SUPPLEMENTARY NOTE

Alternative motif matches—A1 sites

In addition to the stringent and relaxed matching criteria we tested in the main text, we also tested motif matches to A1 sites, which are another common type of hexamer match for microRNAs. A1 sites are those sites matching to nts 2-6 of the mature miRNA with an additional A pairing to nt 1 at the 5' end of the mature miRNA (not necessarily the complementary base). However, we found that including the A1 site resulted in many matches to low-confidence miRNAs (e.g. poorly conserved and having very low expression) for the categorical linear model, while influencing the results of other models very little. Therefore we chose to omit the A1 site matches from our analyses (data not shown).

Plotting the truncated Receiver Operator Curves

To draw a ROC curve, we must be able to define the true positives. In our case, we chose not to draw ROC curves across all possible motifs while using all miRNAs in miRBase as the true positives, since relatively few of them are expected to be expressed in a particular cell type. Furthermore, the methods did not output the same number of motifs—in particular, miReduce outputs many fewer motifs than the other methods. It is not clear how to best draw ROC curves when the methods do not output the same number of predictions.

We thus chose to truncate the ROC curves to the number of motifs to *N* = 20 and *N* = 50, to demonstrate how well the methods perform in the top predictions. The way we truncated the curves produces exactly the same curves as would be obtained by magnifying the top results in a full ROC curve. The remainder of the full ROC curve is expected to approach the random predictor line for all methods and would not give additional information about the performance of the methods. After truncation, we simply scale the X and Y axes to both range from 0 to 1. We caution that this truncated AUC statistic should only be used to compare the different methods to each other and not to the typical baseline value of 0.5 for a random method. To this end, we include in each plot a baseline for a random predictor calculated from the expected true and false positive rate given the total number of hexamers matching miRBase miRNAs and the total number of motifs being tested.

In the truncated ROC curves, we see that the results of the comparisons are robust in that the accuracy of MixMir dominates that of the other methods over almost the entire range of sensitivity settings. We consider this to be the best indicator of MixMir's performance as the AUC values are really only needed for comparison when ROC curves intersect each other..

Analysis of the effects of adding a 3' UTR length covariate

Figure S2 plots the percentage of positive associations against motif rank as a PP-plot. The LM Bin model without 3' UTR length as a covariate had nearly all positive associations across all motifs, but when we added the 3' UTR length covariate, this was altered dramatically. There was a less pronounced effect for MixMir, presumably because the relationship matrix implicitly corrects for this UTR length effect, as genes with longer UTRs (with more motifs present) will have lower relatedness to genes with shorter UTRs.

Additionally, the inclusion of the 3' UTR length covariate partially corrected the skewness of the PP plots observed in Figure 1 (see Figure S2). Notably, the simple linear models became much less skewed. Interestingly, MixMir6* showed no improvement over MixMir6. Further, motif rankings produced by the linear model was substantially different when comparing models with and without the added covariate. This shift was much smaller or non-existent in MixMir (Table S6).

The addition of the 3' UTR length covariate provides a strong correction for the overall percentage of positive coefficients in the simple linear models (Table S7). This brings out the enrichment of positive coefficients in the significant motifs for the linear models, to be more in line with what we observe in the mixed linear models.

However, note that these changes in motif rank did not strongly affect our previous ROC results, as the most highly ranked motifs did not change significantly (data not shown). We thus present our results in the main text without the correction for 3’ UTR length.

SUPPLEMENTARY REFERENCES

1. Friedman, R.C., Farh, K.K., Burge, C.B. and Bartel, D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, **19**, 92-105.
2. van Dongen, S., Abreu-Goodger, C. and Enright, A.J. (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods*, **5**, 1023-1025.
3. Bartonicek, N. and Enright, A.J. (2010) SylArray: a web server for automated detection of miRNA effects from expression data. *Bioinformatics*, **26**, 2900-2901.

