MixMir: microRNA motif discovery from gene expression data using mixed linear models

Liyang Diao[1], Antoine Marcais[2], Scott Norton[1,3] and Kevin C. Chen[1*]

[1] BioMaPS Institute for Quantitative Biology and Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[2] CIRI, International Center for Infectiology Research, Université de Lyon, Inserm, CNRS, Ecole Normale Supérieure, Lyon, France

[3] Department of Mathematics and Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA

*To whom correspondence should be addressed Tel: 1(732)445-1027 ext 40055; Fax: 1(732)445-1147; Email: kcchen@dls.rutgers.edu

ABSTRACT

MicroRNAs (miRNAs) are a class of ~22nt non-coding RNAs that potentially regulate over 60% of human protein-coding genes. MiRNA activity is highly specific, differing between cell types, developmental stages and environmental conditions, so the identification of active miRNAs in a given sample is of great interest. Here we present a novel computational approach for analyzing both mRNA sequence and gene expression data, called MixMir. Our method corrects for 3' UTR background sequence similarity between transcripts, which is known to correlate with mRNA transcript abundance. We demonstrate that after accounting for *k*mer sequence similarities in 3' UTRs, a statistical linear model based on motif presence/absence can effectively discover active miRNAs in a sample. MixMir utilizes fast software implementations for solving mixed linear models which are widely-used in genome-wide association studies (GWAS). Essentially we use 3' UTR sequence similarity in place of population cryptic relatedness in the GWAS problem. Compared to similar methods such as miREDUCE, Sylamer and cWords, we found that MixMir performed better at discovering true miRNA motifs in Dicer knockout CD4+ T-cells, as well as protein and mRNA expression data obtained from miRNA transfection experiments in human cell lines. MixMir can be freely downloaded from https://github.com/ldiao/MixMir.

INTRODUCTION

MicroRNAs (miRNAs) are small (~22nt) non-coding RNAs that post-transcriptionally regulate the expression of protein-coding genes (1). Their impact on gene regulation is the subject of intense study, with over 60% of all human genes estimated to be regulated by miRNAs (2) and some miRNAs potentially regulating hundreds of genes (3,4). Thus computational prediction of active miRNAs and their targets from gene expression data in a particular cellular context is of significant interest, leading to the development of a number of algorithms that analyze miRNAs jointly in the context of sequence and gene expression (5-9).

Animal miRNAs can bind to their targets in a variety of ways, centering on a 6nt region at the 5' end of the mature miRNA (bases 2-7) called the "seed" region (4). Most computational target prediction methods make use of exact Watson-Crick pairing of the seed region, as well as other features such as evolutionary conservation (10,11), co-occurrence of miRNA and RNA-binding protein binding sites (12), mRNA, miRNA, and Argonaute expression levels (5-7,9,13), 3' UTR sequence composition and other mRNA sequence features (5,6,12,13), and protein interaction data (14). A commonly-used computer program for analyzing miRNAs from both gene expression and sequence data is miREDUCE (7), which is based on the REDUCE algorithm for predicting transcription factor motifs (15). Two other published programs, Sylamer (5) and cWords (6), also solve the same problem while explicitly correcting for background sequence composition. The context of the miRNA binding site is known to affect binding efficacy (12), either through mRNA secondary structure or by providing binding sites for other post-transcriptional regulators. Background sequence similarity is also correlated with paralogy and therefore similarity between transcriptional programs. Note that it is not trivial in mammalian genomes to identify promoter or enhancer regions in order to directly correct for transcriptional control of protein coding genes.

Here we present MixMir, a novel method for miRNA motif discovery which, like Sylamer and cWords,

explicitly corrects for background sequence composition, but does so using a mixed linear model framework. In our MixMir implementation, we borrowed computational methods from the genome-wide association studies (GWAS) field for efficiently solving large systems of mixed linear model (MLM) equations. Our method uses the MLM to correct for similarities between 3' UTR sequences, analogous to the way MLMs are used to correct for cryptic relatedness between individuals in GWAS, where such artifacts can lead to high false positive associations (16-18).

We demonstrate the utility of MixMir on a gene expression data set in mouse wildtype and Dicer knockout CD4+25- T-cells (hereafter called CD4+ T-cells when the context is clear). We found that MixMir performed better than miREDUCE, Sylamer, and cWords at finding miRNAs annotated in the miRBase database (19) and highly expressed miRNAs in this cell type. We also confirmed our results on miRNA transfection experiments in human cell lines, both for quantitative proteomics data and microarray data (20). Thus we expect MixMir to be of practical use in helping experimentalists focus attention on the most important miRNAs in a given sample. Importantly, the most active miRNAs (i.e. those that play the biggest role in controlling mRNA expression in a particular cell type) are positively but imperfectly correlated with miRNA expression levels in the cell. Thus, it is not sufficient to simply assay miRNA expression levels in the cell to identify the most active miRNAs (see Discussion).

More broadly, our study suggests that mixed linear models are a powerful tool for motif discovery and highlight the importance of correcting for cryptic similarity in background sequence composition, an observation which might be useful in other motif-finding problems as well. Our MixMir software is open source and freely available from https://github.com/ldiao/MixMir.

MATERIAL AND METHODS

Experimental methods for obtaining the mRNA and microRNA expression data

Mice carrying a floxed Dicer allele in combination with CD4Cre transgene on a mixed C57BL/129 background (21) were maintained under specific pathogen-free conditions. Peripheral CD4+CD25- T cells were sorted on a FACS ARIA (Becton Dickinson) from 6-8 week-old mice and RNA extracted using RNAbee (AMSBio) according to the manufacturer's instructions. 100 nanogram of RNA was used to interrogate the GeneChip Mouse Gene 1.0 ST Array (Affymetrix). We obtained log fold changes in gene expression for 24,601 mRNA transcripts between WT and Dicer KO mouse CD4+ CD25- T cells.

We obtained three data sets of miRNA expression for CD4+CD25- T cells from three independent sources using different technologies. First, from the same cells from which we obtained our mRNA microarray expression data, we also obtained miRNA expression data in terms of the ratio of expression between CD4+CD25+ T-cells and CD4+25- T cells from Cobb *et al*. (22). Second, we obtained miRNA expression data from Jeker *et al.* (23). The authors extracted cells from female FoxP3-GFP-hCre miR-10a deficient mice and measured miRNA expression using Exiqon MirCury V9.2 arrays. Third, we used miRNA expression data from C57BL/6 mice determined by the nCounter miRNA expression assay kit (Nanostring Technologies), from Sommers *et al.* (24). The authors validated the Nanostring nCounter expression results with Exiqon microarrays and Taqman qRT-PCR assays.

pSILAC and microarray data for miRNA transfection experiments

In addition to our Dicer KO data, we also tested MixMir on pSILAC and mRNA microarray data from miRNA transfection experiments from Selbach *et al*. (20). The authors performed transfections by synthetic miRNAs and mock transfections in human HeLa cells for miRNAs let-7b, miR-1, miR-155, miR-16 and miR-30a. Paired pSILAC and microarray measurements were performed at both 8hrs and 32hrs post-transfection. The amount of protein synthesized in miRNA transfections vs. mock transfections was given by the log fold change ratio between the two experiments. Microarray analyses were performed with the Affymetrix Human Genome U133 Plus 2.0 chip. We used the microarray log fold change values taken at 32hrs post transfection for each miRNA transfection experiment.

