Supplementary materials for "MicroRNA transfection and AGO-bound CLIP-seq datasets reveal distinct determinants of miRNA action"

Jiayu Wen, Brian J. Parker, Anders Jacobsen, and Anders Krogh

The Bioinformatics Centre, Department of Biology and the Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark

Contents

List of Figures

List of Tables

List of Algorithms

1 Supplementary methods

1.1 miRNA-target interaction features

We defined a large number of target features that may be relevant for miRNA targeting (listed in Table 1 in the main text).

Seed, site and flank conservation. phastCons scores (phastCons 28way for hg18 assembly; phastCons 30way for mm9 assembly) were used to compute seed conservation, target conservation (entire miRNA target binding site), and flanking region conservation (70 nucleotides upstream and downstream of seed site).

Alignment score, computed by the miRanda program, is a weighted miRNA–target sequence alignment score with greater weights on seed regions, allowing mismatch and GU pairings (see above).

Seed type. Four types of seed match were used as described in [1]: 6mer (miRNA position 2 to 7), 7mer (miRNA position 2 to 8), 7mer-A1 (6mer plus nucleotide A at position 1), and 8mer (7mer plus nucleotide A at position 1). The seed type was encoded as 1 for 6mer, 2 for 7mer and 7mer-A1, 4 for 8mer, and 0 for no perfect 6mer seed match. This encoding approximately corresponds to a linear encoding with respect to expression fold change, as described in [1]. Note that the HITS-CLIP and PAR-CLIP datasets assigned the clusters using a best seed match which could potentially introduce a positive selection bias for this feature, however the consistency between the two CLIP-seq datasets, and given that the observed effects in comparison with transfection were in the opposite direction to the potential bias indicate that the bias is not substantial.

Number of sites were counted for each 3′ UTR with at least one perfect 6-mer match seed.

Seed mismatch and seed GU were counted as the number of mismatches and G-U pairings occurring in positions 2 to 7 inclusive of the mature miRNA.

The ***3′ out-seed pairing*** was calculated as described in [1] by positioning a 4mer match at 13-16 in the 3′ miRNA regions.

We computed ***mono-nucleotide frequency and di-nucleotide frequency*** on the flanking 70 nt. Di-nucleotide frequency was defined as $\rho_{XY} = f_{XY}/f_X \times f_Y$, where f_X and f_{XY} represent mono- and di- nucleotide frequencies, respectively. This measures the deviation of the dinucleotide frequency from mono-nucleotide frequency.

The ***flank AU content*** was calculated on the flanking 30 bp by counting A and T frequencies with a higher weights near the target sites as defined in [1].

To compare the ***target accessibility*** features described in [2], we used the PITA program to generate the interaction energy score ($\Delta\Delta G$) and the miRNA–target hybridization energy (ΔG_{duplex}). The parameter settings were: a seed length of 6 to 8 bp, allowing 1 mismatch and 1 G-U pairing in seed regions. The calculation of ΔG_{open} including a flank of 3 and 15 bp upstream and downstream, respectively.

The ***local structural strand asymmetry*** [3] features of flanking regions were calculated on 70 nucleotides upstream and downstream of the seed site. We used RNAsubopt in the RNAfold package [4] to generate a sample of size 100 from the Boltzmann ensemble of secondary structures and calculated the minimum free energy (MFE) for each structure on the sense (forward) strands. Using the same matched structures as on the forward strands, we used RNAeval to calculate MFEs on the reverse complement strand. The difference between the two strands was defined as ΔG_{fw-rc}, where fw and rc represent the sense (forward) strand and the reverse complement (antisense) strand, respectively. Miklos et al [5] noted that minimum free energy has a smaller variance for stable RNAs. Therefore, the three moments (mean, standard deviation, and 4th moment) of MFE difference distribution were calculated as local structural strand asymmetry features.

The ***local base composition asymmetry*** was estimated on 70 nucleotides upstream and downstream of the seed site using the following features: (G-C) = $(f_G - f_C)/(f_G + f_C)$, (A-U) = $(f_A - f_U)/(f_A + f_U)$, and (G+U) = $\log_2((f_G + f_U)/(f_A + f_C))$. Note that these features are standardized such that the value of 0 indicates no asymmetry across the sense and antisense strands.

1.2 3′ UTR length simulation

We simulated UTRs drawn from an empirical distribution based on the full set of Refseq gene 3′ UTR lengths. For each UTR, we modeled the placement of targets as a simple Poisson process with targets spaced (parameter lambda) 1.27 per kb apart on average; lambda was estimated empirically from the mean distances between targets as predicted by TargetScanS averaged over all known miRNAs. UTRs with 0 targets modeled were removed to

simulate the true positive set. From this simulated true positive set, a negative set with no targets but identical length distribution was created. Finally, a simulated positive set as labelled by a simulated CLIP-seq procedure was created by modeling the detection of targets as a Bernoulli process with parameter p=the theoretical CLIP-seq sensitivity. In the simulation the sensitivity was varied between 0 and 1.

1.3 miRNA target datasets for prediction comparison

We downloaded pre-compiled human target predictions of TargetScanS (release 5.1: http://www.targetscan.org/cgi-bin/targetscan/), PicTar (from UCSC table browser), miRanda (Sept2008 release; http://www.microrna.org/microrna/), miRBase Target (v5: http://www.ebi.ac.uk/enright-srv/microcosm/cgi-bin/targets/v5/download.pl), PITA (version 6 with 3/15 flank: http://genie.weizmann.ac.il/pubs/mir07/index.html), EIMMo (v3, http://www.mirz.unibas.ch/ElMMo3/), and RNA22 on 3´ UTRs (http://cbcsrv.watson.ibm.com/rna22.html). Different prediction programs used different gene identifiers to report their predictions. To make the predictions comparable, we mapped the gene identifiers to gene symbols. Mapping RefSeq to gene symbol was conducted through a mapping table provided by UCSC and mapping Ensemble to gene symbol was through biomaRt identifier conversion. The pooled pSILAC proteomics dataset with down-regulation fold change of greater than 1.4 fold as the positive set (483 genes) and the corresponding negative set (483 genes) were used for prediction comparison. For each pre-compiled data from existing prediction program, the sensitivity was measured as the percentage of miRNA-target interactions predicted from the positive set and 1−specificity was measured as the percentage of miRNA-target interactions predicted from the negative set.

