The Genomic Distribution of Musashi Binding Elements

by

Megan Rowe, B.S.

A Thesis

In

Biological Sciences

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

MASTER OF SCIENCE

Approved

Dr. Caleb Phillips, Ph.D.
Chair of the Committee

Dr. David Ray, Ph.D.

Dr. Natasja van Gestel, Ph.D.

Mark Sheridan, Ph.D.
Dean of the Graduate School

December, 2021

ACKNOWLEDGMENTS

There is an endless list of people that I want to acknowledge dearly for helping me to finish my degree. First, I would like to thank the Texas Tech College of Arts and Sciences and the Biology Department for the funding and supplies necessary for my project. Thank you to my committee, Dr. van Gestel and Dr. Ray, for graciously providing me with their time and guidance throughout my years here. I would also like to extend a massive thank you to my advisor, Dr. Caleb Phillips, for allowing me the opportunity to join his lab, for his support and belief in me, for his time and guidance, for bringing Patrick around the lab, and for invitations to the field trips that felt more like vacations. I want to thank Shannon Phillips, too, for being an absolute joy to be around and laugh with. To my past and present lab mates, Rachael Wiedmeier, Oscar Sandate, Craig Tipton, Preston McDonald, Matthew Fox, and Hendra Sihaloho: thank you so much for all the support, good memories, and countless hours spent together in the lab while we all pushed through our work.

I would also love to thank every last family member and friend of mine, though a thank you doesn't begin to cover my gratitude for each of you. Thank you to my mom and dad for your help, advice, belief in my ability, and for raising me to have that same belief in myself. Thank you to my sister Madison for your friendship, encouragement, coffee runs, and for taking on college in Lubbock with me. Thank you to my aunt Nicole for being my near sole inspiration to go into research in the first place. Thank you to my Nana for all your love and Sunday phone calls. Thank you to Leanne and Taylor for getting me through undergrad and grad school and for helping me to believe that I could take on this degree to begin with. Thank you to Alexis and Dan for your lifelong

friendships and for always being down to celebrate our milestones and achievements. Thank you to Kaki for all of the laughs and for commiserating in our grad school woes while being nearly a thousand miles apart. Thank you so much to Rachael for literally every last bit of our master's experience here at Tech, from Friday night scary movies and dinner, to dog walks at Buddy Holly, to working on exams in the lab until four in the morning, to camping in the Jemez, and so, so much more. Together we make one whole grad student. Thanks also to Roxie, Trixie, and Squeaky for being my honorary pets. And to Cosmo, my best boy and forever travel buddy: maybe I shouldn't have drug you to the other side of Texas for graduate school the day after adopting you, but I'll always be glad that I did. My life wouldn't be nearly the same without you, little shadow.

Finally, I want to send all my love and thanks to my grandpa, Vernon Rowe, and to my dog, Sophie, for supporting me through as much of my master's as you both could before you had to go. Your love and support will forever mean the world to me.

TABLE OF CONTENTS

ABSTRACT

The regulation of gene expression is fundamental to developmental and physiological processes. Such regulation occurs at all steps of the central dogma, and post-transcriptional regulation is thought to be a prominent point in the process. Musashi is an RNA-binding protein that is known to post-transcriptionally regulate genes involved in development and cell fate through interaction with the Musashi-binding element (MBE) in target mRNA 3' UTRs. Given the important ways that Musashi regulates cell fate in humans, this study aims to characterize MBE frequency and distribution and to highlight candidate Musashi targets. The neutral model predicts that, due to their small size (5-7 bp), MBEs commonly occur in the genome resulting in several potential regulatory targets. Moreover, past studies have shown that Musashi's effect on translation is proportional to the number of MBEs in a target mRNA 3' UTR. It was found that genome-wide MBEs occur more often than expected by chance in the 3' UTRs of protein-coding genes. The distribution of MBEs along the length of 3' UTRs of protein-coding genes was determined to be non-random, with MBEs more likely to appear near the 3' end of the 3' UTR, potentially relating to functional interactions between Musashi and poly(A)-binding protein. Finally, genes with counts of MBEs on the 3' UTR that were greater than random expectation had Gene Ontology terms associated with mRNA processing and post-transcriptional regulation of gene expression. This characterized pattern of MBEs in 3' UTR of protein-coding genes is thought to reflect selection for Musashi regulation and suggests a large number of previously unappreciated regulatory targets.

LIST OF FIGURES

CHAPTER 1

THE GENOMIC DISTRIBUTION OF MUSASHI BINDING ELEMENTS

1.1 Introduction

Developmental and physiological processes are governed by the regulation of gene expression (Peter & Davidson, 2015). Commonly known as the central dogma of molecular biology, gene expression is the flow of information from DNA to RNA to functional gene product (Schaefke, Sun, Li, Fang, & Chen, 2018). Each of the somatic cells of a species generally contain the same genetic information but vary in morphology and function due to differential gene expression. Related, differences in morphology and function among orthologous cells and structures across a phylogeny of species are also enabled through mutations creating differential regulation. Because the timing and location of functional gene products are critical in determining cell fate, a great deal of regulation occurs across all steps of the central dogma (Shparberg, Glover, & Morris, 2019).

Gene expression begins in the nucleus with the transcription of RNA from a DNA template. Nascent messenger RNA (mRNA) transcripts complement the DNA template and, barring regulatory inhibition, are then spliced and exported to ribosomes in the cytoplasm where they may be translated into protein. Despite this general proceeding, it is thought that a majority of regulatory control over mRNAs occurs in the nucleus (Glisovic, Bachorik, Yong, & Dreyfuss, 2008). A large component of this regulation occurs post-transcriptionally in ways that either promote or hinder eventual transport to ribosomes. For example, mRNAs can be targeted for degradation via nonsense mediated decay with the insertion of a premature termination sequence (Licatalosi, 2016; Lykke-

Andersen et al., 2014). mRNAs are also commonly regulated via the binding of small RNAs, which can act to silence genes by targeting them for degradation (Moazed, 2009). Further, RNA binding proteins (RBPs), which typically recognize and bind to specific sequences embedded in mRNAs, can regulate cellular localization of mRNAs as well as the promotion or repression of their translation (Licatalosi, 2016). Given the perception that post-transcriptional regulation of mRNAs is a common and diverse phenomenon, characterizing the details of these regulatory processes informs how dynamic regulation is consequential to organismal development, function, and evolution.

Much post-transcriptional regulation is known to occur via interaction at the 3' untranslated region (UTR) of mRNAs (Theil, Herzog, & Rajewsky, 2018). Likely reflecting the importance of 3' UTR mediated post-transcriptional regulation, the average 3' UTR in the human genome comprises about 36% of the total mRNA length (Zhao, Blagev, Pollack, & Erle, 2011). Specific sequences commonly distributed throughout 3' UTRs, generally known as regulatory motifs, are complementary to certain small RNAs or RBPs, which may bind such motifs and elicit their regulatory effect. RBPs in particular identify their cognate RNA motifs through affinity to the RBP's binding domains (Szostak & Gebauer, 2013).