We mapped the pSILAC Uniprot protein IDs to Refseq transcript IDs by downloading an ID mapping table from the Uniprot website. For the different transfection experiments, there were slightly different numbers of proteins with expression values, resulting in a range of the number of protein expression data points with

corresponding 3' UTR sequences from ~3000 - 3600 across all the transfection experiments.

Processing of the 3' UTR and miRNA sequence data

We downloaded all 26,845 mouse RefSeq gene 3' UTR sequences and 40,571 human 3' UTR sequences from the UCSC Genome Browser (version mm10 and hg19, respectively) (25,26). We removed all 3' UTRs of length less than 10nt and retained the longest isoform if there were multiple 3'UTR isoforms. In total, we were able to associate 17,988 unique UTR sequences to their microarray expression values for the mouse Dicer KO dataset, and 22,266 unique UTR sequences to their microarray expression values for each of the Selbach *et al*. miRNA transfection experiments.

We downloaded 1,908 mature mouse miRNA sequences corresponding to ~1200 distinct 6mer seeds and 2,578 mature human miRNA sequences corresponding to ~1500 distinct 6mer seeds from the miRBase database (release 20) (19).

Linear regression model of miRNA targeting

A naive linear model formulation of miRNA targeting is to regress the log fold change in gene expression against the count of the miRNA motif in the 3' UTR:

$$y_i = \beta_0 + \beta_1 miRNA_j + \epsilon_{ij}$$

where y_i is the log fold change in expression level of mRNA *i* and $miRNA_j$ is the number of times the motif for miRNA *j* appears in the 3' UTR of mRNA *i*, the β variables are constants and ϵ_{ij} is a Gaussian error term. We applied this simple linear model for all potential miRNA seeds (i.e. the set of all 4096 hexamers) across all mRNAs. The null hypothesis is that there is no miRNA effect, i.e. $\beta_1 = 0$. If the deviation of the inferred β_1 from 0 is statistically significant, then we say that the presence of the motif for miRNA *j* is significant. We also performed analyses for *k*mers with *k* ranging from 2 to 5 (data not shown). For *k* > 6, the computational demand was high for the mixed linear model so we do not report these results here. Similar computational efficiency issues with long motifs also occur with other motif finding programs, such as miREDUCE.

We implemented two versions of the simple linear model in R: the one given above, where the independent variable is the number of times the motif occurs in a 3' UTR and the one we use to report the results below with a binary variable which represents the presence or absence of a motif (i.e. without count information). The purpose of this categorical version of the simple linear model is for comparison with MixMir.

Mixed linear model of miRNA targeting

Our mixed linear model builds on the simple linear model above by adding an additional random effect to account for pairwise background sequence similarity between 3' UTRs. A random effect is a factor which can be modelled as being drawn from a probability distribution and is commonly used to model hierarchical structures in data (27). Here we take the set of all background *k*mer compositions of all 3' UTRs as the distribution and consider each 3' UTR to be a sample from it. Specifically, our mixed linear model is formulated as:

$$y_i = \beta_0 + \beta_1 miRNA_j + \alpha_i + \epsilon_{ij} \ ,$$

where α_i is the random effect for the *i*th mRNA and $miRNA_i$ is the binary variable representing the presence or absence of the hexamer motif associated with miRNA *j*. Rewriting the above in matrix notation gives:

$$y = M\beta + Z\alpha + \epsilon$$
$$Var(\alpha) = \sigma_g^2 K$$
$$Var(\epsilon) = \sigma_e^2 I$$

Here y represents the $T \times 1$ vector of mRNA expression changes, where T is the number of mRNAs, and Z is an incidence matrix (in our case $Z = I_T$). The most important part of the model is α, a vector of random effects, which incorporates a constraint on the covariance matrix via the $T \times T$ relationship matrix K, which we set to be the pairwise relationship matrix between mRNAs, as discussed below. σ_g^2 and σ_e^2 are called the variance components of the model.

We let $K_{ij} = cor(kmer_i, kmer_j)$, the pairwise Pearson correlation for the fractional *k*mer counts between mRNA *i* and mRNA *j*. The fractional *k*mer counts are obtained by taking the *k*mer counts and scaling by the sum of the total number of motif counts:

$$kmer_i = \{n_{i1}/S\,, n_{i1}/S\,, n_{i2}/S\,, \ldots, n_I/S\}$$

where n_{ik} is the number of times the motif m_k appears in the 3' UTR of mRNA *i*, $M = 4^k$ is the total number of possible *k*mers, and $S = \sum n_{ik}$ is the sum of the motif counts. We tested various *k*mer lengths in the construction of the relationship matrix in the range $k = 2,\ldots,6$.

We used both GEMMA v0.94 (28) and FaST-LMM v2.07 (29) to solve the mixed linear models. GEMMA computes a solution to the maximum likelihood for large mixed linear models and is an "exact" MLM solver in the sense that other fast solvers commonly used in GWAS often make simplifying assumptions specific to GWAS applications that are often not appropriate in the miRNA context. Note that our MLM implementation uses motif presence/absence information and is therefore more comparable to the categorical simple linear model as opposed to the counts simple linear model. This is due to the fact that GEMMA only accepts categorical information as input. Nonetheless, we show below that the binned and counts simple linear models performed very similarly on our test data set.

The differences between FaST-LMM and GEMMA are mainly in the details of their optimization algorithms. Briefly, FaST-LMM reparamaterizes the optimization problem in the mixed linear model to be a function of only a single parameter δ and then performs a spectral decomposition of the relatedness matrix once that can be used to test all motifs (or SNPs in the GWAS problem). GEMMA applies a different computational trick in the optimization step where first and second derivatives are obtained before the eigendecomposition. In principle the rankings of results between the two methods are identical though small differences in the implementations can result in slight differences in the results. Indeed, comparing results for the mouse Dicer KO data between GEMMA and FaST-LMM revealed nearly identical results between the two methods as expected (data not shown). We could not run GEMMA on the miRNA transfection microarray data, while FaST-LMM was able to complete the analyses. Since the results of the two algorithms are expected to be identical, we report results for the Dicer KO data using GEMMA and for the miRNA transfections using FaST-LMM.

Overview of the miREDUCE, Sylamer and cWords programs

miREDUCE takes the log fold change of gene expression between two conditions as input and outputs significant motifs and their associated miRNAs (7). The underlying algorithm is a forward stepwise linear regression, an iterative procedure where at each iteration the motif which minimizes the residual error in a simple linear model is selected and the residuals are taken as the new dependent variable for the following iteration. This continues until a significance threshold set by the user is exceeded. Forward stepwise linear regression procedures are known to suffer from many statistical problems, so the *p*-values from such procedures should be taken only as a general guideline (30). We thus set a liberal cutoff of $p = 0.50$ for miREDUCE for the purpose of comparing the miREDUCE motifs with the linear models described above.

We ran Sylamer via its web-based implementation, Sylarray (5,31). Given a gene list of *N* genes ranked in descending order of differential gene expression, Sylamer computes over- and under-representation of motifs in the top *T* genes vs. the remaining *N* - *T* in the list, as *T* is incremented in bin sizes of *b*. The hypergeometric distribution is used to determine the significance of the enrichment or depletion of a particular motif *m* in the top *T* genes, compared to the rest of the gene list, given that the total number of genes containing the motif in its 3' UTR is K_m. Sylamer corrects for background sequence composition by estimating expected motif counts based on the sequence composition of shorter motifs within each bin and using these values in place of K_m (5). We input a ranked list of genes based on differential expression in the Dicer KO cells.

cWords (6) corrects for background sequence composition using a *k*th order Markov model. The probability that a motif is enriched is calculated using the binomial distribution and the negative log of the probabilities is plotted to show enrichment across all ranked genes (genes with highest differential expression ranked first). Background enrichment values are computed as a sum of all such log probabilities, called the "running sum", and enriched motifs are those that have statistically higher sums than the background. We downloaded cWords from https://github.com/simras/cWords and ran it with word length = 6 and order of Markov background nucleotide model in the range 2-6.

