Bioinformatic description of tRNA-fragments and their targets in human AGO2

Jacob Peacock

Advisor: Andrey Grigoriev

December 19, 2016

1 Introduction

tRNA-fragments (tRFs) have recently been shown to play a regulatory role similar to miRNAs and have been linked to cancer, neurodegeneration and transgenerational epigenetic inheritance. tRFs are loaded into argonaute proteins (AGO) and incorporated into RISC complexes as *guides* for identifying *target* RNAs, suggesting a role of tRFs as similar to that of miRNAs in RNA interference (RNAi) (Anderson 2014). Four AGO proteins are expressed in humans and differential loading of guide miRNAs and tRFs has been observed previously, with AGO2 the least favored for tRF loading (Kumar 2014). To this end, we make a detailed analysis of the tRF guides and their targets in AGO2 for later comparison with AGO4 behavior (Karaiskos 2015).

For this analysis, we use data from photoactivable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP), a high-throughput biochemical technique for sequencing the loaded targets and guides of AGO proteins. Although PAR-CLIP was intended to yield merely a collection of targets and guides, subsequent analysis has found target-guide chimeras directly associating guides with their particular targets (Grosswendt 2014). While recent techniques like CLASH provide these direct associations readily, there remains much to be learned from the more abundantly available PAR-CLIP data.

2 Methods

Approximately 140,000 unique chimeras with a tRF ligated to a target of length > 16 nt were recovered from the PAR-CLIP data of Kishore *et al* (2011). The results for all 6 experiments were combined and the adapter sequence TCGTATGCCGTCTTCTGCTTGT was removed with `fastx_clipper`. Identical reads were collapsed and the number of identical reads recorded an abundance with `fastx_collapser`. Chimeric reads that began with a tRFs were identified for tRFs derived from human nuclear and mitochondrial tRNAs described in gtRNAdb and tRNAdb, producing a liberal set of tRFs by including bioinformatically identified tRNA genes which may or may not be transcribed. All subsequent calculations and plotting were performed using Python and LibreOffice Calc.

2.1 Figure 2

The length of every unique tRF detected in the PAR-CLIP was tallied. The target sequences of all detected tRF-target chimeras were identified using BLAST against human intronic and exonic mRNA, lnRNA, lncRNA, miRNA, rRNA, snRNA and other miscellaneous RNAs. The lengths of those unique tRFs detected in chimeras with identifiable targets were tallied.

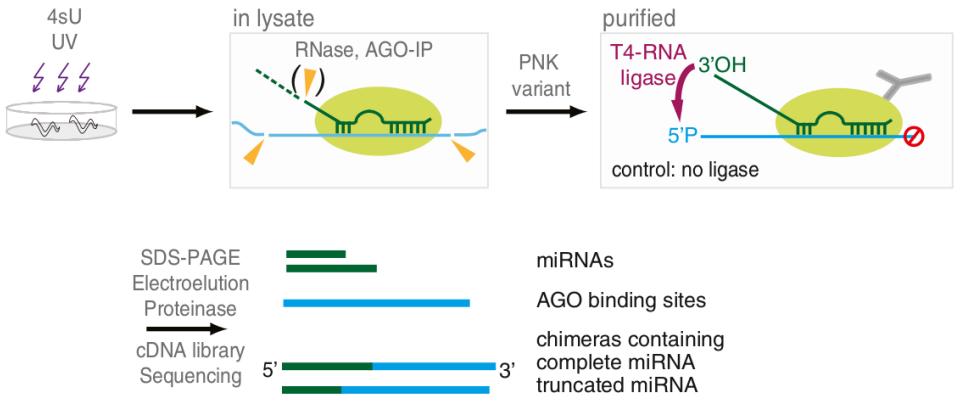

Figure 1: Illustration of the PAR-CLIP protocol, beginning with UV cross-linking of the photoactivable ribonucleoside *in vivo*. After lysing, RNases trim the guide and target sequences and AGO2 is immunoprecipitated. Endogenous ligases next act to ligate the guide and target, producing a chimera that can be isolated by SDS-PAGE and sequenced. Figure adapted from Grosswendt 2014.