One such RBP, Musashi, was first discovered in association with the development of mechanosensory bristles of *Drosophila melanogaster* (Nakamura, Okano, Blendy, & Montell, 1994). A mutation at the gene subsequently named Musashi resulted in a phenotypic variation in which two bristles grew out of a socket rather than one, and the likeness of the bristles to the famous two swords of the samurai Miyamoto Musashi inspired the name (Fox, Park, Koechlein, Kritzik, & Reya, 2015; Nakamura et al., 1994).

Following its initial discovery, Musashi has since been observed across a wide variety of both vertebrates and invertebrates and is characterized as evolutionarily conserved across phylogeny (Ohyama et al., 2012; Okano et al., 2005; Siddall, McLaughlin, Marriner, & Hime, 2006; Zearfoss et al., 2014). Within vertebrates, comparative analysis revealed the Musashi gene underwent a duplication event and now exists in a family of RBP orthologs that include Musashi 1 (*Msi1*) and Musashi 2 (*Msi2*) (Akindahunsi, Bandiera, & Manzini, 2005; Fox et al., 2015). Both Musashi proteins are thought to be pivotal to regulating differentiation and renewing somatic and germ stem cells, and though there are some instances of functional redundancy (Fox et al., 2015), the orthologs are commonly characterized by unique roles. For example, in developing neural tissue, the Musashi proteins are co-expressed, where MSI1 has a role in the differentiation of the central nervous system (Shibata et al., 2012) while MSI2 assists in the self-renewal of neural stem cells (Sakakibara et al., 2002). MSI1 and MSI2 are additionally expressed in spermatogonial stem cells. There, MSI1 proteins are typically found in mitotic gonocytes and MSI2 proteins in meiotic spermatocytes (Sutherland et al., 2014). MSI2 is also highly expressed in hematopoietic stem cells, ensuring the cells maintain an undifferentiated state (Kharas et al., 2010). Overexpression of MSI2 during hematopoiesis is thought to promote mitotic progression and is associated with poor prognosis in patients with myeloid leukemia (Ito et al., 2010; Kharas et al., 2010). Regarding fully differentiated cells, Musashi expression is greatly reduced (Fox et al., 2015), further suggesting its roles in regulation that primarily target developmental processes.

The primary mechanism by which Musashi is thought to regulate gene expression was initially discovered in the Numb/Notch signaling pathway. Here, Musashi represses

Numb expression through direct interaction, thus increasing the expression of Numb agonist, Notch, and promoting cell cycle progression (Imai et al., 2001). Additional lines of evidence suggest that Musashi may act as a translational repressor (Bennett et al., 2016; Imai et al., 2001; Okano et al., 2005) or enhancer (MacNicol, Cragle, & MacNicol, 2011; Phillips, Butler, Fondon, Mantilla-Meluk, & Baker, 2013), a difference that is likely cell type and pathway dependent. Although mechanistic details about how Musashi achieves regulation are incomplete, protein-protein interaction studies indicate Musashi's regulation is contingent upon other actors, such as poly(A)-binding protein (PABP; MacNicol et al., 2011) and translational initiation factors (Cragle et al., 2019; Kawahara et al., 2008).

Musashi proteins have two binding domains, known as RNA recognition motifs, with an affinity for a target sequence called the Musashi-binding element (MBE; Imai et al., 2001). MBEs consist of a 5-7 nucleotide sequence (G/A)U_{1-3}AGU (Imai et al., 2001; Iwaoka et al., 2017; Katz et al., 2014; Ohyama et al., 2012; Zearfoss et al., 2014). mRNA binding experiments in mouse keratinocytes indicate that Musashi has highest affinity for the UAG core of MBEs (Zearfoss et al., 2014), and an in vivo study found that local binding is enhanced if two or more UAGs are present within 50 nucleotides of known MSI2 binding sites (Bennett et al., 2016). In addition, in vitro experiments have shown that the number of MBEs present in 3' UTRs correspond to proportional changes in reporter protein expression (Phillips, Butler, Fondon, Mantilla-Meluk, & Baker, 2013). Current understanding intimates that differences in the number of MBEs present at a locus may correspondingly modify regulation of that locus with effects on inter-individual or inter-specific cellular outcomes.

Apart from Musashi targets identified through tissue-specific pull-down assays or through studies of specific interactions, the genomic distribution of MBEs is uncharacterized. Given the important ways Musashi can regulate cells, a comprehensive description of MBE distribution is expected to inform the potential breadth of Musashi regulatory control and its evolutionary modification. The current study provides this context by quantifying MBE distribution across protein coding genes in the human genome. Under a neutral model of MBE mutational emergence and decay, MBE genomic frequency would match the null expectation of a quantity that is based on MBE motif lengths and 3' UTR lengths. However, considering the current understanding that Musashi regulates specific processes regarding genes influencing development, cell division, and cell differentiation, it is predicted that the overall genomic frequency of MBEs in the 3' UTR of protein-coding genes will be less than expected by chance and thus compatible with a genomic average of purifying selection. Moreover, granting previous evidence that Musashi achieves regulation in part via interaction with PABP (Cragle et al., 2019), it is predicted that the distribution of MBEs on 3' UTRs is nonrandom with MBEs appearing near the 5' ends of UTRs more frequently than 3' ends due to the purging of deleterious motifs that do not benefit the interaction of Musashi and PABP. Though this trend is expected to appear generally, it is thought that a clear MBE distribution bias will be more apparent on larger 3' UTRs than on those small 3' UTRs with less available room for MBE accumulation. Finally, as Musashi has largely been reported regulating somatic and germ cell differentiation and expressed in select stem and cancer cells, it is predicted that genes with an outlying number of MBEs will be

disproportionately enriched for gene ontology (GO) terms related to development, cell division, and cell differentiation.

1.2 Methods and Analysis

1.2.1 Data Access

All genomic data for the GRCh38.p13 version of the *Homo sapiens* genome assembly was accessed from the Ensembl genomic database (Ensembl Genes 102; Howe et al., 2021) using the package biomaRt within RStudio (R version 4.0.3; Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009; RStudio Team, 2020). To access gene annotation data, a biomaRt query was built specifying Ensembl mart, species, filters, and attributes. getBM, one of the various functions within biomaRt, was used to retrieve attributes such as common gene name, Ensembl gene id, gene coordinates, 3' UTR coordinates, chromosome number, gene biotype, and GO information. getSequence was used to download all available gene sequences, including all introns and exons from the 5' end to the 3' end and excluding flanking sequences.

1.2.2 Data Preprocessing

The comprehensive set of genes and related details downloaded from Ensembl was filtered by gene biotype to only include those genes characterized as protein-coding. The resulting subset was filtered a second time to remove any genes that did not have associated 3' UTR location information. For each unique Ensembl gene ID, only the longest transcript was retained for analysis. The length of each gene and each gene's 3' UTR were both calculated by subtracting the end nucleotide position from the start

nucleotide position and adding one. Using a list of gene IDs from the finalized data set, DNA sequences were retrieved from Ensembl for 18,631 unique protein-coding transcripts from the *H. sapiens* reference genome and stored in a separate data frame. Annotated nucleotide location coordinates were then used to identify the 3' UTR sequence from every gene. The 3' UTR sequences were systematically examined for the MBE motif (G/A)$U_{1\text{-}3}$AGU, and all observed counts of the motif within a gene were stored in a data frame along with the associated Ensembl gene ID, the common gene name, the version of the identified motif, the nucleotide coordinates for each motif, and the length of the 3' UTR.

