

University of South Bohemia
Faculty of Science
České Budějovice, Czech Republic
and
Johannes Kepler University
Faculty of Engineering and Natural Sciences
Linz, Austria

**Characterization of oocyte- and embryo-specific 5' untranslated
regions (UTRs) of mouse mRNAs**

Bachelor Thesis

Alexandra Austenová

Supervisor: Mgr. Lenka Gahurová, Ph.D.

České Budějovice

2020

Austenová A., 2020: Characterization of oocyte- and embryo-specific 5' untranslated regions (UTRs) of mouse mRNAs. Bc. Thesis, in English. -75p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

Annotation

To identify, annotate and characterize the oocyte- and embryo-specific 5' UTRs, RNA-seq datasets from various developmental stages of mouse were processed, followed by de novo transcriptome assembly, filtering, and analysis of the length, GC content, presence of uORFs and identification of enriched sequence motifs.

I hereby declare that I have worked on my bachelor 's thesis independently and used only the sources listed in the bibliography. I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full form to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

In České Budějovice, 20.5.2020,

A handwritten signature in cursive script, appearing to read "Austenová", written above a horizontal dotted line.

Signature

Table of Contents

1	ABSTRACT.....	1
2	INTRODUCTION	1
3	BACKGROUND	2
3.1	5' UTR	2
3.2	Role of 5' UTR in Translation.....	4
3.2.1	Translation Initiation.....	4
3.2.2	Upstream AUG Codon.....	5
3.2.3	Secondary Structures	6
3.2.4	microRNA-mediated Gene Regulation.....	7
3.2.5	RNA Binding Proteins	7
3.2.6	Internal Ribosomal Entry Site.....	8
3.2.6.1	IRES Trans-Acting Factors.....	9
3.2.6.2	uORF.....	10
3.2.6.3	RG4 Structures.....	10
3.2.7	Terminal Oligo-Pyrimidine Structures	11
3.2.8	RNA-G Quadruplexes.....	11
3.3	5' UTR and its Effect on Embryonic Development.....	12
4	AIMS.....	14
5	WORKFLOW OVERVIEW.....	15
6	MATERIALS AND METHODS.....	16
6.1	Datasets, Trimming and Mapping.....	16
6.2	De Novo Transcriptome Assembly.....	16
6.3	Generating Annotations of All Genes	17
6.4	Identification of Annotated Genes with Novel Upstream 5' UTRs.....	18
6.5	Removal of False Positives	19
6.6	Association of De Novo Assembled Genes with Officially Annotated Genes.....	20
6.8	Length and GC Content Analysis.	22
6.9	uORFs	22
6.10	Sequence Motifs.....	23
7	RESULTS	24
7.1	De Novo Transcriptome Assembly.....	24
7.2	Identification of genes with Novel Upstream 5' UTRs	27
7.3	Association of De Novo Assembled Genes with the Officially Annotated Genes.....	30

7.4	Identification of Oocyte- and 2C Embryo-specific 5' UTRs.....	33
7.5	Length of Oocyte- and 2C Embryo-specific 5' UTRs.....	34
7.6	GC Content of Oocyte- and 2C Embryo-specific 5' UTRs	39
7.7	uORFs	41
7.8	Sequence Motifs.....	43
8	DISCUSSION	50
9	CONCLUSION.....	54
10	REFERENCES	55
11	APPENDICES	64

1 ABSTRACT

The 5' UTR is a region located upstream of the coding sequence of mRNA. It is known for exerting translational control via interactions of the structures present in the 5' region, and it has a significant impact on the resulting gene expression. The official annotation of 5' UTRs mostly represents the 5' UTRs of mRNAs in somatic tissues. In order to identify oocyte- and embryo-specific 5' UTRs longer than annotated 5' UTRs, we processed the RNA-seq data from various mouse developmental stages including oocytes, preimplantation embryos, and somatic tissues. We performed de novo transcriptome assembly and generated new annotations for all genes. Following a series of filtering and analyses, we identified a list of genes with upstream 5' UTRs that had the potential to be oocyte- and embryo-specific. We observed that the oocyte-specific 5' UTRs were on average 683.8 bp longer than the officially annotated 5' UTRs, the embryo-specific 5' UTRs were on average 746.5 bp longer than the officially annotated 5' UTRs, and both oocyte- and embryo-specific 5' UTRs were more GC-rich than the officially annotated 5' UTRs. We also identified that they contain multiple recognition sites of miRNAs but no obvious binding sites for RNA-binding proteins based on sequence analysis. Finally, we discovered the uORF content to be lower for oocyte- and embryo-specific 5' UTRs than that of the official annotated 5' UTRs.

2 INTRODUCTION

Messenger RNA (mRNA) is an essential component of any proteo-synthetic process. Its primary function is to serve as a template for protein synthesis. It consists of a coding sequence (CDS) containing the information translated into proteins, and 5' and 3' untranslated regions (UTRs) that are non-coding, and hence do not participate directly in the protein production. Instead, the UTRs exert indirect control, for example through binding with specific RNA binding proteins. These interactions have a significant impact on the resulting gene expression, and the interplay between these structures is strictly controlled. During development and cell differentiation, each change in these interactions has the potential to affect crucial biological processes (Jackson et al., 2010; Sonenberg and Hinnebusch, 2009; Gebauer and Hentze, 2004).

Despite their biological importance, the transcriptomes of mammalian oocytes and early embryos are not well annotated due to the low amount of available material. Therefore, the official annotation of 5' UTRs (in genome browsers such as Ensembl, UCSC, or NCBI) mostly represents 5' UTRs of respective mRNAs in somatic tissues.

Nevertheless, it appears that transcriptomes of oocytes and early embryos differ from somatic tissues (Veselovska et al., 2015; Zhang et al., 2014a). This also includes substantial changes in the length and sequence of the 5' UTRs of the mRNAs (Veselovska et al., 2015). For example, mRNAs that are transcribed during the developmental phases tend to have longer than average 5' UTRs (Kozak, 1987). Each developmental phase also requires specific transcription factors and growth factors, and so the appropriate pattern of the 5' UTRs of their mRNAs is of great importance for their precise regulation. Both 5' and 3' UTRs have been known to exert control over posttranscriptional regulation by affecting stability, subcellular localization, and translatability (Jansen, 2001; Lin and Li, 2011).

This project aims to identify oocyte-specific and embryo-specific 5' UTRs that are longer than somatic 5' UTRs. For this purpose, we used publicly available deep RNA-seq data from mouse, as it is a classic mammalian model organism. Furthermore, we characterized the length, GC content, RNA binding proteins, miRNA recognition sites, and uORFs for the oocyte- and embryo-specific 5' UTRs.

3 BACKGROUND

3.1 5' UTR

The 5' UTR refers to the untranslated region that is upstream of the coding sequence on the 5' end of the mature mRNA. Due to its location, it is the ideal candidate region for exerting translational control via interactions of the structures present in the 5' region. These structures include 7-methyl-guanosine cap (7-meG cap), hairpin-like secondary structures, upstream open reading frames, terminal oligopyrimidine tracts, G-quadruplexes, and internal ribosome entry sites (Dvir et al., 2013).

Over the span of evolution, 5' UTRs have undergone a series of elongations, leading to an effective increase in the number of regulatory elements (Chen et al., 2011). This may have had an effect on the emergence of organismal complexity; however, no correlation has

been found between the organismal complexity and the length of the 5' UTR (Chen et al., 2011). The genome average of 3' UTR ranges from 100 bp to 800 bp over diverse taxonomic classes (Pesole et al., 2001; Mignone et al., 2002; Nagalakshmi et al., 2008). Compared to 3' UTR, the length of the 5' UTR in eukaryotes tends to be less variable over diverse taxonomic classes, consisting of an average of 100 to 200 nucleotides (Mignone et al., 2002; Pesole et al., 2001; Lin and Li, 2011). The opposite is true when the focus is on the length of 5' UTR of individual genes in a genome. Here the variation in length ranges greatly from a few nucleotides to thousands (Nagalakshmi et al., 2008; Pesole et al., 2001; Lin and Li, 2011). Strikingly, even as little as a single nucleotide may serve as the 5' UTR for the translation initiation (Hughes and Andrews, 1997).

Still, some trends may be observed. Specifically, distinct lengths of UTRs can be associated with genes with different functions (Lin and Li, 2011). For example, the vertebral transcripts of transcription factors, protooncogenes, growth factors, and receptors tend to all have a 5' UTR that is longer in length than the average cellular 5' UTR (Davuluri et al., 2000; Kozak et al., 1987). This can be interpreted as due to the fact that the greater length corresponds to a greater level of regulation, as was revealed by genome-wide surveys of transcript boundary with microarrays in *Saccharomyces cerevisiae* (Hurowitz and Brown, 2003; David et al., 2006).

Moreover, variations in the 5' UTRs of the transcripts of genes are not only common, but they also act as important switches for gene expression regulation. It has been estimated that 10-18% of genes utilize multiple promoters to express alternative 5' UTRs, whereas 13% of genes in mammalian transcriptome are affected by alternative splicing within their UTRs (Trinklein et al., 2003; Zhang et al., 2003; Carninci et al., 2005).

5' and 3' UTRs also differ in their respective GC content. Overall, the average GC content of 5' UTR is greater than that of 3' UTR sequences (Pesole et al., 2000). Its percentage is rather conserved across species, and its increase is known to affect translation efficiency in an inhibitory way (Pelletier and Sonenberg, 1985; Babendure et al., 2006). Specifically, it is thought to be the culprit behind inefficient scanning by the 43S pre-initiation complex that contributes to an overall lower rate of translation initiation (Taliaferro et al., 2016). Since GC bonds are more difficult to melt than AU hairpins, they confer higher stability per base (Babendure et al., 2006). Consequently, transcripts coding for regulatory proteins have 5' UTRs with high GC content, approximately 70-90% (Kozak, 1991). This

demonstrates the fact that the 5' UTR composition may serve as a mechanism for gene regulation (Babendure et al., 2006).

Moreover, observations suggest an inverse relationship between the length of 5' UTR with respect to its GC content. Accordingly, genes found in heavy isochores (i.e. regions of denser GC content) tend to have 5' UTRs of shorter length (Pesole et al., 1999). Similarly, such correlation has also been observed for the coding sequences and introns (Duret et al., 1995).

Another feature of eukaryotic 5' UTRs is their possession of different types of repeats, including short and long interspersed elements (SINEs and LINEs), simple sequence repeats (SSRs), minisatellites and macrosatellites (Jurka, 1998; Smit, 1999). The occurrence of these is greater in human and rodent mRNAs than in mRNA of other mammalian species (Pesole et al., 2000). For example, 12-15% of the rodent and human 5' UTR consists of repeats, while the 5' UTR of other mammals consists of only approximately 8% of repeats (Pesole et al., 2000). However, it is not known whether these elements confer any functions to the UTR regions (Pesole et al., 2000).

3.2 Role of 5' UTR in Translation

3.2.1 Translation Initiation

The role of 5' UTR in translation regulation is specifically in the step of translation initiation that requires a number of protein factors, especially the eukaryotic initiation factors (eIFs). These factors participate in the initiation process directly by recognizing the cap structure and by allowing the binding of the 40S ribosomal subunit (Weaver et al., 2008) and have the ability to alter the rate of protein synthesis. To properly initiate eukaryotic translation, at least eleven different eIFs are necessary (Hershey and Merrick, 2000).

Classically, the step of translation initiation begins at the 5' end of the transcript that carries the 7-meG cap, when the eIF4E binds to it (Poulin et al., 2000). This leads to the assembly of the eIF4F protein complex that is composed of subunits eIF4E, eIF4A, and eIF4G. Alternatively, eIF4F binds to the 7-meG cap as a complex.

eIF4A acts as an ATP-dependent binding and unwinding machinery of mRNA. The factor is crucial for the binding of the 40S complex, and hence changes in its activity can

have a substantial impact. Such changes have been suggested to be caused by sequence motifs as well as local RNA structures (Wolfe et al., 2014; Rubio et al., 2014; Iwasaki et al., 2016).

eIF4B aids eIF4A with the binding of the complex to mRNA through eIF3 that in turn interacts with the eIF4G subunit of the eIF4F complex (Lamphear et al., 1995). Once bonded, the 40S complex is allowed to start scanning along the mRNA for the AUG start codon that signals the initiation of protein synthesis.

3.2.2 Upstream AUG Codon

The AUG codon is flanked by a non-random sequence that is conserved and well defined for individual species. Interestingly, the AUG codons are not restricted to the CDS and are found also upstream of CDS, in the 5' UTR. It has been observed that in such cases, the 40S ribosomal subunit can adapt accordingly by the employment of a mechanism called leaky scanning.

Leaky scanning occurs when the scanning complexes ignore the most upstream AUG due to its less than optimal surrounding sequence and continue searching for the sequentially second AUG with a stronger context (Araujo et al., 2012; Wang et al., 2004). In the case of vertebrate mRNAs, the context of an AUG codon is the most optimal if it matches the Kozak sequence GCCA/GCCAUGG (Kozak, 2002). Roughly 10% of ribosomes participate in leaky scanning, initiating at a downstream AUG instead (Wang et al., 2004). This mechanism not only allows single mRNA to produce multiple different proteins through the use of alternative AUG codons but also may act as a negative regulator of translation for a proportion of ribosomes (Oyama et al., 2004; Xiong et al., 2001).

An uAUG is defined as an upstream start codon without an in-frame stop codon, resulting in translation of the alternative protein isoform as described in the previous paragraph. However, in situations where there is a stop codon following an uAUG before the main start codon, it forms an upstream ORF (uORF) and can result in a short translated peptide.

Once such short peptide is translated, the 40S ribosomal subunit may either remain bound to the mRNA, resume search for the next AUG start codon, reinitiate translation at a downstream AUG or it may simply unwind and leave (Mignone et al., 2002). Although

translation reinitiation is possible, it is limited by the length of the uORF as well as the stop codon context (Cassan and Rousset, 2001). An uORF longer than 30 codons with intercistronic spacer that is ≤ 50 nt in length prevents the ribosome from reinitiating, in a process known as down-regulation of translation (Luukkonen et al., 1995; Child et al., 1999; Lincoln et al., 1998). It is a common notion that the peptides produced by uORF may hamper the translation initiation process at downstream ORFs due to the ribosome stalling at the end of the uORF (Oyama et al., 2004). However, this notion was challenged in a 2009 study by Calvo, where no correlation was found between the impact of the uORF on the expression of the downstream gene and the distance between the uORF and the coding sequence (CDS) (Calvo et al., 2009; Barrett et al., 2012).

The creation and deletion of uORFs is also often associated with mutations. Such mutations may, therefore, affect the number of protein products, resulting in the development of or predisposition to certain diseases (Wethmar et al., 2010; Chatterjee and Pal, 2009). On average, uORFs reduce protein expression by 30-80% (Calvo et al., 2009).

3.2.3 Secondary Structures

Secondary structures are a common feature of the 5' UTRs, found especially in mRNAs encoding transcription factors, protooncogenes, and growth factors (Araujo et al., 2012). They are characterized by high GC content, as well as highly negative free folding energy (ΔG) (Leppek et al., 2018). Formed by intrastrand interactions, they are capable of effectively blocking protein translation.

Whether a secondary structure causes inhibition or not depends on its position relative to the 7-meG cap structure as well as its free energy. In general, an increase in the expected thermal stability of mRNA leads to a decrease in translation. On average, a stable secondary structure has a free energy of less than -50 kcal/mol (Araujo et al., 2012), and it was observed that the greatest decrease in translation occurs when stabilities increase from ΔG of -25 to ΔG of -35 kcal/mol (Babendure et al., 2006).

The impact of the hairpin position on the translation should not be underestimated. The closer the hairpin is to the 7-meG cap, the more effectively it will block translation (Kozak, 1989). More specifically, hairpins located at positions +4 or less are inhibitory even in case of very stable secondary structures, whereas those found in the positions +31 to +46

did not participate in such inhibition, unless they were weaker than -30 kcal/mol (Babendure et al., 2006; Gray and Hentze, 1994; Pickering and Willis, 2005).

The inhibitory effects of these structures may be counteracted by an increase in the level of eIF4A, the subunit of eIF4F complex that is responsible for unwinding the RNA secondary structures (Svitkin et al., 2001).

3.2.4 microRNA-mediated Gene Regulation

MicroRNAs (miRNAs) are a 20-24 nt long class of highly conserved non-coding RNA molecules involved in the regulation of gene expression (Bartel, 2009; MacFarlane and Murphy, 2010). It is predicted that at least 30% of protein-coding genes in humans are regulated through miRNA (Rajewsky, 2006). These RNA molecules base-pair with sequences in mRNA transcripts, resulting in gene silencing through translational repression and/or degradation of RNA (Bartel, 2009; MacFarlane and Murphy, 2010).

Interestingly, mRNA-miRNA interaction affects gene expression in the early stages of translation (Djuranovic et al., 2012). An increase in the number of mRNA secondary structures allocated near the 5' cap was observed to correlate with miRNA-mediated gene regulation in animals (Gu et al., 2014). The region proximal to the 5' cap, specifically the region 30-50 nt downstream from the 5' cap, acts as the binding platform during the formation of the 43S pre-initiation complex (Araujo et al., 2012). Several 2013 studies hinted that the miRNA represses gene translation through the impairing of the function of the pre-initiation complex (Meijer et al., 2013; Ricci et al., 2013). This leads to the suggestion that the mRNA 5' UTR secondary structures are important for miRNA-mediated gene silencing and that the mRNAs containing unstructured 5' UTRs are not affected by the miRNA repression (Meijer et al., 2013).

3.2.5 RNA Binding Proteins

One of the most important functions of RNA binding proteins (RBPs) is their ability to positively or negatively affect the translational efficiency of a specific mRNA with respective RBP binding sites in their 5' UTR (Moore and Lindern, 2018).

An example of such RBP-mediated regulation is offered by the iron regulatory proteins (IRPs). IRPs respond to intracellular iron concentrations through the regulation of mRNAs that contain an iron-responsive element (IRE); a highly conserved stem-loop structure consisting of approximately 30 nucleotides (Wilkie et al., 2003; Araujo et al., 2012). This regulation is necessary for maintaining balance in concentrations of cellular iron, and mutations within IRE can lead to certain diseases (Girelli et al., 1997).

The translational inhibition of IRP-IRE complexes is mediated in 2 ways. When IRP binds to IRE, it inhibits the translation of a downstream ORF (Stripecke et al., 1994). For this inhibition to be efficient, IRE must be localized in a position proximal to the 7-meG cap, but its distance from the AUG codon does not appear to play a role (Stripecke et al., 1994; Paraskeva et al., 1999). The inhibition is caused due to steric inhibition that prevents the 40S subunit from binding to the mRNA, while the binding of eIFs remains unaffected (Wilkie et al., 2003).

However, in situations when the IRE-IRP complex is in a location more distant from the cap, inhibition does not have an impact on the binding of the 40S subunit. Instead, the complex blocks the 40S from scanning the mRNA for the AUG codon and consequently inhibits translation (Wilkie et al., 2003).

3.2.6 Internal Ribosomal Entry Site

Internal Ribosomal Entry Site (IRES) is found in approximately 10-15% of mammalian mRNA sequences (Spriggs et al., 2008). While most cellular mRNAs undergo cap-based translation initiation, IRES offers an alternative cap-independent mechanism that may be used in cases when cap-based translation is inhibited (Pelletier et al., 1988). Such cases include stress, embryonic development, mitosis, or apoptosis (Komar and Hatzoglou, 2011). Consequently, these IRES-containing mRNAs mostly encode regulatory proteins, especially proto-oncogene products and their receptors (Mignone et al., 2002). These elements have only recently gained attention in the area of developmental biology, where they emerged as crucial regulators of gene expression (Xue et al., 2015).

Compared to the viral IRESs, IRESs of cellular origin appear to be less structured and less stable in terms of Gibbs free energy of the folded mRNA (Komar and Hatzoglou, 2005; Komar and Hatzoglou, 2011; Xia and Holcik, 2009). Computational modeling in *FGF-2*,

BiP, and *VEGF* IRESs revealed the common structure of IRES to be a Y-shaped stem-loop followed by a small stem-loop upstream of the AUG start codon, located within the 5' UTR (Stein et al., 1998; Le and Maizel, 1997). Speculations suggest that the specifics behind the cellular IRES structure aid in binding of the ribosome to the cellular mRNAs, however, this is yet to be proven (Velden and Thomas, 1999; Martineau et al., 2004; Godet et al., 2019). Some cellular IRES were also observed to contain pseudoknots (Quesne and Le, 2001; Jopling et al., 2004).

While the mechanisms of the viral IRES are becoming better understood, not much is known about the function of cellular IRES (Komar and Hatzoglou, 2011). Even though the computational modeling revealed above-mentioned complex structures that are common to contain stem-loops, there has not been a common sequence or structural motif classified that would allow for cellular IRES prediction from an mRNA sequence, except in the case of short RNA sequences such as the *Gtx* 9-nucleotide motif (Le et al., 2003; Xia and Holcik, 2009; Komar and Hatzoglou, 2011; Chappell et al., 2000). Hence in most cases, the existence of the IRES must be proved experimentally (Andreev et al., 2009; Komar and Hatzoglou, 2005).