2.2 Figure 3

Using available annotations of tRNA secondary structures provided by tRNAdb, we annotated the secondary structures of tRF starts and ends. 30% of the ~140,000 unique tRFs were annotated. The fraction of these annotated tRFs starting and ending in each category are shown in Figure 3(A). Categories with zero tRFs are not shown. This same set of tRFs is recategorized in Figure 3(C). All tRFs with a terminus in the 5' or 3' acceptor stem or CCA end and length > 32 were classified as tRNA halves. All shorter tRFs starting in the 5' acceptor stem were classified as 5' tRFs, while those ending in the 3' acceptor stem or CCA end were classified as 3' tRFs. (D, A)-tRFs and (A,T)-tRFs were classified as in Figure 3(A). All tRFs not in the previous classes were classified as Other.

2.3 Figure 4

For all detected tRF-target chimeras, the target sequences were aligned to human mRNAs with up to 3 mismatches using Bowtie. This allowed us to detect T to C conversions produced by the RNA-protein crosslinks during PAR-CLIP as a mismatch in the alignment, with a C in the experimental results where a T was expected. For each mRNA, the number of such mismatches found in targeted mRNA sequences was tallied at each position, yielding a profile of mismatches. A 41-nt window centered at the most frequently converted position was extracted from each mRNA. When the mRNA was not long enough to admit such a 41-nt window, the window was truncated. The complementarity of the reverse complement of the putative seed sequence CCAGGGA was tested at each position along each window. Positions with perfect complementarity were tallied across all windows, with the windows aligned at the most abundant mismatch.

2.4 Figure 5

The target sequences of all detected tRF-target chimeras were BLASTed against human intronic and exonic mRNA, lnRNA, lncRNA, miRNA, rRNA, snRNA and other miscellaneous RNAs. The abundance of the chimeras containing a particular target sequence was summed and all target sequence abundances of a particular class of RNAs summed. These total abundances are displayed for each class of RNAs.

2.5 Figure 6

The target sequences of all detected tRF-target chimeras were BLASTed against the human genome. Promoter regions were identified as those 2000 nt preceding a know transcription start site. All target sequences with matches starting in identified promoter regions were collected, yielding 506 targets. From this subset of target sequences, we excluded 23 targets matching perfectly to multiple locations of the genome, 29 targets previously treated in Figure 5 and 45 targets which fell into the promoter region of multiple distinct transcription sites. For the remaining 409 promoter targets, the exact distance from the start of the transcription site was determined, taking into account the varying orientation of transcription sites. The promoter targets were binned by their distance from the transcription start and the number of promoter targets with a positive and negative orientation on the genome displayed for each bin.

3 Results

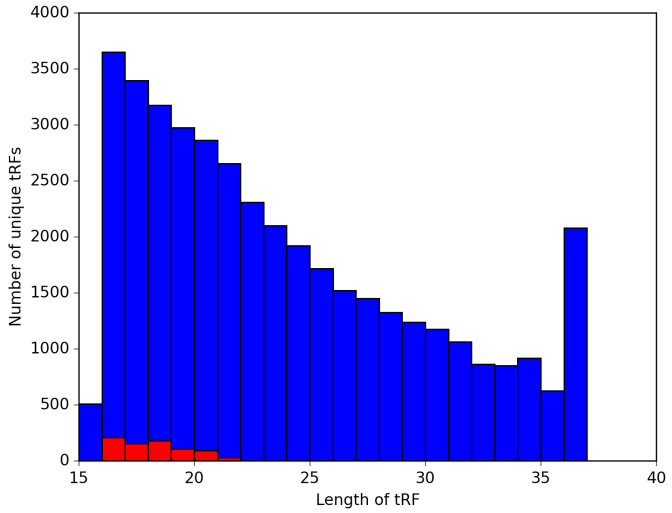

Figure 2: Length distribution of all unique tRFs detected in the Kishore PAR-CLIP experiments. In red at the bottom, is that subset of tRFs found ligated to identifiable targets.

We observed ∼40,000 unique tRFs of lengths between 16 and 36 nt in the PAR-CLIP data. Those tRFs of length > 32 nt are outside the range usually identified as tRFs (Kumar 2014) and are better characterized as tRNA halves. The roughly linear trend in sizes (Fig 2) is concordant with random RNase degradation of tRFs during the PAR-CLIP protocol. The deviation from linearity at 36 nt is most likely an artifact of the sequencing read length of 40 nt, effectively limiting the observable length of a tRF. Lastly, in red is that fraction of tRFs present in chimeras for which the target sequence could be annotated. This suggests our analysis of tRF targets will be representative of only a small slice of all tRFs.