1.2.3 MBE Frequency

The random expectation of MBE frequency on a typical 3' UTR was calculated using the formula $\sum_{i=5}^{7} \frac{3'UTR\ length}{4^i}$, where *i* represents the various MBE kmer lengths: 5, 6, or 7 nucleotides. This calculation was conducted for each of the unique protein-coding transcripts to obtain the number of MBEs one would expect to find at random (rounded to the nearest whole number). The number of observed MBEs on each transcript's 3' UTR was determined by assigning a motif pattern in R, (G|A)(T|TT|TTT)AGT, to represent all three kmer variations, and then by passing that motif through every gene sequence with the stringr (Wickham, 2019) function str_locate_all to retrieve the counts and nucleotide locations of all present MBEs.

With the observed and expected MBE frequencies, a chi-squared goodness of fit test was conducted for each gene with the formula $\chi^2 = \frac{(O-E)^2}{E}$. The pchisq function in

the stats package (R Core Team, 2020) was used to calculate p-values from the chi-squared values for each gene. Input for the function was one degree of freedom and lower.tail = FALSE. Any chi-squared values of NaN were replaced with a value of 0 as they were associated with genes that had both zero expected MBEs and zero observed MBEs.

Genes were further organized for analysis by their chi-squared results. If the number of MBEs on a transcript's 3' UTR had more observed MBEs than expected and had a $p < 0.05$, it was categorized as "more than expected." If the number of MBEs on the 3' UTR had fewer observed MBEs than expected and had a $p < 0.05$, it was categorized as "less than expected." If the p-value was greater than 0.05, the gene was categorized as "not significant." Simple linear regression was used to investigate if the proportion of genes having more MBEs than expected by chance was dependent upon 3' UTR size class. For the chi-squared and the linear regression analyses, size classes with fewer than three genes were dropped from visualization.

1.2.4 MBE Positioning

Base pair location information for all genes and 3' UTRs was initially downloaded from Ensembl at a whole-genome scale: with the first nucleotide positioned at the beginning of chromosome 1 and the final nucleotide positioned at the end of chromosome 22. Using these genomic start and end base pair positions for both the genes and 3' UTRs, base pair locations were scaled down such that each gene began at nucleotide 1 and ended with the position of the last nucleotide of the 3' UTR. As noted

prior, all protein-coding gene sequences were inspected for any of the three MBE variations and their scaled base pair locations on the 3’ UTR were also recorded.

To better understand the physical distribution of MBEs on the 3’ UTR, the start and end base pair positions of all 3’ UTRs were transformed to a 0-100% scale. The first base pair coordinate of each MBE was also transformed to a percentage respective of the 3’ UTR the MBE appeared on. Every 3’ UTR was then divided into even 25% quartiles and its MBEs were assigned the quartile number (1-4) in which they fell. Quartile 1 begins on the 5’ end of the 3’ UTR, and quartile 4 ends on the 3’ end of the 3’ UTR. Genes were grouped into one of 329 size class bins (each bin sized to 100 base pairs) according to the length of their 3’ UTRs. All 329 size classes were examined independent of one another with the aim of keeping MBE distribution analysis informed by variation in 3’ UTR lengths.

For all genes in each 100 base pair bin, a binomial test was performed on each quartile. Passed through each iteration of the binom.test function from the stats package was the number of observed MBEs from all genes within the bin and quartile of interest, the total number of observed MBEs across the 3’ UTR of all genes within the size bin of interest, and the random expected frequency of MBEs in the quartile of interest (always $p = 0.25$).

1.2.5 MBEs and Gene Ontology

Genes with an outlying number of MBEs were determined by using the interquartile range (IQR) of the chi-squared values calculated prior to find the inner- and outer-upper fences. For the genes that had observed MBEs on the 3’ UTR when there

were no expected observations, the chi-squared value was infinity. These genes were separated from the principal analysis and subset to be investigated on their own. The summary function from the R base package was run on the chi-squared values from the remaining genes. The IQR was calculated using the first and third quartile values in the formula $IQR = Q_3 - Q_1$. The inner-upper fence was found by multiplying the IQR by 1.5 and adding the third quartile value. The outer-upper fence was found by multiplying the IQR by 3 and adding the third quartile value. Chi-squared values in the data set that were larger than either of the fence values were considered outliers. The outlying genes were recorded and further categorized by whether they had more observed MBEs than their expectation or less observed MBEs.

The groups of Ensembl gene names retained for GO analysis were those in the inner-upper fence with more MBEs than were expected, those in the inner-upper fence with less MBEs than were expected, those in the outer-upper fence with more MBEs than were expected, those with chi-squared values of infinity, those with at least one MBE on the 3' UTR, and those with zero MBEs on the 3' UTR. The online analysis tool PANTHER was used to detect the statistical overrepresentation of GO biological processes in all sets of outlying genes (Mi et al., 2021). Each of the six lists of Ensembl gene IDs were analyzed independently against the comprehensive background list of unique protein-coding genes obtained in the first steps of analysis. The statistical test option run was Fisher's Exact with the False Discovery Rate (FDR) correction. Returned from the analyses were the top GO biological processes ranked by fold enrichment and with FDRs < 0.05.

Within each significant gene ontology PANTHER returned the gene IDs from the candidate list assigned to that GO term. These genes were linked to PANTHER Protein Class functional classifications. The protein classes are distinct, biologically informative, and complement GO terms (Mi et al., 2021). All protein classes that fell under each child GO term were tallied to determine the primary biological functions of the genes within each candidate list. The protein classes within each GO term were also manually assessed for functional similarities amongst the classes and amongst the roles Musashi is already known to have a part in.

1.3 Results

1.3.1 MBE Frequency

A total of 12,057 from the 18,631 protein-coding genes (64.7%) had at least one MBE present on the 3' UTR. From the chi-squared values, 119 genes had significantly fewer MBEs than the expectation and 3,861 genes had significantly more MBEs than the expectation (with the threshold for significance being $p < 0.05$). The ratios of genes with more than expected, less than expected, and non-significant results were visualized at a finer 3' UTR size scale (100 base pair size classes) as seen in Figure 1.1. A general trend of an increasing percentage of genes with more than expected MBEs as 3' UTR size increased is evident in this figure. This relationship was formally evaluated using linear regression (Figure 1.2) through which it was found that 3' UTR size class significantly explained the proportion of genes with more MBE than expected by chance $R^2 = 0.599$, $p < 0.0001$).

