes are spliced. By splicing the final mRNA does not contain all the bases of the DNA that was transcribed (called precursor mRNA), the removed parts are called introns and the remaining parts spliced together are called exons. On each gap between the exons, which are called exon-exon junctions, there are splicing proteins. These can act as a signal pathway. When the ribosome scan over these splicing proteins they are removed, such that if there are still splicing proteins left after ribosomes are halted, it means that not all exons were read by the ribosome. Many mRNAs have exon-exon junctions downstream of the uORF stop sites. In these cases, uORF translation can lead to NMD and reduced gene expression (if the downstream CDS remains fully or partially untranslated).

If the NMD pathway is not activated and the uORF stops before the CDS, the ribosome can reinitiate at the CDS start codon. On the other hand, if the stop codon of the uORF is inside the CDS as shown in figure 1 b), the ribosome cannot reinitiate, since it's already past the CDS start.

The number of uORFs in leader sequences have been shown to be less than what is to be expected by chance [6,12]. At the same time, an increased presence of uORFs in genes related to regulation, like transcription factors, have also been reported [6,12]. A comprehensive catalogue of human uORFs should therefore contain information about these different parameters and features to be able to predict whether the uORF is translated or not.

Next generation sequencing

Even though a possible uORF exists on the genomic sequence in proximity to the 5' leader it does not mean that it will be transcribed into mRNA. Some genes produce alternative products called splice-isoforms or are exclusively transcribed in certain tissues. That means certain uORFs only exist in tissues with the required variant of the 5' leader. To understand among others these variations between tissues, several experimental methods have been developed. In this project I made use of 3 next generation sequencing (NGS) technologies to identify variants of 5' leaders and their translation. By combining these, I could estimate accurately which uORFs are translated (active) in a given tissue or cell-type.

Most NGS technologies use the fact that different experimental methods retain certain regions of the DNA or RNA being investigated. These regions that are then mapped to locations on the genome are called reads. The number of reads mapped to a given position, provides information about how strong the signal is at that location. See figure 2 for typical read distributions of the relevant NGS technologies.

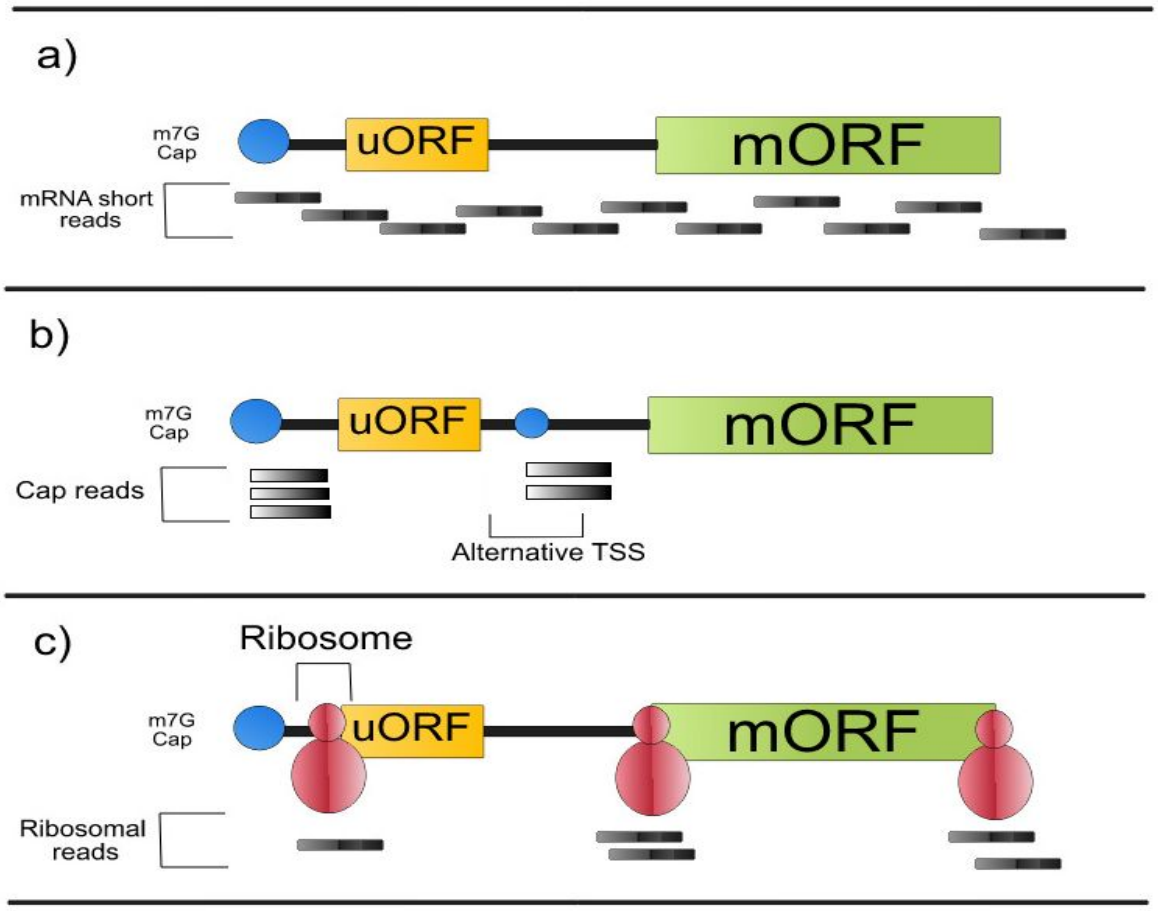

Figure 2: *Three NGS technologies and their typical read distributions a) RNA-sequencing, a method to catch small reads of RNA which can be aligned to the genome, this gives evidence for which mRNA isoforms are expressed in tissues. b) CAGE, a method to catch location of the 5' end of the mRNA transcript. They are a special case, since it is the start position of the reads that is important, not the whole read. c) Ribosome-sequencing, a method to identify locations of ribosomes on the transcripts, providing evidence of translation.*

RNA sequencing

RNA-sequencing (RNA-seq) is a NGS technology that gives a snapshot of the current mRNAs in the cell. Several experimental protocols exists, for both mRNA and total RNA levels in the cell. For mRNA this is done by removing the DNA in the cell by DNase, and then extracting the mRNA by their poly A tails, converting the RNA back to complementary DNA (cDNA), which is then sequenced and mapped back to the genome. This can be used to identify which mRNAs are present in the cell, as well as their relative abundance in a sample.

CAGE

RNA-seq experiments have some drawbacks, one of them is that it is not good at capturing the exact 5' end (the transcription start site) of the mRNA. A complementary sequencing

technology has therefore been developed called cap analysis gene expression (CAGE) to address this problem. CAGE extracts short reads of the 5' ends of the mRNA transcripts and sequences only these, providing a deep sampling of the transcription start site landscape of each sample. Allowing the discrimination of leaders with alternative 5' caps, that can occur when, for example, different tissues are regulated by different promoters.

For the study of uORFs, the usefulness of CAGE lies in the fact that some tissues will contain longer or shorter leaders than other tissues. Therefore different uORFs can potentially be found in those tissues. In this way CAGE enables us to map tissue specific uORF usage.

Ribosome profiling

To be able to identify which uORFs are actually translated, ribosome profiling can be used. Ribosome profiling (ribo-seq), is a NGS technology that provides a snapshot of the mRNAs that are currently being translated. The protocol is similar to the RNA-seq protocol, the difference is that all mRNAs are digested such that only positions where ribosomes are currently positioned are retained. These positions are called ribosome protected fragments.

Since the protocol takes some time to complete it is desirable to prevent the ribosomes from moving further along on the mRNA, for accurate determination of the boundaries of the translated region. To accomplish this, samples can be treated with chemicals such as cycloheximide, flash frozen or exposed to other treatments that halts translation.

While most mRNA fragments not protected by ribosomes are expected to be washed away during the protocol, RNA-binding proteins and RNA structural elements can produce the same protective environment as the translated mRNA and may be accidentally captured and sequenced [13]. Because of this, sophisticated methods are needed to separate signal from noise. These methods will be discussed in the Methods section.

The ribosome-derived fragments from ribosome profiling have lengths around 28-34, since that is the number of nucleic bases spanned by the ribosome. When aligned with a transcriptome reference (a map of genomic coordinates for mRNAs), the corresponding reads will span more nucleic bases than just the codon it is currently translating, although the ribosome is known to read one codon at a time. Finding this specific codon within the read of length 28-34 is called p-site shifting.

uORFs translation level and biological function

Most uORFs are thought to act by limiting the number of ribosomes reaching the CDS. So despite being translated it does not follow that the peptide resulting from uORF translation is biologically functional. Most uORF peptides are likely degraded with small effects on the cell, if any at all [14]. Many of the reported regulatory effects from uORFs have been linked with specific cellular conditions or events. For example the regulatory effect of uORFs the ATF4 gene are dependent on eIF4G concentration. It is therefore important to separate translation

rate and biological function. To predict translation is easier, since ribosome profiling (in combination with RNA-seq) allows quantification of the translation rate of a given RNA sequence. While regulatory functions that affects phenotypes may depend on more than just translation of the uORF. For the sake of clarity our definitions are as follows: 1) a *uORF* is simply a sequence delimited by a start codon and an in-frame stop codon beginning in the 5' leader; 2) a *translated uORF* is a uORF that exhibits feature profiles in ribo-seq and RNA-seq experiments similar to those observed over canonical coding sequences and 3) a *functional uORF* is a uORF that regulates translation leading to a measurable cellular phenotype. This gives a concise definition of the goal for this thesis, I want to find translated uORFs so that a potential set of functional uORFs can be experimentally validated from this set and show differential usage between tissues.

Examples of functional uORFs

Global uORF features have been investigated in several species. uORFs have on average been found to modestly repress the translation of the CDS, leading to a ~ 15 - 30 % down regulation [15]. The repressiveness also increases with the number of uORFs in the leader sequence. The distance between uORF ends and CDS starts, also correlates with reduced CDS translation rate [16]. It has been estimated that around half of all transcripts in the human genome contain translated uORFs [17], and only a small part of these have been experimentally validated. Most of these validated uORFs have been catalogued in the database uORFdb [18]. However, uORFdb contains only 166 uORFs from the human genome, and only a part of these are actually validated. The articles referenced in uORFdb mostly focus on specific uORFs, that have been extensively investigated, like the uORFs in the ATF4 gene and the BRCA1 gene.

The ATF4 (activating transcription factor 4), is a gene that codes for a transcriptional regulation product. It has three experimentally validated functional uORFs in its leader sequence [10], see figure 3. uORF 1 positively regulates the translation rate of the CDS, while uORFs 2 & 3 negatively regulate the CDS. Two different scenarios happen in translation of this transcript. In the first scenario, shown in figure 3 a), the ribosome scans to and translates uORF 1. The distance is then too short for the ribosome to reinitiate at uORFs 2 and 3, allowing translation of the CDS. In the second scenario, uORF 1 is skipped leading instead to translation of uORFs 2 and 3. Since uORF 3 overlaps the CDS, this downregulates the CDS expression. Under certain stress situations in the cell, the translation rate of uORF 1 is upregulated, this make the CDS translation rate increase. This is a complex example of how uORFs can regulate the CDS and contain uORFs that both up and downregulate the CDS, even though on average uORFs are found to downregulate the CDS.

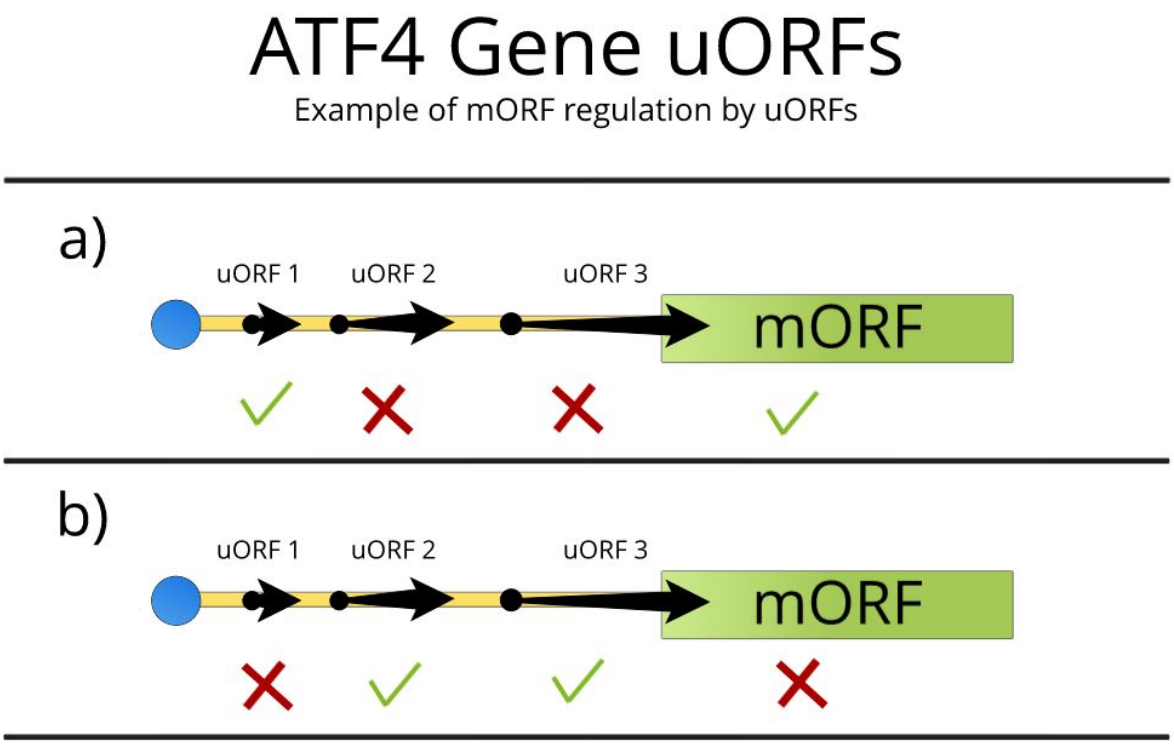

Figure 3: *Illustration of regulation by uORFs on the mRNA from the ATF4 gene. a) In the first scenario uORF 1 is translated, which leads to a too short distance for the ribosome to reinitiate at uORF 2 and 3. As a result the mORF is translated. b) In the second scenario uORF 1 is blocked, because of changes in eIFa concentration, uORF 2 and 3 are translated. The third uORF ends inside the mORF, which means the mORF will not be translated. Ensembl transcript ID: ENST00000396680.*

Prior work in detecting uORFs

Several studies have looked at specific uORFs or specific metrics like start codon, distances, and periodicity [19], to infer whether a given uORF is actively translated and potentially functional.