1.4 False discovery rate estimation

FDR estimation allows choosing a threshold for a good balance between sensitivity and specificity. Since miRNA seed sites often occur by chance, it is important to estimate the false discovery rate (FDR) of genome-wide miRNA target predictions. To do so, we calculated the significance (p-values) of each predicted target through comparison with background distributions of prediction posterior probability. The background distributions were derived by predicting targets for each miRNA on a large sample of randomly shuffled 3´ UTR sequences. In more detail, we calculated p-values for all predicted targets using a permutation test: a set of 5000 3´ UTRs were randomly sampled and randomly shuffled while keeping the same base compositions to simulate the null distribution. The miRNA sequences were used to predict targets against the shuffled 3´ UTR sequences by the same prediction procedure to generate a null distribution for each miRNA. For each miRNA, the proportion of hits with prediction posterior probability greater than the corresponding unshuffled prediction posterior probability was used to estimate a p-value. The raw p-values for each miRNA model were adjusted to compute the FDR of the final prediction set (Benjamini-Hockberg [6]).

References

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

[2] M Kertesz, N Iovino, U Unnerstall, U Gaul, and E Segal. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39:1278–1284, 2007.

[3] J Wen, B J Parker, and G F Weiller. In silico identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula. *In Silico Biol*, 7:485–505, 2007.

[4] I L Hofacker, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie*, 125:167–188, 1994.

[5] I Miklós, I M Meyer, and B Nagy. Moments of the Boltzmann distribution for RNA secondary structures. *Bull Math Biol.*, 67:1031–1047, 2005.

[6] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57:289–300, 1995.

2 Supplementary figures

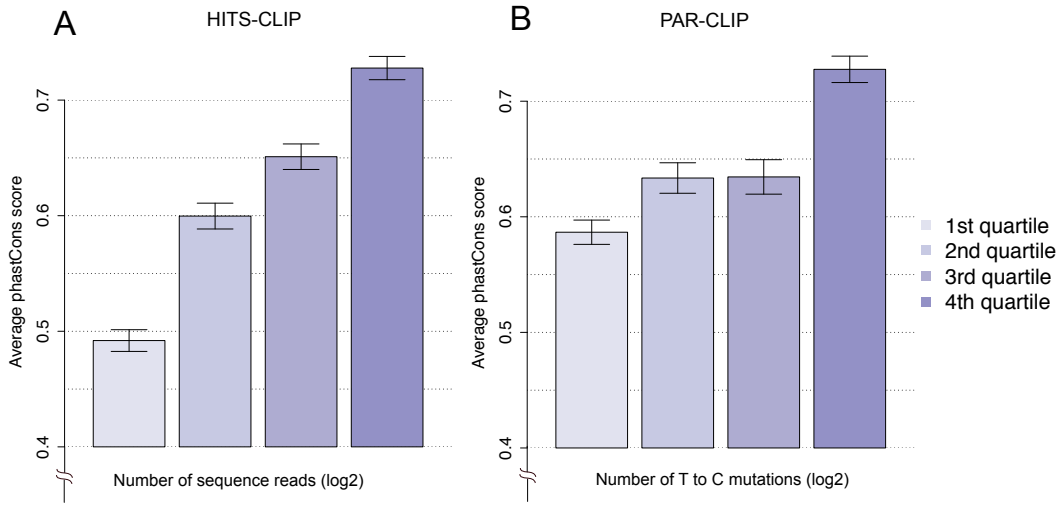

Figure 1: Association of seed phastCons conservation score and AGO binding affinity in the CLIP-seq datasets. The distribution for each quartile of sequence reads (x-axis) and mean values (standard error) of conservation phastCons score for each expression quartile (y-axis) are plotted. Number of sequence reads for (A) HITS-CLIP (BC≥2, tag≥2) and number of T to C mutations for (B) PAR-CLIP (T to C ≥1) were used as a measure of binding affinity. Graph shows a higher number of cross-linked miRNPs correlated with higher seed conservation.

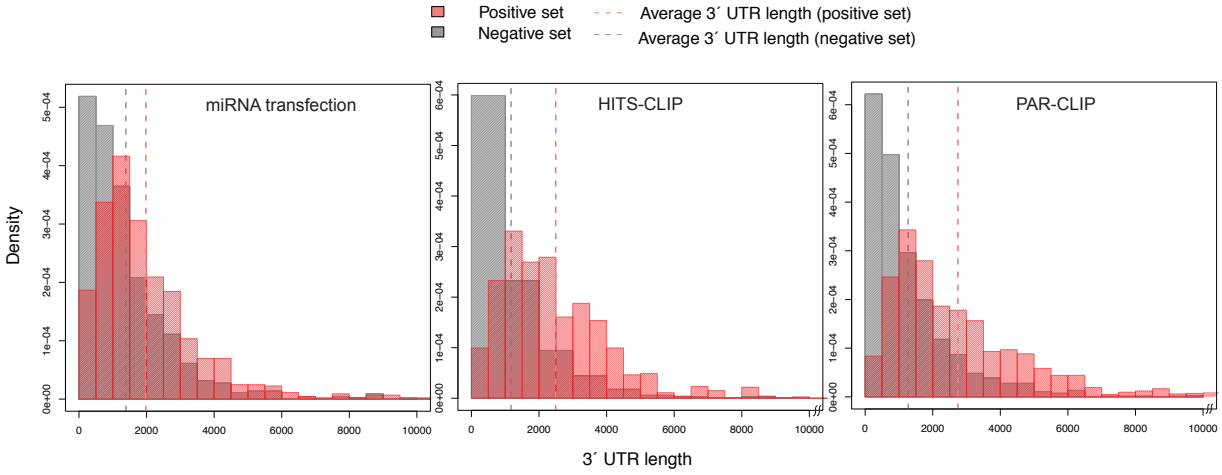

Figure 2: Comparison of 3´ UTR length distributions in three datasets.