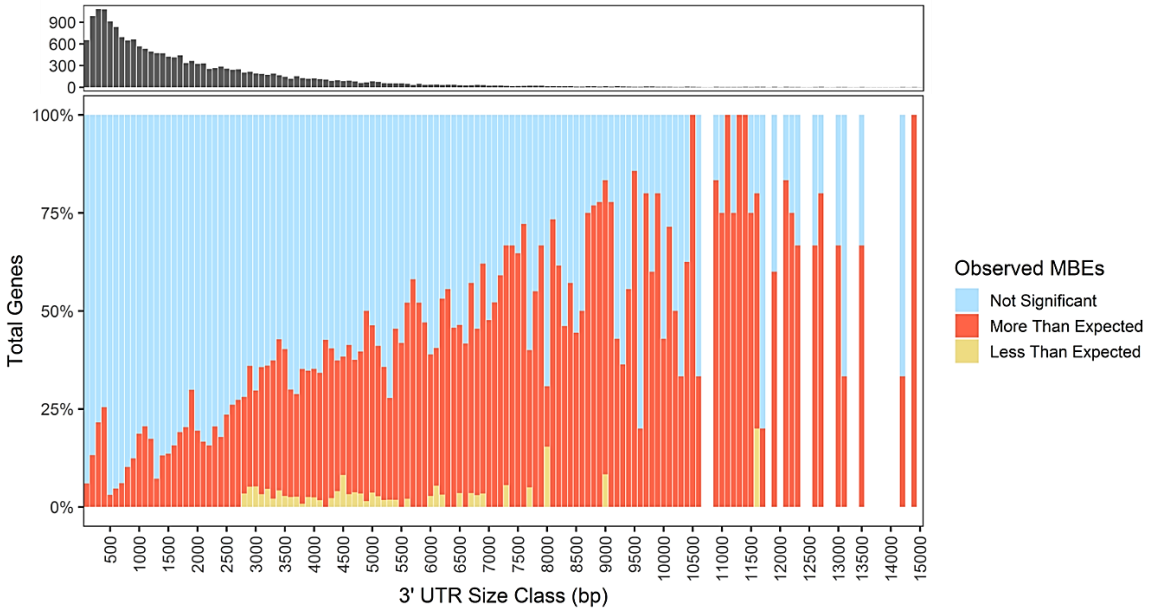

Figure 1.1 Trends in the proportion of genes per size class with more MBEs than expected on the 3' UTR, less MBEs than expected, and genes with an insignificant number of MBEs. Each size class bar was filled using p-value results from the chi-squared calculations. The bar graph at the top displays the number of genes found in each size class.

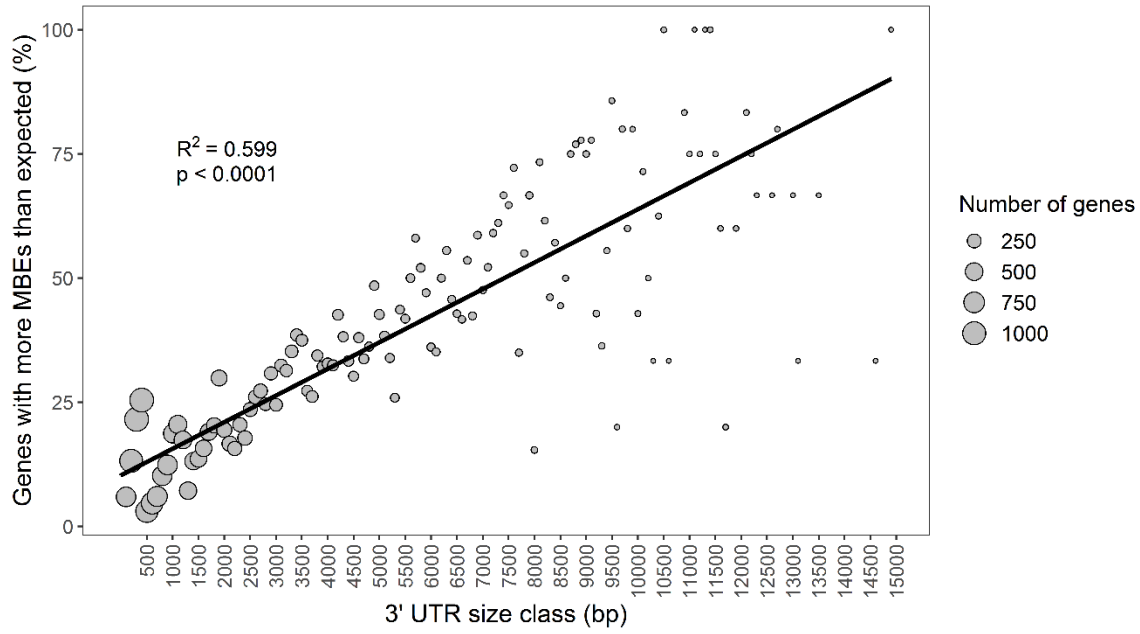

Figure 1.2 Regression analysis of 3' UTR length and the proportion of genes with more MBEs than the expectation across 18,626 samples. The number of genes in each size class is represented by the size of the point.

1.3.2 MBE Distribution

The distribution of MBE along the length of 3' UTRs was next evaluated with respect to random expectation. The proportion of total observed MBEs occurring in each 3' UTR quartile for each 3' UTR size class is expected be 25% if the location of MBE is random. Contrary to the random expectation, it was found that for 71 of the 329 size classes (Figure 1.3), quartile 1 had significantly less than 25% of observed MBE ($p < 0.05$), and 4 size classes with significantly more than 25% of observed MBE. Quartile 2 had four size classes that had MBE rates significantly lower than 25%, and 14 size classes that had MBE rates significantly above 25%. Quartile 3 had two size classes that had MBE rates significantly lower than 25%, and 17 size classes significantly above 25%. Quartile 4 had eight size classes that had MBE rates significantly lower than 25%, and 33 size classes significantly above 25%. Overall, MBEs were observed less often on the 5' end of the 3' UTR, though this trend did attenuate with increasing 3' UTR size.

