# Signatures of RNA binding proteins globally coupled to effective microRNA target sites

Anders Jacobsen (1, 3), Jiayu Wen (1), Debora S. Marks (2,*), Anders Krogh (1,*)

(1) The Bioinformatics Centre, Department of Biology, and the Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark
(2) Systems Biology Department, Harvard Medical School, Boston, MA, USA
(3) Present address: Computational and Systems Biology Center, Memorial Sloan-Kettering Cancer Center, NY, USA

(*)These authors contributed equally to this work

Correspondence to: deboramarks@gmail.com, krogh@binf.ku.dk

## Supplementary Discussion

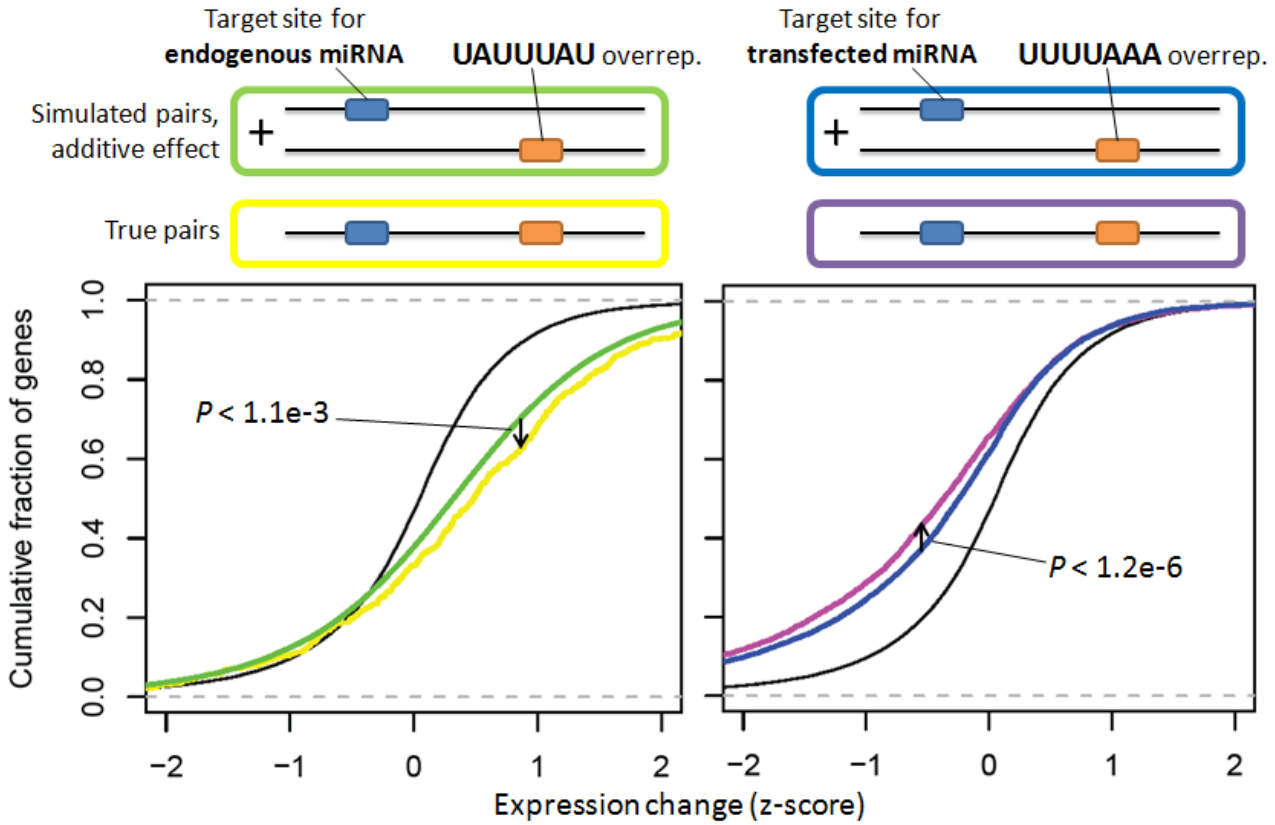
## Contents



## Figures

## Testing cooperativity of ARE/URM motifs and endogenous/transfected miRNA target sites

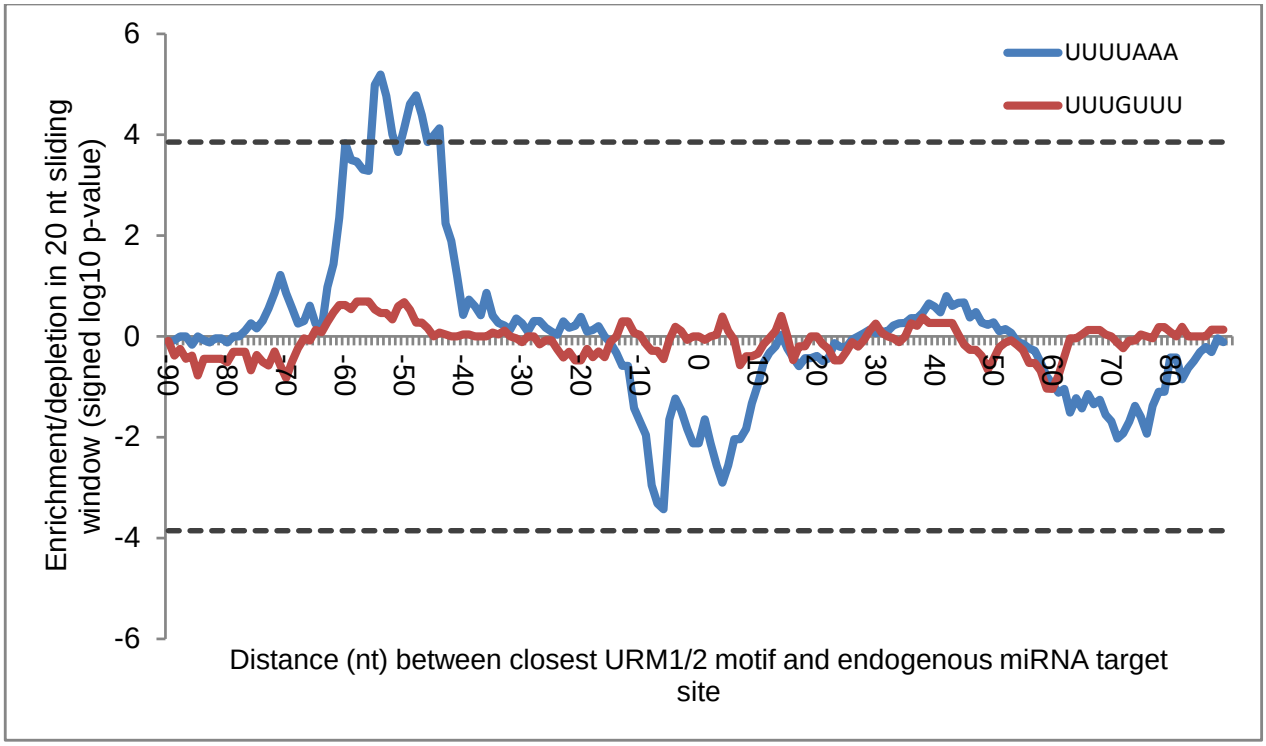
We simulated expression changes of genes having non-cooperative occurrences of (i) ARE motif overrepresentation ($P < 0.05$) and endogenous microRNA target sites, (ii) URM1 motif overrepresentation ($P < 0.05$) and transfected microRNA target sites (see Methods). To accommodate for the greater variance of the simulated compounded distribution (due to microarray measurement noise), the expression changes of *true pair* genes were summed with the expression change of randomly selected genes in the experiment. Comparing the distribution of expression changes for *true pairs* and *simulated pairs* revealed that (i) genes with true co-occurrences of ARE overrepresentation and endogenous miRNA target sites were more *up*-regulated than expected if they had additive effects ($P < 1.1e-3$, KS one sided test), (ii) genes with true co-occurrences of URM1 overrepresentation and transfected miRNA target sites were more down-regulated than expected if they had additive effects ($P < 1.2e-6$, KS one sided test). We noted that the test for cooperativity were even more significant without the modification in variance (adding random expression changes) of the *true pairs* distribution ($P < 7.1e-9$ and $P < 3.7e-8$ for ARE and URM1 motifs respectively).