Supplementary Tables

Method	**LM Bin**	**MixMir2**	**MixMir3**	**MixMir4**	**MixMir5**
MixMir2	0.8876				
MixMir3	0.8101	0.9677			
MixMir4	0.6336	0.8397	0.9363		
MixMir5	0.3407	0.5356	0.6517	0.8012	
MixMir6	0.2293	0.3624	0.4428	0.5643	0.8832

Table S1. Comparison of MixMir result similarities with the simple linear model (LM Bin). Pairwise Pearson correlation of all motif ranks. We saw that the degree of rank similarity between the LM and MixMir results varied directly with the length of the *k*mer used to construct the relationship matrix.

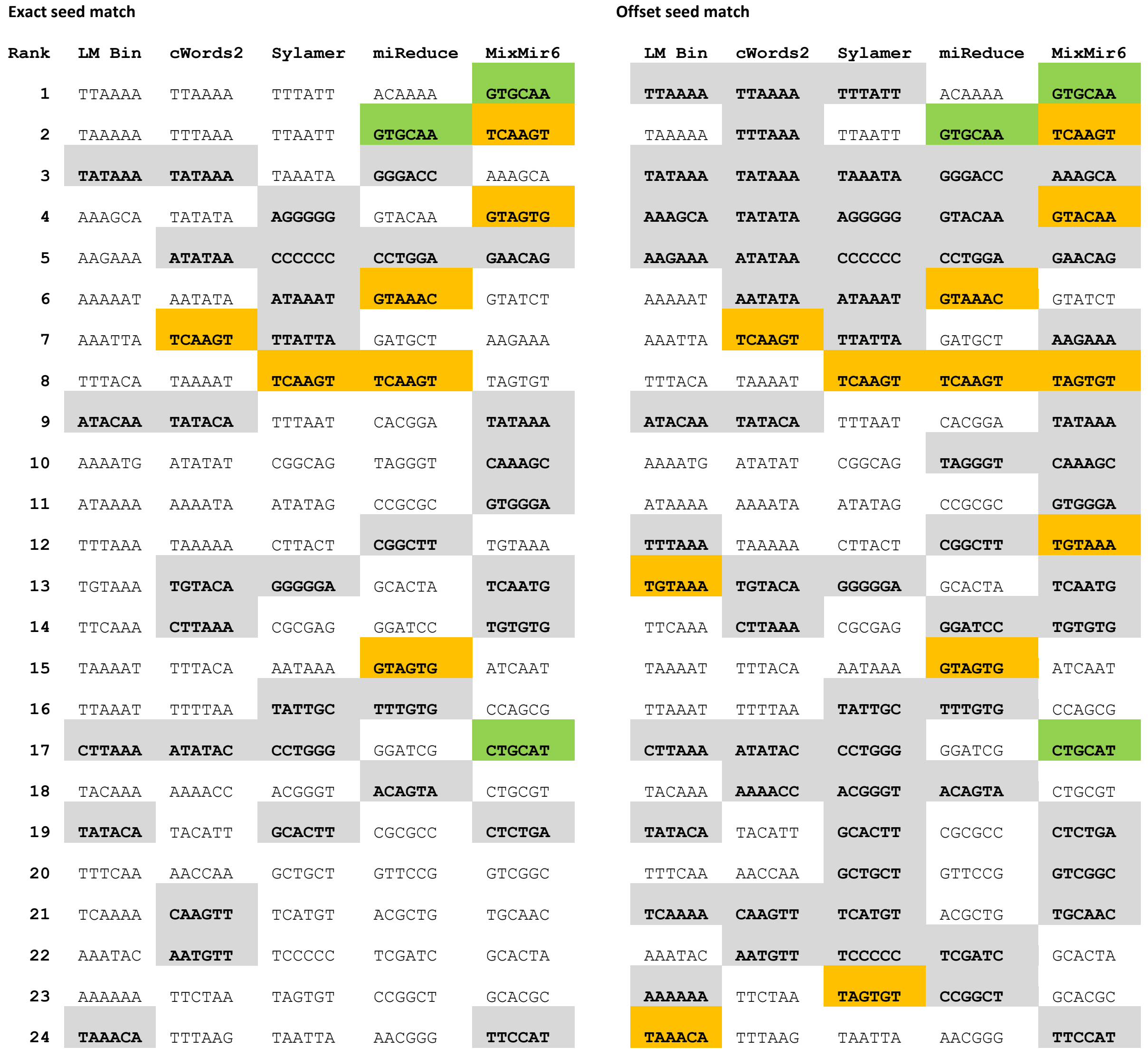
Method	**LM Bin**	**cWords2**	**cWords3**	**cWords4**	**cWords5**
cWords2	0.7978				
cWords3	0.8611	0.9335			
cWords4	0.9287	0.8682	0.9312		
cWords5	0.9616	0.8476	0.9020	0.9610	
cWords6	0.9616	0.8475	0.9019	0.9609	1.000

