Supplementary Material: Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions

M. van Iterson S. Bervoets E.J. de Meijer H.P. Buermans
P.A.C. 't Hoen R.X. Menezes and J.M. Boer

1 The integrated analysis method

We propose an integrated analysis of miRNA and mRNA expression based on the global test [1]. The global test is a generalization for testing the global null hypothesis of a linear (or generalized linear) regression model $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ when the number of features exceeds the number of samples ($p >> n$).
In our integrated analysis the linear model with only an intercept is tested against the alternative model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. Here $\mathbf{y}_{n\times 1}$ represents the expression profile of a certain miRNA and $\mathbf{X}_{n\times p}$ the expression profiles of the predicted mRNA targets for that miRNA. The number of targets p is generally larger than the number of samples n.

A useful interpretation of the global test for the linear model is as a sum of squared covariances between predictors and responses (see section 5 of [1]). Consider the sample covariance, $r_{y,x}$ between a miRNA expression profile $\mathbf{y}_{n\times 1}$ and a single target $\mathbf{x}_{n\times 1}$ given by:

$$r_{y,x} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}_n)(x_k - \bar{x}_n) = \frac{(\mathbf{x} - \bar{\mathbf{x}}_n)^T(\mathbf{y} - \bar{\mathbf{y}}_n)}{n-1}, \tag{1}$$

where $\bar{y}_n$ and $\bar{x}_n$ denote the sample means of miRNA and mRNA expression profiles, $\bar{\mathbf{y}}_n$ and $\bar{\mathbf{x}}_n$ are vectorized versions (note that $r_{y,x} = r_{x,y}$). For multiple mRNA profiles $\mathbf{X}_{n\times p}$ the $p \times 1$ vector of the sample covariances, $\mathbf{r}_{y,\mathbf{X}}$ can be expressed as:

$$\mathbf{r}_{y,\mathbf{X}} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y}_n)(X_{kj} - \bar{X}_j) = \frac{(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{y} - \bar{\mathbf{y}})}{n-1}. \tag{2}$$

Note that this expression is valid even when the number of targets exceeds the number of samples $p > n$, and again $\mathbf{r}_{y,\mathbf{X}}^T = \mathbf{r}_{\mathbf{X},y}$. Goeman *et al.* [1] proposed

the following test statistics proportional to the squared sample covariance

$$\frac{(\mathbf{y}-\bar{\mathbf{y}}_n)^T\mathbf{X}\mathbf{X}^T(\mathbf{y}-\bar{\mathbf{y}}_n)}{(\mathbf{y}-\bar{\mathbf{y}}_n)^T(\mathbf{y}-\bar{\mathbf{y}}_n)} \propto \mathbf{r}_{y,\mathbf{X}}^T\mathbf{r}_{y,\mathbf{X}}. \quad (3)$$

The global test can be used not only for a continuous response, as explained here, but also for binary, multi-class, count and survival, and adjustment for confounders is possible [2, 3]. Moreover, they derived an asymptotic distribution for the test statistics, which is exact in case the linear model with normal errors is correct [3]. See Supplementary Figure S3 for a comparison of z-scores from the global test with pairwise correlations between three miRNAs and their predicted targets.

2 Prioritization of microRNAs and their targets: quantitative comparison of global test, correlation and lasso

2.1 Motivation

Several methods have been proposed to jointly analyse microRNA (miRNA) and mRNA expression data, taking also into account *in silico* predicted targets. Some of them are based on the strength and direction of the Pearson correlation coefficient between the expression profiles the miRNAs and targets [4]. Others have used lasso [5], a penalized regression algorithm, for example using miRNA expression profiles as responses and targets expression profiles as covariates [6, 7]. However, a quantitative comparison between these methods in this context is still lacking.

Here we perform a quantitative comparison between our proposed method, based upon global test, with using correlation or lasso. We do this by considering for each miRNA all mRNAs that are predicted as possible targets, as given by the overlap of three target prediction tools; TargetScan, MicroCosm and PITA. Firstly, we show how results using these three approaches are obtained and can be related in general, when a large number of samples is available. Secondly, we assess sensitivity and specificity of each method using a subsampling approach where the full data set is taken as the truth. The prostate cancer data described in the Methods section of the Main text is used in these comparisons.

2.2 Theoretical background

Correlations are statistics that are used to measure association between pairs of variables. The Pearson correlation is in particular often used to measure linear correlation. There exists a statistical test, due to [8], to assess how likely a Pearson correlation coefficient is to be observed if the variables are in fact uncorrelated. If that is the case, a function of the computed correlation is distributed as a Student's t-statistic with $n-2$ degrees of freedom, where n represents the

sample size. In the context of our problem, one correlation coefficient and its p-value is produced per miRNA:mRNA pair.