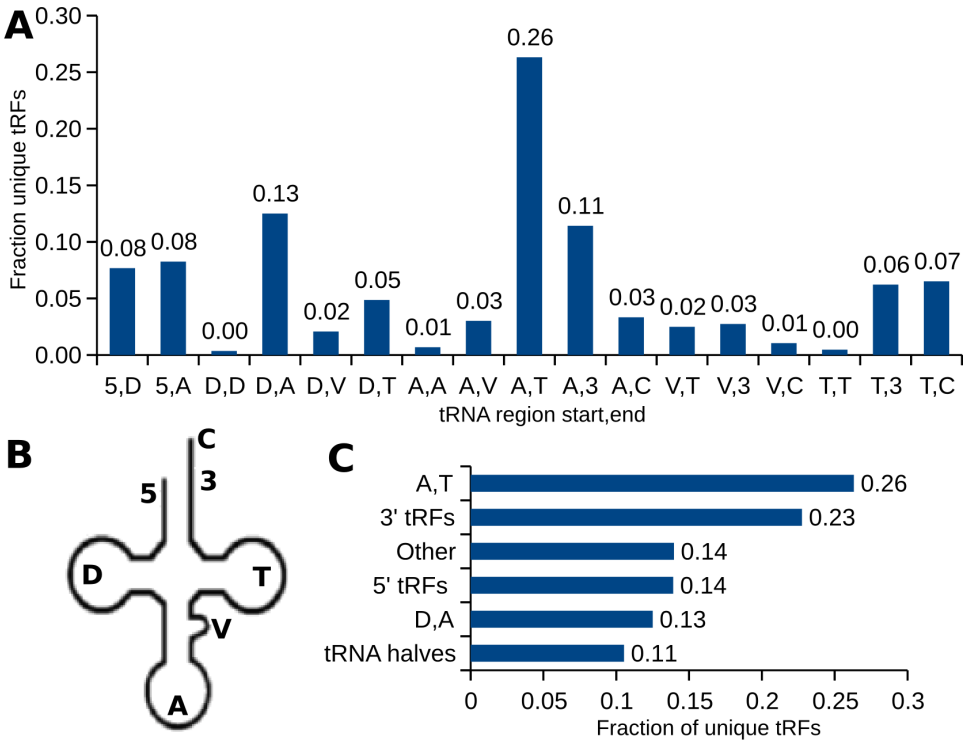

Figure 3: The derivation of tRFs across tRNAs differs from the canonically described tRF distribution. **(A)** Both (D,A)-tRFs and (A,T)-tRFs account for a significant portion of all observed tRFs and of those tRFs not classified as halves, 5' or 3' tRFs.**(B)** Single letter codes were assigned to each major secondary structure of the tRNA: *5*' acceptor stem, *D*-loop, *A*nti-codon loop, *V*ariable loop, *T*-loop, *3*' acceptor stem and *C*CA end were used as indicated. **(C)** Canonical tRFs still account for a significant portion of our observed tRFs.

Using the available tRNA secondary structure annotations, we classified 30% of the unique tRFs by their origin and terminus on their parent tRNA (Fig 3A). An unexpectedly high fraction of tRFs originated in the anti-codon loop and terminated in the T-loop (A,T 25%). We further classified tRFs into established categories of 5' and 3'-tRFs, deriving from the 5' and 3' end of the tRNA, respectively, as well as identifying tRNA halves (Fig 3C). The canonical categories of 3' and 5'-tRFs and tRNA halves account for 48% of the fragments we observed.

The unusually high portion of (A,T)-tRFs may be an artifact of RNase degradation or related to the greater proportional length of these regions relative to the rest of the tRNA. Some attempts were made to normalize these fractions to the relative length represented in each category (not shown). These efforts are severely hindered by the highly non-uniform distribution of tRFs across tRNAs. A handful of tRNAs account for a large fraction of the unique tRFs and such normalization would require weighting to account for this. While these obstacles could be overcome, it seems likely that the tRFs represented in PAR-CLIP are incomplete and do not accurately represent the actual forms of tRFs. Instead, these "fragment of fragments" might better be interpreted as bioinformatic identifiers for their progenitor tRF rather than the *in vivo* sequence of tRFs.

