06/05/2020

BSc 3 - Solving Biological Problems that Require Mathematics
Université de Lausanne

Comparative Analysis of Codon Usage in Human mRNA and lncRNA

Juno KIM, Aurélien ROBADEY, Augustin ROLAND, Varjany VASHANTHAKUMAR

Supervisor: Anneke BRÜMMER

Introduction

The aim of this project was to analyse mRNA (**m**essenger RNA) and lncRNA (**l**ong **n**on-**c**oding RNA) sequences in order to understand why lncRNAs are not translated into proteins even though they are similar in structure to mRNAs. We analysed the codon and codon pair usage in those two RNA types and looked for features that are distinct.

lncRNAs are defined as RNA-transcripts that are longer than 200 nucleotides and are not translated into proteins. Unlike mRNAs, they can be found within and outside the nucleus. They play important roles in biological processes, such as gene transcription regulation, post-translational regulation, epigenetics, etc. For example, the lncRNA *Xist* is responsible for the inactivation of one of the X-chromosome in female placental mammals. Nevertheless, the way lncRNAs work and their roles are still poorly understood. Our project aims to increase the general understanding of this biological feature.

Methods

We used RNA sequence data from the GENCODE project. For the purpose of this project, we were interested in analysing the human mRNA genome and lncRNA sequences. We used Python and the Biopython package to read and study these data.

To identify the mRNA coding sequences, we used the CDS (**C**oding **D**NA **S**equence) numbers. These are specific to each sequence and indicate the first and the last nucleotides that are translated. In lncRNAs, we extracted the longest ORF (**O**pen **R**eading **F**rame) for each sequence. The ORFs were defined as sequences starting with an ATG and ending with a STOP codon.

The project consisted of two main parts (Figure 1). The first part was the study of (single) codon usage in mRNA and lncRNA sequences. For that, we cut the sequences into pieces of three nucleotides (codons). The second part of the project consisted in looking at the codon context (codon pairs).

We defined the codon pairs as overlapping sequences of length six, e.g. the sequence 5'-ATGCCATAG-3' would be cut into 5'-ATGCCA-3' and 5'-CCATAG-3'. All sequences of the mRNA and lncRNA databases were treated as described, only skipping sequences that could not be defined clearly.

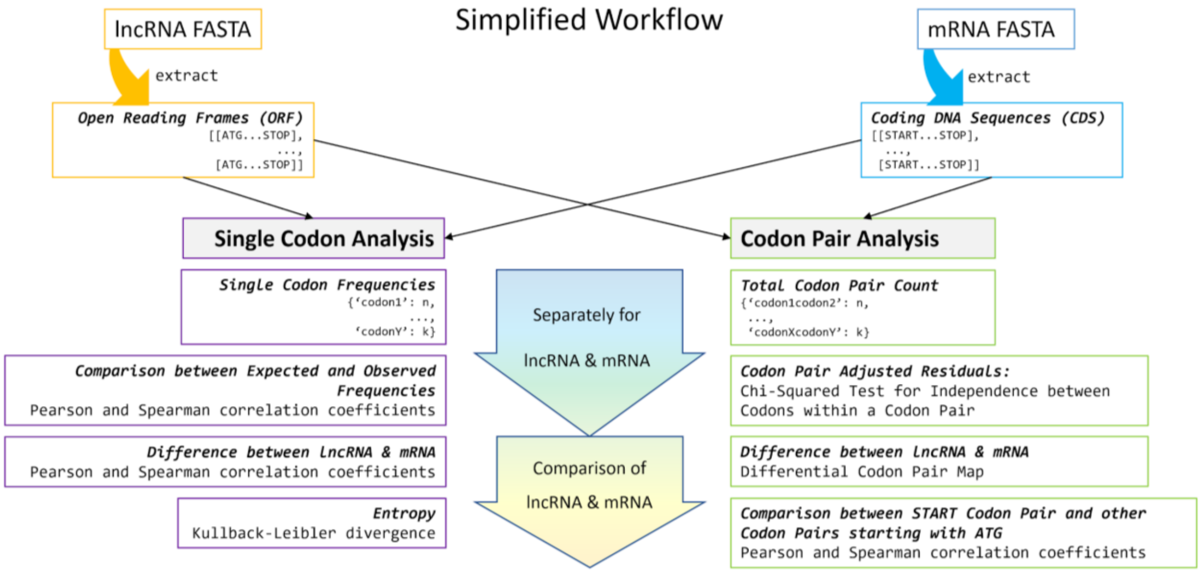
Simplified Workflow
lncRNA FASTA
extract
Open Reading Frames (ORF)
[[ATG...STOP],
...,
[ATG...STOP]]
mRNA FASTA
extract
Coding DNA Sequences (CDS)
[[START...STOP],
...,
[START...STOP]]
Single Codon Analysis
Single Codon Frequencies
{'codon1': n,
...,
'codonY': k}
Comparison between Expected and Observed Frequencies
Pearson and Spearman correlation coefficients
Difference between LncRNA & mRNA
Pearson and Spearman correlation coefficients
Entropy
Kullback-Leibler divergence
Separately for
lncRNA & mRNA
Comparison of
lncRNA & mRNA
Codon Pair Analysis
Total Codon Pair Count
{'codon1codon2': n,
...,
'codonXcodonY': k}
Codon Pair Adjusted Residuals:
Chi-Squared Test for Independence between Codons within a Codon Pair
Difference between LncRNA & mRNA
Differential Codon Pair Map
Comparison between START Codon Pair and other Codon Pairs starting with ATG
Pearson and Spearman correlation coefficients