RESULTS

Comparison of different MixMir and cWords parameters on the CD4+ T-cell microarray data

We started by testing five different settings of the *k*mer length in MixMir on the mouse Dicer-knockout CD4+ T-cell microarray data set (Methods). MixMir uses a linear model of miRNA targeting, similar to previous models such as miREDUCE, but adds an additional relationship matrix in a mixed linear model framework to correct for background sequence composition of 3' UTRs (Methods). The kmer length determines the construction of the relationship matrix in MixMir (Methods). We tested values of *k* from 2 to 6 and refer to these models as MixMir2 to MixMir6. For values of k above 6, the estimate of the correlation matrix became inaccurate because of the limited total amount of 3' UTR sequence in the genome and the running time of the implementation was slow, so we did not consider higher values of k further (Discussion).

We considered two types of mixed linear model - categorical and count linear models (Methods) – but found that the results were very similar so we used the categorical linear model for the remainder of the analysis. The value of *k* had a strong effect on the similarity of MixMir to the two baseline linear models we tested in which gene expression is simply regressed on motif counts or presence/absence (Methods), where similarity was defined by the Pearson correlation of the *p*-values of the motifs tested (data not shown). The similarity of MixMir to the linear models dropped as *k* increased in MixMir, with MixMir6 having little similarity with either of the linear models. These patterns were similar if we computed the Pearson correlation of motif ranks instead of motif *p*-values (Table S1).

We compared percentile-percentile (PP) plots of all of the MixMir models to each other and to the linear models, to determine how skewed the *p*-values were across all motifs for each method (Figure 1). Under the null hypothesis of no association, we expect the curves to fall on the diagonal line y=x, so the presence of skew away from this line is indicative of false-positive associations. These plots clearly showed that the MixMir models which performed the most correction of the *p*-values were MixMir5 and MixMir6. Note that this analysis implicitly assumes that there are relatively few highly active miRNAs in any particular cell type compared to the total number of possible miRNA seed sequences (in this case 4096 hexamers), an assumption we believe to be generally true biologically. Therefore, we selected MixMir5 and MixMir6 to represent the mixed linear model results in comparison with the other methods in our analysis.

We initially tested *k* = 2 to 6 for cWords to perform model selection as we did with MixMir, which we refer to as cWords2 to cWords6. We also compared these results to the simple linear models as we did with MixMir2 to MixMir6 (Results). We observed that as *k* increased, there was increasing similarity of cWords to the simple linear models, with cWords2 and cWords3 being the most different (Table S2). In fact we saw a dramatic drop in performance for *k*=5 and *k*=6, with many fewer matches to miRBase miRNAs and CD4+ T cell highly expressed miRNAs (see Methods; data not shown) than for *k*=2, 3, and 4. This result suggests that there was not enough data to learn a higher order Markov model in cWords.

Furthermore, we produced PP plots for the cWords2 through cWords6. These plots showed that there was a significant discrepancy between observed and expected *p*-values, similar to the simple linear models, suggesting a relatively high false positive rate for cWords on our data set. Little improvement was gained by using any *k*mer background correction (Figure S1), so we did not use this as a criterion for model selection, and relied instead on performance on the CD4+ T-cell data set alone. Thus, for further analyses, we retained just cWords2, as cWords3 performed similarly, and cWords4, 5, and 6 were very similar to the linear models.

MixMir5 and MixMir6 had the highest accuracy according to ROC curves

To compare MixMir to the previous motif discovery methods, we tested a total of five methods: the two simple linear models based on motif presence/absence (LM Bin) and motif counts (LM Count) which we take as our baseline methods, MixMir5, MixMir6, cWords2, and miREDUCE. For all of the linear models, all possible motifs were ranked by p-value; for miREDUCE, we set the p-value cutoff to be 0.5, resulting in 39 motifs returned (see Methods for a discussion of this choice of p-value cutoff). We omitted Sylamer because it returned only three overrepresented and three underrepresented motifs with corresponding miRNAs, making it difficult to compare with the other methods, which discovered much longer lists of motifs. We discuss the results from Sylamer separately (Supplementary Note).

We compared the significant hexamer motifs found by each method to miRNAs in miRBase (Methods). We performed two matching procedures to the miRNA sets. First, in our stringent matching criterion, we

considered a hexamer a match to a particular miRNA only if it matches the seed sequence of a mature miRNA. Second, in our relaxed matching criterion, we allowed the hexamers to match to any of three positions starting at nucleotides 1, 2, or 3 from the 5' end of the mature miRNA. We included offset match positions 1 and 3 in order to include all possible types of marginal binding site matches (1), including the potential for extensive complementarity through nts 1-8. This relaxed criterion also allows for shifts in the discovered motifs, which are common in practical applications of motif-finding algorithms to biological data. In general we expect to see more false positives when including matches to offset seed sequences, so for all comparisons we considered both the results from the stringent and the relaxed matching criterion. Additional A1 type site matches (i.e. a match in the first position to an A instead of the complementary base) are discussed in Supplementary Notes.

We present results for the two matching criteria using truncated receiver operating characteristic (ROC) curves (Methods) and analyze the results by computing an area-under-the-curve (AUC) value for each truncated ROC curve (Figure 2). Briefly we constructed the truncated ROC curves by taking the top 39 ranked motifs of each method, with true positives taken to be matches to any miRNA in miRBase (see Supplementary Note for details). We chose a cutoff of 39 because miREDUCE returned this number at a liberal cutoff of $p < 0.5$ (Methods). Since the other methods returned a *p*-value for all possible motifs, we made them comparable by selecting the top 39 motifs output by each method. We chose to truncate the full ROC curve, which is typically constructed over all possible 6-mer motifs, both because the methods did not return the same number of predictions and because we believe that focusing attention on only the top motifs is a more biologically meaningful comparison since only a few motifs are likely to be biologically relevant (i.e. only a small fraction of all possible miRNAs in the database are actually expressed in a cell). It is important both that truncating the ROC curve does not change the ranking of the methods and that we believe our results are robust in that the ROC curves for MixMir dominate the other curves over essentially the entire range of sensitivity settings (Figure 2). However, the truncated AUC value should not be interpreted as a typical AUC with a baseline value of 0.5 for a random method (see the next section for a more appropriate baseline value in our setting).

We present both ROC curves for relaxed motif matching as well as for stringent motif matching (Figure 2). We found that the truncated AUC values for the two simple linear models were low and that they found fewer miRNAs than both versions of MixMir, cWords2, and miREDUCE. Furthermore, cWords2 and miREDUCE did not perform much better than the simple linear models in terms of the truncated AUC. All four of those methods performed worse than both MixMir5 and MixMir6, which were more accurate over almost the entire range of sensitivity values. This effect was most dramatic when we used the strict motif matching criterion to position 2 only. These results suggest that MixMir more accurately identifies motifs corresponding to the exact miRNA seed region. The top 39 motifs and their corresponding miRNAs in miRBase are given in Table S4.