The lasso [5] is a method for fitting regression models so as to select the most relevant covariates to explain the outcome via shrinkage. The resulting fit yields thus an interpretable and parsimonious function of the covariates. Because its aim is interpretability, it may eliminate covariates that are associated with outcome, after retaining a single one that represents that association. In addition, the amount of shrinkage must be estimated empirically, so that different fits are not necessarily comparable. Furthermore, each lasso fit yields at most n non-zero coefficients, or n non-zero variables are selected, where n represents the sample size. This is both sensible and desirable for representation and interpretability purposes, but not if the objective is to find all (or as many as possible) active mRNA targets, especially in small studies when just tens of samples are available - some miRNAs may have hundreds of possible mRNA targets [9]. Finally, no statistical significance can be assigned to the resulting fit, so prioritization of miRNAs according to how likely they are of regulating mRNAs in the context under study is not straightforward.

The global test [1] was proposed as a test to assess association between an outcome and a set of variables. It can be seen as a test for the variance of a random effect relating the variable set and the outcome. As such, it can be put in the same context as ridge regression, another well-known penalized regression algorithm. However, as it focuses on testing the association, rather than on describing the association as ridge regression, it eliminates the need for estimating the amount of shrinkage. In addition, as the name says, it is a test and thus produces p-values which enable objective prioritization of miRNAs. Furthermore, for each miRNA the test statistic can be decomposed into the independent contributions of each mRNA towards the final statistic, allowing for prioritization of the mRNAs too. The global test statistic corresponds to the average of these independent, per miRNA:mRNA pair, test statistics.

Under the null hypothesis that there is no association between miRNA and the many mRNAs' expression, the global test statistic has asymptotically a distribution that is a weighted sum of independent chi-square distributions, each with 1 degree of freedom. If the test is used with a linear model, rather than a generalized linear model, this holds exactly. In particular, for a single miRNA:mRNA pair the test statistic has a chi-square distribution, under a linear model and the null hypothesis, and the test statistic itself is a function of the data via the Pearson correlation. Since the statistical test statistic for a Pearson correlation can be approximated by the normal distribution for large sample sizes, its square is also asymptotically distributed as a chi-square. With appropriate standardizations, thus, these two methods lead asymptotically to the same p-value for each miRNA:mRNA pair. So, in the context of linear models, the global test could be seen as an extension of Pearson correlation testing to the case where multiple miRNA:mRNA pairs are considered.

2.3 General comparison

Here we give an overview of results obtained with the three methods, using the prostate cancer data set involving 139 prostate cancer samples for which both miRNA and mRNA expression profiles are available (a total of 267 miRNAs and 20035 mRNAs were used). For the global test, we computed for each miRNA the global test statistic and its p-value, as well as the test statistics for each miRNA:mRNA pair involved. The Pearson correlations for the relevant miRNA:mRNA pairs were also computed. Finally, for each miRNA lasso was fitted using the predicted mRNA targets.

Figure 4 of the Main text summarizes these results. The miRNAs are ordered according to the global test statistic, with increasing significance from left to right, and the vertical line separates not significant (left) from significant (right) associations, all according to the global test. The stacked points represent the global test statistics separately for each target, coloured according to significance (black if significant, i.e. p-value < 0.001 after multiple testing correction using Benjamini-Hochberg's FDR; grey otherwise). The size of each point reflects the absolute correlation coefficient. Two things are immediately evident from the figure. Firstly, the ordering given by the correlation (square size) is indeed the same as the ordering given by the pairwise global test statistics (square height), as expected due to the relationship between the Pearson correlation coefficient and the per-pair global test statistic. This is evident by the fact that, for each vertical column of points, squares increase in size from bottom to top. Secondly, the global test significance helps separating the miRNAs according to the average association they display with target mRNAs: significant test results may be yielded by a handful of particularly large correlations, or by a larger group of medium-sized correlations. Making such distinctions is not straightforward using pairwise results alone.

Still in Figure 4 of the Main text, red squares indicate miRNA:mRNA pairs selected by lasso, so that in each vertical column of squares the red ones represent mRNA targets with a non-zero lasso-regression coefficient. It is clear that lasso does not always prioritize large correlation, with some mRNAs with small (and not significant) correlation being selected by lasso (red squares amongst mostly grey ones), and some miRNA:mRNA pairs displaying large correlation and zero lasso-regression coefficient (large black squares on the top). This can be explained by the fact that lasso aims at an interpretable and parsimonious representation, selecting one of any possible set of correlated mRNAs associated with the outcome miRNA, and thus neglecting individually relevant miRNA:mRNA pairs. In addition, lasso does not help to prioritize miRNAs playing a significant role in mRNA regulation: red squares are observed on both sides of the vertical line.

2.4 Assessing consistency in sensitivity and specificity

2.4.1 Subsampling

In order to compute a method's sensitivity (true positive rate) and specificity (1 - false positive rate), the truth must be known. In the context of miRNA:mRNA expression associations, existing databases include but a small number of validated targets to date, with this number further reduced if interest lies in one specific tissue or condition. In contrast, lists of predicted mRNA targets per miRNA amount to at least tens, if not hundreds of mRNAs per target prediction tool. Thus restricting results lists to pairs already validated greatly limits conclusions. A simulation study is normally used in such situations, but given that this complex biological problem is not yet well understood by biologists, involving many-to-many relations and various other issues, it is hardly possible to come up with a simulation setup anywhere near realistic. Furthermore, the three methods considered here are well-understood in the literature of statistical methods (see 'Theoretical background' of the previous section), so there is not a great deal to be learned from such a simulation study. What still lacks is an understanding of these methods' consistency for studies of various sizes.