Table S2. Comparison of cWords result similarities against simple linear model. Pairwise Pearson correlation of all motif ranks. Here we saw an opposite effect of what we observed with MixMir (Table S1): as we increase *k* in cWords, the results became closer to those of the linear models, with cWords2 and cWords3 producing the most different results. cWords5 and cWords6 were nearly identical in motif ranking.

Method	**Rank**	**Motif**	**miRNAs matched**
LM Bin	13	TGTAAA	[1]mmu-miR-30b-5p, [1]mmu-miR-30c-5p, [1]mmu-miR-30e-5p
	24	TAAACA	[3]mmu-miR-30b-5p, [3]mmu-miR-30c-5p, [3]mmu-miR-30e-5p
cWords2	7	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	26	TGTAAA	[1]mmu-miR-30b-5p, [1]mmu-miR-30c-5p, [1]mmu-miR-30e-5p
	30	TTCAAG	[1]mmu-miR-26a-5p, [1]mmu-miR-26b-5p
	35	TAGTTT	[1]mmu-miR-19a-3p
	44	GTGCAA	[2]mmu-miR-19a-3p
	47	AGCAGC	[2]mmu-miR-15b, [2]mmu-miR-195a-5p
Sylamer	8	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	23	TAGTGT	[3]mmu-miR-142-3p
miREDUCE	2	GTGCAA	[2]mmu-miR-19a-3p
	6	GTAAAC	[2]mmu-miR-30b-5p, [2]mmu-miR-30c-5p, [2]mmu-miR-30e-5p

	8	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	15	GTAGTG	[2]mmu-miR-142-3p
MixMir6	1	GTGCAA	[2]mmu-miR-19a-3p
	2	TCAAGT	[2]mmu-miR-26a-5p, [2]mmu-miR-26b-5p
	4	GTAGTG	[2]mmu-miR-142-3p
	8	TAGTGT	[3]mmu-miR-142-3p
	12	TGTAAA	[1]mmu-miR-30b-5p, [1]mmu-miR-30c-5p, [1]mmu-miR-30e-5p
	17	CTGCAT	[2]mmu-miR-20a-3p
	37	TTCAAG	[1]mmu-miR-26a-5p, [1]mmu-miR-26b-5p

Table S3. Performance of each method on miRNA expression data from mouse CD4+ T-cells. The number in square brackets refers to the position of the 6-mer match in the mature miRNA (position 2 is the exact seed match). Selected miRNAs shown are those which are also highly expressed in one of two experimental data sets.