Figure 1: The simplified workflow of this project. The project consisted of two separate parts, namely the Single Codon Analysis and the Codon Pair Analysis.

Single Codon Analysis

During the single codon analysis we mainly worked with codon frequencies, which facilitate comparisons between mRNA and lncRNA sequences.

In order to know whether the observed single codon and codon pair frequencies were biased, we shuffled the sequences randomly. This allowed us to compare our observations with what would be expected if the nucleotides were randomly distributed.

Codon Pair Analysis

During the codon pair analysis we worked with 61×64 contingency tables and absolute numbers (counts) of codon pairs (Table 1). The contingency tables only have 61 row variables due to the impossibility of the three stop codons (TAA, TAG, and TGA) appearing in the first position of a pair (causing a break).

In order to detect biases in the usage of certain codon pairs, we used methods described by Moura et al.[2] These are based on a Pearson's Chi-squared test for independence of two variables and the calculation of adjusted residuals $d_{i,j}$ for each codon pair.

2nd codon→ 1st codon↓	AAA	···	TTT	row totals
AAA	$n_{1,1}$	$n_{1,j}$	$n_{1,64}$	$n_{1.}$
⁝	$n_{i,1}$	$n_{i,j}$	$n_{i,64}$	$n_{i.}$
TTT	$n_{61,1}$	$n_{61,j}$	$n_{61,64}$	$n_{61.}$
column totals	$n_{.1}$	$n_{.j}$	$n_{.64}$	N

Table 1: General form of the contingency tables used for the codon pair analysis.

Adjusted residuals $d_{i,j}$ are defined as

$$d_{ij} = \frac{r_{ij}}{\sqrt{v_{ij}}} \qquad \text{where } r_{ij} = \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{N}\right)}{\sqrt{\frac{n_{i\bullet} n_{\bullet j}}{N}}} \qquad v_{ij} = \left(1 - \frac{n_{i\bullet}}{N}\right)\left(1 - \frac{n_{\bullet j}}{N}\right) \qquad \text{and}$$

and are derived from the test-statistic $\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} {r_{ij}}^2$

Therefore, the test-statistic is 0 if and only if all cells of the contingency table contain observed values that are equal to the expected values under the assumption of independence ($r_{i,j} = 0$, $\forall$ i, j). The adjusted residuals $d_{i,j}$ are useful for identifying cells that are responsible for the rejection of independence, i.e. identifying codon pairs that are either overrepresented or underrepresented relative to their expected occurrences when codons are randomly distributed. Furthermore, Haberman showed that P($-3 < d_{i,j} < 3$) ≈ 0.9973, meaning that for a 99.73% confidence level, the pair (*1st codon$_{(i)}$*, *2nd codon$_{(j)}$*) is responsible for the rejection of independence if $|d_{i,j}| \geq 3$.[3]

In order to compare the codon pair usage between lncRNA and mRNA, we calculated the absolute differences between the adjusted residuals of overlapping cells (= identical codon pairs in lncRNA and mRNA) and ranked them according to their differences.

Results and Discussion

Single Codon Analysis

We first compared the codon frequencies between the real and shuffled mRNA sequences. We observed that they differ greatly (Figure 2). Whereas the codon frequencies in shuffled sequences seem rather even, we observe big differences between codon frequencies in real sequences. These two datasets are not correlated (Figure 3), indicating that codon usage in mRNA is biased, which can be explained by the fact that mRNAs code for proteins and therefore encode highly specific primary structures.