To enable assessment of how consistent sensitivity and specificity measurements are in a realistic setting, we used here a subsampling approach on the prostate cancer dataset. This involved drawing different-sized subsets of the data proportional to 40%, 80% and 90% of the total number of samples. For each method, results obtained on each subset were compared to the results for the full dataset, assumed to represent thus the truth. In this way it is possible to calculate 'true' positive and 'false' positive rates per method. For each subset size, 25 subsets were used in order to obtain a sense of the variability of the estimates. Note, however, that the rates here computed are relative to the method, so care must be taken when comparing rates between methods.

2.4.2 Pearson correlation and the global test

For each miRNA, our approach yields a p-value derived from the global test. This indicates evidence or not for the miRNA's role in regulating expression levels of mRNAs predicted as its targets. The list of p-values computed using all samples is first produced. Then, for each subset, p-values are again computed for the same miRNAs and their lists of mRNA targets. Using different significance thresholds, we compute how many miRNAs are found to be significant or not, in the subset and in the full dataset. The Benjamini-Hochberg multiple testing correction procedure was applied to each generate list of p-values. So true and false positive rates here refer to proportions of miRNAs found in the full dataset that were also found in a subset, and proportions not found in the full dataset that were however declared significant in a subset, respectively.

Pearson correlation p-values were also produced for all relevant miRNA:mRNA pairs. We then needed to choose a rule to summarize results per miRNA: we declared that a miRNA played a role in mRNA expression regulation if its largest correlation yielded a p-value below a threshold.

Results are grouped per threshold used (significance level) and per subsample size (see Figure S4 A and B). In each case, a boxplot is made of true positive rates (coloured boxes) and false positive rates (white boxes with only contours coloured). It is clear from the figure that both sensitivity and specificity increase as either the subset size (given by the box colour) or the significance level increase, as expected. Also, the global test and Pearson correlation selections of miRNAs yield relatively the same sensitivity with respect to the full dataset if either 80% or more of the samples are chosen. When only 40% of the samples are used, the global test displays a larger drop in sensitivity than correlation. This is a consequence of taking as a summary the smallest p-value of all correlations per miRNA, which also yields more false positives: while the global tests yields a proportion of false positives in agreement with the significance level taken, selected miRNAs by Pearson correlation include many more false positives than expected. This illustrates well the advantage of taking a set of mRNAs together to perform the association test: only consistent associations results are found, and false positives are kept under the correct control level. Pairwise association measures, such as the Pearson correlation, are more likely to be influenced by individual pairs with large associations, which may well arise by chance.

Here a summary of results per miRNA had to be chosen for the Pearson correlation coefficients. Other alternative summaries could be used. However, any other arbitrary choice would likely have disadvantages due to first evaluating results per miRNA:mRNA pair, then summarizing per miRNA. Indeed, the global test has been shown to display more power against pairwise association testing, for alternative hypotheses often of practical interest [10]. Thus, we are confident that the conclusions above can be qualitatively extended to other methods involving summaries in two stages.

2.4.3 Lasso and the global test

Lasso regression results cannot be summarized per miRNA in a meaningful, practical way. All lasso fits yield at least one non-zero coefficient, so each miRNA always has at least one mRNA left in the regression fit. The number of non-zero coefficients remaining in the fit may also not be of relevance. So, for this comparison, we decided to focus on pairs of miRNA:mRNA selected, where for each miRNA only mRNAs that are predicted targets are considered, as before.

For each miRNA, the true positive rate represents the proportion of miRNA:mRNA pairs with a global test p-value below a threshold using the subset, compared with those found when using the full dataset. Similarly, the false positive rate is the proportion of pairs found in the subset, that were not found using the full dataset. As argued in section "Theoretical background", results for the Pearson correlation are identical to those of the global test (Figure S4 C). For the lasso, no threshold can be used, so each subset yields a list of mRNAs with non-zero coefficients, which is compared with the full dataset list to yield true and false positive rates (Figure S4 D).

The trend of increasing true positive rate is still seen for increasing subset sizes for both the global test and lasso. Also, false positive rates remain under

control for global test results, in spite of splitting the results per pair. Lasso keeps its false discoveries very much under control, compared to pairs found using the full dataset. So, we can conclude that both methods yield consistent results.

It is, however, surprising that true positive rates do not go above 80% when results per pair are considered, in contrast with when results per set are evaluated. Indeed, if we compare the graphs for the global test only (Figures S4 A and C), we can see that once it is used considering gene sets, true positive rates are reassuringly consistent with those of the full data set, but they drop considerably when using miRNA:mRNA pairs, for any choice of threshold and significance level . This illustrates another advantage of using our approach with gene sets: it yields results that are more consistent with the full data set.

2.5 Conclusion

The results of the quantitative comparison strengthen the choice of the global test for our approach for integrated analysis of miRNA and mRNA expression data. The global test has similar sensitivity compared to Pearson correlation but as important better specificity at an useful range of significance levels. For the prioritization of miRNAs lasso does not perform well, furthermore, the selected targets are not consistent in highly correlated targets.