**Supplementary figure 1: Cooperativity of ARE/URM motifs and endogenous/transfected miRNA target sites**

# Spatial bias between endogenous microRNA target sites and URM motifs in PAR-CLIP data

We analyzed data generated by the PAR-CLIP method (Hafner et al. 2010) for local spatial bias between endogenous miRNA target sites (7mer seed site of miR-15/16, miR17/93 and miR-19) and each of the URM and ARE motifs in genes bound by AGO in HEK293 cells. Given a set of motif and endogenous miRNA target site instances in a specific 3'UTR, we recorded the shortest absolute distance between any motif and predicted miRNA target site instance. We estimated expected distances by permuting all motif and miRNA target site instances 100 times in each 3'UTR. Spatial enrichment and depletion of motifs relative to miRNA target sites was evaluated by a two-tailed binomial test in sliding windows of 20 nt. We found a significant enrichment of URM1 motifs approximately 50 nt upstream of miRNA target sites (dashed line corresponds to Bonferroni corrected significance threshold). URM2 motifs also showed a subtle non-significant enrichment in this interval.



## Word enrichment analysis of crosslinked AGO binding regions in PAR-CLIP data

Using data generated by the PAR-CLIP method (Hafner et al. 2010), we analyzed words enriched in 17319 3'UTR sequence clusters bound by AGO in HEK293 cells. We evaluated overrepresentation (presence/absence) of all 7mers in these 41 nt sequence clusters relative to dinucleotide shuffled sequences. Overrepresentation was quantified using a z-score statistic comparing observed occurrences to the distribution of expected occurrences estimated from 500 cohorts of dinucleotide shuffled sequences and two-tailed Bonferroni corrected P-values are estimated from the z-scores. In consistency with the original paper we found strong enrichment of 7mers corresponding to seed sites of the most highly expressed endogenous miRNAs. We also found significant enrichment of ARE and URM1 motifs. However, URM2 occurrences were not found more often than would be expected from dinucleotide frequencies in the crosslinked sequence clusters.

| Rank | Word | Z-score | Observed Occurrences | Expected occurrences | P-value | Annotation |
|---|---|---|---|---|---|---|
| **1** | UGCUGCU | 45.06 | 602 | 103+-11.1 | 0 | miR-15/16 |
| **2** | GCACUUU | 39.75 | 437 | 76+-9.1 | 0 | miR-17/93/20a/106b |
| **3** | GCUGCUA | 38.00 | 306 | 47+-6.8 | 0 | miR-15/16 |
| **4** | UUGCUGC | 34.49 | 367 | 81+-8.3 | 1.3e-256 | miR-15/16 |
| **5** | CACUUUA | 33.32 | 355 | 75+-8.4 | 6.3e-239 | miR-17/93/20a/106b |
| **6** | AUGCUGC | 31.25 | 322 | 61+-8.3 | 1.7e-209 | hsa-miR-103,hsa-miR-107 |
| **7** | UGCACUU | 31.15 | 411 | 97+-10.1 | 3.9e-208 | miR-17/93/20a/106b/519 |
| **8** | GUGCAAU | 28.17 | 234 | 47+-6.6 | 9.6e-171 | miR-92a/32 |
| **9** | UUUGCAC | 26.87 | 335 | 83+-9.3 | 3.6e-155 | miR-19a/19b |
| **10** | GCUGCUU | 25.00 | 286 | 74+-8.5 | 1.1e-133 | miR-15/16 |
| **...** | | | | | | |
| **110** | UAUUUAU | 9.04 | 186 | 99+-9.6 | 2.6e-15 | ARE |
| **126** | UUUUAAA | 8.55 | 277 | 169+-12.6 | 2.0e-13 | URM1 |
| **7882** | UUUGUUU | 1.38 | 190 | 173+-11.9 | 1 | URM2 |

## Change in expression of ARE binding proteins after microRNA transfections

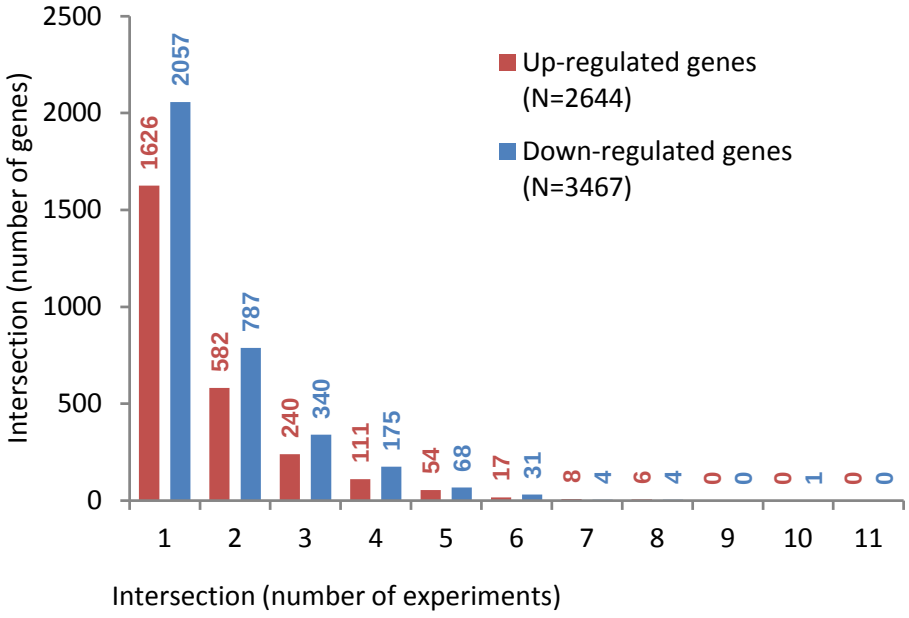
We tested if the change in expression of ARE-BPs themselves are the cause of the ARE motif mediated up-regulation. We evaluated if 16 known ARE-BPs [1] were significantly down- or up-regulated after each transfection in HeLa cells. We found that none of the proteins alone had a consistent mRNA expression change pattern which could be coupled directly to the general up-regulation of ARE genes (see table below). Neither is the ARE BP expression data consistent with combinations of ARE BP expression changes causing the constant up-regulation signal, as random combinations of expression changes should lead to both up- and down-regulation of ARE genes in the different experiment.