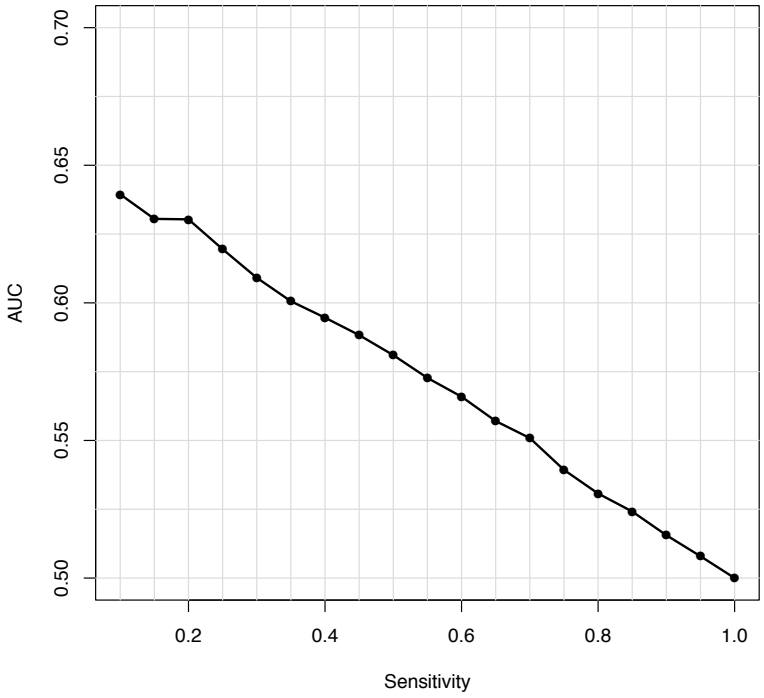

Figure 3: 3′ UTR length simulation. Shown are the simulated theoretical CLIP-seq sensitivity (x-axis) vs AUC of simulated 3′ UTR length as a feature (y-axis). Data for the simulated 3′ UTRs shows that as sensitivity drops from an ideal of 1 to 0, the AUC of 3′ UTR length feature increases from the null of 0.5 to a maximum of 0.64.

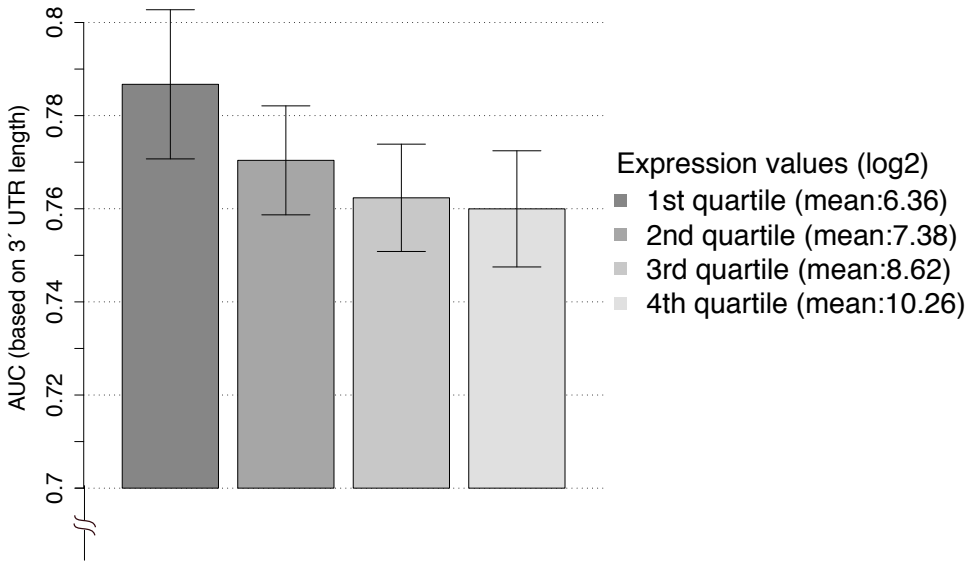

Figure 4: Correlation of mRNA expression values and the predictive power of 3′ UTR length. The distribution for each quartile of mRNA expression level (see Methods) (x-axis) and discriminability of 3′ UTR length feature for each expression quartile are plotted (y-axis) (Error bars show standard errors of AUCs). Mean mRNA expression values for each quartile are shown in brackets. Graph shows a negative correlation between mRNA expression level and 3′ UTR length.