According to the type of factors or elements that interact with cellular IRES structures, we can distinguish between IRES trans-acting factors (ITAFs), uORFs, and RNA G-quadruplex (RG4) structures (Leppek et al., 2018).

3.2.6.1 IRES Trans-Acting Factors

ITAFs, many of which belong to a group of heterogeneous ribonucleoproteins, represent a group of RBPs that model the activity of IRESs (Komar and Hatzoglou, 2005; Lewis and Holcik, 2008; Spriggs et al., 2005). Specifically, the IRES-ITAF interactions not only contribute to the stability of the IRES itself, but they are also capable of inducing a conformational change of the IRES RNA (Leppek et al., 2018). These structural changes are necessary for the recruitment of ribosome subunits without the presence of the 7-meG cap (Leppek et al., 2018). The precise mechanism behind the activation of cellular IRES by ITAFs is not known (Bradshaw and Stahl, 2016). However, the ITAF binding to cellular IRESs can be illustrated via PTB. PTB is an ITAF of the *APAF1* IRES. Together with its neuronal variant nPTB, PTB and nPTB have 2 binding sites (Mitchell et al., 2003). The major

one is located in the 3' loop part of the *APAF1* IRES, while the other binds to a purine-rich loop found in an upstream IRES domain (Mitchell et al., 2003).

3.2.6.2 uORF

Alternatively, short uORFs may affect the IRES structures via ribosome stalling (Somers et al., 2013). This may result in both activation and repression of the IRES activity (Yaman et al., 2003; Fernandez et al., 2005; Bastide et al., 2008; Fernandez et al., 2002). The relationship between uORFs and IRESs is exceptionally crucial since it actively participates in the regulation of translation when it comes to the areas of differentiation and cell growth (Yaman et al., 2003; Fernandez et al., 2005; Chen et al., 2014; Bastide et al., 2008).

An example can be provided by uORFs found upstream of *CAT1* and *FGF9* IRESs, both of which are found in mRNAs that encode regulators of differentiation and cell growth (Leppek et al., 2018). In case of *CAT1* mRNA (an arginine-lysine transporter), following amino acid starvation, the uORF translation causes structural remodeling that results in the unfolding of the 5' UTR inhibitory structures (Fernandez et al., 2000; Yaman et al., 2003; Fernandez et al., 2005; Fernandez et al., 2002). Moreover, the remodeling also facilitates a switch to a translationally active state of the IRESs (Yaman et al., 2003; Fernandez et al., 2005; Fernandez et al., 2002).

The *FGF9* (fibroblast growth factor 9) illustrates the opposite phenomenon. In normal conditions, the translation of the uORF upstream of the IRES is greater than the translation of the *FGF9* mRNA. However, in the case of hypoxia, the FGF9 protein levels increase due to a switch to IRES dependent translation (Fernandez et al., 2002).

3.2.6.3 RG4 Structures

In addition to ITAFs and uORFs, the RG4s also aid in the process of cap-independent translation. The exact importance of these structures is yet to be further examined; however, it is known that the RG4s are fully functional parts of the IRESs themselves (Morris et al., 2010; Cammas et al., 2015). According to the in vitro footprinting and structure mapping, amongst its greatest attributes is its ability to recruit the 40S ribosomal subunit

(Bhattacharyya et al., 2015). This makes it possible to proceed with translation in a cap-independent manner.

3.2.7 Terminal Oligo-Pyrimidine Structures

Terminal oligo-pyrimidine (TOP) genes belong to a special set of eukaryotic cis-regulatory genes that participate in translational control. The term directly refers to a 4-15 bp long oligo-pyrimidine tract at the 5' end of RNAs encoding ribosomal proteins or translation elongation factors (EFs) (Amaldi and Pierandrei-Amaldi, 1997; Meyuhas et al., 1996; Meyuhas and Hornstein, 2000). There is conflicting evidence whether the TOP motif has to be directly following the 5' cap, or if it can be also localized more downstream within the 5' UTR, or contain few purine bases within polypyrimidine tract (Thoreen et al., 2012; Levy et al., 1991). Despite their role in translational regulation, the number of TOP-containing genes is not well documented. Furthermore, it appears that other genes not directly associated with translation also contain TOP or TOP-like motif and their translation might be regulated in the same way as of TOP-containing ribosomal and EFs mRNAs (Thoreen et al., 2012; Hsieh et al., 2012).

In normal conditions, most mRNAs are bound by polysomes, suggesting they are actively translated (Amaldi and Pierandrei-Amaldi, 1997; Meyuhas et al., 1996; Meyuhas and Hornstein, 2000; Krichevsky et al., 1999). However, if a cell experiences starvation or specific chemical treatment, most TOP mRNAs change into the inactive state through the release of their ribosomes, whereas the state of non-TOP mRNAs remains unchanged (Yamashita et al., 2008). TOP-containing RNAs are translated through cap-dependent translation initiated by binding of eIF4E to the 5' cap. mTOR signaling, for example during starvation, prevents binding of eIF4E to the cap, resulting in the inhibition of translation of these RNAs at the growth arrest of cells (Thoreen, 2017).

3.2.8 RNA-G Quadruplexes

Certain G-rich RNA sequences are capable of forming stable structures known as G-quadruplexes (Kumari et al., 2008). These structures are thought to be present in many mRNAs and participate in the regulation of the level of translation of their host gene, as well

as inhibition of translation by small molecules (Bugaut and Balasubramanian, 2012). The way in which the regulation is exerted is similar to the way seen in stable RNA hairpin structures in the 5' UTRs (Bugaut and Balasubramanian, 2012). More specifically, it has been reported that the presence of RNA-G quadruplexes within mRNA 5' UTRs corresponds to a decrease in translation efficiency as compared to mRNAs with lacking or incomplete G-quadruplexes (Schaeffer et al., 2001). However, in some cases the existence of these structures has been proved to act oppositely, i.e. promoting translation, just as it has been seen in an analysis by Bonnal et al. in 2003 where such structure found within an IRES was revealed as the determinant of IRES activity (Bugaut and Balasubramanian, 2012; Bonnal et al., 2003).

It has also been suggested that single nucleotide polymorphisms (SNPs) within 5' UTR G-quadruplex-forming sequences can lead to differential translational activity between individuals (Beaudoin and Perreault, 2010).

3.3 5' UTR and its Effect on Embryonic Development

The prenatal development of the mouse, a classical mammalian model organism, begins with a single fertilized egg, a zygote. Subsequently, the cell undergoes a series of cell divisions combined with differentiation until a complex multicellular organism is formed in a span of approximately 19 days (Brust et al., 2015). This is a tightly controlled process, and adequate control is crucial for the proper development of the organism. 5' UTR-regulated translation plays an important role in spatio-temporal regulation of protein expression.

One such example, relevant to mouse, is offered by the retinoic acid receptor $\beta 2$ (*Rar $\beta 2$*). This 461 bp long 5' UTR containing five short uORFs is known for exerting specific regulatory programs that control tissue specificity (Zimmer et al., 1994; Reynolds et al., 1996; Sonawane et al., 2017). Generally, the presence of uORFs positively or negatively regulates translation of the main ORF. In the case of *Rar $\beta 2$* , uORFs affect translation both positively and negatively. For example, a mutation in uAUG2 results in a reduced reporter expression, hence it appears this uORF has rather stimulating effects on translation (Velden and Thomas, 1999). Following termination at the uAUG2 stop codon, ribosomes may reinitiate one nucleotide upstream at uAUG4 (Velden and Thomas, 1999). uORF4 is known for its important function in tissue specificity. Unlike the uORF2, the uORF4 inhibits the

expression of *Rarβ2* in heart and brain in regular conditions; however, if a mutation occurs in the uORF4, this leads to induced expression in both tissues (Velden and Thomas, 1999).

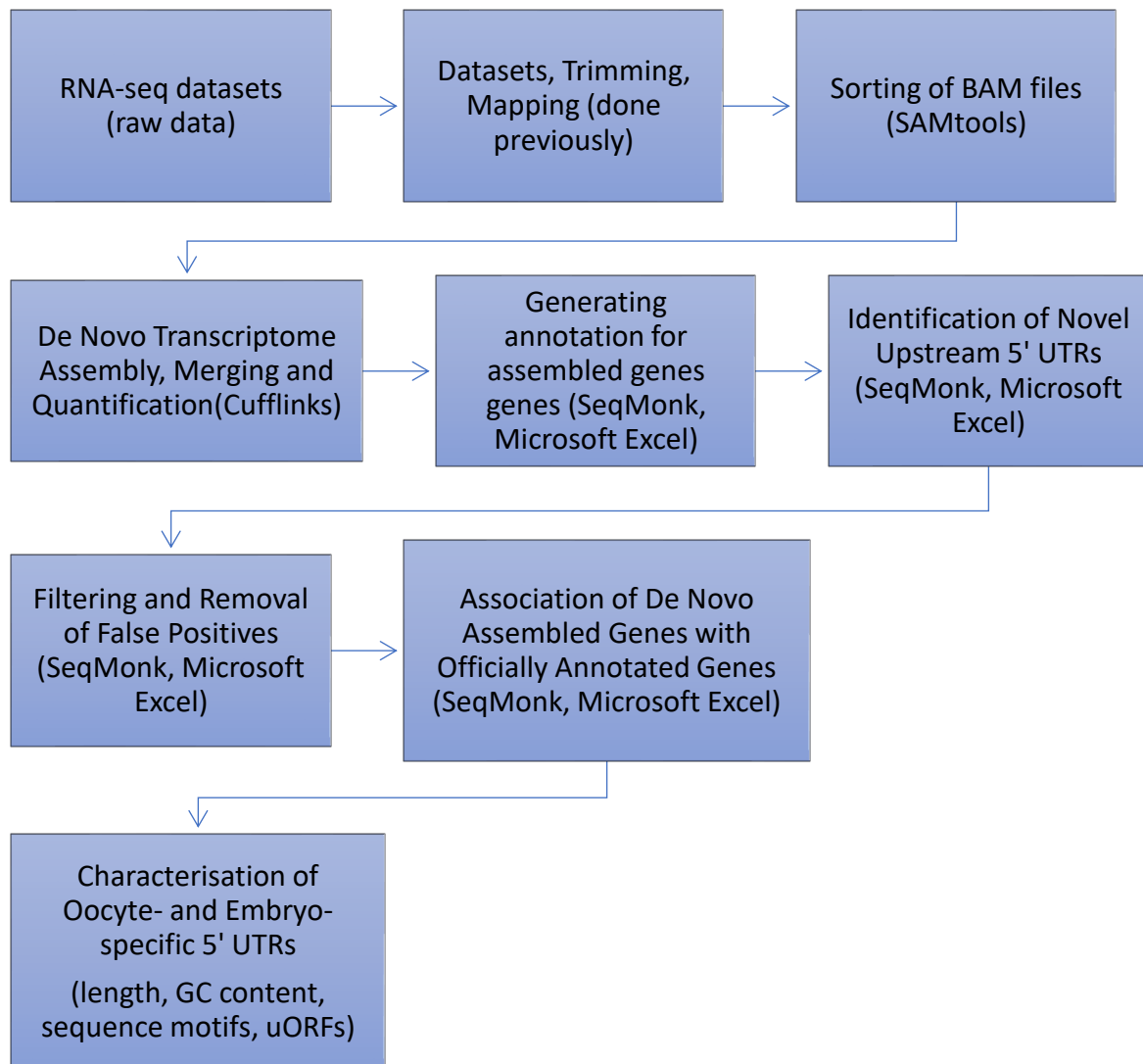
Further proof of the impact that the 5' UTR uAUGs have on the translation efficiency can be provided by inspecting the 5' UTR of proto-oncogene *C-mos*. *C-mos* is well-known for its regulatory function during spermatid development. Its transcription produces testicular and ovarian mRNAs with 5' UTRs that differ both in length and in their structure. While the ovarian mRNA, consisting of 80 nucleotides, only somewhat inhibits translation, the testicular mRNA of 3000 nucleotides and four uAUGs inhibits the translation strongly and effectively (Steel et al., 1996). Hence, just like in the case of the *Rarβ2*, the uAUGs of *C-mos* also exert translational control.

Even though the transcriptomes of oocytes and preimplantation embryos differ from somatic mRNAs (Veselovska et al., 2015; Zhang et al., 2014a) including different lengths of 5' UTRs (Veselovska et al., 2015), the features and potential biological effect of these UTRs in early development were not studied yet. Therefore, in this thesis, we aimed to identify which genes have longer 5' UTRs in oocytes or preimplantation embryos and bioinformatically characterize these UTR extensions to shed more light on their potential biological relevance.

4 AIMS

- To perform a de novo transcriptome assembly
- To identify and annotate oocyte- and embryo-specific 5' UTRs starting upstream of the canonical 5' UTRs from RNA-seq data in mouse oocytes and preimplantation embryos
- To characterize the length and GC content of oocyte- and embryo-specific 5' UTRs
- To investigate the presence of uORFs in oocyte- and embryo-specific 5' UTRs
- To identify enriched sequence motifs in oocyte- and embryo-specific 5' UTRs and to investigate whether they match recognition motifs of RNA binding proteins binding sites and miRNAs

5 WORKFLOW OVERVIEW



6 MATERIALS AND METHODS

6.1 Datasets, Trimming and Mapping

We used RNA-seq datasets representative of the model organism mouse (*Mus musculus*). The datasets were obtained from the NCBI Gene Expression Omnibus database with accession codes: GSE70116 (Veselovska et al., 2015), GSE98150 (Wang et al., 2018), GSE75957 (Andergassen et al., 2017). Respective files were provided in the fastq format from the European Nucleotide Archive. A summary of datasets used can be seen in Table 1.

In order to produce high quality reads free of low-quality bases and adapters, the Trim Galore program (www.bioinformatics.babraham.ac.uk/projects/trim/galore/) v0.4.1 was used with default parameters. We used the command “trim_galore --paired *fastq.gz”.

The trimmed reads were mapped to the indexed mouse genome of GRCm38 via the Hisat2 (Kim et al., 2015; Pertea et al., 2016) v2.0.5 with parameters specifying the maximum and minimum penalties for soft-clipping per base (--sp) and modifying the output to be compatible with the de novo transcriptome assembly using Cufflinks (--dta-cufflinks).

The output file from Hisat2 with mapped reads (Sequence Alignment Map (sam) file) was converted to Binary Alignment Map (bam) file using SAMtools view function of SAMtools v1.3.1 (Li, 2011; Li et al., 2009). Datasets download, trimming, and mapping was performed previously in the laboratory.

The mapped reads provided in the form of a bam file were sorted by leftmost coordinates. SAMtools sort function was used from SAMtools v1.3.1. The command used for sorting was: “samtools sort -o output_sorted.bam input.bam”.

6.2 De Novo Transcriptome Assembly

To perform de novo transcriptome assembly from the selected datasets (Table 1), the Cufflinks program (Roberts et al. 2011a; Roberts et al. 2011b; Trapnell et al., 2010; Trapnell et al., 2013) v2.2.1 was used in the reference annotation-based transcript (RABT) mode (specified by the parameter -g). The transcriptome annotation that was used as a baseline for the assembly was Mus_musculus.GRCm38.94.chr.gtf downloaded from the Ensembl genome browser. The RABT transcriptome assembly was performed using default parameters (the command used was “cufflinks -g Mus_musculus.GRCm38.94.chr.gtf -u --library-type xxx -o

output_folder reads_sorted.bam”). The specification of library types (option --library-type) used for the definition of the strand specificity was as follows: fr-secondstrand for d10 and d15 oocyte datasets from Veselovska et al. (2015); fr-firststrand for d5 and GV oocyte datasets from Veselovska et al. (2015) and all datasets from Andergassen et al. (2017); fr-unstranded for all datasets from Wang et al. (2018).

The next step was the merging of all annotations from individual oocyte growth stages, and individual replicates for each embryonic stage and for each somatic tissue into conclusive final annotations. For this purpose, the Cuffmerge function from Cufflinks v2.2.1 was used. Merging was also performed collectively to obtain single conclusive annotation for all embryonic stages together and all somatic tissues together. The command used was “cuffmerge xxx.txt”, where the respective txt file contained a list of gtf files to be merged.

6.3 Generating Annotations of All Genes

First, we quantified annotations in merged gtf files with one sorted bam file using the Cufflinks function from Cufflinks v2.2.1 (the command “cufflinks -G xxx_merged.gtf -u --library-type -o output_folder xxx_sorted.bam”). Each of the newly generated output_folders contained among other files also files genes.fpkm_tracking and isoforms.fpkm_tracking, containing information about genomic coordinates (chromosome, start and end positions) and expression levels of genes and their transcript isoforms, respectively (each gene consists of one or more transcript isoforms). From the genes.fpkm_tracking files, we extracted the information about the name (gene_id) and genomic coordinates of each gene for all merged gtf files and saved it in a .txt format with a tab delimiter preference.

In order to obtain strand information for genes (whether they are encoded on + or - DNA strand), a series of steps was performed using SeqMonk v1.45.4 and Microsoft Excel 2016 (MSO: 16.0.12624.20422). In Seqmonk, we imported individual merged gtf files as annotations and random gtf or bam file as reads (we needed something imported as reads to be able to perform analysis within Seqmonk even if the imported read datasets were not used for the quantification). For each imported annotation, probes were defined as mRNAs of respective annotation using feature probe generator. We unchecked the option for removing exact duplicates and created probes with the rest of the options as default (probes over feature from -0 to +0 bp). We selected fixed value quantitation, as we did not need the actual

quantitation values. The output .txt file was generated as an annotated probe report including information about transcript (mRNA) name and strand. This data was combined with the data from the isoforms.fpk_tracking file described in the previous paragraph containing information about transcript name and respective gene name. Using Microsoft Excel, we alphabetically ordered the data (SORT function) and matched strand information from the Seqmonk output file with the gene name from isoforms.fpk_tracking file through common transcript names (MATCH and LOOKUP functions). Using this approach, we generated complete annotation of genes for each merged gtf file (gene name, chromosome, start, end, strand). These gene annotations were saved as .txt files.

6.4 Identification of Annotated Genes with Novel Upstream 5' UTRs

To identify novel 5' UTRs with transcriptional start site (TSS) upstream of the canonical TSS in oocyte and embryo datasets, we first generated an annotation of all annotated genes except genes for short non-coding RNAs. This was performed in Seqmonk v1.45.4 by generating probes using the RNA-Seq quantitation pipeline option with the selected option “merge transcript isoforms”. These probes were imported as annotation named “annotated_genes”. These genes served as a baseline for finding novel 5' UTRs with upstream TSS.

We imported annotated_genes and gene annotations from our merged assemblies as reads. To find out which annotated_genes appear to start from an upstream TSS, we defined probes upstream of the annotated genes from -1000 to -100 bp, and performed the read count quantitation with modified default settings as follows: Count reads on the same strand as the probe; uncheck “Correct for total read count”; uncheck “Log transform count”. The quantification of annotated_genes as reads gave us the information if the upstream region overlapped another annotated gene on the same strand. The quantification of our assembled annotations imported as reads counted whether the upstream region overlaps a gene in our assembly on the same strand, potentially the same gene as from which the upstream region was determined, but starting from an upstream TSS.

To remove all upstream regions overlapping annotated_genes on the same strand, we performed the filtering of values within Seqmonk. We retained only upstream regions with value 0 (filter was for a value between 0 and 0) for a dataset consisting of annotated_genes

imported as reads. This was performed to remove false positives as such upstream regions would appear as potentially being a novel UTR starting from upstream TSS in the following analysis.

The read count quantifications with our assembled annotations as reads and with the upstream regions after filtering step as annotation were exported from Seqmonk. For each assembly, if the count was >0 it suggested that in that particular dataset there was the same strand gene overlapping the upstream region of the annotated gene. To obtain the list of genes with their upstream region overlapped by same strand gene in at least one of our assemblies, we applied the Microsoft Excel SUM function per line (each line contained quantitation values from all assemblies for the upstream region of one gene), followed by sorting of the summed values from smallest to largest using Microsoft Excel SORT function, and selecting upstream regions with a summed value greater than 0. Names of genes with these upstream regions were copied and pasted into the Seqmonk search option (“find named features”) in annotated_genes and saved as an annotation track named “annotated_genes_with_upstream”.

6.5 Removal of False Positives

Next, we wanted to remove annotated genes with previously identified upstream 5' UTR based on the read count described in chapter 6.4. in which the reads in our assembled annotations corresponded to the novel 3' elongation of an upstream same strand annotated gene or an independent novel same strand gene, instead of novel upstream 5' UTR of the same gene. To achieve this, we reimported the data containing our assembled genes (option import data from visible data stores) but with a filter to import only reads overlapping the annotation of annotated_genes_with_upstream. Next, probes were defined upstream of annotated_genes_with_upstream from -1000 to -100 bp without the option of the removal of exact duplicates. Read count quantitation was performed with reimported reads, counting reads on the same strand as the probe. Then, we performed filtering on values to retain upstream regions with value 1 or greater in at least one of the data stores, as they should be upstream regions of annotated genes with novel upstream 5' UTR. Genes with such upstream regions were named as filtered_annotated_genes_with_upstream.

