Transcriptome-wide analysis of microRNA-mRNA correlations in unperturbed tissue transcriptomes identifies microRNA targeting determinants.

Juan Manuel Trinidad Barnech[123], Rafael Sebastian Fort[13], Guillermo Trinidad Barnech[4], Beatriz Garat[1], Maria Ana Duhagon[12*]

1 Laboratorio de Interacciones Moleculares, Facultad de Ciencias, UDELAR, Iguá 4225, 11400 Montevideo, Uruguay
2 Departamento de Genética, Facultad de Medicina, Universidad de la República, Montevideo, C.P. 11800, Uruguay
3 Departamento de Genómica, Instituto de Investigaciones Biológicas Clemente Estable, Av. Italia 3318, Montevideo CP 11600, Uruguay
4 Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, CP 11300, Uruguay

* Corresponding author: mduhagon@fcien.edu.uy/mduhagon@fmed.edu.uy (M.A.D.); Tel.: +598-2-525-8618 (ext. 7237), +598-2-924-3414 (ext. 3468)

Abstract

MicroRNAs are small RNAs that regulate gene expression through complementary base pairing with their target mRNAs. Given the small size of the pairing region and the large number of mRNAs that each microRNA can control, the identification of biologically relevant targets is difficult. Since current knowledge of target recognition and repression has mainly relied on *in vitro* studies, we sought to determine if the interrogation of gene expression data of unperturbed tissues could yield new insight into these processes. The transcriptome-wide repression at the microRNA-mRNA canonical interaction sites (seed and 3′-supplementary region, identified by sole base complementarity) was calculated as a normalized Spearman correlation (Z-score) between the abundance of the transcripts in the PRAD-TCGA tissues (RNA-seq and small RNA-seq data of 546 samples). Using the repression values obtained we confirmed established properties or microRNA targeting efficacy, such as the preference for gene regions (3'UTR>CDS>5'UTR), the proportionality between repression and seed length (6mer<7mer<8mer) and the contribution to the repression exerted by the supplementary pairing at 13-16nt of the microRNA. Our results suggest that the 7mer-m8 seed could be more repressive than the 7mer-A1, while they have similar efficacy when they interact using the 3'-supplementary pairing. Strikingly, the 6mer+suppl sites yielded normalized Z-score of repression similar to the sole 7mer-m8 or 7mer-A1 seeds, which raise awareness of its potential biological relevance. We then used the approach to further characterize the 3´-supplementary pairing, using 39 microRNAs that hold repressive 3'-supplementary interactions. The analysis of the bridge between seed and 3'-supplementary pairing site confirmed the optimum +1 offset previously evidenced, but higher offsets appear to hold similar repressive strength. In addition, they show a low GC content at position 13-16, and base preferences that allow the selection of a candidate sequence motif. Overall, our study demonstrates that transcriptome-wide analysis of microRNA-mRNA correlations in large, matched RNA-seq and small-RNA-seq data has the power to uncover hints of microRNA targeting determinants operating in the *in vivo* unperturbed set. Finally, we made available a bioinformatic tool to analyze microRNA-target mRNA interactions using our approach.

Keywords: microRNA, seed, 3´-supplementary, transcriptome, RNA-seq, small RNA-seq, TCGA, offset, bridge, loop.

Introduction

MicroRNAs are small, single-stranded RNAs that shape the expression of messenger RNAs (mRNAs) and are critical for physiological homeostasis and development in metazoans. It was shown that microRNAs influence all developmental processes and diseases, including cancer (Bartel, 2018). In microRNA targeting, the seed region (nt 2–7 or 2–8, from the microRNA 5' end) is the primary determinant for targeting efficacy and specificity, which interact with 3' untranslated region (UTR) (Bartel, 2009) and coding sequence (CDS) of target mRNAs (Grimson et al., 2007; Hafner et al., 2010; Helwak et al., 2013; Moore et al., 2015; Schnall-Levin et al., 2010). Since the seed region is only composed of 6–7 contiguous nucleotides, such sites often appear in the transcriptome but are not functionally associated to microRNA mediated repression (Baek et al., 2008; Selbach et al., 2008; Yue et al., 2009). Despite several years of effort to better understand the underlying factors and contexts involved in efficient repression of mammalian mRNAs by microRNAs, it is still challenging to identify functionally relevant interactions (Bartel, 2018).

It has been demonstrated that nucleotides outside of the seed of the microRNA can contribute to target recognition (Broughton et al., 2016; Grimson et al., 2007; Helwak et al., 2013; Moore et al., 2015; Salomon et al., 2015; Wahlquist et al., 2014; Wee et al., 2012). Pairing to the microRNA´s 3' region, particularly involving microRNA nucleotides 13–16, can supplement pairing to the seed region (Bartel, 2009). However, this 3'-supplementary pairing has shown little influence on site affinity and efficacy (Grimson et al., 2007; Salomon et al., 2015; Wee et al., 2012), and only about 5% of the seed-matched regulatory sites undergoing purifying selection appear to involve it (Friedman et al., 2009).

Recent work has resolved the crystallography of human Ago2 with microRNA-122, uncovering microRNA-mRNA interactions through the 2-8 seed and the 3'-supplementary region (nucleotides 13-16) of the microRNA (Schirle et al., 2014; Sheu-Gruttadauria et al., 2019; Sheu-Gruttadauria et al., 2019). Contrary to previous reports, they demonstrate that optimal supplementary interactions can increase target affinity. They also show that the seed and the 3´-supplementary region can be bridged by an unstructured target loop in the mRNA of 1–15 nucleotides. Even when central pairing is available, Ago2 molecules maintain a stable conformation that avoids central pairing (Sheu-Gruttadauria et al., 2019). It was suggested that a possible reason for differences in the repression level between the global analyzes and those of specific microRNAs may be due to a “dilution” effect (Xiao & Macrae, 2020). Examining targets in batch may have masked the small number of targets undergoing strong supplementary repression among a larger number that are predicted to make supplementary interactions but weakly (Xiao & Macrae, 2020).

Because members of a microRNA family typically have identical 5' sequences but divergent 3' sequence regions they provide an ideal source for assessing contributions of the 3'-supplementary region to specificity and function (Broughton et al., 2016; Moore et al., 2015; Xiao & Macrae, 2020). Interestingly, experimental approaches show that the notion that most individual microRNA families have only subtle or redundant biological processes seems to apply uniquely to nematodes, in that a different conclusion has emerged for both flies and mammals (Bartel, 2018). Single-cell reporters on human cell lines have demonstrated that

some seed sites with 3'-supplementary pairing can direct microRNA family-specific regulation (Moore et al., 2015). Many reports illustrate the potential for microRNA 3′-pairing to enable high-affinity binding and demonstrate that the benefit of this pairing varies, depending on the microRNA sequence and the 3′-pairing architecture. Understanding how these features modulate the use of 3′-pairing will be possible only after acquiring many more measurements with multiple microRNA sequences (Mcgeary et al., 2021).