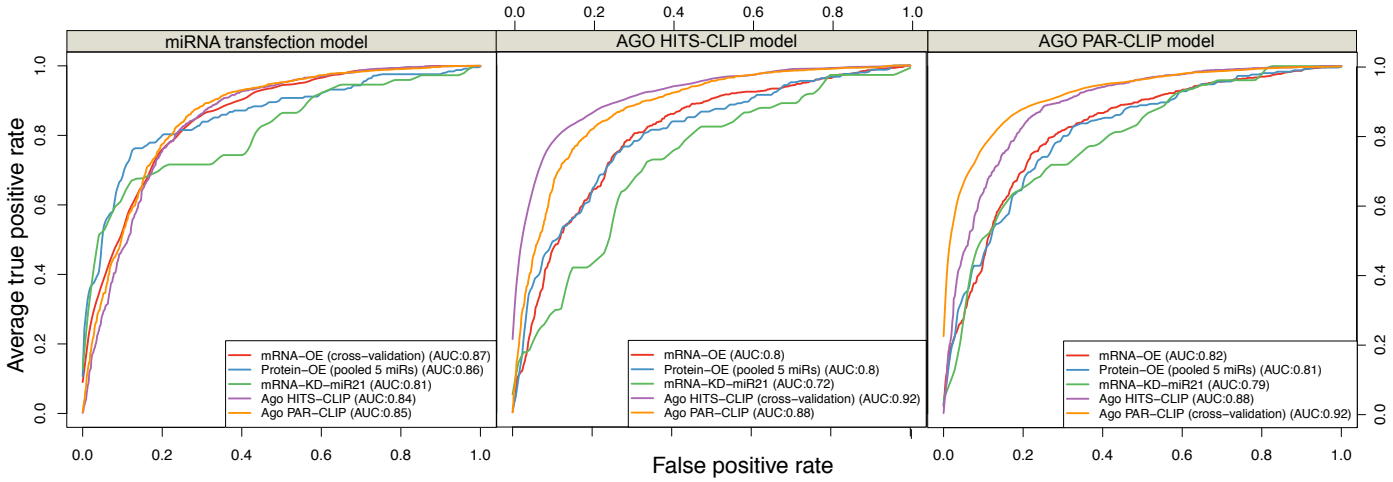

Figure 5: Comparison of predictive models trained on miRNA transfection, HITS-CLIP, and PAR-CLIP experiments. Plotted are receiver operating characteristic (ROC) curves for the three predictive models with a combination of all features that were trained on these three datasets and validated either using leave-one-miRNA family-out cross-validation or on independent datasets as indicated. AUC values are shown in the legend. Additionally two independent datasets were also used for validation: protein-OE (pSILAC proteomic dataset measuring protein level changes after transfecting 5 miRNAs), and mRNA-KD-miR21 (miR-21 knock–down microarray expression data). Refer to main paper for discussion.

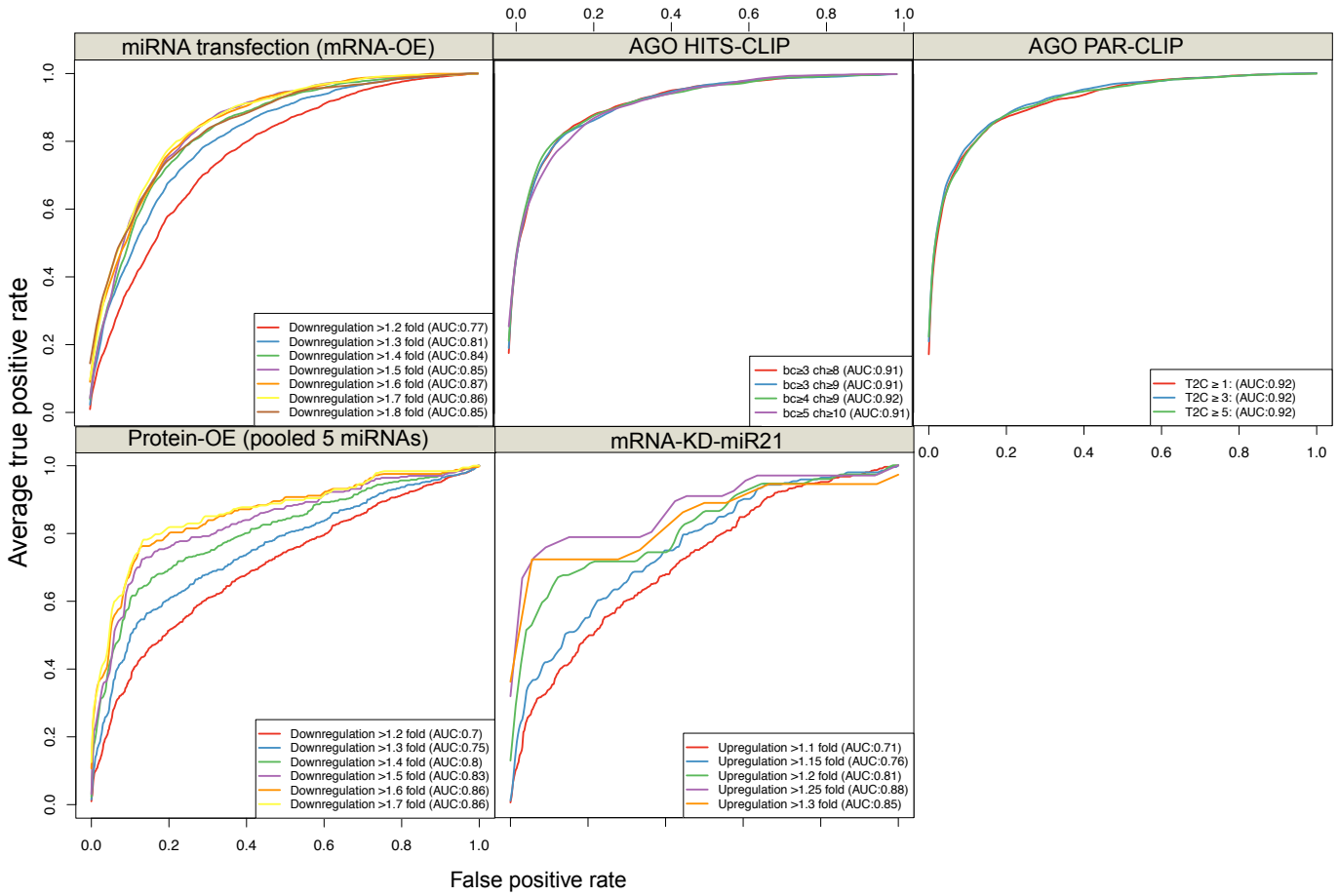

Figure 6: ROC curves and AUCs using varied thresholds for defining down- or up-regulation as the response variable. The top panel shows the predictive performance of models trained on 12 miRNA transfection, HITS-CLIP and PAR-CLIP datasets, respectively, using a leave-one miRNA family-out cross-validation. The bottom panels show the predictive performance of the model trained on the transfection dataset and tested on the proteome and miR-21 knock-down datasets.