	Exact seed match					Offset seed match				
Rank	**LM Bin**	**cWords2**	**Sylamer**	**miReduce**	**MixMir6**	**LM Bin**	**cWords2**	**Sylamer**	**miReduce**	**MixMir6**
1	TTAAAA	TTAAAA	TTTATT	ACAAAA	**GTGCAA**	**TTAAAA**	**TTAAAA**	**TTTATT**	ACAAAA	**GTGCAA**
2	TAAAAA	TTTAAA	TTAATT	**GTGCAA**	**TCAAGT**	TAAAAA	**TTTAAA**	TTAATT	**GTGCAA**	**TCAAGT**
3	**TATAAA**	**TATAAA**	TAAATA	**GGGACC**	AAAGCA	**TATAAA**	**TATAAA**	**TAAATA**	**GGGACC**	**AAAGCA**
4	AAAGCA	TATATA	**AGGGGG**	GTACAA	**GTAGTG**	**AAAGCA**	**TATATA**	**AGGGGG**	**GTACAA**	**GTACAA**
5	AAGAAA	**ATATAA**	**CCCCCC**	**CCTGGA**	**GAACAG**	**AAGAAA**	**ATATAA**	**CCCCCC**	**CCTGGA**	**GAACAG**
6	AAAAAT	AATATA	**ATAAAT**	**GTAAAC**	GTATCT	AAAAAT	**AATATA**	**ATAAAT**	**GTAAAC**	GTATCT
7	AAATTA	**TCAAGT**	**TTATTA**	GATGCT	AAGAAA	AAATTA	**TCAAGT**	**TTATTA**	GATGCT	**AAGAAA**
8	TTTACA	TAAAAT	**TCAAGT**	**TCAAGT**	TAGTGT	TTTACA	TAAAAT	**TCAAGT**	**TCAAGT**	**TAGTGT**
9	**ATACAA**	**TATACA**	TTTAAT	CACGGA	**TATAAA**	**ATACAA**	**TATACA**	TTTAAT	CACGGA	**TATAAA**
10	AAAATG	ATATAT	CGGCAG	TAGGGT	**CAAAGC**	AAAATG	ATATAT	CGGCAG	**TAGGGT**	**CAAAGC**
11	ATAAAA	AAAATA	ATATAG	CCGCGC	**GTGGGA**	ATAAAA	AAAATA	ATATAG	CCGCGC	**GTGGGA**
12	TTTAAA	TAAAAA	CTTACT	**CGGCTT**	TGTAAA	**TTTAAA**	TAAAAA	CTTACT	**CGGCTT**	**TGTAAA**
13	TGTAAA	**TGTACA**	**GGGGGA**	GCACTA	**TCAATG**	**TGTAAA**	**TGTACA**	**GGGGGA**	GCACTA	**TCAATG**
14	TTCAAA	**CTTAAA**	CGCGAG	GGATCC	**TGTGTG**	TTCAAA	**CTTAAA**	CGCGAG	**GGATCC**	**TGTGTG**
15	TAAAAT	TTTACA	AATAAA	**GTAGTG**	ATCAAT	TAAAAT	TTTACA	AATAAA	**GTAGTG**	ATCAAT
16	TTAAAT	TTTTAA	**TATTGC**	**TTTGTG**	CCAGCG	TTAAAT	TTTTAA	**TATTGC**	**TTTGTG**	CCAGCG
17	**CTTAAA**	**ATATAC**	**CCTGGG**	GGATCG	**CTGCAT**	**CTTAAA**	**ATATAC**	**CCTGGG**	GGATCG	**CTGCAT**
18	TACAAA	AAAACC	ACGGGT	**ACAGTA**	CTGCGT	TACAAA	**AAAACC**	**ACGGGT**	**ACAGTA**	CTGCGT
19	**TATACA**	TACATT	**GCACTT**	CGCGCC	**CTCTGA**	**TATACA**	TACATT	**GCACTT**	CGCGCC	**CTCTGA**
20	TTTCAA	AACCAA	GCTGCT	GTTCCG	GTCGGC	TTTCAA	AACCAA	**GCTGCT**	GTTCCG	**GTCGGC**
21	TCAAAA	**CAAGTT**	TCATGT	ACGCTG	TGCAAC	**TCAAAA**	**CAAGTT**	**TCATGT**	ACGCTG	**TGCAAC**
22	AAATAC	**AATGTT**	TCCCCC	TCGATC	GCACTA	AAATAC	**AATGTT**	**TCCCCC**	**TCGATC**	GCACTA
23	AAAAAA	TTCTAA	TAGTGT	CCGGCT	GCACGC	**AAAAAA**	TTCTAA	**TAGTGT**	**CCGGCT**	GCACGC
24	**TAAACA**	TTTAAG	TAATTA	AACGGG	**TTCCAT**	**TAAACA**	TTTAAG	TAATTA	AACGGG	**TTCCAT**