Most experimentally verified uORFs have ATG as their start codon. uORFs that do not have ATG as their start codon are called, non-ATG uORFs. Some of these have been shown to be evolutionary conserved between species and are estimated to be functional [19].

There are very few examples of similar efforts to what I wanted to achieve. The most recent example is an article from McGillivray *et al*.[20]. The article describes a prediction of uORFs in the human transcriptome using a small set of 3 ribo-seq datasets.

There are currently no accurate annotations for differential uORF usage in human tissues. This makes it problematic to prioritize candidate functional uORFs for experimental validation.

Aims of thesis

In this thesis I have constructed a comprehensive catalogue of human uORFs and provided a standardized way to classify and identify them. I have furthermore annotated their regulation across 1863 human transcriptomes in 122 tissues. The pipeline was optimized for fast and efficient searches in NGS libraries, allowing the catalogue to be easily extended in the face of new data. It has been included as an R package with the latest Bioconductor release and features well documented code, that can be reused and extended by others.

The primary challenge of the thesis was to find an efficient way of identifying translated, active uORFs and the biological features that distinguish them from other elements in 5' leaders. To accomplish this I catalogued all possible uORFs across all transcriptomes and based on this built a classifier to identify the translated uORFs based on experimental data. I therefore had two sets of uORFs, candidate uORFs identified purely by sequence, and uORFs predicted to be translated. First, I trained a classifier using experimental data for translation and transcription: ribo-seq and rna-seq. I used CDSs and 3'UTR regions to learn the features of translated ORFs based on these datasets. I then applied this model to obtain a high confidence set of translated uORFs. I used this high-confidence set to build a second classifier relying only on sequence features of these translated uORFs, which could be applied to all tissue-specific transcriptomes for which ribosome profiling data was not available. Finally, I validated my findings with experimentally verified uORFs and compared the final predictions of active uORFs with the ones reported by McGillivray *et al.*

The primary goals can be summarised as follows

1. Develop an efficient and easy to use tool for finding uORFs.
2. Make a database containing all identified uORFs .
3. Predict translated uORFs from features of CAGE, ribo-seq and RNA-seq.
4. Predict uORFs across all transcriptomes based on sequence features of (3)
5. Describe tissue variance of uORF usage in humans

Methods

The atlas presented in this thesis was made with different bioinformatics tools, several of which were made by me as part of the project. This chapter will describe the relevant steps that went into making these tools and the resulting atlas. I will focus on the relevant tools and methods used in a biological and computer science perspective.

The primary methodical steps were:

1. Preparing datasets (CAGE, ribo-seq and RNA-seq)
2. Reannotating the 5' leader sequences using CAGE.
3. Identifying putative uORFs in these new leaders.
4. Calculating metrics for all putative uORFs.
5. Training classifiers to identify the subset of translated/active uORFs.
6. Comparing my predictions to experimentally validated uORFs.

Overview of Datasets

Datasets used in this thesis are the following:

The genome reference for the human genome was: Genome Reference Consortium Human Build 38 (GRCh38 patch 79).

The transcriptome annotation (the mRNA locations on the genome) was the ensembl gene reference from GRCh38.

The CAGE libraries used were from the Fantom 5 projects (1863 files in total), a collection of of 164 cell lines, primary cells and tissues.

The ribo-seq libraries (103 experiments in total) were downloaded from different experiments listed in supplements.

The RNA-seq libraries (43 experiments in total) were downloaded from different experiments listed in supplements.

GEO accession numbers for experiments used, can be found in supplementary table 1.

Preparing datasets

The different datasets needed to be filtered and preprocessed, to remove data that were not usable for our purposes. CAGE experiments belong to different tissues, we filtered out all experiments that that did not have at least 2 experiments for its specific tissue-type.

For the ribo-seq and RNA-seq, only pairs of experiments that were from the same tissues, were kept. That means ribo-seq tissue types that did not match any RNA-seq experiment by tissue were filtered out.

I then created the table containing the data that would be used in the thesis.

For the CAGE data:

1. A table for the experiments with the experiment id, file location and tissue type.
2. For each CAGE experiments, filter out all reads that does not have a least one duplicate in the experiment. This was done to remove noise.

For ribo-seq and RNA-seq data:

1. A table for each experiment with experiment ID, file location, tissue type and treatment type (haringtonin, flash freezing and cycloheximide)
2. A matching table, where only ribo-seq and RNA-seq experiments with the same cell-line and treatment type are accepted. From the original 103 ribo-seq and 43 RNA-seq, I created valid 35 matching pairs of ribo-seq and RNA-seq.

If available, samples from cancerous and healthy cell lines for the same tissue were matched.

Finding uORFs

To find the uORFs, we developed an R / c++ package called ORFik, available on Bioconductor (https://bioconductor.org/packages/release/bioc/html/ORFik.html). The package features a pipeline that can be run in one function to: 1) reassign TSS' and annotate 5' leaders based on CAGE, 2) find all ORFs and 3) calculate ribo-seq and RNA-seq features from these. It can be run on any eukaryotic and even circular prokaryotic genomes. The inspiration for ORFik was to search for uORFs, but it could also be used to identify other ORF such as novel genes and micropeptides.

The procedure in ORFik to find uORFs is the following:

For each leader sequence in the transcriptome, change the TSS to the strongest CAGE-peak, see figure 4. The strongest CAGE-peak must be a position with more than 1 CAGE-read and a maximum of 1000 bases upstream of the original TSS. The peak can also be downstream of the original TSS, see TSS search space in figure 4.

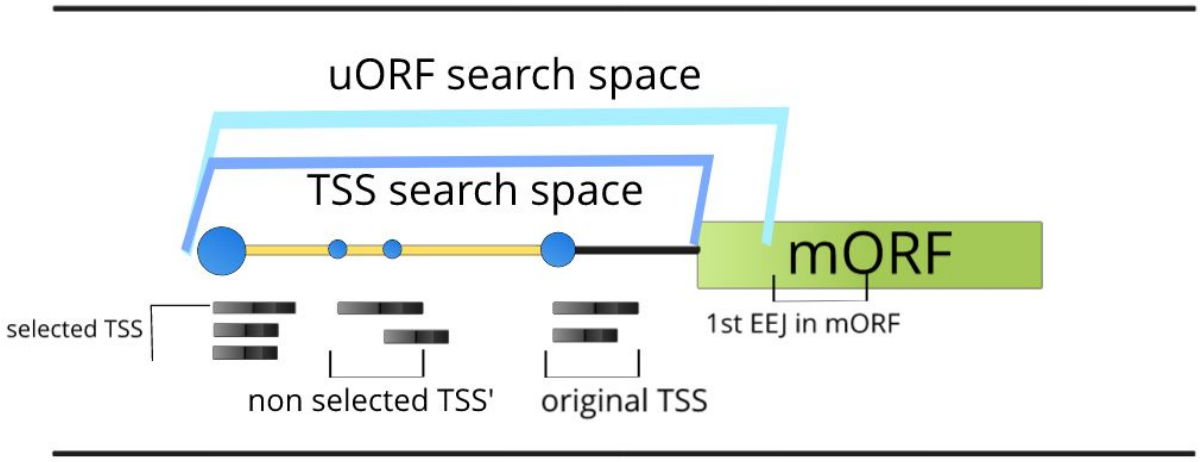

Figure 4: *The search space definition for uORFs and TSS selection in leader sequences. The new TSS is the position with most CAGE-reads (must be more than 1) in the search space. If no reads are present, the original TSS is used. EEJ is the exon-exon boundary in the mORF, the end point of the search space for uORFs. The search space of the tss spans 1000 bases upstream of the annotated TSS until the end of the original leader.*

The CAGE data used are from the FANTOM 5 project [21] whose purpose was to make a definitive atlas over all promoters in humans. The leaders are also extended downstream to include the first CDS exon, to allow uORFs to overlap the cds. Finding the reassigned leaders is done by the function reassignTSSbyCage() in ORFik. The function takes as input the Original 5' leaders and the CAGE file, together with filtering options specified above.

On these reassigned leaders, I searched for ORFs. This must be done efficiently, since there are so many leaders to search for uORFs in. A total of 1863 different CAGE libraries and 78423 transcripts with 5' leaders in the human genome. I made a c++ implementation of the Knuth–Morris–Pratt (KMP) string search algorithm. That indexes hit locations of start codons and stop codons, see figure 5.

A start codon was defined as the set of the start codon ATG and all 1 base variations:
{ATG, CTG, TTG, GTG, AAG, AGG, ACG, ATC, ATA, ATT}

Stop codons was defined as the set:
{TAA, TAG, TGA}

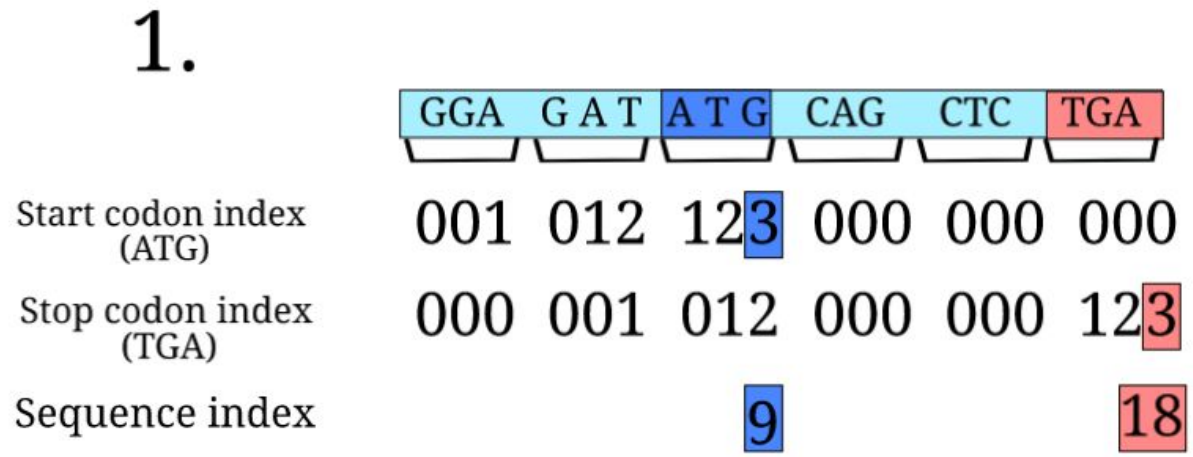

Figure 5: *Finding start and stop codons in fasta file by the KMP algorithm. The number of hits of the substring (start/stop codon) on the transcript is counted. When the value reaches 3, it means there is a full match. Seen as dark blue and red for start and stop codon. The start site in transcript coordinates is for this example 9-2 = 7, stop site is 18-2 = 16.*

All hits of start and stop codons on the transcripts were indexed, and pairs of start and stop codons were returned that were in frame (that is the modulus 3 distance between them is 0), and obey the rule that each start codon can only be used with its closest in frame stop codon downstream. Lastly the positions of the uORFs must be converted to genomic coordinates. For instance. in figure 5, the start codon is on position 9. This is relative to the transcript, the genomic coordinates most likely is something different.

DNA is double stranded, that means it consists of two paired strands containing the same complementary information. These two strands are represented as the positive strand and the negative strand. Therefore a mapping to genomic coordinates is done in opposite directions for the two strands. The example equation for a single exon leader sequence is shown under.

Sequence index: position of start codon given by KMP algorithm, see figure 5.
TSS: Genomic coordinate of TSS.

For mRNAs on positive strand:

$$Genomic\ coordinates = TSS + sequence\ index$$

For mRNAs on negative strand:

$$Genomic\ coordinates = TSS - sequence\ index$$

Finding uORFs is done with the function findMapORFs() in ORFik. The function findMapORFs takes as input the leaders FASTA sequences, together with start and stop codon definitions, and the minimum uORF length of 12 bases.

The function findMapORFs was computed for each of the 1863 leader definitions, found in the previous step. We made ORFik as a general tool used in ORF prediction. I therefore

made a pipeline wrapper around ORFik for running ORFik in parallel specifically for uORFs. The uORFome pipeline I made utilising ORFik, see figure 6, is available at: (https://github.com/Roleren/UORFome).