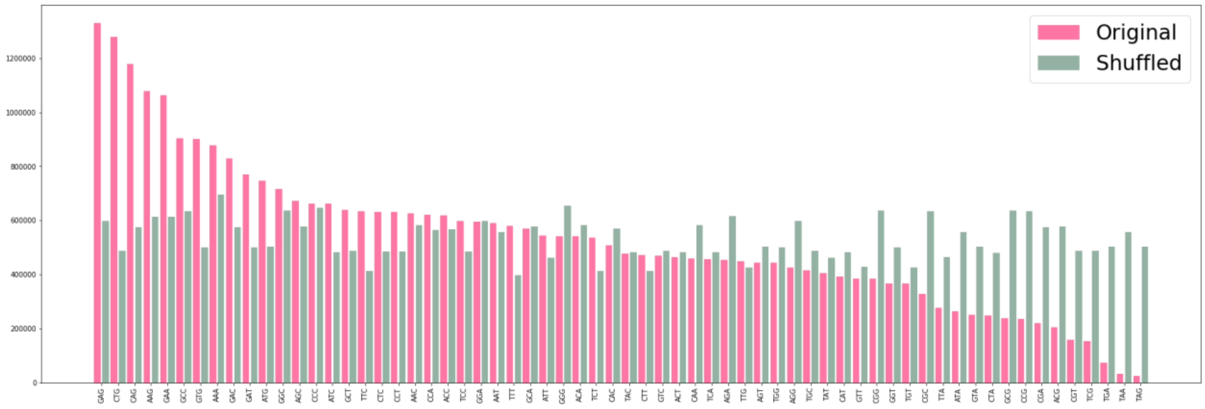

Figure 2: Codon usage comparison between observed (Original) and randomly shuffled (Shuffled) mRNA sequences.

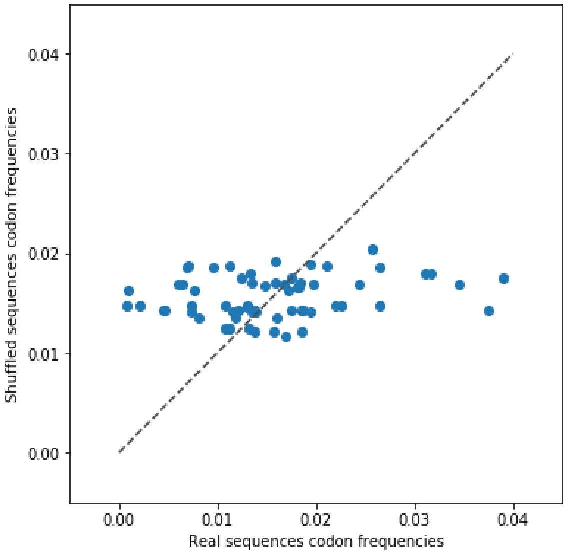

Figure 3: Scatter plot of codon frequencies between real and shuffled mRNA sequences. Pearson: *r* = 0.213, *p* > .05; Spearman: *ρ* = 0.173, *p* > .05.

The same comparisons were performed on lncRNAs and we observed a similar bias. Some codons are overrepresented or underrepresented compared to what is expected when the sequences are randomly shuffled (Figure 4). Codon usage in real and shuffled sequences do not correlate (Figure 5). Given that lncRNAs do not code for proteins, this is interesting information. Indeed, it seems that their sequence still has some kind of importance.