Randomizing the similarity matrices shows that MixMir is not overfitting the data

We considered the possibility that the performance of MixMir might be simply due to the increased complexity of the model versus the linear models (i.e. the additional similarity matrix) or perhaps even the particular algorithm used to optimize the likelihood of the mixed linear model. To control for this possibility, we ran MixMir using randomized data. We permuted both transcripts and motif count data while preserving the similarity matrix a total of 20 times, for both MixMir5 and MixMir6, and plotted the results alongside the original data (Figure 3). The average truncated AUROCs for the randomized MixMir methods was significantly smaller than for the original data for both stringent and relaxed matching to miRNAs. This result implies that the improved performance of MixMir over the other linear models was not merely due to an increase in the number of parameters of the mixed linear model. In addition the randomized trials also show that the baseline truncated AUC value for a random method is lower than the 0.5 value typically used in a conventional AUC value.

Validation of our computational results using experimental data sets of miRNA expression

As discussed above, one issue with the above analysis is that miRNAs are generally tissue-specific (32), and so comparing the predicted motifs to all miRNAs in miRBase, while informative, may not be the most biologically meaningful representation of their performance. We therefore further validated our results using miRNA expression levels in CD4+ T cells determined in an independent experiment by Jeker *et al.* (23). We found that miREDUCE, MixMir5, and MixMir6 all discovered miR-30, miR-26b, and miR-142-3p as highly active miRNAs in

this cell type. Overall, both MixMir models ranked these true motifs higher than miREDUCE, while the simple linear models and cWords found fewer matches to miRNAs expressed in this cell type (Table S3). These results are consistent with our previous analysis of the truncated ROC curves on the full miRBase miRNA data set.

Because miRNA quantification can sometimes vary between technologies, we further confirmed our analysis using two other miRNA expression data sets by Sommers *et al*. (24) and Cobb *et al*. (22). The former measured miRNA expression using the nCounter system (Nanostring Technologies) in CD4+ CD25- T-cells. The latter used miRNA microarrays to compare miRNA expression profiles between CD4+ CD25- and CD4+ CD25+ T-cells and the experiments were performed in the same laboratory and on the same cells from which we obtained the mRNA microarray data used in our analysis.

Several, but not all, of the most highly-expressed miRNAs (where we took only unique, exact seed sequence motifs from the miRNAs) in each of the three data sets overlapped. Between the Jeker *et al.* and Sommers *et al.* data, 5 of the top 10 exact seeds of highly expressed miRNAs overlapped, corresponding to miR-30b, miR-142, miR-16, let-7 and miR-29a. Similarly, the Jeker *et al*. and Cobb *et al*. data shared 4 of the top 10 exact seeds, corresponding to miR-30b, miR-142, miR-16 and let-7. The Sommers *et al*. and Cobb *et al.* data shared a slightly different set of highly expressed miRNAs, including miR-181a and miR-106a and missing miR-142 and miR-29a. This is consistent with differences between the studies, including the particular labs, quantification technologies and the comparison between two cell types in the case of the Cobb *et al*. data.

Nonetheless, when we took the highly expressed miRNAs found by all three methods, we found that these shared highly-expressed miRNAs tended to be discovered by MixMir, with the other motif-finding methods also performing quite well, but not as accurately (Table S4). Of the discovered highly-expressed miRNAs, miR-142 (both 3p and 5p) is particularly interesting as it has previously been found to be highly expressed in CD4+ CD25- T-cells and it plays a significant biologically role in regulating cAMP (33). These results suggest that MixMir tends to rank true miRNAs higher in the ranked list than other motif-finding methods, an important consideration for experimental groups that might only have the resources to validate a few top candidate miRNAs. It also shows that we were able to discover biologically meaningful results in our mouse Dicer-knockout CD4+ T cell data set.

MixMir and miREDUCE correct for AU bias in the motifs discovered

It is known that there is often an AU bias in computationally discovered motifs when using microarray data (34). The AU content in the 3' UTRs used in our analyses was 55.9%, while the average AU content in the miRNA seed sequences from miRBase was 48.8%. However, the motifs discovered by the simple linear models had very high average AU content, suggesting that their high false positive rate was partially due to discovering elements representing the AU-rich background sequence (Table 1). MixMir5 and MixMir6 motifs both had average AU content similar to that in the background 3' UTR sequence, suggesting that the correlation matrix component of MixMir successfully corrected for the AU bias. Consistent with this idea, as we altered the correlation matrix used in MixMir from $k = 2$ to $k = 6$, we observed a linear decrease in the average AU content of motifs as k increases (data not shown). The miREDUCE results also showed a strong correction for AU bias and even had a lower average AU content than the background 3' UTRs. cWords, on the other hand, had motif AU composition similar to those of the simple linear models, which was very high and was not significantly changed by altering the value of k. Taken together these results showed that simple linear models suffered from high AU bias, but this bias was corrected by MixMir and miREDUCE. Although miREDUCE does not have an explicit correction for 3' UTR base composition, it likely implicitly performs this correction by finding a motif highly correlated with background composition and then finding the residuals with respect to that motif to identify the remaining motifs. We observed this phenomenon in our data in practice, where miREDUCE often found an AU-rich motif as the most significant motif. We discuss the possible reasons for the AU bias in the Discussion section.

MixMir corrects for 3' UTR length and the discovered motifs are enriched for positive effects

We expect the coefficient of the fixed effect (i.e. the motif effect) to be positive if the motif represents the seed sequence of an active miRNA since miRNAs almost always downregulate their targets and a positive effect corresponds to higher expression in the Dicer KO. At first we observed that this was overwhelmingly true across

all motifs and methods and especially in the simple linear models (Table 2). We omitted cWords from this analysis because it does not produce association values to determine the direction of the regulatory effect.

However we reasoned that the overall very high enrichment of positive effects across all motifs in the simple linear models might be an artifact due to the inherent relationship between 3' UTR length and motif count, because longer sequences have a higher probability of containing any given motif, simply by chance. Thus an mRNA that is repressed due to a miRNA motif would also induce a similar correlation for all other motifs found in that 3' UTR. To test this hypothesis, we included 3' UTR length as a covariate to test how it would affect the direction of the miRNA effect. A full discussion of this the 3' UTR length effect can be found in the Supplementary Note. Briefly, the 3' UTR length covariate strongly shifted the *p*-values of motifs found by the simple linear models, which resulted in the PP plots for the simple linear models being significantly less skewed (Supplementary Material). These results suggest that an additional reason for the higher performance of MixMir compared to the simple linear models is that MixMir implicitly corrects for 3' UTR length using the relatedness matrix. After correcting for 3' UTR length, we found that the percentage of positive effects across motifs remained high but not artificially high. This is consistent with our biological intuition that while most significant motifs should have positive effects, some significant motifs will appear to have negative effects due to the indirect effects that are not captured by our steady-state microarray expression measurements. In any case, since we found that the additional length covariate did not change the rankings of the top 39 motifs in any of the linear methods, we did not use it for the comparisons between methods presented above.

MixMir performs at least as well as miREDUCE in recovering true seed sequences in miRNA overexpression experiments

Finally, in addition to testing MixMir on our mouse Dicer-knockout data, we also tested our algorithm on miRNA transfection data from human cell lines, to demonstrate that our results are not particular to the mouse microarray data set. We tested both microarray and quantitative protein expression data obtained from Selbach *et al.* (20) (see Methods), and compared our results to those obtained from the same data using miREDUCE (Table 3).