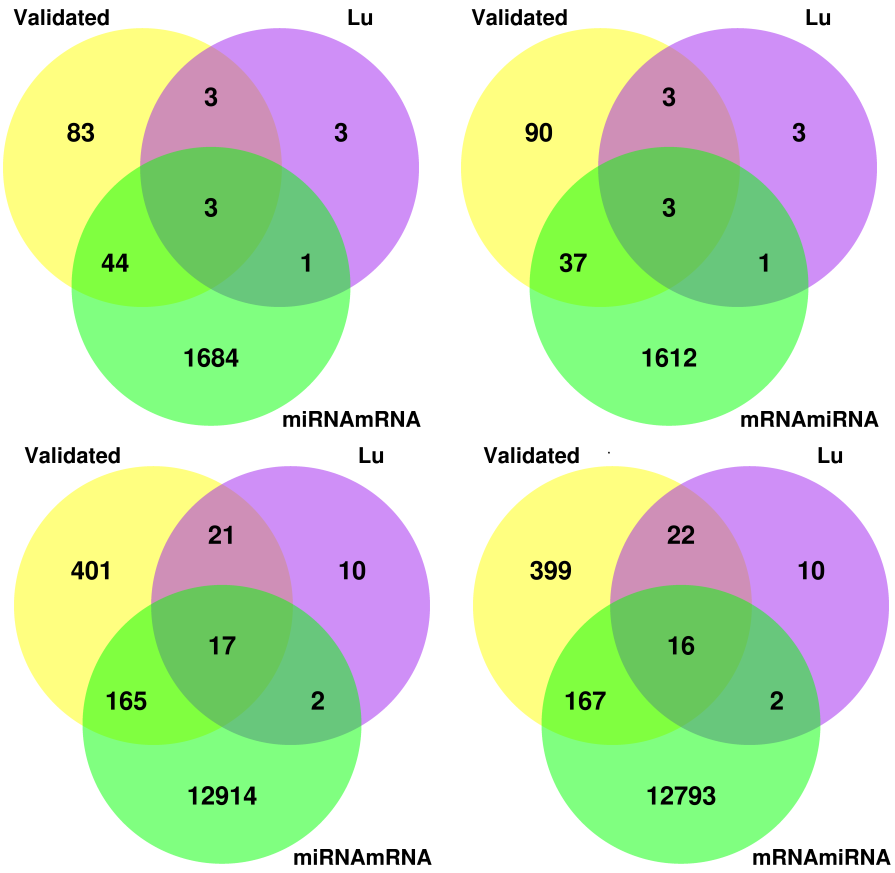

Figure S1: Venn diagrams showing the overlap between the significantly predicted miRNA:mRNA pairs with validated pairs from TarBase, miRTarbase and miRecords and with manually collected pairs specifically for prostate cancer by Lu *et al.*. In **A** the significantly predicted miRNA:mRNA pairs are obtained by predicting miRNA expression using the predicted targets from the strict overlap between TargetScan, PITA, microCosm. In **B** the reversed model (mRNAmiRNA) is used again with strict overlap of predicted targets. In **C** the same model as in **A** is used except now using partial overlap between TargetScan, PITA, microCosm and **D** is similar to **C** with the partial overlap.

