Supplementary material:
Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation

Jean Hausser, Afzal Pasha Syed, Biter Bilen, Mihaela Zavolan

Contents

1 Supplementary methods

1.1 Processing of quantitative proteomics and microarray data

Computational analysis of two-channel Agilent microarrays from Linsley et al. (2007) and Grimson et al. (2007) We downloaded the processed, differential expression data that was associated with the study of Linsley et al. (2007) from the Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`, accessions GSE6838 and GSE8501), and we selected those experiments where the same miRNA was tranfected both in HCT116 and DLD-1 cells. These involved microarray measurements 24h after transfection of let-7c, miR-103, miR-106b, miR-141, miR-15a, miR-16, miR-17, miR-192, miR-200a, miR-20a or miR-215 miRNAs (GEO accessions: GSM156546, GSM156550, GSM156545, GSM156549, GSM156543, GSM156576, GSM156532, GSM156541, GSM156534, GSM156542, GSM156580, GSM156544, GSM156547, GSM156551, GSM156548, GSM156552, GSM156553, GSM156555, GSM156554, GSM156556, GSM156557, GSM156558). We also downloaded the SOFT-formatted file of probe-to-transcript mappings provided by the authors and we selected only those probes that were associated with RefSeq transcript for subsequent analysis. We further used all the experimental data sets provided in the study of Grimson et al. (2007).

Computational analysis of one-channel Affymetrix microarrays from Selbach et al. (2008), Baek et al. (2008) (miR-223 dataset) and Hausser et al. (2009) We downloaded the CEL files of mRNA expression data sets reported in the study of Selbach et al. (2008) from `http://psilac.mdc-berlin.de/download/` and we further used the miR-124 and miR-7 over-expression and Ago2/EIF2C2 RIP-Chip data from our previous publication (Hausser et al., 2009) (GEO accession: GSE14537). We imported the CEL files into the R software (`www.R-project.org`) using the BioConductor affy package (Gentleman et al., 2004). Probe intensities were corrected for optical noise, adjusted for non-specific binding and quantile normalized with the gcRMA algorithm (Wu et al., 2004).

Per gene $\log_2$ fold change were obtained through the following procedure. We first fitted a lowess model of the probe $\log_2$ fold change using the probe AU content. We used this model to correct for the technical bias of AU content on probe-level $\log_2$ fold change reported by Elkon and Agami (2008). Subsequently, probe set-level $\log_2$ fold changes were defined as the median probe-level $\log_2$ fold change. Probe sets with more than 2 probes mapping ambiguously (more than 1 match) to the genome were discarded, as were probe sets that mapped to multiple genes. We then collected all remaining probe sets matching a given gene, and averaged their $\log_2$ fold changes to obtain an expression change per gene. For sequence analyses, we selected for each gene the longest RefSeq transcript with 5' UTR, CDS and 3' UTR annotation. Finally, we considered all genes for which at least one probeset was called present in the transfection experiments as expressed, and we only used these genes in subsequent analyses.

Computational analysis of two-channel human Agilent microarrays from Karginov et al. (2007) and Baek et al. (2008) We downloaded the mRNA expression data associated with the study of Baek et al. (2008) from the GEO database (accession GSE11968). The data (text file output of the Agilent scanner) of Karginov et al. (2007) was kindly provided to us by Ted Karginov. After extracting the rProcessedSignal, gProcessedSignal, LogRatio, rIsWellAboveBG, gIsWellAboveBG fields for each probe, keeping only probes for which both gIsWellAboveBG and rIsWellAboveBG flags were true in all experiments, we quantile-normalized the green and red channel intensities (rProcessedSignal and gProcessedSignal fields) of all experiments together and computed probe-level $\log_2$ fold changes. All probes mapping to multiple genes were then discaded, and we estimated the $\log_2$ fold change per gene as the average over the probes associated with that gene. Finally, for each gene we selected for further sequence analysis the longest RefSeq transcript with 5' UTR, CDS and 3' UTR annotation corresponding to that gene.

Computational analysis of SILAC assay from Baek et al. (2008) We downloaded the data provided by the authors in the supplementary material of the paper and used it without any other post-processing.

Computational analysis of pSILAC assay from Selbach et al. (2008) We downloaded the "all peptide evidence" flat file from `http://psilac.mdc-berlin.de/download/`, mapped all peptides in the pSILAC data set against the RefSeq Protein database (Jun, 20th 2011) with wu-blastp 2.0 (seed word length of 5), discarding alignments with gaps or with more than one mismatch. We further discarded peptides that mapped ambiguously to more than one protein and then computed per-protein $\log_2$ fold changes for all proteins credited with at least 3 peptides fold changes measurements.

1.2 Properties of miRNA binding sites

We computed 32 properties to quantify the structure and the sequence context of miRNA binding sites.

seed Eopen Structural accessibility of the miRNA seed-binding region, defined in terms of the energy necessary to open the secondary structure of the target in the region binding positions 1-8 of the miRNA.

site Eopen Structural accessibility of the miRNA-binding site, similarly defined in terms of the energy required to open the secondary structure of the target in a region of 20 nucleotides, anchored at the 3' end by the seed-complementary region.

seed Eduplex Energy of hybridizing the miRNA seed region to the mRNA.

target site Eduplex Energy of hybridization between positions 1 to 20 of the miRNA and the target site.

3' region Eduplex Energy of the hybrid formed between bases 9 to 20 of the miRNA and the 12 nucleotides upstream of the seed complementary region of the mRNA.

pos. 9–12 Eduplex, pos. 13–16 Eduplex, pos. 13–20 Eduplex The differences $\Delta G_u - \Delta G_c$ between the minimum binding free energy ΔG_u of the full mRNA-miRNA duplex and the binding free energy ΔG_c of the same duplex under the constraint that the nucleotides 9–12, 13–16 or 13-20 of miRNA are unpaired, respectively.

seed Einteraction The free energy of hybridization of the miRNA seed region defined as $\Delta G = \Delta G_o + \Delta G_h$, where ΔG_o is the energy required to open the secondary structure of the target in the seed-complementary region, and ΔG_h is the energy of the hybrid formed between the seed and the seed-complementary site.

target site Einteraction The free energy of hybridization of the miRNA to the target, similarly defined as $\Delta G = \Delta G_o + \Delta G_h$, where ΔG_o is the energy required to open the secondary structure of the target in the miRNA-binding region of 20 nucleotides anchored at the seed (as described above), and ΔG_h is the energy of the hybrid formed between the miRNA and the miRNA-complementary site.

Flanks A, C, G and U contents were defined as the proportions of A, C, G and U nucleotides within 50 nucleotides upstream and 50 nucleotides downstream of the miRNA binding site of 20 nucleotides, anchored downstream by the seed-matching region.

Domain A, C, G and U contents were defined as the proportions of A, C, G and U nucleotides within the CDS or 3' UTR harboring the miRNA binding site.

Transcript A, C, G and U contents were defined as the proportions of A, C, G, U and A+U nucleotides in the transcript harboring the miRNA binding site.

Transcript and domain length were obtained from the RefSeq sequence and annotation.

Distances to stop codon, domain start, domain stop and boundary are the number of nucleotides that separate the closest position of the 1-8 seed match from the 3' most position of the coding domain, domain start (i.e. start codon for CDS binding sites or first position of the 3' UTR for 3' UTR binding sites), domain stop (stop codon for CDS sites, 3' end of the mRNA for 3' UTR sites), and the smallest of the two latter distances.

Relative distances to domain start, domain stop and boundary were computed by dividing the distances to domain start, domain stop and boundary by the length of the domain (CDS or 3' UTR) in which the miRNA binding site is located.

ElMMo is the posterior probability that a seed complementary region is under evolutionary selective pressure described in Gaidatzis et al. (2007).

1.3 Expected number of genes co-targeted in the CDS and in the 3' UTR as a function of the total number of sites

Supplementary Figure 9 shows the number of transcripts in which a motif occurs in both CDS and 3' UTR, for all possible 8mers, as a function of the total number of occurrences of the 8mers in the transcriptome.

Here we derive a simple model for the number of transcripts that are expected to have a motif in both CDS and 3' UTRs (we call these transcripts co-targeted) under the hypothesis that motifs are distributed independently in CDS and 3' UTR. Suppose we have s motif occurrences to distribute along the CDS and 3' UTRs of $n = 18430$ representative transcripts. To simplify, we neglect 5' UTRs which are much shorter than CDS and 3' UTRs. On average, the fraction of genes co-targeted in the CDS and in the 3' UTR reflects the probability $P(c > 0 \wedge u > 0|n, s)$ that a motif occurs at least once in the CDS and at least once in the 3' UTR of the same gene, where c and u are the number of motif occurrences in the CDS and in the 3' UTR respectively. We now assume that the CDS and 3' UTR are targeted independently, which implies that we now have s sites to distribute on n CDS and n 3' UTRs. The average CDS is slightly longer than the average 3' UTR on average (1705 nucleotides vs 1334 nucleotides, see Supplementary Figure 19), meaning that we expect a fraction $\alpha = \frac{1705}{1705+1334} \simeq 56\%$ of the s sites to be located in the CDS. Therefore,

$$\begin{aligned}
P(c > 0 \wedge u > 0|n, s) &= P(c > 0|n, \alpha s)P(u > 0|n, (1-\alpha)s) && (1)\\
&= [1 - P(c = 0|n, \alpha s)]\,[1 - P(u = 0|n, (1-\alpha)s)] && (2)\\
&= \left[1 - \left(1 - \frac{1}{n}\right)^{\alpha s}\right]\left[1 - \left(1 - \frac{1}{n}\right)^{(1-\alpha)s}\right] && (3)
\end{aligned}$$

For the last step, we assumed that the number of motif occurrences in a single CDS or 3' UTR follows a binomial distribution. In other words, the probability $P(c = 0|n, \alpha s)$ that a given motif with a total of αs CDS occurrences does not appear in a given CDS is the probability not to draw an event of probability $\frac{1}{n}$ in a series of αs draws:

$$P(c = 0|n, \alpha s) = \left(1 - \frac{1}{n}\right)^{\alpha s}$$

Using the same reasoning for 3' UTRs, we obtain Equation 3.