In the present study, we evaluate the contribution of the microRNA canonical seeds and 3'-supplementary region to target mRNA repression through the analyzing of all the evolutionary conserved microRNAs and all mRNA transcripts expressed in prostate tissue samples of the PRAD-TCGA cohort (Agarwal et al., 2015). MicroRNA-mRNA interaction strength was evaluated by the Spearman correlation coefficient of the expression of both partners obtained from the RNA-seq transcriptomes of the PRAD-TCGA. Overall, our investigation suggests that the comparative analysis of RNA-seq transcriptomes of mRNAs and small RNAs of a large number of unperturbed tissues has the power to analyze microRNA-mRNA interactions allowing the confirmation of previous rules and the discovery of potential novel variables of the interaction.

Methods and Data Sets

microRNAs sequences and conservation scores were obtained from TargetScan 7.2 database (Agarwal et al., 2015). For this analysis, we chose microRNAs with the highest "Family Conservation score" (score 2 defined by TargetScan) expressed in the TCGA-PRAD small RNA data (firebrowser.org), hereafter denominated "conserved microRNAs" (Supplementary Table 1) . The mRNA fasta sequences of all *Homo sapiens* protein coding genes were obtained using BioMart (Ensembl GRCh38.13). In the analysis we selected microRNAs and mRNAs that are detected in at least 80% of the samples (Supplementary Figure 1, Supplementary Table 1, see complete pipeline analysis in Figure 1), representing 143 microRNAs and 16786 mRNAs. Using an *in-house* script written in Python programming language (available in github), we identified all the sequences complementary to all the canonical microRNA seeds of the 143 microRNAs as defined in TargetScan. We use the TargetScan nomenclature for microRNA-mRNA interaction types, which comprise 6mer, 7mer-A1, 7mer-m8 and 8mer microRNA seeds (Supplementary Table 2 and Supplementary Figure 2). The 3'-supplementary region of a mRNA-microRNA interaction refers to four mRNA nucleotides complementary to microRNA nucleotides at positions 13-16. These two sequence modules (seed and 3'-supplementary region, see Figure 2) can be separated by a variable length nucleotide bridge that we set at a maximum length of 15 nucleotides based on previous functional studies (Sheu-Gruttadauria et al., 2019).
Using a Python regular expression (available in github), we found the mRNA sites for the 8 types of microRNA interactions indicated above. For each gene, we chose the splice variant with the highest number of putative microRNA interactions. In addition, since individual mRNAs can bear more than one complementary site for a given microRNA, in order to assess the repressive contribution of the different seed types, for each of the microRNA we selected the mRNAs that presented only one predicted microRNA-mRNA interaction, pondering the CDS, 3'UTR and 5'UTR independently. As a result, the computed mRNAs only have a single microRNA site at the region considered for every given microRNA analyzed, thus avoiding the

co-occurrence of the sites. Of the 16786 mRNAs and 143 microRNAs analyzed, 1,116,787 unique mRNA sites were identified for 193203, 443,323 and 480,261 at 5'UTR, CDS and 3'UTR respectively (Supplementary Table 2, Supplementary Figure 2). Normalized Spearman correlations coefficient (Z-score centered on microRNA) for all microRNA-mRNA pairs identified were calculated with the Python´s Pandas library (McKinney Wes, 2010) using TCGA RNA-seq (mRNAs) and small RNAseq (microRNAs) normalized data (546 tissue samples) obtained from FireBrowse.

Results

Identification of microRNA-mRNA interactions sites in the PRAD-TCGA transcriptome based on nucleotide base complementarity.

Given the tissue specificity of microRNA expression we performed this study within a single tissue type, and we chose the PRAD cohort because it is large and extensively studied for microRNAs (Abeshouse et al., 2015; Lin et al., 2021; Wei et al., 2020; Yang et al., 2019; Ye et al., 2018). We investigated the expression of the conserved human microRNAs (221) and all the human transcripts for protein coding genes annotated in Ensemble (20440), studied in the 545 samples of the PRAD-TCGA cohort (20531) (see pipeline analysis in Figure 1). In order to increase the signal/noise ratio we filtered out poorly expressed transcripts, arbitrarily defined as those undetected in ≥ 20% of the samples, which narrowed down the list of microRNAs and mRNA to 143 and 16786 respectively (Supplementary Table 1 and Supplementary Figure 1). Aiming to compare the repressive power of different microRNA pairing sites, we studied 6mer, 7mer-A1, 7mer-m8 and 8mer microRNA seeds with or without the 3'-supplementary region of 4 nucleotides downstream the seed (hereafter denominated 6mer+suppl, 7mer-A1+suppl, 7mer-m8+suppl and 8mer+suppl) (Figure 2A). We determined all the mRNA sites having perfect complementarity to these 8 canonical microRNA pairing sites, presuming they represent putative microRNA binding sites. This analysis was performed independently for each of the three regions of the mRNAs (5'UTR, CDS, and 3'UTR). The examination of seed sequence complementarity between the 143 microRNA and the 16786 mRNA on the PRAD/TCGA cohort identified 3,065,731 putative microRNA interaction sites on the mRNAs (285,237 at 5'UTR, 1,226,568 at CDS, 1,553,926 at 3'UTR). Since each mRNAs can bear more than one putative microRNA site, in subsequent calculations we only included the interactions involving mRNAs with a single site (of any type) per microRNA, considering each of the three mRNA regions independently. We iterated this analysis for each of the 143 selected microRNAs, obtaining a total of 1,116,787 sites (193,203 at 5'UTR, 443,323 at CDS, 480,261 at 3'UTR), corresponding to an average microRNA interaction of 1351±621 at the 5'UTR, 3100±748 CDS and 3358±602 at the 3'UTR) (Supplementary Table 2). As expected, the number of predicted interactions identified decreases with the size of the seed and the incorporation of the 3'-supplementary region for the 8 interaction types in the three regions studied (Supplementary Figure 2).