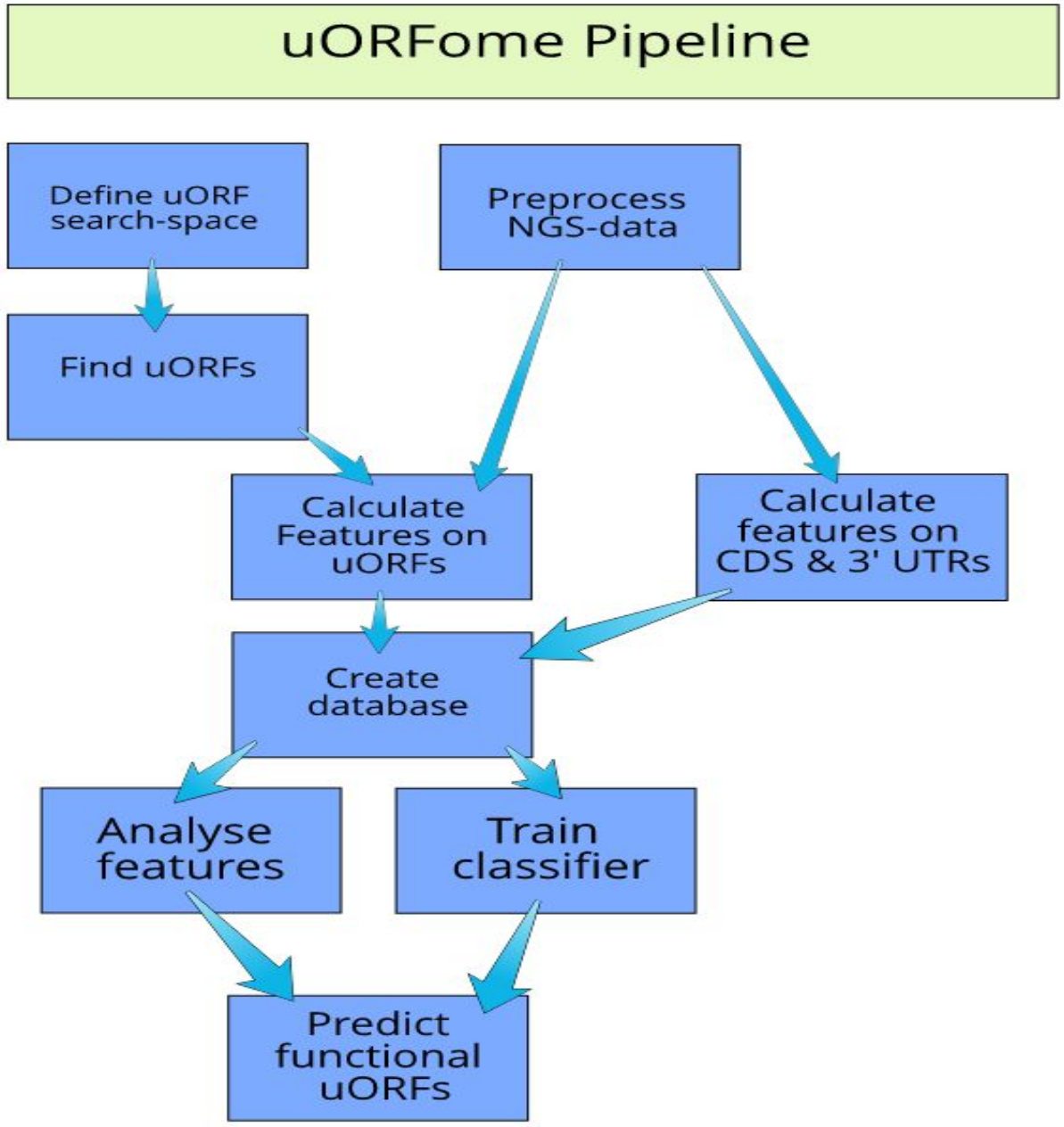

Figure 6: *The uORFome pipeline. A flowchart showing the general steps taken to create the uORFome atlas. The defined search space of the uORFs shown in figure 5, was searched for uORFs using ORFik. The processed data were used to calculate features on the uORFs. The features of the CDS and 3'UTRs are calculated the same way. A database was created with all the results. From the database I analysed the data for patterns and trained a classifier. I lastly predict functional uORFs on this classifier.*

Constructing a uORF Database

Since the size of the datasets returned is quite big, there was need for an efficient way of storing and querying the data. An SQLite database was created for this purpose. The database consists of several tables of data, which can extracted efficiently.

The 4 initial tables added into the database after finding the uORFs where:

1. A table of all the uORFs found and their genomic coordinates, chromosome and strand.
2. For each uORF: Ensembl transcript id for the transcript it came from, and Gene id from the gene the transcript came from.
3. For each CAGE experiment (1863 in total), which uORFs does the leader sequences contain. A logical (True/False) table.
4. For each tissue from CAGE data (122 in total), which uORFs does the leader sequences contain. A logical (True/False) table.

A tissue was defined to have support for a specific uORF if two experiments from that tissue contained the specific uORF. Out of the 164 tissues, 42 were rejected, because they only had 1 experiment (so they could never have two experiments supporting a uORF). Leaving a total of 122 tissues in the database. From this you can query for uORFs in specific tissues, or check which tissues a uORFs is supported by CAGE.

The primary key of the database is called the orf-identifier. It is a unique identifier that specifies an ORF.

It is of the syntax:
“chromosome, strand, start width”

An example from chromosome 1 with single exon uORF is:
chr1,-,9938553 33

For multiple exon uORFs a repetitive syntax is given for start and width per exon.
Each exon is split by a semicolon.
“chromosome, strand, start_1 width_1;start_2 width_2”

An example from chromosome 1, with two exons is:
chr1,-,112792 15;55281 396

All tables related to uORFs will use this primary key, so all the uORF tables have the same number of rows as in the primary key table. That is the total number of uORFs.

Calculating Features of uORFs

After the completion of the initial database, features and metrics could be extracted and calculated from the uORFs. I comprehensively catalogued and implemented all features that have been described in scientific articles having a possible relevance for ORF translation [22,23,24,25–29,30]. Since these are the features used in my model, description for all the features will be shown here.

Ribo-seq FPKM

FPKM (Fragments Per Kilobase per Million reads) is a normalization on the number of overlapping fragments per region of interest. For ribo-seq reads of uORFs this is the number of reads per uORF normalized to the length of the uORF and the ribo-seq library size. This is a measure of translation. [9]

nReads: number of reads

$$FPKMribo = (nReadsOrf \times 10\hat{}9) / (orfLength \times librarySize)$$

RNA-seq FPKM

This FPKM normalization uses the RNA-seq reads to find the number of overlaps per transcript. It is defined as number of reads per transcript normalized to the length of the transcript and the RNA-seq library size. This is a measure of RNA expression.

Tx: transcript

$$FPKMrna = (nReadsTx \times 10\hat{}9) / (orfLength \times librarySize)$$

Translation efficiency

Translational efficiency (TE) is normalization between ribo-seq FPKM and RNA-seq FPKM to find a more accurate value for the translation rate per uORF. Since mRNAs are transcribed at different rates, the ribo-seq fpkm does not explain how efficient a uORF is translated. To find a relative and less biased comparison, ribo-seq fpkm can be normalized by the RNA-seq fpkm, so that the relative transcription rate is included.

$$TE = FPKMribo / FPKMrna$$

ORFscore

Since the ribosome reads three bases at a time, the ribo-seq reads, which map the position of the ribosome, should have a higher accumulation in the frame of the ORF (frame 0), compared to the two other frames (frame 1 and frame 2). ORFscore is a function of the number of reads in each of the three frames.

$$frameTotal: (nReads0 + nReads1 + nReads2) / 3$$

$$frame0: (nReads0 - frameTotal)\hat{}2 / frameTotal$$
$$frame1: (nReads1 - frameTotal)\hat{}2 / frameTotal$$
$$frame2: (nReads2 - frameTotal)\hat{}2 / frameTotal$$

If: frame0 < frame1 or frame0 < frame 2

$$ORFscore = -log(frame0 + frame1 + frame2)$$

else:

$$ORFscore = log(frame0 + frame1 + frame2)$$

The ORFscore is negative if frame1 or frame2 have more reads than frame0.

Kozak Sequence score

The ribosome binds certain areas of the mRNA better than others, these strongly binding sequences called Kozak sequences can be compared between the start site region of each uORF, to understand how strongly the uORFs binds the ribosome. Using the experimentally verified reference Kozak sequence for human, represented as a position frequency matrix (PFM), I made a position weight matrix (PWM) used to score to ribosome binding strength of each uORF [29].

promoter: a string from the genomic alphabet [ATCGN] in the TIS region region of ORF. Given TIS position as position 0, the Kozak region is defined as {-9:-1, 3:5}. Start codon at {0:2} is excluded.

len: length of promoter, same as number of columns in the PFM.
[index]: index accessor of vector with promoter region
[row, column] index accessor of matrix, there are 5 rows, the genomic alphabet [ATCGN]

$$Kozak\ Score = \sum_{i=1}^{len} PWM(PFM[, i], promoter[i])$$

It should be noted that the start codon is not part of this Kozak sequence score, as that would bias the search towards ATG uORFs, since these have the highest Kozak scores. A new member of the DNA base set is added here, called N. Used in genome assemblies when the actual base is unknown.

Entropy

To see how the ribo-seq reads distribute over all the ORF codons, a read distribution entropy score is used. It calculates logarithmic variance over the distribution of reads. It provides

evidence for certain biases in read coverage. An examples would be all reads only overlapping the start codon.

codonSum: sum of reads per codon
N: sum of reads in ORF
len: length of ORF
cLen: number of codons, that is (len / 3).
i: sequence from (1 … cLen)

$$X = codonSum[i] / N$$

$$Hx = \sum_{i=1}^{cLen} X[i] \times log2(X[i])$$

$$Mx = 1 / cLen \times log2(1 / cLen)$$

$$entropy = Hx / Mx$$

Disengagement Score

Defined as the ratio of reads over the ORF, by reads over the remaining downstream part of the transcript.

Down stream is defined as the space: [ORFStop+1, TxStop]

$$disengagementScore = nReadsORF / nReadsTxDownStream$$

Inside Outside Score (IO score)

Defined as the ratio of reads over the ORF by reads on the rest of the transcript.

Outside is defined as the length: [TxStart, ORFStart-1] + [ORFStop+1, TxStop]

$$IO = nReadsORF / nReadsTxOutside$$

Ribosome release score (RRS)

To see how strongly the ribosome translate the ORF compared to the trailer, RRS can be used. It is defined as the number of ribo-seq reads over the ORF divided by the number of reads over the trailer of the transcript that contains the ORF. This is normalized by length. It is similar to disengagement score.

$$RRS = (nReadsORF / ORFlength) / (nReadsTrailer / trailerLength)$$

Ribosome stalling score (RSS)

To see how strongly the ribosome stalls on the uORF stop codon, RSS can be used. It is defined as the number of ribo-seq reads over the uORF stop codon divided by the number of reads over the whole uORF. This is normalized by lengths.

nReadsORFStopCodon: number of ribo-seq reads over the uORF stop codon.

$$RSS = (nReadsORFStopCodon/3)/(nReadsORF / ORFlength)$$

Distance to CDS

The distance between the stop site of the uORF and the start site of the CDS can not be calculated directly in genomic coordinates. Transcript coordinates is needed, to remove exon-intron boundaries.

stopORF: stop site of ORF in transcript coordinates
startCDS: start site of CDS in transcript coordinates

$$distToCds = startCDS - stopORF$$

The distance is negative if ORF stops inside the CDS.

Relative frame to CDS

The distance between the uORF and the CDS can be used to calculate the frame of the uORF compared to the CDS.
mod: modulus operation.

$$orfFrame = (distToCds - 1) \bmod 3$$

If orfFrame is 0, it means the ORF is in frame with the CDS.

Fraction length

To find the relative size of the uORF compared to the transcript it belongs to. It is defined as the ratio of the uORF length by the transcript length.

$$fracLength = uORFLength / TxLength$$

uORF rank order

A transcript can have several uORFs, and since the ribosome scans from 5' to 3', an ordering of the uORFs in the transcript can be created. The order in which the uORFs occur in the transcript can be computed by sorting the start sites of each uORF in the transcript. This is called the rank order of the uORFs.

All of these features were implemented into our package (ORFik).

For calculating the features a valid set of matching ribo-seq and RNA-seq experiments were used. These grouped to 5 tissues / cell types.
Tissues of ribo-seq and RNA-seq (35 matched pairs in):
{ Brain, kidney, fibroblast, prostate, Ovary }

The table of GEO matchings can be found in supplements.

The ribo-seq also needed to be pre processed, shifting the reads to the p-site, as described in the NGS section of the introduction. This was done using the program Shoelaces with automatic shifting of the different ribo-seq read lengths. The algorithm used by Shoelaces was integrated into ORFik as the function shiftFootprints(), so no installation of shoelaces is needed in future use.

Calculations for all ribo-seq, RNA-seq and sequence features were calculated for uORFs, CDS' and trailers. These were stored as tables in the database. Some features could not be calculated for CDS' and trailers, e.g. distance to CDS for the CDS is always 0. The primary key for accessing CDS and trailer tables was the ensembl id of the transcript it belonged to.

Prediction of Translated uORFs

To predict which uORFs are translated, I chose to make a pipeline of combining two random forest classifiers. Described in figure 7. The motivation for this choice was that I only had 5 tissues from ribo-seq (35 samples), while I had 122 CAGE tissues (1863 samples). I also did not want to bias the model towards sequence features like the start codon etc. By first creating a model for ribo-seq features I could ensure that I identified translated ORFs and could then extract sequence features from these that did not depend on ribo-seq. This enabled us to predict for any tissue as long as there is 5' leader annotation available.

Since the amount of validated uORFs in uORFdb is so small, I could not train the first random forest on uORFs. I therefore chose to use CDS' as positive set and trailers as negative set. Using the open source big data analysis package H2O, from H2O.ai, I trained a random forest with 8 of the CDS and trailers ribo-seq and RNA-seq features.

The features were:
{floss, TE, ORF-score, entropy, inside-outside score, disengagement score, ribosome-release score and ribosome-stalling score}.

The random forest prediction model was chosen to output regression instead of classification, to give more control on cut off between predicted translating / non-translating. The input data was split into 60% for training and 40% for validation, using 10 cross validations.