6.6 Association of De Novo Assembled Genes with Officially Annotated Genes

In the next step, we aimed to associate the de novo assembled genes with the officially annotated genes for cases where our analysis indicated that elongation in the 5' UTR could have occurred. To achieve this, we wanted to identify which de novo assembled gene overlaps each annotated gene with potentially prolonged 5' UTR on the same strand, and which overlaps -1000 bp to -100 bp upstream regions of these genes.

De novo assembled genes for each dataset were split into two files (xxx_plus.txt and xxx_minus.txt) according to the strand (plus or minus) of the assembled genes. We then imported each of the xxx_minus.txt and xxx_plus.txt files as annotations into Seqmonk. Probes were defined as filtered_annotated_genes_with_upstream described in the previous chapter without the removal of exact duplicates. Following fixed value quantitation, we generated an annotated probe report, where we annotated probes first with overlapping xxx_minus.txt and xxx_plus.txt files. This generated overlapping_minus_xxx.txt and overlapping_plus_xxx.txt files, respectively.

Next, we repeated the process but with probes being generated upstream of annotated_genes_with_upstream_filtered from -1000 to -100 bp. After annotation of probes with overlapping xxx_minus.txt and xxx_plus.txt, we obtained result files upstream_minus_xxx.txt and upstream_plus_xxx.txt, respectively. Each of the resulting files was exported and modified in Excel to only contain filtered_annotated_genes_with_upstream or their upstream regions of the relevant strand, i.e. minus for files with overlap with the de novo assembled genes on the minus strand, and plus for files with overlap with the de novo assembled genes on the plus strand.

We first combined the information from Overlapping (overlapping_xxx.txt) and Upstream files (upstream_xxx.txt) for respective strands. The first 6 columns from each file were retained and saved in a separate file. These columns included information regarding Probes (the gene), Chromosome, Start, End, Strand, and Feature. We renamed the column Feature to “Body” for overlapping files as it corresponded to the overlap with the gene body of the annotated gene, and to “Upstream” for upstream files, as the overlap corresponded to the region upstream of the annotated gene.

Subsequently, we sorted both sets of data, and following the deletion of the “_upstream” from the upstream Probes, we combined the data into one dataset, removing redundant data in the process. As a result, we obtained a table that consisted of columns titled

Probe (the name of the annotated gene from filtered_annotated_genes_with_upstream), Chromosome, Start, End, Strand, Body (showing the overlap of gene body of the annotated gene with same strand de novo assembled gene), Upstream (showing the overlap of the upstream region of the annotated gene with same strand de novo assembled gene).

We utilized the Microsoft Excel function IF in order to uncover where there was a match in the name of the de novo assembled gene in Body and Upstream columns. We filtered the data to only have the cases where there was a match found for Body and Upstream names of the de novo assembled gene. This gave us a list of annotated genes that are prolonged on 5' ends (by more than 100 bp) in our assemblies with names of associated de novo assembled genes.

6.7 Identification of Oocyte- and 2C Embryo-specific 5'UTRs

We then focused on oocyte and 2C embryo prolonged 5'UTRs. We took the genes that had a 5' prolongation either in the oocyte or 2C dataset, or both, and we searched whether they have an upstream TSSs in the somatic datasets. We created 2 tables - one for plus genes and one for minus genes. Each contained information about the officially annotated gene name, coordinates, whether there was an upstream TSS or not in the oocytes, 2C embryos, and any of the somatic datasets. We removed all genes with upstream TSS in at least one somatic dataset to only have oocyte-, 2C- or oocyte, and 2C-specific upstream TSSs.

From files with the annotations of the de novo assembled genes for each dataset, we extracted the start coordinates of these prolonged genes in the oocytes and 2C embryos, creating a table with the name of the annotated gene, its coordinates, names of de novo assembled gene associated with this annotated gene in oocytes and 2C if there were any, and start coordinates of these de novo assembled genes. Using this information, we quantified the length of the prolongations in oocytes and 2C compared to the official annotation. For cases when the gene was prolonged in both oocyte and 2C embryos, we also quantified the difference in these prolongations. The mean and median values of the differences in the prolongations were quantified in Microsoft Excel 2016. For further analysis, we selected only genes with prolongation up to 500 bp, up to 1000 bp, and up to 3000 bp, as these most likely correspond to unspliced prolonged 5' UTRs.

6.8 Length and GC Content Analysis.

Mean and median values of 5' UTR prolongations up to 500 bp, 1000 bp, and 3000 bp in the oocytes or 2C embryos were quantified in Microsoft Excel 2016. Histograms of 5' UTR prolongations (see Appendix 3 and 5) and a horizontal notched boxplot were generated in R language (see Appendix 4).

Then, using the table we prepared in the previous chapter, we generated annotations of the 5' UTR prolongations in the oocytes and 2C embryos up to 500 bp, 1000 bp, and 3000 bp, i.e. the regions between the upstream oocyte or 2C TSS and TSS in the official annotation. In addition, we downloaded annotated 5' UTR annotation from the UCSC genome browser to serve as a control.

These seven annotations were used as an input for a python script that computes the GC content (generated previously in the laboratory, see Appendix 1). The input for the script was provided in a .txt format with annotation of genomic coordinates and fasta files with mouse genomic sequence split into individual chromosomes. Due to the size of the control file, we processed only 5' UTRs of mRNAs encoded on chromosome 1. Violin plot depicting GC content of respective annotations was generated in R language (see Appendix 6).

6.9 uORFs

In order to analyze the uORF abundance of oocyte and 2C embryo 5' UTRs, we first generated sequences of the 5' UTR prolongations in the oocytes and 2C embryos up to 500 bp, 1000 bp and 3000 bp using a python script (previously generated in the laboratory, see Appendix 2). The input for the script was provided as a raw mouse genome sequence split into individual chromosomes, an annotation of genome coordinates of regions of interest, and the list of names of regions from the annotation for which we would like to obtain the sequence. The output of the script comprises sequences of regions of interest in fasta format.

We used the output sequences to search for ORFs – these would correspond to uORFs. For this purpose, we utilized GenScript's ORF Finder tool (Stothard, 2000) with the settings as follows: ORFs can begin with ATG; Search for ORFs in reading frame 1,2, and 3 on the direct strand; Only return ORFs that are at least 5 codons long; Use the standard (1) genetic code. As an input, we provided a random selection of 46 sequences from the official 5' UTR annotation to serve as a control, 39 sequences from oocytes with upstream 5' UTR

shorter or equal to 3000 bp, and 39 sequences from 2C embryos with upstream 5' UTR shorter or equal to 3000 bp. For oocyte and 2C embryo sequences, we picked 13 random sequences up to 200 bp in length, 13 random sequences from 200 to 1000 bp in length, and 13 sequences from 1000 to 3000 bp in length. A notched boxplot with jitter plot and outliers was generated, plotting ORFs per 100 bp in 3 categories (see Appendix 7). Furthermore, we generated a scatterplot for mean and median per 100 bp in 12 categories (for oocyte, 2C embryo and official annotation sequences up to 200 bp in length, from 200 to 500 bp in length, from 500 to 1000 bp in length, and from 1000 to 3000 bp in length) (see Appendix 8).

6.10 Sequence Motifs

To identify enriched sequence motifs in the oocyte- and 2C embryo-specific 5' UTRs, we used the MEME tool v 5.1.0 (Bailey and Gribskov, 1998). We aimed to discover whether the oocyte- and 2C-specific 5' UTRs contain any significantly enriched recognition sites for RNA binding proteins or miRNAs compared to the annotated 5' UTRs.

We performed the analysis on 6 datasets categorized according to the length and type (oocyte-specific 5' UTRs shorter or equal to 500 bp, oocyte-specific 5' UTRs shorter or equal to 1000 bp, oocyte-specific 5' UTRs shorter or equal to 3000 bp, 2C-specific 5' UTRs shorter or equal to 500 bp, 2C-specific 5' UTRs shorter or equal to 1000 bp, 2C-specific 5' UTRs shorter or equal to 3000 bp). A fasta file with sequences of each of the selected datasets generated in the previous chapter was used as a primary sequence input, and the annotated 5' UTR sequences were used as control sequences input.

We modified the settings as follows: Differential Enrichment mode as the motif discovery mode; Any Number of Repetitions (anr) as the site distribution; 5 as the number of motifs; check search given strand only.

The result motifs that showed significant p-value (below 0.1) then served as input for the Tomtom tool v5.1.1 (Gupta et al., 2007) to find if our motifs were significantly similar to known motifs acting as binding sites of RNA binding proteins or as recognition motifs of miRNAs. We modified the default Tomtom settings by checking the „Do not score the reverse complements of target motifs“. To search for motifs within RNA binding protein binding sites, we specified the motif database to be RNA (DNA-encoded) and Ray2013 Mus Musculus (DNA-encoded), and for motifs within miRNA recognition sites, we specified the

motif database to be miRBase Single Species microRNA (DNA-encoded) and *Mus_musculus_mmu* (DNA-encoded).

7 RESULTS

7.1 De Novo Transcriptome Assembly

In order to produce full annotation of all transcripts, including novel isoforms of known genes, and to compare differences in the 5' UTRs across development in mouse, 51 publicly available RNA-seq datasets were used, corresponding to oocytes (Veselovska et al. 2015), six developmental stages of cell lineages in preimplantation embryonic development (Wang et al. 2018) and seven adult somatic tissues (Andergassen et al. 2017) (listed in Table 1). The datasets of oocytes consisted of growing and fully grown oocytes, specifically growing oocytes of postnatal day 5 (d5), 10 (d10), and 15 (d15), and fully grown germinal vesicle (GV) oocytes. Stages of preimplantation development were represented by two-cell (2C) embryos, four-cell (4C) embryos, eight-cell (8C) embryos, morula, the inner cell mass (ICM), and trophectoderm (TE). The adult somatic tissues were represented by the brain, liver, lung, leg muscle, heart, spleen, and thymus tissues. These specific datasets were selected because of the depth of their sequencing, and for oocyte and somatic datasets also strand specificity (we did not find strand-specific preimplantation embryo datasets) as we aimed to obtain the datasets of the best possible quality.

Once the datasets were downloaded, appropriate trimming was applied, quality check took place, followed by mapping to the GRCm38 mouse genome. This was performed previously in the laboratory. Next, mapped reads were sorted using SAMtools.

Subsequently, we performed the de novo transcriptome assembly using the Cufflinks program on the respective datasets. The resulting annotations were merged via the Cuffmerge function from Cufflinks into a complete transcriptome annotation for each developmental stage or tissue, and together for all preimplantation embryo datasets and all somatic tissues.

Merged gtf files (listed in Table 2) contain annotations of mRNA transcripts. However, for our purposes, we needed the annotation of genes to compare them to the official gene annotation. We, therefore, quantified the annotations with one sorted bam file

using Cufflinks, which gave us the genomic coordinates of genes in the annotation (chromosome, start, end).

To discover whether the genes were encoded on a + or - DNA strand, we utilized Seqmonk – we exported mRNA information from our merged assemblies, which contained strand information for each mRNA. Cufflinks quantification output contained also information about which mRNA corresponds to which gene, therefore, using mRNA names we could match strand information from Seqmonk with gene names. This allowed us to generate a complete annotation of genes for each merged gtf file.

Table 1: Summary of the datasets used for the analysis.

Publication	Cell type	Number of replicates	Accession code
Veselovska et al. 2015	d5 oocytes	1	GSE70116
	d10 oocytes	1	
	d15 oocytes	1	
	GV oocytes	1	
Wang et al. 2018	2C embryo	4	GSE98150
	4C embryo	4	
	8C embryo	3	
	morula embryo	2	
	E3.5 - ICM	4	
	E3.5 - TE	4	
Andergassen et al. 2017	adult_brain	4	GSE75957
	adult_liver	4	
	adult_heart	4	
	adult_lung	4	
	adult_leg_muscle	4	
	adult_spleen	4	
	adult_thymus	2	

Table 2: mRNA count for each of the respective datasets.

Dataset	File name	mRNA count
Wang et al., 2018	2C_merged.gtf	111245
	4C_merged.gtf	108101
	8C_merged.gtf	89623
	morula_merged.gtf	84518
	ICM_merged.gtf	118718
	TE_merged.gtf	90610
	embryo_merged.gtf*	167987
Veselovska et al., 2015	oocytes_merged.gtf*	69203
Andergassen et al., 2017	brain_merged.gtf	85339
	heart_merged.gtf	54882
	leg_merged.gtf	60366
	liver_merged.gtf	59175
	lung_merged.gtf	73351
	spleen_merged.gtf	89283
	thymus_merged.gtf	68666
	adult_merged.gtf*	82624

(*) final merged gtf file

7.2 Identification of genes with Novel Upstream 5' UTRs

We aimed to identify novel upstream 5' UTRs in oocytes, preimplantation embryos, and as a control in somatic tissues. In order to identify novel upstream 5' UTRs, we began by generating an annotation of all known annotated genes with the exception of genes for short non-coding RNAs (“annotated_genes”). The number of these genes was 35579. Since one of our aims was to discover the annotated_genes that have a TSS upstream of the canonical TSS (which we would then further filter to have only genes with prolonged 5' UTRs), we defined probes in Seqmonk upstream of the annotated genes from -1000 to -100 bp, followed by a read count quantitation using annotated genes, as well as our assembled genes as reads. The quantification was performed, because we were interested whether the upstream region is overlapped by the same strand assembled gene, suggesting it might be the same gene but with an upstream TSS and, therefore, potentially longer 5' UTR. The quantification of reads being annotated genes was performed to exclude genes with upstream regions overlapped by the same strand annotated genes located upstream of the analyzed gene (Figure 1).

After quantification, we first filtered genes to have a value of 0 in the quantification of annotated genes as reads, in order to remove false positives i.e. all upstream regions overlapping annotated_genes on the same strand. There were 31257 genes after the filtering step. Had these not been removed, these regions would have appeared as having an upstream novel 5' UTR starting from an upstream TSS in the following analysis, even though that would not be the case (Figure 2).

We then filtered the genes with upstream regions with read count of each assembly imported as reads that exhibited a value >0 . This indicated that the upstream region of the annotated gene might be overlapped by an elongated novel 5' UTR of the same gene. This was done for the oocyte dataset, together for all preimplantation embryo datasets (if at least one of the datasets had value 1 or higher), and together for all somatic tissue datasets (if at least one of the datasets had value 1 or higher). However, there was still a chance that there were false-positive findings, specifically in situations when the upstream region of the annotated gene was overlapped by 3' non-annotated extension of same strand annotated gene (Figure 3) or another same strand novel gene separate from the gene in our assembly corresponding to the annotated gene (Figure 4).

To remove such genes, we created an annotation containing all genes retained after the previous two filtering steps. The names of the genes remaining after the previous two filtering steps were copied and pasted into the Seqmonk search option in annotated_genes

and saved as an annotation track “annotated_genes_with_upstream”. We then re-imported reads (being our assembled genes) selecting only reads overlapping annotated_genes_with_upstream on the same strand. Using this approach, we should remove reads corresponding to the assembled genes not directly overlapping annotated genes with potentially prolonged 5’ UTR generated after the first two filtering steps. Then, we defined probes again -1000 to -100 bp upstream of the “annotated_genes_with_upstream” and performed read count quantitation counting re-imported filtered reads in the same strand as the probe. We continued by performing filtering on values. This was done in order to retain upstream regions with value 1 and more as they should correspond to the upstream regions of annotated genes with novel upstream 5’ UTR. Regions that exhibited values of 0 were discarded from the analysis, as they likely referred to false-positive findings described above (Figures 3 and 4).

The names of adequate genes were copied and pasted into the Seqmonk search option in annotated_genes_with_upstream, and saved as an annotation track “filtered_annotated_genes_with_upstream”. More information regarding the numbers of genes after each filtering is in Table 3.

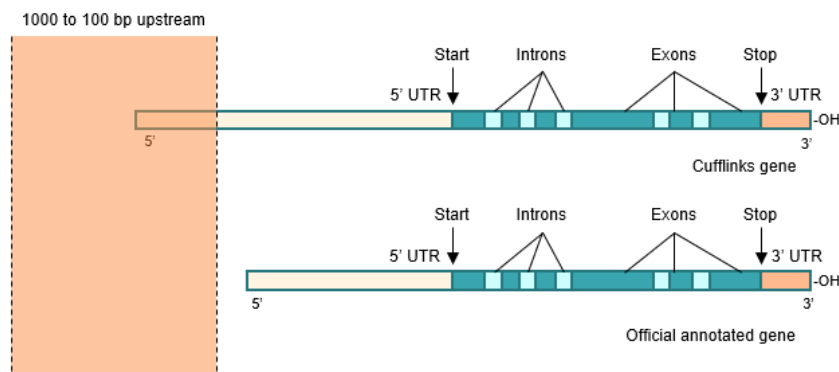


Figure 1: Read count quantitation would reveal count of 1 for a novel upstream 5’ UTR.

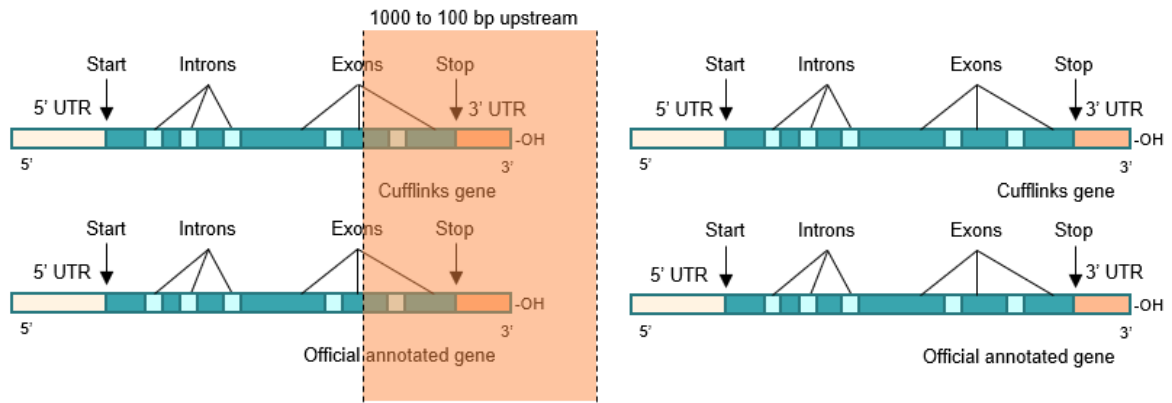


Figure 2: Read count quantitation would reveal count of 1 for a different gene on the same strand if it was found in the region -1000 bp to -100 bp upstream of studied genes.

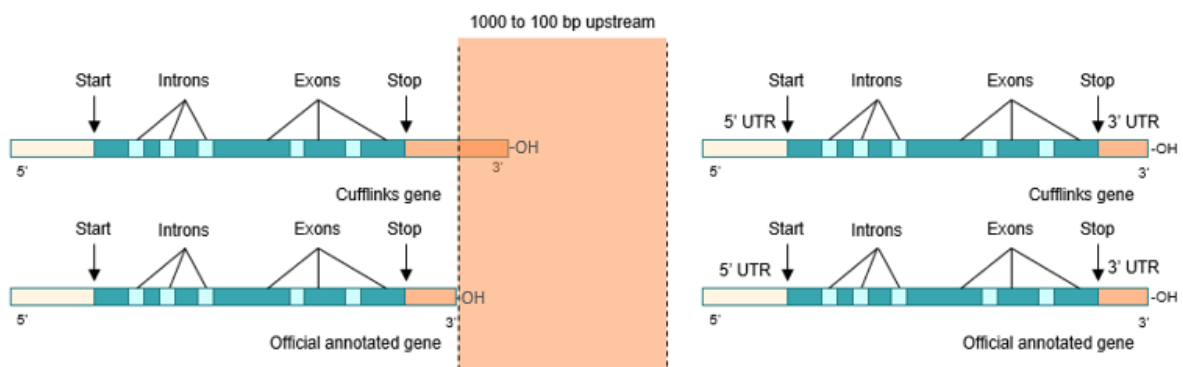


Figure 3: Read count quantitation would yield count of 1 for a novel 3' UTR.

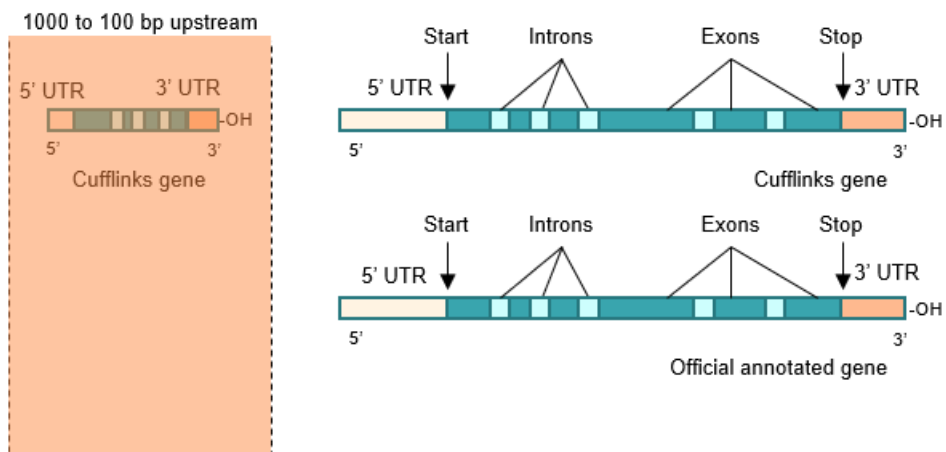


Figure 4: Read count quantitation would yield count of 1 if there was a separate gene found in the upstream region.