We found that both miREDUCE and MixMir were able to find the exact seed sequence for all the quantitative proteomics data sets. This is not surprising because unlike the Dicer-knockout scenario where many microRNAs were perturbed, the transfection experiment perturbs one microRNA very strongly and therefore is expected to produce much less noisy expression data.

In addition, we found that MixMir was also able find the exact seed sequence or an offset seed sequence (in the case of let-7b) of the transfected miRNA as precisely the most significant motif for each of the microarray experiments, while miREDUCE failed to do so for two out of five of the transfection experiments: let-7b and miR-16. In these two cases, miREDUCE ranked the seed sequence of the transfected miRNA as a significant match, but not the most significant in the list. We also found that MixMir was able to find many offset seed matches - all 3 offset seed sequences and one A1 match were found generally within the top 10 motifs. Additionally, we found motifs further downstream of the miRNA seed sequence for let-7b (rank 17, miRNA nts 12-17), miR-155 (rank 5, nts 4-9), and miR-16 (rank 16, nts 9-14), which may be suggestive of noncanonical binding in these miRNAs (35,36). The center of miR-16 has also been suggested to be involved in binding to AU-rich elements (37) although this result has been challenged (38). Since miREDUCE is a useful tool in experimental labs for validating that a transfection experiment actually worked and MixMir improves on miREDUCE slightly for several experiments, we believe this is an additional practical use of MixMir as well.

DISCUSSION

In conclusion, we have presented MixMir, a novel method for microRNA (miRNA) motif discovery from sequence and gene expression data. Our method corrects for pairwise sequence similarities between 3' UTRs that could confound a motif finding algorithm in a way that is fundamentally different from previous approaches to this problem (e.g. cWords, Sylamer). We applied MixMir to a microarray dataset from wild-type and Dicer knock-out (KO) mouse CD4+CD25- T cells collected by one of the authors. Since Dicer is required for miRNA biogenesis, we expect that Dicer KO cells do not contain any miRNAs and indeed this point was validated by quantitative PCR for selected miRNAs, showing a greater than 90% decrease in the knock-out (unpublished results). We

found that MixMir was more accurate in finding active miRNAs in these cells than three other similar published methods, miREDUCE, cWords and Sylamer, as well as a simple linear regression model we used as a baseline for comparison. We validated our computational predictions using three independent biological data sets consisting of miRNA expression measurements in this cell type quantified by either miRNA microarrays or single molecule imaging using the nCounter system (Nanostring Technologies).

Importantly we found that miRNA activity was highly but not perfectly correlated with miRNA abundance in the cells, so it is not sufficient to simply measure miRNA expression levels in a cell type to determine the miRNAs that play the largest role in shaping global gene expression in those cells. For example, as in similar analyses for transcription factors, miRNAs could be highly abundant but not highly active in repressing mRNA expression due to their sub-cellular localization or the presence of competing RNA species that could sequester the miRNAs from their mRNA targets (39). Another possibility is that miRNAs may have differential efficiency of loading into the RISC complex or of targeting mRNAs, and certain mRNAs may not be efficiently repressed by miRNAs due to the presence of either stable RNA secondary structures occluding the miRNA binding site or the binding of additional *trans*-acting factors. An interesting biological finding from our analysis is that the miRNAs that we found to be the most active in CD4+ CD25- T cells were in fact exactly the miRNAs that were more differentially expressed between these cells and CD4+ CD25+ T cells, based on previously published data from the same cell type (22).

To confirm the performance of MixMir on additional data sets, we tested MixMir against miREDUCE on five miRNA transfection experiments in HeLa cells, using both microarray and pSILAC quantitative proteomics data previously published by Selbach *et al*. (20). In all transfection experiments, for the microarray data, MixMir performed better than miREDUCE, ranking the exact seed sequence of the miRNA transfected first. In all transfection experiments, for the pSILAC data, both MixMir and miREDUCE found the exact seed sequence as the top candidate motif. This demonstrates the generality of MixMir on different types of data (microarray and pSILAC) and species (human and mouse). It also demonstrates the utility of MixMir in a context where miREDUCE is often used in practice – to verify that a miRNA transfection experiment was carried out successfully.

In our miRNA targeting model, we made several assumptions similar to previous methods, like miREDUCE. First, we searched over non-degenerate *k*mer motifs only. Although this does not rule out the possibility of detecting degenerate motifs, it probably biases our search towards non-degenerate seed matches. Although we searched for several published types of degenerate motifs such as G-bulge sites and imperfect sites in our data, we found only a few cases of such sites. We note that many of the analyses of non-canonical miRNA motifs have been performed on Ago HITS-CLIP or PAR-CLIP data and therefore represent biochemical binding events of the miRNAs, which are not necessarily perfectly correlated with repression that is detectable at the mRNA level. Similar observations hold for ChIP-seq data on transcription factors where biochemical binding does not necessarily produce transcription of the target gene. Second, we searched over motifs in 3' UTRs only. This choice was based on previous results in the literature but can be easily changed to examine other sequences, such as coding sequences or 5' UTRs, by users of MixMir. Third, our model assumes that the miRNA regulatory effect is additive, which is supported by previous evidence (1) but is still an approximation to biological reality.

Our approach to the motif discovery problem borrows an idea from genome-wide association studies (GWAS), namely that cryptic relatedness between individuals acts as a confounding factor that causes simple linear models to detect many false positive associations. In GWAS, cryptic relatedness is captured by a kinship matrix representing pairwise similarities between individuals. In the miRNA motif discovery problem, we considered background nucleotide composition similarity, which may affect miRNA binding in a variety of ways. It may affect binding site accessibility (12), represent other *cis*-regulatory sites for RNA-binding proteins, or simply be a correlate of paralogy – consider for instance ribosomal genes that are very similar and have similar expression patterns (e.g. due to similar transcriptional regulation) but are not affected by miRNA targeting (40). Such signals can confound a motif finder based on a simple linear model if sequence similarity is not corrected. In particular, we found that the relatedness matrix corrected for high AU content of the 3' UTRs. This observation could be due to the presence of AU-rich elements, which are known to be involved in mRNA regulation, other AU-rich motifs for *trans*-acting factors or more open secondary structures in the 3' UTR that might increase the

efficiency of miRNA binding.

We constructed a relatedness matrix analogous to the kinship matrix by representing *k*mer content similarity between 3' UTRs, which implicitly accounts for 3' UTR length. Our finding that $k = 5$ and $k = 6$ provided the most correction of the results are intuitive, as this choice of *k* corrects for motifs of the same length as the seed sequences for which we are searching and synergistic interactions between nearby miRNA binding sites and RNA binding protein binding sites have been previously documented (12,41,42). It is possible that we are also computing an approximation to the alignment score of the 3' UTRs and that global similarity of 3' UTRs is more important than the presence of short, 5-6 nt motifs, but we consider this possibility unlikely because very few pairs of 3' UTRs should have any meaningful sequence alignment at all. Most significantly, MixMir5 and MixMir6 were able to correctly implicate significant hexamer motifs associated with both known miRNAs as well as with highly expressed miRNAs in our dataset, as indicated using the area under the truncated receiver operating characteristic (AUROC). In particular, on the data sets we tested, MixMir performed better than a current state-of-the-art method of motif discovery, miREDUCE (7) over the entire range of sensitivity settings considered.