25	**ATCAAA**	**ACTTAA**	**CCCTCA**	TTCTAT	AAGCAT	**ATCAAA**	**ACTTAA**	**CCCTCA**	TTCTAT	AAGCAT
26	ATTAAA	TGTAAA	CGAAGC	CCGTAA	TCTGCG	ATTAAA	**TGTAAA**	CGAAGC	CCGTAA	**TCTGCG**
27	AAACAT	AAATAT	CCGTTT	GCTCCG	CAACGG	**AAACAT**	AAATAT	CCGTTT	**GCTCCG**	**CAACGG**
28	**ATACAT**	**AACTGC**	GCACGC	TCGCTC	**GCTGGC**	**ATACAT**	**AACTGC**	GCACGC	TCGCTC	**GCTGGC**
29	**AAAGAA**	ATGTAC	TCCCAT	GAACGC	TTAGTA	**AAAGAA**	ATGTAC	**TCCCAT**	GAACGC	TTAGTA
30	AAAAGA	TTCAAG	ACGAAT	CGATGC	AACGGG	AAAAGA	**TTCAAG**	ACGAAT	CGATGC	AACGGG
31	AAAAGC	TTTCAA	**TATATG**	CGATGG	TCAAAC	**AAAAGC**	TTTCAA	**TATATG**	**CGATGG**	**TCAAAC**
32	AATAAA	AAACAA	**CTACCC**	**CGTTGG**	**ATCAAA**	AATAAA	**AAACAA**	**CTACCC**	**CGTTGG**	**ATCAAA**
33	AAAACA	ATAAAA	TTTAAG	TGGTCC	GGAACA	**AAAACA**	ATAAAA	TTTAAG	**TGGTCC**	GGAACA
34	AAAATT	**TTCCTA**	GGTGAG	**ATACAC**	AATGCA	AAAATT	**TTCCTA**	**GGTGAG**	**ATACAC**	**AATGCA**
35	AAATAA	TAGTTT	**GGAGGG**	**AATCTC**	CAAACG	**AAATAA**	**TAGTTT**	**GGAGGG**	**AATCTC**	CAAACG
36	AAAATA	AAATGT	GGTAAT	ACGAGA	**GGCAGC**	AAAATA	AAATGT	GGTAAT	ACGAGA	**GGCAGC**
37	GAAAAA	**TACATA**	**AGTATT**	AAAGCG	TTCAAG	GAAAAA	**TACATA**	**AGTATT**	AAAGCG	**TTCAAG**
38	TTTTAA	**ATACAT**	TTATTT	CTACGT	**TTCAGC**	TTTTAA	**ATACAT**	TTATTT	CTACGT	**TTCAGC**
39	CAAAAC	**ATACAA**	ACGCGT	**CACTTA**	AGCGCA	CAAAAC	**ATACAA**	ACGCGT	**CACTTA**	**AGCGCA**
40	AAAAAG	ACCAAG	TGAAAC	**TTCTTC**	AAAGTT	AAAAAG	ACCAAG	**TGAAAC**	**TTCTTC**	**AAAGTT**
41	TAAATA	AATCAA	**TTGCTC**	AACCGA	AACCGA	**TAAATA**	**AATCAA**	**TTGCTC**	AACCGA	AACCGA
42	AGAAAA	TTAAGT	GCGTTC	CGGTAT	GAACAC	AGAAAA	TTAAGT	**GCGTTC**	CGGTAT	GAACAC
43	ATTTAA	TTAAGA	**AGGGAG**	TTATCG	**GGGGCA**	ATTTAA	**TTAAGA**	**AGGGAG**	TTATCG	**GGGGCA**
44	**AATTTA**	**GTGCAA**	TTTATA	CGCATA	TTCGGC	**AATTTA**	**GTGCAA**	TTTATA	CGCATA	TTCGGC
45	AAATGT	AAAGCA	TAAATT	CGGACG	AATAAG	AAATGT	**AAAGCA**	TAAATT	CGGACG	AATAAG
46	**TGAAAA**	TGATTT	CGTTCA	**GCTGCG**	CGCTCA	**TGAAAA**	**TGATTT**	CGTTCA	**GCTGCG**	**CGCTCA**
47	AACATA	**AGCAGC**	**GCTGGG**	CGGCGG	CCCATA	**AACATA**	**AGCAGC**	**GCTGGG**	**CGGCGG**	CCCATA
48	AATGCA	ATTCAA	**ATATCA**	TTCGCA	TACCAT	**AATGCA**	ATTCAA	**ATATCA**	TTCGCA	**TACCAT**
49	TTACAA	TACAAT	AAATAA	ACCTGT	**TTATTG**	**TTACAA**	TACAAT	**AAATAA**	**ACCTGT**	**TTATTG**
50	TAAATT	**AACAAA**	ACATGT	TTCGGC	AGTAGT	TAAATT	**AACAAA**	**ACATGT**	TTCGGC	**AGTAGT**