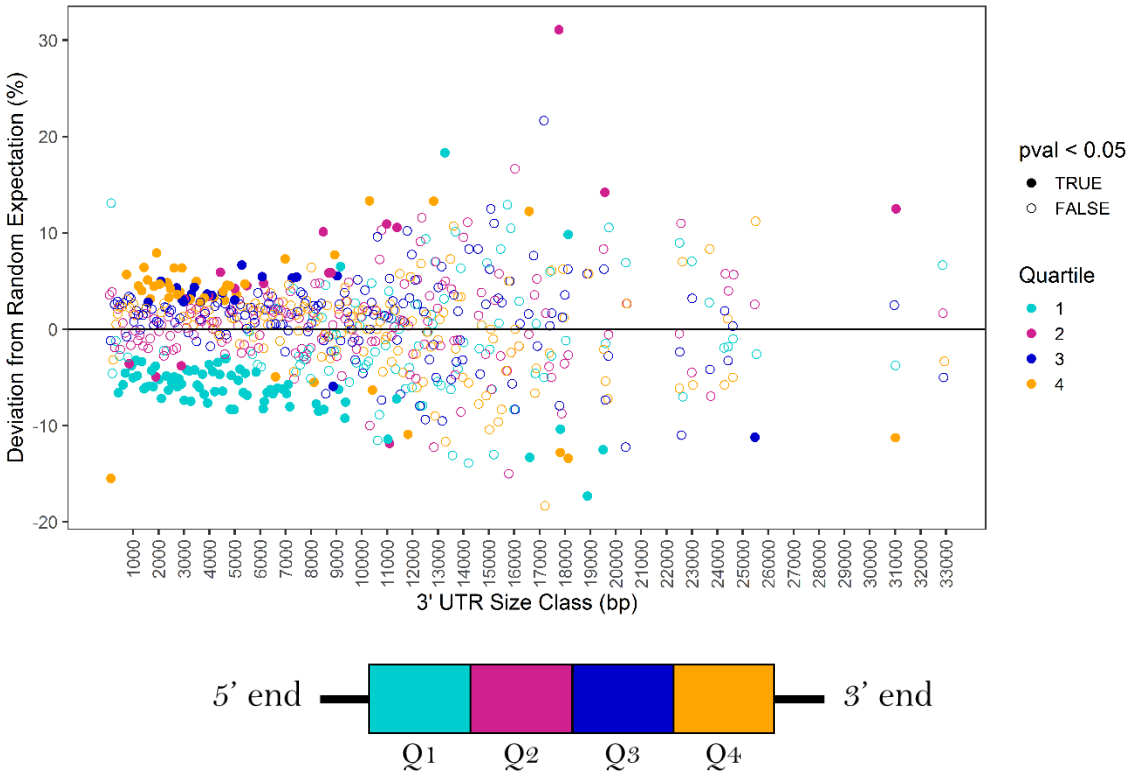

Figure 1.3 Distribution trends of MBEs on the 3' UTRs of genes across 178 size classes. As shown in the illustration at the bottom, quartile 1 begins on the 5' end of the 3' UTR and quartile 4 ends on the 3' end of the 3' UTR. Deviation from random expectation was calculated using the success values from binomial tests.

1.3.3 MBEs and Gene Ontology

After isolating the genes with outlying chi-squared values, there were 2,429 genes above the inner-upper fence with more MBEs than expected, 21 genes above the inner-upper fence with fewer MBEs than expected, 1801 genes above the outer-upper fence with more MBEs than expected, and 676 genes with chi-squared values of infinity. There were no genes above the outer-upper fence with fewer MBEs than expected. Further, the inner-upper fence had a comprehensive set of genes that encapsulated the genes from the outer-upper fence, so only those 2,429 genes were used to investigate GO term

enrichment. The 21 genes above the inner-upper fence with fewer MBEs than expected did not have any significantly enriched GO terms.

Enriched biological process GO terms from PANTHER that were associated with genes above the inner-upper fence, with more MBEs than expected, were related to homophilic cell adhesion via plasma membrane molecules (FDR = 0.019), mRNA processing (FDR = 0.01), RNA splicing (FDR = 0.043), post-transcriptional regulation of gene expression (FDR = 0.021), cellular localization (FDR = 0.019), macromolecule modification (FDR = 0.018), and cellular macromolecule metabolic process (FDR = 0.037). Homophilic cell adhesion via plasma membrane molecules (Figure 1.4) had the highest gene count in PANTHER Protein Class cadherin (n = 38). Within mRNA processing (Figure 1.5), the PANTHER Protein Classes (n = 17) with the highest gene counts were RNA splicing factor (n = 30) and RNA metabolism protein (n = 6). For post-transcriptional regulation of gene expression (Figure 1.6) the PANTHER Protein Classes (n = 28) with the highest gene counts were RNA metabolism protein (n = 10), translation initiation factor (n = 8), mRNA polyadenylation factor (n = 8), and RNA splicing factor (n = 7).

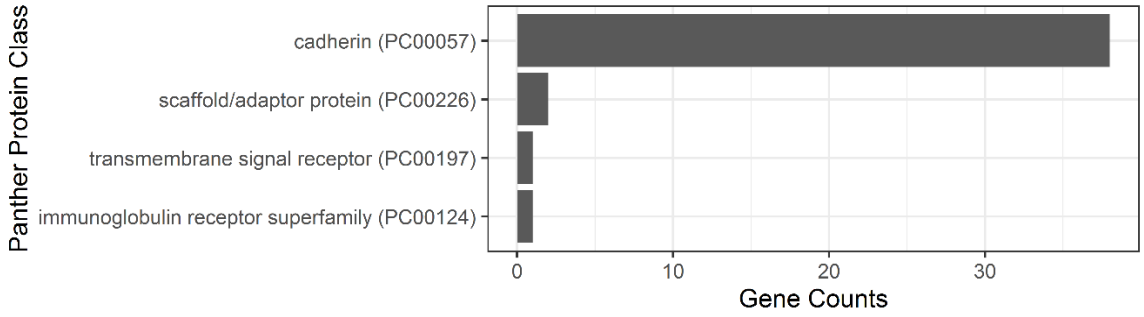

Figure 1.4 Counts of the number of genes labeled with various PANTHER Protein Classes associated with the homophilic cell adhesion via plasma membrane molecules GO term.

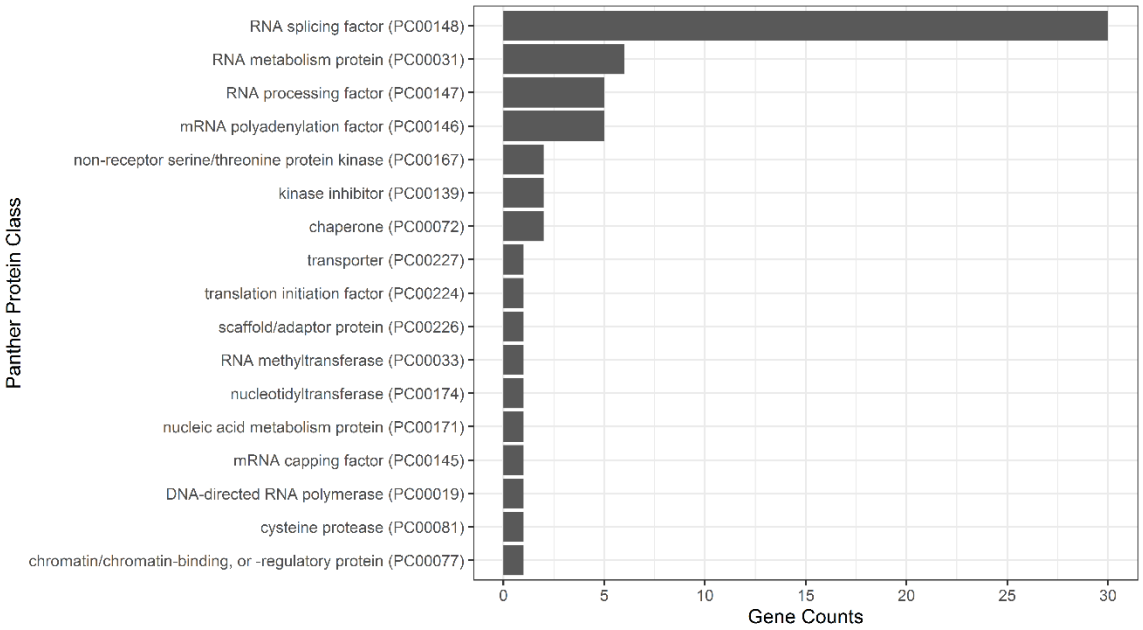

Figure 1.5 Counts of the number of genes labeled with various PANTHER Protein Classes associated with the mRNA processing GO term.