1.4 Establishment of a stable cell line with inducible hsa-miR-124 expression

We PCR amplified the primary hsa-miR-124-2 (miRBase accession MI0000444) and cloned it into PGEM-T Easy vector (Promega). The insert was sequenced and subsequently cloned into pRTS-1 vector (Bornkamm et al., 2005) replacing the luciferase gene at SfiI restriction sites. We then established a stable cell line with miR-124-2-pRTS-1 (miR-124 cell line) by transfecting HEK293T cells with the miR-124-2-pRTS-1 plasmid and selecting for colonies after two weeks of culture in the presence of Hygromycin B (100μg/ml Calbiochem). Selected colonies were subsequently propagated in presence of Hygromycin B. The cell line was tested for miR-124-2 expression after the addition of doxycycline (1μg/ml) with Northern blot (Supplementary Figure S16). Small RNA Northern was performed as previously described (Pall and Hamilton, 2008) with a minor change: we used the conventional TBE buffer instead of a MOPS-NaOH buffer.

1.5 Reporter constructs for hsa-miR-124 CDS targets

We chose two miR-124 targets — *LPCAT3* (mRNA RefSeq ID: NM_005768.5) and *PSEN1* (mRNA RefSeq ID: NM_000021.3) — that fulfilled the following criteria

- were isolated in the Ago2-CLIP experiment that Hafner et al. (2010) performed upon miR-124 transfection
- were located on an expressed mRNA according to the mRNA expression data from the same study. Crosslinked regions and mRNA expression data were download from the ClipZ server (Khorshid et al., 2011).
- the crosslinked region was located in the mRNA coding region.
- the crosslinked region contained a miR-124 site predicted to be under strong selection pressure by the ElMMo algorithm ($P > 0.75$, see Gaidatzis et al. (2007) and the present paper).

- the 3' end of the miR-124-seed complementary sequence was located less than 200nt away from the STOP codon of the CDS.

Two short CDS fragments containing the miR-124 binding site were PCR amplified and cloned into dual luciferase psiCHECK2 vector (Promega). The inserts were cloned in-frame immediately upstream of the stop codon of the renilla luciferase (Supplementary Figure S16, panels B, C and D), with the overlap extension PCR method employing four different sets of primers. To obtain the mutant constructs, we used again the overlap extension PCR method to introduce synonymous mutations in the miR-124 binding site.

Primers for WT constructs

```
LPCAT3
for0: ttcatgggttactccatgact
rev0: tgtgaaaagggagacgagta
for1_inser: gcgtgctgaagaacgagcagcttggccacatcttcttcctg
rev1_ins: ggctcgagcgatcgcctagaattattccatcttctttaacttc
rev2_caggaagaagatgtggccaagctgctcgttcttcagcacgc
for3_gaagttaaagaagatggaataattctaggcgatcgctcgagc
for: agtcctgggacgagtggcctga
rev: ccccgagccgcctccgaatg

PSEN1
for0: gtgttctggttggtaaagcc
rev0: ttgttagatgtggacacagg
for1_inser: gcgtgctgaagaacgagcagaccatagcctgtttcgtagcca
rev1_ins: ggctcgagcgatcgcctagaactagatataaaattgatggaa
rev2_tggctacgaaacaggctatggtctgctcgttcttcagcacgc
for3_ttccatcaattttatatctagttctaggcgatcgctcgagc
for: agtcctgggacgagtggcctga
rev: ccccgagccgcctccgaatg
```

Primers for Mut constructs

```
LPCAT3
for: agcctactattcatactaccgtacattcacaaagcaatggtgcc
rev: ggcaccattgctttgtgaatgtacggtagtatgaatagtaggct

PSEN1
for: tagccatattaattggtttgtgtttgacattattactccttgcca
rev: tggcaaggagtaataatgtcaaacacaaaccaattaatatggcta
```

1.6 Luciferase assays

miR-124 cells were split in 24 well plates for both luciferase and qRT-PCR assays. Both mutant and wildtype psiCHECK2 (200ng) reporter constructs were transfected with lipofectamine 2000 (Invitrogen) reagent according to the manufacturers protocol. After 24hrs, we collected the cells and measured renilla and firefly activities with dual luciferase reporter assay system (Promega) and a luminometer (Centro LB960; Berthold Technologies). The firefly luciferase was used as internal control.

1.7 Quantitative real-time PCR

We extracted total RNA with the TRI reagent (Sigma) according to the manufacturers protocol, and then applied DNase digestion with RQ1DNase (Promega) followed by phenol-chloroform purification and cDNA synthesis with SuperScriptIII (invitrogen) reverse transcriptase according. We measured mRNA levels with the Step One Plus real-time PCR system (Applied Biosystems) employing Power SYBR Green PCR Master Mix (Applied Biosystems). Firefly expression was used as an internal control. Primers were designed such as to avoid the amplification of Plasmid DNA (primers sequence are available upon request).

1.8 Statistical analysis of luciferase and qPCR measurements upon hsa-miR-124 induction

Three independent biological replicates were performed (including cell seeding, RNA extraction, luciferase readouts), with three technical replicates for each biological replicate. For each measurement, we first normalized the data — qPCR CT and luciferase activity — to the firefly luciferase, so as to obtain $\log_2$ fold changes with respect to the internal control. For each independent biological replicate, we then computed the mean and Standard Error on the Mean (SEM) across technical replicates. Measurements from Dox-treated and -untreated cells of the same experiment were then compared to obtain mean $\log_2$ fold changes and SEM in luciferase activity and mRNA levels upon miR-124 induction. Finally, for each of the four constructs LPCAT3 WT, LPCAT3 Mut, PSEN1 WT and PSEN1 Mut, we computed $\log_2$ fold changes and SEM in luciferase activity and mRNA levels by aggregating the three independent experiments.

References

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

Bazzini, A. A., Lee, M. T., and Giraldez, A. J., 2012. Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science (New York, N.Y.)*, **336**(6078):233–237.

Bornkamm, G. W., Berens, C., Kuklik-Roos, C., Bechet, J.-M., Laux, G., Bachl, J., Korndoerfer, M., Schlee, M., Hölzel, M., Malamoussi, A., *et al.*, 2005. Stringent doxycycline-dependent control of gene activities using an episomal one-vector system. *Nucleic acids research*, **33**(16):e137.

Elkon, R. and Agami, R., 2008. Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. *PLoS computational biology*, **4**(10):e1000189.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M., 2007. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, **8**(1):69.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.*, 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10):R80.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F., 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, **312**(5770):75–9.

Grimson, A., Farh, K., Johnston, W., Garrett-Engele, P., Lim, L. P., and Bartel, D. P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, **27**(1):91–105.

Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P., 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**(7308):835–840.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-c., Munschauer, M., *et al.*, 2010. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, **141**(1):129–141.

Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., and Zavolan, M., 2009. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome research*, **19**(11):2009–20.

Karginov, F. V., Conaco, C., Xuan, Z., Schmidt, B. H., Parker, J. S., Mandel, G., and Hannon, G. J., 2007. A biochemical approach to identifying microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(49):19291–6.

Khorshid, M., Rodak, C., and Zavolan, M., 2011. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic acids research*, **39**(Database issue):D245–52.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M., 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, **8**(7):559–564.

Linsley, P., Schelter, J. M., Burchard, J., Kibukawa, M., Martin, M., Bartz, S., Johnson, J., Cummins, J., Raymond, C., Dai, H., *et al.*, 2007. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Molecular and Cellular Biology*, **27**(6):2240.

Pall, G. S. and Hamilton, A. J., 2008. Improved northern blot method for enhanced detection of small RNA. *Nature protocols*, **3**(6):1077–84.

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

Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., and Spencer, F., 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, **99**(468):909–917.

2 Supplementary figures

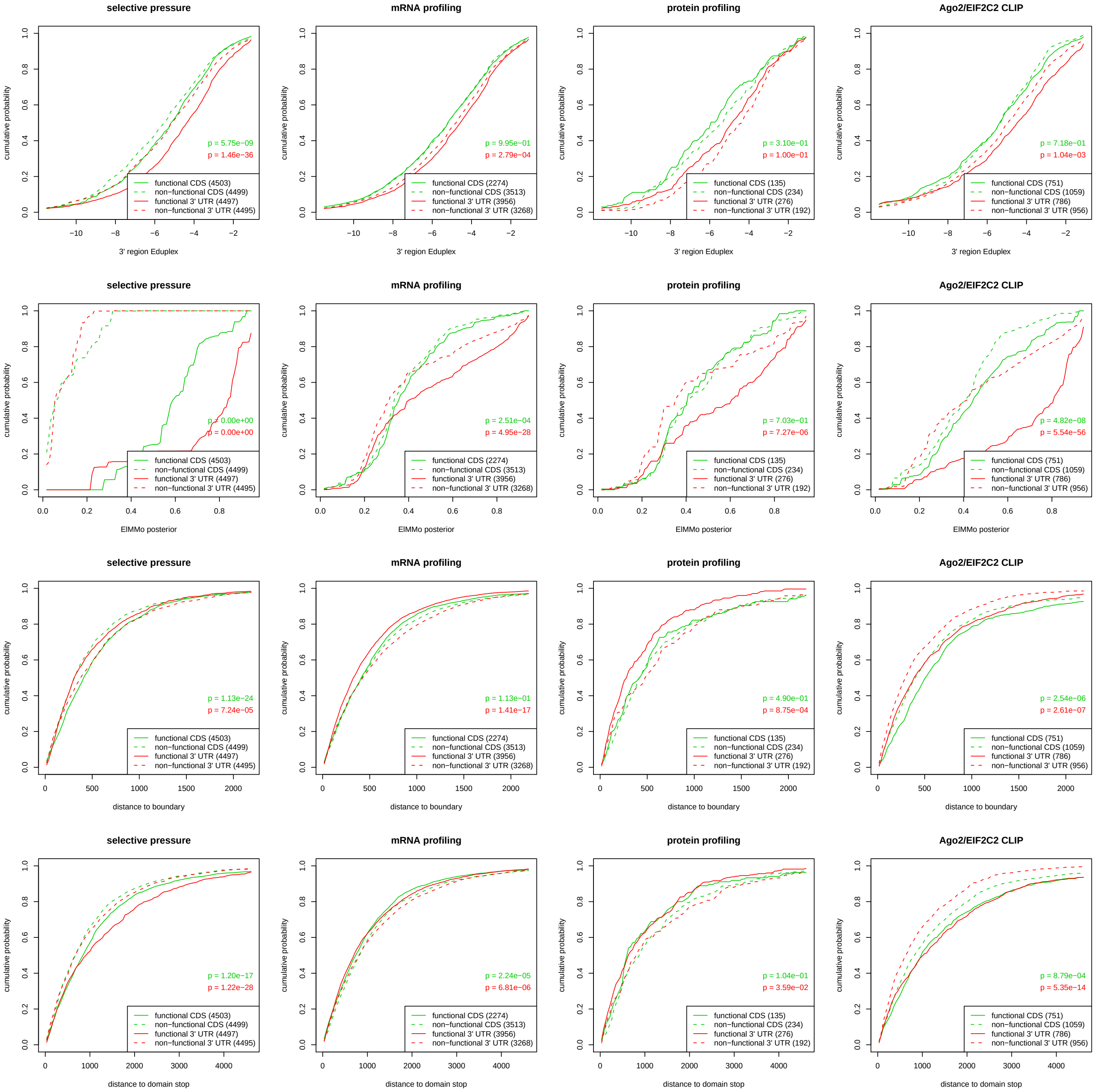

Figure 1: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3' UTR, part 1/7. Sites were obtained through comparative genomics analyses, mRNA and protein profiling upon miRNA transfection, and Ago2/EIF2C2 CLIP experiments (columns). Shown are the empirical cumulative density functions of different binding site properties (rows), analyzed separately for CDS (green) and 3' UTR (red) sites. Full and dotted lines correspond to functional and non-functional binding sites. Finally, the reported p-values compare the distribution of binding site properties in functional sites vs non-functional sites using Wilcoxon's rank sum test (CDS sites in green, 3' UTR sites in red).