Table 3: The number of identified genes with an upstream region in every step of filtration. Pre-filtering refers to the number of identified regions prior to filtering. First filtering refers to regions retained following the removal of false positives, where read count value was greater than 0. Second filtering refers to “*annotated_genes_with_upstream*”. Third filtering refers to “*filtered_annotated_genes_with_upstream*”. These likely correspond to the elongated novel 5’ UTRs.

	Oocyte	Embryo	Somatic
Pre-filtering	35579	35579	35579
First filtering	31257	31257	31257
Second filtering	3551	6435	4470
Third filtering	3080	6126	4101

7.3 Association of De Novo Assembled Genes with the Officially Annotated Genes

In the next step of our analysis, we wanted to associate *filtered_annotated_genes_with_upstream* defined in the previous chapter with the overlapping de novo assembled genes. This was done in order to find out which de novo assembled genes correspond to the prolonged annotated genes at their 5’ UTR.

In order to do this, we split the annotations for de novo assembled genes depending on the DNA strand on which they are encoded (plus or minus genes). In Seqmonk, we first made probes over all *filtered_annotated_genes_with_upstream* and identified overlaps with plus and minus genes from the de novo assembled genes from all datasets. Then, we repeated it but with probes over upstream regions -1000 bp to -100 bp of the *filtered_annotated_genes_with_upstream*. From the resulting files, we removed annotated genes on the other strand, i.e. plus-strand annotated genes in files with overlaps with the minus de novo assembled genes and vice versa.

We needed to divide the genes into plus and minus groups because otherwise, we would not be able to associate genes properly in the next steps of the analysis. It is because during the generation of annotated probe report with the overlapping de novo assembled genes, we cannot specify that the de novo assembled gene must have the same strand as the annotated gene. In case that there is an overlap of annotated genes with one de novo

assembled gene on the same strand and one on the opposite strand, we could obtain the wrong gene name as the name of the overlapping de novo assembled gene.

We then looked for genes with matching names of the associated de novo assembled genes for Body and Upstream. This group represented those annotated genes with properly associated de novo assembled gene which is the same gene as the annotated gene but with upstream TSS. If the names were not matching, it corresponded to the more complicated situations with overlapping or closely located same strand genes where the precise annotation of prolonged 5' UTR would be more complicated. In some cases, there was a value "null" instead of the de novo assembled gene name. It was in situations where there was no overlap of the annotated gene itself or its upstream region with the same strand de novo assembled gene in that particular dataset. It occurred in cases when multiple genes had the same name and only one had prolonged 5' UTR, while the remaining had instead „null“ value. This was the case in the oocyte de novo assembled genes. For embryonic and somatic genes, we worked with all annotated genes that were prolonged in at least one of the datasets. Therefore, in cases when the gene did not have a prolonged 5' UTR in that particular dataset or was not assembled at all, there was a „null“ value. We quantified the number of genes in each category (Table 4). The highest number of annotated genes with prolonged 5' UTR was found in oocytes and early embryos (2C, 4C). From the somatic datasets, the highest number of prolongations was found in the brain, followed by lung and spleen (Table 4).

Table 4: Association of de novo gene with the official gene. Number of easily associated genes corresponding to a match in de novo assembled names and no “null“ value; the number of genes corresponding to a mismatch in de novo assembled names and no “null“ value; the number of genes without upstream 5’ UTR with “null“ value in one or both names.

Dataset	Total no. of genes	No. of genes with matching de novo assembled names	No. of genes with mismatching de novo assembled names	No. of genes without upstream 5’ UTR
Oocyte	3080	3012	61	7
2C	6126	2951	123	3052
4C	6126	2384	174	3568
8C	6126	1322	101	4703
Morula	6126	1135	78	4913
ICM	6126	1391	131	4604
TE	6126	1275	99	4752
Embryo	6126	5041	229	856
Brain	4101	1500	131	2470
Heart	4101	467	56	3578
Leg	4101	823	72	3206
Liver	4101	856	88	3157
Lung	4101	1431	124	2546
Spleen	4101	1300	141	2660
Thymus	4101	588	77	3436
Full Somatic	4101	3299	160	642

7.4 Identification of Oocyte- and 2C Embryo-specific 5' UTRs

We then focused on the genes that had an upstream 5' UTR either in the oocyte or in 2C dataset, or both, and we tried to discover whether they had an upstream 5' UTR also in the somatic datasets. Those genes that were found in 2C embryo, oocyte and somatic upstream 5' UTR could not be considered oocyte- or 2C embryo-specific 5' UTRs.

The total number of annotated genes with an upstream 5' UTR in oocyte and/or 2C dataset was 4974. After removing genes with upstream 5' UTR also detected in somatic tissues, the number of genes was 3803. Although it is possible that an upstream 5' UTR identified in the somatic genes would be different from that found in oocytes and embryos, we decided to remove these cases to avoid false-positive findings. The number of annotated genes exclusively in oocyte upstream 5' UTRs (1513), or exclusively 2C embryonic upstream 5' UTRs (1622), was much higher than the number of genes with both upstream 5' UTRs (668). The results are summarized in Table 5.

Table 5: Comparison of oocyte and 2C embryonic 5' UTRs for plus and minus strands. Only those genes that were present in the upstream 5' UTRs of both oocytes and 2C embryos and had no upstream in somatic tissues were selected for further analysis.

	Plus strand		Minus strand	
	Including those with upstream 5' UTR in somatic tissues	Does not have an upstream 5' UTR in somatic tissues	Including those with upstream 5' UTR in somatic tissues	Does not have an upstream 5' UTR in somatic tissues
Has upstream 5' UTR in oocyte, does not have an upstream 5' UTR in 2C embryo	1065	791	958	722
Has upstream 5' UTR in 2C embryo, does not have an upstream 5' UTR in oocyte	1026	853	935	769
Has upstream 5' UTR in both 2C embryo and oocyte	508	341	482	327

7.5 Length of Oocyte- and 2C Embryo-specific 5' UTRs

Only genes that had an upstream 5' UTR in oocytes or 2C embryos and lacked an upstream in all somatic tissues were selected for further analysis, with the plus and minus strands pooled together. We were first interested in how much longer these 5' UTRs are compared to the annotated 5' UTRs.

The observed elongations in the upstream 5' UTR region in oocytes and 2C embryos proved to be very long in certain genes compared to the official 5' UTR. This occurs due to splicing (either alternative splicing within the coding sequence with upstream first exon, or splicing within the 5' UTR region), and, therefore, the length of elongation accounts also for introns. Since we aimed to focus exclusively on elongations without splicing events, we discarded all elongations greater than 3000 bp in length and split results into 3 categories for both oocytes and 2C embryos - less than or equal to 500 bp, less than or equal to 1000 bp, and less than or equal to 3000 bp (Figure 5).

Compared to the official annotations, the oocyte upstream 5' UTRs were longer by 227.6 bp on average for elongations smaller or equal to 500bp. Similarly, the 2C embryo elongations were longer by 252.8 bp on average for elongations smaller or equal to 500 bp. As the sample set increased to include elongations smaller or equal to 3000 bp, the differences between oocyte and 2C embryo upstream 5' UTR became more distinct. We observed the mean elongation for oocytes to be 683.8 bp, while it was 746.5 bp for 2C embryos. Generally, the elongations of 5' UTRs were rather short, with a median of 340.5 bp for oocytes and 475 bp for 2C embryos for UTR elongations up to 3000 bp (Figures 6 and 7).

We then focused on genes that had elongated 5' UTRs in both oocytes and 2C embryos. According to our analysis, even though fewer genes are elongated in 2C embryos than in oocytes, the elongations of 5' UTR are greater by 1 545 206 bp in 2C embryos as opposed to oocytes. It is important to note that in 269 of all 5' UTR elongations (oocyte and 2C embryo), this extension was likely caused by mRNA retention from the oocyte, as these elongations spanned less than 200 bp. The average elongation of 2C embryos was found to be 6544.05 bp longer than the average elongation in oocytes. A detailed overview of oocyte- and embryo-specific 5' UTRs is shown in Table 6 and Figure 8.

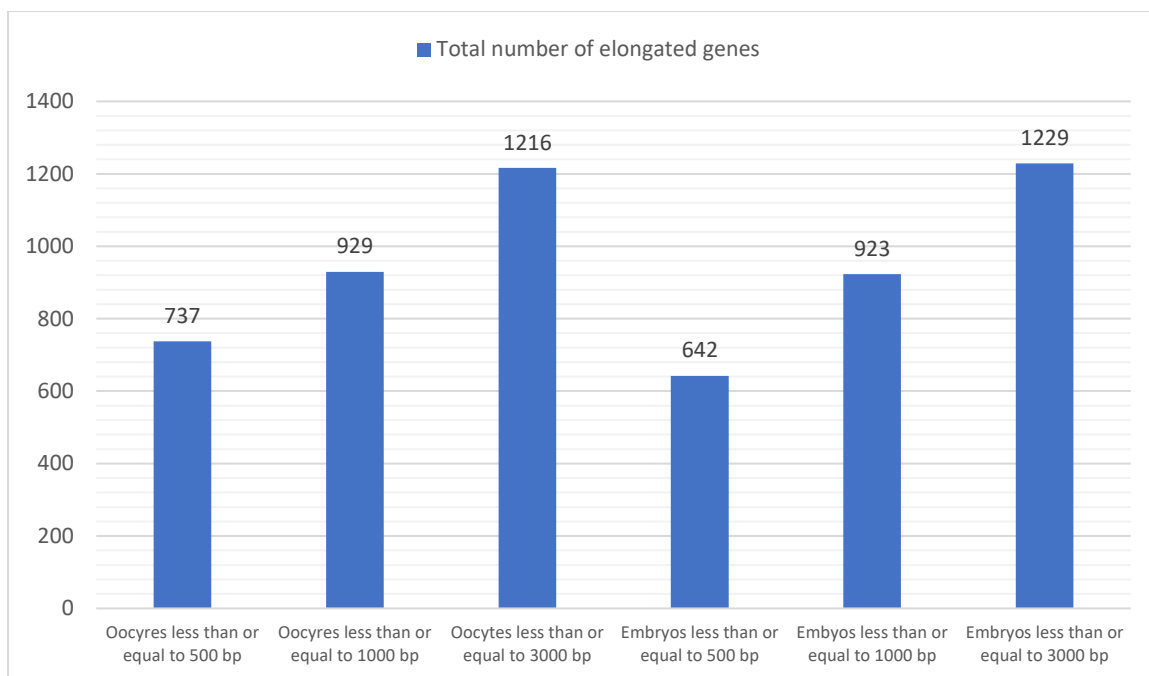


Figure 5: Total number of elongated genes observed in oocyte- and 2C embryo- 5' UTRs compared to the official upstream 5' UTRs.

Table 6: Comparison of 5' UTR elongations of oocytes and 2C embryos.

	Oocyte	2C Embryo
Total elongation	2 364 928 bp	3 910 134 bp
Number of elongated genes	362	299
Smallest elongation in size	1 bp	1 bp
Greatest elongation in size	181 713 bp	227 642 bp
Mean	6 532.95 bp	13 077 bp
Median	258.5 bp	605 bp
Number of genes where elongations < 200bp	170	99

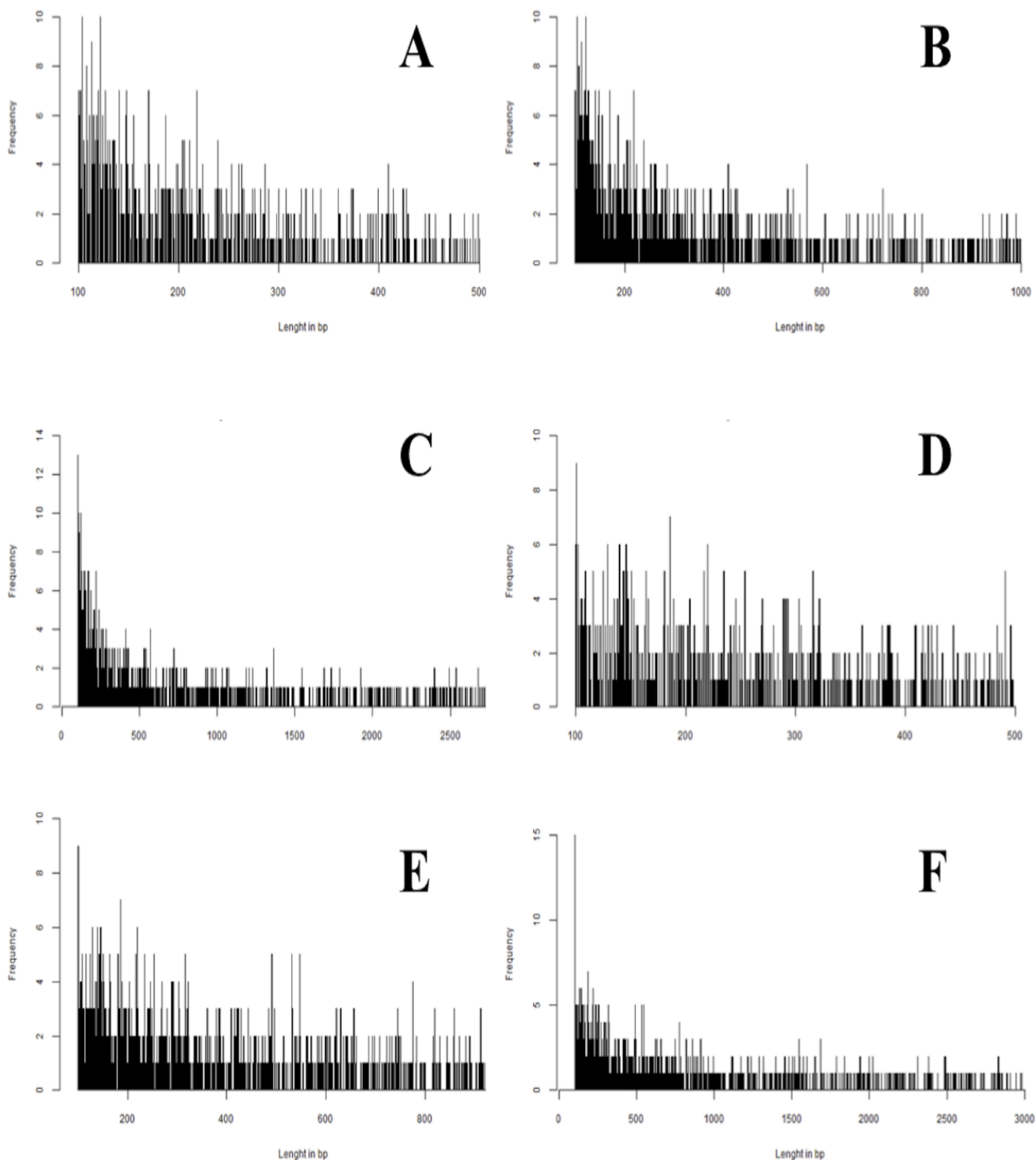


Figure 6: Histograms of oocytes and 2C embryos in categories according to length of 5' UTR elongation compared to the official annotated 5' UTR. Plotted are frequencies of respective lengths. A: Oocyte frequencies for elongations smaller or equal than 500 bp. B: Oocyte frequencies for elongations smaller or equal to 1000 bp. C: Oocyte frequencies for elongations smaller or equal to 3000 bp. D: 2C Embryo frequencies for elongations smaller or equal to 500 bp. E: 2C Embryo frequencies for elongations smaller or equal to 1000 bp. F: 2C Embryo frequencies for elongations smaller or equal to 3000 bp.

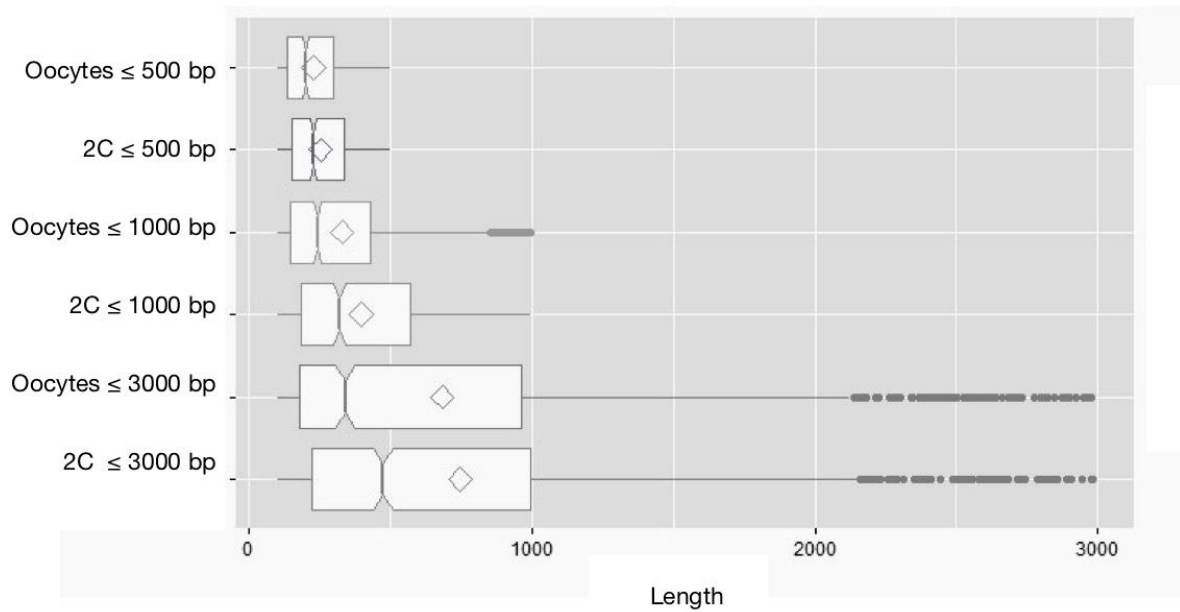


Figure 7: Comparison of 5' UTR elongations of oocytes and 2C embryos via boxplots. From top to bottom: Oocyte with 5' UTR elongation smaller or equal to 500 bp; 2C embryo with 5' UTR elongation smaller or equal to 500 bp; oocyte with 5' UTR elongation smaller or equal to 1000 bp; 2C embryo with 5' UTR elongation smaller or equal to 1000 bp; oocyte with 5' UTR elongation smaller or equal to 3000 bp; 2C embryo with 5' UTR elongation smaller or equal to 3000 bp. Each of the “boxes” in the boxplots shows the interquartile range. The vertical line splitting the “boxes” corresponds to median, and the notch depicts the confidence interval around the median. Mean is depicted as a diamond symbol.