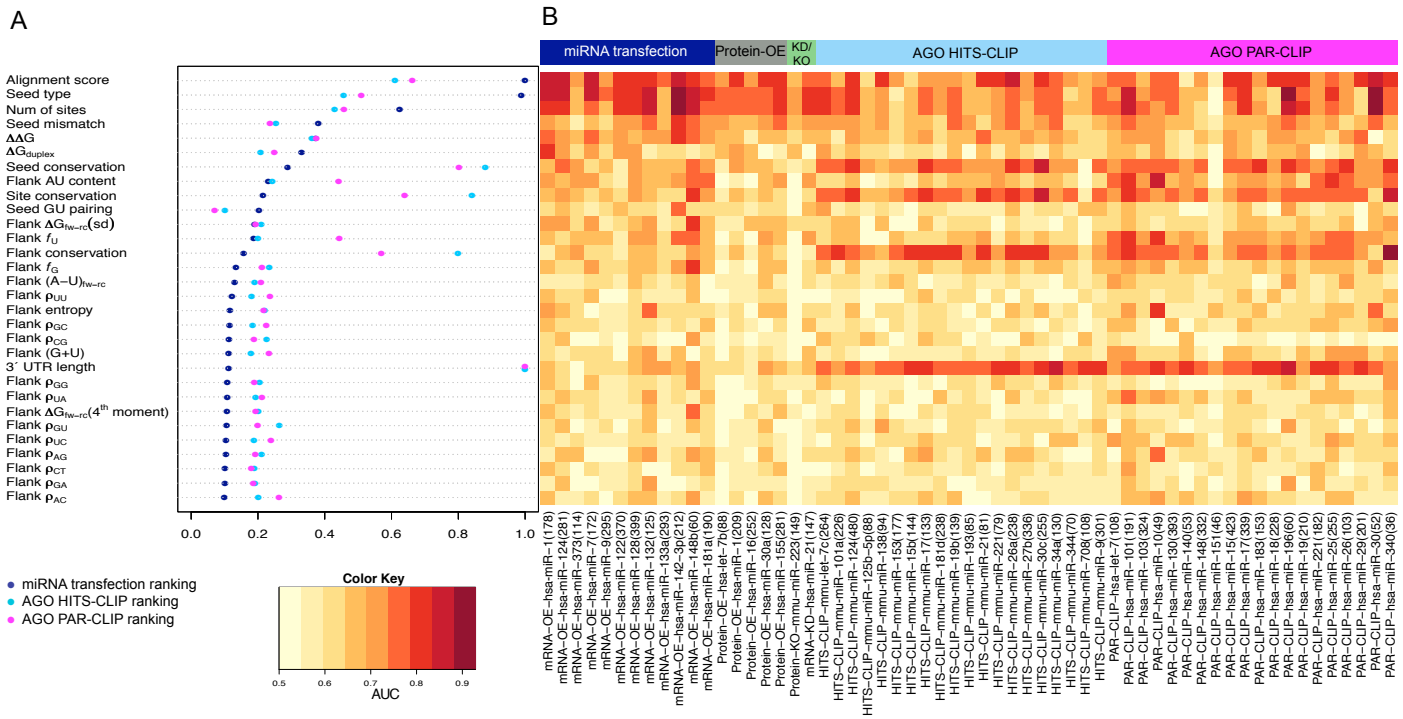

Figure 7: Feature importance ranking and predictive power for miRNA transfection/inhibition, HITS-CLIP, and PAR-CLIP datasets. The same feature ranking analyses as Figure 1 in the main text except that to ensure a completely unbiased feature comparison analysis, the best scoring putative seed match was used to define the target sites in both positive and negative set of all datasets (as opposed to utilizing the CLIP-seq clustering information to define the target sites. Thus this analysis trades off sensitivity for unbiasedness). (A) Feature importance rankings were evaluated using the Gini impurity criterion (normalized between 0 to 1) from Random Forest classification for three datasets separately. Top 30 features are shown: miRNA transfection data (dark blue, in decreasing order), HITS-CLIP (light blue), and PAR-CLIP (pink).