Table S4. The top 50 motifs from each of the following methods, along with their miRNA matches in miRBase: LM Bin, cWords2, miReduce, MixMir6. Left: Matches to exact seed sequence. Right: Matches allowing offset seed sequences. Light grey backgrounds indicate a match to miRBase, Orange indicates a match to a highly expressed miRNA found by all three experimental data sets (Jeker *et al.*, Sommers *et al.*, and Cobb *et al*). Green indicates a miRNA found by only the Cobb *et al.* data set. We take as highly expressed the miRNAs corresponding to the top ten unique motifs in each dataset.

Method	**% AU in motif**
MixMir2	75.0
MixMir3	71.0
MixMir4	66.0
MixMir5	53.67
cWords3	85.67
cWords4	84.33
cWords5	86.33

cWords6	86.67

Table S5. Percentage of A and U nucleotides in the top 50 motifs returned. As *k* decreases from 6 to 5, we see a decrease in the percentage of AUs.

Method	LM Bin*	MixMir6	MixMir6*
LM Bin	0.5666	0.2293	0.2227
LM Bin*		0.3097	0.3088
MixMir6			0.9993

Table S6. Pairwise Pearson correlations of motif rank, comparing LM Bin and MixMir. While motif rank was considerably changed by adding the UTR length covariate to the linear model, MixMir changed much less.

Method	**Number of significant motifs (p < 0.05)**	**Percent of positive coefficients**	**Percent positive coefficients (overall)**
LM Bin*	1792	75.56%	61.04%
MLM6*	439	86.33%	66.75%

Table S7. After incorporating a covariate for 3' UTR length (methods including the covariate marked with an asterisk), we found that the number of positive coefficients overall dropped significantly, particularly for the simple linear model (LM Bin). Similar to Table 2 in the manuscript, the number of significant motifs in the first column is determined by a cutoff of p < 0.05. The second column shows the percentage of motifs from the first column which have positive coefficients, and the third column shows the percentage of all motifs which have positive coefficients. Notably, the overall percentage of positive coefficients has dropped considerably for the linear model. However, MixMir6 has changed very little.

	E15.5		**E16.5**	
	Rank	**miRNAs**	**Rank**	**miRNAs**
MixMir	**2**	[1]miR-34b-3p, [1]miR-34c-3p	**1**	[1]miR-34b-3p, [1]miR-34c-3p
	5	[2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p	**3**	[2]let-7d-5p, [2]miR-202-3p
	8	[3]let-7d-5p, [3]let-7g-5p	**16**	[3]let-7d-5p, [2]miR-196a-5p
			31	[2]miR-30e-5p
			39	[3]miR-193a-3p, [3]miR-193b-3p
miREDUCE	**1**	[3]let-7d-5p, [3]let-7g-5p	**1**	[2]let-7d-5p, [2]miR-202-3p
	4	[1]miR-672-3p	**4**	[2]miR-362-3p, [1]miR-672-3p
Sylamer	**2**	[3]let-7d-5p, [3]let-7g-5p	**34**	[3]miR-10a-3p
	10	[2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p	**43**	[3]miR-18a-5p
	23	[2]miR-22-5p		
cWords	**1**	[2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p	**3**	[2]let-7d-5p, [2]miR-202-3p
	2	[3]let-7d-5p, [3]let-7g-5p	**9**	[2]miR-107-3p
			35	[3]miR-193a-3p
			46	[3]miR-34b-3p, [3]miR-34c-3p