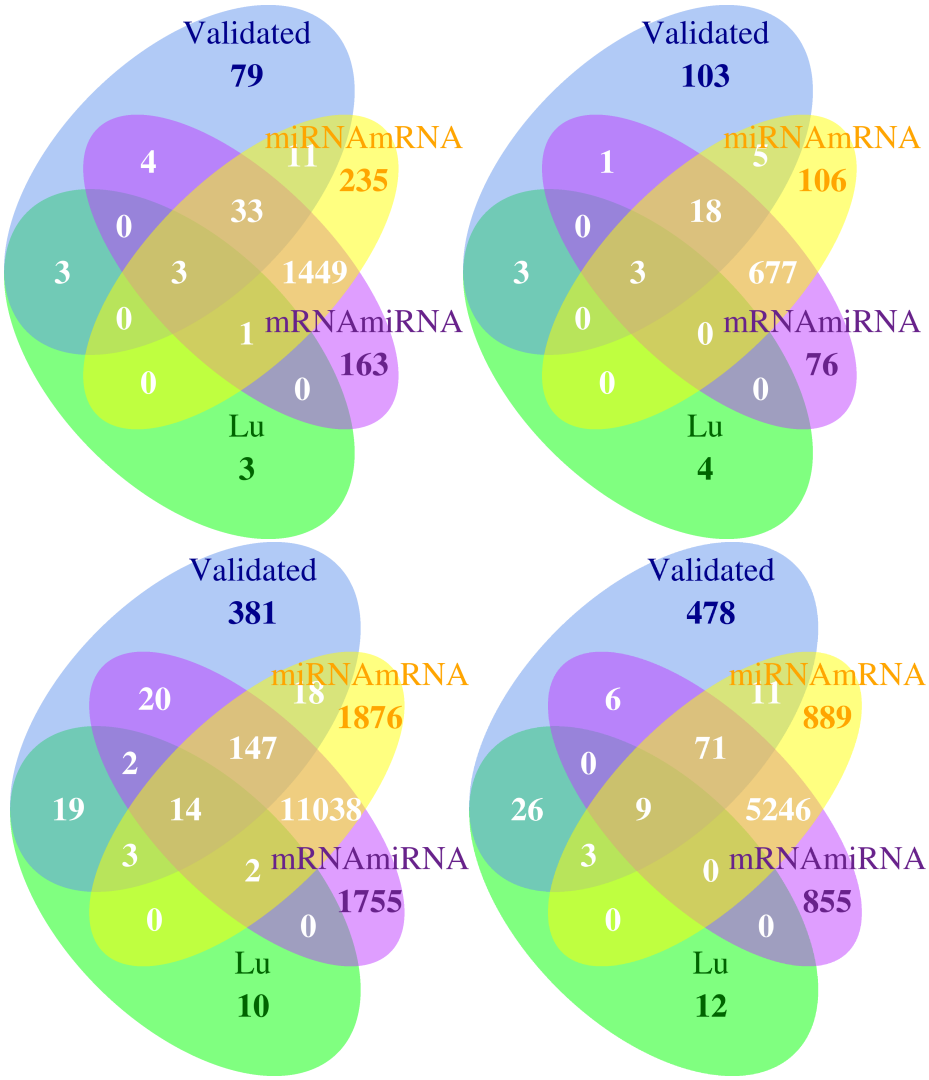

Figure S2: Venn diagrams showing the overlap between the significantly predicted miRNA:mRNA pairs using original (miRNAmRNA) and the reverse (mRNAmiRNA) models with validated pairs from TarBase, miRTarbase and miRecords and with manually collected pairs specifically for prostate cancer by Lu *et al.*. In **A** and **B** the significantly predicted miRNA:mRNA pairs are obtained using the predicted targets from the strict overlap between TargetScan, PITA, microCosm, where in **B** only negatively associated miRNA:mRNA pairs are shown. In **C** and **D** partial overlap between TargetScan, PITA, microCosm was used, where in **D** only negatively associated miRNA:mRNA pairs are shown. Actually, **A** combines Supplementary Figure 2a and 2b and **C** Figures 2c and 2d. **B** and **D** are similar to **A** and **C** except only negatively associated pairs are shown.

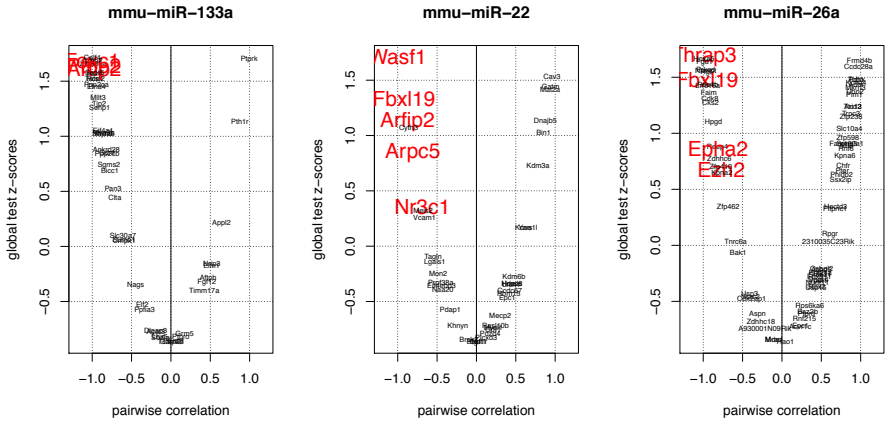

Figure S3: Comparison between the global test z-scores, representing the individual contributions for each target to the test statistic, and the pairwise miRNA:mRNA correlations. For each miRNA the overlapping targets between three prediction tools were used. Global test z-scores are proportional to the square of the pairwise correlation between miRNA and mRNAs. In red are the validated targets. **A** mmu-miR-133a with validated targets *Foxc1*, *Ptbp2* and *Arfip2*, **B** mmu-miR-26a with validated targets: *Epha2*, *Ezh2*, *Thrap3* and *Fbxl19* and **C** mmu-miR-22 with validated targets: *Wasf1*, *Arpc5*, *Nr3c1*, *Arfip2* and *Fbxl19*.