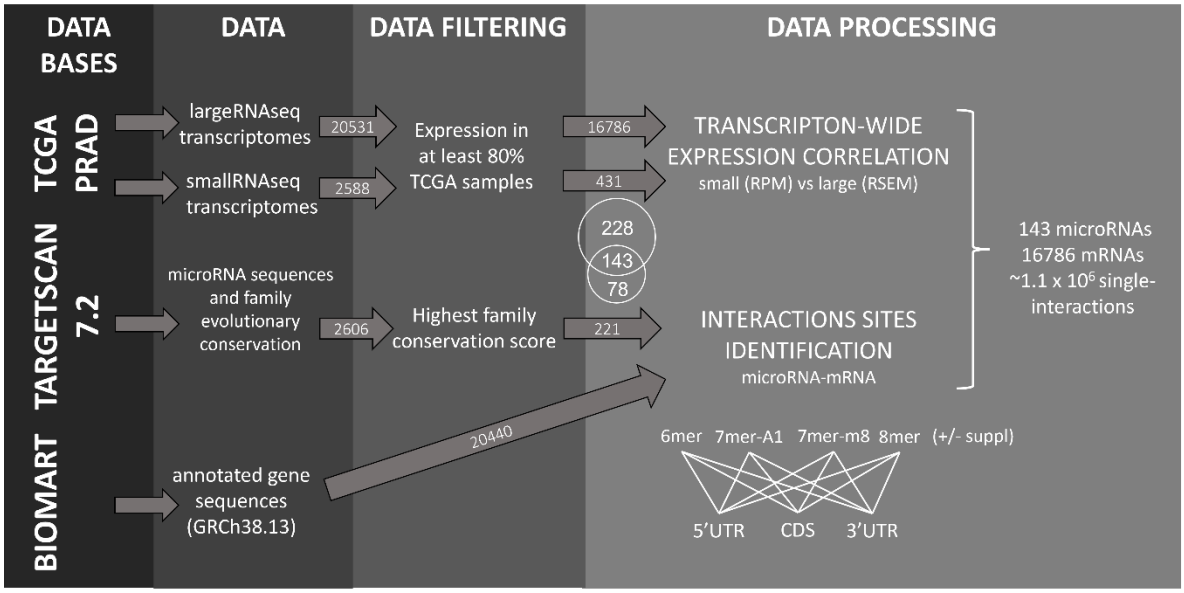

Figure. 1 Data analysis pipeline

Inference of microRNA activity from microRNA-mRNA expression correlation in the PRAD-TCGA transcriptomes.

The availability of matched RNAseq and small-RNAseq and transcriptomes of hundreds of tissues as a result of global cancer initiatives, allow the high-throughput transcriptome-wide assessment of microRNA-mRNA correlations. Our approach evaluates the strength of microRNA-mRNA interaction by the Spearman correlation coefficient of the expression of both molecules calculated from the RNAseq data of the 546 patient tissues (normal and tumor) of the PRAD-TCGA cohort (Figure 1). We calculated the correlation of the mRNA transcripts and the conserved microRNAs expressed in at least 80% of the samples.
Since the microRNAs can exert different effects on the expression of their target genes, we included all the correlations (positive and negative) between the mRNA-microRNA pairs identified. Nevertheless, in the absence of consensus about the threshold for biological significant correlation expected in tissues for a direct microRNA repression we included all the correlations.

Validation of microRNA repressive mechanisms using microRNA activity inferred from microRNA-mRNA expression correlations in PRAD-TCGA.

We initially sought to evaluate if the sole correlation of microRNA-mRNA pairs abundance in the tissues validates established microRNA repression rules. The computation of a Z-score of microRNA-mRNA correlation for each type of seed, showed that repression levels increase with the length of the seed, i.e. the number of nucleotide bases involved in the interaction (6mer< 7mer-m8 < 7mer-A1 < 8mer) (Figure 2B and Supplementary Figure 3) (Table 1 shows statistically significant differences FDR<0.05 and Supplementary Table 3 shows all the comparisons), confirming one previously established rule of microRNA targeting (Friedman et al., 2009; Grimson et al., 2007; Hafner et al., 2010). The statistical differences are more significant for intra 3'UTR (average FDR= E-25) followed by intra CDS (average FDR= E-16) comparison. Meanwhile, at the 5'UTR only 6mer/7mer-A1 and 7mer-A1/7mer-m8 differences are significant (FDR 1.14 E-02 and 2.93 E-03 respectively). As expected, although the

significance of the differences is partially influenced by the number of interactions in each dataset, it is clear that the mRNA target repression inferred from the correlations increases with the length of microRNA/mRNA pairing (Figure 2B and 2C). The hierarchy of site efficacy is as follows: 8mer > 7mer-m8 > 7mer-A1 > 6mer (Grimson et al., 2007; Nielsen et al., 2007) (Figure 2B). Additionally, in contrast to the previous results (Friedman et al., 2009; Grimson et al., 2007), the 7mer-A1 seed interaction is significantly more repressive than the 7mer-m8 in the three regions of the mRNA studied and in comparison to the 6mer (whereas the 7mer-m8 shows no difference with the 6mer seed).