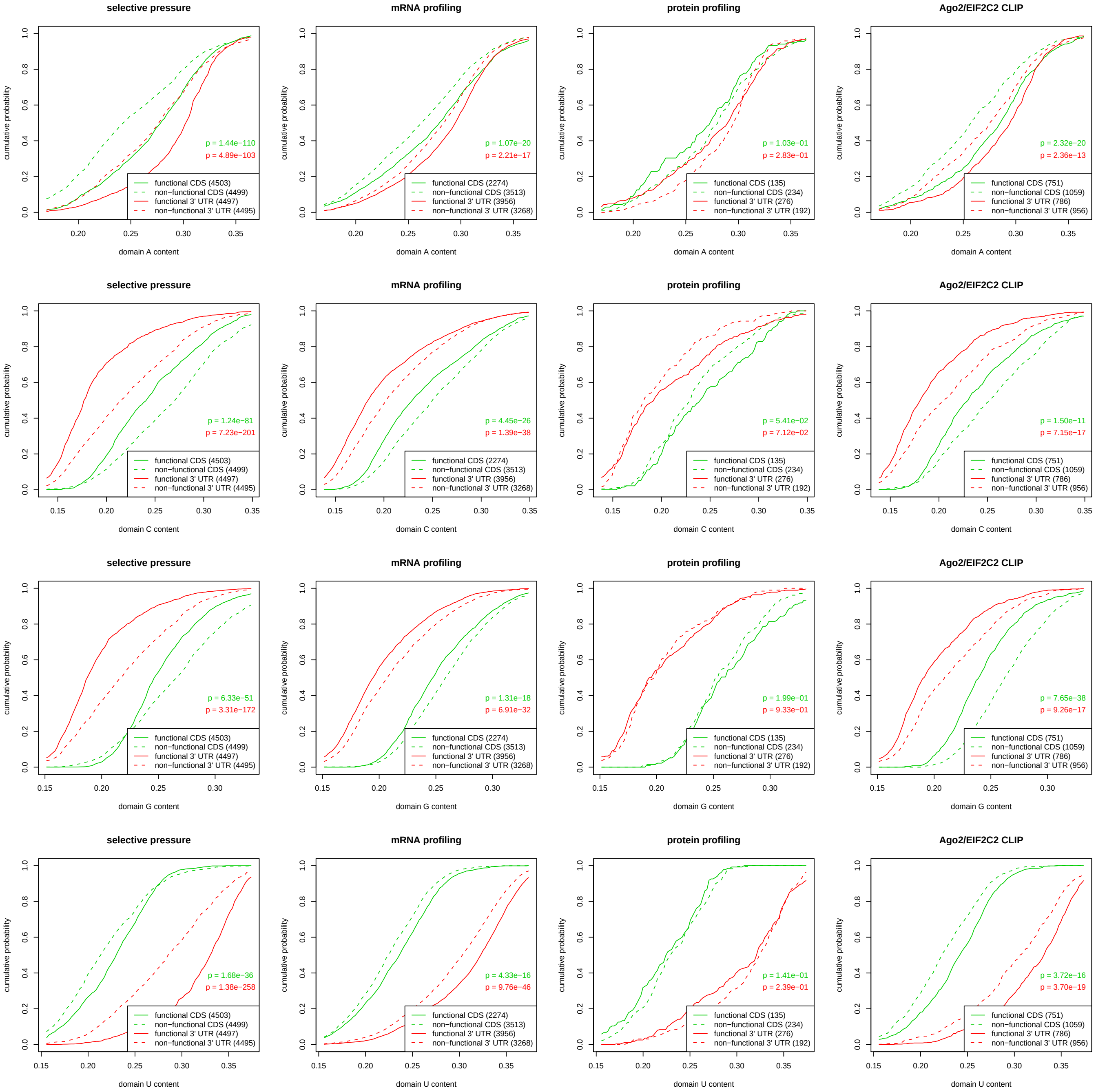

Figure 2: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3' UTR, part 2/7.

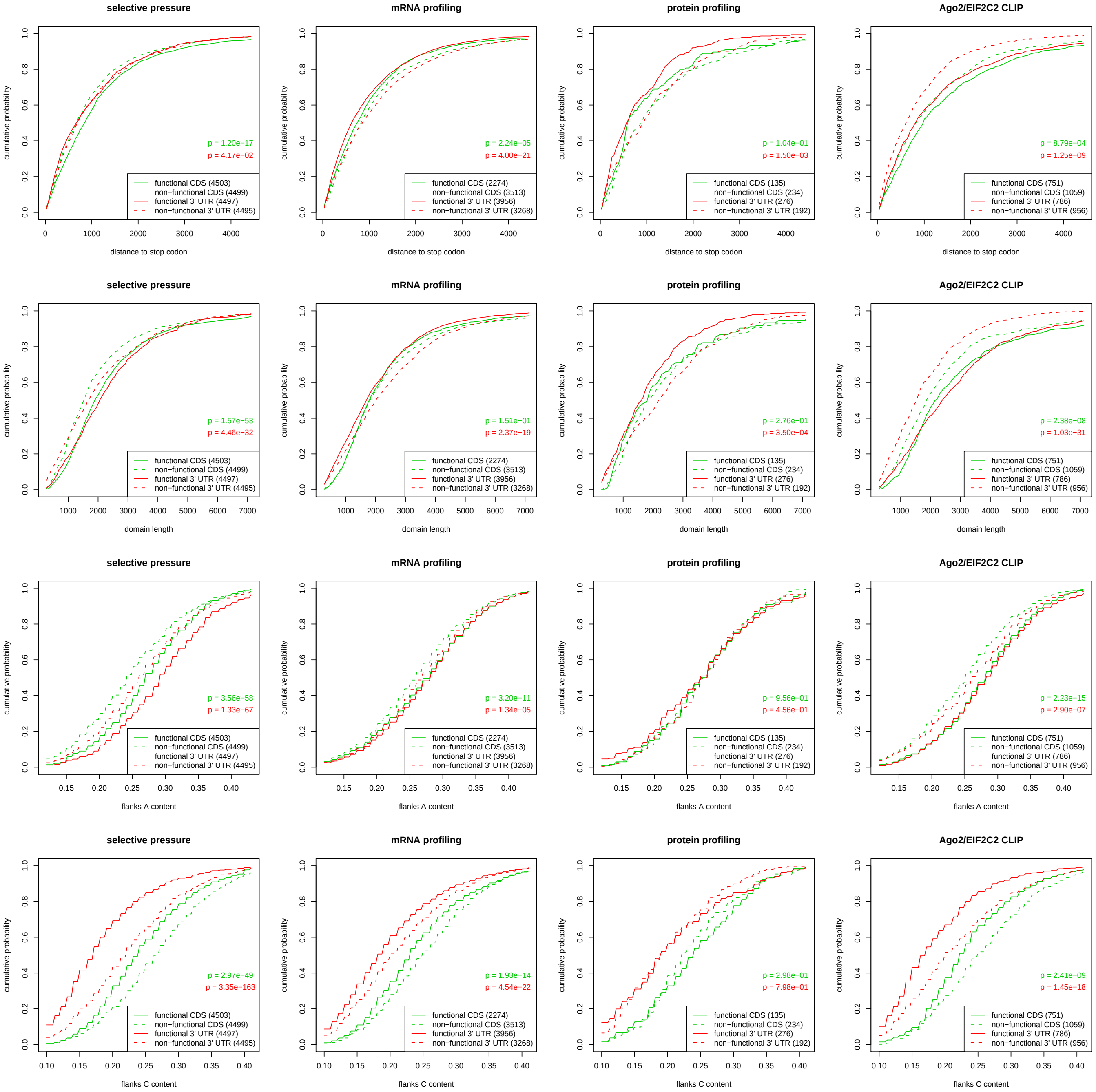

Figure 3: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3' UTR, part 3/7.