| ARE-BP | Expression percentile | miR-181a | miR-9 | miR-128 | miR-132 | miR-133a | miR-142 | miR-148 | miR-122 | miR-7 | miR-124 | miR-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AUF1 / HNRNPD** | 84 | - | + | -- | - | - | - | - | - | - | -- | -- |
| **AUH** | 72 | - | -- | + | + | + | + | + | ++ | + | + | - |
| **BRF1** | 72 | + | + | + | + | + | + | - | + | + | + | + |
| **CUGBP1** | 99 | - | - | + | - | - | + | + | + | - | - | + |
| **GAPDH** | 98 | + | - | + | + | + | - | - | + | - | + | + |
| **HNRNPA1** | 99 | - | - | + | - | - | + | + | + | - | - | + |
| **Hsp70 / HSPA4** | 89 | + | - | + | + | + | + | + | + | - | -- | ++ |
| **HuB / ELAVL2** | 24 | - | + | - | - | + | - | + | + | - | ++ | ++ |
| **HuC / ELAVL3** | 16 | + | + | + | + | - | - | - | - | - | - | - |
| **HuD / ELAVL4** | 6 | + | - | + | + | - | + | - | + | + | - | + |
| **HuR / ELAVL1** | 41 | + | -- | - | + | - | - | - | - | + | + | ++ |
| **KSRP / KHSRP** | 62 | - | + | - | + | -- | - | + | - | + | + | - |
| **NCL** | 99 | + | - | + | + | - | + | + | + | - | + | + |
| **PAIP2** | 93 | - | -- | -- | -- | + | - | + | + | + | + | + |
| **TIA1** | 73 | - | - | - | - | - | - | + | + | + | - | - |
| **TTP / ADAMTS13** | 57 | - | - | - | + | + | - | - | + | - | + | + |

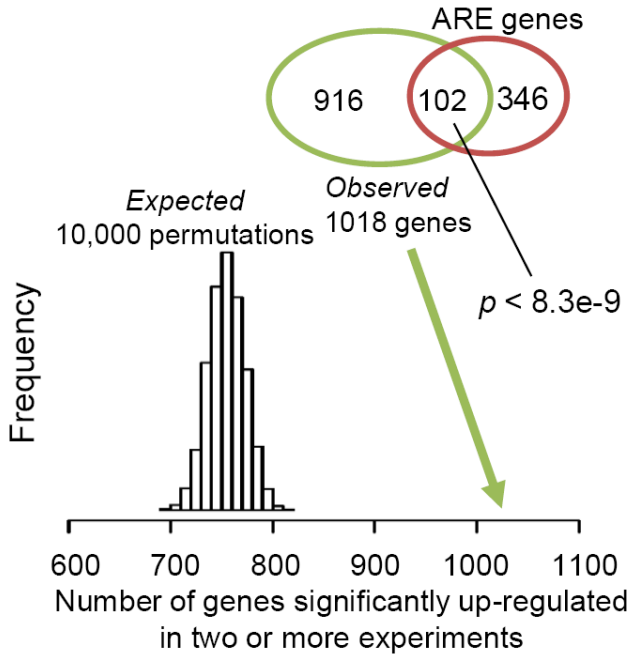
**Expression changes of ARE binding proteins:** mRNA expression changes of 16 known ARE-BPs in the 11 HeLa miRNA transfections [1]. The column “Expression percentile” indicates the relative expression level (1-100) of the given ARE-BP in HeLa cells (estimated from the control samples in the 11 expression profiles). Non-significant down-regulation is indicated by “-“, and up-regulation by “+”. Significant ($p<0.05$) down or up-regulation is indicated by “--" and “++” respectively.

## Genes differentially regulated across multiple miRNA transfection experiments

For each of the 11 different miRNA transfections in HeLa cells, we selected the significantly ($P < 0.05$) up or down regulated genes. In the 11 experiments, the up-regulated sets consisted of 225, 435, 376, 257, 670, 175, 483, 376, 107, 819 and 507 genes, while the down-regulated sets consisted of 371, 556, 626, 546, 849, 357, 516, 522, 233, 762 and 609 genes. The union of genes significantly up-regulated ($p<0.05$) in at least one of the 11 experiments comprised 2644 different genes (3467 genes were down-regulated). The figure below shows the size of the intersections among these 11 gene sets. However, although we see significant overlap between pairs of experiments (P < 1e-4, permutation test, see Supplementary figure 3 below), we do not see any genes consistently up-regulated across most the experiments. Inspecting the overlap of these genes, only 31 (40) genes were significantly up (down) regulated in at least half of the 11 experiments, and only 1 gene is significantly down regulated in more than 9 experiments.




**Supplementary figure 2: genes differentially regulated across multiple miRNA transfection experiments**




**Supplementary figure 3: A significant number of genes up-regulated in two or more experiments**

**A significant number of genes up-regulated in two or more experiments**: 1018 genes were significantly (*P* < 0.05) up-regulated in two or more of the 11 HeLa transfection experiments. The null distribution for this overlap was estimated by randomly sampling (N=10,000) an identical number of up-regulated genes in each experiment (see methods). The observed overlap of 1018 genes is highly significant (*P* < 0.0001). A significant fraction of genes up-regulated in two or more experiments have ARE over-representation (*P* < 8.3e-9, Fisher's exact test, *N* = 7756 genes expressed in HeLa cells).

## Correcting for higher order sequence composition bias

We analyzed the possibility that correlations of shorter words after miRNA transfections (i.e. 2-4mers) could explain the correlations observed for long words (i.e. 6-7mers). Word correlations were therefore also evaluated in the context of higher order sequence composition bias. Our method provides a straight-forward correction for such bias by measuring word overrepresentation in a sequence relative to expected word occurrences in *k*-mer shuffled sequences. We used the uShuffle algorithm which generates uniform random permutations while preserving exact *k*-let counts [2].