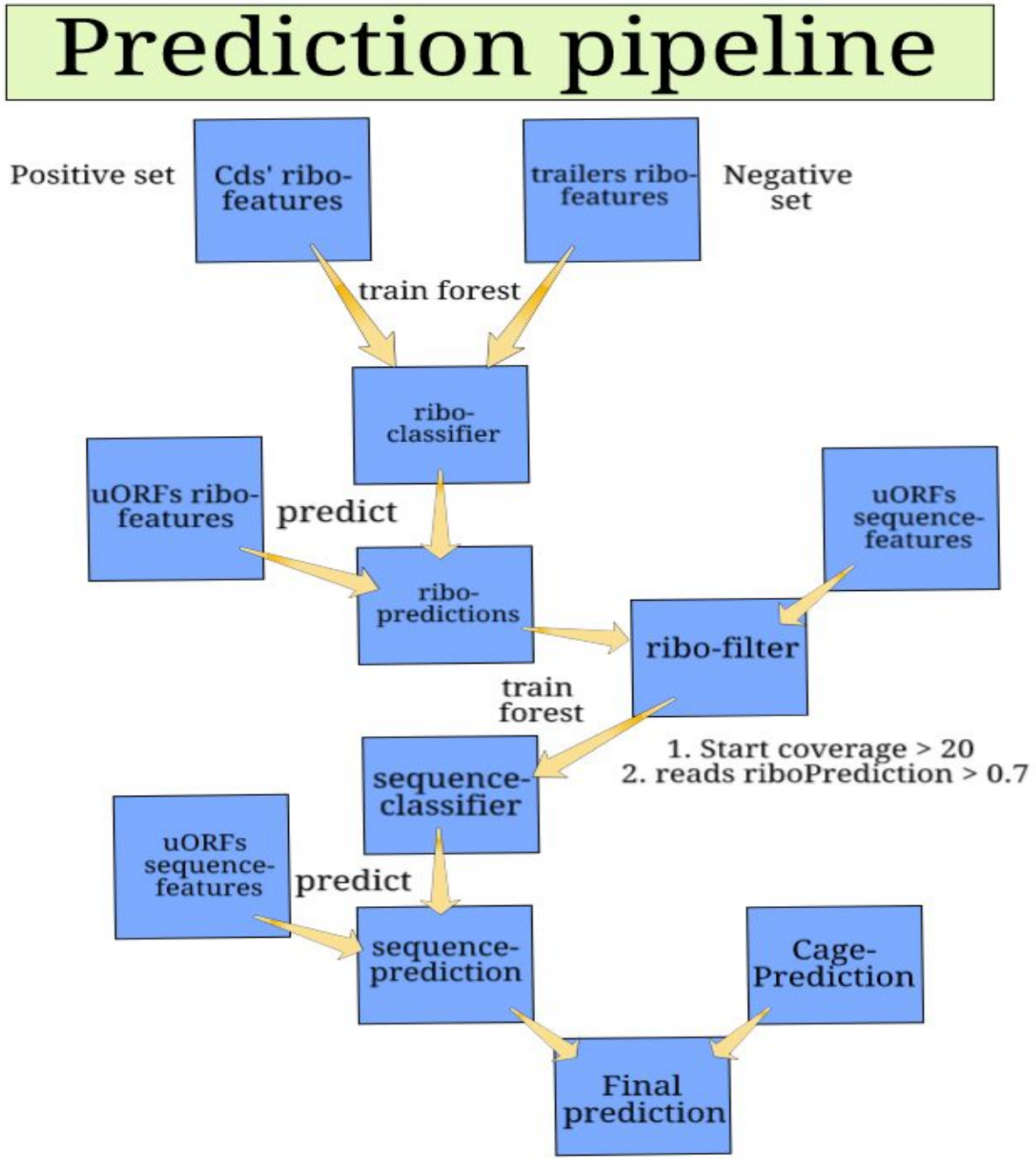

Figure 7: *Representation of prediction pipeline. This is a detailed model of the training and prediction steps in the uORFome pipeline shown in figure 6. First a Random forest model is trained from CDS' ribo-seq features as positive set and trailers ribo-seq features as negative set. Then the uORFs ribo-seq features are inserted into this classifier, and uORFs are predicted. These predictions are sent into the ribo-filter, that only accepts the uORFs that were predicted with more than 0.7 (70% certainty). Also the Ribo filter takes in the sequence features of the uORFs, the uORFs are grouped by stop codon position, such that all uORFs with same stop codon position are in the same group. One uORF per group is chosen, as the biggest with more than 20 reads on the start codon. A sequence classifier is trained on these as positive, and a random sampling as negative set. All candidate uORFs were then predicted on by their sequence features. Lastly CAGE-predictions are combined with this predicted to set give the resulting uORF prediction.*

I then predicted translated uORFs using the same 8 ribo-seq/RNA-seq features for uORFs instead of CDS' and trailers.

The second random forest was trained with the sequence features of the uORFs. I used the predictions from the ribo-seq model as input for the second random forest. A filter was added to reduce false positives. All uORFs was grouped by stop codon, so that all uORFs with the same stop codon were in the same group. If the group had a uORF that had more than 20 reads overlapping the start codon and the ribo-seq prediction had a probability of at least 70 %, it would be inserted in the positive training set. A random set 3 times bigger than the positive set was chosen from the remaining uORFs not in the positive training set.

The splitting and training parameters of the data was equal to the first.

By this model I predicted on all candidate uORFs, by their sequence features:
{ distance to CDS, length of uORF, fraction length, Kozak score, relative frame to CDS, if uORF overlaps CDS, uORF rank order}

A prediction for each of the 5 ribo-seq tissues was made, and stored in the database.
It is important to note that since I had 5 tissues of ribo-seq, 5 models were made. One for each ribo-seq tissue combined with sequence features. The final sequence model is made by combining these 5 tissue specific models into one. They were included by taking the union of the 5 sets.

model#: The sequence model by tissue

$$Final\ sequence\ model\ =\ model1\ \cup\ model2\ \cup\ model3\ \cup\ model4\ \cup\ model5$$

Lastly to get the final prediction, I combined the results from the sequence prediction and our earlier CAGE leader annotation (the logical tables of uORFs per Tissue), to give a final atlas of which uORFs were predicted in each of the 122 tissues. Since both tables were represented as logicals, a simple intersection operation gave the final prediction, one uORF-prediction per tissue.

$$Predicted\ uORFs\ =\ Final\ sequence\ model\ \cap\ uORF\ predicted\ by\ CAGE$$

Validating uORF prediction

As an example validation, I tested to see if the best experimentally verified uORFs from the literature was found by our prediction. To do this I accessed uORFdb, and found articles on experimental validation of uORFs in the ATF4 [10], the ABCC2 [31]and the ADH5 gene [32]. The reason I only choose three genes, was because I had to manually backtrack the original papers of the uORF predictions, analysing their results. There are no list of genomic coordinates of the uORFs in uORF db, only the number of uORFs per transcript, and most of the articles did not include genomic positions of the uORFs.

The procedure I followed was the same for all three genes, here are the steps explained for ATF4.

The ATF4 gene (Ensembl transcript id: ENST00000396680) has 3 well documented uORFs. One of them are only 6 base pairs long, so it would not be found by our uORF prediction, since it filters out all uORFs smaller than 12 base pairs. The article with the experiments on ATF4 was published in 2000, november. An older transcript annotation was used, so the leaders were not the same length as the current model (GRch38 patch 79). I then transferred the coordinates of the uORFs in the the original article to our coordinates (the older annotation had a 4 base pair extension in the beginning compared to new annotation). By comparing the start sites of the uORFs, as shown in figure 8. From this I could see if our database found any uORFs with hits on those locations for the two remaining uORFs in the ATF4 gene.

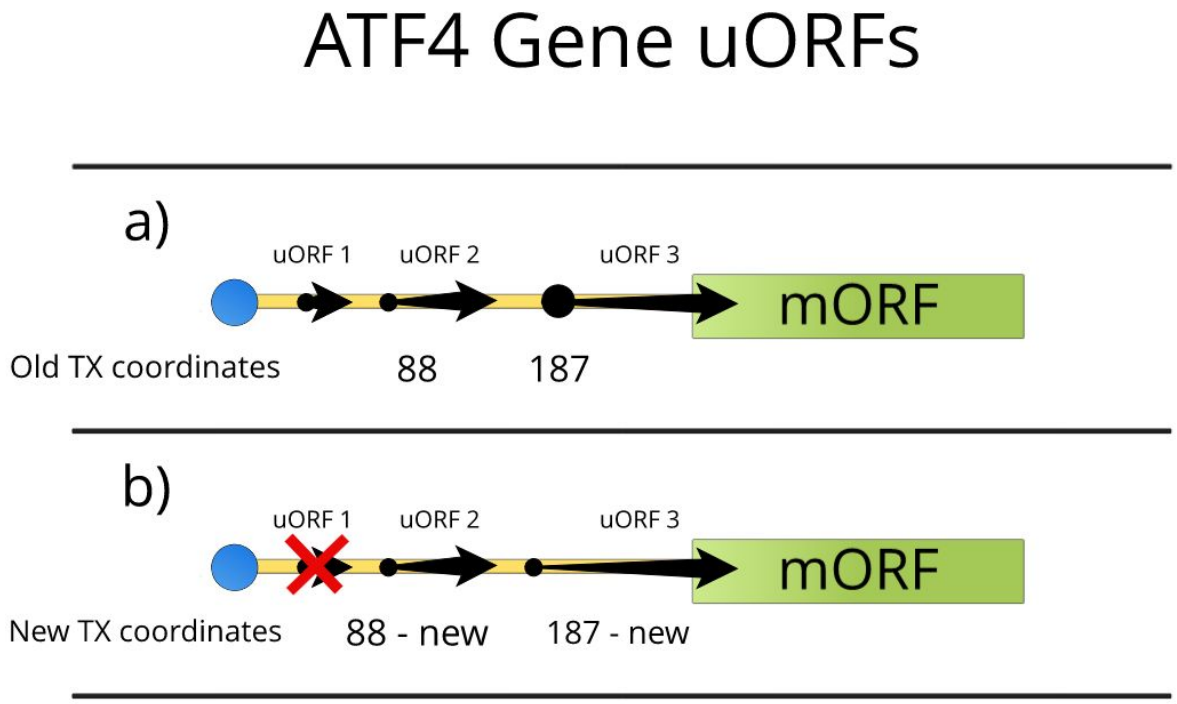

Figure 8: *Finding coordinates of uORFs in the ATF4 gene from old annotation to new annotation. a) old annotation by original experimental article. b) Our transferred coordinates for the new annotation. uORF 1 must be deleted, since it is only 6 base pairs long. For ATF4 the new transcript annotation was shifted by -4, so the transcript coordinates of uORF 2 for GRch38 annotation would be 88-4 = 84.*

We also tried to validate uORFs from the ABCC2 gene (transcript ENST00000370449) and the ADH5 gene (transcript ENST00000626055) the same way as ATF4, backtracking the articles to the original genomic positions of the uORFs.

Comparison of Predicted uORFs

Finally I did a comparison between our predicted translated uORFs and the predicted translated uORFs in the Bayesian classifier from McGillvary et al., explained in the introduction. I downloaded the supplements and found the table of predicted uORFs (supplements table 5 of top 10 % predicted uORFs). I then converted this to a bed file and used the ncbi assembly conversion tool: NCBI Genome Remapping Service, since they used an older assembly version. Conversion was done from hg19 to hg38. From the total 18880 uORFs in theirs list, 13040 uORFs were converted to our assembly. I then compared these uORFs to see how many were also in our prediction. McGillvary's pipeline is based on a much smaller dataset, but in the results and discussion section I will compare and discuss these.

Results

In this section I will present the findings of my uORF classifier and the differential regulation of uORFs across tissues. All the data visualized in the results are either stored in the database or modification based on that data, of which all the code is available.

The results will be divided in the following 3 subjects:

1. Predictions and annotation of translated uORFs
2. Results from 5' leader reannotation using CAGE data
3. Differential uORF usage across tissues

In addition I will show a comparison with the pipeline of McGillvary et al. and an analysis of the uORFs in the ATF4 gene.

Finding uORFs from leader sequences

Searching the human leader sequences for uORFs with the considered start codons (ATG and near-cognate ones), I found 2,242,885 unique potential uORFs. These will be called candidate uORFs, to separate them from our predicted set of active uORFs.
On average a leader sequence had 12 uORFs, and each gene had 115 uORFs. The number of candidate uORFs is much higher than expected from the literature, which is likely a result of the broad start codon definition of 10 possible start codons, in addition to including overlapping uORFs with the same stop codon. This set of 2,242,885 unique uORFs, represented the set of candidates for our prediction model.

Feature	Value
Number of unique candidate uORFs	2,242,885
Average uORFs per transcript	12 ∓ 27
Average uORFs per Gene	115 ∓ 212
Average uORF length	113 ∓ 132
Average number of exons per uORF	1.19 ∓ 0.45

Table 1: *Features of the set of all candidate uORFs. The number of unique uORFs, the average number of uORFs per transcript and gene is shown, together with average length and number of exons per uORF. It is important to remember these are the uORFs that are candidates for prediction of translation, not the prediction it self.*

Change in TSS for leader sequences and uORF usage

To verify that using CAGE gives variability in uORF usage in tissues, we did several validations. From the CAGE data there was 122 tissues with at least one replicate in the 1863 experiments. Using the 1863 CAGE-experiments the TSS of the 5' leader sequences were reassigned and stored in the database. By comparing to the original leader annotation, I checked how leaders were affected in each tissue.

Figure 9A) shows the number of candidate uORFs per tissue in the top 20 tissues with the most uORFs. The Turquoise part of the bars represent the set of uORFs that are present in all of the tissues. The red part of the bars represent the remaining uORFs found in the tissue. Each tissue contained on average 1,117,106 uORFs (48.5% of all candidate uORFs), and the maximum amount of uORFs was found in blood, which had a total of 1,594,398 uORFs (65.8 % of all candidate uORFs).

In figure 9 B) the changes in leader lengths over all tissues are shown. The figure shows how the leader size of the 78423 leaders in the human genome change on average over the 122 tissues. A positive value means on average over all tissues, this leader increases its length by CAGE reannotation. On average a leader increased by 16 bases with a standard deviation of 138 bases. The high standard deviation reflects the differential uORF usage in tissues.