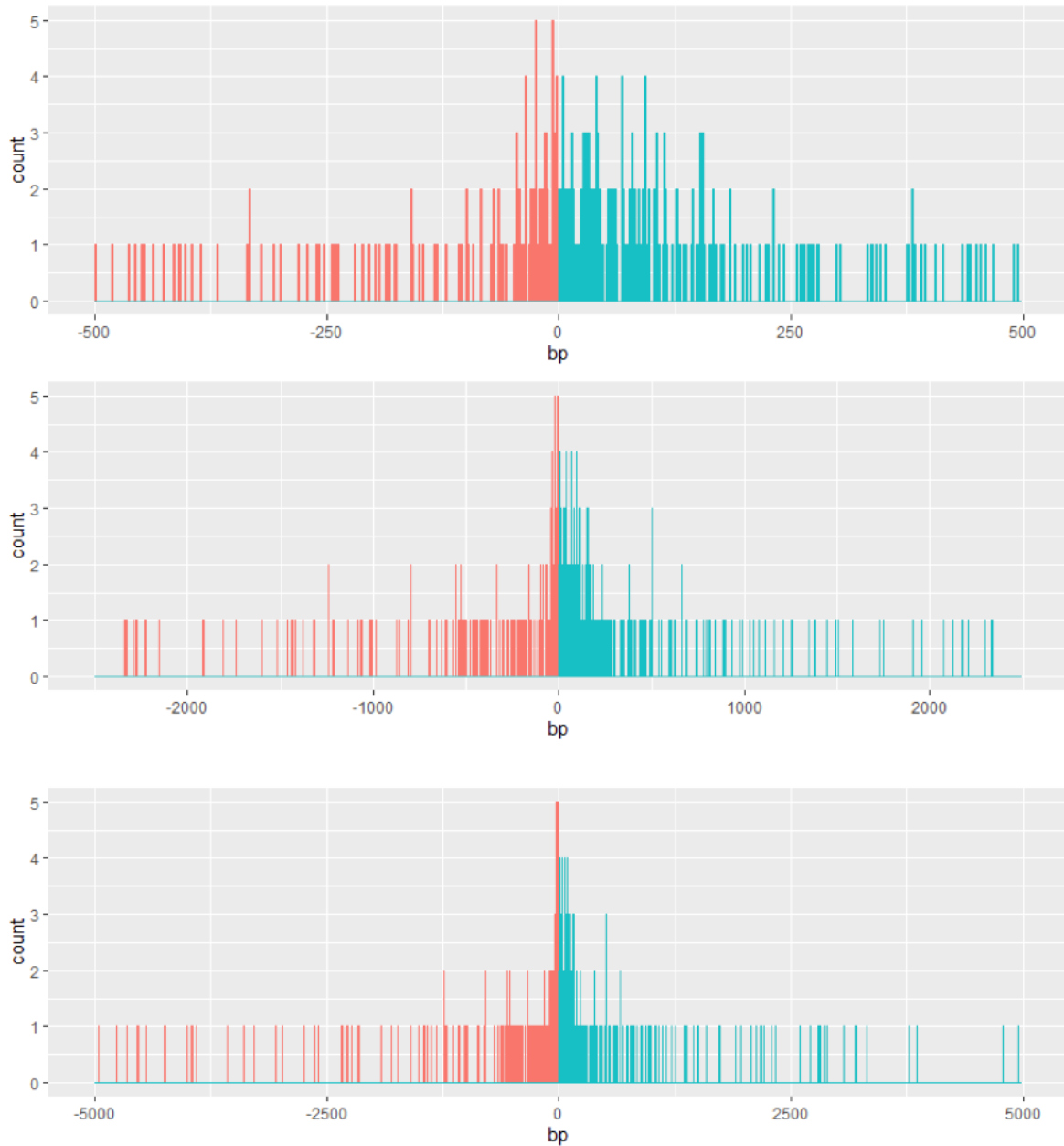


Figure 8: Comparison of 5' UTR elongations of oocytes and 2C embryos via a histogram. Negative values (red) correspond to 2C embryo elongations of 5' UTR. Positive values (blue) correspond to oocyte-specific elongations of 5' UTR. Plotted frequency (count) versus length (bp). From top to bottom: Elongations less than 500 bp in size. Elongations less than 2500 bp in size. Elongations less than 5000 bp in size.

7.6 GC Content of Oocyte- and 2C Embryo-specific 5' UTRs

We wanted to compare GC content of oocyte- and 2C embryo-specific 5' UTR extensions, as higher GC content is generally associated with tighter translational regulation. For this and following analysis, we selected all 5' UTR elongations in the oocytes (not elongated in somatic tissues), including those genes that are also elongated in 2C embryos, and 5' UTR extension specific for 2C embryos (as those also elongated in the oocytes are to some extent mRNAs originating in the oocytes stored after fertilization and might not be transcribed in 2C embryos themselves). To calculate GC content, we used custom made GC script generated previously in the laboratory (Appendix 1), using the genomic coordinates of 5' UTR extensions and raw genomic sequence. As a control, we used annotated 5' UTRs from chromosome 1.

As expected, 2C embryo-specific, and particularly oocyte-specific 5' UTR extensions were more GC rich than the official 5' UTR. Interestingly, the 2C embryo 5' UTR GC content (an average of 51.09%, a median of 49.66%) was more similar to the GC content of the official 5' UTRs (an average of 50.63%, a median of 48.81%) than to the oocyte 5' UTR GC content (an average of 55.59%, a median of 53.56%) (Figure 9).

In order to find out whether our results were statistically significant, we performed the Mann-Whitney U nonparametric test for both oocytes and 2C embryos in categories of less than or equal to 500 bp of elongations, less than or equal to 1000 bp of elongations, and less than or equal to 3000 bp of elongations (Table 7). At a .05 significance level, we can conclude that the oocytes and the official annotated 5' UTRs are nonidentical populations, as well as the 2C embryos and the official annotated 5' UTRs are nonidentical populations.

Table 7: Results of the statistical Mann-Whitney U test.

Tissue	W	p-value
Oocyte with an upstream 5' UTR less than or equal to 500 bp	732 248	< 2.2e-16
Oocyte with an upstream 5' UTR less than or equal to 1000 bp	1 028 177	< 2.2e-16
Oocyte with an upstream 5' UTR less than or equal to 3000 bp	1 605 439	< 2.2e-16
2C Embryo with an upstream 5' UTR less than or equal to 500 bp	737 123	4.041e-14
2C Embryo with an upstream 5' UTR less than or equal to 1000 bp	1 115 969	2.524e-09
2C Embryo with an upstream 5' UTR less than or equal to 3000 bp	1 537 453	0.0003847

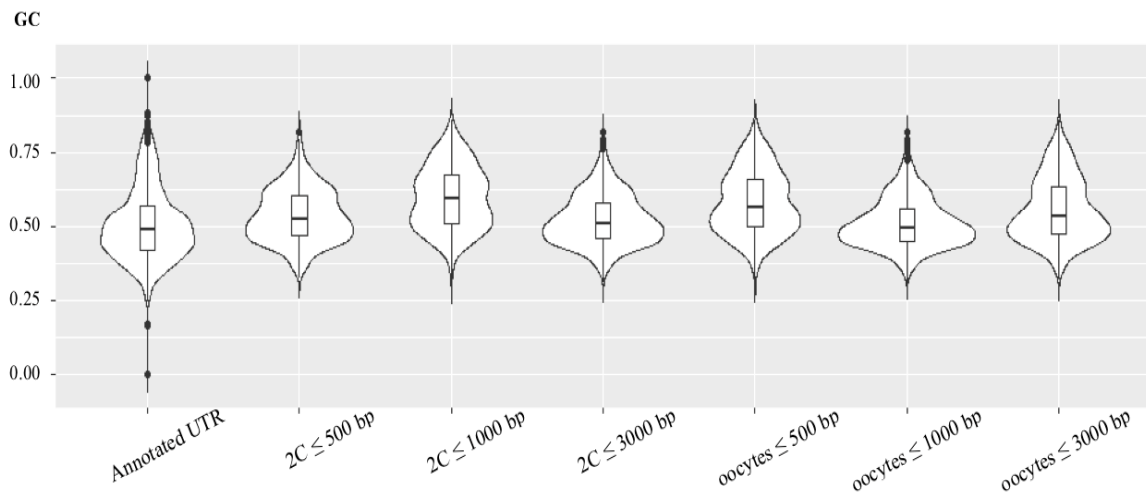


Figure 9: Comparison of GC content of oocytes and 2C embryos expressed as a violin plot. From left to right: Annotated UTR, 2C embryo with 5' UTR smaller or equal to 500 bp. 2C embryo with 5' UTR smaller or equal to 1000 bp. 2C embryo with 5' UTR smaller or equal to 3000 bp. Oocyte with 5' UTR smaller or equal to 500 bp. Oocyte with 5' UTR smaller or equal to 1000 bp. Oocyte with 5' UTR smaller or equal to 3000 bp. The horizontal line corresponds to median, displayed within the quartile range ("box") of each respective dataset.

7.7 uORFs

We wanted to compare uORFs occurrence in 5' UTR extensions in oocytes and 2C embryos compared to the annotated 5' UTRs as they represent one of the means of translational regulation within 5' UTRs. We first generated sequences of the 5' UTR extensions (for the same regions as in the previous chapter) including control annotated 5' UTRs, using a custom-made script generated previously in the laboratory (Appendix 2). Then, we analyzed uORF content using program ORF Finder, aiming to identify all ORFs with length 5 codons or more. Because the program can analyze only with one sequence, we randomly selected 46 sequences from the official 5' UTR annotation to serve as a control, 39 sequences from oocytes with upstream 5' UTR shorter or equal to 3000 bp, and 39 sequences from embryos with upstream 5' UTR shorter or equal to 3000 bp. For oocyte and 2C embryos, we selected 13 random sequences up to 200 bp in length, 13 random sequences from 200 to 1000 bp in length, and 13 sequences from 1000 to 3000 bp in length. The numbers of found ORFs were transformed into the number of ORFs per 100 bp of the UTR sequence. On average, fewer uORFs were found in oocyte and 2C embryo 5' UTR extensions compared to the annotated 5' UTRs annotation (Figures 10 and 11). The difference was

especially pronounced for oocytes that were observed to have on average 0.29059 fewer uORFs per 100 bp compared to the official 5' UTRs. In contrast, 2C embryos had on average only 0.028483 fewer uORFs per 100 bp compared to the official 5' UTRs.

Median per 100 bp reflected the same trend that uORFs between 2C embryos and the official 5' UTR annotations were more similar than uORFs between oocytes and annotated 5' UTRs (Figures 10 and 11). While the control annotated 5' UTRs had a median of 0.9009334 uORFs per 100 bp, oocytes had a median of 0.623053 uORFs per 100 bp, and 2C embryos a median of 0.846262 uORFs per 100 bp.

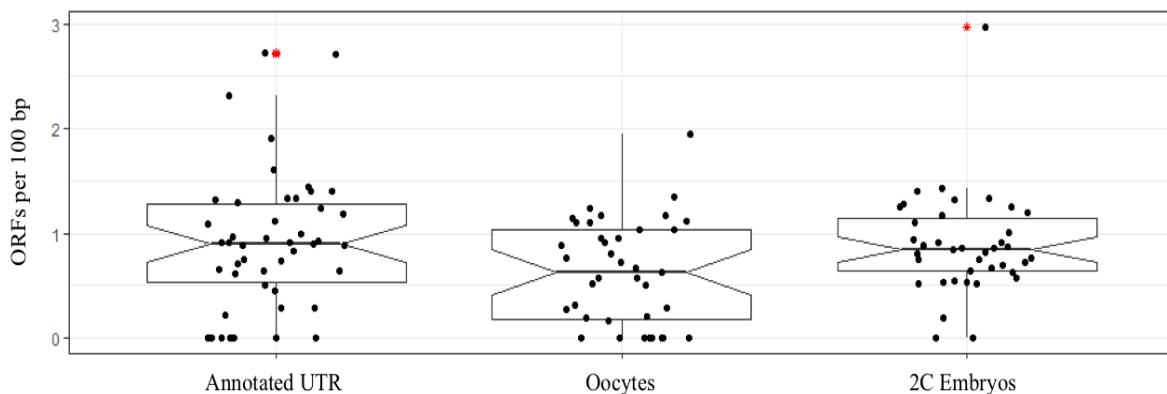


Figure 10: A boxplot with jitter plot and outliers. From left to right: ORFs per 100 bp in the official annotated 5' UTRs. ORFs per 100 bp in oocytes. ORFs per 100 bp in 2C embryos. Outliers denoted as red stars. Each of the “boxes” in the boxplots depicts the interquartile range. The horizontal line corresponds to median, and the notch depicts the confidence interval around the median.

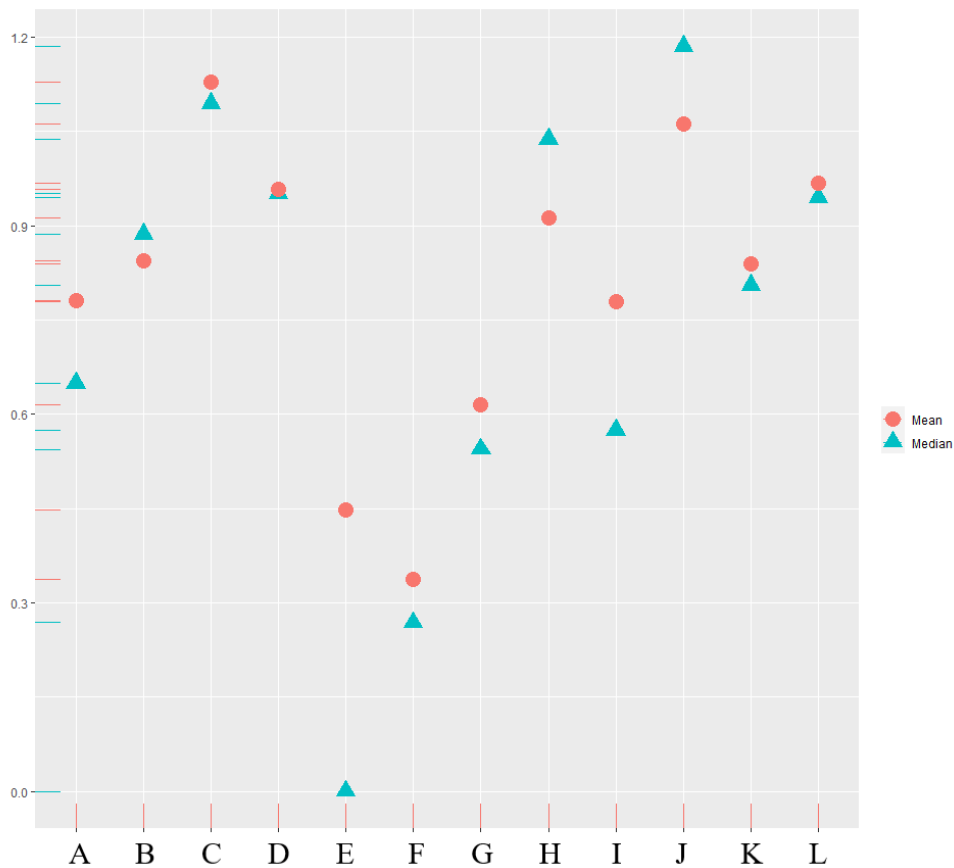


Figure 11: Scatterplot with ORF average and median per 100 bp for 12 categories. Mean is denoted as a red circle; median is denoted as a blue triangle. A: Control official annotated 5' UTR for ORFs shorter than 200 bp. B: Control official annotated 5' UTR for ORFs between 200-500 bp. C: Control official annotated 5' UTR for ORFs between 500-1000 bp. D: Control official annotated 5' UTR for ORFs between 1000-3000 bp. E: Oocyte 5' UTR for ORFs shorter than 200 bp. F: Oocyte 5' UTR for ORFs between 200-500 bp. G: Oocyte 5' UTR for ORFs between 500-1000 bp. H: Oocyte 5' UTR for ORFs between 1000-3000 bp. I: 2C embryo 5' UTR for ORFs shorter than 200 bp. J: 2C embryo 5' UTR for ORFs between 200-500 bp. K: 2C embryo 5' UTR for ORFs between 500-1000 bp. L: 2C embryo 5' UTR for ORFs between 1000-3000 bp.

7.8 Sequence Motifs

In this section, we wanted to identify enriched sequence motifs in oocyte and 2C 5' UTR elongations and check whether they are significantly similar to binding sites of RBPs or recognition motifs of miRNAs. To achieve this, we used sequences of 5' UTR elongations and control annotated 5' UTRs in fasta format generated in the same way as in the previous chapter. We generated sequences for 5' UTR elongations up to 500 bp, up to 1000 bp, and up to 3000 bp. These sequences were submitted into the MEME sequence analysis tool, with sequences of annotated 5' UTRs as a control, to find enriched sequence motifs. Motifs with

significant p-values (below 0.1) were further submitted to the Tomtom program that searched for the significant similarity of our identified motifs with known binding motifs of RBPs and miRNA recognition sites.

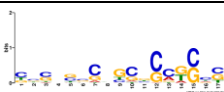

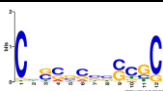

The oocyte- and embryo-specific 5' UTRs with elongations shorter or equal to 3000 bp contain 136 miRNA binding sites. In contrast, no RBP binding sites were identified for oocyte and embryo-specific 5' UTRs with elongations shorter or equal to 3000 bp. Interestingly, we observed 3 RBP binding sites in oocyte- and embryo-specific 5' UTRs with elongations shorter or equal to 500 bp. The same trend was observed for oocyte- and embryo-specific 5' UTRs with elongations shorter or equal to 1000 bp. All of the identified RBP binding sites serve as binding sites for 2 proteins – KHDRBS1 and RBM38.

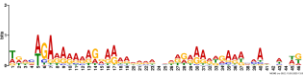
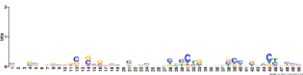
Overall, more miRNA binding sites were observed in oocytes than in 2C embryos. While oocytes carry 105 miRNA binding sites situated on 2 motifs, 2C embryos were found to have 33 miRNA binding sites situated on 1 motif. The motif with the most binding sites for miRNA is Motif2 of the oocyte 5' UTRs with elongations shorter or equal to 3000 bp, with 64 identified miRNA binding sites in total. For the embryo dataset, most miRNA binding sites were discovered in Motif1 of embryo 5' UTRs with elongations shorter or equal to 1000 bp, with 36 observed binding sites for miRNA. The motif with the smallest number of binding sites for miRNA in oocyte- and embryo-specific 5' UTRs is Motif1 of the embryo 5' UTRs with elongations shorter or equal to 500 bp, carrying 7 miRNA binding sites.

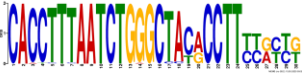
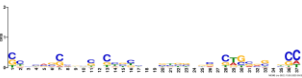


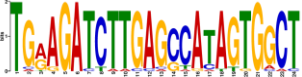
The greatest number of identified motifs is in 2C embryo 5' UTR elongations shorter or equal to 1000 bp (5 motifs). The number of identified motifs in oocytes is 2 for every 5' UTR elongation.



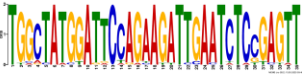


Overall, the results of this section suggest that 5' UTR elongations might serve as recognition motifs for miRNAs, resulting in translational downregulation, and potentially as binding sites for RBPs. Motif counts, sequences, logos, and binding sites for RBPs and miRNAs are below in Table 8 (miRNA) and Table 9 (RBPs).

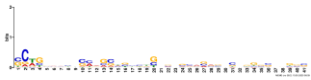
Table 8: A list of motif counts, sequences, logos, and miRNA binding sites found in oocyte- and embryo-specific 5' UTRs.

Dataset	Motif	Sequence	Logo	miRNA binding sites
Oocyte < 500 bp	Motif1	CBSNSSCHG CBCCKSCS		MIMAT0019336, MIMAT0020637, MIMAT0027790, MIMAT0016990, MIMAT0028070, MIMAT0028413, MIMAT0020615, MIMAT0020644, MIMAT0025092, MIMAT0020641, MIMAT0027770, MIMAT0017030, MIMAT0017039, MIMAT0013803, MIMAT0022357, MIMAT0027910, MIMAT0014934, MIMAT0027860, MIMAT0025583, MIMAT0027994, MIMAT0014939, MIMAT0003499, MIMAT0011213, MIMAT0027942, MIMAT0031405, MIMAT0020630, MIMAT0014841
	Motif2	TTTTCTCAA T TAGTGAT CAAGGGG AAA		MIMAT0022359, MIMAT0027751, MIMAT0017084, MIMAT0000648, MIMAT0004854, MIMAT0025136, MIMAT0003173, MIMAT0017261, MIMAT0000655
Oocyte < 1000 bp	Motif1	CHSSCSBS CGC		MIMAT0019336, MIMAT0020637, MIMAT0007879, MIMAT0020641, MIMAT0020644, MIMAT0025583, MIMAT0003892, MIMAT0027860, MIMAT0029815, MIMAT0017039, MIMAT0031405, MIMAT0027773, MIMAT0029838, MIMAT0029849, MIMAT0016990, MIMAT0027788, MIMAT0020630, MIMAT0028068, MIMAT0027838, MIMAT0005860, MIMAT0027862
	Motif2	GGGAGG*		MIMAT0027875, MIMAT0017342, MIMAT0014951, MIMAT0003120, MIMAT0027747, MIMAT0027855, MIMAT0027995, MIMAT0028039, MIMAT0003127, MIMAT0004781, MIMAT0027719, MIMAT0027805, MIMAT0027983, MIMAT0028023, MIMAT0028053, MIMAT0004862,

				MIMAT0019136, MIMAT0000240, MIMAT0027977, MIMAT0028037
Oocyte < 3000 bp	Motif1	WDDVWGA RARARAGR ARAADDRRA NHDDRRRA RDGARWV MADVWNW K		MIMAT0027815, MIMAT0027793, MIMAT0027871, MIMAT0028130, MIMAT0027989, MIMAT0029871, MIMAT0028021, MIMAT0027693, MIMAT0029801, MIMAT0027799, MIMAT0027923, MIMAT0027921, MIMAT0027931, MIMAT0007868, MIMAT0027987, MIMAT0028390, MIMAT0004862, MIMAT0025087, MIMAT0027829, MIMAT0022985, MIMAT0028132, MIMAT0027751, MIMAT0027853, MIMAT0014956, MIMAT0027893, MIMAT0022690, MIMAT0028091, MIMAT0027837, MIMAT0027927, MIMAT0028001, MIMAT0017208, MIMAT0028063, MIMAT0027947, MIMAT0003488, MIMAT0029805, MIMAT0028027, MIMAT0004704, MIMAT0028093, MIMAT0027823, MIMAT0028033, MIMAT0000668
	Motif2	SNNVVNDG SVDSNSMR VVVNSHBB NNBBHBCT SNNBHBCY NBNBCYBB SV		MIMAT0020637, MIMAT0017030, MIMAT0020615, MIMAT0007878, MIMAT0000569, MIMAT0028076, MIMAT0020641, MIMAT0027832, MIMAT0031405, MIMAT0019336, MIMAT0028133, MIMAT0028004, MIMAT0028046, MIMAT0004861, MIMAT0003499, MIMAT0003731, MIMAT0027900, MIMAT0027918, MIMAT0007876, MIMAT0011213, MIMAT0027910, MIMAT0029836, MIMAT0017039, MIMAT0025112, MIMAT0014852, MIMAT0027864, MIMAT0017059, MIMAT0029890, MIMAT0027788, MIMAT0005460, MIMAT0027870, MIMAT0016990, MIMAT0027940, MIMAT0027838, MIMAT0025110, MIMAT0031419, MIMAT0031403, MIMAT0027994, MIMAT0027816, MIMAT0004782,