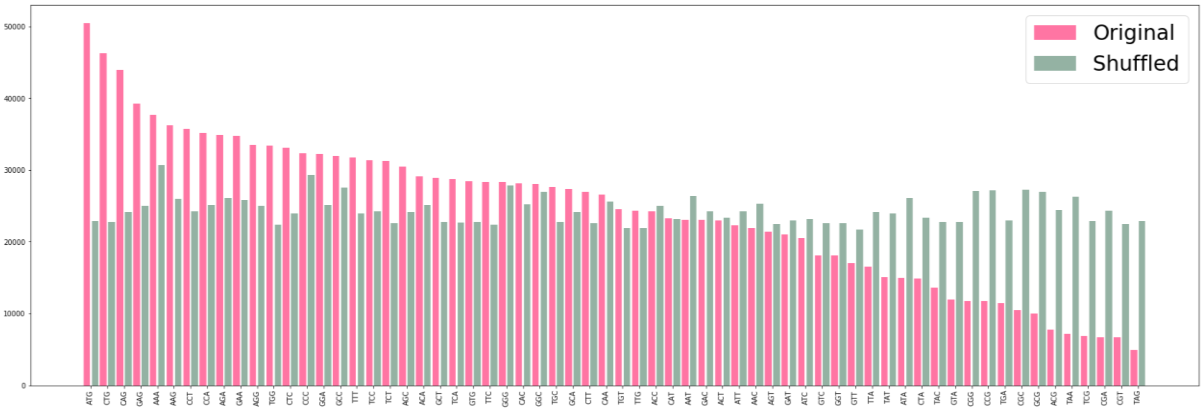

Figure 4: Codon usage comparison between observed (Original) and randomly shuffled (Shuffled) lncRNA sequences.

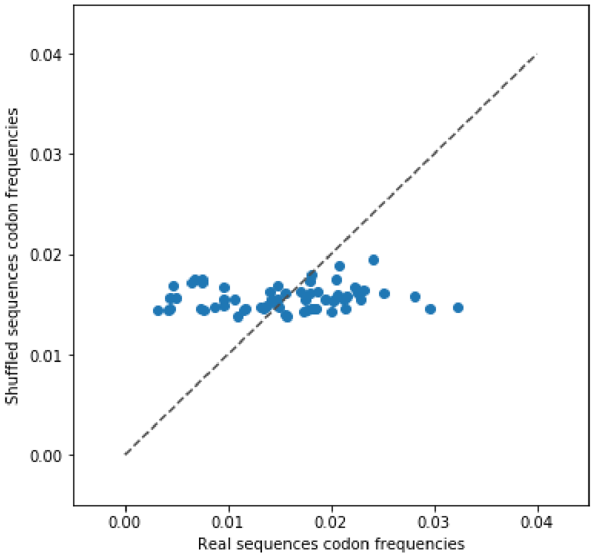

Figure 5: Scatter plot of codon frequencies between real and shuffled lncRNA sequences. Pearson: $r = 0.106$, $p > .05$; Spearman: $\rho = 0.143$, $p > .05$.

After separately analysing the mRNA and lncRNA sequences, we compared them to each other. We found that their codon frequencies correlate quite well (Figure 6). Therefore, even if lncRNAs are not translated into proteins, their codon frequencies seem to be very similar to the ones we observe in mRNAs. This could be explained by the fact that some lncRNAs were derived from mRNAs in the course of evolution and vice-versa. In addition, some lncRNAs can overlap mRNAs, which could contribute to the observation of similar codon frequencies.

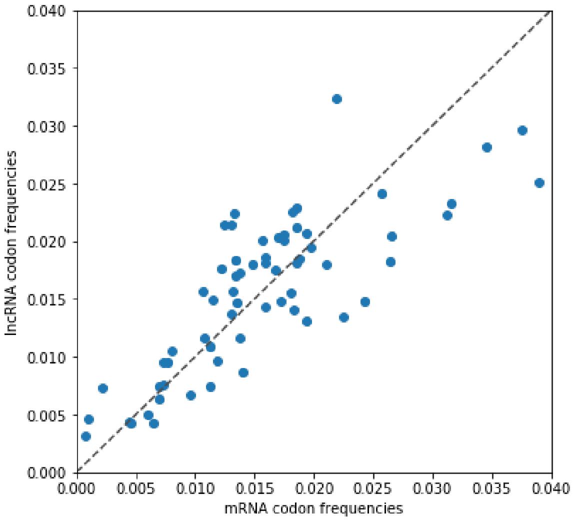

Figure 6: Scatter plot of codon frequencies between mRNA and lncRNA.
Pearson: r = 0.799, $p \ll .05$; Spearman: ρ = 0.775, $p \ll .05$.

Codon Pair Analysis

To gain a first overview of the codon pair counts and their distributions, we generated codon pair cluster maps for both, lncRNA and mRNA (Figure 7). Since mRNA data is much larger than lncRNA data and since we worked with counts, rather than with normalized frequencies, we generally observe higher counts in mRNA (compare color bars). We further note that in mRNA, only a few codon pairs are very common (small cluster in the upper left corner), whereas in mRNA the codon pair counts seem to be more narrowly distributed.