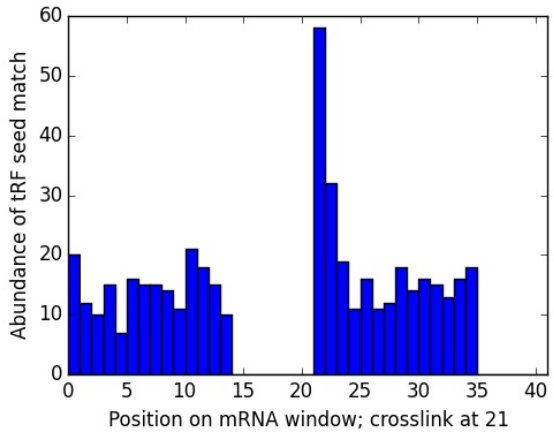

Figure 4: For a tRF derived from $\text{tRNA}^{\text{Glu}}_{\text{GTC}}$, we observe a peak in complementarity to target mRNA sequences immediately subsequent protein-mRNA crosslinks, indicative of a tRF seed region. The hypothesized seed sequence is CCAGGGA. Note that with a fixed T at position 21, complementarity is not possible at positions 14–21, explaining the gap at those positions.

In analogy with the miRNA guide sites that recognize and bind complementary targets, we identify such seed regions in tRFs based on the PAR-CLIP data. Experiments with miRNAs have shown that protein-target crosslinks, which change a T to a C in the target sequence, immediately precede seed sites. This is possibly due to protective effects of the guide-target complementarity or the structure of the AGO complex. We demonstrate the activity of a representative tRF seed site in Fig 4 by aligning all mRNA targets at their crosslinks and checking for complementarity of the posited tRF seed at each position. With the crosslink at position 21, we see a peak in complementarity at position 22 as expected, indicating a functional seed region of the tRF.

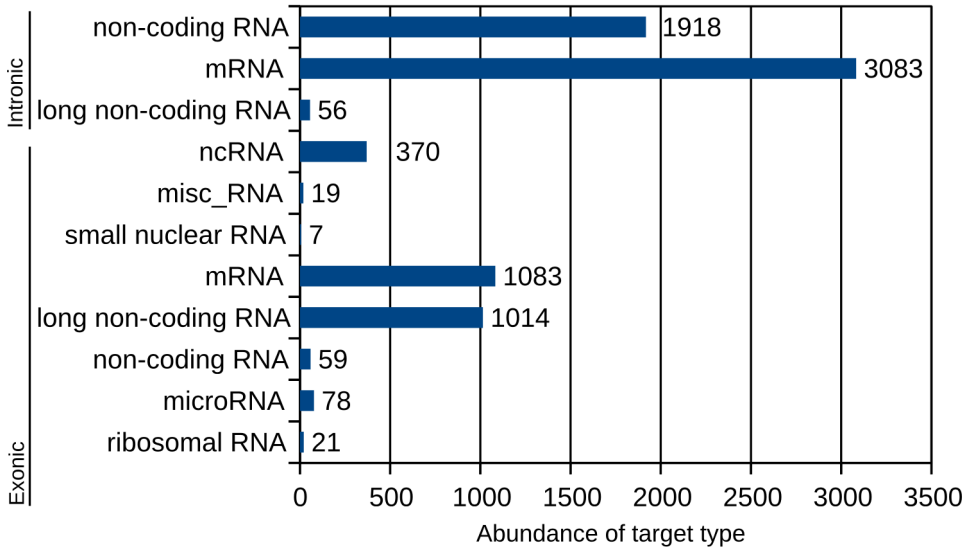

Figure 5: Intronic mRNAs and ncRNAs dominate the identified targets of tRFs, while exonic mRNAs and lncRNAs constitute another significant portion of targets. Also present are a small number of rRNA, miRNA and snRNAs with less clear functions as targets.

Analyzing the sequences targeted by tRFs, we identified novel intronic mRNA targets in surprising quantities, as well as conventional exonic mRNA targets alongside lncRNA, rRNA and miRNA targets (Fig 5). The presence of intronic mRNAs suggests several hypotheses: (1) contamination during the experiment, (2) AGO complexes are functioning in the nucleus, or (3) intronic mRNAs are exported from the nucleus. The first hypothesis has been rejected based on PAR-CLIP experiments where exogenous contaminating RNAs were deliberately introduced and shown to be incorporated at rates less than 2% (Broughton 2016). The likely case is a combination of the second and third hypotheses, since RNAi has been shown to be active in the nucleus (Gagnon 2014) and introns have been shown to be exported from the nucleus after splicing (Hesselberth 2013). Note that the "targeting" of miRNAs by tRFs should draw our attention to the symmetry of the experiment: these reads could just as well be interpreted as miRNAs targeting tRFs, as has been previously reported.