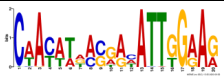

				MIMAT0029843, MIMAT0035714, MIMAT0029902, MIMAT0014862, MIMAT0028438, MIMAT0028413, MIMAT0027896, MIMAT0028000, MIMAT0027936, MIMAT0028048, MIMAT0027806, MIMAT0031427, MIMAT0007879, MIMAT0028070, MIMAT0027718, MIMAT0027846, MIMAT0028050, MIMAT0000549, MIMAT0009446, MIMAT0027830, MIMAT0014941, MIMAT0022986, MIMAT0020644, MIMAT0007867
2C embryo < 500 bp	Motif1	CACCTTTA ATCTGGGC TACRCCTT YYRYYYK		MIMAT0020605, MIMAT0004572, MIMAT0000235, MIMAT0025164, MIMAT0031408, MIMAT0027343, MIMAT0022364
	Motif2	SCHSVSSHB BSNCYBYH NNSYBBNB SNCWSVVS NBCC		MIMAT0019336, MIMAT0029799, MIMAT0020637, MIMAT0031405, MIMAT0029902, MIMAT0020641, MIMAT0020615, MIMAT0028438, MIMAT0027816, MIMAT0004821, MIMAT0035714, MIMAT0031425, MIMAT0003499, MIMAT0014862, MIMAT0029878, MIMAT0028133, MIMAT0004853, MIMAT0027712, MIMAT0014934, MIMAT0004187, MIMAT0028068, MIMAT0011213, MIMAT0016990, MIMAT0017245, MIMAT0003892
	Motif3	CWAYATDA SGAHATTG GAAG		MIMAT0025151, MIMAT0009407, MIMAT0029896, MIMAT0017018, MIMAT0003181, MIMAT0029871, MIMAT0004533, MIMAT0000667, MIMAT0004634, MIMAT0025099
	Motif4	ATTAAAGG TGTGGTG		MIMAT0029876, MIMAT0027833, MIMAT0022371, MIMAT0025108, MIMAT0017328, MIMAT0019356, MIMAT0004869, MIMAT0029853, MIMAT0019136
	Motif5	TGRAGATC TTGAGCCA TAGTGGCT		MIMAT0027991, MIMAT0027963, MIMAT0031398, MIMAT0027807,

				MIMAT0000159, MIMAT0000669, MIMAT0004704, MIMAT0029826
2C embryo < 1000 bp	Motif1	CHRGVRND GVVVRGSM NDNVVDS HVNSDBYY YYNNVNSY BBNSNS		MIMAT0020630, MIMAT0025178, MIMAT0031405, MIMAT0014939, MIMAT0029890, MIMAT0025583, MIMAT0031409, MIMAT0027755, MIMAT0027754, MIMAT0014858, MIMAT0005460, MIMAT0029856, MIMAT0027730, MIMAT0003499, MIMAT0015646, MIMAT0014900, MIMAT0016997, MIMAT0027758, MIMAT0007876, MIMAT0019336, MIMAT0028046, MIMAT0020621, MIMAT0020637, MIMAT0007867, MIMAT0029868, MIMAT0025166, MIMAT0014901, MIMAT0017030, MIMAT0028413, MIMAT0014917, MIMAT0029802, MIMAT0014899, MIMAT0025110, MIMAT0009459, MIMAT0016982, MIMAT0025144
	Motif2	THATCTKG RCTATRBY T		MIMAT0001091, MIMAT0020605, MIMAT0003499, MIMAT0000235, MIMAT0000745, MIMAT0004665, MIMAT0003511, MIMAT0004643, MIMAT0022378, MIMAT0029856
	Motif3	TGGCTATG GATTCCAG AAGATTGA ATCTCCGA GTT		MIMAT0000131, MIMAT0000655, MIMAT0004843, MIMAT0027847, MIMAT0027753, MIMAT0004891, MIMAT0004623, MIMAT0029829
	Motif4	CACACCTT		MIMAT0025132, MIMAT0025091, MIMAT0027104, MIMAT0027904, MIMAT0029856, MIMAT0003461, MIMAT0028080, MIMAT0031401, MIMAT0027343, MIMAT0014898
	Motif5	TGCTGGAG AC		MIMAT0029801, MIMAT0028425, MIMAT0000659, MIMAT0000656, MIMAT0024857, MIMAT0028415, MIMAT0027699, MIMAT0028449, MIMAT0000144, MIMAT0020614, MIMAT0025108, MIMAT0029855, MIMAT0028419, MIMAT0025083,

				MIMAT0027970, MIMAT0028393, MIMAT0004839, MIMAT0027979, MIMAT0014877
2C embryo < 3000 bp	Motif1	SCWGBVVB NSMDGSHB BBWSNVNS VRVVBNSN KGSSNNGSS		MIMAT0017179, MIMAT0024862, MIMAT0011213, MIMAT0028054, MIMAT0022362, MIMAT0000240, MIMAT0027970, MIMAT0027819, MIMAT0027881, MIMAT0000649, MIMAT0029902, MIMAT0003898, MIMAT0014960, MIMAT0027846, MIMAT0025089, MIMAT0027983, MIMAT0000540, MIMAT0027790, MIMAT0014862, MIMAT0017039, MIMAT0014939, MIMAT0028135, MIMAT0004643, MIMAT0017009, MIMAT0003458, MIMAT0028392, MIMAT0017246, MIMAT0000556, MIMAT0009391, MIMAT0027988, MIMAT0000597, MIMAT0009405, MIMAT0000385

**Evaluated as very similar to other earlier specified motifs and may be biasing the results*

Table 9: A list of motif counts, sequences, logos, and RBP binding sites found in oocyte- and embryo-specific 5' UTRs.

Dataset	Motifs	Sequence	Logo	RBP binding site
Oocyte < 500bp	Motif2	TTTTCTCAATTA GTGATCAAGGG GGAAA		KHDRBS1
Oocyte < 1000 bp	Motif2	GGGAGG*		RBM38
2C embryo < 500 bp	Motif3	CWAYATDASGA HATTGGAAG		KHDRBS1
	Motif4	ATTAAAGGTGTG GTG		KHDRBS1, RBM38

**Evaluated as very similar to other earlier specified motifs and may be biasing the results.*

8 DISCUSSION

The official annotation of 5' UTRs in genome browsers mostly represents the 5' UTRs of respective mRNAs in somatic tissues. However, recent publications (Veselovska et al., 2015; Zhang et al., 2014a) revealed that transcriptomes of oocytes and early embryos appear to differ from somatic tissues.

In order to explore the oocyte- and embryo-specific 5' UTRs, we processed publicly available RNA-seq datasets from various developmental stages of mouse, including oocytes, preimplantation embryos, and adult somatic tissues. We inspected appropriate literature and databases to produce an extensive knowledge base for our research.

Through our work, we identified novel upstream 5' UTRs of annotated genes in oocytes, embryos, and adult somatic tissues. The highest number of novel upstream 5' UTRs was identified in oocytes and early embryos (2C, 4C), followed by brain, lung, and spleen from the somatic datasets. We speculate that at least some of the novel upstream 5' UTRs that we discovered in early embryos (especially 2C) could represent retained oocyte mRNA after fertilization. Alternatively, it could have been newly transcribed mRNA from the same upstream promoter as in the oocytes.

It is important to note that some 5' UTRs might have been misannotated during the de novo transcriptome assembly, for example, 5' UTRs might not be fully annotated in cases when transcripts with longer 5' UTR has low expression, and therefore the prolonged 5' UTR is either missing or is divided into an array of same strand monoexonic genes in the proximity of 5' end of the gene. In addition, there might be problems with the annotation in the case of bidirectional promoters in datasets without strand specificity, where the gene transcribed in the opposite direction could be misannotated as prolongation of 5' UTR of the first gene because of the lack of strand information of the RNA-seq reads. The preimplantation embryonic datasets are not strand-specific. Consequently, the assemblies from embryonic data may not be as accurate.

Afterwards, we compared the novel upstream 5' UTRs found in 2C embryos and oocytes. The number of embryo-specific upstream 5' UTRs found was higher than the number of oocyte-specific 5' UTRs, although the difference was quite small (109 genes). Interestingly, the number of genes found in both 2C embryos and oocytes was much lower, with only 668 genes in total.

We characterized those genes with an upstream 5' UTRs in oocytes and 2C embryos that lacked an upstream in all somatic tissues. A number of these elongations was very long compared to the official annotated 5' UTRs. We discarded such elongations that were longer than 3000 bp, as they probably represented alternative splicing isoforms.

The fact that we identified oocyte- and embryo-specific 5' UTRs that were longer in length than the annotated 5' UTRs is not surprising, not only because exclusively 5' UTRs displaying elongations were analyzed, but also since it was previously observed that the greater length of 5' UTR corresponds to a greater level of regulation (Hurowitz and Brown, 2003; David et al., 2006). Following fertilization, the cell divides and differentiates, and appropriate control is crucial for proper development of the organism, hence adequate regulation is necessary. Interestingly, the elongations of 5' UTRs of 2C embryos were observed to be greater in length than those in oocytes. However, majority of the elongations was rather short, with medians of 340.5 bp for oocytes and 475 bp for 2C embryos, compared to the annotated 5' UTRs.

We further inspected those genes that were elongated in 5' UTRs in both 2C embryos and oocytes. We discovered that the total length of elongations in 5' UTRs of 2C embryos was 39.518% greater compared to the total elongations of 5' UTRs in oocytes. This is interesting when combined with the fact that the upstream 5' UTR elongations in 2C embryo were 50.042% longer on average compared to the oocyte elongations. We discovered that 269 of these 5' UTR elongations in 2C embryos were most likely retentions of oocyte mRNA as the difference between the oocyte and 2C elongation was shorter than 200 bp and might, therefore, represent the same misannotated 5' UTR.

Through our analysis, we discovered that oocyte- and embryo-specific 5' UTRs were more GC rich than the official annotated 5' UTRs. As the GC content is greater, it is likely that the upstream regions contained secondary structures. The GC content of embryo-specific 5' UTRs was only 0.46% greater than the GC content of the official 5' UTRs, while the GC content of oocyte-specific 5' UTRs was 4.9% greater than the GC content of the official 5' UTRs. The GC content of both oocyte- and embryo-specific 5' UTRs proved to belong to statistically different populations than the GC content of the official annotated 5' UTRs. The greater GC content for oocyte- and embryo-specific 5' UTRs matched our expectations, since, generally, GC bond confers greater stability per base than an AU hairpin and therefore the translation can be more precisely regulated (Babendure et al., 2006). The early

developmental stages require strictly controlled translation as translation timing could be crucial in pattern formation and differentiation.

We also discovered that on average, fewer uORFs were found in oocyte- and 2C embryo-specific upstream 5' UTRs compared to the annotated 5' UTRs. The difference was greater for oocyte-specific 5' UTRs than for embryo-specific 5' UTRs. However, these were just the elongations of the annotated 5' UTRs, therefore, the translation of respective mRNAs is affected by uORFs from both the elongation and the annotated 5' UTRs. As Calvo previously stated that uORFs reduce the protein expression on average by 30-80% (Calvo et al., 2009), extra uORFs present in the 5' UTRs elongations might cause translational downregulation on top of the effect of uORFs in the annotated 5' UTRs. On the other hand, the effect of uORFs being distant from the main AUG codon is not clear. In addition, uORFs can result in the translation of short peptides (Ji et al., 2015), which might affect oogenesis or early development.

We identified 136 new recognition sites for miRNA in oocyte- and embryo-specific 5' UTRs for elongations shorter or equal to 3000 bp. This possibly infers tighter regulation of genes via mRNA-miRNA interaction, likely through translational repression and/or degradation of RNA (Bartel, 2009; MacFarlane and Murphy, 2010). Probably the most likely scenario, due to their position in 5' UTR, is that the miRNA could block translational initiation. It is not surprising that such a large number of sites was identified since it is predicted that a great portion of protein-coding genes (30% in humans) is regulated through miRNAs (Rajewsky, 2006). Nevertheless, the probability of identified miRNA to regulate mRNAs should be further explored by analyzing the expression profile of individual miRNAs to find out whether they are expressed in relevant developmental stages and to analyze their stoichiometry with the respective mRNAs (i.e. if they are capable of imposing regulation).

In contrast to miRNA recognition sites, not many RBP binding sites were discovered. We identified the binding sites for two RBPs – RBM38 and KHDRBS1. RBM38 is an RBP that serves as a target of the p53 family and regulates the expression of p53 through mRNA translation (Zhang et al., 2014b). Mice that are RBM38-deficient are susceptible to spontaneous tumor development, express hematopoietic defects, and are vulnerable to accelerated aging (Zhang et al., 2014b).

KHDRBS1 (KH domain containing, RNA binding, signal transduction associated protein 1) functions as an adapter protein in signal transduction cascades and has a role in

G2-M progression in the cell cycle (Kim et al., 2016). When homozygously mutated, the gene prevents mice from experiencing age-related bone loss and formation of fatty bone marrow, leaving males infertile (Demontiero et al., 2011). These two RBPs were therefore not associated with oogenesis or preimplantation development yet, therefore, we would first have to check whether they are expressed in the relevant developmental stages, before exploring them further.

An alternative approach to identifying RBP binding sites would be to use RBPmap (Paz et al., 2014). While this tool provides all of the motifs of all RBPs and outputs their location within a given sequence, it offers a lot of redundant information (e.g. proteins binding to 3' UTRs). RBPmap also does not evaluate the results statistically. Another way would be to use the DREME tool (Bailey, 2011) to search for discriminative regular expression motif elicitation. Still, it is not guaranteed that any relevant RBP binding sites would be discovered, since RBPs in 5' UTRs often have binding sites with unknown sequence binding motifs.

This project will serve as a basis for future bioinformatic analyses of 5' UTRs as well as experimental functional analyses of 5' UTRs of selected genes. The analysis could be extended to other selected mammalian species to offer an overview of the oocyte- and embryo-specific 5' UTRs across species. Alternatively, the data could be analyzed to obtain information about the transposable elements, intra-UTR splicing, or novel 5' UTRs resulting from alternative splicing of the coding sequence, or insights into the secondary structures found within the 5' UTRs elongations. In addition, the analysis will be continued by pairwise comparison of annotated and novel 5' UTRs of the same genes to obtain more insights into the effect of the 5' UTRs elongations. Experimentally, we could analyze the effect of prolonged 5' UTRs on transcription, localization, stability, and rate of translation of mRNAs of selected candidate genes and its potential implications for early mammalian development.

9 CONCLUSION

The 5' UTRs of oocytes and early embryos appear to be different from the 5' UTRs found in somatic tissues, that represent the official annotation of 5' UTR. In order to identify oocyte- and embryo-specific upstream 5' UTRs, we processed RNA-Seq datasets from mouse oocytes, preimplantation embryos, and somatic tissues. We identified novel upstream 5' UTRs for oocytes, preimplantation embryos, and adult somatic tissues.

We focused on 5' UTRs specific for oocytes and 2C embryos. We obtained the means and medians for the total length of elongations of oocyte- and 2C embryo-specific 5' UTRs as compared to the official annotated 5' UTR. We further characterized the differences in length for genes that were found in both oocyte and 2C embryo 5' UTRs. Our findings matched our assumption that we would find oocyte and embryo-specific 5' UTRs that were longer, as greater length tends to indicate a mechanism of tighter control that is crucial for the proper development of an organism.

The characterization was further extended to the analysis of GC content and uORFs of oocyte- and embryo-specific 5' UTRs. We observed that the GC content was higher for oocyte- and embryo-specific 5' UTRs than for the official annotated 5' UTRs. Interestingly, a lower number of uORFs was identified for oocyte- and embryo-specific 5' UTRs, and we speculate that it is due to an increase in protein expression.

Moreover, we identified 136 recognition sites for miRNAs and no binding sites for RBP in oocyte- and embryo-specific 5' UTRs with elongations smaller or equal to 3000 bp. 68.5714% more miRNA recognition sites were observed in oocytes than in 2C embryos. Further analysis of RBP binding sites is necessary in order to draw a conclusion.

Our data, python scripts, and results could further be used to increase the depth of characterization for oocyte- and embryo-specific 5' UTRs. It could also serve as a guideline for such characterization to be extended to other mammalian species, and analyses of other features, such as transposable element content, presence of secondary structures, etc. One could also develop upon the elements of this research project that were not addressed – e.g. the characterization of genes that has a 5' UTR that was longer than 3000 bp. Moreover, these results confirmed the claim laid by recent publication (Veselovska et al., 2015; Zhang et al., 2014), as the oocyte- and embryo-specific 5' UTRs were found to differ from the official annotated 5' UTRs.

10 REFERENCES

- Amaldi, F., and P. Pierandrei-Amaldi. "TOP Genes: a Translationally Controlled Class of Genes Including Those Coding for Ribosomal Proteins." *Prog Mol Subcell Biol.*, vol. 18, 1997, pp. 1–17., doi:10.1007/978-3-642-60471-3_1.
- Andergassen, D., et al. "Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression." *Elife*, vol. 6, 2017, doi: 10.7554/eLife.25125.
- Andreev, D. E., et al. "Differential Contribution of the m7G-Cap to the 5' End-Dependent Translation Initiation of Mammalian MRNAs." *Nucleic Acids Research*, vol. 37, no. 18, 2009, pp. 6135–6147., doi:10.1093/nar/gkp665.
- Araujo, P. R., et al. "Before It Gets Started: Regulating Translation at the 5' UTR." *Comparative and Functional Genomics*, vol. 2012, 2012, pp. 1–8., doi:10.1155/2012/475731.
- Babendure, J.F., et al. "Control of mammalian translation by mRNA structure near caps" *RNA*, vol. 12, no. 5, 2006, pp. 851–861., doi: 10.1261/rna.2309906.
- Bailey, T. L. "DREME: Motif Discovery in Transcription Factor ChIP-Seq Data." *Bioinformatics*, vol. 27, no. 12, 2011, pp. 1653–1659., doi:10.1093/bioinformatics/btr261.
- Bailey, T. L., and M. Gribskov. "Combining Evidence Using p-Values: Application to Sequence Homology Searches." *Bioinformatics*, vol. 14, no. 1, 1998, pp. 48–54., doi:10.1093/bioinformatics/14.1.48.
- Barrett, L. W., et al. "Regulation of Eukaryotic Gene Expression by the Untranslated Gene Regions and Other Non-Coding Elements." *Cellular and Molecular Life Sciences*, vol. 69, no. 21, 2012, pp. 3613–3634., doi:10.1007/s00018-012-0990-9.
- Bartel, D. P. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell*, vol. 136, no. 2, 2009, pp. 215–233., doi:10.1016/j.cell.2009.01.002.
- Bastide, A., et al. "An Upstream Open Reading Frame Within an IRES Controls Expression of a Specific VEGF-A Isoform." *Nucleic Acids Research*, vol. 36, no. 7, 2008, pp. 2434–2445., doi:10.1093/nar/gkn093.
- Beaudoin, J.-D., and J.-P. Perreault. "5-UTR G-Quadruplex Structures Acting as Translational Repressors." *Nucleic Acids Research*, vol. 38, no. 20, 2010, pp. 7022–7036., doi:10.1093/nar/gkq557.
- Bhattacharyya, D., et al. "An Independently Folding RNA G-Quadruplex Domain Directly Recruits the 40S Ribosomal Subunit." *Biochemistry*, vol. 54, no. 10, 2015, pp. 1879–1885., doi:10.1021/acs.biochem.5b00091.
- Bonnal, S., et al. "A Single Internal Ribosome Entry Site Containing a G Quartet RNA Structure Drives Fibroblast Growth Factor 2 Gene Expression at Four Alternative Translation Initiation Codons." *J. Biol. Chem.*, vol. 278, no. 41, 2003, pp. 39330–39336., doi:10.1074/jbc.M305580200.
- Bradshaw, R. A., and P. Stahl. "Evidence for Cellular mRNA IRESs." *Encyclopedia of Cell Biology*, 1st ed., vol. 3, Elsevier, 2016, pp. 313–313.
- Brust, V., et al. "Lifetime Development of Behavioural Phenotype in the House Mouse (*Mus Musculus*)." *Frontiers in Zoology*, vol. 12, no. Suppl 1, 2015, doi:10.1186/1742-9994-12-s1-s17.