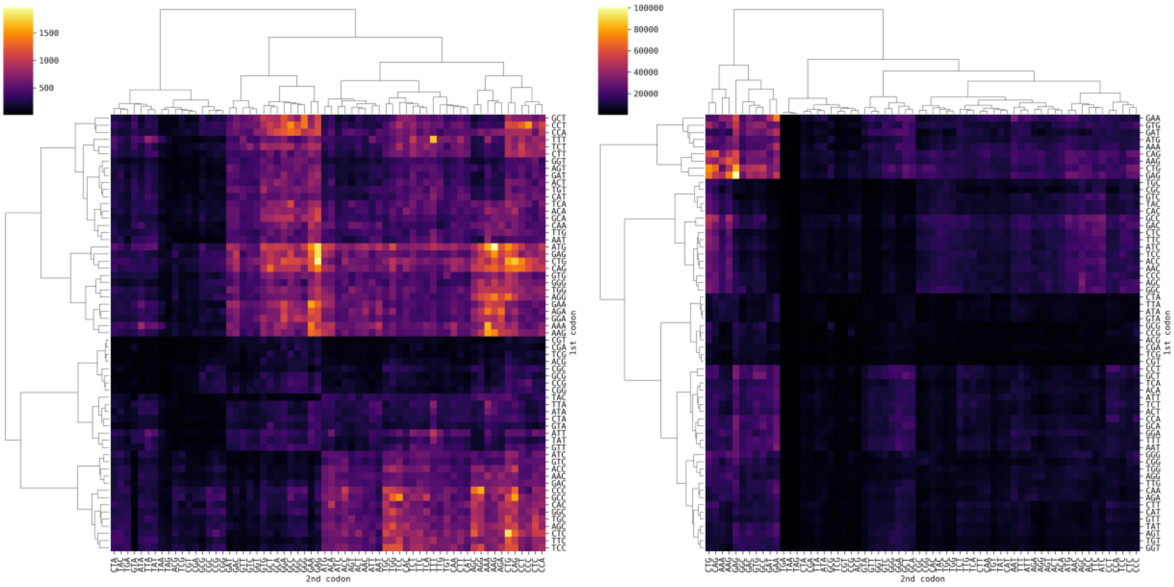

Figure 7: Total codon pair count cluster maps of lncRNA (left) and mRNA (right).
Clustering method: Ward's minimum variance method.

After calculating the adjusted residuals for each codon pair, we looked at their distributions in lncRNA and mRNA (Figure 8). The grey area corresponds to the number of codon pair residuals that do not reject independence (unbiased), whereas the coloured areas represent residuals that indicate a certain bias. Both datasets show a normal distribution with a few outliers, but most importantly we noticed that in mRNA the biases seem to be stronger, due to a decrease in the size of the grey area and an increase in variation of the residual values (x-axis) with respect to the lncRNA histogram.

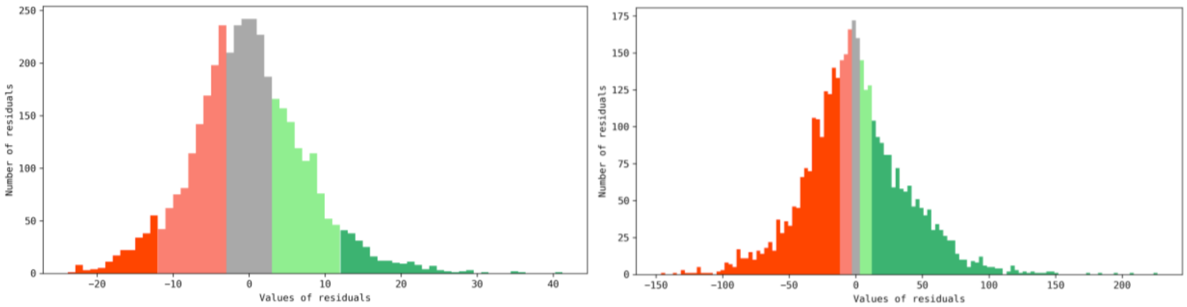

Figure 8: Distribution histograms of the adjusted residuals of codon pairs in lncRNA (left) and mRNA (right). The grey area highlights values for $d_{i,j}$ between -3 and 3, that do not reject independence. The coloured areas show residuals that are significantly biased (99.73% *CI* [-3, 3]), green indicating an overrepresentation, red an underrepresentation. The light colours were arbitrarily chosen for values between -12 and 12 and their sole purpose is to facilitate the direct comparison between the two plots.

In order to identify the biased codon pairs and their principal features, we generated two cluster maps with a diverging color scheme (Figure 9).