LM Bin	**9**	[2]let-7d-5p, [2]let-7g-5p, [2]miR-202-3p	**3**	[1]miR-34b-3p, [1]miR-34c-3p
	13	[3]let-7d-5p, [3]let-7g-5p	**9**	[3]miR-193a-3p, [3]miR-193b-3p

Table S8. Comparison of all methods in analyses of adrenal cortex Dicer knockout data for mouse embryos at stages E15.5 and E16.5. We present matches to miRNAs found experimentally down-regulated in the Dicer KO compared to WT adrenal cortex samples, broken down for E15.5 and E16.5 separately. As in Table 4 in the manuscript, only the top 50 motifs returned by each method were analyzed.

Supplementary Figures

Figure S1. PP plot comparing the performance of cWords for various values of *k*, which determines the length of the words being corrected for. Very little correction for *p*-value skew was detected any value of *k*.

Figure S2. PP plots for various statistical methods after inclusion of a fixed 3' UTR length covariate, where the inclusion of a * denotes the amended model, showing that the LM Bin model improve dramatically. Little improvement was seen in MixMir6, suggesting that the length covariate was already implicitly corrected for.

Figure S1

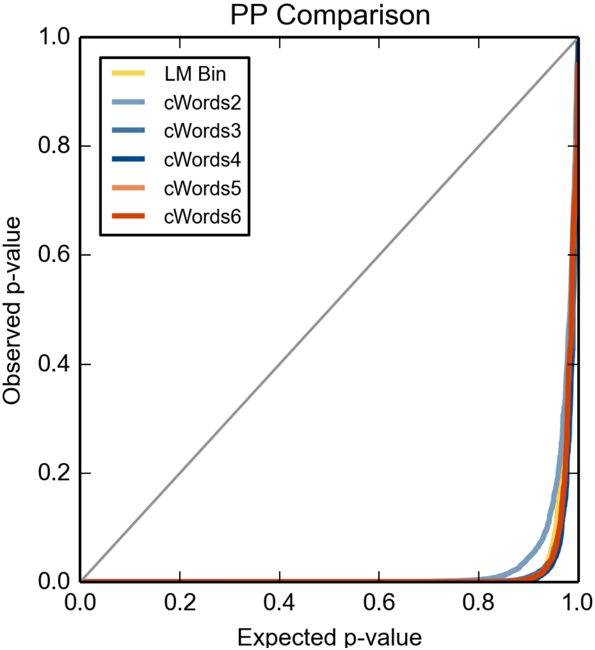
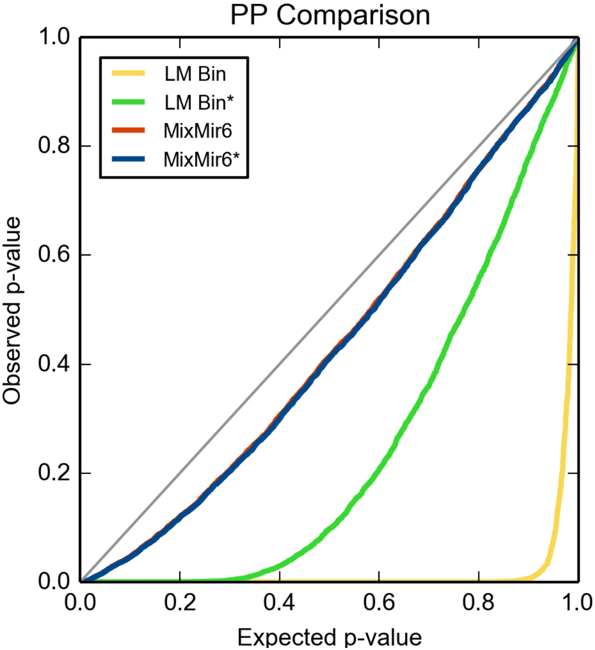

PP Comparison
LM Bin
cWords2
cWords3
cWords4
cWords5
cWords6
Observed p-value
Expected p-value
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Figure S2


PP Comparison
LM Bin
LM Bin*
MixMir6
MixMir6*
Observed p-value
Expected p-value
0.0
0.2
0.4
0.6
0.8
1.0