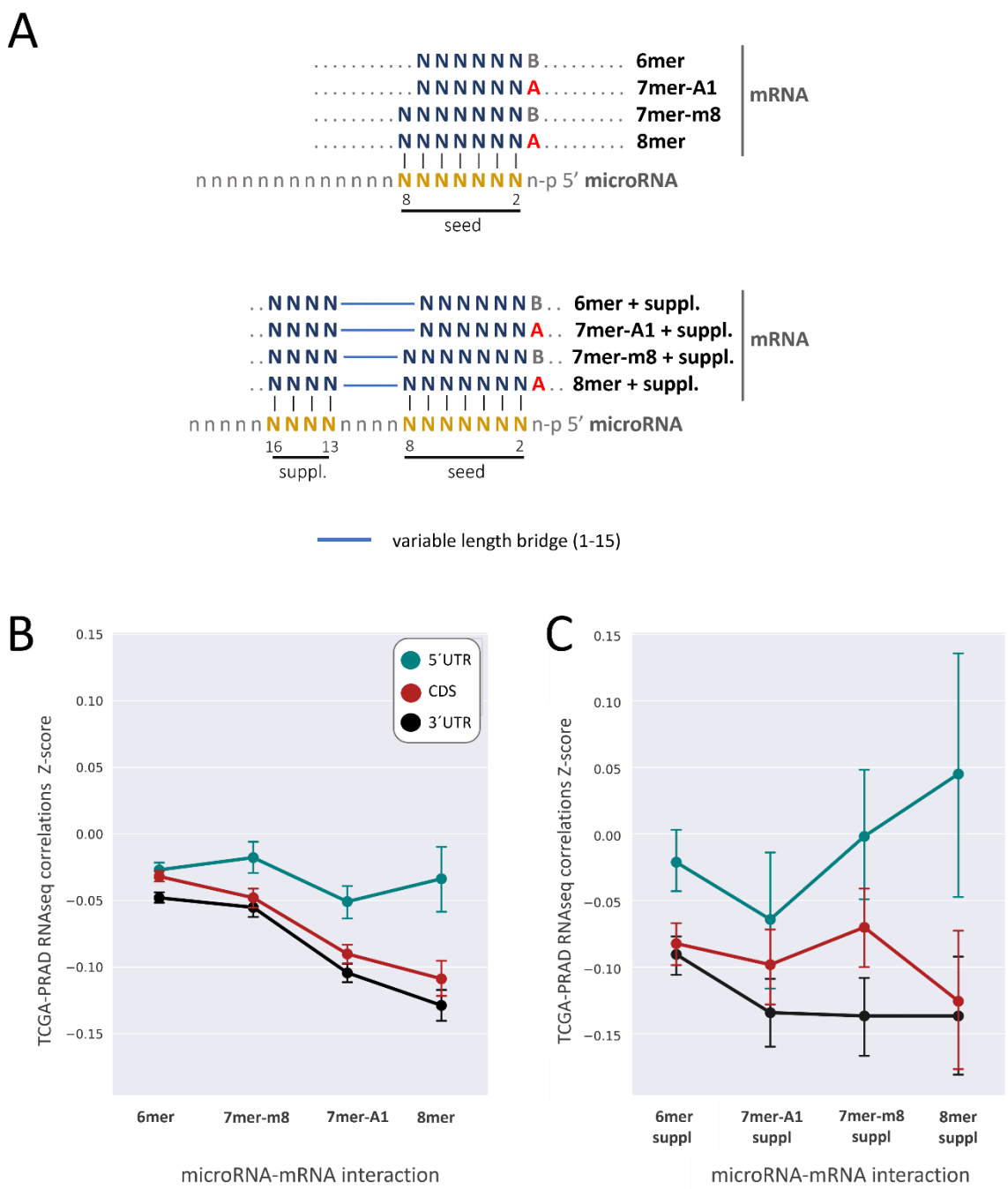

Figure 2. Repression of mRNA-target interactions with different pairing sites. The average Z-score correlations between the 143 conserved microRNAs and their predicted target mRNA are represented for each of

the CDS, 5'UTR and 3'UTR. Vertical bars represent the confidence interval (95%). Letter A in the sequences correspond to Adenine and Letter B in the sequences correspond to the other three nucleotides without Adenine.

Table 1

		FDR		
Group1	**Group2**	**5´UTR**	**CDS**	**3´UTR**
6mer	7mer-m8	n.s.	6.09E-03	n.s.
6mer	7mer-A1	1.14E-02	2.36E-42	4.82E-45
6mer	8mer	n.s.	1.74E-25	9.23E-35
7mer-m8	8mer	n.s.	2.24E-13	4.41E-23
7mer-A1	7mer-m8	2.93E-03	3.94E-14	9.48E-21
7mer-A1	8mer	n.s.	n.s.	1.24E-02
6mer	6mer suppl	n.s.	2.84E-08	9.90E-07
6mer	7mer-m8 suppl	n.s.	n.s.	3.66E-08
6mer	7mer-A1 suppl	n.s.	1.71E-04	2.48E-09
6mer	8mer suppl	n.s.	1.23E-02	3.72E-03
7mer-m8	7mer-m8 suppl	n.s.	n.s.	1.74E-06
7mer-m8	8mer suppl	n.s.	n.s.	1.52E-02
6mer suppl	7mer-m8	n.s.	2.84E-03	6.92E-04
6mer suppl	8mer	n.s.	n.s.	1.74E-03
7mer-A1 suppl	7mer-m8	n.s.	2.20E-02	2.43E-07

Table 2

		FDR			
Group1	**Group2**	**6mer**	**7mer-A1**	**7mer-m8**	**8mer**
3UTR	5UTR	5.98E-09	2.69E-14	4.34E-07	7.30E-12
3UTR	CDS	2.62E-08	1.37E-02	n.s.	n.s.
5UTR	CDS	n.s.	6.25E-08	6.82E-05	2.06E-07

		FDR			
Group1	**Group2**	**6mer suppl**	**7mer-A1 suppl**	**7mer-m8 suppl**	**8mer suppl**
3UTR	5UTR	2.81E-06	3.62E-02	1.14E-05	1.29E-03
3UTR	CDS	n.s.	n.s.	5.11E-03	n.s.
5UTR	CDS	7.30E-05	n.s.	n.s.	4.49E-03

Figure 2C plots the result of an identical analysis carried out for the interaction sites involving seeds with 3'-supplementary region possibilities. Again, most of the comparisons show a tendency of more repression for longer seeds, which is more notoriously at the CDS and 3'UTR; yet the significance of the difference is above FDR 0.05, which can be partially attributed to the small number of events. Nevertheless, a comparison of sole seeds and seeds plus 3'-supplementary regions for all the sites shows that the supplementary interaction mostly contributes to the repression, particularly at the 3'UTR and the CDS (Figure 2B and 2C and Supplementary Figure 3), while only the 6mer (CDS and 3'UTR) and 7mer-m8 (3'UTR) differences are statistically significant in this analysis (FDR < 0.05) (Table 1). This finding

supports the contribution of the supplementary 13-16 nt microRNA-mRNA interaction to the microRNA repression that has been proposed using *in vitro* approaches (Grimson et al., 2007; Helwak et al., 2013; Moore et al., 2015; Sheu-Gruttadauria et al., 2019; Xiao & Macrae, 2020). Additional validation of established rules of microRNA mechanism comes from the observation that the relative magnitude of the repression of the microRNA/mRNA interactions is 3'UTR>CDS>5'UTR (Figure 2B-C), which is supported by the FDR values of comparisons among the sites in the three different regions (Table 2). This finding corroborates that in unperturbed tissue microRNA repression likely operates mostly at the 3'UTR and CDS compared to the 5'UTR (Grimson et al., 2007; Hafner et al., 2010; Helwak et al., 2013).

Insight into the characteristics of microRNA-mRNA interaction involving beneficial supplementary pairing regions.

Our previous results drew our attention to the relevance of the 3'-supplementary region interaction for all types of seed, thus we sought to investigate if the analysis of the microRNA activity inferred from the tissue transcriptomes could uncover new rules of microRNA/mRNA pairing. For the following analyzes, we focus on the seed plus 3'-supplementary sites of 6mer+suppl (3'UTR and CDS) and 7mer+suppl. (3'UTR) datasets, since they showed statistically significant increase in target repression compared to sole seeds (Table 1); therein, we selected the individual microRNAs that exert more repression in those targets bearing 3'-supplementary interaction than in those without pairing possibilities, as determined by a T-test of Z-score differences (p-value ≤ 0.05). Overall, we found 39 repressive microRNAs in one or more of the three datasets studied and 18 microRNAs with the opposite behavior. If each dataset is analyzed separately, the number of microRNAs that follow or oppose the repressive behavior are: 19 vs 11 (6mer 3'UTR), 16 vs 4 (7mer-m8 3'UTR) and 21 vs 8 (6mer CDS). The 39 repressive microRNAs have no significant difference in expression (Supplementary Figure 4) or number of mRNA targets (Supplementary Figure 5) than the total 143 microRNAs analyzed (T-test p-value ≤0.05).

High-throughput approaches showed that nucleotides 8-13 of the microRNA are not paired in the AGO-microRNA-target complex (Hafner et al., 2010). Consistently, it was recently shown that the seed and the 3'-supplementary region could be bridged by an unstructured target loop of 1-15 nucleotides in the AGO2 complex, even when central complementary bases are available (Sheu-Gruttadauria et al., 2019). Moreover, 8mer+suppl interaction with bridges up to 10 nucleotides long in the target RNA exert more repression than 8mer sole seeds. The same study proposes that interactions involving 3'-supplementary regions with high GC content, can be established using bridges of up to 15 nt (Sheu-Gruttadauria et al., 2019). Since the bridge between the seed and the 3'-supplementary region of the microRNA is 5 nt for the 6mer+suppl (8-12 nt) and 4 nt for 7mer-m8+suppl (9-12 nt), the same lengths at the mRNA bridge means no loop formation in either of the two molecules (zero offset). Longer and shorter mRNA bridge lengths imply loop formation in the mRNA (positive offset values) or microRNA (negative offset values) respectively (Figure 3A).