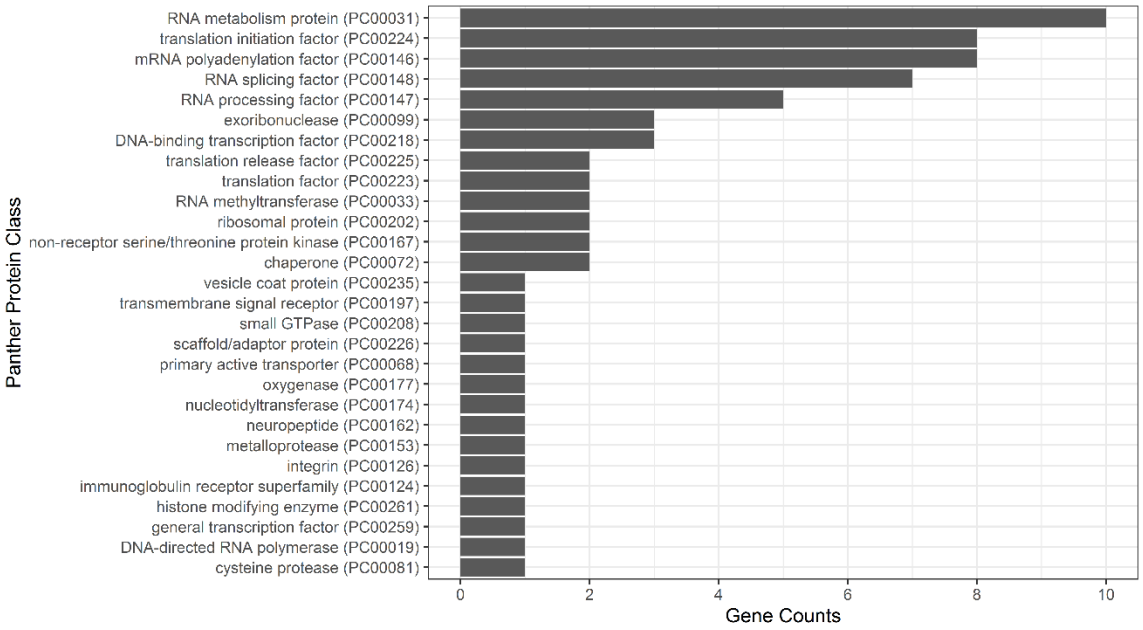

Figure 1.6 Counts of the number of genes labeled with various PANTHER Protein Classes associated with the post-transcriptional regulation of gene expression GO term.

1.4 Discussion

Post-transcriptional regulation is a powerful component in the control of cell fate, and the ways that RBPs can influence cells via post-transcriptional interaction is an active field of research (Kanellopoulou & Muljo, 2018). Based on the results presented herein, the RBP Musashi may play a regulatory role governing a more extensive array of genes than has been previously appreciated. Whereas the small size of the MBE provides the potential for it to frequently occur at random, MBEs were predicted to occur in human protein-coding 3' UTRs significantly less than random and confined to specific genes and pathways pertaining mostly to organismal development and stem cell differentiation, the cellular functions for which Musashi is most well associated. Even so, 64.7% of the 18,631 genes considered had at least one MBE present on the 3' UTR. Moreover, 3,861

of the 3,980 genes (97%) that had a significant number of binding elements contained more (rather than less) MBEs on the 3’ UTR than was expected by chance. These results suggest that it is either more likely that expected for MBE accumulation to be tolerated, that the number of Musashi targets is higher than previously appreciated, or potentially some combination of both scenarios.

The potential that the greater than expected frequency of MBEs can be explained by a combination of tolerance/neutrality and a large number of functional targets is supported by the observation that increasing 3’ UTR size predicted an increasing proportion of genes with significantly more MBE than expected by chance (Figure 1.1); e.g., as 3’ UTRs approached a length of 5,000 base pairs, the number of genes with significantly more MBEs than expected was often near or above 50%. This trend is interpreted such that with more genomic capacity on larger 3’ UTRs, there is more space for the accumulation of neutral/non-functional MBEs (presumably because they are distant from co-factor binding domains, see below) in addition to the functional subset that is selected for a given gene.

It was predicted that MBEs would preferentially occur toward the 5’ end of 3’ UTRs because a more 5’ position was hypothesized to increase likelihood of non-functionality, and because 3’ origination of new MBEs was expected to generally be selected against as they would alter established regulatory networks. Contrary to this expectation, analysis demonstrated that MBE distributions across the majority of 3’ UTR size classes occurred less often in the 5’ end of the 3’ UTR and more often in the 3’ end of the 3’ UTR (Figure 1.3). Based on the meaningful interactions that Musashi has with PABP (Cragle et al., 2019), it is possible that the considerable number of MBEs

positioned nearest the poly(A) tail are functional. Following the working hypothesis that proximity to 3' end of 3' UTRs corresponds to increasing functional potential, the MBEs identified on the 5' end of the 3' UTR, especially in smaller 3' UTRs, may not be far enough away from the poly(A) tail to be non-functional, but at the same time may be too far away to provide correct Musashi regulatory positioning. Consequently, as 5' positioned MBEs may not be capable of efficiently participating in protein-protein interaction at the poly(A) tail, selection against these MBE would explain their less than expected rate of occurrence nearer 5'. This model is supported by the observation that preferential accumulation of MBEs nearer 3' ends of 3' UTRs dissipated with increased 3' UTR length, suggesting the sufficient distance from the end of genes affords the presence of 5' positioned non-functional MBEs.

Expression of Musashi has been associated with processes of cell differentiation (Shibata et al., 2012), stem cell renewal (Sakakibara et al., 2002), mitotic gonocytes, and meiotic spermatocytes (Sutherland et al., 2014), all indicating the important role Musashi plays in organismal development and cell cycle progression. Based on this knowledge it was predicted that genes with the most MBEs in their 3' UTRs would fall into gene ontologies closely related to organismal development. When the outlier genes were analyzed for GO term enrichment, key terms included homophilic cell adhesion, mRNA processing, and post-transcriptional regulation of gene expression. Whereas cell adhesion and cell migration are fundamental to adhesion and cell migration during embryogenesis, within the homophilic cell adhesion GO term 'cadherin' (Figure 1.4) was the most common PANTHER Protein Class, a type of protein known for important developmental roles (Halbleib & Nelson, 2006). Genes categorized by mRNA processing, and post-

transcriptional regulation of gene expression were chiefly described by PANTHER Protein Classifications of RNA splicing factor (Figure 1.5) and RNA metabolism protein, translation initiation factor, mRNA polyadenylation factor, RNA splicing factor, and RNA processing factor (Figure 1.6), respectively. Because such biological processes are fundamental to development and normal physiology, the role of Musashi potentially regulating these targets in development specifically is ambiguous. However, these enrichments clarify that many of the genes in which MBEs are appearing more often than expected by chance are genes that function in RNA regulation, which is the known role of Musashi.