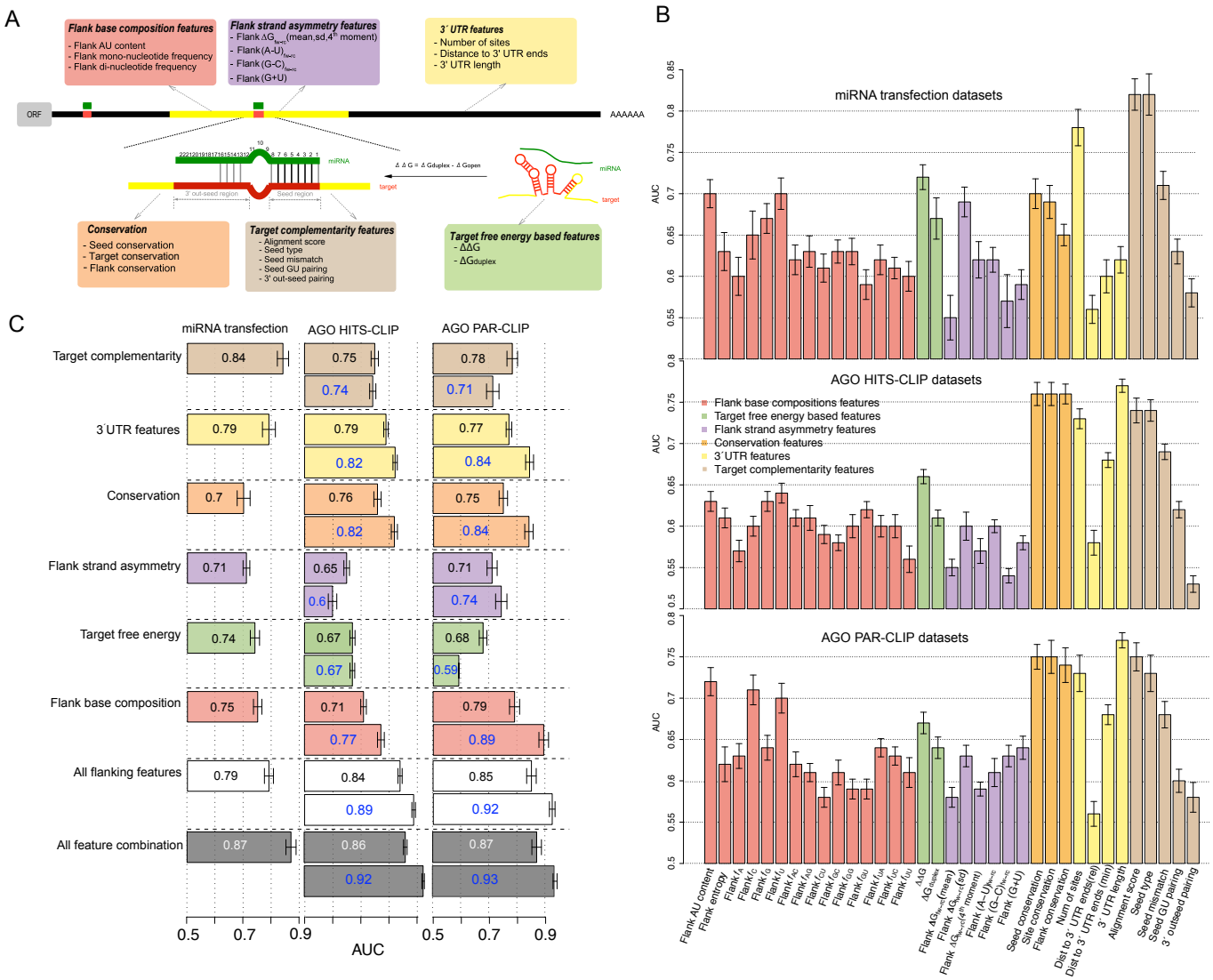

Figure 8: Predictive power of both individual miRNA-target features and feature combinations. The same feature analyses as Figure 2 in the main text except that it is based on a completely unbiased feature comparison analysis where the best scoring putative seed match sites were used in all datasets. (A) miRNA-interaction features are grouped into six categories by their sequence, structure, or positional characteristics (listed in the colored box). (B) Barplot shows univariate feature AUC for the positive set vs. negative sets in miRNA transfection, HITS-CLIP, and PAR-CLIP datasets. The error bars show standard errors of AUC. (C) Comparison of combined predictive performance for each feature category. AUCs for feature combinations in each of six categories are shown in bars for all three datasets (leave-one miRNA family-out cross-validation; Random forest classifier). Two additional feature combinations are also shown: the combination of all flanking features and the combination of all features. All flanking features includes flanking nucleotide composition, target free energy based features, flanking strand asymmetry, flanking region conservation, and relative/minimum distance to 3´ UTR ends. For comparision, bars with AUCs in blue color also lists the feature category prediction performance where features in the positive set were extracted from the CLIP-seq cluster sites including the potentially biased seed-features ("target complementarity", "number of sites" (in 3´ UTR feature category) and target free energy).