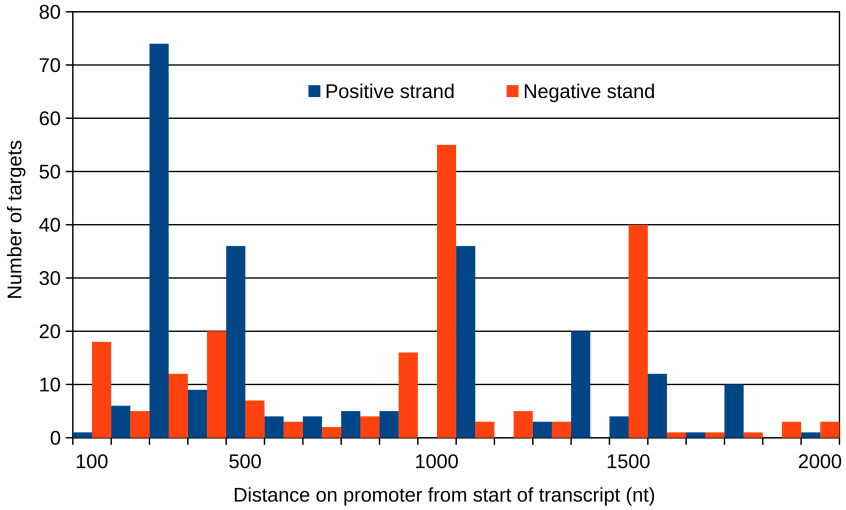

Figure 6: tRFs targeting occurs in 3 clusters on gene promoters. Targeting occurs on both the positive and negative (right red columns and left blue, respectively) strand in all 3 clusters.

Finally, we searched for tRF target sequences that mapped to 2000 nt preceding transcription start sites. Sequences in these regions may serve as promoters for the subsequent gene. In Fig 6, we show the distribution of the start positions of the ∼400 promoter targets relative the transcription start of the respective genes. We observe 3 peaks in targeting, spaced approximately 500 nt along the promoter regions. In each of these peaks, we observe targeting on both the positive and negative strand, with some parts of the promoter showing exclusive or near exclusive targeting of one strand or the other. Interestingly, for 65% of the sites, the orientation of the targeted promoter sequence did not agree with the orientation of the gene itself.

4 Conclusion

Our current studies of tRF targeting in AGO2 using PAR-CLIP data are limited by read length and RNase degradation. The tRF sequences recovered by PAR-CLIP and related protocols where the target sequence is prone to degradation are best interpreted as identifiers for their respective tRFs, rather than the tRFs themselves. We observed a novel and unexpected preponderance of intronic targets of tRFs. Further exploration of this phenomena, particularly whether such intronic sequences are targeted by miRNAs, may yield insight into the still poorly understood role of post-splicing introns. The apparent targeting of rRNAs and miRNAs remains to be explained and might be posited as yet another layer of regulatory mechanism. Finally, our study of promoter targeting requires a biological explanation for the observed distribution of promoter sites and orientations.

5 References

1. Anderson, P., & Ivanov, P. (2014). TRNA fragments in human health and disease. FEBS Letters, 588(23), 4297-4304. doi:10.1016/j.febslet.2014.09.001

2. Broughton, J. P., & Pasquinelli, A. E. (2016). A tale of two sequences: MicroRNA-target chimeric reads. Genetics Selection Evolution Genet Sel Evol, 48(1). doi:10.1186/s12711-016-0209-x

3. Gagnon, Keith T. et al. "RNAi Factors Are Present and Active in Human Cell Nuclei." Cell Reports 6.1 (2014): 211221. www.cell.com. Web.

4. Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Rajewsky, N. (2014). Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. Molecular Cell, 54(6), 1042-1054. doi:10.1016/j.molcel.2014.03.049

5. Hesselberth, Jay R. "Lives That Introns Lead after Splicing." Wiley interdisciplinary reviews. RNA 4.6 (2013): 677691. PubMed. Web.

6. Karaiskos, Spyros, Ammar S. Naqvi, Karl E. Swanson, and Andrey Grigoriev. "Age-driven Modulation of TRNA-derived Fragments in Drosophila and Their Potential Targets." Biology Direct Biol Direct 10.1 (2015). Web.

7. Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., & Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nature Methods Nat Meth, 8(7), 559-564. doi:10.1038/nmeth.1608

8. Kumar, P., Anaya, J., Mudunuri, S. B., & Dutta, A. (2014). Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biology BMC Biol, 12(1). doi:10.1186/s12915-014-0078-0