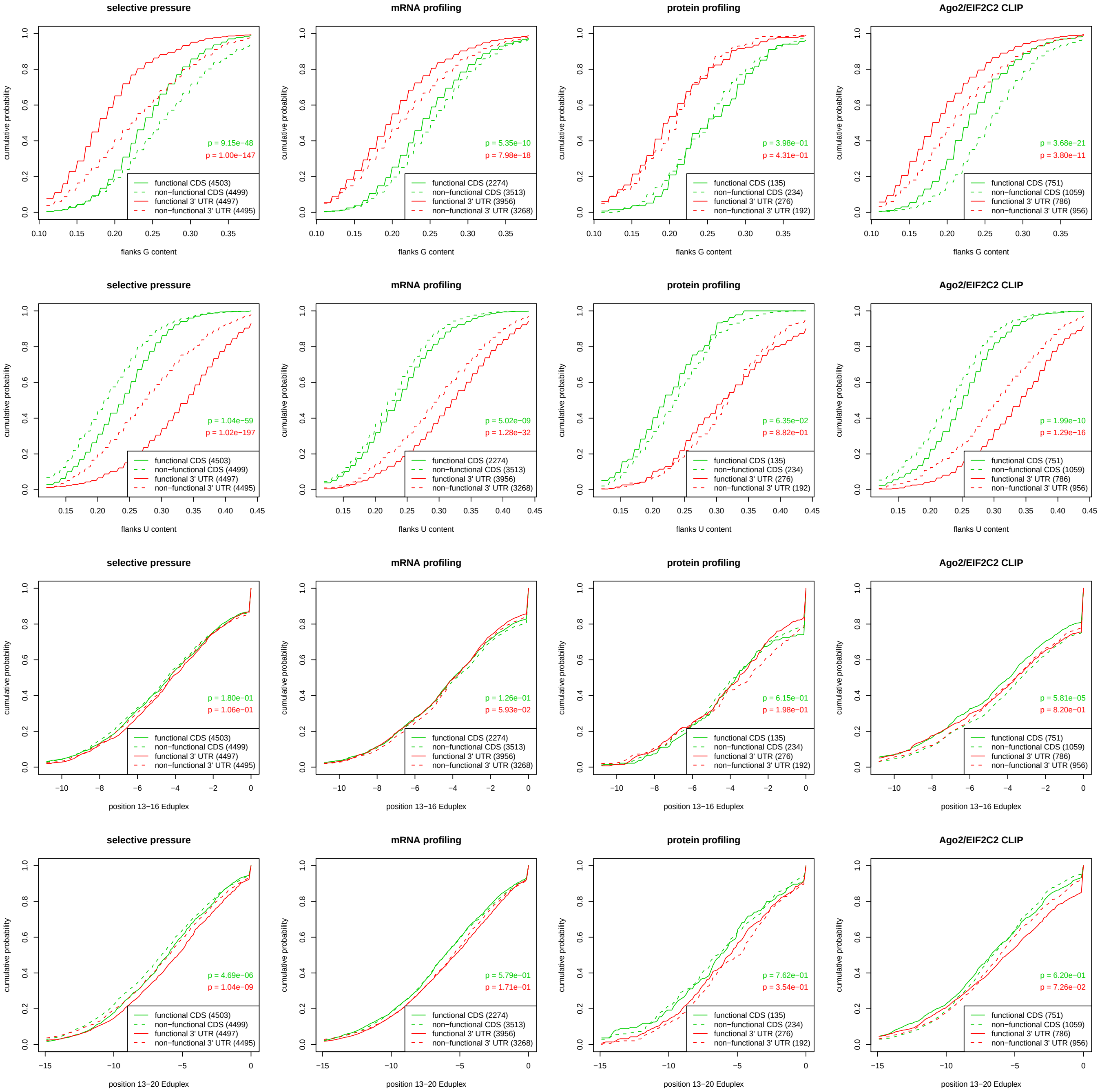

Figure 4: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3' UTR, part 4/7.

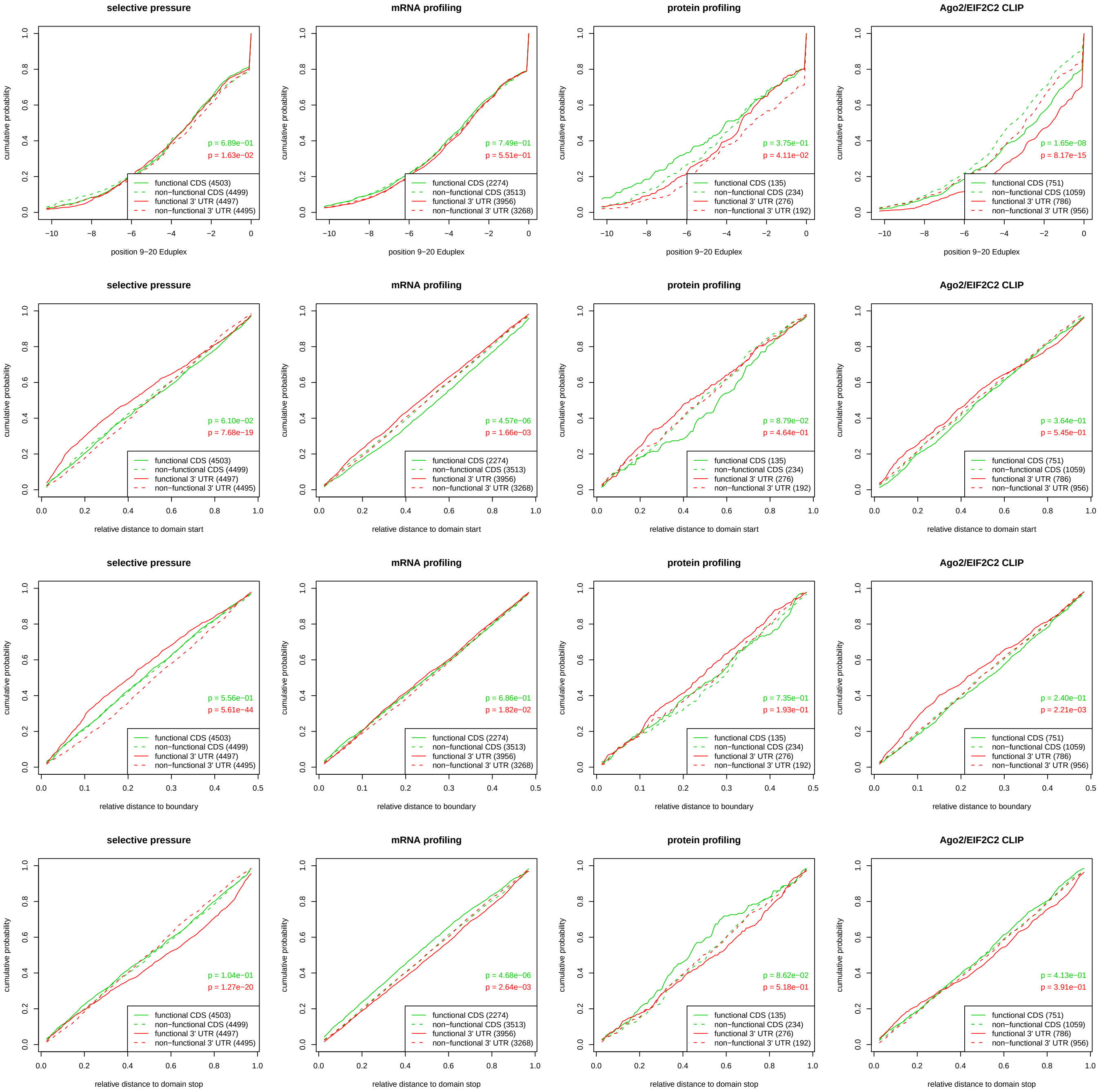

Figure 5: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3' UTR, part 5/7.

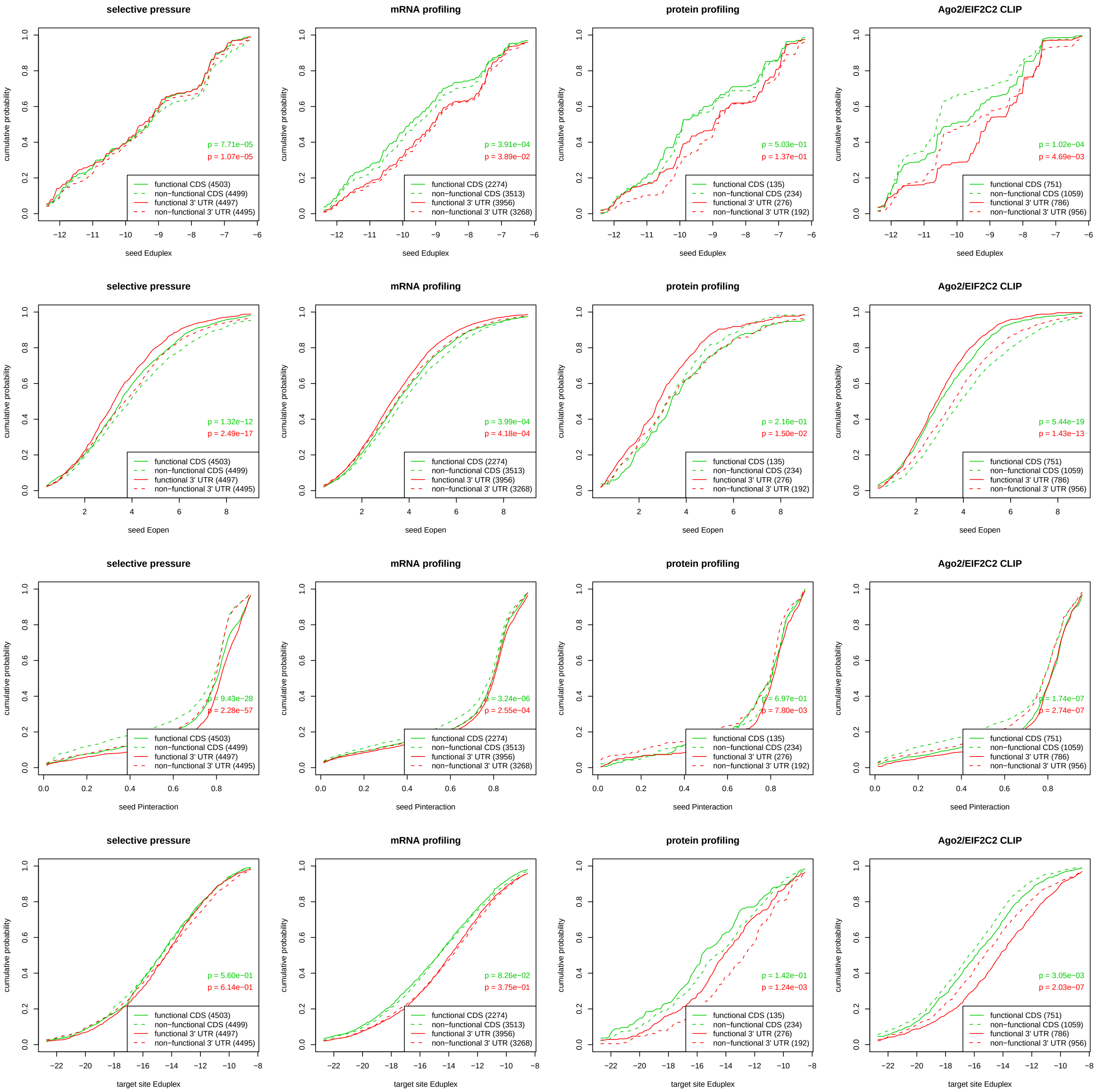

Figure 6: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3’ UTR, part 6/7.