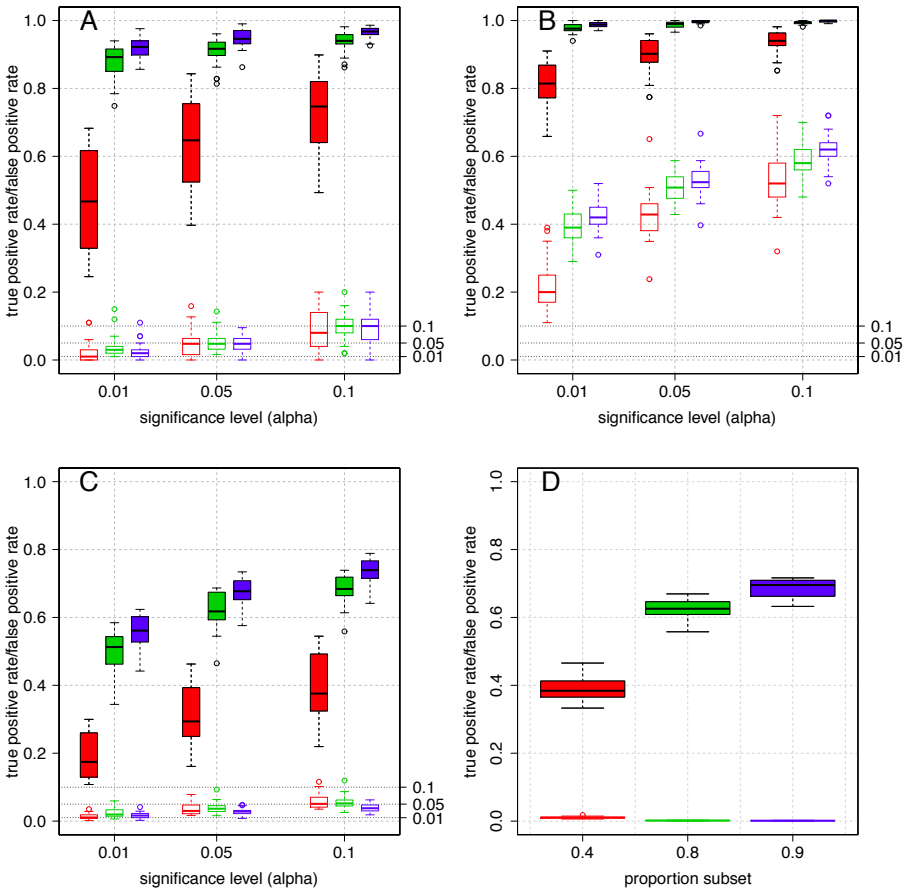

Figure S4: Assessment of consistency in sensitivity and specificity using global test, correlation and lasso for the integrated analysis of miRNA and mRNA expression data. A subsampling approach was used using the prostate cancer dataset. For each method, results obtained on subsets were compared to the results on the full dataset, assumed to represent the truth. The boxplots summarize true positive rates (solid) and false positive rates (transparent) of 25 subsets sampled from the full dataset for three subsets sizes 40%, 80% and 90% proportional to the original size of the data indicated in red, green and blue. True positive rates and false positive rates were calculated at three different significance levels 0.01, 0.05 and 0.1 using the Benjamini-Hochberg multiple testing correction (x-axis). Assessment of consistency in sensitivity and specificity for the selection of miRNAs **A** global test and **B** Pearson correlation (not possible for lasso). **C** and **D** present the assessment of consistency in sensitivity and specificity for the selection of targets respectively, for global test and Pearson correlation (results are identical) and lasso. See Supplementary Section "Prioritization of miRNAs and their targets: quantitative comparison of global test, correlation and lasso" for details.

Table S1: Overview of targets for mmu-miR-133a, mmu-miR-22 and mmu-miR26a that were validated using the luciferase assay, including information on the primers that were used for the cloning. The start position of the seed region is relative to the UTR start position of the target gene.

Target	**Entrez**	**Chr**	**miRNA**	**Validated**	**Seed :**		**Primer**[a] **:**		
					sequence	start	sequence	position	
Whsc2	24116	5	miR-133a	Care *et al.* [11]	UUGGUC	369	GCCATTTCTCTGGAGAGTTTAGGC GTTTGAAATGTTTACAACTGTAC	33,898,574–33,898,597[b]	33,898,242–33,898,264
Arfip2	76932	7		-[c]		470	CACTCCCTGGCCAATGGCAT AAAGGTTTATTCCTAGTGCTAG	105,635,631–105,635,652	105,635,940–105,635,959
Ptbp2	56195	3		-		979	GCCATCCTTAGTTTGTAATTAAG TGTTCAAATAAATTTGACCTGTAG	119,719,201–119,719,224	119,719,504–119,719,526
Foxc1	17300	13		-		524	GTAAGTTTCTTGCGTTCAGAG ATGATAGAAGGAGATTAATAC	31,809,488–31,809,509	31,809,181–31,809,202
Arpc5	67771	1	miR-22	-	AGCUGC	422	GACCCTTTCATAACCATGTC AGCATCATCTGGAGGCAAG	152,774,540–152,774,560	152,774,853–152,774,872
Fbxl19	233902	7		-		673	GAGGAGCAATTGGGGATCCGAGTG CCCATCTTTAAGCTGACATTTCC	127,768,896–127,768,920	127,768,418–127,768,441
Nr3c1	14815	18		-		584	ATGCATGGAAACCTGAAAAA AATTCCCCATGGAAGCAGA	39,414,050–39,414,069	39,413,726–39,413,745
Wasf1	83767	10		-		426	ACCTTAATTTTTCCCCCTGG CCTTTAAGAGAATTCAACACTACAAGC	40,938,327–40,938,353	40,938,052–40,938,071
Arfip2	76932	7		-		377	CACTCCCTGGCCAATGGCAT AAAGGTTTATTCCTAGTGCTAG	105,635,631–105,635,652	105,635,940–105,635,959
Epha2	13836	4	miR-26a	Wong *et al.* [12]	UCAAGU	25	GGTGGTACAGATGTCCAACG TCCAAGTTTCCCAGGCTCAGG	141,328,668–141,328,688	141,324,367–141,324,386
Ezh2	14056	6		Wong *et al.* [12]		247	GGAAATCCCTTGACATCTACTACC TCAACAACAAGTTCAAGTATTC	47,530,528–47,530,553	47,530,275–47,530,296
Fbxl19	233902	7		-		809	GAGGAGCAATTGGGGATCCGAGTG CCCATCTTTAAGCTGACATTTCC	127,768,896–127,768,920	127,768,418–127,768,441
Fbxl19	233902	7		-		780	GAGGAGCAATTGGGGATCCGAGTG CCCATCTTTAAGCTGACATTTCC	127,768,896–127,768,920	127,768,418–127,768,441
Thrap3	230753	4		-		1211	GTTGAAACATTTTCAGATGT CATCTGCCACTTCATTTATTG	126,164,086–126,164,106	126,164,370–126,164,389