Seeking to assess if the Z-score of microRNA repression withdrawn from the tissue PRAD-TCGA data was influenced by the offset of the microRNA/mRNA interaction, we analyzed the repression exerted by the 39 repressive microRNAs using the different offsets defined by base complementarity with their putative mRNA targets. As expected, the microRNA/mRNA interactions using 3'-supplementary region pairing which are not associated with repression (control) have higher Z-scores compared with the 39 repressive microRNAs, regardless of the seed type and gene location or the offset (Figure 3B, left panels). Meanwhile, the

supplementary interactions of the 39 repressive microRNAs show heterogeneous repression depending on the offset (Figure 3B, right panels). For the 6mer interactions occurring at the 3'UTR a gradual loss of repression is observed for longer negative offsets, indicating that the length of the unpaired loop between the seed and the 3'-supplementary region is relevant for the repression and displaying a negative loop-length/repression dependency. The increase in repression by shorter microRNA loops reaches its maximum at +1 offset, a pattern common to the three datasets. On the contrary the length of mRNA loops (offsets +1 to +10) is not proportional to the repressiveness of the interaction, as discrete lengths have equal or more activity than the +1 offset and others have significantly less repressive activity. A similar pattern of repression is observed for the two other datasets (6mer CDS and 7mer-m8 3'UTR), though the differences are smoother, probably due to the smaller number of interactions composing them. In addition, the integrated analysis of the three datasets supports this finding (Supplementary Figure 5).

The most repressive interactions occur for between position -2 and +8 of the mRNA loop (-logFDR ~5), with an optimum at +1, which is consistent with published experimental studies (Mcgeary et al., 2021; Sheu-Gruttadauria et al., 2019).

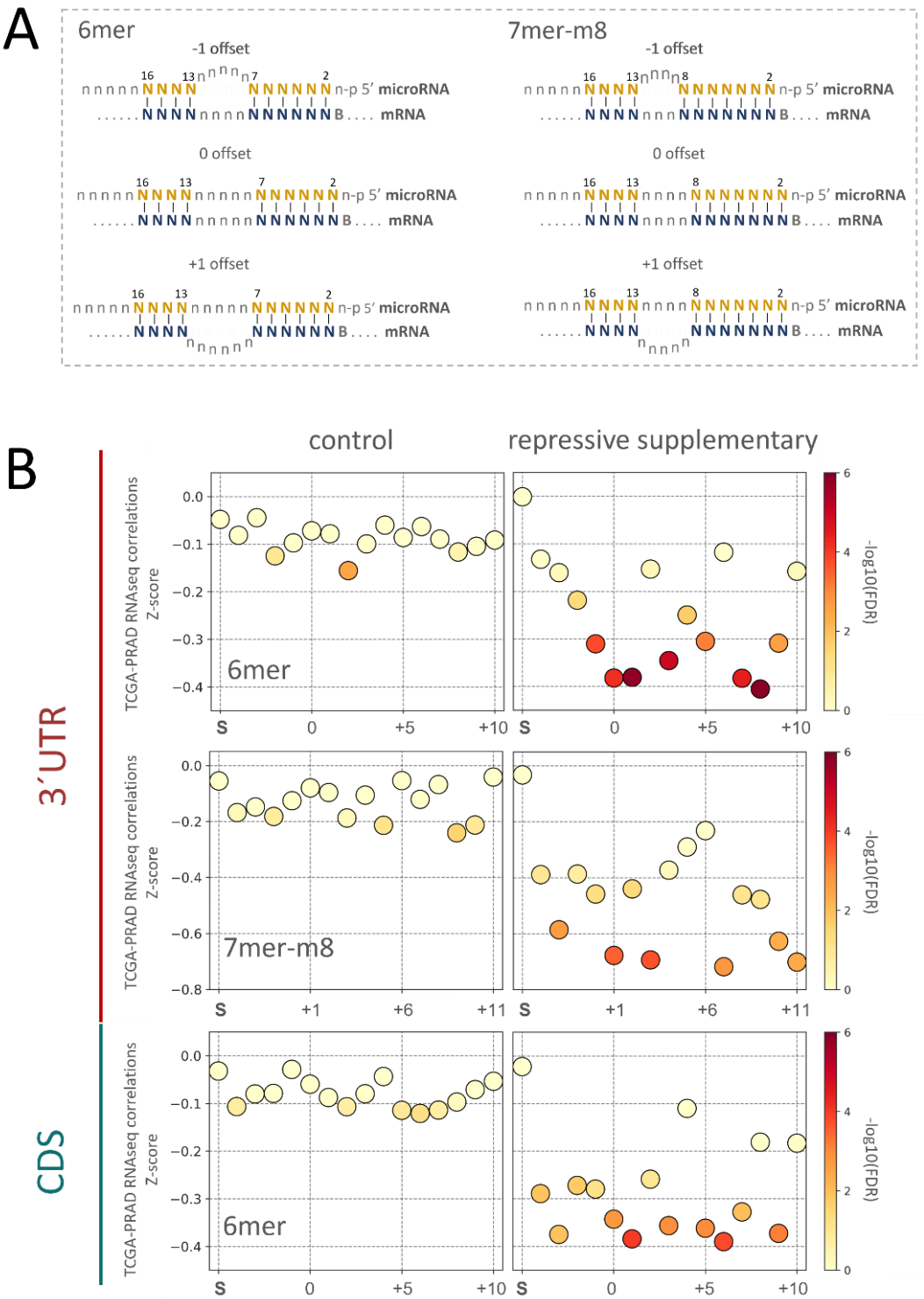

Figure 3. Contribution of the mRNA bridge length to the microRNA repression. A) Graphic representation. The bridge region is defined as the nucleotides from the seed to the 3'-supplementary region in this case considering a microRNA seed from position 2-7 (6mer) or 2-8 (7mer-m8) and microRNA supplementary from position 13-16, leaving a 5 nt (6mer) or 4 nt (7mer-m8) region between. Therefore, these values in the mRNA bridge means no loop formation in either of the two molecules denominated as zero offset. Higher values imply loop formation in the mRNA (positive offset values) and lower values in the microRNA (negative offset values). **B)** Analysis of bridge length effect on mRNA target repression (Z-transformed score) using the TCGA transcriptomes and the 39 microRNAs with repressive supplementary interactions and the 143 total microRNAs included in this study. The regions of the gene and the type of sides analyzed are indicated. The -log10(FDR) values of the

differences between each seed+suppl length and the respective sole seed interaction are shown as indicated by the color intensity bar.