- Bugaut, A., and S. Balasubramanian. "5-UTR RNA G-Quadruplexes: Translation Regulation and Targeting." *Nucleic Acids Research*, vol. 40, no. 11, 2012, pp. 4727–4741., doi:10.1093/nar/gks068.
- Calvo, S. E., et al. "Upstream Open Reading Frames Cause Widespread Reduction of Protein Expression and Are Polymorphic among Humans." *Proceedings of the National Academy of Sciences*, vol. 106, no. 18, 2009, pp. 7507–7512., doi:10.1073/pnas.0810916106.
- Cammas, A., et al. "Stabilization of the G-Quadruplex at the VEGF IRES Represses Cap-Independent Translation." *RNA Biology*, vol. 12, no. 3, 2015, pp. 320–329., doi:10.1080/15476286.2015.1017236.
- Carninci, P., et al. "The Transcriptional Landscape of the Mammalian Genome." *Science*, vol. 309, no. 5740, 2005, pp. 1559–1563., doi:10.1126/science.1112014.
- Cassan, M., and J.-P. Rousset. "UAG Readthrough in Mammalian Cells: Effect of Upstream and Downstream Stop Codon Contexts Reveal Different Signals." *BMC Mol Biol.*, vol. 2, no. 3, 2001, doi:10.1186/1471-2199-2-3.
- Chappell, S. A., et al. "A 9-Nt Segment of a Cellular mRNA Can Function as an Internal Ribosome Entry Site (IRES) and When Present in Linked Multiple Copies Greatly Enhances IRES Activity." *Proceedings of the National Academy of Sciences*, vol. 97, no. 4, 2000, pp. 1536–1541., doi:10.1073/pnas.97.4.1536.
- Chatterjee, S., and J. K. Pal. "Role of 5'- and 3'-Untranslated Regions of MRNAs in Human Diseases." *Biology of the Cell*, vol. 101, no. 5, 2009, pp. 251–262., doi:10.1042/bc20080104.
- Chen, C.-H., et al. "The Plausible Reason Why the Length of 5 Untranslated Region Is Unrelated to Organismal Complexity." *BMC Research Notes*, vol. 4, no. 1, 2011, doi:10.1186/1756-0500-4-312.
- Chen, T.-M., et al. "Overexpression of FGF9 in Colon Cancer Cells Is Mediated by Hypoxia-Induced Translational Activation." *Nucleic Acids Research*, vol. 42, no. 5, 2014, pp. 2932–2944., doi:10.1093/nar/gkt1286.
- Child, S. J., et al. "Translational Control by an Upstream Open Reading Frame in the HER-2/Neu Transcript." *Journal of Biological Chemistry*, vol. 274, no. 34, 1999, pp. 24335–24341., doi:10.1074/jbc.274.34.24335.
- David, L., et al. "A High-Resolution Map of Transcription in the Yeast Genome." *Proceedings of the National Academy of Sciences*, vol. 103, no. 14, 2006, pp. 5320–5325., doi:10.1073/pnas.0601091103.
- Davuluri, R. V., et al. "CART Classification of Human 5' UTR Sequences." *Genome Research*, vol. 10, no. 11, 2000, pp. 1807–1816., doi:10.1101/gr-1460r.
- Demontiero, O., et al. "Aging and Bone Loss: New Insights for the Clinician." *Therapeutic Advances in Musculoskeletal Disease*, vol. 4, no. 2, 2011, pp. 61–76., doi:10.1177/1759720x11430858.
- Djuranovic, S., et al. "MiRNA-Mediated Gene Silencing by Translational Repression Followed by mRNA Deadenylation and Decay." *Science*, vol. 336, no. 6078, 2012, pp. 237–240., doi:10.1126/science.1215691.

- Duret, L., et al. “Statistical Analysis of Vertebrate Sequences Reveals That Long Genes Are Scarce in GC-Rich Isochores.” *Journal of Molecular Evolution*, vol. 40, no. 3, 1995, pp. 308–317., doi:10.1007/bf00163235.
- Dvir, S., et al. “Deciphering the Rules by Which 5-UTR Sequences Affect Protein Expression in Yeast.” *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, 2013, doi:10.1073/pnas.1222534110.
- Fernandez, J., et al. “Internal Ribosome Entry Site-Mediated Translation of a Mammalian mRNA Is Regulated by Amino Acid Availability.” *Journal of Biological Chemistry*, vol. 276, no. 15, 2000, pp. 12285–12291., doi:10.1074/jbc.m009714200.
- Fernandez, J., et al. “Regulation of Internal Ribosome Entry Site-Mediated Translation by Eukaryotic Initiation Factor-2 α Phosphorylation and Translation of a Small Upstream Open Reading Frame.” *Journal of Biological Chemistry*, vol. 277, no. 3, 2002, pp. 2050–2058., doi:10.1074/jbc.m109199200.
- Fernandez, J., et al. “Ribosome Stalling Regulates IRES-Mediated Translation in Eukaryotes, a Parallel to Prokaryotic Attenuation.” *Molecular Cell*, vol. 17, no. 3, 2005, pp. 405–416., doi:10.1016/j.molcel.2004.12.024.
- Gebauer, F., and M. W. Hentze. “Molecular Mechanisms of Translational Control.” *Nature Reviews Molecular Cell Biology*, vol. 5, no. 10, 2004, pp. 827–835., doi:10.1038/nrm1488.
- Girelli, D., et al. “Hereditary Hyperferritinemia-Cataract Syndrome Caused by a 29-Base Pair Deletion in the Iron Responsive Element of Ferritin L-Subunit Gene.” *Blood*, vol. 90, no. 5, 1997, pp. 2084–2088., doi:10.1182/blood.v90.5.2084.
- Godet, A.-C., et al. “IRES Trans-Acting Factors, Key Actors of the Stress Response.” *International Journal of Molecular Sciences*, vol. 20, no. 4, 2019, pp. 924., doi:10.3390/ijms20040924.
- Gray, N. K., and M. W. Hentze. “Regulation of Protein Synthesis by mRNA Structure.” *Molecular Biology Reports*, vol. 19, no. 3, 1994, pp. 195–200., doi:10.1007/bf00986961.
- Gu, W., et al. “The Role of RNA Structure at 5' Untranslated Region in MicroRNA-Mediated Gene Regulation.” *Rna*, vol. 20, no. 9, 2014, pp. 1369–1375., doi:10.1261/rna.044792.114.
- Gupta, S., et al. “Quantifying Similarity between Motifs.” *Genome Biology*, vol. 8, no. 2, 2007, doi:10.1186/gb-2007-8-2-r24.
- Hershey, J. W. B., and W. C. Merrick. “The Pathway and Mechanism of Initiation of Protein Synthesis.” Cold Spring Harbor Laboratory Press, vol. 39, 2000, doi:10.1101/0.33-88.
- Hsieh, A. C., et al. “The Translational Landscape of MTOR Signalling Steers Cancer Initiation and Metastasis.” *Nature*, vol. 485, no. 7396, 2012, pp. 55–61., doi:10.1038/nature10912.
- Hughes, M. J., and D. W. Andrews. “A Single Nucleotide Is a Sufficient 5' Untranslated Region for Translation in an Eukaryotic in Vitro System.” *FEBS Letters*, vol. 414, no. 1, 1997, pp. 19–22., doi:10.1016/s0014-5793(97)00965-4.
- Hurowitz, E. H., and P. O. Brown. “Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*.” *Genome Biology*, vol. 5, no. 1, 2003, doi:10.1186/gb-2003-5-1-r2.

- Iwasaki, S., et al. “Rocaglates Convert DEAD-Box Protein eIF4A into a Sequence-Selective Translational Repressor.” *Nature*, vol. 534, no. 7608, 2016, pp. 558–561., doi:10.1038/nature17978.
- Jackson, R. J., et al. “The Mechanism of Eukaryotic Translation Initiation and Principles of Its Regulation.” *Nature Reviews Molecular Cell Biology*, vol. 11, no. 2, 2010, pp. 113–127., doi:10.1038/nrm2838.
- Jansen, R.-P. “MRNA Localization: Message on the Move.” *Nature Reviews Molecular Cell Biology*, vol. 2, no. 4, 2001, pp. 247–256., doi:10.1038/35067016.
- Ji, Z., et al. “Many LncRNAs, 5’ UTRs, and Pseudogenes Are Translated and Some Are Likely to Express Functional Proteins.” *ELife*, vol. 4, 2015, doi:10.7554/elife.08890.
- Jopling, C. L., et al. “L-Myc Protein Synthesis Is Initiated by Internal Ribosome Entry.” *Rna*, vol. 10, no. 2, 2004, pp. 287–298., doi:10.1261/rna.5138804.
- Jurka, J. “Repeats in Genomic DNA: Mining and Meaning.” *Curr. Opin. Struct. Biol.*, vol. 8, no. 3, 1998, pp. 333–337., doi:10.1016/S0959-440X(98)80067-5.
- Kim, D., et al. “HISAT: a fast spliced aligner with low memory requirements.” *Nat. Methods*, vol. 12, 2015, pp. 357–360., doi: 10.1038/nmeth.3317.
- Kim, S., et al. “Khdrbs1 - KH domain containing, RNA binding, signal transduction associated 1 (house mouse).” *Nucleic acids research*, vol. 47, 2016, PubChem, <https://pubchem.ncbi.nlm.nih.gov/gene/Khdrbs1/mouse>.
- Komar, A. A., and M. Hatzoglou. “Cellular IRES-Mediated Translation.” *Cell Cycle*, vol. 10, no. 2, 2011, pp. 229–240., doi:10.4161/cc.10.2.14472.
- Komar, A. A., and M. Hatzoglou. “Internal Ribosome Entry Sites in Cellular MRNAs: Mystery of Their Existence.” *Journal of Biological Chemistry*, vol. 280, no. 25, 2005, pp. 23425–23428., doi:10.1074/jbc.r400041200.
- Kozak, M. “An Analysis of 5-Noncoding Sequences from 699 Vertebrate Messenger RNAs.” *Nucleic Acids Research*, vol. 15, no. 20, 1987, pp. 8125–8148., doi:10.1093/nar/15.20.8125.
- Kozak, M. “An Analysis of Vertebrate MRNA Sequences: Intimations of Translational Control.” *The Journal of Cell Biology*, vol. 115, no. 4, 1991, pp. 887–903., doi:10.1083/jcb.115.4.887.
- Kozak, M. “Circumstances and Mechanisms of Inhibition of Translation by Secondary Structure in Eucaryotic MRNAs.” *Molecular and Cellular Biology*, vol. 9, no. 11, 1989, pp. 5134–5142., doi:10.1128/mcb.9.11.5134.
- Kozak, M. “Pushing the Limits of the Scanning Mechanism for Initiation of Translation.” *Gene*, vol. 299, no. 1-2, 2002, pp. 1–34., doi:10.1016/S0378-1119(02)01056-9.
- Krichevsky, A. M., et al. “Translational Control of Specific Genes during Differentiation of HL-60 Cells.” *J. Biol. Chem.*, vol. 274, no. 20, 1999, pp. 14295–14305., doi:10.1074/jbc.274.20.14295.
- Kumari, S., et al. “Position and Stability Are Determining Factors for Translation Repression by an RNA G-Quadruplex-Forming Sequence within the 5’ UTR of TheNRASProto-Oncogene†.” *Biochemistry*, vol. 47, no. 48, 2008, pp. 12664–12669., doi:10.1021/bi8010797.

- Lamphear, B. J., et al. "Mapping of Functional Domains in Eukaryotic Protein Synthesis Initiation Factor 4G (eIF4G) with Picornaviral Proteases." *Journal of Biological Chemistry*, vol. 270, no. 37, 1995, pp. 21975–21983., doi:10.1074/jbc.270.37.21975.
- Le, S., and J. V. Maizel. "A Common RNA Structural Motif Involved in the Internal Initiation of Translation of Cellular MRNAs." *Nucleic Acids Research*, vol. 25, no. 2, 1997, pp. 362–369., doi:10.1093/nar/25.2.362.
- Le, S.-Y., et al. "Discovering Well-Ordered Folding Patterns in Nucleotide Sequences." *Bioinformatics*, vol. 19, no. 3, 2003, pp. 354–361., doi:10.1093/bioinformatics/btf826.
- Leppek, K., et al. "Author Correction: Functional 5' UTR mRNA Structures in Eukaryotic Translation Regulation and How to Find Them." *Nature Reviews Molecular Cell Biology*, vol. 19, no. 10, 2018, pp. 673–673., doi:10.1038/s41580-018-0055-5.
- Levy, S., et al. "Oligopyrimidine Tract at the 5' End of Mammalian Ribosomal Protein MRNAs Is Required for Their Translational Control." *Proceedings of the National Academy of Sciences*, vol. 88, no. 8, 1991, pp. 3319–3323., doi:10.1073/pnas.88.8.3319.
- Lewis, S. M., and M. Holcik. "For IRES Trans-Acting Factors, It Is All about Location." *Oncogene*, vol. 27, no. 8, 2008, pp. 1033–1035., doi:10.1038/sj.onc.1210777.
- Li, H. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." *Bioinformatics*, vol. 27, 2011, pp. 2987–93., doi: 10.1093/bioinformatics/btr509.
- Li, H., et al. "1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools." *Bioinformatics*, vol. 25, 2009, pp. 2078–9., doi: 10.1093/bioinformatics/btp352d.
- Lin, Z., and W.-H. Li. "Evolution of 5' Untranslated Region Length and Gene Expression Reprogramming in Yeasts." *Molecular Biology and Evolution*, vol. 29, no. 1, 2011, pp. 81–89., doi:10.1093/molbev/msr143.
- Lincoln, A. J., et al. "Inhibition of CCAAT/Enhancer-Binding Protein Alpha and Beta Translation by Upstream Open Reading Frames." *J. Biol. Chem.*, vol. 273, no. 16, 1998, pp. 9552–9560., doi:10.1074/jbc.273.16.9552.
- Luukkonen, B. G., et al. "Efficiency of Reinitiation of Translation on Human Immunodeficiency Virus Type 1 MRNAs Is Determined by the Length of the Upstream Open Reading Frame and by Intercistronic Distance." *Journal of Virology*, vol. 69, no. 7, 1995, pp. 4086–4094., doi:10.1128/jvi.69.7.4086-4094.1995.
- MacFarlane, L.-A., and P. R. Murphy. "MicroRNA: Biogenesis, Function and Role in Cancer." *Current Genomics*, vol. 11, no. 7, 2010, pp. 537–561., doi:10.2174/138920210793175895.
- Martineau, Y., et al. "Internal Ribosome Entry Site Structural Motifs Conserved among Mammalian Fibroblast Growth Factor 1 Alternatively Spliced MRNAs." *Molecular and Cellular Biology*, vol. 24, no. 17, 2004, pp. 7622–7635., doi:10.1128/mcb.24.17.7622-7635.2004.
- Meijer, H. A., et al. "Translational Repression and eIF4A2 Activity Are Critical for MicroRNA-Mediated Gene Regulation." *Science*, vol. 340, no. 6128, 2013, pp. 82–85., doi:10.1126/science.1231197.

- Meyuhas, O., and E. Hornstein. “Translational Control of TOP MRNAs.” Cold Spring Harbor Laboratory Press, vol. 39, 2000, doi:<http://dx.doi.org/10.1101/0.671-693>.
- Meyuhas, O., et al. “Translational Control of Ribosomal Protein MRNAs in Eukaryotes.” Cold Spring Harbor Laboratory Press, 1996, doi:<http://dx.doi.org/10.1101/0.363-364>.
- Mignone, F., et al. “Untranslated Regions of MRNAs.” *Genome Biology*, vol. 3, no. 3, 2002, <http://genomebiology.com/2002/3/3/reviews/0004>.
- Mitchell, S. A., et al. “The Apaf-1 Internal Ribosome Entry Segment Attains the Correct Structural Conformation for Function via Interactions with PTB and Unr.” *Molecular Cell*, vol. 11, no. 3, 2003, pp. 757–771., doi:10.1016/s1097-2765(03)00093-5.
- Moore, K. S., and M. V. Lindern. “RNA Binding Proteins and Regulation of mRNA Translation in Erythropoiesis.” *Frontiers in Physiology*, vol. 9, no. 910, 2018, doi:10.3389/fphys.2018.00910.
- Morris, M. J., et al. “An RNA G-Quadruplex Is Essential for Cap-Independent Translation Initiation in Human VEGF IRES.” *Journal of the American Chemical Society*, vol. 132, no. 50, 2010, pp. 17831–17839., doi:10.1021/ja106287x.
- Nagalakshmi, U., et al. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.” *Science*, vol. 320, no. 5881, 2008, pp. 1344–1349., doi:10.1126/science.1158441.
- Oyama, M., et al. “Analysis of Small Human Proteins Reveals the Translation of Upstream Open Reading Frames of MRNAs.” *Genome Research*, vol. 14, no. 10b, 2004, pp. 2048–2052., doi:10.1101/gr.2384604.
- Paraskeva, E., et al. “Ribosomal Pausing and Scanning Arrest as Mechanisms of Translational Regulation from Cap-Distal Iron-Responsive Elements.” *Mol Cell Biol.*, vol. 19, no. 1, 1999, pp. 807–816., doi:10.1128/mcb.19.1.807.
- Paz, I., et al. “RBPmap: a Web Server for Mapping Binding Sites of RNA-Binding Proteins.” *Nucleic Acids Research*, vol. 42, no. W1, 2014, doi:10.1093/nar/gku406.
- Pelletier, J., and N. Sonenberg. “Insertion Mutagenesis to Increase Secondary Structure within the 5' Noncoding Region of a Eukaryotic mRNA Reduces Translational Efficiency.” *Cell*, vol. 40, no. 3, 1985, pp. 515–526., doi:10.1016/0092-8674(85)90200-4.
- Pelletier, J., et al. “Cap-Independent Translation of Poliovirus mRNA Is Conferred by Sequence Elements within the 5 Noncoding Region.” *Molecular and Cellular Biology*, vol. 8, no. 3, 1988, pp. 1103–1112., doi:10.1128/mcb.8.3.1103.
- Perteau, M., et al. “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.” *Nat. Protoc.*, vol. 11, 2016, pp. 1650–1669., doi:10.1038/nprot.2016.095.
- Pesole, G., et al. “Isochore Specificity of AUG Initiator Context of Human Genes.” *FEBS Letters*, vol. 464, no. 1-2, 1999, pp. 60–62., doi:10.1016/s0014-5793(99)01675-0.
- Pesole, G., et al. “Structural and Functional Features of Eukaryotic mRNA Untranslated Regions.” *Gene*, vol. 276, no. 1-2, 2001, pp. 73–81., doi:10.1016/s0378-1119(01)00674-6.
- Pesole, G., et al. “The Untranslated Regions of Eukaryotic MRNAs: Structure, Function, Evolution and Bioinformatic Tools for Their Analysis.” *Briefings in Bioinformatics*, vol. 1, no. 3, 2000, pp. 236–249., doi:10.1093/bib/1.3.236.

- Pickering, B. M., and A. E. Willis. “The Implications of Structured 5' Untranslated Regions on Translation and Disease.” *Seminars in Cell & Developmental Biology*, vol. 16, no. 1, 2005, pp. 39–47., doi:10.1016/j.semcdb.2004.11.006.
- Poulin, F., et al. “Mechanism of Translation Initiation in Eukaryotes.” *Madame Curie Bioscience Database [Internet].*, Landes Bioscience, 2000, www.ncbi.nlm.nih.gov/books/NBK6597/.
- Quesne, J. P., and C. Le, et al. “Derivation of a Structural Model for the c-Myc IRES, Edited by J. Karn.” *Journal of Molecular Biology*, vol. 310, no. 1, 2001, pp. 111–126., doi:10.1006/jmbi.2001.4745.
- Rajewsky, N. “MicroRNA Target Predictions in Animals.” *Nature Genetics*, vol. 38, no. S6, 2006, doi:10.1038/ng1798.
- Reynolds, K., et al. “Regulation of RAR Beta 2 mRNA Expression: Evidence for an Inhibitory Peptide Encoded in the 5-Untranslated Region.” *The Journal of Cell Biology*, vol. 134, no. 4, 1996, pp. 827–835., doi:10.1083/jcb.134.4.827.
- Ricci, E. P., et al. “MiRNA Repression of Translation in Vitro Takes Place during 43S Ribosomal Scanning.” *Nucleic Acids Research*, vol. 41, no. 1, 2013, pp. 586–598., doi:10.1093/nar/gks1076.
- Roberts, A., et al. “Identification of novel transcripts in annotated genomes using RNA-Seq.” *Bioinformatics*, vol. 27, 2011a, pp 2325–2329., doi: 10.1093/bioinformatics/btr355.
- Roberts, A., et al. “Improving RNA-Seq expression estimates by correcting for fragment bias.” *Genome Biol.*, vol. 12, no. R22, 2011b, doi: 10.1186/gb-2011-12-3-r22.
- Rubio, C. A., et al. “Transcriptome-Wide Characterization of the eIF4A Signature Highlights Plasticity in Translation Regulation.” *Genome Biology*, vol. 15, no. 10, 2014, pp. 1–19., doi:10.1186/s13059-014-0476-1.
- Schaeffer, C., et al. “The Fragile X Mental Retardation Protein Binds Specifically to Its mRNA via a Purine Quartet Motif.” *EMBO J*, vol. 20, no. 17, 2001, pp. 4803–4813., doi:10.1093/emboj/20.17.4803.
- Smit, A. F. A. “Interspersed Repeats and Other Mementos of Transposable Elements in Mammalian Genomes.” *Curr. Opin. Genet. Dev.*, vol. 9, no. 6, 1999, pp. 657–663., doi:10.1016/S0959-437X(99)00031-3.
- Somers, J., et al. “A Perspective on Mammalian Upstream Open Reading Frame Function.” *Int. J. Biochem. Cell Biol.*, vol. 45, no. 8, 2013, pp. 1690–1700., doi:10.1016/j.biocel.2013.04.020.
- Sonawane, A. R., et al. “Understanding Tissue-Specific Gene Regulation.” *Cell Reports*, vol. 21, no. 4, 2017, doi:10.1101/110601.
- Sonenberg, N., and A. G. Hinnebusch. “Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets.” *Cell*, vol. 136, no. 4, 2009, pp. 731–745., doi:https://doi.org/10.1016/j.cell.2009.01.042.
- Spriggs, K. A., et al. “Internal Ribosome Entry Segment-Mediated Translation during Apoptosis: the Role of IRES-Trans-Acting Factors.” *Cell Death & Differentiation*, vol. 12, no. 6, 2005, pp. 585–591., doi:10.1038/sj.cdd.4401642.
- Spriggs, K. A., et al. “Re-Programming of Translation Following Cell Stress Allows IRES-Mediated Translation to Predominate.” *Biology of the Cell*, vol. 100, no. 1, 2008, pp. 27–38., doi:10.1042/bc20070098.