[a] First line represent the forward primer sequence and position second line the reverse.
[b] Primers genomic positions were obtained by BLAT (UCSC genome browser) genome build GRCm38/mm10.
[c] To our knowledge these are not experimentally validated.

Table S2: Overview of miRNA mmu-miR-22 targets with strict overlap between the three prediction tools TargetScan, microCosm and PITA, including the individual P-value for association with mmu-miR-22 expression. The targets in bold are those that were used in validation experiments.

Negative associations			Positive associations		
Entrez	Symbol	P-value	Entrez	Symbol	P-value
83767	***Wasf1***	**0.00415**	12391	*Cav3*	0.03830
233902	***Fbxl19***	**0.07967**	67092	*Gatm*	0.05633
76932	*Arfip2*	0.12357	232087	*Mat2a*	0.06313
19159	*Cyth3*	0.13766	56323	*Dnajb5*	0.12344
67771	***Arpc5***	**0.18739**	30948	*Bin1*	0.14616
14815	***Nr3c1***	**0.31383**	104263	*Kdm3a*	0.21728
17536	*Meis2*	0.32297	66240	*Kcne1l*	0.36748
22329	*Vcam1*	0.34188	107271	*Yars*	0.36959
21345	*Tagln*	0.45586	216850	*Kdm6b*	0.52230
16852	*Lgals1*	0.47281	27281	*Hrasls*	0.54816
67074	*Mon2*	0.51303	70315	*Hdac8*	0.54844
230596	*Prpf38a*	0.54665	380916	*Lrch1*	0.55348
108112	*Eif4ebp3*	0.55761	234964	*Ccdc67*	0.57421
67877	*Naa20*	0.57213	229700	*Rbm15*	0.58935
231887	*Pdap1*	0.66938	13831	*Epc1*	0.60787
219094	*Khnyn*	0.76520	17257	*Mecp2*	0.69782
75770	*Brsk2*	0.90336	276952	*Rasl10b*	0.75741
12226	*Btg1*	0.97314	16918	*Mycl1*	0.78415
			18285	*Odf1*	0.78714
			245469	*Pdzd4*	0.82422
			239318	*Plcxd3*	0.87587
			56349	*Net1*	0.96627
			12978	*Csf1r*	0.97297

Table S3: Overview of miRNA mmu-miR-133a targets with strict overlap between the three prediction tools TargetScan, microCosm and PITA, including the individual P-value for association with mmu-miR-133a expression. The targets in bold are those that were used in validation experiments. The underlined targets could not be cloned.

Negative associations			Positive associations		
Entrez	Symbol	P-value	Entrez	Symbol	P-value
108013	*Celf4*	0.00224	19272	*Ptprk*	0.00593
17300	***Foxc1***	**0.00895**	19228	*Pth1r*	0.12300
13017	*Ctbp2*	0.00962	216190	*Appl2*	0.35621
29813	*Zfp385a*	0.01417	78593	*Nrip3*	0.47936
56195	***Ptbp2***	**0.02202**	243312	*Elfn1*	0.48227
76932	***Arfip2***	**0.02642**	216549	*Aftph*	0.52637
56526	*Sept6*	0.03019	14167	*Fgf12*	0.54326
23873	*Faim*	0.03089	21854	*Timm17a*	0.57684
19671	*Rce1*	0.04030	108071	*Grm5*	0.82700
13345	*Twist2*	0.04046	19277	*Ptpro*	0.87367
19052	*Ppp2ca*	0.05338	13803	*Enc1*	0.93127
13639	*Efna4*	0.05490	226896	*Tcfap2d*	0.99817
70122	*Mllt3*	0.07573	99326	*Garnl3*	0.99828
21873	*Tjp2*	0.08822			
223870	*Senp1*	0.09540			
13681	*Eif4a1*	0.14082			
17925	*Myo9b*	0.14631			
66940	*Shisa5*	0.14670			
17886	*Myh9*	0.14858			
105522	*Ankrd28*	0.18061			
14573	*Gdnf*	0.18734			
19053	*Ppp2cb*	0.19079			
74442	*Sgms2*	0.21377			
83675	*Bicc1*	0.22581			
72587	*Pan3*	0.26695			
12757	*Clta*	0.28853			
66500	*Slc30a7*	0.39067			
22218	*Sumo1*	0.40243			
66588	*Cmpk1*	0.40671			
217214	*Nags*	0.55364			
69257	*Elf2*	0.63642			
76787	*Ppfia3*	0.66704			
242667	*Dlgap3*	0.80598			
224530	*Acat3*	0.80903			
16873	*Lhx5*	0.86312			
56389	*Stx5a*	0.86889			

Table S4: Overview of miRNA mmu-miR-26a targets with strict overlap between the three prediction tools TargetScan, microCosm and PITA, including the individual P-value for association with mmu-miR-26a expression. The targets in bold are those that were used in validation experiments. The underlined targets could not be cloned.