We analyzed how GC content affected the interaction between microRNA with a repressive supplementary and the complete set. We observed a clear separation of the GC content between the total 143 microRNAs and the microRNAs with a repressive supplement in the stretch between position 13-16 nt of the microRNA, which shows a significantly lower GC content for the three sites analyzed (Figure 4A). In addition, positions 4, 5, 6 and 19 of 6mer sites have a GC content deviated from the total. The analysis of the base composition at each position shows a preference for base A at positions 13-15 of the 3'-supplementary pairing region of the microRNA and for base T at position 16 (Figure 4C-D) that are not observed for the total 143 microRNAs. There is also a global preference for an A at position 5. Thus, the base composition by position is similar among the three types of seed. The logo represented in Figure 4B, that is supported by 16 microRNAs (E-value =4.7e-15), resulted of the comparison the full length 39 miRNA sequence with the remaining 143 miRNAs using the Discriminative Mode of the Motif Discovery of The MEME Suite (Timothy L et al, 2009).
Although the enrichment of base T and A at position 1 and 2 of the all the microRNA observed in Figure 4C has been previously described in mammals and other species (Friedman et al., 2009; Wang, 2013), the enriched logo shows an overrepresentation of these bases.

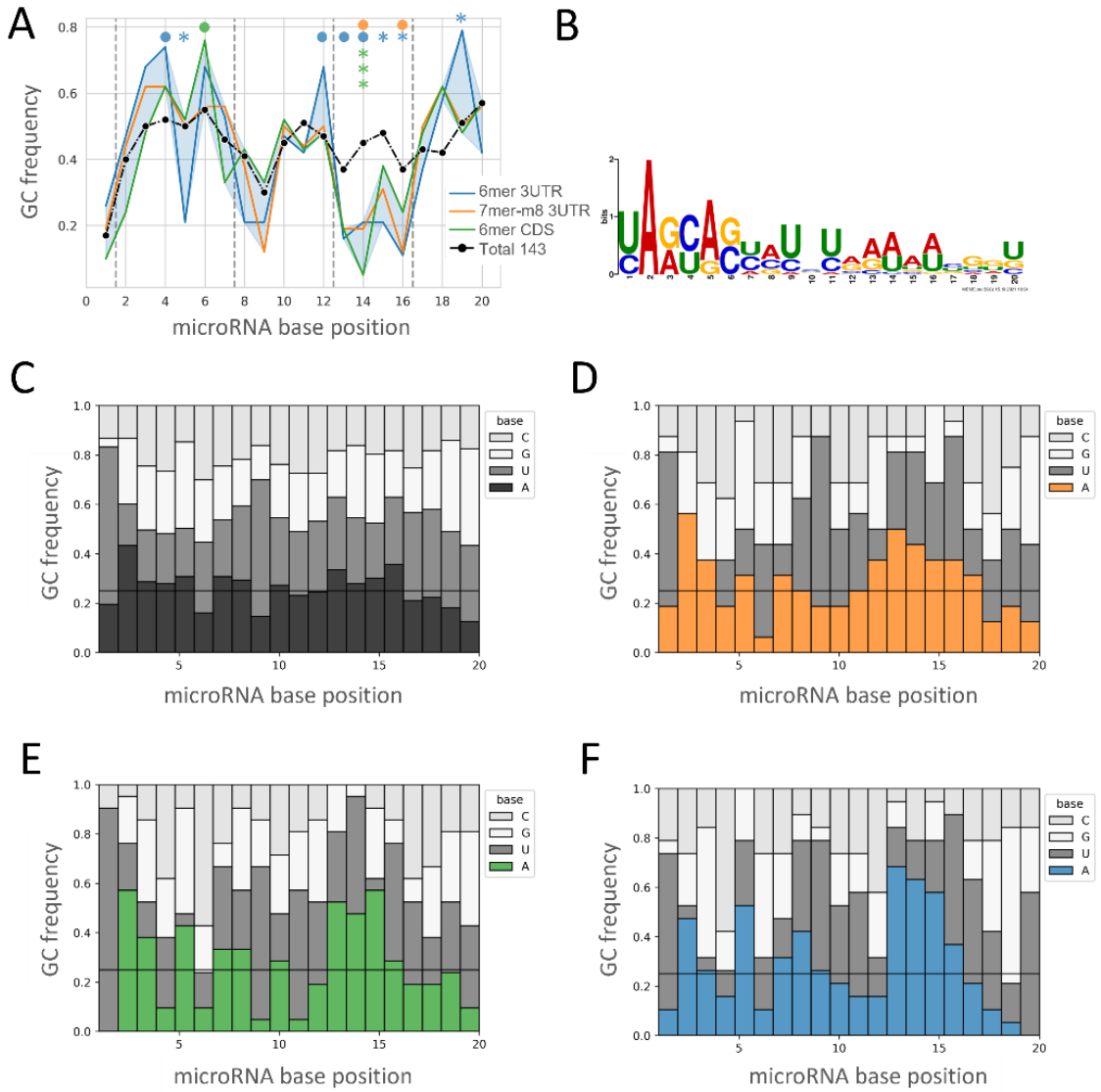

Figure 4. Nucleotide composition of the microRNAs with repressive supplementary interactions. A) Distribution of GC content along the positions of the microRNAs with the three repressive supplementary interactions indicated (colored) versus the whole set of microRNAs analyzed (black). Fisher exact test was performed to compare each of the represented sites with the total 143 microRNAs (● < 0.1, * < 0.05, ***<0.001). **B)** microRNA site identified by the MEME suite (Discriminative Mode), using the 39 versus all the 143 microRNAs. The motif is represented by 16 of the 39 microRNAs (E-value =4.7e-15). **C-F)** Base composition at each microRNA position. Site types are represented by colors as indicated in A (B:blue/6mer 3'UTR, C: orange/7mer-m8 3'UTR and D: green/6mer CDS, E: all microRNAs). Horizontal line represents the ¼ frequency expected for a random nucleotide distribution.

Finally, we developed two Windows compiled executable scripts in order to offer a user-friendly tool for accessing our data or searching for seed+suppl interaction in other datasets. One of the scripts was designed to search all the mRNA targets for a given microRNA and the other to search all microRNAs that interact with a given mRNA transcript. The software, manual and files are available in github.