In addition to miREDUCE, we examined two recent methods, Sylamer (5,32) and cWords (6) (Methods), both of which correct for 3' UTR length and compositional biases. While Sylamer was able to identify three highly expressed miRNAs in the mouse CD4+ T-cell data, one match was due to an offset seed match and another was an A1 match. cWords did not perform as well on our data set, returning motifs with very strong AU bias, most of which do not appear to be high-confidence miRNAs in miRBase. These results suggest that background nucleotide composition similarity can strongly affect the ability of a linear model to uncover true motifs, but also that the way in which we correct for background composition can dramatically alter the results. Unlike Sylamer and cWords, MixMir utilizes the expression fold change values instead of just the ranks. Additionally, MixMir makes pairwise comparisons of the entire 3' UTR sequences, thus performing a more direct comparison of sequence context, rather than comparing motif vs. background composition within each 3' UTR like the other methods.

The MixMir software is freely available online along with the GEMMA software that solves the mixed linear model equations. FaST-LMM, another fast mixed linear model solver, can be used in place of GEMMA and can be obtained freely online (see the MixMir README file with the software for details). One limiting factor in our approach is the total amount of 3' UTR sequence available to construct large correlation matrices for long kmers. Increasing the kmer length to 7mers or higher would make the correlation matrix very sparse and difficult to estimate accurately. Another drawback of MixMir is its relative computational inefficiency: we exhaustively analyzed all 6mers but if we wanted to exhaustively analyze all 7- or 8mers, the runtime and memory requirements for GEMMA and FaST-LMM would make the computation too inefficient for practical use in our experience. However, miREDUCE suffers from a similar problem of computational inefficiency for values of k greater than about 6, so this is not an issue unique to MixMir.

Finally, we note that our mixed linear model approach is not limited to solving the miRNA motif discovery problem. Like the REDUCE software, MixMir can probably also be applied to other regulatory element motif detection problems, such as transcription factor and RNA binding protein motif prediction, by varying the type of sequence input and gene expression fold change input. For example, REDUCE was originally applied to transcription factors but was later applied to miRNAs in miREDUCE (7), RNA binding proteins in matrixREDUCE (43), degenerate transcription factor motifs in fREDUCE (44) and ChIP-chip data. We believe that MixMir can be similarly applied to many of these types of data and possibly also other data types such PAR-clip (45) as well .

ACKNOWLEDGEMENT

We thank Matthias Merkenschlager for providing laboratory space for producing the experimental data used in this study and helpful discussions. We also thank Marc Friedlaender, Dominic Gruen, Alexander Schliep and Jinchuan Xing for comments on this work.

FUNDING

This work was partially funded by the National Institutes of Health (R00HG004515 to K.C.C.). S.N. was supported by the Rutgers DIMACS REU program during the time that this work was performed.

REFERENCES

1. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215-233.
2. 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.
3. Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S. and Johnson, J.M. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769-773.
4. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.
5. 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.
6. Rasmussen, S.H., Jacobsen, A. and Krogh, A. (2013) cWords - systematic microRNA regulatory motif discovery from mRNA expression data. *Silence*, **4**, 2.
7. Sood, P., Krek, A., Zavolan, M., Macino, G. and Rajewsky, N. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A*, **103**, 2746-2751.
8. Lu, Y., Zhou, Y., Qu, W., Deng, M. and Zhang, C. (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, **27**, 2406-2413.
9. Stanhope, S.A., Sengupta, S., den Boon, J., Ahlquist, P. and Newton, M.A. (2009) Statistical use of argonaute expression and RISC assembly in microRNA target identification. *PLoS Comput Biol*, **5**, e1000516.
10. Lewis, B.P., Burge, C.B. and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.
11. Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat Genet*, **37**, 495-500.
12. Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, **27**, 91-105.
13. Huang, J.C., Frey, B.J. and Morris, Q.D. (2008) Comparing sequence and expression for predicting microRNA targets using GenMiR3. *Pac Symp Biocomput*, 52-63.
14. Le, H.S. and Bar-Joseph, Z. (2013) Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation. *Bioinformatics*, **29**, i89-97.
15. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet*, **27**, 167-171.
16. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. and Eskin, E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709-1723.
17. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, **42**, 348-354.
18. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, **44**, 821-824.
19. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, **39**, D152-157.
20. Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58-63.
21. Cobb, B.S., Nesterova, T.B., Thompson, E., Hertweck, A., O'Connor, E., Godwin, J., Wilson,

C.B., Brockdorff, N., Fisher, A.G., Smale, S.T. *et al.* (2005) T cell lineage choice and differentiation in the absence of the RNase III enzyme Dicer. *J Exp Med*, **201**, 1367-1373.

22. Cobb, B.S., Hertweck, A., Smith, J., O'Connor, E., Graf, D., Cook, T., Smale, S.T., Sakaguchi, S., Livesey, F.J., Fisher, A.G. *et al.* (2006) A role for Dicer in immune regulation. *J Exp Med*, **203**, 2519-2527.
23. Jeker, L.T., Zhou, X., Gershberg, K., de Kouchkovsky, D., Morar, M.M., Stadthagen, G., Lund, A.H. and Bluestone, J.A. (2012) MicroRNA 10a marks regulatory T cells. *PLoS One*, **7**, e36684.
24. Sommers, C.L., Rouquette-Jazdanian, A.K., Robles, A.I., Kortum, R.L., Merrill, R.K., Li, W., Nath, N., Wohlfert, E., Sixt, K.M., Belkaid, Y. *et al.* (2013) miRNA signature of mouse helper T cell hyper-proliferation. *PLoS One*, **8**, e66709.
25. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.
26. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.
27. Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance components*. Wiley, New York.
28. Zhou, X.a.S.M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, **44**, 821-824.
29. Listgarten, J., Lippert, C. and Heckerman, D. (2013) FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat Genet*, **45**, 470-471.
30. Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2 ed. Springer-Verlag.
31. Bartonicek, N. and Enright, A.J. (2010) SylArray: a web server for automated detection of miRNA effects from expression data. *Bioinformatics*, **26**, 2900-2901.
32. Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401-1414.
33. Huang, B., Zhao, J., Lei, Z., Shen, S., Li, D., Shen, G.X., Zhang, G.M. and Feng, Z.H. (2009) miR-142-3p restricts cAMP production in CD4+CD25- T cells and CD4+CD25+ TREG cells by targeting AC9 mRNA. *EMBO Rep*, **10**, 180-185.
34. Elkon, R. and Agami, R. (2008) Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. *PLoS Comput Biol*, **4**, e1000189.
35. Shin, C., Nam, J.W., Farh, K.K., Chiang, H.R., Shkumatava, A. and Bartel, D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell*, **38**, 789-802.
36. Martin, H.C., Wani, S., Steptoe, A.L., Krishnan, K., Nones, K., Nourbakhsh, E., Vlassov, A., Grimmond, S.M. and Cloonan, N. (2014) Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. *Genome Biol*, **15**, R51.
37. Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Di Padova, F., Lin, S.C., Gram, H. and Han, J. (2005) Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, **120**, 623-634.
38. Helfer, S., Schott, J., Stoecklin, G. and Förstemann, K. (2012) AU-rich element-mediated mRNA decay can occur independently of the miRNA machinery in mouse embryonic fibroblasts and Drosophila S2-cells. *PLoS One*, **7**, e28907.
39. Salmena, L., Poliseno, L., Tay, Y., Kats, L. and Pandolfi, P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353-358.
40. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. and Cohen, S.M. (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, **123**, 1133-1146.