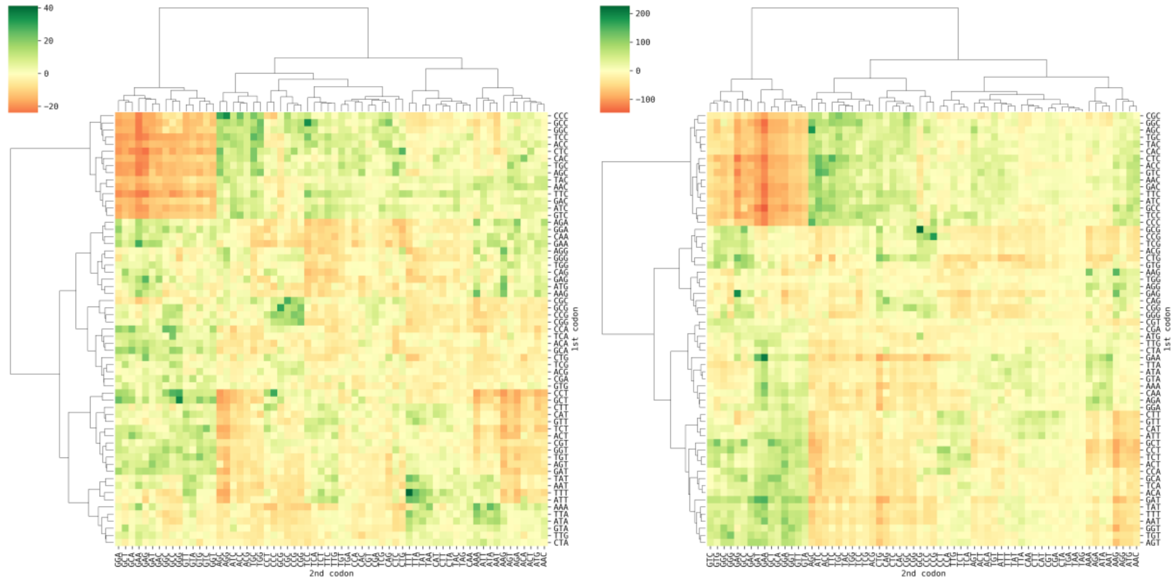

Figure 9: Cluster maps of the adjusted residuals of codon pairs in lncRNA (left) and mRNA (right). Red cells indicate codon pairs that are underrepresented, whereas green cells represent codon pairs that are overrepresented. Clustering method: Ward's minimum variance method.

On both adjusted residual cluster maps we notice a rather prominent red cluster in the upper left corner (Figure 9). A closer look reveals that all codon pairs responsible for that cluster, are of the general form 5'-nnC-Gnn-3' (codon pairs with CpG in positions 3-1) and thus represent a class of codon pairs that are consistently underrepresented in both mRNA and lncRNA. This finding is consistent with other observations made in previous studies.[4] Tulloch et al. suggest that this underrepresentation might be due to DNA methylation-induced mutations in the nucleus.[5]

In order to find differences between mRNA and lncRNA codon pair usage, we generated a differential codon pair cluster map based upon the absolute differences between the adjusted residuals (Figure 10). We noticed that there seems to be a considerable amount of differences in the usage of codon pairs that have either a GAA or a GAT as their second codon (two vertical orange bands). These codons interestingly correspond to the two negatively charged amino acids, glutamic acid (Glu) and aspartic acid (Asp), respectively. However, also other codon pairs seem to be differently used in lncRNA and mRNA, which are highlighted by the yellowish cells, e.g. 5'-GCG-GCG-3'.