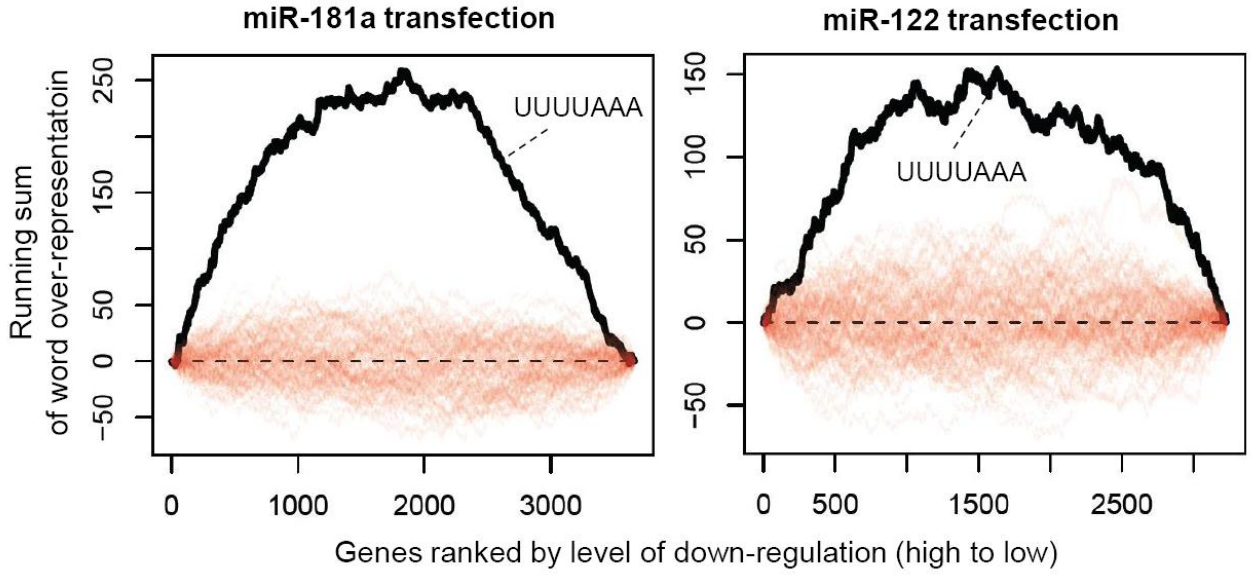
We evaluated word correlations for all 21.760 4, 5, 6 and 7-mer words using mono, di or tri-nucleotide shuffled sequences (*N*=5000) (see supplementary table S16-19 for di and tri-nucleotide results). In the 11 miRNA transfections in HeLa cells we specifically analyzed the correlations of the two top-correlating 7-mer words, UAUUUAU correlating with up-regulation and UUUUAAA correlating with down-regulation. Both words displayed a similar trend in this analysis (see table X and Y below). A) The words ranked in top-5 (*FDR*<$10^{-4}$) independent of the sequence composition background used. B) Di and tri-nucleotide shuffling slightly improved the overall word rank compared to mono-nucleotide shuffling. C) Using di and tri-nucleotide shuffling, the two 7-mer words ranked 1. Strikingly, no shorter words (4, 5 or 6-mers) had higher correlations.

| Sequence shuffling | Rank of UAUUUAU | Mean rank (N=21 760 words) of UAUUUAU across all experiments | UAUUUAU *FDR* estimated from permutation of word-correlation matrix (N=10.000) [3] |
|---|---|---|---|
| **1-mer (mono)** | 4 | 710 | <$10^{-4}$ |
| **2-mer (di)** | 1 | 243 | <$10^{-4}$ |
| **3-mer (tri)** | 1 | 247 | <$10^{-4}$ |

| Shuffling | Rank of UUUUAAA | Mean rank (N=21 760 words) of UUUUAAA across all experiments | UUUUAAA *FDR* estimated from word-correlation matrix permutation [3] |
|---|---|---|---|
| **1-mer (mono)** | 3 | 40 | <$10^{-4}$ |
| **2-mer (di)** | 1 | 53 | <$10^{-4}$ |
| **3-mer (tri)** | 1 | 62 | <$10^{-4}$ |

Overall, the results of this analysis demonstrate that the correlations of the UAUUUAU and UUUUAAA words cannot be explained by bias of shorter words in the transfection experiments. While the same analysis was not carried out for the other top-ranked correlating motifs identified in this study, it is reasonable to assume that the same result will hold for most of these motifs.