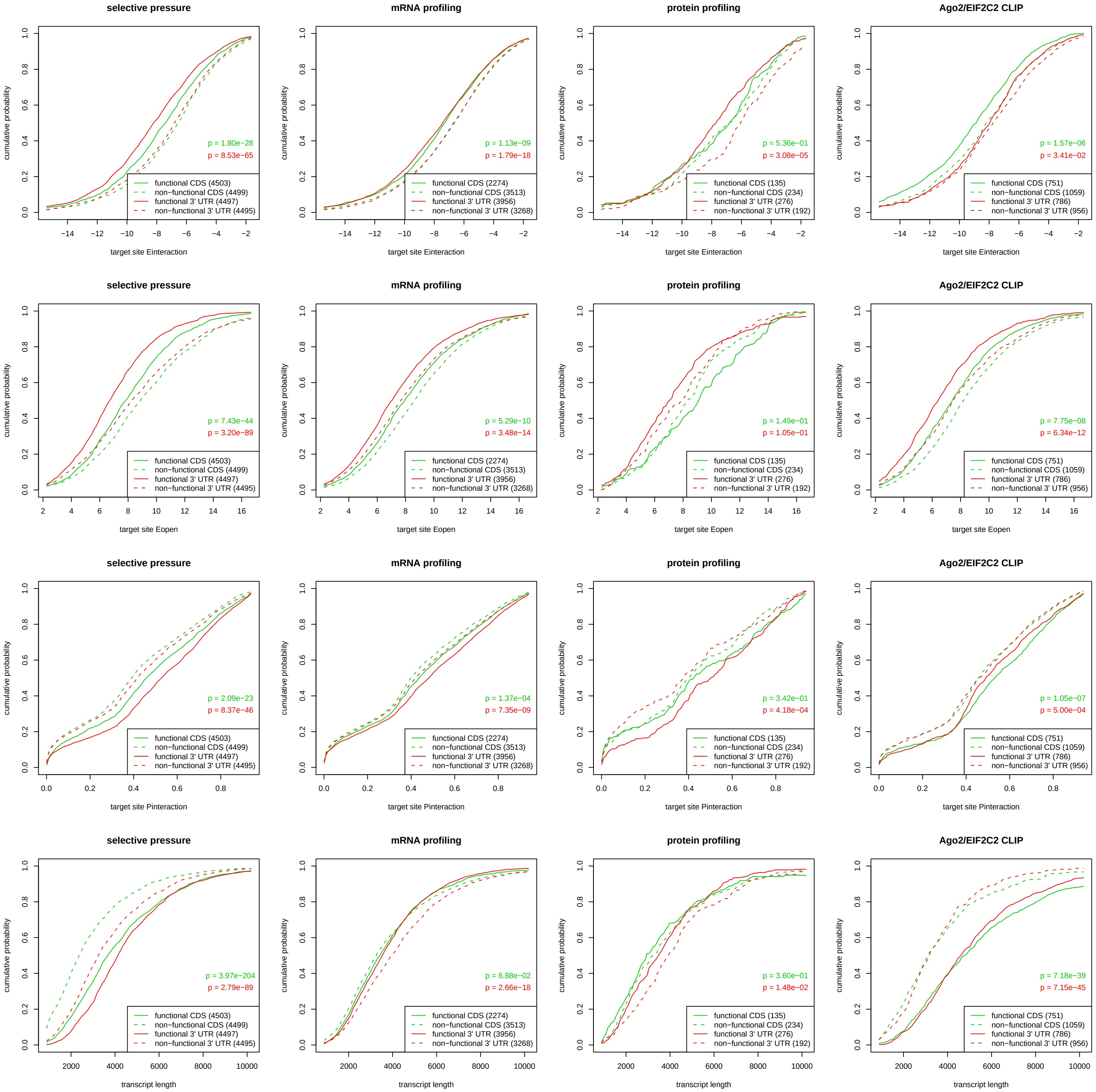

Figure 7: Comparison of properties of functional and non-functional miRNA-complementary sites located in the CDS and in the 3' UTR, part 7/7.

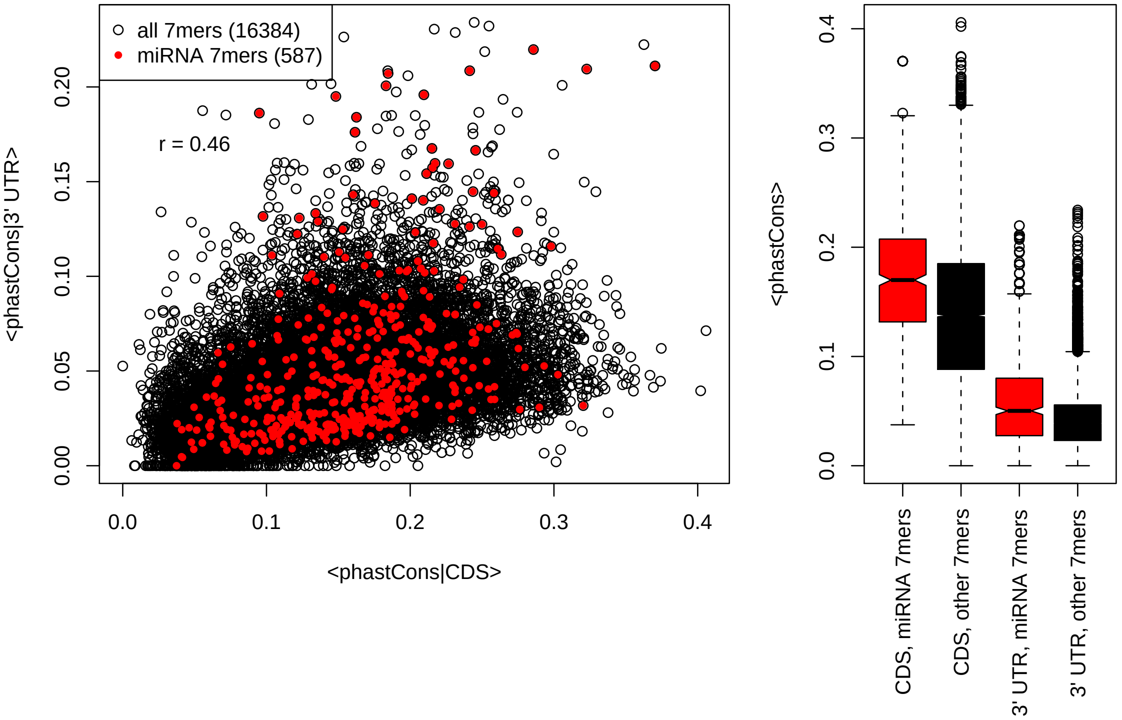

Figure 8: Left panel: Positive correlation between phastCons conservation scores of CDS- and 3' UTR-located occurrences of individual 7mers. Each dot represents a 7mer, with 7mers that are complementary to the 2-8 positions of known miRNAs shown in red and other 7mers shown in black. The Pearson correlation for all 7mers is reported on the figure. Right panel: 7mers that are complementary to miRNAs are significantly more conserved than those that are not ($p < 10^{-15}$ in Wilcoxon's rank sum test for CDS as well as 3' UTR). The box-plots summarize average phastCons scores of miRNA-related 7mers (red) and non-miRNA-related 7mers (black).

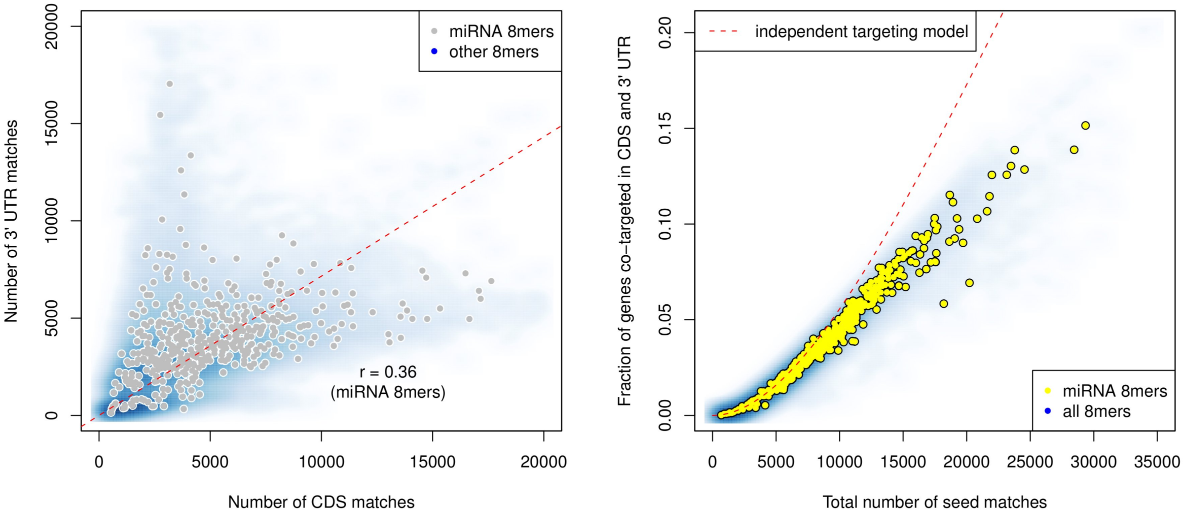

Figure 9: Left panel: Scatter plot of the number of occurrences of individual 8mers in CDS versus 3' UTRs of transcripts. 8mers that are complementary to miRNA seed regions are colored in grey while the density of the scatter for all 8mers is shown in blue. Right panel: The number of transcripts that have matches for 8mer motifs in both the CDS and 3' UTR as a function of the total number of matches for these motifs across the transcriptome. Only one representative mRNA was used per gene. This was the longest RefSeq transcript with annotated 5' UTR, CDS snd 3' UTR regions associated with the gene. An 8mer motif was considered to have a match in a transcript if the transcript contained the reverse complement of positions 1-7, 2-8, or 1-8 of the 8mer motif. 8mers corresponding to miRNA seed regions are colored in yellow while the density of the scatter for all 8mers appears in blue. The number of representative transcripts that can be simultaneously targeted in the CDS and in the 3' UTR correlates very well with the frequency of the miRNA seed match in the transcriptome ($r = .93$) because the more frequent a motif is, the likelier it is that a transcript contains the motif both in the CDS and in the 3' UTR. For 8mers of low frequency (less than 10'000 occurrences), the number of transcripts in which the motif occurs in both CDS and 3' UTR can be explained by a model that assumes that CDS and 3' UTR are targeted independently (red dashed curve, see subsection 1.3 below). Very abundant motifs (with more than 10'000 occurrences in the transcriptome) have a very specific composition and tend to occur much more frequently in one type of region (CDS and 3' UTR), thus the number of transcripts in which the motif occurs in both regions is lower than expected. A/U-rich occur with much higher frequency in 3' UTRs while CDS are more enriched in motifs such as AAGAA, AGAAG, CAGCA, GAAGA, GAGGA, GCUGC, GGAGG, UGAAG and AGCAG. Consistently, there is a fork in the left panel of the Figure for motifs with 5'000 CDS and 5'000 3' UTR occurrences and more. The same pattern is also visible when performing the analysis with ElMMo-predicted binding sites (Figure 2A in the main text).