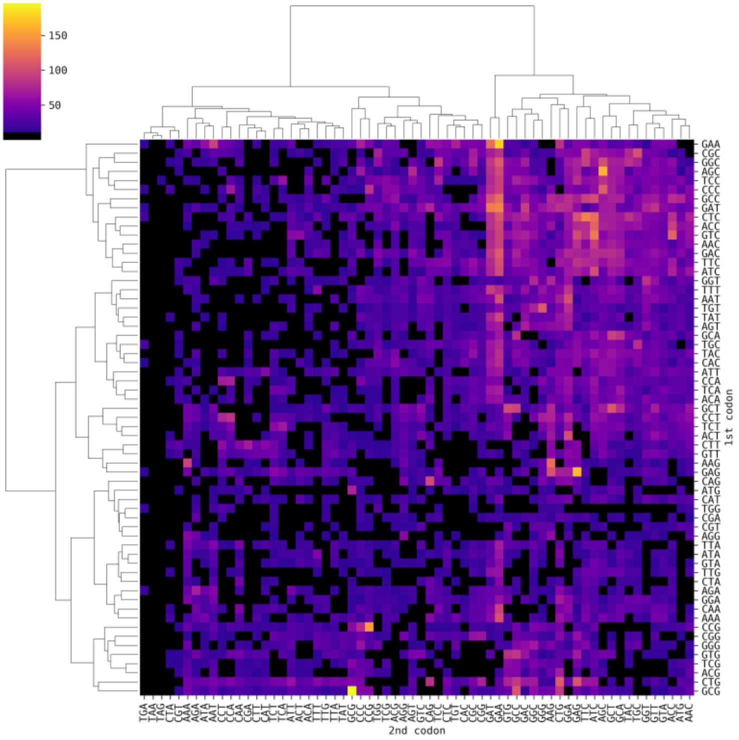

Figure 10: Differential codon pair cluster map of lncRNA and mRNA. Black cells correspond to |□$d_{i,j}$| < 15 (arbitrary threshold) and represent codon pairs that occur at a similar frequency in both, lncRNA and mRNA.

Next, we were interested in determining whether the most important differences identified on the differential codon pair cluster map, are due to opposing biases (e.g. overrepresented in lncRNA, but underrepresented in mRNA) or rather due to differences in the sheer strength of a certain bias.

To get a general idea, we created a ranking of the top ten most differently biased codon pairs in mRNA and lncRNA (Table 2). We noticed that among these top ten codon pairs, all adjusted residuals were positive in both lncRNa and mRNA. Moreover, these codon pairs seem to be significantly ($d_{i,j} > 3$) preferred by both, however much stronger so by mRNA. Even when looking at the 100 most differently biased codon pairs, no opposing biases were found. However, among those 100 codon pairs, a total of four codon pairs showed no bias in lncRNA ($-3 < d_{i,j} < 3$), while being quite strongly overrepresented in mRNA.

($84 < d_{i,j} < 96$). Those four codon pairs are 5’-CGC-TTC-3’ (Arg>*Phe*), 5’-TTT-GAT-3’ (*Phe*>Asp), 5’-TTT-GGA-3’ (*Phe*>Gly), and 5’-GGC-TAC-3’ (Gly>*Tyr*). Strikingly, all four codon pairs contain a codon that translates into an aromatic amino acid (in italic). This bias might be an indication that the context of the aromatic amino acids is of some importance in proteins.

Looking at Table 2, we noticed another interesting fact. Six out of the ten codon pairs are homodimers, which are among the most overrepresented codon pairs in both mRNA and lncRNA. However, further analyses are required, in order to understand the broader context of those homodimers, such as whether they form long repeats (homopolymers) or are involved in specific protein domains or secondary structures.

Rank	Absolute Difference	Codon Pair	Dipeptide	Adjusted Residual lncRNA	Adjusted Residual mRNA
1	195.983	GCG > GCG	Ala > Ala	30.518	226.501
2	177.201	GAA > GAA	Glu > Glu	28.943	206.144
3	167.838	GAG > GAG	Glu > Glu	26.150	193.988
4	161.706	AGC > AGC	Ser > Ser	20.617	182.323
5	152.639	CCG > CCG	Pro > Pro	21.622	174.260
6	136.856	GAT > GAA	Asp > Glu	12.969	149.825
7	136.231	CTC > TTC	Leu > Phe	11.489	147.720
8	131.995	GAT > GAT	Asp > Asp	13.080	145.075
9	130.191	GTC > ATC	Val > Ile	15.971	146.162
10	123.802	CTC > ATC	Leu > Ile	8.292	132.094

Table 2: Top ten ranking of codon pairs that show the greatest difference in bias between lncRNA and mRNA.

Codon pairs starting with ATG

We compared the frequencies of the 5’-ATG$_{\text{START}}$-nnn-3’ codon pairs at the beginning of sequences with the frequencies of 5’-ATG-nnn-3’ codon pairs that occur within the sequences in mRNAs and lncRNAs. We hypothesized that in mRNAs, the first pair of codons might be biased, because it is involved in the start of the translation. The ribosome needs an ATG codon to start translating an mRNA sequence. As the second codon is also inside the ribosome at the start of translation, it might also play a role in this process.