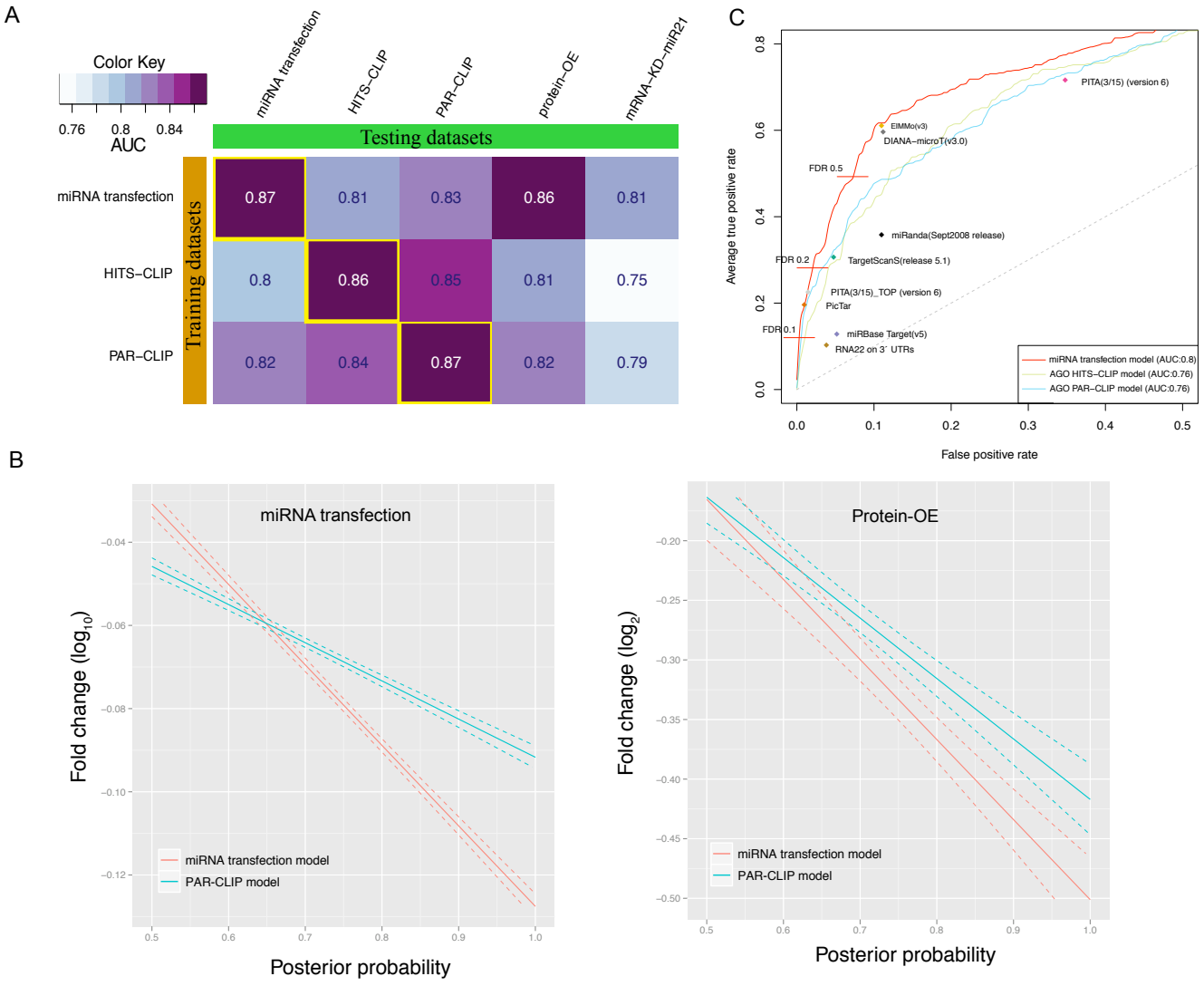

Figure 9: Comparisons of predictive models trained on miRNA transfection, HITS-CLIP, and PAR-CLIP datasets. The same model comparison as Figure 5 in the main text except based on a completely unbiased feature comparison analysis where the best scoring putative seed match sites were used in all datasets. (A) The heatmap shows the predictive performance (measured in AUC) of models using all features, trained and validated on different datasets. Diagonal cells (yellow framed) show the AUCs for leave-one miRNA family-out cross-validation of the models. Off-diagonal cells show the model trained on the row dataset and tested on the column dataset. (B) Comparisons of the correlation of expression change and posterior probability of prediction generated by the miRNA transfection-trained model (red) and the PAR-CLIP-trained model (blue). The two models were applied to miRNA transfection and protein-OE datasets without thresholding on fold change. Linear regression lines were fitted to all predicted targets with posterior probability ≥ 0.5 and dashed lines are the 95% confidence interval of the fitting. (C) The performance of the model trained on the miRNA transfection data using all features and, for comparison, several current target prediction programs. All were evaluated on the protein-OE dataset. The ROC curves (red) show the true positive rate (sensitivity) versus false positive rate (1−specificity) for the positive set (note: fold change threshold $\geq$ 1.4-fold) vs. the negative set (low fold change) classifications. Short red horizontal lines on ROC curves marked predictions with a FDR limit of 50%, 20%, and 10% (FDR estimation in supplementary methods), showing the trade-off between sensitivity and specificity. Colored dots represent maximum sensitivity (percentage of miRNA-target interactions predicted from the positive set) versus 1−specificity (percentage of miRNA-target interactions predicted from the negative set) of predicted miRNA-target interactions for each corresponding prediction program. The CLIP-seq models evaluated on the same data are also shown.

3 Supplementary tables

	Testing datasets					
	RIP-chip (CV)	miRNA transfection	HITS-CLIP	PAR-CLIP	protein-OE	mRNA-KD-miR21
RIP-chip model	0.85	0.71	0.77	0.82	0.69	0.69

Table 1: Prediction performance of RIP-chip model

Category	Region	Features	P-values (significance of AUC difference)		
			miRNA transfection vs HITS-CLIP	miRNA transfection vs PAR-CLIP	PAR-CLIP vs HITS-CLIP
Conservation	70 nt	Conservation of flanking region.	4.2E-22	3E-12	9.9E-4
	seed	Conservation of seed.	4.1E-11	2.3E-9	0.27
	target	Conservation of entire target sequence	1.8E-15	7.1E-6	2.8E-5
Target complementarity	target	miRNA-target sequence alignment score	4.4E-7	9.3E-6	0.28
	seed	6mer, 7mer-m8, 7mer-A1, and 8mer [1]	4.4E-7	3.8E-9	0.28
	seed	Seed mismatch/GU pairing	0.4	0.15	0.1
	out-seed	3´ out-seed pairing [1]	8.3E-4	0.33	2.6E-6
3´ UTR features	3´ UTR	3´ UTR length	1E-14	1.2E-10	0.09
	3´ UTR	Number of target sites	2.3E-8	1.5E-5	0.11
	3´ UTR	Relative distance to 3´ UTR ends	3E-29	1.8E-30	0.4
	3´ UTR	Minimum distance to 3´ UTR ends	1.6E-10	1.5E-7	0.13
Composition in target site flanking regions	70 nt	Mononucleotide frequencies	4.4E-3	2.4E-13	3.4E-7
	70 nt	Dinucleotide frequencies	0.01	0.01	0.4
	70 nt	AU content [1]	1.1E-6	0.02	2.4E-19
	70 nt	Base compositional entropy	2.5E-11	2.9E-8	0.15
Target free energy based features	target	Free energy loss $\Delta\Delta G$ that indicates target site accessibility [2]	3.6E-3	1.8E-2	0.3
	target	ΔG_{duplex} miRNA-target hybridization energy	8.3E-7	1.1E-5	0.31
Flanking strand asymmetry	70 nt	Folding free energy difference ΔG between two strands. Mean, standard deviation and 4th moment [3]	5E-13	1.4E-15	3.1E-1
	70 nt	Base asymmetry bias, A vs. U and G vs. C [3]	5.8E-5	7.3E-2	7.5E-3
	70 nt	G+U content [3]	2.2E-7	0.19	2.2E-6

Table 2: miRNA target and contextual features used in this study with significance of AUC difference of a feature between datasets (P-values, estimated by Hanley-McNeil test).