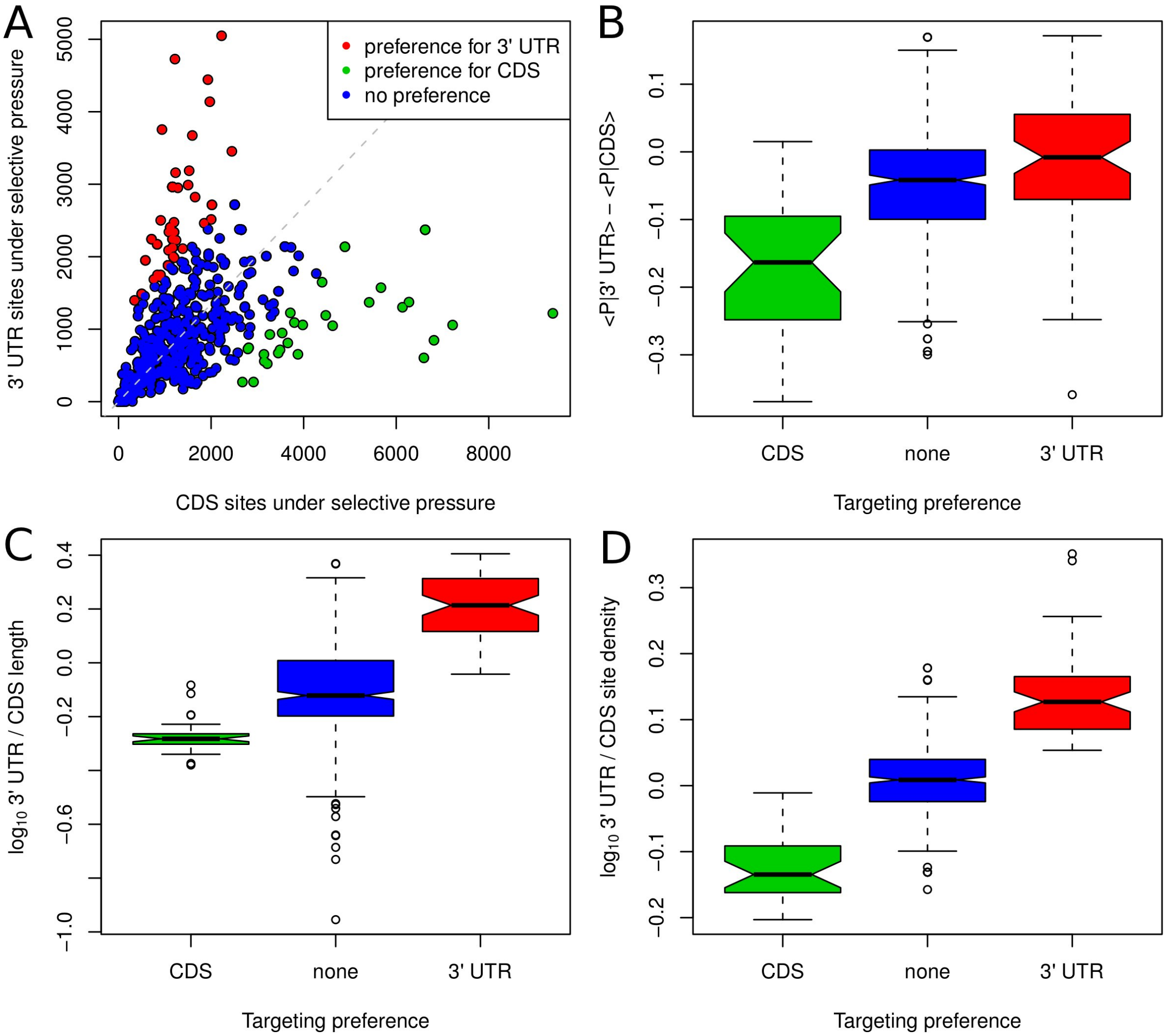

Figure 10: The preference of different miRNAs to target the CDS or the 3' UTR stems from a combination of factors. A: Scatter plot of the estimated number of binding sites under selection pressure in the CDS against the estimated number of binding sites under selection pressure in the 3' UTR. Each dot represents one miRNA and the number of binding sites under selection pressure was obtained by summing the ElMMo posteriors of all predicted target sites. The gray dashed line represents the scaling between the number of CDS and 3' UTR sites, defined as the line that goes through the origin and maximizes the projected variance. The preference of each miRNA to target a specific region (CDS or 3' UTR) can be quantified by the signed distance to the dashed gray line. The miRNAs for which this distance is largest and positive have a preference to target 3' UTRs and appear as red dots while miRNAs with largest negative distance to the line mostly target the CDS and appear as green dots. The remaining miRNAs appear as blue dots. The preference for CDS (or conversely, for 3' UTR) targeting by a miRNA can be influenced by three means. First, the average ElMMo posterior per site may be higher in the CDS than in the 3' UTR (see panel B). Second, the miRNA may target mRNAs with long CDS and short 3' UTRs, leading to a larger number of CDS sites (panel C). Finally, the density of sites may be higher in the CDS compared to 3' UTR (panel D). As shown by panels B, C and D, all three factors appear to contribute to the CDS vs 3' UTR targeting preference. B: Distribution of the difference of average 3' UTR and CDS ElMMo posterior of sites for miRNAs that preferentially target the CDS, 3' UTRs, or without targeting preference. C: 3' UTR to CDS length ratio of mRNAs carrying binding sites to miRNAs that preferentially target the CDS, 3' UTRs, or without targeting preference. D: $\log_{10}$ 3' UTR to CDS site density (defined as the number of sites per length unit) ratio of binding sites to miRNAs that preferentially target the CDS, the 3' UTR, or without targeting preference.

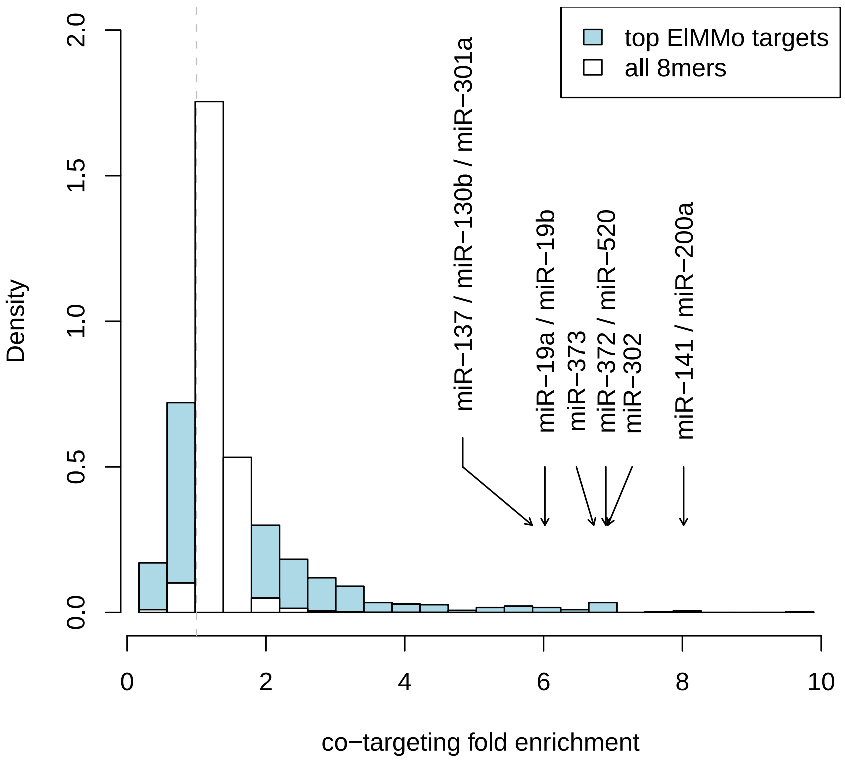

Figure 11: Fold enrichment in transcripts that are targeted by a given miRNA in both CDS and 3' UTR relative to what would be expected if the sites were independently distributed (see Methods in the main text). Blue bars represent fold enrichments obtained from the top 250 CDS- and top 250 3' UTR-located binding sites of known miRNAs predicted by ElMMo. As a comparison (white bars), we determined the distribution of fold enrichments in transcripts that contain 1-7, 2-8 or 1-8 matches to all 65536 possible 8mers in both CDS and 3' UTR.