## UUUUAAA word correlations after two miRNA transfections

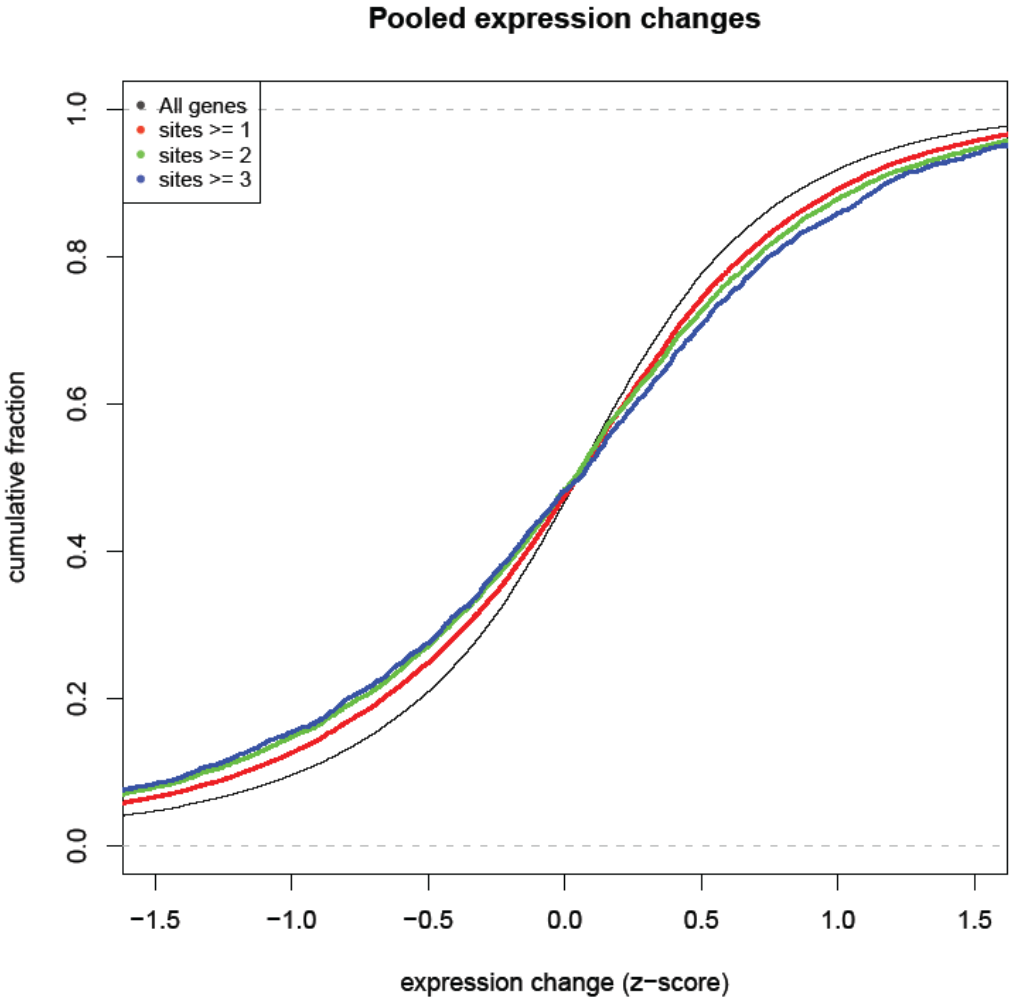



**Supplementary figure 4: UUUUAAA word correlations in single experiments**

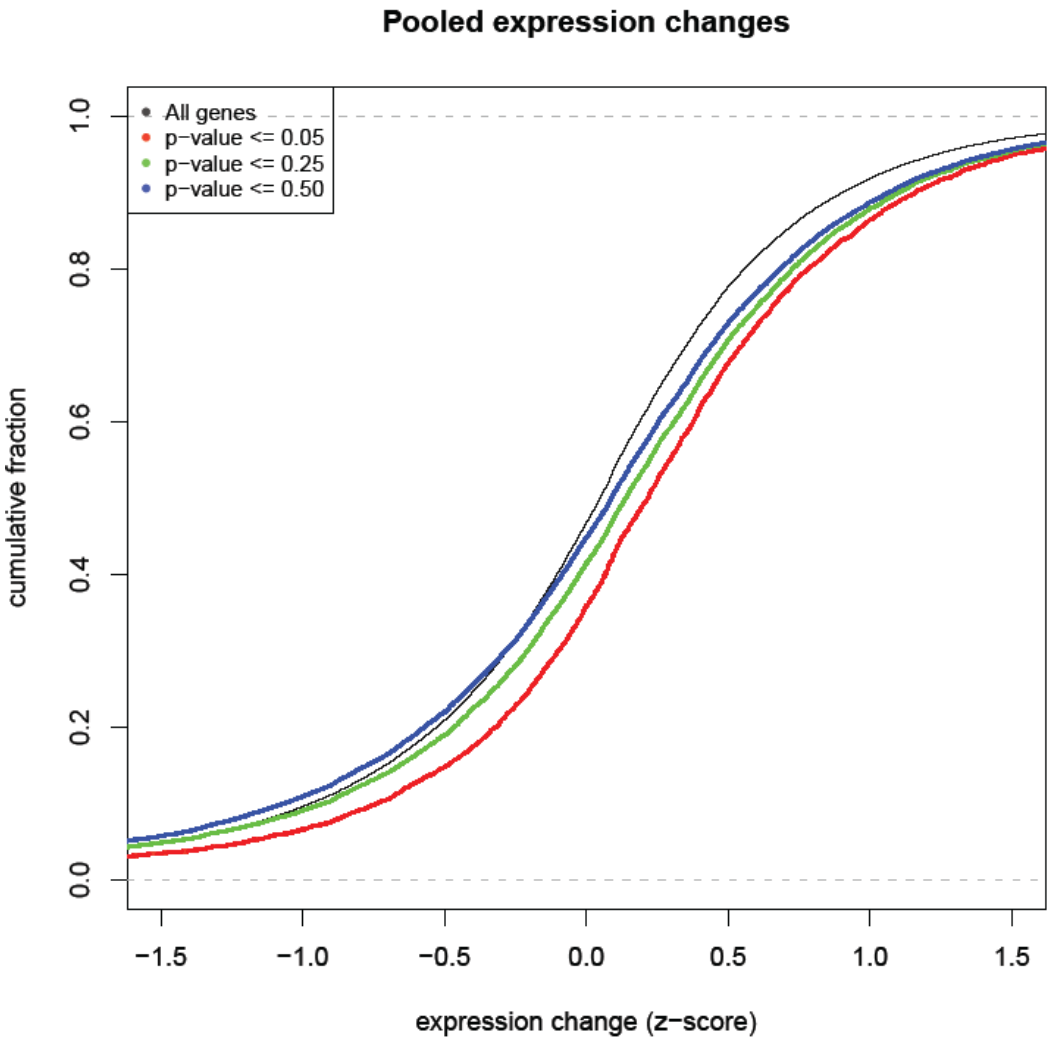
The UUUUAAA over-representation in 3'UTRs of up-regulated genes is plotted after transfection of two different miRNAs in HeLa cells. The black line is the UUUUAA running sum (left to right) of word over-representation scores in the ranked list of 3'UTRs (from high down-regulation to zero expression change), the red lines are running sums from 500 random permutations (100 shown in plot) of the UUUUAAA word scores (see methods). miR-181a is included as it has the highest UUUUAAA overrepresentation (rank 10, *FDR* < 1e-7) among the 11 different miRNA transfections in HeLa cells, while miR-122 transfection has the weakest effect (rank 108, *FDR* < 1e-7).

## Defining ARE genes by UAUUUAU overrepresentation

We analyzed different strategies to translate overrepresentation of ARE-stability motifs into a set of genes. Initially we tested genes defined by presence of the UAUUUAU word in the 3'UTR or using a threshold for the number of sites. While we found that genes with more occurrences of UAUUUAU were more up-regulated, we also found that the level of down-regulation correlated with the number UAUUUAU words. In contrast, UAUUUAU overrepresentation, measured by number of occurrences relative to shuffled sequences, more consistently defined genes up-regulated after transfections (Supplementary figure 6). One interpretation of this result is that UAUUUAU overrepresentation more accurately predicts functional ARE-sites that just using the site counts. Similarly, the observation that both genes up or down-regulated after miRNA transfections have long 3'UTRs (discussed below), implies that any word is more likely to be found in up/down regulated genes by chance compared to non-regulated genes.