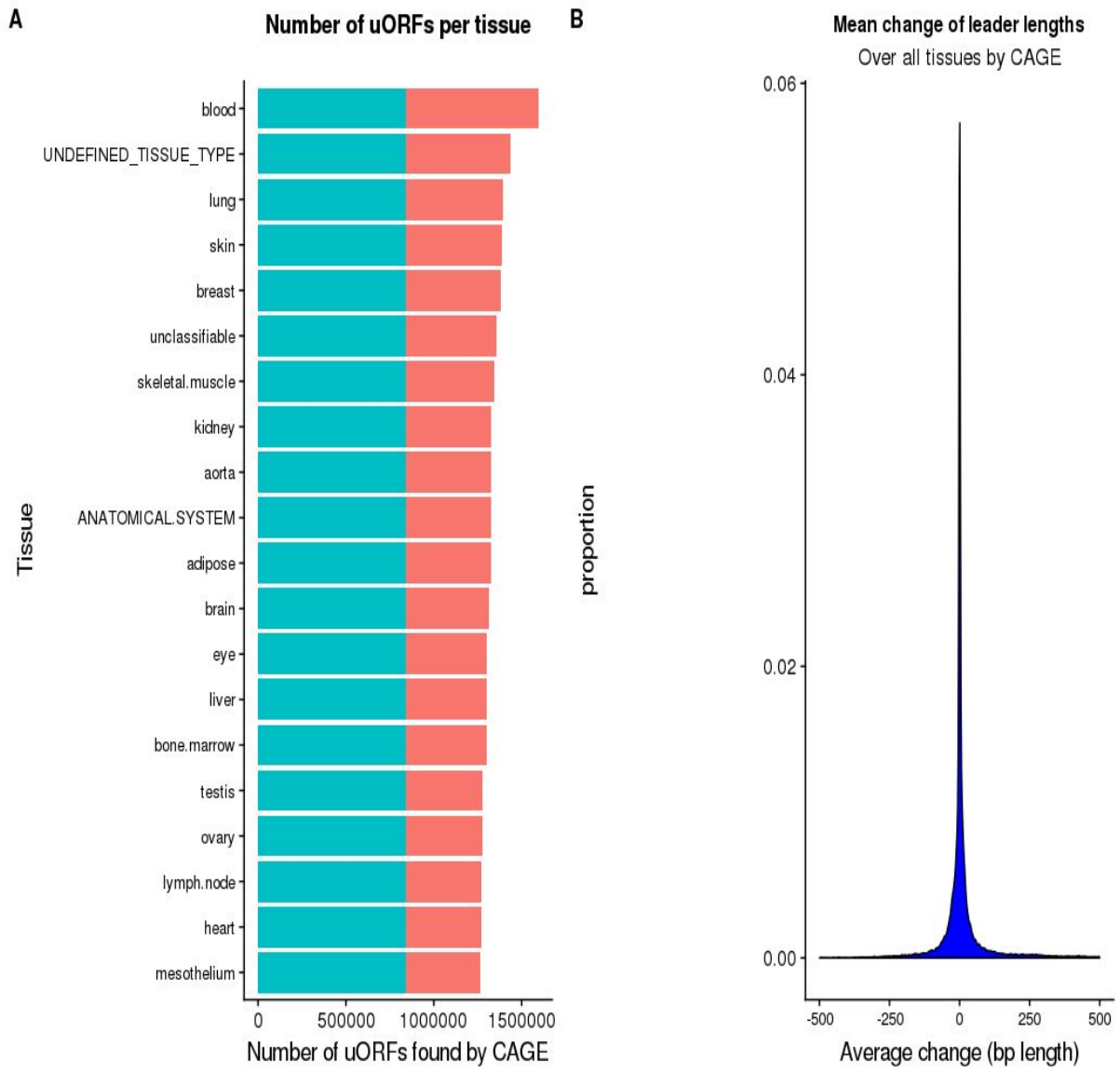

Figure 9: **A)** Number of uORFs by sequence per tissue. Figure showing top 20 tissues according to number of uORFs. Turquoise colour represents uORFs that are included in all tissues, while red represents the uORFs that varies between tissues. **B)** *Change of leader sequence lengths from CAGE data between tissues. X-axis is cut on -500,500 for clarity. A few leaders changed their length by as much as -10,000.*

The annotated leaders had an average length of 236 ∓ 264, and the CAGE leaders were on average extended by 16 bases upstream to an average leader length of 252 ∓ 297. Only an average of a 2855 leaders did not change its TSS per tissue. This shows that CAGE have a strong effect on the TSS position in the tissues.

I also tried to analyse read depth per CAGE library and the number of reads on the genomic locations that were assigned as the new TSS. The libraries contained on average 251,362 reads that must be distributed for the 78k leaders in the human genome. The genomic position of the newly assigned TSS had an average of 19 reads, (the median was 5). See table 2.

Feature	Value
Average change in leader length from CAGE	16 ∓ 138 (6.7 %) increase in length
CAGE-reads on new TSS (quantile summary)	**Min. 1st Qu. Median Mean 3rd Qu. Max.** 2 3 5 19 10 6,625
Average Transcripts affected by CAGE	75,568 (96.3 %)
Average Genes affected by CAGE	21,215 (81.5 %)
Average number of reads per CAGE library	251,362

Table 2: *Modifications of leader sequences by CAGE data. Features are average per CAGE-experiment. By affected, it is meant that the TSS is reassigned.*

Predicting translated uORFs

The previous section describes the result from cataloguing all candidate uORFs, but only a small subsets of these are translated and 'active'. The final prediction was made of a combination of two random forest classifiers. The results from each step will be shown in the details below.

Ribo-seq model:

In the ribo-seq classifier, based on CDS and trailers, one classifier was made for each of the 5 different tissues from the ribo-seq data. The random forest models gave a total of 74,847 uORFs predicted as translated. The tissue with most uORFs was fibroblast, with 33,859 uORFs found. See table 2 column 2 for mean squared error values for the ribo-seq model.

Sequence and CAGE model:

From these 5 ribo-seq models, I trained 5 sequence models. I used the predictions from the previous models to classify the uORFs based on sequence features. With these 5 new models, I overlapped all of them with the CAGE tables for those tissue (existence tables true/false).
The final model was then defined as all uORFs predicted by any of these 5 models. When combining the random forest models with the CAGE tissue data, I found a total of 21766 uORFs predicted. With an average of 13,966 uORFs (64.1% of positively predicted) per tissue. There are 9,997 (45.9% of the total number of predicted) uORFs present in all tissues. See table 2 column 3 for mean squared error values for the Sequence model.

The mean squared error (MSE) for the two models are shown in table 2. I will use the name active for predicted translated uORFs.

Tissue	Ribo-seq model	Sequence model
Brain	0.019	0.156
Fibroblast	0.020	0.149
kidney	0.018	0.164
Ovary	0.029	0.152
Prostate	0.031	0.165

Table 3: *MSE in random forest models per tissue. Each model have 5 tissue variants, the sequence model of brain is trained on ribo-seq model of brain and so on.*

There is a clear difference in the confidence levels in the models, since the first model is trained on CDS and second on uORFs.
The important features for this prediction will be described next.

Features of the two classifiers

After running the classifier pipeline, a relative importance was found in the model, for how important each feature was for the classifier. The ribo-seq features are described in table 4, and sequence features in table 5.

Rank of Feature	Feature name	Relative importance
1	ORFscore	1.00
2	Entropy	0.34
3	Translational efficiency	0.30
4	IO score*	0.29
5	Disengagement score	0.13
6	Floss	0.10
7	RSS**	0.09
8	RRS***	0.05

Table 4: *Relative importance ranking of uORF ribo-seq features from prediction. It can be noted that this is the average importance from the 5 ribo-seq models. *Inside outside score. **Ribosome stalling score. *** Ribosome release score.*

As seen by the importance of the ribo-seq features in table 4, the periodic features ORFscore and entropy have the highest impact on the model. It should be noted that features related to the end region of the uORF are less important, including RSS, RRS and disengagement score.

Rank of Feature	Feature name	Relative importance
1	Distance to CDS	1.00
2	Kozak sequence score	0.82
3	uORF order in Tx	0.74
4	Start codon	0.70
5	Fraction length*	0.69
6	Length of uORF	0.61
7	Stop Codon	0.22
8	In frame with CDS	0.21
9	uORF overlapping CDS	0.04

Table 5: *Relative importance ranking of uORF sequence features from prediction. It can be noted that this is the average importance from the 5 sequence models. *Fraction length, the size of the uORF relative to the transcript*

For the sequence features, distance to CDS is the most important feature. This means the distance from uORF stop site to CDS start site is the strongest indicator of translation for the uORFs. This is a somewhat surprising finding since this information is not likely to be available to a scanning ribosomes. The Kozak sequence score, which describes the initiation context of the uORFs, is, as expected, a strong feature with a relative importance of 0.82. Another interesting finding is that while distance to CDS is important, overlap with the CDS is the least important feature.

To get a clear view of differences between our predicted translated subset and the uORF candidate set, I compared metrics between the sets. I checked how the most important features varied between the predicted translated uORFs and the set of candidate uORFs. Start codon usage was one of the features with the largest change in its distribution, as seen in figure 10 A). The change in the distribution is strongly skewed towards increased ATG usage. The frequency of the low quality start codons AAG and AGG are almost non-existent in the predictions and a good indication that our classification is of high quality. Furthermore, our predicted distribution is in accordance with earlier finding of a ATG/CTG bias in uORFs [33]. I also checked the Stop codon usage, figure 10 C/D). Compared to the start codon usage, there was no significant difference between candidate set of uORFs and predicted active uORFs. The Kozak sequence score of CDS', active uORFs and predicted non-active uORFs is shown in figure 10 E). The active uORFs have a Kozak score more similar to the CDS,

compared with the predicted non-active ones. This is again indicating that our ribo-seq predictions gives predictions of high quality. The distance between the stop site of the uORF and the start site of the CDS is shown in figure 10 F). It is interesting to note that predicted active uORFs are much closer to the CDS, even though overlapping the CDS was an uninformative feature in the sequence model. This relationship between distance to CDS and overlap with CDS will be revisited in the discussion.

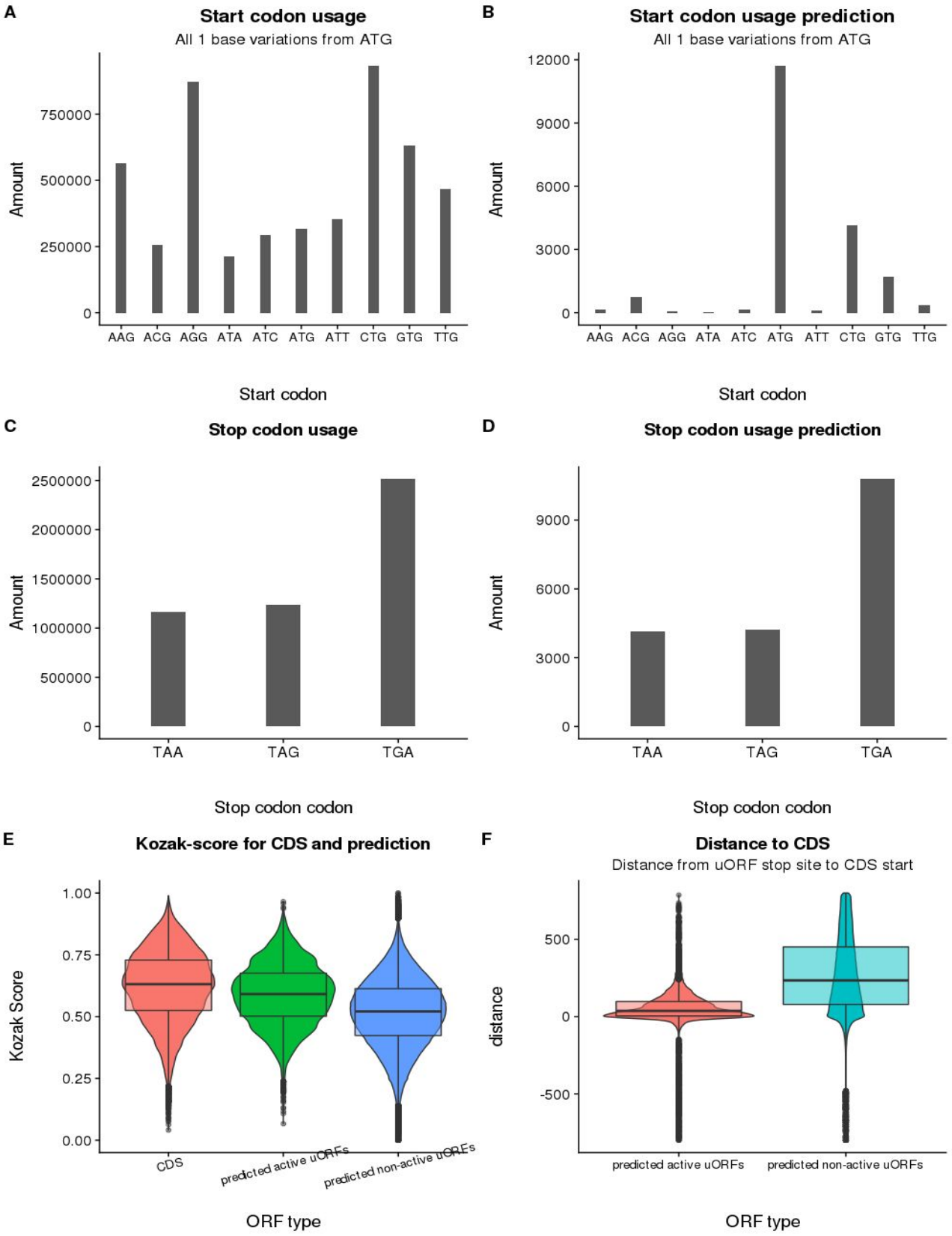

Figure 10: **A)** *Start codon usage in all candidate uORFs.* ***B)*** *Start codon usage in uORFs predicted to be active.* ***C)*** *Stop codon usage in all candidate uORFs.* ***D)*** *Stop codon usage in all uORFs predicted active.* ***E)*** Kozak score comparison between cds and prediction. Predicted active uORFs and CDS' have a more similar Kozak distribution than predicted

non-active uORFs. **F)** The distance between the stop site of the uORF and the start site of the CDS.

As a final test for bias in the model, I checked the correlation for each ribo-seq feature with the length of the uORFs. Only the RRS score had a significant correlation, of 72 %. Since RRS is the number of reads of uORF divided by number of reads on the trailer normalized by length this is mainly a result of very few ribo-seq reads overlapping the trailer. This score will therefore be highly dependent on the length of the uORF.

Effect of translated uORFs on CDS

A way to measure the effect of uORFs on the CDS is to see how the presence of uORFs affects the translational efficiency (TE) of the CDS. Figure x describes the variance in TE for CDS' in transcripts that contains predicted translated uORFs and CDS with no predicted translated uORFs. Figure 11 shows a small, but significant difference in TE of the CDS between active and non-active uORFs. The graph has a cutoff for included CDS', such that all CDS' with RNA-seq FPKM < 0.5 is filtered out.