41. Saetrom, P., Heale, B.S., Snøve, O., Aagaard, L., Alluin, J. and Rossi, J.J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res*, **35**, 2333-2342.
42. Jacobsen, A., Wen, J., Marks, D.S. and Krogh, A. (2010) Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Res*, **20**, 1010-1019.
43. Foat, B.C., Morozov, A.V. and Bussemaker, H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141-149.
44. Wu, R.Z., Chaivorapol, C., Zheng, J., Li, H. and Liang, S. (2007) fREDUCE: detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics*, **8**, 399.
45. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M. *et al.* (2010) PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp*.

TABLE AND FIGURES LEGENDS

Table 1. AU content of motifs discovered by the different methods. Simple linear models and cWords2 returned motifs with very high AU content. Both MixMir and miREDUCE had substantially lower average AU content, closer to the background 3' UTR base composition.

Table 2. Percentage of significant motifs that have positive coefficients in the linear model. The number of significant motifs in the first column is determined by a cutoff of $p < 0.05$. The percentage of motifs from the first column which are positive (i.e., the percentage of significant coefficients which are positive) is given in the second column. The third column is the percentage of all motifs which have positive coefficients, not limited to those which have been found to be significant.

Table 3. Comparison of miREDUCE and MixMir in analyses of miRNA transfection experiments, for (a) pSILAC quantitative proteomics and (b) mRNA microarray data. For columns two and three, the first number is the rank of the true miRNA seed sequence, and the number in parentheses is the match position, i.e. "2" represents an exact seed match, while "1" and "3" represent offset matches. "NA" indicates that the true motif was not found, either in the miREDUCE results with $p < 0.5$, or the top 20 motifs returned by MixMir. miREDUCE was run with a *p*-value cutoff of 0.5.

Figure 1. PP plot comparing the performance of MixMir with five different values of the kmer length *k* which defines the way in which the relationship matrix was constructed. We found that higher values of *k* were better at correcting for skewness in the PP plots (i.e. false positive predictions).

Figure 2. Truncated ROC curves comparing the six methods examined and their performance in ranking motifs from miRNAs in miRBase (see Supplementary Note). Left: Results when allowing for offset seed sequence matching, Right: Results when restricting to exact seed matches only.

Figure 3. Truncated ROC curves comparing MixMir5 and MixMir6 to the same models when run on randomized data. Matrices were randomized by permuting the connections between expression and genotypic data, while maintaining the same relationship matrix to preserve the distribution of relationship coefficients. Truncated AUC values of the randomized data were lower than for the true data, suggesting that simply incorporating a relationship matrix does not inflate the accuracy of the mixed linear model. Left: Results when allowing for offset seed sequence matching, Right: Results when restricting to exact seed matches.

TABLES

Method	% AU in motif
LM Bin	88.89
LM Count	88.46
cWords2	84.62
miREDUCE	44.02
MixMir5	55.98
MixMir6	52.56

Table 1.

Method	Number of significant motifs	Percent of positive coefficients	Percent positive coefficients overall
LM Bin	3863	99.97%	99.34%
LM Count	3848	100%	99.34%
MixMir5	654	98.17%	78.13%
MixMir6	430	90.23%	68.29%

Table 2.

pSILAC	miREDUCE	MixMir
let-7b	**1**[2]	**1**[2], **2**[1], **3**[3]
miR-1	**1**[2]	**1**[2], **4**[3], **11**[1]
miR-155	**1**[2]	**1**[2], **2**[3], **6**[1]
miR-16	**1**[2]	**1**[2], **2**[3], **3**[1]
miR-30a	**1**[2]	**1**[2]

Table 3a.

Microarray	miREDUCE	MixMir
let-7b	**6**[2], **22**[3]	**1**[3], **2**[2], **5**[1]
miR-1	**1**[2], **10**[3]	**1**[2], **2**[3], **3**[1], **6**[A1]
miR-155	**1**[2]	**1**[2], **3**[3]
miR-16	**3**[2], **29**[1]	**1**[2], **2**[3], **17**[1]
miR-30a	**1**[2]	**1**[2], **17**[3]

Table 3b.

Figure 1

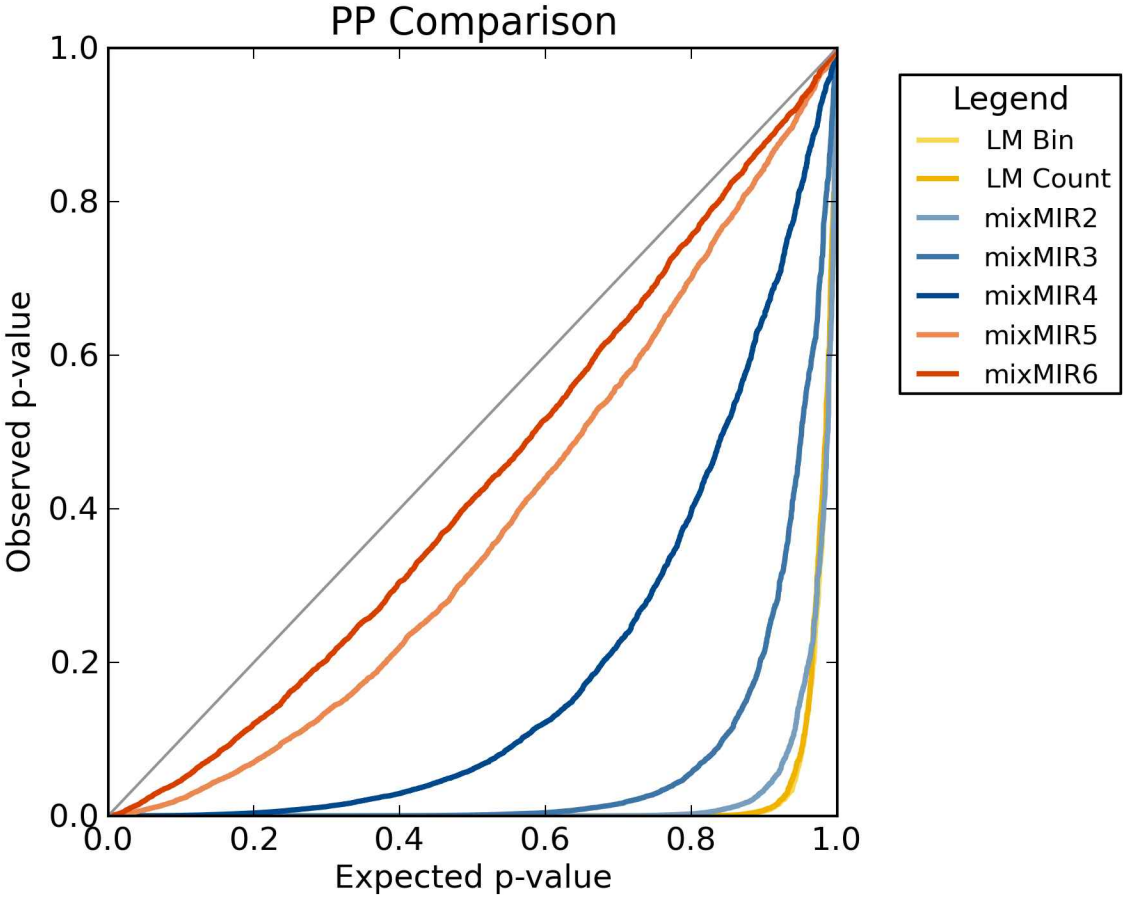

PP Comparison
Observed p-value
Expected p-value
0.0
0.2
0.4
0.6
0.8
1.0
Legend
LM Bin
LM Count
mixMIR2
mixMIR3
mixMIR4
mixMIR5
mixMIR6