**Supplementary figure 5: ARE genes defined by number of sites in 3'UTRs.**



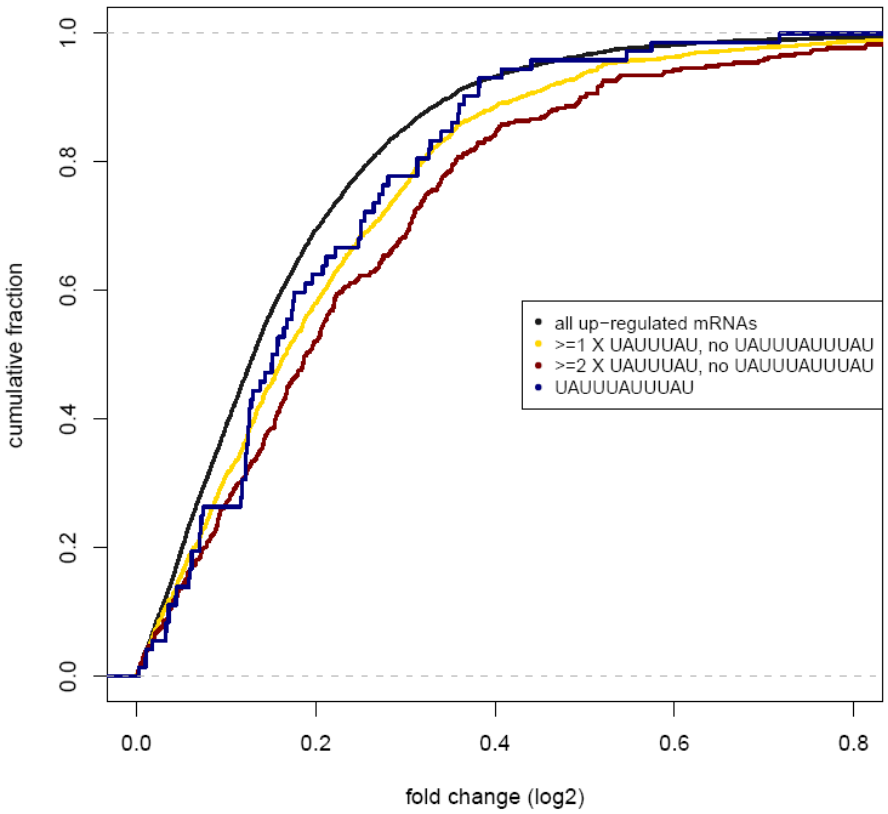

**Supplementary figure 6: ARE genes defined by UAUUUAU overrepresentation in 3'UTRs.**

## Analysis of the UAUUUAU sequence context in 3'UTRs of up regulated mRNAs

We investigated the possibility that the UAUUUAU overrepresentation was due to a longer over-represented motif. Analysis of miR-124 transfection (24 hours) showed that isolated word instances were more important for up-regulation than tandem repeat variants (Supplementary figure 7). Furthermore, we found that isolated UAUUUAU words were more important for up-regulation than tandem/overlapping repeats of UAUUUAU (Supplementary figure 7).

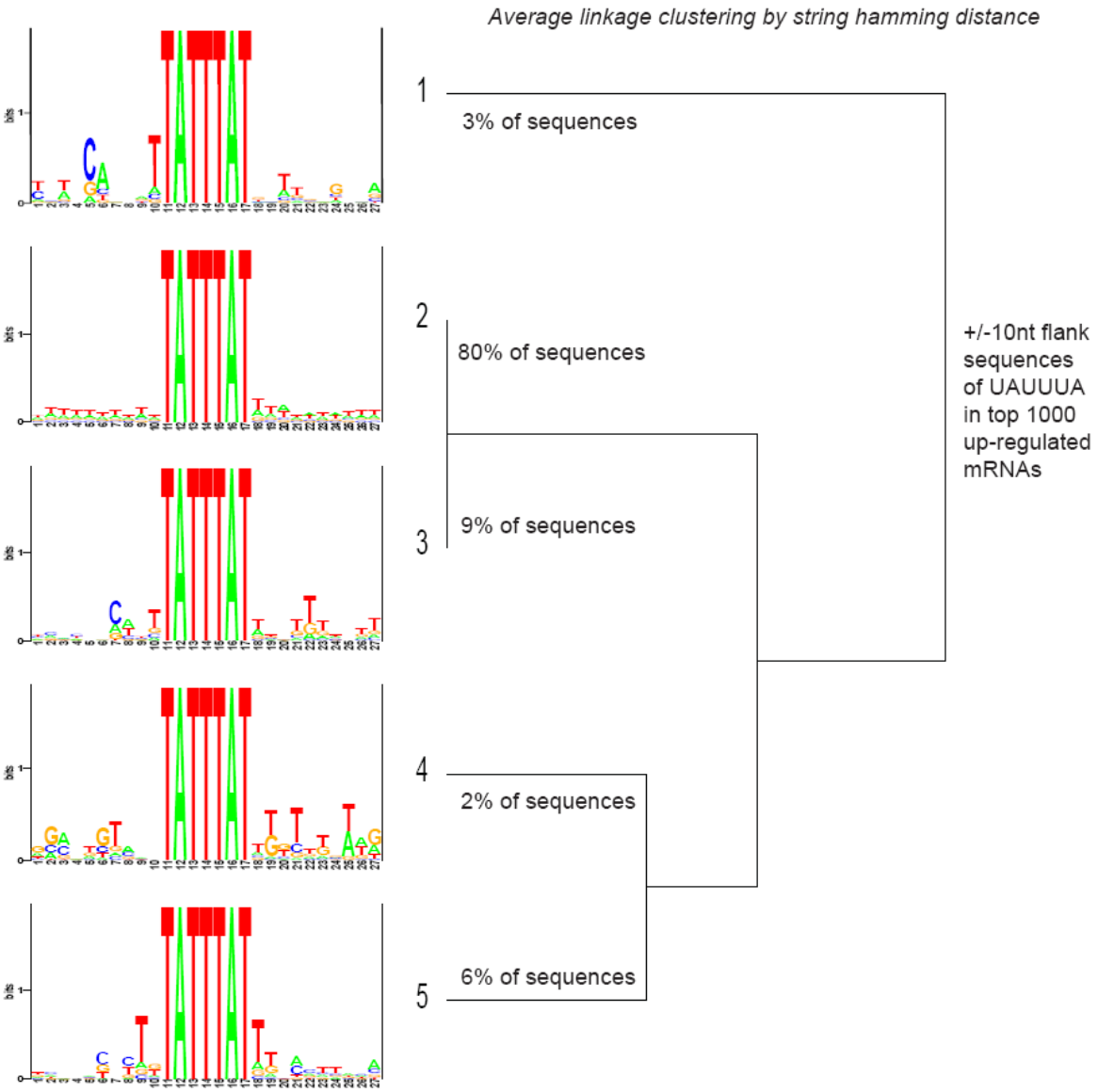
Next, we analyzed the sequences flanking the UAUUUAU word (+/- 10 nucleotides) in the top 1000 up-regulated mRNAs 24 hours after transfection of miR-124. Clustering of these sequences showed that a major fraction (80%) of the UAUUUAU words resided in AU rich regions but we found no cluster of sequences with noteworthy position specific nucleotide preferences outside the UAUUUAU word (Supplementary figure 8).