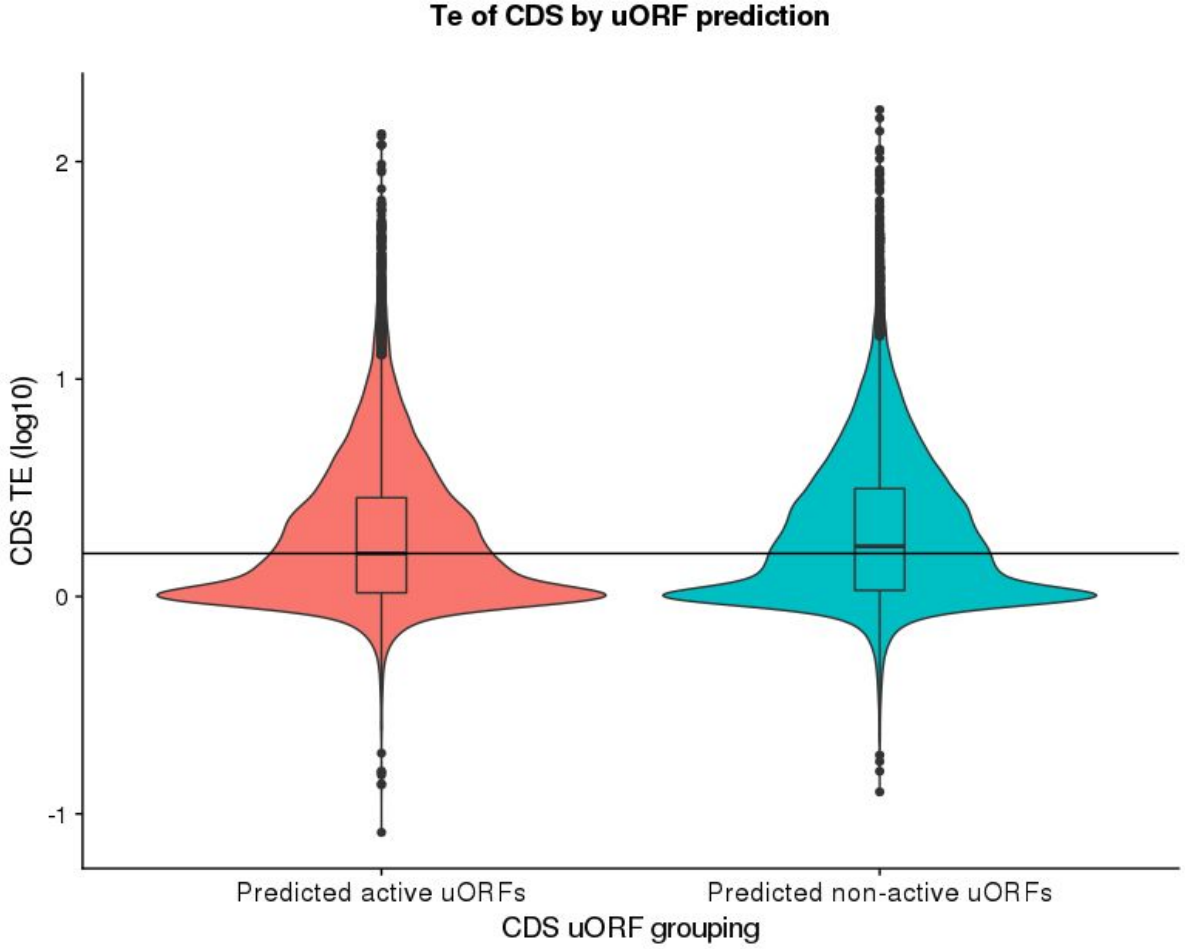

Figure 11: *Variation in CDS translational efficiency by cds' with predicted active uORFs vs predicted non-active uORFs. The CDS te values are of log scale. There is a statistical significant difference between the predicted active and non-active group with a p-value of*

1.0e-11 (Welch Two Sample t-test). The graph shows a filtered version where all CDS' with RNA-seq FPKM < 0.5 are filtered out. This filtering does not affect the conclusion in any way.

Tissue variance

Even though 21,766 uORFs were predicted to be translating, the variance of uORF usage between tissues were high with a standard deviation of 1,072 uORFs between tissues. In figure 12 A) the number of uORFs predicted for 20 tissues are shown. These are the 20 tissues with most uORFs predicted. In figure 12 B), the overlap of uORF usage between ovary and brain is shown. They have an overlap of 11523 uORFs (76.7 % of all predicted uORFs in ovary, and 81.1 % of all predicted uORFs in brain).

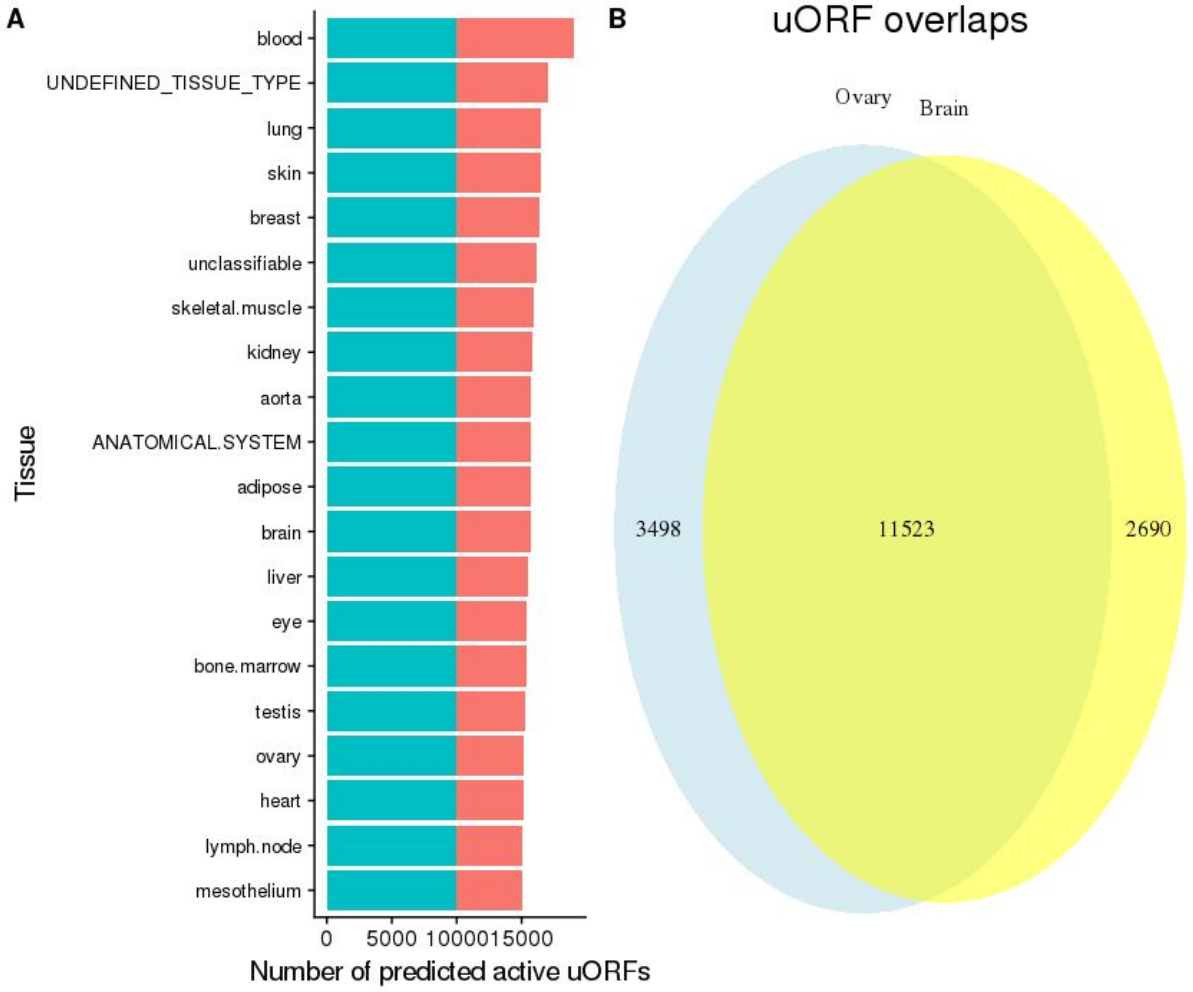

Figure 12: A) Number of predicted uORFs per tissue. Figure showing top 20 tissues according to number of uORFs. Turquoise colour represents uORFs that are included in all tissues, while red represents the uORFs that varies between tissues. *B) Overlap between prediction of uORFs in Ovary and Brain tissues. They have an overlap of 11523 uORFs.*

As for the ranking of the tissues with most predicted active uORFs (figure 12A), the results showed a similar orderings as the candidate uORFs by tissues in figure 9 A). While blood is the top tissue in both rankings, some tissues like heart change their rank position.

uORF variance between cancerous and healthy cell-lines.

To see if there were any variance between uORF usage in healthy and cancerous cell lines in the same tissues. We compared differential uORF usage between health brain cell-lines and glioblastoma (a type of brain cancer), see figure 13.

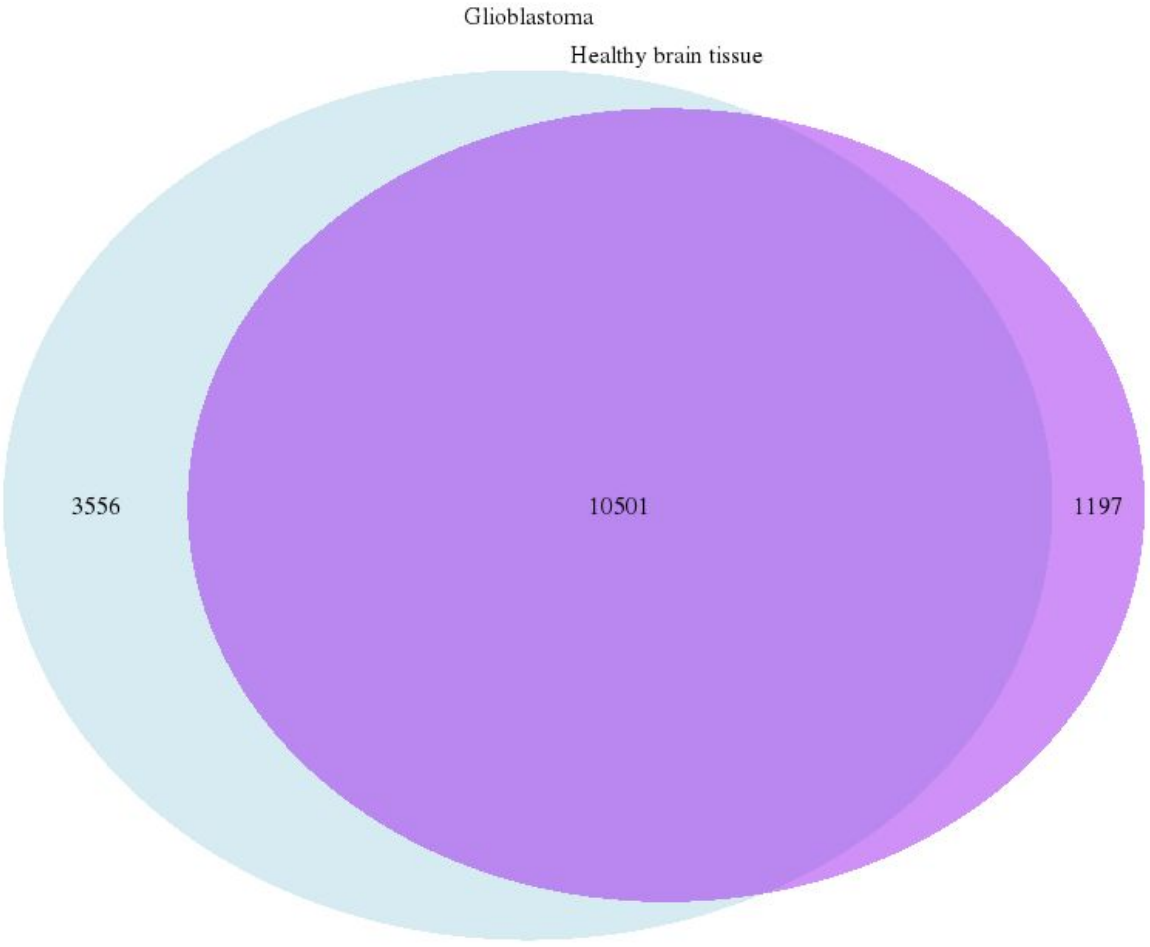

Figure 13: *Overlap between predicted active uORFs in glioblastoma and healthy brain tissue. An overlap of 10,501 uORFs (89.8 % of all prediction uORFs in healthy brain tissue and 74.7 % of all predicted active uORFs in Glioblastoma)*

The Glioblastoma sample had a lot more predicted uORFs compared to healthy brain tissue. Glioblastoma have 14,057 predicted uORFs compared to 11,698 uORFs in the healthy brain tissue. The number of uORFs not shared by the tissues were therefore different, at 3,556 for glioblastoma. I will talk about the implications of these differences in the discussion.

Comparison with other uORF predictions

To assess how the prediction corresponds to other studies of uORF prediction, I compared our results to that of McGillvary et al. From their total number of 18880 predicted translated uORFs in the GRCh19 annotation, I used the NCBI assembly conversion tool to recover 13040 of them in the newest human annotation (GrCH38). All of which were included in our

initial catalog of candidate uORFs. When comparing predicted active uORFs, the intersection was reduced to 1675 uORFs contained in both sets (12 % of all predicted active uORFs by McGillvary et al.). See figure 14 A).

I hypothesized that a reason for this small intersection, could be a difference in start codon usage between the classifiers. I therefore compared the start codon usage between the predictions, shown in figure 14 B).