Category	Region	Features	AUCs			P-values (significance of AUC difference)		
			miRNA trans-fection	HITS-CLIP	PAR-CLIP	miRNA transfection vs HITS-CLIP	miRNA transfection vs PAR-CLIP	PAR-CLIP vs HITS-CLIP
Conservation	70 nt	Conservation of flanking region.	0.65	0.76	0.74	1.5E-12	1.1E-8	8.3E-2
	seed	Conservation of seed.	0.70	0.76	0.75	9E-5	1.2E-3	0.27
	target	Conservation of entire target sequence	0.69	0.76	0.75	5.2E-6	1.1E-4	0.27
Target complementarity	target	miRNA-target sequence alignment score	0.82	0.74	0.75	1.8E-9	1.2E-7	0.27
	seed	6mer, 7mer-m8, 7mer-A1, and 8mer [1]	0.82	0.74	0.73	1.8E-9	1.2E-11	0.27
	seed	Seed mismatch/GU pairing	0.76	0.73	0.73	4.1E-2	4E-2	0.4
	out-seed	3′ out-seed pairing [1]	0.58	0.53	0.58	4.1E-3	0.4	3.4E-4
3′ UTR features	3′ UTR	3′ UTR length	0.66	0.77	0.77	8.3E-13	7.2E-13	0.4
	3′ UTR	Number of target sites	0.78	0.73	0.73	5.1E-4	4.9E-4	0.4
	3′ UTR	Relative distance to 3′ UTR ends	0.56	0.58	0.56	0.19	0.4	0.13
	3′ UTR	Minimum distance to 3′ UTR ends	0.60	0.68	0.68	1.5E-6	1.4E-6	0.4
Composition in target site flanking regions	70 nt	Mononucleotide frequencies	0.73	0.62	0.70	9E-13	4.8E-2	5.3E-10
	70 nt	Dinucleotide frequencies	0.71	0.69	0.76	0.16	1E-3	6E-9
	70 nt	AU content [1]	0.70	0.63	0.72	1.2E-5	0.16	1.1E-12
	70 nt	Base compositional entropy	0.63	0.61	0.62	0.18	0.33	0.3
Target free energy based features	target	Free energy loss $\Delta\Delta G$ that indicates target site accessibility [2]	0.72	0.66	0.67	1.3E-4	1.2E-3	0.29
	target	ΔG_{duplex} miRNA-target hybridization energy	0.67	0.61	0.64	2.8E-4	6.2E-2	2.7E-2
Flanking strand asymmetry	70 nt	Folding free energy difference ΔG between two strands. Mean, standard deviation and 4th moment [3]	0.69	0.60	0.64	2.2E-8	2E-3	3.3E-3
	70 nt	Base asymmetry bias, A vs. U and G vs. C [3]	0.64	0.60	0.66	1.8E-2	0.18	7.1E-6
	70 nt	G+U content [3]	0.63	0.58	0.64	3.3E-3	0.33	9.3E-6

Table 3: miRNA target and contextual features used in this study and their predictive power measured in AUC based on unbiased feature analysis.
The same as Table 1 in the main text except that it is based on a completely unbiased feature comparison analysis where the best scoring putative seed match was used to define target sites in all datasets. P-values of significance of AUC differences of a feature between datasets are shown (estimated by Hanley-McNeil test).

Algorithm 1 3′ UTR simulation in R

```
library(caTools)

#############################################################################

get_auc <- function(data) {
        auc = colAUC(subset(data,select=-1), data[,1], plotROC=F)
        auc
}

#############################################################################

lambda = 1.27/1000
sample_size = 20000

human_utrlen = read.table(file="human_utrlen.txt",header=T)$utrlen

AUCs = NULL
for (prob in seq(0.1,1,0.05)) {
        s1 = human_utrlen[sample(1:length(human_utrlen), sample_size,replace=T)]

        dd_pos = NULL
        for (i in 1:length(s1)){
                #Possion distribution to get number of targets per UTRs
                u1 = rpois(1,lambda=s1[i]*lambda)
                # modeling the detection of targets as Bernoulli process
                u2 = rbinom(1, u1, prob)
                dd_pos = rbind(dd_pos,cbind(u1,u2))
        }
        dd_pos = cbind(dd_pos,s1)
        dd_pos = dd_pos[dd_pos[,"u1"]!=0,]

        dd_neg = dd_pos
        dd_neg[,"u2"]=0

        dd = data.frame(rbind(cbind(response=1,dd_pos), cbind(response=0,dd_neg))
)
        dd = subset(dd,select=-u1)
        colnames(dd) = c("response","nsites","utr3len")
        dd$response = as.factor(dd$response)
        dd = transform(dd, clip_response = as.numeric(as.logical(dd$nsites)) )
        dd$clip_response = as.factor(dd$clip_response)

        auc_full = get_auc(dd[,c("clip_response","utr3len")])

        aucs = cbind(prob, auc=auc_full)
        AUCs = rbind(AUCs, aucs)

}
colnames(AUCs) = c("sensitivity","utrlen")

print(AUCs)
```