We then tested if there was a bias in the UAUUUAU flanking sequences specific to up-regulated mRNAs. We used the Two Sample Logo [4] method to calculate position specific nucleotide enrichment or depletion relative to UAUUUAU flanking sequences in down-regulated mRNAs. We found a weak U- enrichment bias extending 3 nucleotides upstream and downstream of UAUUUAU in the top 1000 up-regulated mRNAs 24 hours following miR-124 transfection in HepG2 cells (Supplementary figure 9, top). Moving further away from the core motif, a number of upstream positions show weak enrichment of G and depletion of A/U. The same analysis was done using the union of the top 200 up regulated mRNAs from each of the 11 different miRNA transfections in HeLa cells relative to the union of the top 200 down regulated mRNAs. The U enrichment bias is now only present immediately downstream of UAUUUAU but we find the G/U bias in the upstream region even more pronounced (Supplementary figure 9, bottom).



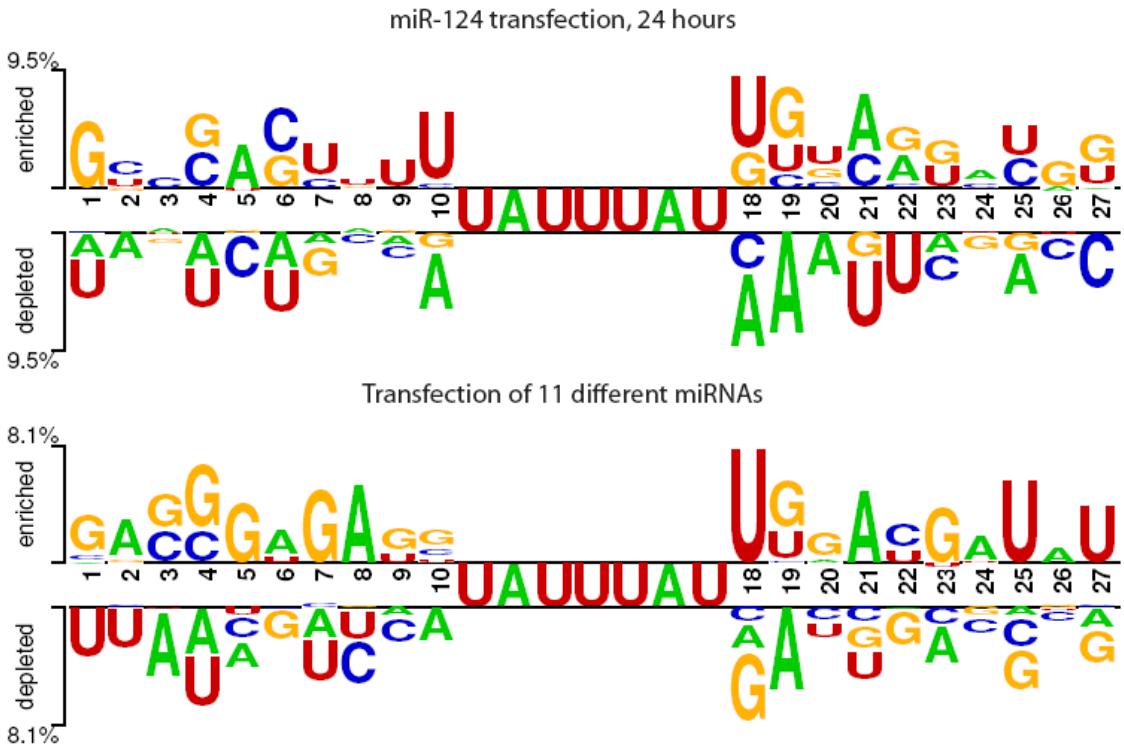

**Supplementary figure 7: Analysis of UAUUUAU tandem repeats**

**Isolated UAUUUAU words are more important than tandem repeats for up regulation:** We analyzed if UAUUUAU words occurring in tandem repeat structures contribute to up regulation of mRNAs. The black line shows the background cumulative expression change distribution of all up regulated mRNAs 24 hours after miR-124 transfection in HepG2 cells. Also shown is the cumulative expression change distribution of mRNAs with UAUUUAU in tandem repeat structure (blue), 3'UTRs containing >=1 isolated UAUUUAU word (yellow) and 3'UTRs containing >=2 isolated UAUUUAU words (red). While the tandem repeats show up regulation relative to background we observe that the set of mRNAs with 3'UTRs that only have isolated instances of UAUUUAU words have higher expression changes.




**Supplementary figure 8: Clustering of UAUUUAU flanking sequences**

**Flanking sequences of UAUUUAU in up regulated mRNAs:** We analyzed the sequences flanks of the UAUUUAU word in top 1000 upregulated mRNAs 24 hours after transfection of miR-124. 545 occurrences of UAUUUAU were found in 3'UTRs of the top 1000 up-regulated mRNAs. Flanks were extracted and the hamming distance was computed for all pairs of sequences. The sequences were clustered using average linkage hierarchical clustering and the hamming distance measure. The clustering dendrogram was cut to extract 5 clusters and sequence logos showing the information content at each position were computed [5].