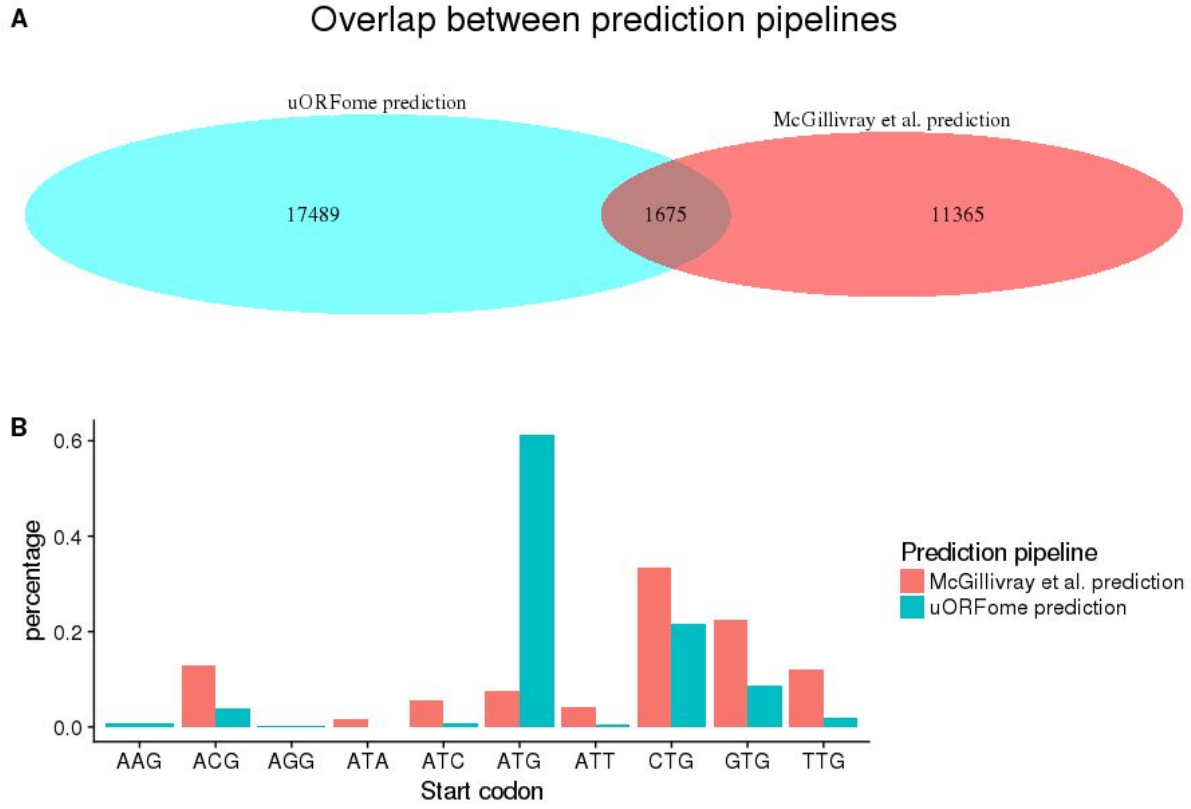

Figure 14: A) *Overlap between prediction pipelines. A total overlap of 1675 uORFs.* ***B)*** *Start codon distribution in uORFome prediction and McGillivray prediction. Height of each bar is the percentage of total the total prediction set, scale is set so 0.2 = 20%. There is a clear difference in number of ATG's and CTG's used. * A few of the uORFs in the uORFome prediction was filtered out in the comparison, because of a difference in definition of uORF for the two predictions.*

As shown in figure 14 B) the start codon distributions differ. While my uORFome pipeline have many more ATGs, *McGillivray's* pipeline have more CTGs. It can also be noted that they choose to exclude AAG and AGG from the pipeline. The comparison will be continued in the discussion.

Example of validating uORF predictions

Very few uORFs have been experimentally validated. One example of validation is in the ATF4 gene that harbors 3 uORFs. Out of these only 2 uORFs can be found by our pipeline (

since the first uORF is only 6 base pairs long). Both of these are predicted to be translated (table 6). I also checked our prediction set for any predicted false positive uORFs in ATF4. The result showed that there were 25 candidate uORFs in the ATF4 gene, but only the 2 experimentally verified uORFs were predicted as translated by our prediction. That is out of 25 candidate, all 25 were predicted correctly (2 as translated, and 23 and non translated).

Another example is the ABCC2 gene, where experimental validation has validated 1 translated uORF. This uORF is found in my prediction set, but I also found 2 additional uORFs, these however were not checked in the experiment so it is not validated whether these are false positives or novel translated uORFs.
Finally, for the ADH5 gene, the experimental validation found 2 candidate uORFs out of which one was validated in experiments. The other one was not tested for translation. Our prediction is in accordance with this and the uORF they found to be translated I also predict to be translated.

Gene symbol	uORF ID	Predicted translated	Experimentally validated
ATF4	chr22,+,39520586 6	✗*	✓
ATF4	chr22,+,39520648 12	✓	✓
ATF4	chr22,+,39520747 5 ;39521354 175	✓	✓
ABCC2	chr10,+,99782699 48	✓	✓
ABCC2	chr10,+,99782708 39	✓	Not checked in experiment
ABCC2	chr10,+,99782740 69	✓	Not checked in experiment
ADH5	*chr4,-,99088703 33*	✓	✓

Table 6: Example of comparison between predicted uORFs in database and experimental validation. The 5th and 6th uORF (ABCC2) were not checked in the article for signs of translation, so there is no data to say our prediction is correct or not for these. It can be of interest to see that the uORF 5 and 6 have the same stop codon position. * This uORF is only 6 bases long, while our pipeline only checks uORFs that are at least 12 bases long.

As a visual confirmation, I created a browser window with tracks from the ATF4 gene and its uORFs. In figure 15 the tracks of the uORFs in ATF4, overlapped with a displayed ribo-seq library from a fibroblast cell-line. All three uORFs have a strong ribo-seq signal over their start codon supporting the translation of these uORFs. Furthermore, the ribo-seq signal of

uORF 1 is stronger than the ribo-seq signal of the CDS. Something that points to how important these uORFs might be for these transcripts. From the overlap of uORF 3 with the CDS it is difficult to know what part of the ribo-seq signal over uORF 3 is from the CDS and what is from the uORF itself. This creates a potential bias in the model.

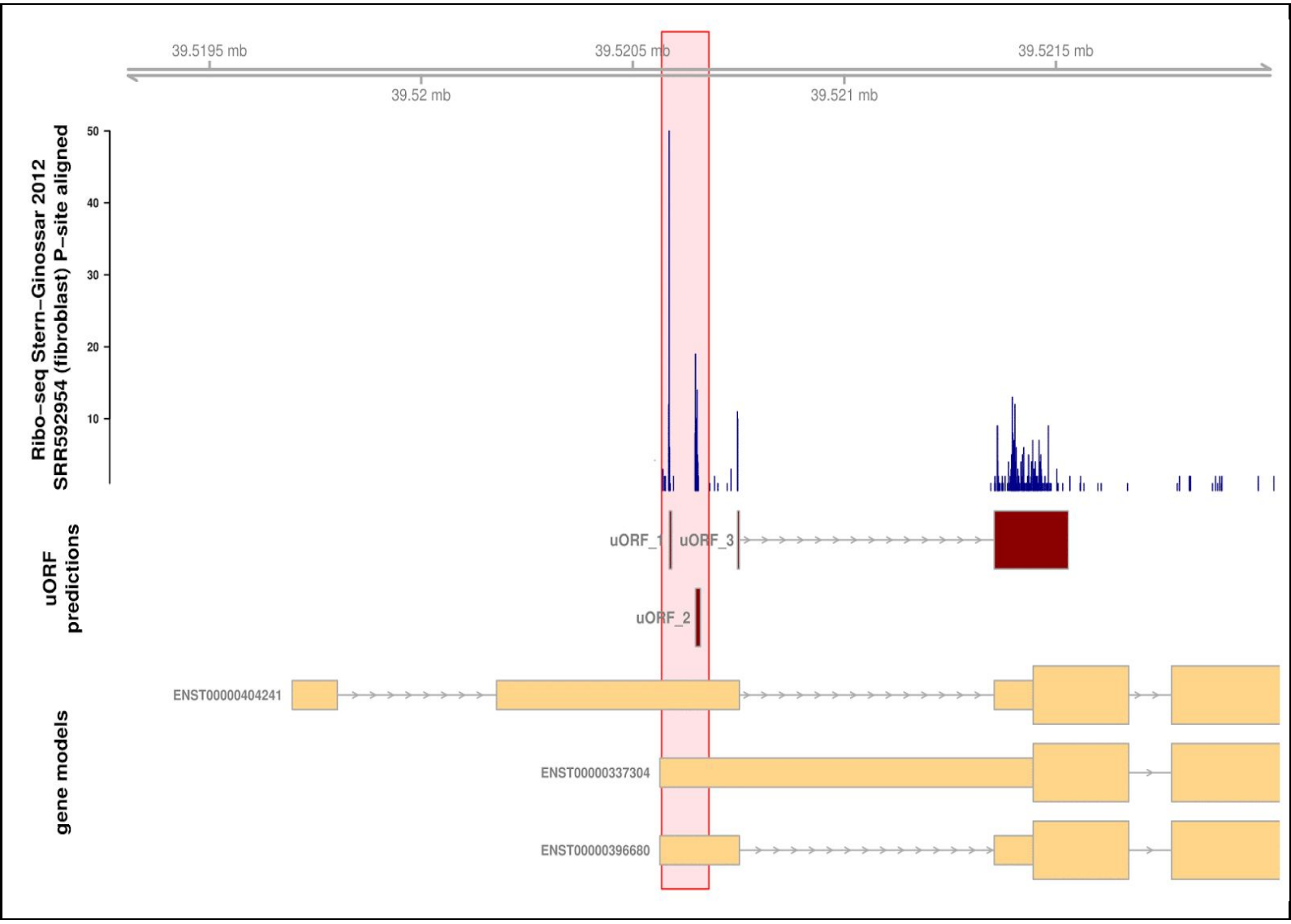

Figure 15: *Browser snapshot of ATF4 uORFs. The tracks are combined with ribo-seq from fibroblast (SRR592954). The ribo-seq Forward track shows the distribution of ribo-seq reads , (in blue). The uORF predictions track shows the genomic positions of the uORFs. The gene models show the three different transcripts of the ATF4 gene. The beige and dark red rectangles are exons, the grey arrow bars are introns.*

Here, I have shown results from my uORFome pipeline using ORFik and other tools. In the next chapter I will discuss these interesting findings and their implications.

Discussion

In this thesis I have developed a method for identifying uORFs both from ribo-seq and sequence based metrics and used it to comprehensively catalogue the human uORFs and the tissue-specific regulation of these. Specifically, I used the combination of these models to predict translated uORFs on the entire FANTOM5 data set, encompassing the largest resource of 5' leader annotation across tissues, primary cells and cell lines. I have compared variance in uORFs usage between tissues and between healthy and cancerous cell lines. Finally, I have compared my prediction with the set of uORFs identified by McGillivray *et al.* and investigated experimentally validated uORFs, focusing on the ATF4 gene. The small validation set of uORFs I found from uORFdb is not sufficient to provide a statistical validation for our prediction, but can serve as a proof of concept that from the set of 2.2 million candidate uORFs, the predicted set of 21,766 (0.1 % of the candidate set) presents a functionally relevant subset. This suggests that my method for uORF identification has the potential to extend the current uORF annotation significantly and provides a sound basis for further experimental validations.

Here I will discuss general thoughts on the results, the limitations of the approach and potential improvements.

CAGE and 5' leader annotation

A major computational task addressed in this thesis, is the efficient identification of uORFs in a large sequence reference. The KMP-algorithm created for uORFs in ORFik can search the entire human transcriptome for uORFs in a matter of seconds. Since I include non-canonical start codon ORFs in our definition of uORF, our candidate set of uORFs becomes very large.

By integrating published ribo-seq and RNA-seq libraries from a number of human tissues and cell-lines with precise annotation of transcription start sites from the FANTOM5 atlas, I was able to map the human uORFome and its variance across tissues at an unprecedented scale. The contribution of precise TSS annotations to inform differential uORFome usage is depicted in figure 9. The number of uORFs predicted to be active in all tissues was 9,997 (45.9% of the total number of predicted uORFs). This suggests that a substantial number is under regulatory control and could have an impact on the translation of the CDSs.

Prediction of uORFs

From the results, it is shown that the tissues that have more candidate uORFs, also have more predicted active uORFs. The correlation in the number of uORFs in the candidate set

by tissues in figure 9 A) and the number of predicted active uORFs per tissue in figure 12 A) is clear. The more candidate uORFs a tissue has, the more predicted uORFs it should get.

As can be seen from the metrics, figure 10, our predicted uORFs have features closer to what would expect from translated uORFs, compared to the uORF candidate set. The codons known not to be efficient initiation sites (AAG/AGG) are almost entirely absent in the set of predicted active uORFs shown in figure 10 B).

Our classifiers used many different features, some of which are more likely biologically relevant than others. The variable importance of features derived in the classification step (shown in table 4 & 5), allowed me to rank the features according to their relative importance. Based on these results, features relying on periodic signals like ORFscore and entropy were better predictors than pure counting or ratio metrics. It is also interesting that the metrics that focus on the downstream region of the uORF, like Ribosome Stalling Score and Ribosome Release Score, are weighted as less important in the classifiers, as can be seen in table 4. I hypothesize that good periodic feature scores are harder to generate by random chance especially given signal noise, and are therefore better predictors of biologically relevant uORFs.

Somewhat surprisingly, the distance of a candidate uORF to its associated downstream CDS was one of the most informative features. A statistical reason for this results, is that the importance ranking is affected when features used to train the model correlate. The distance from uORF stop site to CDS start site is always negative for uORFs going into the CDS. In some sense the model therefore does not need overlap with CDS as a feature, it is already included in the distance feature.

I hypothesize that since the information about the distance to the CDS is unavailable to scanning and translating ribosomes, its importance is likely caused by some indirect effect like evolutionary constraints. For instance it could indicate that proximal uORFs have a higher usage of ATG as start codon, since our model biases towards ATG uORFs (figure 10 A/B). From an evolutionary perspective, it might be that uORFs are being “tested” for functionality through translation, because some will be close enough to the CDS to have an effect. As depicted for the ATF4 gene in figure 15 the uORF that is actually regulating the CDS, is uORF 3., i.e.the uORF that overlaps the CDS, while the two others merely regulate the 3rd one.

Comparison to other uORF predictions

The set of predicted active uORFs had little overlap with the uORF prediction of *McGillivray* et al. As shown in figure 14 A), the overlap is only 12 %. This could be related to differences in how translated uORFs were identified. McGillivray *et al*. mainly focus on sequence-based features and amino acid content, whereas my uORFome pipeline stresses the importance of ribo-seq and RNA-seq derived features. It can be noted that the pipeline created by McGillivray *et al.* uses a peptide database to get expression levels of different uORFs in

tissues instead of CAGE. A drawback of using peptide databases instead of CAGE is that uORFs upstream of the original TSS annotation can never be included, since they are not searched for in the peptide databases. uORFs have been found not to be conserved on the amino acid level, making it less useful to use amino acids metrics in the model [34]. As explained in the introduction section, uORFs primary functional mechanism is through releasing ribosomosomes. There is no functional peptide needed to make this happen, which McGillivray et al use as their primary metric of uORF function.