A great deal of genes are regulated post-transcriptionally, including those genes responsible for post-transcriptional regulation. For example, there are various instances in which RBP participation in autogenous regulation have been studied (Imig, Kanitz, & Gerber, 2012; Müller-McNicoll, Rossbach, Hui, & Medenbach, 2019), and indeed MSI2 3’ UTRs contain MBEs. As genes with an abundance of 3’ UTR MBEs are enriched for RNA regulatory ontologies, it appears that Musashi could be disproportionately targeting genes that are similarly involved in post-transcriptional regulation. This general phenomenon has been described in previous studies that consider the relationships between RBPs and mRNAs. Searching for the mRNA targets of various RBPs in *Saccharomyces cerevisiae* revealed that many RBPs bind to mRNAs that encode proteins with biological functionality like that of the RBP (Gerber, Herschlag, & Brown, 2004; Hogan, Riordan, Gerber, Herschlag, & Brown, 2008). Considering this, the broader classification of RNA regulatory proteins as the enriched list of candidate Musashi targets

supports the notion that post-transcriptional regulation is a complex and highly-networked process (Dassi, 2017) for which Musashi is a participant.

Overall, the results from this study indicate that Musashi may have a much larger influence on a wider array of target genes and biological pathways than was previously known. The comprehensive understanding from past studies have painted the Musashi RBP as a specialized regulator of a narrow subset of genes. However, considering that nearly 65% of protein-coding genes in the *H. sapiens* genome contain a minimum of one MBE per 3' UTR, and that 97% of those genes with a significant number of MBEs per 3' UTR had more MBEs than would be expected by chance, current results suggest that Musashi may play a regulatory role targeting a large fraction of the protein-coding genome. This assertion is also supported by other non-random observations including the preference for observed MBEs to occur towards the 3' end of 3' UTRs where they are anticipated to more likely be functional, as well as significant ontological enrichment among genes with the greatest excess of MBEs. Although previous work indicates that changing the number of MBE will change translational output, expression measurements and consideration of cellular context will be needed to understand how adding or deleting MBEs influences translation of specific targets.

CHAPTER 2

CONCLUSION

2.1 Conclusions and Future Directions

Past studies of Musashi have identified binding mechanisms, target gene interactions, and biological pathways that Musashi is involved in, though none have characterized the distribution patterns of the Musashi binding element before this study. There is still much to be discovered about the regulatory role that Musashi plays across the genome, but preliminary results suggest that the protein's reach may extend far beyond what is currently known.

It would be a valuable addition to expand the methods described here out in a phylogenetic context to determine the evolutionary conservation of Musashi binding element presence on the 3' UTR in other mammals. Comparisons across mammals could be based on calculating both the percentage of lineages with 3' UTR-embedded MBEs for each ortholog as well as the total number of MBEs in each lineage's ortholog. Non-random overlap in ontological enrichment across species could also be assessed by identifying those orthologs with an outlying number of MBEs.

Results produced from this study could also be used to further evaluate candidate Musashi targets by using tissue-specific RNA Immunoprecipitation Sequencing (RIP-seq). As RIP-seq maps the sites on a transcript to which RNA binding proteins are bound, this method could be applied in developing and adult tissue to place Musashi's regulation of thousands of potential targets into a cellular context.

A third experimental direction to take would involve using site-directed mutagenesis to measure the expression of a gene when the core sequence of 3' UTR

MBEs is altered. The resulting gene expression rates from this process could demonstrate the individual and/or additive effect of MBEs that are available for Musashi's binding.

Finally, a genome-wide comparison of MBE frequency on 3' UTRs, other regions of protein-coding genes (coding exons, introns, 5' UTRs), and non-coding genomic regions could reveal other ways in which Musashi imposes genomic control.

LITERATURE CITED

Akindahunsi, A. A., Bandiera, A., & Manzini, G. (2005). Vertebrate 2xRBD hnRNP proteins: A comparative analysis of genome, mRNA and protein sequences. *Computational Biology and Chemistry*, *29*(1), 13–23. https://doi.org/10.1016/j.compbiolchem.2004.11.002

Bennett, C. G., Riemondy, K., Chapnick, D. A., Bunker, E., Liu, X., Kuersten, S., & Yi, R. (2016). Genome-wide analysis of Musashi-2 targets reveals novel functions in governing epithelial cell migration. *Nucleic Acids Research*, *44*(8), 3788–3800. https://doi.org/10.1093/nar/gkw207

Cragle, C. E., MacNicol, M. C., Byrum, S. D., Hardy, L. L., Mackintosh, S. G., Richardson, W. A., … MacNicol, A. M. (2019). Musashi interaction with poly(A)-binding protein is required for activation of target mRNA translation. *Journal of Biological Chemistry*, *294*(28), 10969–10986. https://doi.org/10.1074/jbc.RA119.007220

Dassi, E. (2017). Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins. *Frontiers in Molecular Biosciences*, *0*(SEP), 67. https://doi.org/10.3389/FMOLB.2017.00067

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*, 3439–3440.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*, 1184–1191.

Fox, R. G., Park, F. D., Koechlein, C. S., Kritzik, M., & Reya, T. (2015). Musashi signaling in stem cells and cancer. *Annual Review of Cell and Developmental Biology*, *31*, 249–267. https://doi.org/10.1146/annurev-cellbio-100814-125446

Gerber, A. P., Herschlag, D., & Brown, P. O. (2004). Extensive Association of Functionally and Cytotopically Related mRNAs with Puf Family RNA-Binding Proteins in Yeast. *PLOS Biology*, *2*(3), e79. https://doi.org/10.1371/JOURNAL.PBIO.0020079

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008, June 18). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, Vol. 582, pp. 1977–1986. https://doi.org/10.1016/j.febslet.2008.03.004

Halbleib, J. M., & Nelson, W. J. (2006). Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. *Genes & Development*, *20*(23), 3199–3214. https://doi.org/10.1101/GAD.1486806

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., & Brown, P. O. (2008). Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *PLoS Biology*, *6*(10), 2297–2313. https://doi.org/10.1371/JOURNAL.PBIO.0060255

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/NAR/GKAA942

Imai, T., Tokunaga, A., Yoshida, T., Hashimoto, M., Mikoshiba, K., Weinmaster, G., … Okano, H. (2001). The Neural RNA-Binding Protein Musashi1 Translationally Regulates Mammalian numb Gene Expression by Interacting with Its mRNA. *Molecular and Cellular Biology*, *21*(12), 3888–3900. https://doi.org/10.1128/mcb.21.12.3888-3900.2001

Imig, J., Kanitz, A., & Gerber, A. P. (2012). RNA regulons and the RNA-protein interaction network. *Biomolecular Concepts*, *3*(5), 403–414. https://doi.org/10.1515/bmc-2012-0016

Ito, T., Kwon, H. Y., Zimdahl, B., Congdon, K. L., Blum, J., Lento, W. E., … Reya, T. (2010). Regulation of myeloid leukaemia by the cell-fate determinant Musashi. *Nature*, *466*(7307), 765–768. https://doi.org/10.1038/nature09171

Iwaoka, R., Nagata, T., Tsuda, K., Imai, T., Okano, H., Kobayashi, N., & Katahira, M. (2017). Structural Insight into the Recognition of r(UAG) by Musashi-1 RBD2, and Construction of a Model of Musashi-1 RBD1-2 Bound to the Minimum Target RNA. *Molecules (Basel, Switzerland)*, *22*(7). https://doi.org/10.3390/molecules22071207