Discussion

The interaction between a microRNAs and their target mRNAs can be inferred from the regulation of the target mRNA and/or protein levels provoked by the enforced modulation of the microRNA abundance in experimental settings (Baek et al., 2008; Grimson et al., 2007; Hafner et al., 2010; Helwak et al., 2013; Selbach et al., 2008). Despite the contribution to target recognition, these comprehensive experimental approaches have the disadvantage of being limited by using few microRNAs *in vitro*. Yet, this approach allows the study of the *in vitro* activity of individual microRNAs in a genome-wide or a single target gene approach. In addition, evidence of the microRNA regulation *in vivo* has been withdrawn from the correlation between microRNA/mRNA pairs abundance in unperturbed cells or tissues, which is usually used as a criterion to support the microRNA regulation of single targets (Elton & Yalowich, 2015). Since Cancer Genome projects involving multi-omics methods have provided matched RNAseq and small-RNAseq of hundreds of tissues, they allow the high-throughput transcriptome-wide assessment of microRNA-mRNA expression correlations. Here we sought to investigate if the sole analysis of microRNA-mRNA correlations in large datasets is able to validate established microRNA-mRNA pairing rules, which may in turn provide an indication of its putative power to uncover novel findings as well. We thus performed a transcriptome-wide approach measuring the correlation between all the mRNA transcripts and all the conserved microRNAs expressed in the PRAD-TCGA cohort.

Our strategy is solely based on perfect Watson-Crick base complementarity at the 2-7 or 2-8 positions of the microRNA seed region to assign candidate microRNA interaction sites on the target mRNAs. Advantages of this criteria comprise the inclusion of CDSs and 5'UTRs regions, and the lack of bias towards any of the requirements used by the diverse specific target prediction algorithms (target site conservation, free energy, site accessibility, and less frequent the target-site abundance, the local AU content, GU wobble in the seed match or unknown machine learning features) (Peterson et al., 2014).

Another advantage of our approach is the use of the endogenous level of the microRNA-mRNA partners, since concerns have been raised about artifactual observation of the perturbation based approaches (Xiao & Macrae, 2020). Meanwhile, limitations of the study are the exclusion of seed-site interactions with mismatches and the co-occurrence of multiple seed sites of microRNA regulation at the mRNAs. In addition, the lack of inclusion of common target prediction rules may decrease the biological meaning of the findings. Nevertheless, the choice to analyze only the conserved microRNAs favors the compliance to mechanistically conserved microRNA-target interaction properties.

Our initial analysis allowed the confirmation of well-established rules or microRNA-mRNA target regulation, such as a repression preference at gene regions (3'UTR>CDS>5'UTR), the strength of different seed length (6mer<7mer<8mer), the contribution of the supplementary pairing at 13-16 of the microRNA to the repression (Bartel, 2009, 2018). It is worth noting that, in terms of inferred repression, the 6mer+suppl interaction is not significantly different from the 7mer-A1, which raises the awareness of the possibly underestimated relevance of the 6mer interaction. Although the 6mer site per se is known to be less specific for repression than longer seeds, and is classified as a "marginal site" (Agarwal et al., 2015; Grimson et al., 2007), the addition of a supplementary 4 nucleotide base pairing (microRNA position 13-16) may confer a greater Argonaut mediated repressive activity. Additionally, and contrary to previous reports (Bartel, 2009; Grimson et al., 2007; Nielsen et al., 2007), we found that the 7mer-A1 seed is significantly more repressive than the 7mer-m8 in the three transcript regions

analyzed. To understand if this unexpected finding is due to the nature of the models analyzed (*in vitro* vs *in vivo*), to the bioinformatic pipelines, or another source variation, further investigation is needed.

The functionality of the 3'-supplementary pairing in microRNA-mRNA interactions was described several years ago (Grimson et al., 2007) and further insight in the molecular basis of this interaction was recently achieved (Mcgeary et al., 2021; Sheu-Gruttadauria et al., 2019; Xiao & Macrae, 2020). Unlike the first studies that used a 2-6 nt bridge length, recent publications showed that an up to 15 nt long loop on the target mRNA can bridge the seed and the 3'-supplementary region (Sheu-Gruttadauria et al., 2019). The different views of the relative contribution of the supplementary pairing to the target repression have been ascribed to the different methodological strategies of each analysis (Xiao & Macrae, 2020). A comprehensive understanding of these microRNA/target interactions restrictions are needed to improve current target prediction algorithms. In this context, we thought our approach would be useful to contribute to the understanding of the supplementary 13-16 nt interaction, because it can overcome the limitation of few 3′-supplementary sites in the transcriptome by the aggregation of multiple interactions, allowing the reliable identification of features of the supplementary interaction, such as the bridge length, base preferences, or repression strength.

We found that the use of supplementary nucleotide pairing potentiates the repressor activity of 6mer and 7mer-m8 seeds in at least 39 microRNAs. These interactions appear to be functional in the 3'UTR and coding sequence of mRNAs targets. The analysis of bridge length contribution to repression using our pipeline yields results consistent with those reported in structural (Sheu-Gruttadauria et al., 2019) and high throughput (Mcgeary et al., 2021) *in vitro* studies, where the optimal length is near or at the +1 offset. Moreover, the pattern of repression for offset larger than +1 (until +11 was analyzed) suggest that the 3´-supplementary pairing can be equally repressive with higher offsets, although the heterogeneity of the repression per offset would be compatible with heterogeneous 3' pairing microRNA architecture (Mcgeary et al., 2021). It is likely that more data, i.e., more transcriptomes and additional tissues, would be needed to increase the power of the analysis.

We also identified sequence patterns that are enriched in the 39 repressive supplementary microRNAs, including a low GC content with a higher A proportion at the 3'-supplementary region and a preferred sequence motif, whose relevance needs to be validated by additional methods. Previously, a low GC content in this region was associated to a lower affinity of the interaction *in vitro* evaluated using a specific microRNA-RNA target pair (Sheu-Gruttadauria et al., 2019; Xiao & Macrae, 2020), thus the discrepancy may be due to the inclusion of more microRNAs in their cellular context in our analysis or to still not understood factors.

Overall, our study applies a novel strategy to investigate the microRNA-mRNA target interaction by withdrawing patterns from expression correlations genome wide in a large unperturbed transcriptomic tissue dataset. The findings confirm existing knowledge, which supports the utility of the approach, so we used it to get a deeper understanding of the 3'-supplementary region.

Due to this realization of the relevance of 6mer+suppl site and the length of the bridge between the seeds and supplementary sites, and the need of user-friendly tools including these interactions, we built two Windows compiled executable scripts that predict the interactions defined in our current pipeline which is available at github.

Author Contributions: Conceptualization, M.A.D., R.S.F.; methodology, R.S.F, J.M.T.; software, R.S.F, J.M.T., G.T.; formal analysis, M.A.D., R.S.F, J.M.T.; data curation, R.S.F, J.M.T.; writing—original draft preparation, J.M.T. and M.A.D.; writing—review and editing M.A.D., R.S.F., J.M.T.; project administration, M.A.D. and B.G.; funding acquisition, M.A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Comisión Sectorial de Investigación Científica (CSIC-UDELAR, Uruguay) (Research Grants M.A.D. I+D 2016 #487 and I+D 2020 # 566) and the Programa para el Desarrollo de las Ciencias Básicas (PEDECI-BA-MEC, Uruguay, Annual Aliquots and Equipment Aid Programs of). Postgraduate Students fellowships for J.M.T. and R.F. were provided by the f the University of the Republic (UDELAR), the Comision Academica de Posgrado (CAP-CSIC-UDELAR) and the Agencia Nacional de Investigación e Innovación (ANII).