- Steel, L. F., et al. “Elements in the Murine c-Mos Messenger RNA 5'-Untranslated Region Repress Translation of Downstream Coding Sequences.” *Cell Growth Diff.*, vol. 7, no. 10, 1996, pp. 1415–1424., <https://www.ncbi.nlm.nih.gov/pubmed/8891345>.
- Stein, I., et al. “Translation of Vascular Endothelial Growth Factor mRNA by Internal Ribosome Entry: Implications for Translation under Hypoxia.” *Molecular and Cellular Biology*, vol. 18, no. 6, 1998, pp. 3112–3119., doi:10.1128/mcb.18.6.3112.
- Stothard, P. “The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences.” *BioTechniques*, vol. 28, no. 6, 2000, pp. 1102–1104., doi:10.2144/00286ir01.
- Stripecke, R., et al. “Proteins Binding to 5 Untranslated Region Sites: a General Mechanism for Translational Regulation of MRNAs in Human and Yeast Cells.” *Molecular and Cellular Biology*, vol. 14, no. 9, 1994, pp. 5898–5909., doi:10.1128/mcb.14.9.5898.
- Svitkin, Y. V., et al. “The Requirement for Eukaryotic Initiation Factor 4A (eIF4A) in Translation Is in Direct Proportion to the Degree of mRNA 5' Secondary Structure.” *RNA*, vol. 7, no. 3, 2001, pp. 382–394., doi:10.1017/s135583820100108x.
- Taliaferro, J. M., et al. “RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation.” *Molecular Cell*, vol. 64, no. 2, 2016, pp. 294–306., doi:10.1016/j.molcel.2016.08.035.
- Thoreen, C. C. “The Molecular Basis of mTORC1-Regulated Translation.” *Biochemical Society Transactions*, vol. 45, no. 1, 2017, pp. 213–221., doi:10.1042/bst20160072.
- Thoreen, C. C., et al. “A Unifying Model for mTORC1-Mediated Regulation of mRNA Translation.” *Nature*, vol. 485, no. 7396, 2012, pp. 109–113., doi:10.1038/nature11083.
- Trapnell, C., et al. “Differential analysis of gene regulation at transcript resolution with RNA-seq.” *Nat. Biotechnol.*, vol. 31, 2013, pp. 46–53. doi: 10.1038/nbt.2450.
- Trapnell, C., et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.” *Nat. Biotechnol.*, vol. 28, 2010, pp. 511–515. doi:10.1038/nbt.1621.
- Trinklein, N. D., et al. “Identification and Functional Analysis of Human Transcriptional Promoters.” *Genome Research*, vol. 13, no. 2, 2003, pp. 308–312., doi:10.1101/gr.794803.
- Velden, A. W. V. D., and A. A. M. Thomas. “The Role of the 5' Untranslated Region of an mRNA in Translation Regulation during Development.” *The International Journal of Biochemistry & Cell Biology*, vol. 31, no. 1, 1999, pp. 87–106., doi:10.1016/s1357-2725(98)00134-4.
- Veselovska, L., et al. “Deep Sequencing and De Novo Assembly of the Mouse Oocyte Transcriptome Define the Contribution of Transcription to the DNA Methylation Landscape.” *Genome Biology*, vol. 16, no. 1, 2015, pp. 209., doi:10.1186/s13059-015-0769-z.
- Wang, C., et al. “Reprogramming of H3K9me3-Dependent Heterochromatin during Mammalian Embryo Development.” *Nature Cell Biology*, vol. 20, no. 5, 2018, pp. 620–631., doi:10.1038/s41556-018-0093-4.
- Wang, X.-Q., and J. A. Rothnagel. “5-Untranslated Regions with Multiple Upstream AUG Codons Can Support Low-Level Translation via Leaky Scanning and Reinitiation.”

- Nucleic Acids Research, vol. 32, no. 4, 2004, pp. 1382–1391.,
doi:10.1093/nar/gkh305.
- Weaver, R. F. *Molecular Biology*. 4th ed., McGraw-Hill, 2008.
- Wethmar, K., et al. “Upstream Open Reading Frames: Molecular Switches in (Patho)Physiology.” *BioEssays*, vol. 32, no. 10, 2010, pp. 885–893.,
doi:10.1002/bies.201000037.
- Wilkie, G. S., et al. “Regulation of mRNA Translation by 5'- and 3'-UTR-Binding Factors.” *Trends in Biochemical Sciences*, vol. 28, no. 4, 2003, pp. 182–188.,
doi:10.1016/s0968-0004(03)00051-3.
- Wolfe, A. L., et al. “RNA G-Quadruplexes Cause eIF4A-Dependent Oncogene Translation in Cancer.” *Nature*, vol. 513, no. 7516, 2014, pp. 65–70., doi:10.1038/nature13485.
- Xia, X., and M. Holcik. “Strong Eukaryotic IRESs Have Weak Secondary Structure.” *PLoS ONE*, vol. 4, no. 1, 2009, doi:10.1371/journal.pone.0004136.
- Xiong, W. “Regulation of CCAAT/Enhancer-Binding Protein-Beta Isoform Synthesis by Alternative Translational Initiation at Multiple AUG Start Sites.” *Nucleic Acids Research*, vol. 29, no. 14, 2001, pp. 3087–3098., doi:10.1093/nar/29.14.3087.
- Xue, S., et al. “RNA Regulons in Hox 5' UTRs Confer Ribosome Specificity to Gene Regulation.” *Nature*, vol. 517, no. 7532, 2015, pp. 33–38., doi:10.1038/nature14010.
- Yaman, I., et al. “The Zipper Model of Translational Control.” *Cell*, vol. 113, no. 4, 2003, pp. 519–531., doi:10.1016/s0092-8674(03)00345-3.
- Yamashita, R., et al. “Comprehensive Detection of Human Terminal Oligo-Pyrimidine (TOP) Genes and Analysis of Their Characteristics.” *Nucleic Acids Research*, vol. 36, no. 11, 2008, pp. 3707–3715., doi:10.1093/nar/gkn248.
- Zhang, J., et al. “Mice Deficient in Rbm38, a Target of the p53 Family, Are Susceptible to Accelerated Aging and Spontaneous Tumors.” *Proceedings of the National Academy of Sciences*, vol. 111, no. 52, 2014b, pp. 18637–18642.,
doi:10.1073/pnas.1415607112.
- Zhang, K., et al. “Identification and Functional Analysis of Long Non-Coding RNAs in Mouse Cleavage Stage Embryonic Development Based on Single Cell Transcriptome Data.” *BMC Genomics*, vol. 15, no. 1, 2014a, pp. 845., doi:10.1186/1471-2164-15-845.
- Zhang, T., et al. “Multiple Variable First Exons: A Mechanism for Cell- and Tissue-Specific Gene Regulation.” *Genome Research*, vol. 14, no. 1, 2003, pp. 79–89.,
doi:10.1101/gr.1225204.
- Zimmer, A., et al. “Tissue Specific Expression of the Retinoic Acid Receptor-Beta 2: Regulation by Short Open Reading Frames in the 5-Noncoding Region.” *The Journal of Cell Biology*, vol. 127, no. 4, 1994, pp. 1111–1119., doi:10.1083/jcb.127.4.1111.

11 APPENDICES

- **Appendix 1:** Python script used for GC/CpG content evaluation (developed by Nikolas Tolar)
- **Appendix 2:** Python Script for data extraction (developed by Nikolas Tolar)
- **Appendix 3:** R script to produce a histogram with a specified y-axis limit
- **Appendix 4:** R script to produce horizontal notched boxplots with means
- **Appendix 5:** R script to produce a histogram with specified x-axis boundaries
- **Appendix 6:** R script to produce a violin plot with median and quartile
- **Appendix 7:** R script to produce a notched boxplot with jitter plot, outliers displayed in red
- **Appendix 8:** R script to produce a scatterplot of means and medians for a selection of 12 datasets

Appendix 1. Python script used for GC/CpG content evaluation (developed by Nikolas Tolar).

Files name: GC_CpG_content_calculator.py

Language: Python

Description: A python script that serves for the evaluation of the GC and CpG content.

Input file: A .txt file with annotation of genomic coordinates and fasta files with mouse genomic sequences split into individual chromosomes.

Output file: A new .txt file with quantified GC and CpG values.

```
# GC/CpG content evaluation tool @Nikolas Tolar JCU 2019

# the program takes annotation file and raw data file as input and creates
# new file with previous annotation with GC and CpG values calculated for
# individual regions

###          --- Editable part ---

annotation_name = "utr.txt"
raw_data_name = "mus.fa"
output_name = "utr_GC.txt"

GC_switch = 1
CpG_switch = 0

'''
    Legend:

annotation_name - the name of the file with the annotation
raw_data_name - the files with data should start with the name/number of a
chromosome,
                followed by common name (no white space between the chromosome name
                and the name)
                - the raw data variable holds the common name part
output_name - the name of the output file

GC/CpG_switch - by setting the value to 0 you deactivate the switch
                - by setting the value to 1 you activate the switch

'''

###          --- Do not touch me part ---

output = open(output_name,'a')
anot_data = open(annotation_name,'r')

anot_line = anot_data.readline()
output.write(anot_line[:-1] + '\tGC_content' + '\tCpG_content \n')
print(anot_line[:-1] + '\tGC_content' + '\tCpG_content')
anot_line = anot_data.readline()
```

```

#main cycle goes thru the lines of annot and computes target values
while anot_line != '':
    anot_line_s = anot_line.split('\t')
    char_count = 0
    gc_count = 0
    cpG_count = 0

    data_file_name = anot_line_s[0] + raw_data_name
    data_file = open(data_file_name,'r')
    data_file.readline()
    data = data_file.read()
    data = data.replace('\n','')
    data = data[int(anot_line_s[1])-1:int(anot_line_s[2])]
    data_file.close()

    #controlling the output line format
    if anot_line[len(anot_line)-1] == '\n':
        output_line = anot_line[:-1]
    else:
        output_line = anot_line

    #GC switch part
    if GC_switch == 1:

        for char in data:
            if char in ['G','C','g','c','A','T','a','t','U','u','N','n']:
                char_count += 1
            if char in ['G','C','g','c']:
                gc_count += 1

        gc_cont = round(gc_count/len(data)*100,2)
        output_line = output_line + '\t' + str(gc_cont) + "%"

    #CpG switch part
    if CpG_switch == 1:

        for n in range(len(data)-1):
            if data[n] in ['C','c'] and data[n+1] in ['G','g']:
                cpG_count = cpG_count + 1

        cpG_density = round(cpG_count/(len(data)/2)*100,2)
        output_line = output_line + '\t' + str(cpG_density) + "%"

    output_line = output_line + '\n'

    #common part
    print(output_line,end='')
    output.write(output_line)
    anot_line = anot_data.readline()

anot_data.close()
output.close()

```

Appendix 2. Python Script for data extraction (developed by Nikolas Tolar).

Files name: non_gtf_search_tool_fixed_index.py

Language: Python

Description: A python script that serves for extraction of sequences that correspond to the extra upstream regions.

Input file: Raw mouse genome sequence split into individual chromosomes, an annotation of genome coordinates of regions of interest, and a list of names of regions of interest from the annotation for which we would like to obtain the sequence.

Output file: Sequences of regions of interest in fasta format.

```
### NON GTF

### Nikolas Tolar data extraction tool, at JCU 2019

# ----- Editable part -----

genes_name = 'mus.fa'
annotation_name = 'utr_forseq.txt'
output = open('utr_sequences','a')
query = open('utr_list.txt')
merge = 0

'''
    HINT: always edit strings in between the '' symbols

    genes_name = files containing raw DNA sequence - file names should follow the
                pattern Xiiii where X is number/letter of chromosome and
                iiii is the actual name that is shared with all other files.

                Variable genes_name holds the part iiii that is shared

    annotation = file containing names of probes and corresponding locations etc.

    output_file = name of the file the results will save into (if existing then
results will append, otherwise new file will be created)

    transcript_name = name of target transcript

    output_header = header of output file (FASTA format)

    merge = 1 means that the probes will be merged (connected) together
           0 means that the probes will be separated
'''

# ----- Do-not-touch-me part -----

def caller(value,neg,k=0):
    ret = ''

    if neg == 0:
```



```

        ret = ret + '_positive_strand_oc_' + str(k) + '\n'
    else:
        ret = ret + '_negative_strand_oc_' + str(k) + '\n'

    return ret

def translate_read_back(string):

    string_new = string[len(string)-1:0:-1] + string[0]

    string_new = string_new.replace('A','R')
    string_new = string_new.replace('T','A')
    string_new = string_new.replace('R','T')

    string_new = string_new.replace('C','F')
    string_new = string_new.replace('G','C')
    string_new = string_new.replace('F','G')

    return string_new

def data_extraction(text, gene_pool):

    start = int(text[2])
    stop = int(text[3])

    segment = gene_pool[start-1:stop]

    return segment

def insert_newlines(string, every=60):
    lines = []

    for i in range(0, len(string), every):
        lines.append(string[i:i+every])

    ret = '\n'.join(lines)
    return ret

def get_exons(genes_name, annotation_name, query, merge):

    transcript_name = query.readline().strip('\n')
    while transcript_name != '':

        annotation = open(annotation_name)
        neg = 0
        res_exons = ''
        res_list = []

        while True:

            text = annotation.readline()
            if text == '':
                break
            if transcript_name in text:
                text = text.split()
# accesing correct chromosome file
                genes = open(text[1]+genes_name)
                genes.readline()
                gene_pool = genes.read()
                gene_pool = ''.join(gene_pool.split())
                genes.close()

```

```

        if text[4] == '-':
            neg = 1

        if merge == 1:
            res_exons = res_exons + data_extraction(text, gene_pool)

        elif merge == 0:
            res_list.append(data_extraction(text, gene_pool))

    if merge == 1:

        if neg == 1:
            res_exons = translate_read_back(res_exons)

        res_exons = insert_newlines(res_exons)

        message = caller(merge, neg)

        print('>_' + transcript_name + message + res_exons + '\n')
        output.write('>_' + transcript_name + message + res_exons + '\n\n')

    else:

        for n in range(len(res_list)):
            message = caller(merge, neg, n)

            if neg == 1:
                res = '>_' + transcript_name + message +
                insert_newlines(translate_read_back(res_list[n]))

            else:
                res = '>_' + transcript_name + message +
                insert_newlines(res_list[n])

            print(res + '\n')
            output.write(res + '\n\n')

    annotation.close()
    transcript_name = query.readline().strip('\n')

get_exons(genes_name, annotation_name, query, merge)

output.close()
query.close()

```

Appendix 3. R script to produce a histogram with a specified y-axis limit.

Files name: basic_histogram.R

Language: R

Description: R script to produce a histogram with a specified y-axis limit. Script output can be found in Figure 6.

Input file: Selected data copied from a periodic table spreadsheet.

Output file: Basic histogram.

```
#basic histograms

combined<-read.delim('clipboard')# load data into the variable

# create a histogram, specifying the Dataset, adjusting the y-axis
dimensions that display frequency, the x-axis dimensions that display the
length in bp, and 'breaks' (2882 for oocyte datasets, 2794 for embryo
datasets)
y<-hist(combined$Length[combined$Type=='oc500'], ylim=c(0,11), breaks=2882,
xlim=c(100,500))
```

Appendix 4. R script to produce horizontal notched boxplots with means.

Files name: horizontal_notched_boxplot.R

Language: R

Description: R script to produce horizontal notched boxplots with means. Script output can be found in Figure 7.

Input file: Selected data copied from a periodic table spreadsheet.

Output file: A horizontal notched boxplot.

```
#Horizontal notched boxplot of oocyte- and 2C embryo elongations in
contrast to the official annotated UTRs presented with respective means

library(ggplot2) # import library

elongVSofficial<-read.delim('clipboard') # load data into the variable

# create a notched boxplot with means depicted as diamonds
pp<-ggplot(elongVSofficial, aes(x=Dataset, y=bp)) + geom_boxplot(notch=TRUE)
+ stat_summary(fun.y=mean, geom='point', shape=23, size=4)

pp+coord_flip() # flip coordinates
```

Appendix 5. R script to produce a histogram with specified x-axis boundaries.

Files name: hist_elongation.R

Language: R

Description: R script to produce a histogram with specified x-axis boundaries. Script output is found in Figure 8.

Input file: Selected data copied from a periodic table spreadsheet.

Output file: A histogram of oocyte- and 2C embryo elongations.

```
# Histogram of respective oocyte- and 2C embryo-specific 5' UTR genes'
elongations in 3 dimensions

library(ggplot2) # import library

elong<-read.delim('clipboard') # load data into the variable

# create a histogram of elongations, specify dataset of origin through
color
q<-ggplot(elong,aes(x=bp,color=Dataset))+geom_histogram(binwidth = 1)

# adjust the displayed x-axis, through adjustments we create 3 histograms
for 3 respective dimensions
q+xlim(-2500,2500)
```

Appendix 6. R script to produce a violin plot with quartile and median.

Files name: violin_plot.R

Language: R

Description: R script to produce a violin plot with median and quartile to display GC content. Script output found in Figure 9.

Input file: Selected data copied from a periodic table spreadsheet.

Output file: A violin plot with quartile and median.

```
# Violin plot with median and quartile to display GC content for respective
datasets

library(ggplot2) # import library

GCcontent<-read.delim('clipboard') # load data to plot into the variable

# create the violin plot of GC content for respective Dataset without
trimming the tails
p<-ggplot(GCcontent,aes(x=Dataset, y=GC))+geom_violin(trim=FALSE)

# add median and quartile to each plot
m<-p+geom_boxplot(width=0.1)
m # draw the plot
```

Appendix 7. R script to produce a notched boxplot with jitter plot.

Files name: notched_boxplot_jitter.R

Language: R

Description: R script to produce a notched boxplot with jitter plot, outliers displayed in red. Script output found in Figure 10.

Input file: Selected data copied from a periodic table spreadsheet.

Output file: A notched boxplot with jitter plot and outliers.

```
# Notched boxplot with jitter plot and outliers for plotting ORFs per 100
bp for 3 categories.

library(ggplot2) # import library

ORFno<-read.delim('clipboard') # load data into the variable

head(ORFno)

#create a notched boxplot with outliers depicted as red stars
v<-ggplot(ORFno, aes(x=Dataset, y=No_per_100bp))+
geom_boxplot(notch=TRUE,outlier.colour="red", outlier.shape=8,outlier.size=1)

#add jitter plot and adjust position
v+geom_jitter(shape=16, position=position_jitter(0.2))
```

Appendix 8. Scatterplot of mean and median.

Files name: scatter.R

Language: R

Description: R script to produce a scatterplot of mean and median for a selection of 12 datasets. Script output found in Figure 11.

Input file: Selected data copied from a periodic table spreadsheet.

Output file: A scatterplot for mean and median

```
# Scatterplot for mean and median per 100 bp for ORFs in 12 datasets
library(ggplot2) #import library
ORFs<-read.delim('clipboard') #load data into the variable ORFs
head(ORFs)

#create the graph, each point represents the mean for a given dataset, each
triangle represents the median of the respective dataset, grouping
according to the type of statistical evaluation
ggplot(ORFs, aes(x=DATASET, y=VALUE, shape=STATISTIC, color=STATISTIC,
fill=STATISTIC))+geom_point(size=5)
+geom_rug() #adds a rug
```