Kanellopoulou, C., & Muljo, S. A. (2018). Posttranscriptional (Re)programming of Cell Fate: Examples in Stem Cells, Progenitor, and Differentiated Cells. *Frontiers in Immunology*, *0*(APR), 715. https://doi.org/10.3389/FIMMU.2018.00715

Katz, Y., Li, F., Lambert, N. J., Sokol, E. S., Tam, W. L., Cheng, A. W., … Burge, C. B. (2014). Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *ELife*. https://doi.org/10.7554/eLife.03915

Kawahara, H., Imai, T., Imataka, H., Tsujimoto, M., Matsumoto, K., & Okano, H. (2008). Neural RNA-binding protein Musashi1 inhibits translation initiation by competing with eIF4G for PABP. *The Journal of Cell Biology*, *181*(4), 639–653. https://doi.org/10.1083/jcb.200708004

Kharas, M. G., Lengner, C. J., Al-Shahrour, F., Bullinger, L., Ball, B., Zaidi, S., … Daley, G. Q. (2010). Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nature Medicine*, *16*(8), 903–908. https://doi.org/10.1038/nm.2187

Licatalosi, D. D. (2016). Roles of RNA-binding Proteins and Post-transcriptional Regulation in Driving Male Germ Cell Development in the Mouse. *Advances in Experimental Medicine and Biology*, *907*, 123–151. https://doi.org/10.1007/978-3-319-29073-7_6

Lykke-Andersen, S., Chen, Y., Ardal, B. R., Lilje, B., Waage, J., Sandelin, A., & Jensen, T. H. (2014). Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes and Development*, *28*(22), 2498–2517. https://doi.org/10.1101/gad.246538.114

MacNicol, M. C., Cragle, C. E., & MacNicol, A. M. (2011, January 1). Context-dependent regulation of Musashi-mediated mRNA translation and cell cycle regulation. *Cell Cycle*, Vol. 10, pp. 39–44. https://doi.org/10.4161/cc.10.1.14388

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, *49*(D1), D394–D403. https://doi.org/10.1093/NAR/GKAA1106

Moazed, D. (2009). Small RNAs in transcriptional gene silencing and genome defence. *Nature*, *457*(7228), 413–420. https://doi.org/10.1038/nature07756

Müller-McNicoll, M., Rossbach, O., Hui, J., & Medenbach, J. (2019). Auto-regulatory feedback by RNA-binding proteins. *Journal of Molecular Cell Biology*, *11*(10), 930–939. https://doi.org/10.1093/JMCB/MJZ043

Nakamura, M., Okano, H., Blendy, J. A., & Montell, C. (1994). Musashi, a neural RNA-binding protein required for drosophila adult external sensory organ development. *Neuron*. https://doi.org/10.1016/0896-6273(94)90460-X

Ohyama, T., Nagata, T., Tsuda, K., Kobayashi, N., Imai, T., Okano, H., … Katahira, M. (2012). Structure of Musashi1 in a complex with target RNA: the role of aromatic stacking interactions. *Nucleic Acids Research*, *40*(7), 3218–3231. https://doi.org/10.1093/nar/gkr1139

Okano, H., Kawahara, H., Toriya, M., Nakao, K., Shibata, S., & Imai, T. (2005). Function of RNA-binding protein Musashi-1 in stem cells. *Experimental Cell Research*. https://doi.org/10.1016/j.yexcr.2005.02.021

Peter, I. S., & Davidson, E. H. (2015). Genomic Control Process: Development and Evolution. In *Genomic Control Process: Development and Evolution*. https://doi.org/10.1016/C2012-0-02817-7

Phillips, C. D., Butler, B., Fondon, J. W., Mantilla-Meluk, H., & Baker, R. J. (2013). Contrasting Evolutionary Dynamics of the Developmental Regulator PAX9, among Bats, with Evidence for a Novel Post-Transcriptional Regulatory Mechanism. *PLoS ONE, 8*(2). https://doi.org/10.1371/journal.pone.0057649

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.r-project.org/

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. Retrieved from http://www.rstudio.com/

Sakakibara, S. ichi, Nakamura, Y., Yoshida, T., Shibata, S., Koike, M., Takano, H., … Okano, H. (2002). RNA-binding protein Musashi family: Roles for CNS stem cells and a subpopulation of ependymal cells revealed by targeted disruption and antisense ablation. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15194–15199. https://doi.org/10.1073/pnas.232087499

Schaefke, B., Sun, W., Li, Y. S., Fang, L., & Chen, W. (2018). The evolution of posttranscriptional regulation. *Wiley Interdisciplinary Reviews: RNA*. https://doi.org/10.1002/wrna.1485

Shibata, S., Umei, M., Kawahara, H., Yano, M., Makino, S., & Okano, H. (2012). Characterization of the RNA-binding protein Musashi1 in zebrafish. *Brain Research*, *1462*, 162–173. https://doi.org/10.1016/j.brainres.2012.01.068

Shparberg, R., Glover, H., & Morris, M. B. (2019). Modelling mammalian commitment to the neural lineage using embryos and embryonic stem cells. *Frontiers in Physiology*, Vol. 10. https://doi.org/10.3389/fphys.2019.00705

Siddall, N. A., McLaughlin, E. A., Marriner, N. L., & Hime, G. R. (2006). The RNA-binding protein Musashi is required intrinsically to maintain stem cell identity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(22), 8402–8407. https://doi.org/10.1073/pnas.0600906103

Sutherland, J. M., Fraser, B. A., Sobinoff, A. P., Pye, V. J., Davidson, T.-L., Siddall, N. A., … McLaughlin, E. A. (2014). Developmental Expression of Musashi-1 and Musashi-2 RNA-Binding Proteins During Spermatogenesis: Analysis of the Deleterious Effects of Dysregulated Expression1. *Biology of Reproduction*, *90*(5). https://doi.org/10.1095/biolreprod.113.115261

Szostak, E., & Gebauer, F. (2013). Translational control by 3'-UTR-binding proteins. *Briefings in Functional Genomics*, *12*(1), 58–65. https://doi.org/10.1093/bfgp/els056

Theil, K., Herzog, M., & Rajewsky, N. (2018). Post-transcriptional Regulation by 3′ UTRs Can Be Masked by Regulatory Elements in 5′ UTRs. *Cell Reports*, *22*(12), 3217–3226. https://doi.org/10.1016/j.celrep.2018.02.094

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. *R Package Version 1.4.0*. Retrieved from https://cran.r-project.org/package=stringr

Zearfoss, N. R., Deveau, L. M., Clingman, C. C., Schmidt, E., Johnson, E. S., Massi, F., & Ryder, S. P. (2014). A conserved three-nucleotide core motif defines musashi RNA binding specificity. *Journal of Biological Chemistry*, *289*(51), 35530–35541. https://doi.org/10.1074/jbc.M114.597112

Zhao, W., Blagev, D., Pollack, J. L., & Erle, D. J. (2011). Toward a systematic understanding of mRNA 3′ untranslated regions. *Proceedings of the American Thoracic Society*, *8*(2), 163–166. https://doi.org/10.1513/pats.201007-054MS