References

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., Auman, J. T., Balasundaram, M., Balu, S., Barbieri, C. E., Bauer, T., Benz, C. C., Bergeron, A., Beroukhim, R., Berrios, M., … Zmuda, E. (2015). The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, *163*(4), 1011–1025. https://doi.org/10.1016/j.cell.2015.10.025

Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *ELife*, *4*(AUGUST2015), 1–38. https://doi.org/10.7554/eLife.05005

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature*, *455*(7209), 64–71. https://doi.org/10.1038/nature07242

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bartel, D. P. (2018). *Metazoan MicroRNAs*. https://doi.org/10.1016/j.cell.2018.03.006

Broughton, J. P., Lovci, M. T., Huang, J. L., Yeo, G. W., & Pasquinelli, A. E. (2016). Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Molecular Cell*, *64*(2), 320–333. https://doi.org/10.1016/j.molcel.2016.09.004

Elton, T. S., & Yalowich, J. C. (2015). Experimental procedures to identify and validate specific mRNA targets of miRNAs. *EXCLI Journal*, *14*, 758. https://doi.org/10.17179/EXCLI2015-319

Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. https://doi.org/10.1101/gr.082701.108

Grimson, A., Farh, K. K. H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, *27*(1), 91–105. https://doi.org/10.1016/j.molcel.2007.06.017

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., & Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, *141*(1), 129–141. https://doi.org/10.1016/j.cell.2010.03.009

Helwak, A., Kudla, G., Dudnakova, T., & Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, *153*(3), 654–665. https://doi.org/10.1016/j.cell.2013.03.043

Lin, B. B., Lei, H. Q., Xiong, H. Y., Fu, X., Shi, F., Yang, X. W., Yang, Y. F., Liao, G. L., Feng, Y. P., Jiang, D. G., & Pang, J. (2021). MicroRNA-regulated transcriptome analysis identifies four major subtypes with prognostic and therapeutic implications in prostate cancer. *Computational and Structural Biotechnology Journal*, *19*, 4941–4953. https://doi.org/10.1016/J.CSBJ.2021.08.046

Mcgeary, S. E., Bisaria, N., & Bartel, D. P. (2021). Pairing to the microRNA 3′ region occurs through two alternative binding modes, with affinity shaped by nucleotide identity as well as pairing position. *BioRxiv*, 2021.04.13.439700. https://doi.org/10.1101/2021.04.13.439700

McKinney Wes. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference (SCIPY 2010)*, *41*(3), 51–56.

Moore, M. J., Scheel, T. K. H., Luna, J. M., Park, C. Y., Fak, J. J., Nishiuchi, E., Rice, C. M., & Darnell, R. B. (2015). miRNA–target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nature Communications*, *6*(May), 8864. https://doi.org/10.1038/ncomms9864

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., & Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA (New York, N.Y.)*, *13*(11), 1894–1910. https://doi.org/10.1261/RNA.768207

Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L., & Congdon, C. B. (2014). Common features of microRNA target prediction tools. *Frontiers in Genetics*, *5*(FEB), 23. https://doi.org/10.3389/FGENE.2014.00023/BIBTEX

Salomon, W. E., Jolly, S. M., Moore, M. J., Zamore, P. D., & Serebrov, V. (2015). Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. *Cell*, *162*(1), 84–95. https://doi.org/10.1016/j.cell.2015.06.029

Schirle, N. T., Sheu-Gruttadauria, J., & MacRae, I. J. (2014). Structural basis for microRNA targeting. *Science*, *346*(6209), 608–613. https://doi.org/10.1126/science.1258040

Schnall-Levin, M., Zhao, Y., Perrimon, N., & Berger, B. (2010). Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3′UTRs. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(36), 15751–15756. https://doi.org/10.1073/pnas.1006172107

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., & Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature*, *455*(7209), 58–63. https://doi.org/10.1038/nature07228

Sheu-Gruttadauria, J., Pawlica, P., Klum, S. M., Wang, S., Yario, T. A., Schirle Oakdale, N. T., Steitz, J. A., & MacRae, I. J. (2019). Structural Basis for Target-Directed MicroRNA Degradation. *Molecular Cell*, *75*(6), 1243-1255.e7. https://doi.org/10.1016/j.molcel.2019.06.019

Sheu-Gruttadauria, J., Xiao, Y., Gebert, L. F., & MacRae, I. J. (2019). Beyond the seed: structural basis for supplementary microRNA targeting by human Argonaute2. *The EMBO Journal*, e101153. https://doi.org/10.15252/embj.2018101153

Wahlquist, C., Jeong, D., Rojas-Muñoz, A., Kho, C., Lee, A., Mitsuyama, S., Van Mil, A., Jin Park, W., Sluijter, J. P. G., Doevendans, P. A. F., Hajjar, R. J., & Mercola, M. (2014). Inhibition of miR-25 improves cardiac contractility in the failing heart. *Nature*, *508*(7497), 531–535. https://doi.org/10.1038/nature13073

Wang, B. (2013). Base composition characteristics of mammalian mirnas. *Journal of Nucleic Acids*, *2013*. https://doi.org/10.1155/2013/951570

Wee, L. M., Flores-Jasso, C. F., Salomon, W. E., & Zamore, P. D. (2012). Argonaute divides Its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, *151*(5), 1055–1067. https://doi.org/10.1016/j.cell.2012.10.036

Wei, J., Yin, Y., Deng, Q., Zhou, J., Wang, Y., Yin, G., Yang, J., & Tang, Y. (2020). Integrative Analysis of MicroRNA and Gene Interactions for Revealing Candidate Signatures in Prostate Cancer. *Frontiers in Genetics*, *11*, 176. https://doi.org/10.3389/FGENE.2020.00176/BIBTEX

Xiao, Y., & Macrae, I. J. (2020). Robust differential microRNA targeting driven by supplementary interactions in vitro. *RNA*, *26*(2), 162–174.

https://doi.org/10.1261/rna.072264.119
Yang, Y., Jia, B., Zhao, X., Wang, Y., & Ye, W. (2019). miR-93-5p may be an important oncogene in prostate cancer by bioinformatics analysis. *Journal of Cellular Biochemistry*, *120*(6), 10463–10483. https://doi.org/10.1002/jcb.28332
Ye, Y., Li, S. L., & Wang, S. Y. (2018). Construction and analysis of mRNA, miRNA, lncRNA, and TF regulatory networks reveal the key genes associated with prostate cancer. *PLOS ONE*, *13*(8), e0198055. https://doi.org/10.1371/JOURNAL.PONE.0198055
Yue, D., Liu, H., & Huang, Y. (2009). Survey of Computational Algorithms for MicroRNA Target Prediction. *Current Genomics*, *10*(7), 478–492. https://doi.org/10.2174/138920209789208219