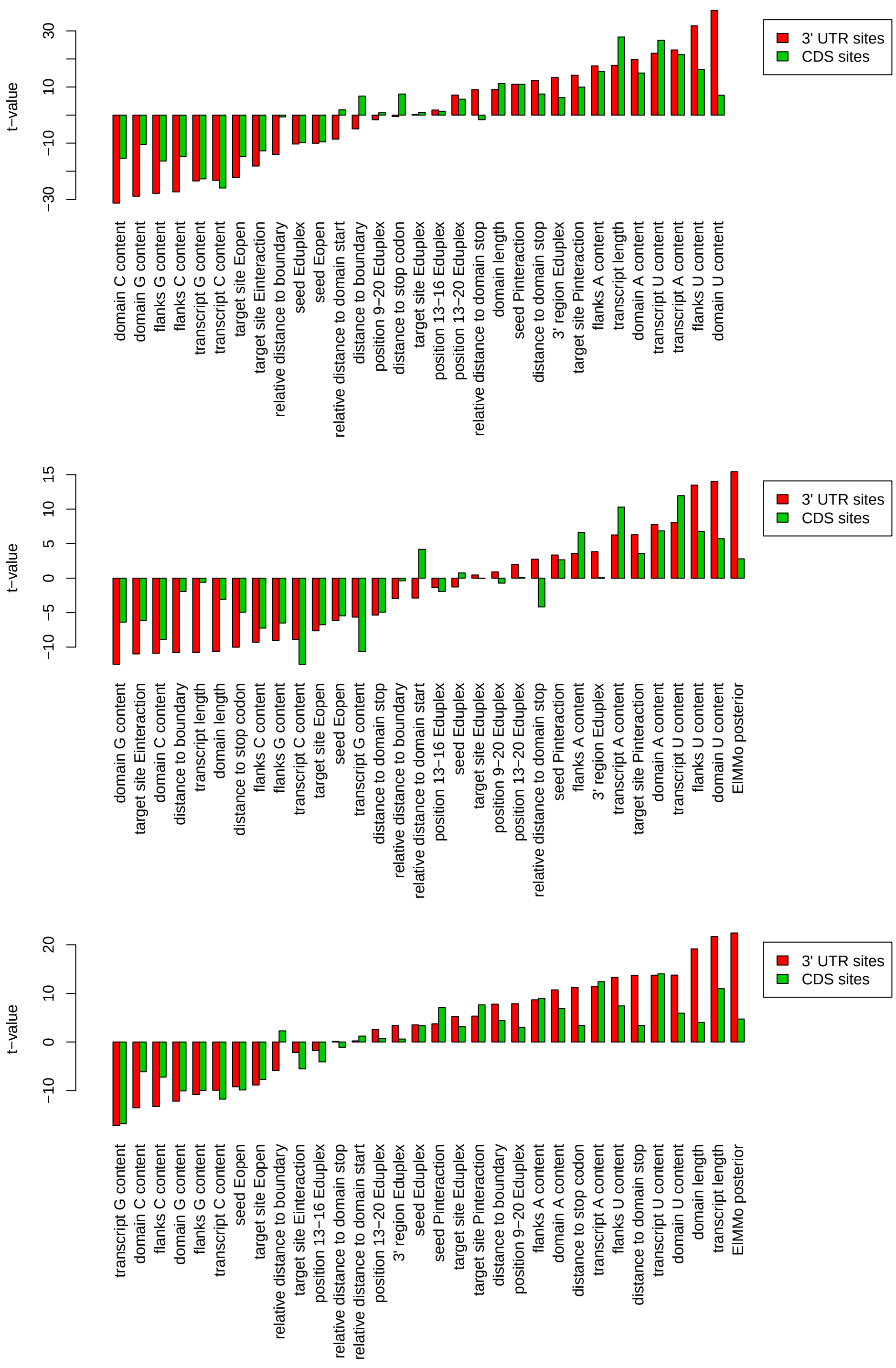

Figure 12: Binding sites in CDS and 3' UTRs share common sequence and structure properties. Plotted are t-values comparing the values taken by different sequence and structure properties in functional vs non-functional sites. Functional sites were defined either as sites inferred to be under evolutionary selection pressure by the ElMMo algorithm (top panel), sites effective in mRNA destabilization after miRNA transfection (middle panel), or sites that had strong enrichment in Ago2/EIF2C2 binding (bottom panel).

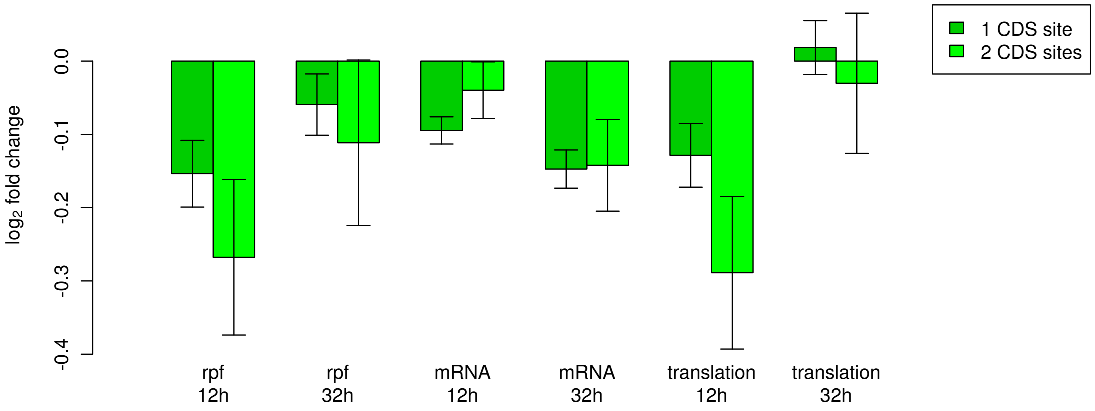

Figure 13: Changes in ribosome protected fragments (rpf), expression levels and translation of mRNAs with one or two seed matches in the CDS following the transfection of miR-155 and miR-1 by Guo et al. (2010).

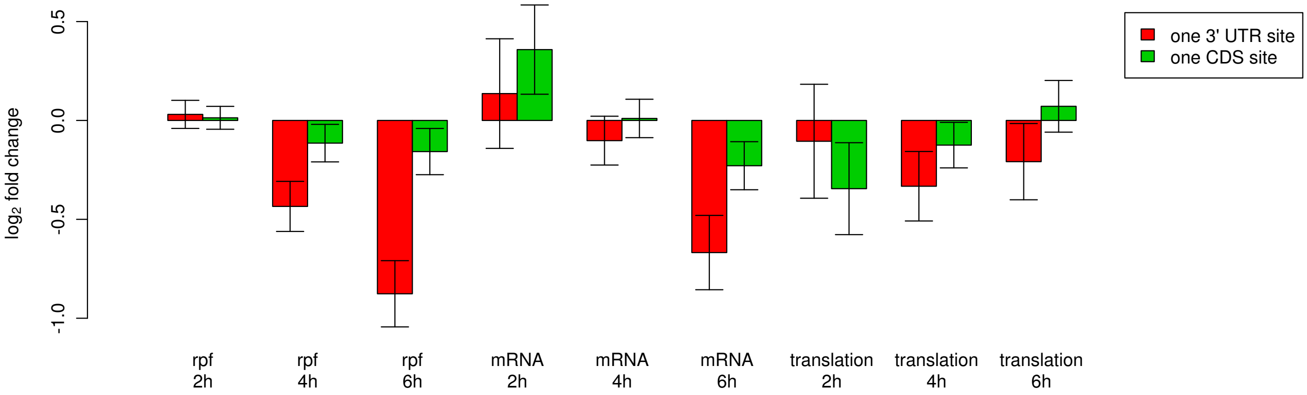

Figure 14: dre-miR-430 binding sites that are located in CDS transiently inhibit the translation during maternal-zygotic transition in zebrafish. dre-miR-430 is expressed at the onset of zygotic genome transcription and induces the clearance of maternal mRNAs Giraldez et al. (2006). The figure shows $\log_2$ fold changes in ribosome protected fragments (rpf) and mRNA levels (mRNA) 2, 4 and 6 hours post fertilization in wild-type zebrafish compared to Dicer knock-out, as determined by Bazzini et al. (2012). Changes in translation were estimated from the difference between changes in rpf and changes mRNA levels. mRNAs with precisely one seed match to dre-miR-430 in the CDS and no seed match in the 3' UTR were analyzed separately from mRNAs with precisely one seed match in the 3' UTR and no seed match in the CDS. Fold changes were determined relative to the average fold change of mRNAs with no seed matches.

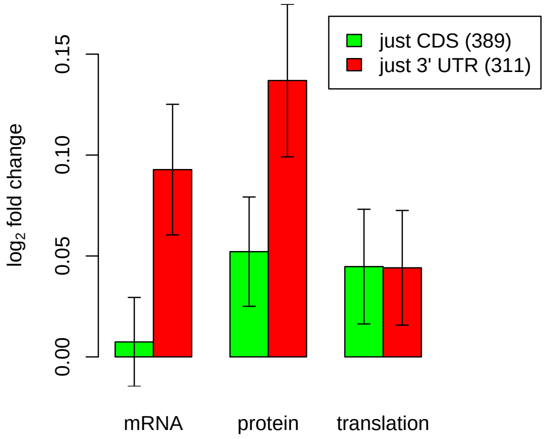

Figure 15: Translation of mRNAs carrying CDS-located binding sites is significantly up-regulated following mmu-miR-223 knock-out in mouse neutrophils. The figure shows $\log_2$ fold changes in mRNA and protein levels in mmu-miR-223 knock-out compared to wild type, computed from the microarray and proteomics measurements of Baek et al. (2008). Changes in translation were estimated from the difference between changes in protein levels and changes mRNA levels. mRNAs with precisely one seed match to mmu-miR-223 in the CDS and no seed match in the 3' UTR were analyzed separately from mRNAs with precisely one seed match in the 3' UTR and no seed match in the CDS. Fold changes were determined relative to the average fold change of mRNAs with no seed matches.