Compared to the predicted by McGillvary *et al.* our prediction finds a much higher fraction of ATG uORFs. I sought to avoid start codon biases by intentionally excluding this feature from the first stage of model training which included mainly features derived from RNA-seq and ribo-seq datasets. However, as a large proportion of annotated CDS begin with an ATG as start codon, the model could have implicitly captured this property from other correlated features. From the sequence prediction model on uORFs it can be seen that it is distance to CDS and not start codon that is the most important feature for the random forest model. MacGillivray *et al.,* use a large set of protein expression level features that are used in the final model, while my model focuses more on ribo-seq features like ORFscore. So the low overlap in the predictions might come from McGillvary's use of protein expression level metrics.

General discussion

A primary focus of this thesis have been tissue specific uORF usage. This was done using CAGE. However, there is potentially a bias in our method of finding the best CAGE tag per leader sequence. If a leader sequence has two equally strong CAGE tags, (same number of reads), the most upstream one is chosen by default. This biases our leaders towards longer lengths. It is also possible that if two equally strong CAGE tags exists,there are actually two different leader variants in the same gene. A leader sequence could also have one isoform that is highly transcribed, and one that is less transcribed. In our pipeline, the downstream CAGE tag will always be filtered out on a per experiment basis. I tried to alleviate this problem by using as many replicates in the tissues as I could find. Since I have a total of 1863 CAGE experiments, on a per tissue level at least 2 experiments should contain the uORF.

There is also a potential issue in allowing several uORFs to have the same genomic stop codon position but different start codons positions. The strongest ribo-seq feature in our prediction was found to be ORFscore. A set of uORFs all sharing the same stop codon position will all be in the same reading frame, therefore their ORFscore will correlate. If the first uORF is always translated, it means the other can never be translated (since they share stop codon), but our pipeline will most likely predict several of these as translated. A possible solution would be to always choose the longest uORF, but this could lead to functional small uORFs being lost in the model.

There are also some potential statistical issues with some of the features, e.g. ORFscore is a scoring of the periodicity of reads on the first frame relative to the two others. On a small uORF, e.g. 4 codons, the random chance of getting a good ORFscore is higher than on a bigger uORF. A 3 codon uORF only need reads to hit on the positions {1,4,7}, to get a good score. While a longer uORF must hit on {1,4,7,10,13,..}. The smaller the uORF is, the fewer reads it needs to get a good score. As a defense for using these metrics, looking at the figure 15 for uORF tracks in the ATF4 gene. uORF1 has a very strong spike on the start codon. That means the ORFscore will be very high. This shows the point of why the filter between my ribo-seq and sequence classifiers filters by reads over the start codon, as seen in figure 7.

Translational efficiency is one of the most popular features for representing predicted translation rate. A problem with TE is that it is combined by ribo-seq and RNA-seq experiments that are not from the same cell. Usually they are from biological replicates of the same cell-line using similar experimental protocols. The variance in mRNA expression levels between these two biological replicates could give the matching a low quality, e.g. the time of day or extracellular environments of the two cells. By using large amounts of replicates in this pipeline, I have tried to address this issue.

As seen from my validation on the experimentally verified ATF4 uORFs in table 6, uORF1 was not in the candidate set of uORFs. The reason was that the length of the uORF was only 6 bases. To avoid too many false positives I excluded all uORFs with size less than 12 bases. Future analyses of the data presented here could incorporate statistical methods to remove noise in ribo-seq datasets [35] . This could be implemented into ORFik, to make a prediction for uORFs down to just a start codon and stop codon (6 bases).

From my definition of a uORF, it must either use ATG or a one base variation of ATG as start codon. There is a possibility that some leader sequences translate regions where another codon is used as start codon than my set, these will not be found [36].

An important distinction must be noted between translation and function (an effect on phenotype). In this thesis I have predicted translation, not function. The article of McGillivray et al. with a similar uORF prediction, claims function with the title: “A comprehensive catalog of predicted functional upstream open reading frames in humans” [20]. McGillivray et al. also differentiate between function and translation, by stating: “Study of uORF translation and function was historically limited to the experimental evaluation of individual uORFs”. In the claim of function, McGillivray *et al.’s* article states: “Measurement of comparative frequency of mutation among uORF start codons was taken as a measure of evolutionary conservation and functional significance of predicted positive uORFs”. McGillivray et al. also performed a blast search of the predicted translated uORFs for peptide products in the THISP database, in addition to a search for single nucleotide variants to verify possible mutations in uORFs. This was used to claim functionality of uORFs. However, this leads to a misleading definition of function in the title, what they can claim is that they have made a catalogue with evidence of translation, from ribo-seq and protein databases. Only a small set of these are additionally possibly functional uORFs by mutation databases of single nucleotide variants.

Mutations in the start codon of uORFs can lead to health issues [37]. McGillivray et al. use of single nucleotide variants (SNV) is interesting. They search the SNV databases for hits on the start codons of the predicted translated uORFs. Combining the knowledge of a predicted translated uORF and a possible SNV's at the start codon of that uORF gives a possibility find uORFs that can lose its start codon in cancerous cell-lines. I could do a similar approach in my prediction. An even stronger method would be to combine the SNV data with an evolutionary conservation score for each uORF. That is, if the uORF is highly conserved between species, and cancerous cell lines have mutations in those uORFs, it would be a stronger indicator of possible functional change.

To investigate within tissue variance of predicted active uORFs, I tested two cell-lines in brain. The results showed that Glioblastoma had 3556 uniquely predicted active uORFs relative to the healthy brain tissue. This gives an example of possible translational regulation between cell-lines and offers interesting prospects for further analyses.

Conclusion and future prospects

This project can be seen as a stepping stone to a better understanding of uORFs at the transcriptome-wide level, but several improvements are possible to make the results more robust.

Apart from challenges in data integration and interpretation, the prediction of uORFs and analysis of large genome scale datasets requires specialized software that can effectively exploit compute resources and allow e.g. for parallelization of tasks.

ORFik was made with this in mind, and is able to handle large data sets. The main bottlenecks are the Bioconductor packages GenomicRanges and GenomicFeatures, that have not implemented methods for big data extraction in all our needed cases. As an example is the method countOverlaps() in the package GenomicFeatures, which on a ribo-seq set of 5GB running on 2.2 million candidate uORFs, can fill up 2TB of ram quite quickly. These big datasets had to be split in smaller sets and then combined together. I have contributed with recommendations for future improvements on these packages and discussed these with the team at the Bioconductor core so this is likely to improve in the future.

An important improvement would be a more statistically rigorous selection of CAGE peaks when assigning new TSSs. There have been attempts of similar ideas, like the bioconductor package cageR [38] where they utilise clustering and noise modelling of the CAGE reads to predict the TSS per genomic window (window specified by user). This package was not used by my pipeline, but their ideas could be implemented in a similar manner into ORFik.

ORFik contains several ribo-seq metrics described in this thesis like TE and ORFscore, since this is a collection of features used in the scientific community, the metrics should be able to reduce the noise in the ribo-seq libraries. Still, improvements in filtering could be made. Some packages have tried to implement stronger filters for periodic noise of ribo-seq reads, like the python package ribodeblur [39]. This would lead to a stronger predictive power, especially for small uORFs.

This atlas can be used as a basis for validation experiments and lead researchers towards possible explanations for observed uORF regulation. To have an atlas with tissue variance could potentially save time before deciding what experiment to run. ORFik also makes it easy to make pipelines for new datasets and species, creating an easy to use and standardized tool for cataloguing uORFs and deriving the rules of uORF-mediated regulation.

References

1. Das, G. *et al.* Role of 16S ribosomal RNA methylations in translation initiation in Escherichia coli. *EMBO J.* **27,** 840 (2008).
2. Ambrogelly, A., Palioura, S. & Söll, D. Natural expansion of the genetic code. *Nat. Chem. Biol.* **3,** 29 (2006).
3. Kozak, M. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 8301–8305 (1990).
4. Bertram, G., Innes, S., Minella, O., Richardson, J. & Stansfield, I. Endless possibilities: translation termination and stop codon recognition. *Microbiology* **147,** 255–269 (2001).
5. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Mol. Cell* **62,** 462–471 (2016).
6. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biol.* **3,** reviews0004.1 (2002).
7. Calvo SE, E. al. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19372376. (Accessed: 31st May 2018)
8. Barbosa, C., Peixeiro, I. & Romão, L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genet.* **9,** (2013).
9. Grant CM, E. al. Requirements for intercistronic distance and level of eukaryotic initiation factor 2 activity in reinitiation on GCN4 mRNA vary with the downstream... - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/8139562. (Accessed: 29th May 2018)
10. Regulated Translation Initiation Controls Stress-Induced Gene Expression in Mammalian Cells. *Mol. Cell* **6,** 1099–1108 (2000).

11. Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F. & Dietz, H. C. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat. Genet.* **36,** 1073–1078 (2004).
12. A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **45,** 1690–1700 (2013).
13. Choy, J. Y. H., Boon, P. L. S., Bertin, N. & Fullwood, M. J. A resource of ribosomal RNA-depleted RNA-Seq data from different normal adult and fetal human tissues. *Sci Data* **2,** 150063 (2015).
14. Wethmar, K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA* **5,** 765–778 (2014).
15. Anna M McGeachy, N. T. I. Starting too soon: upstream reading frames repress downstream translation. *EMBO J.* **35,** 699 (2016).
16. Roy, B. *et al.* The h subunit of eIF3 promotes reinitiation competence during translation of mRNAs harboring upstream open reading frames. *RNA* **16,** 748–761 (2010).
17. Lee S, E. al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22927429. (Accessed: 31st May 2018)
18. Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M. A. & Leutz, A. uORFdb--a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.* **42,** D60–7 (2014).
19. Spealman P, E. al. Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/29254944. (Accessed: 30th May 2018)
20. McGillivray, P. *et al.* A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.* **46,** 3326–3338 (2018).
21. The FANTOM Consortium, Pmi, T. R. & (dgt), C. A promoter-level mammalian

expression atlas. *Nature* **507,** 462 (2014).

22. Barbosa, C., Peixeiro, I. & Romão, L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genet.* **9,** e1003529 (2013).
23. Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8,** 1365–1379 (2014).
24. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
25. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
26. Zhang, S. *et al.* Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning. *Cell Syst* **5,** 212–220.e6 (2017).
27. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33,** 981–993 (2014).
28. Young, S. K. & Wek, R. C. Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *J. Biol. Chem.* **291,** 16927–16935 (2016).
29. Grzegorski, S. J., Chiari, E. F., Robbins, A., Kish, P. E. & Kahana, A. Natural Variability of Kozak Sequences Correlates with Function in a Zebrafish Model. *PLoS One* **9,** e108475 (2014).
30. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251 (2013).
31. Zhang, Y., Zhao, T., Li, W. & Vore, M. The 5'-untranslated region of multidrug resistance associated protein 2 (MRP2; ABCC2) regulates downstream open reading frame

expression through translational regulation. *Mol. Pharmacol.* **77,** 237–246 (2010).

32. Posttranscriptional Regulation of Human ADH5/FDH and Myf6 Gene Expression by Upstream AUG Codons. *Arch. Biochem. Biophys.* **386,** 163–171 (2001).
33. Spealman, P. *et al.* Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res.* **28,** 214–222 (2018).
34. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35,** 706 (2016).
35. Aspden, J. L. *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife Sciences* **3,** e03528 (2014).
36. RAN translation—What makes it run? *Brain Res.* **1647,** 30–42 (2016).
37. The implications of structured 5′ untranslated regions on translation and disease. *Semin. Cell Dev. Biol.* **16,** 39–47 (2005).
38. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43,** e51 (2015).
39. Wang H, E. al. Using the Ribodeblur pipeline to recover A-sites from yeast ribosome profiling data. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/m/pubmed/29330118/. (Accessed: 30th May 2018)
40. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Scientific Data* **4,** 170107 (2017).
41. Gonzalez, C. *et al.* Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J. Neurosci.* **34,** 10924–10936 (2014).
42. Rutkowski, A. J. *et al.* Widespread disruption of host transcription termination in HSV-1 infection. *Nat. Commun.* **6,** 7126 (2015).
43. Stern-Ginossar, N. *et al.* Decoding human cytomegalovirus. *Science* **338,** 1088–1093 (2012).

44. Sidrauski, C., McGeachy, A. M., Ingolia, N. T. & Walter, P. The small molecule ISRIB reverses the effects of eIF2α phosphorylation on translation and stress granule assembly. *Elife* **4,** (2015).
45. Hsieh, A. C. *et al.* The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* **485,** 55–61 (2012).
46. Stumpf, C. R., Moreno, M. V., Olshen, A. B., Taylor, B. S. & Ruggero, D. The Translational Landscape of the Mammalian Cell Cycle. *Mol. Cell* **52,** 574–582 (2013).

Supplements

CAGE libraries

For information of the CAGE data, see the FANTOM5 project:
FANTOM5 OSC CORE (contact: Al Forrest for more information) [40].

Ribo-seq and RNA-seq libraries

GEO accession numbers for all ribo-seq and RNA-seq data sets used in this thesis, including all reference articles for experiments [41,42,43,23,43,44,45,46].

Ribo-seq	Ribo-seq	RNA-seq	RNA-seq
GSM1495244	GSM1020244	GSM1606106	GSM869046
GSM1495249	GSM1020247	GSM1606099	GSM869038
GSM1495245	GSM1020246	GSM1606100	GSM869044
GSM1495250	GSM1020249	GSM1606109	GSM869036
GSM1495246	GSM1331345	GSM1606110	GSM869042
GSM1495251	GSM1331349	GSM1606111	SRS476841
GSM1495247	GSM1331344	GSM1606112	SRS476840
GSM1495252	GSM1331348	GSM1606113	SRS476849
GSM1464095	GSM1331343	GSM1606114	SRS476848
GSM1464101	GSM1331347	GSM1606107	SRS476845
GSM1444166	GSM1331342	GSM1606108	SRS476844
GSM1444180	GSM1331346	GSM869041	SRS476843
GSM1020235	GSM1606101	GSM869047	SRS476842
GSM1020234	GSM1606102	GSM869039	SRS476851
GSM1020238	GSM1606103	GSM869045	SRS476850
GSM1020236	GSM1606104	GSM869037	SRS476847
GSM1020237	GSM1606105	GSM869043	SRS476846
GSM1020245		GSM869040	

Supplementary table 1: *The GEO accession numbers of all Ribo-seq and RNA-seq experiments used for uORF metrics and creating the uORF classifier. There are 35 pairs of ribo-seq and RNA-seq experiments, 70 experiments in total. The first listed ribo-seq experiment links to the first listed RNA-seq experiment etc.*