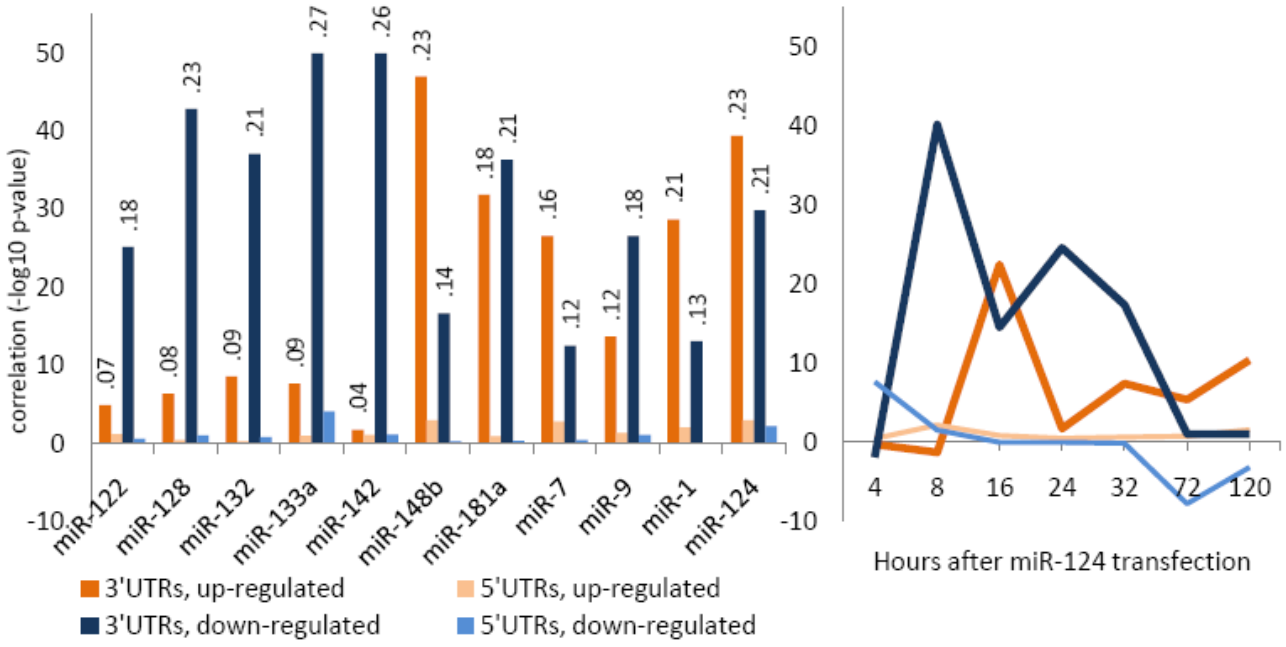

**Supplementary figure 9: Sequence bias of UAUUUAU flanks in upregulated genes**

**Bias in UAUUUAU flanks in up regulated mRNAs**: We analyzed the bias of the +/-10 nucleotide sequence flanks of UAUUUAU in up regulated mRNAs relative to down regulated mRNAs using Two Sample Logo [4]. In the top figure we compared sequence flanks in top 1000 up regulated mRNAs 24 hours following miR-124 transfection in HepG2 cells to the sequence flanks in the top 1000 down regulated mRNAs. In the bottom figure we did a combined analysis of the 11 different miRNA transfections in HeLa cells. Sequence flanks in the union of the top 200 up regulated mRNAs in each experiment were analyzed relative to flanks in the union of down regulated genes. Multiple positions show weak nucleotide preferences across all experiments: position 1 (G/U), 4 (GC/AU), 18 (U/G), 19 (GU/A), 21 (A/U), 23 (GU/CA), 25 (U) and 27 (U).

## mRNA expression change correlates with 3'UTR length after miRNA transfections

We investigated the sequence length of the 3'UTR as a general property that could relate to mRNA expression changes in a miRNA transfection experiment. The Pearson correlation coefficient between the 3' UTR length and absolute mRNA expression change was calculated for up and down-regulated genes separately for each transfection experiment. 5' UTR length correlations were included as a control. We found highly significant positive correlations ($0.12<r<0.27$, $1E\text{-}12<p<1E\text{-}50$) between 3'UTR length and absolute expression change of down-regulated genes in the 11 HeLa transfection experiments (Supplementary figure 10 left). Up-regulated genes also had significant correlations ($0.07<r<0.23$, $1E\text{-}5<p<1E\text{-}46$) in 10 of the 11 transfection experiments and in 3 of the experiments (mir-7, miR-1 and miR-124) the correlations were stronger than for down-regulated genes. The correlations observed for 5'UTRs were generally not significant. We observed the same trend in the miR-124 transfection time-series study where there was a strongly significant positive correlation between 3'UTR length and absolute expression change of down-regulated genes at the time points between 8 and 32 hours (Supplementary figure 10, right). The correlations were strikingly temporal: no correlation in down regulated genes before 4 hours and after 72 hours and a strong positive peak correlation for up-regulated mRNAs at 16 hours following transfection.

Few studies have directly analyzed the length of 3'UTRs in relation to miRNA regulation. Early computational analysis suggested that some mRNAs (denoted "anti-targets"), such as ribosomal genes, have 3'UTRs that are depleted in miRNA target sites and avoid targeting by having short 3'UTRs [6-8]. More recently it has been shown that proliferating cells avoid miRNA targeting by expressing shortened 3'UTR isoforms [9] and that human cortical gene expression variability correlate with the number of miRNA target sites in a 3'UTR and the 3'UTR length [10]. Even though numerous studies have analyzed the effects of over-expression of specific miRNAs, to the best of our knowledge, there are no reports on the relation between 3'UTR length and mRNA expression change in such an experiment. We find that a common characteristic in all experiments is that up and down-regulated mRNAs have significantly longer 3'UTRs than all other mRNAs and this correlation is not present for 5'UTRs. These results suggest that 3'UTR-mediated regulation of mRNA stability characterize a sizable fraction of the expression changes, both up and down-regulation, generated from miRNA transfections.




**Supplementary figure 10: 3'UTR length correlates with expression change after miRNA transfections**

**3'UTR length correlates with expression change after miRNA transfections:** The spearman correlation between 3'UTR (5'UTR shown as control) length and absolute expression change of up and down-regulated mRNAs respectively is plotted for 11 miRNA transfection experiments 24 hours following transfection in HeLa cells (left side of plot). The sign of the correlation significance (y-axis) corresponds to the sign of the computed correlation coefficient. The correlation coefficients are shown above each bar. The right side of the plot shows the same correlations calculated for each of the 7 time points following miR-124 transfection.

# References


1. Barreau C, Paillard L, Osborne HB: **AU-rich elements and associated factors: are there unifying principles?** *Nucleic Acids Res.* 2005, **33**:7138–715010.1093/nar/gki1012.

2. Jiang M, Anderson J, Gillespie J, Mayne M: **uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts**. *BMC Bioinformatics.* 2008, **9**:19210.1186/1471-2105-9-192.

3. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J: **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis**. *Bioinformatics* 2006, **22**:2825-282710.1093/bioinformatics/btl476.

4. Vacic V, Iakoucheva LM, Radivojac P: **Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments**. *Bioinformatics* 2006, **22**:1536-153710.1093/bioinformatics/btl151.

5. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**:6097-6100.

6. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM: **Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution**. *Cell* 2005, **123**:1133-1146.

7. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP: **The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution**. *Science* 2005, **310**:1817-182110.1126/science.1121158.

8. Sood P, Krek A, Zavolan M, Macino G, Rajewsky N: **Cell-type-specific signatures of microRNAs on target mRNA expression**. *Proc. Natl. Acad. Sci. U.S.A* 2006, **103**:2746-275110.1073/pnas.0511045103.

9. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB: **Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites**. *Science* 2008, **320**:1643-164710.1126/science.1155390.

10. Zhang R, Su B: **MicroRNA regulation and the variability of human cortical gene expression**. *Nucl. Acids Res.* 2008, :gkn43110.1093/nar/gkn431.