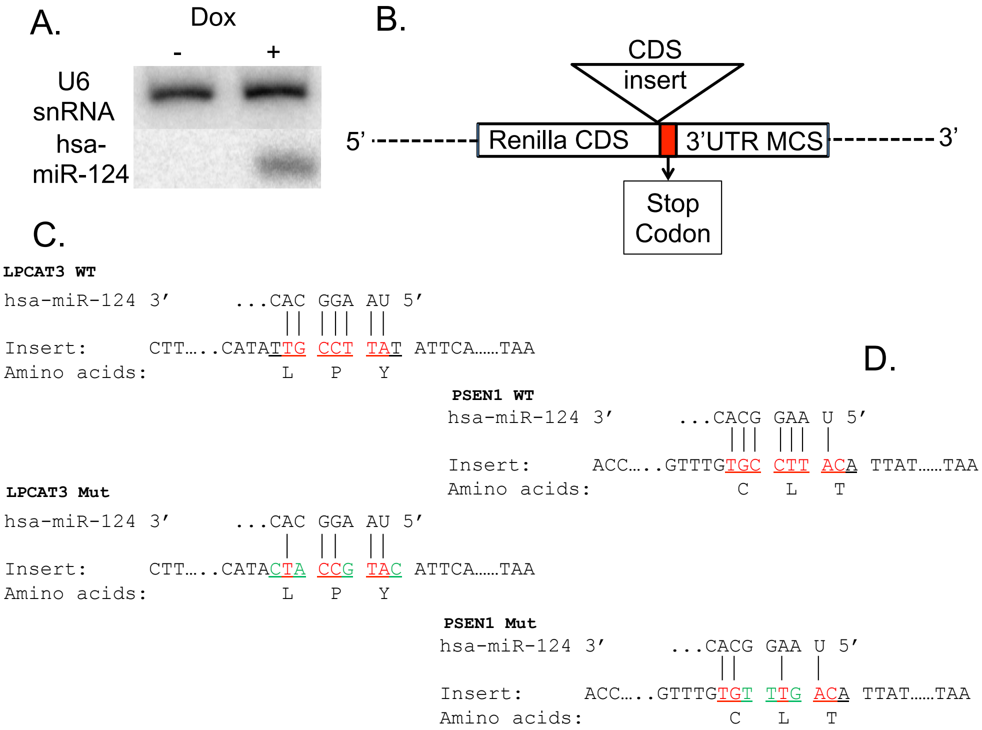
A.
Dox
-
+
U6 snRNA
hsa-miR-124
B.
CDS insert
5'
Renilla CDS
3'UTR MCS
3'
Stop Codon
C.
LPCAT3 WT
hsa-miR-124 3' ...CAC GGA AU 5'
Insert: CTT.....CATATTG CCT TAT ATTCA......TAA
Amino acids: L P Y
LPCAT3 Mut
hsa-miR-124 3' ...CAC GGA AU 5'
Insert: CTT.....CATACTA CCG TAC ATTCA......TAA
Amino acids: L P Y
D.
PSEN1 WT
hsa-miR-124 3' ...CACG GAA U 5'
Insert: ACC.....GTTTGTGC CTT ACA TTAT......TAA
Amino acids: C L T
PSEN1 Mut
hsa-miR-124 3' ...CACG GAA U 5'
Insert: ACC.....GTTTGTGT TTG ACA TTAT......TAA
Amino acids: C L T

Figure 16: Design of the experiment to measure the effect of miR-124 on protein and mRNA levels. A: hsa-miR-124 expression is induced by doxycyclin. Cells with an episomal hsa-miR-124 expression vector were split in a 24-well plate and induced with doxycycline (1μg/mL). After 24 hrs cell were collected and small RNA northern was performed. The same blot was hybridized with U6 snRNA probe as a internal control. B: Short inserts containg hsa-miR-124 binding sites from the *LPCAT3* or *PSEN1* genes were cloned upstream of the STOP codon of the Renilla luciferase. C: Constructs were generated using from the wildtype (WT) sequence of the hsa-miR-124-complementary site in *LPCAT3* or by mutating 4 positions in the seed complementary region. The mutations were designed such that they did not change the encoded amino acid sequence. D: Idem for *PSEN1*.

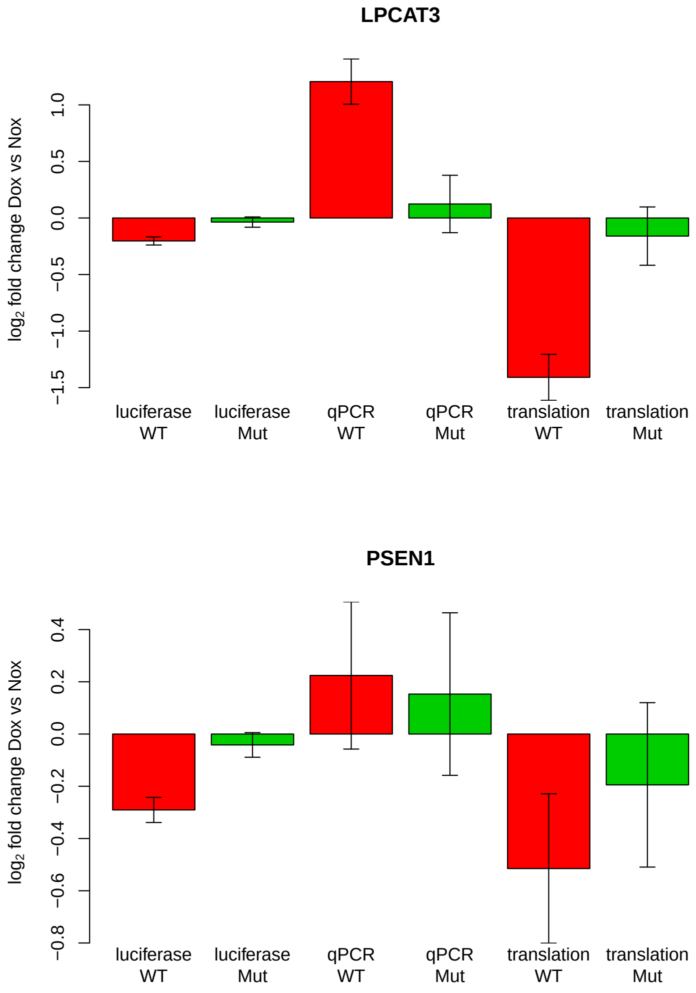

Figure 17: Translation of mRNAs with CDS-located binding sites is significantly repressed following induction of the cognate miRNA. We constructed a stable HEK293 cell line containing an episomal hsa-miR-124 expression vector that can be induced with doxycyclin (see Supplementary Methods). We also constructed two reporter psiCHECK-2 plasmids in which wildtype (WT) miR-124 binding sites from the *LPCAT3* or *PSEN1* genes were inserted upstream of the STOP codon of the renilla luciferase. As negative controls, we constructed two plasmids with 3-4 point mutations in the binding site (Mut). At time t=0h, cells were transfected with one of the plasmids and doxycyclin was added to the medium to induce hsa-miR-124 expression. The figure shows $\log_2$ fold changes in luciferase activity measurements and mRNA levels (quantified by qPCR) 24h post-transfection in doxycyclin-induced cells vs non-induced cells. Changes in translation were estimated from the difference between changes in protein levels and changes mRNA levels. Error bars represent Standard Errors of the Mean (SEM) from three independent experiments with three technical replicates each. The luciferase activities of cells transfected with WT plasmids were repressed for both constructs ($p < 10^{-8}$, one-sided t-test) while cells transfected with Mut plasmids were not ($p > 0.18$). Furthermore, mRNA levels were unaffected ($p > 0.60$) except in the WT LPCAT3 construct whose expression was up-regulated approximately two fold. We computed the change in translation from the difference between the change in luciferase activity and the change in mRNA level upon miRNA induction. In both WT constructs, translation appears to be repressed ($p < 10^{-11}$ for LPCAT3, $p = 0.036$ for PSEN1) while repression is lost with both Mut plasmids ($p > 0.26$).

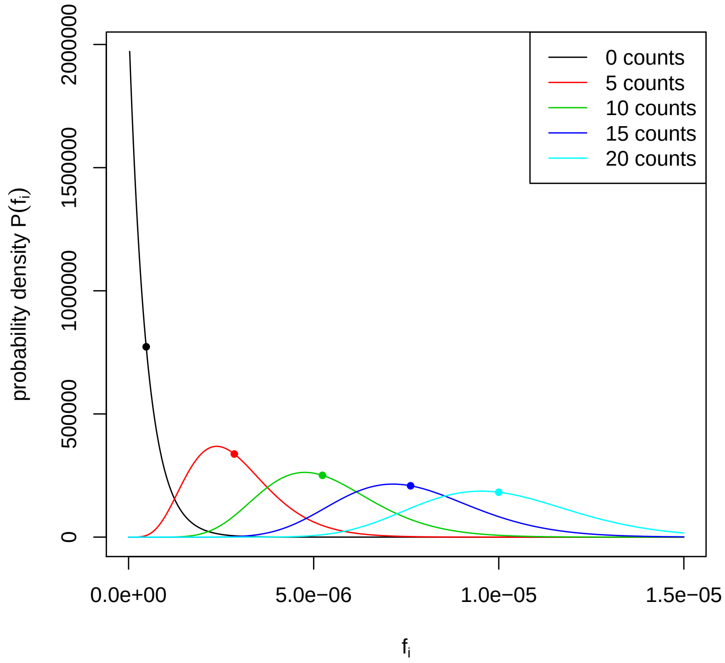

Figure 18: Posterior probability density distribution of the relative mRNA abundance f_i for an mRNA of average length, given that 0, 5, 10, 15 or 20 mRNAseq reads mapped to this mRNA. The total library size was set to 2.1 million reads, which was the library size in the mRNAseq experiments of Kishore et al. (2011). The dots represent the expected value $< f_i >$ of each probability density distribution. Note that probability densities are large ($>> 1$) in the neighborhood of the expected value.

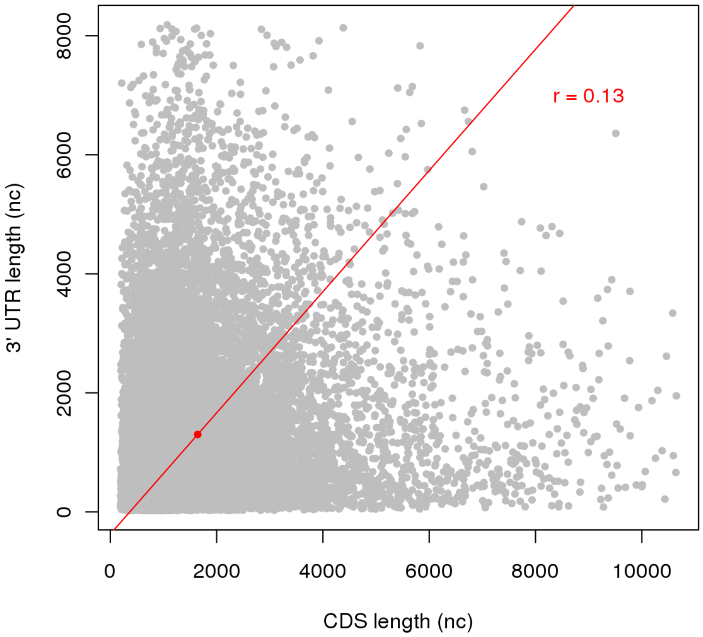

Figure 19: Comparison of CDS and 3' UTR domain lengths of human RefSeq mRNAs. The red line represents the first principal component, the red dot represents the mean CDS and 3' UTR length (1706 and 1334 nucleotides, respectively).