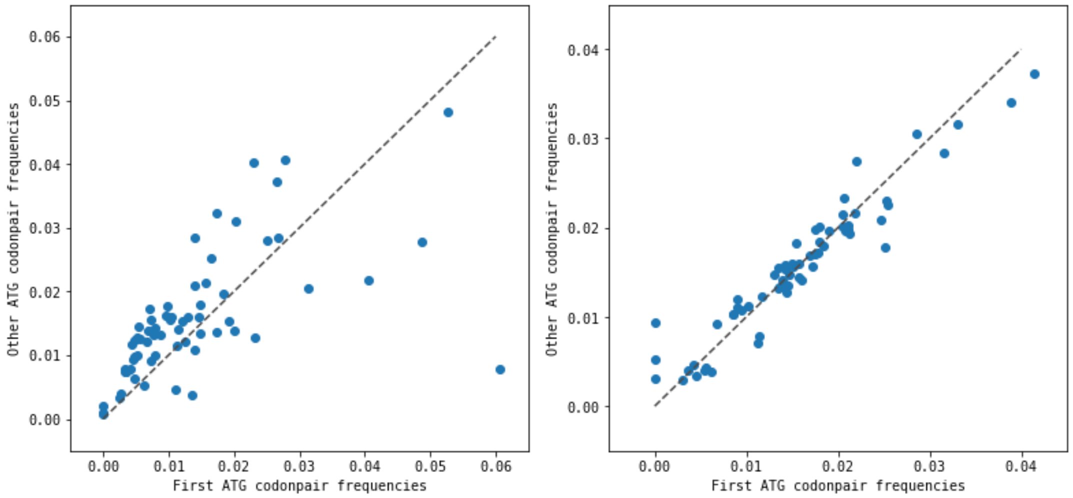

Figure 11: Codon pair usage comparison between codon pairs with an ATG as first codon located at the beginning of a sequence (ATG = START codon) and other codon pairs with an ATG as first codon located within sequences. mRNA (left): Pearson: $r = 0.620$, $p \ll .05$; Spearman: $\rho = 0.749$, $p \ll .05$.
lncRNA (right): Pearson: $r = 0.956$, $p \ll .05$; Spearman: $\rho = 0.961$, $p \ll .05$.

To answer this question, we compared, in mRNAs and in lncRNAs, the codon pair usage in the first pair of codons and in the other pairs of codons starting with ATG and located elsewhere in the sequences. We observed that in both mRNAs and lncRNAs, these two datasets are correlated (Figure 11). Nevertheless, we noticed an interesting difference: in mRNAs, there is much more variance than in lncRNAs, and the correlation coefficients are lower. We think this could be explained by the fact that mRNAs are translated into proteins. In this dataset, some pairs of codons are much more frequent at the beginning of the sequence than elsewhere in the sequence. This bias might be due to some translation reason, and it would be really interesting to see which are these overrepresented codons.

Conclusion

At the end of this project, we investigated numerous features of single codons and codon pair usage in mRNAs and lncRNAs. First, we observed that they were strongly biased compared to what is expected by chance in both types of transcripts. Then, we discovered that the single codons frequencies correlate in mRNAs and in lncRNAs, which shows that they are unexpectedly similar. We also obtained an interesting result by comparing the sequences' first codon pair with other codon pairs starting with "ATG" located elsewhere. Indeed, in lncRNAs, those two datasets are nearly perfectly correlated, whereas in mRNAs there is much more unexplained variance.

There are many ways we could go further with this project. We could investigate the clusters we identified during the codon pair analysis, and compare them between mRNAs and lncRNAs. For example, we noticed that the codon pairs containing the GC dinucleotide at the position 3 and 4 were evenly underrepresented. We think it would be interesting to investigate this further to understand it.

During this project, we only looked at the codons and codon pairs in the DNA context. To understand better the biological meaning of our results, we could study the broader context of preferred codon pairs, and the corresponding amino acid. This could allow us to better understand the effects of such biases on proteins.

References

1. *Long non-coding RNA*, Wikipédia, consulted on the 05.06.2020 https://en.wikipedia.org/wiki/Long_non-coding_RNA
2. *Epigenetics,* M.C. Gambetta, University of Lausanne
3. Moura G, Pinheiro M, Silva R *et al.* Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biol*. 6 R28 (2005).
4. Haberman SJ. Analysis of residuals in cross-classified tables. *Biometrics*. 29: 205-220 (1973).
5. Buchan JR, Aucott LS, Stansfield I. tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Research*. Vol. 34, Issue 3: 1015–1027 (2006).
6. Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *Elife*. (2014).