Figure 2

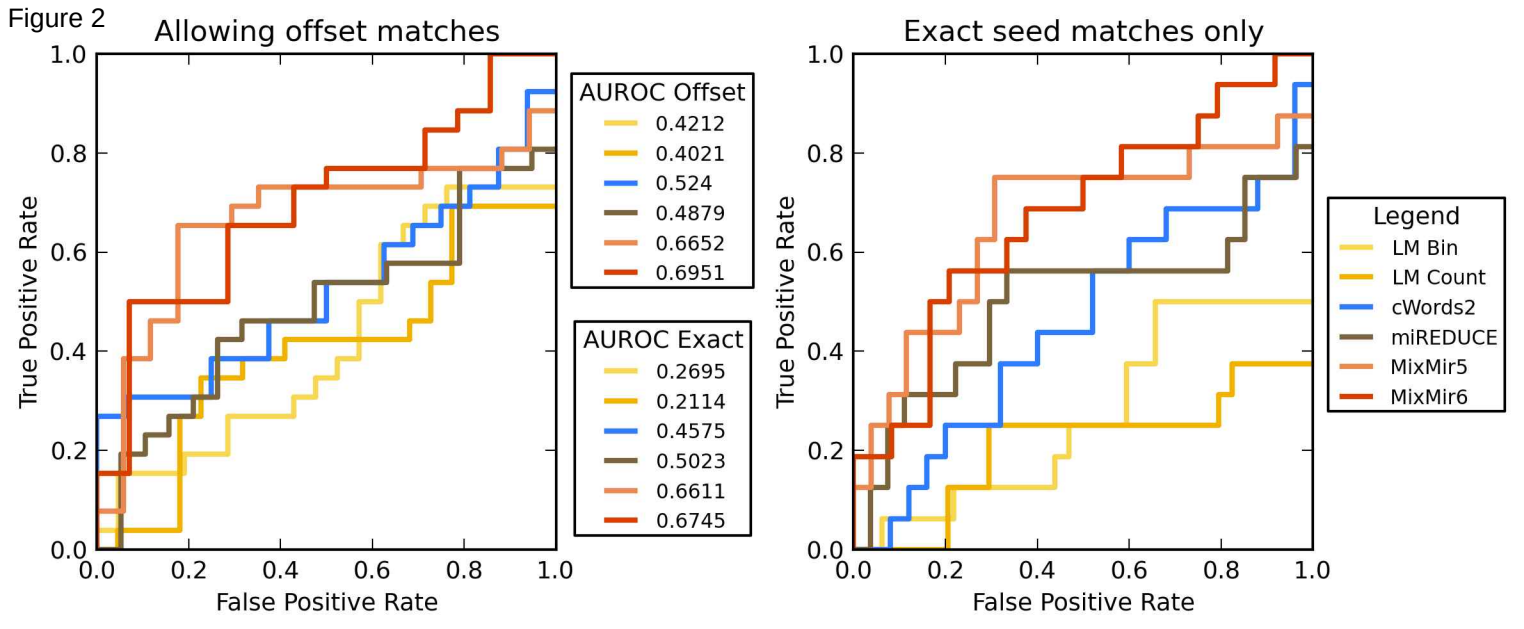
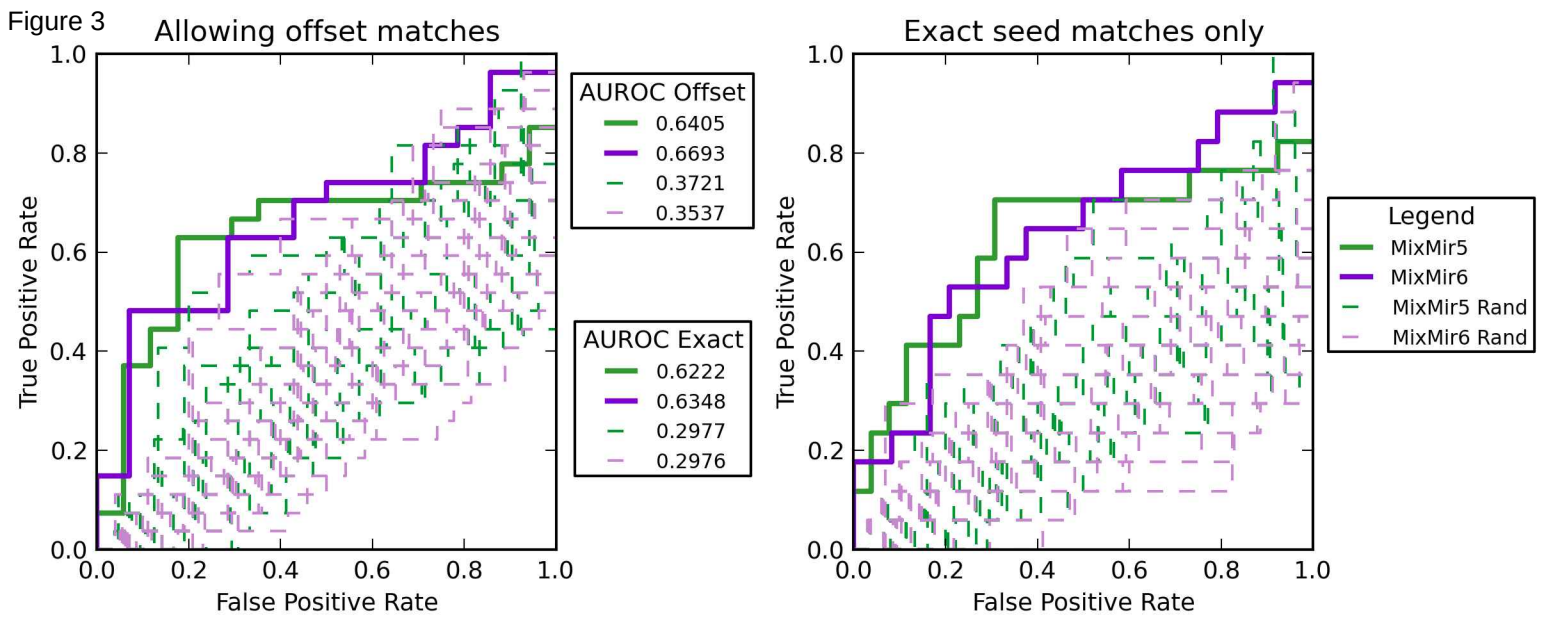

Allowing offset matches
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUROC Offset
0.4212
0.4021
0.524
0.4879
0.6652
0.6951
AUROC Exact
0.2695
0.2114
0.4575
0.5023
0.6611
0.6745
Exact seed matches only
Legend
LM Bin
LM Count
cWords2
miREDUCE
MixMir5
MixMir6

Figure 3


Allowing offset matches
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUROC Offset
0.6405
0.6693
0.3721
0.3537
AUROC Exact
0.6222
0.6348
0.2977
0.2976
Exact seed matches only
Legend
MixMir5
MixMir6
MixMir5 Rand
MixMir6 Rand