Negative associations			Positive associations		
Entrez	Symbol	P-value	Entrez	Symbol	P-value
230753	***Thrap3***	**0.00975**	232288	*Frmd4b*	0.01550
15402	*Hoxa5*	0.01171	215814	*Ccdc28a*	0.02566
14163	*Fgd1*	0.01818	27402	*Pdhx*	0.04878
18753	*Prkcd*	0.03200	22057	*Tob1*	0.04902
59027	*Nampt*	0.03455	60613	*Kcnq4*	0.05639
15234	*Hgf*	0.03738	11535	*Adm*	0.05794
233902	***Fbxl19***	**0.04957**	22652	*Mkrn3*	0.06351
18578	*Pde4b*	0.05925	320538	*Ubn2*	0.07083
68732	*Lrrc16a*	0.06011	18712	*Pim1*	0.07706
23873	*Faim*	0.07397	234875	*Ttc13*	0.09932
264064	*Cdk8*	0.08502	77044	*Arid2*	0.10003
66197	*Cks2*	0.09182	22065	*Trpc3*	0.11484
15446	*Hpgd*	0.13005	30928	*Zfp238*	0.12068
72549	*Reep4*	0.18209	231290	*Slc10a4*	0.14205
13836	***Epha2***	**0.18782**	213753	*Zfp598*	0.16269
66980	*Zdhhc6*	0.20504	229488	*Fam160a1*	0.17340
52708	*Zfp410*	0.22480	320213	*Senp5*	0.17368
14056	***Ezh2***	**0.23016**	231051	*Mll3*	0.17701
16647	*Kpna2*	0.23713	74132	*Rnf6*	0.18469
242466	*Zfp462*	0.31819	16650	*Kpna6*	0.19984
233833	*Tnrc6a*	0.41264	231600	*Chfr*	0.22044
12018	*Bak1*	0.44428	19212	*Pter*	0.23132
235441	*Usp3*	0.59153	208177	*Phldb2*	0.24115
217154	*Stac2*	0.60107	99167	*Ssx2ip*	0.25254
13445	*Cdk2ap1*	0.61613	76608	*Hectd3*	0.31757
66695	*Aspn*	0.69023	71795	*Pitpnc1*	0.32328
503610	*Zdhhc18*	0.73335	19893	*Rpgr*	0.39004
77128	*A930001N09Rik*	0.78505	227446	*2310035C23Rik*	0.41007
17532	*Mras*	0.89128	228983	*Osbpl2*	0.49724
100019	*Mdn1*	0.90387	74159	*Acbd5*	0.49863
			192285	*Phf21a*	0.51123
			22241	*Ulk1*	0.51187
			17125	*Smad1*	0.51643
			225055	*Fbxo11*	0.52365
			238130	*Dock4*	0.53668
			106369	*Ypel1*	0.53955
			58242	*Nudt11*	0.54688
			71069	*Stox2*	0.56330
			14479	*Usp15*	0.56787
			67071	*Rps6ka6*	0.64564
			407823	*Baz2b*	0.67234
			19266	*Ptprd*	0.68593
			71673	*Rnf215*	0.71344
			13831	*Epc1*	0.76169
			269275	*Acvr1c*	0.76807
			15112	*Hao1*	0.95523

Table S5: Number of predicted miRNA:mRNA pairs, unique miRNAs and unique mRNAs for the prostate cancer data using the additional features of our approach a) strict versus partial overlap, b) reversing the model. For both approach we also summarized the results when only looking at the negative associations.

	Strict Overlap				Partial Overlap			
	Original model		Reversed model		Original model		Reversed model	
	all	negative	all	negative	all	negative	all	negative
pairs	1732	809	1653	775	13098	6229	12978	6187
miRNA	175	154	203	172	223	212	286	269
mRNA	1128	588	961	523	4845	3086	3876	2642

Table S6: Overview of miRNA target pairs with strict overlap between the three databases TargetScan, Microcosm and PITA using the original model strict_overlap_original_model.txt

Table S7: Overview of miRNA target pairs with strict overlap between the three databases TargetScan, Microcosm and PITA using the reversed model strict_overlap_reversed_model.txt

Table S8: Overview of miRNA target pairs with partial overlap between the three databases TargetScan, Microcosm and PITA using the original model partial_overlap_original_model.txt

Table S9: Overview of miRNA target pairs with partial overlap between the three databases TargetScan, Microcosm and PITA using the reversed model partial_overlap_reversed_model.txt

These tables are available from `http://www.humgen.nl/bioinf/iterson_et_al_2013_suppl_tables6-9.zip`

Table S10: Similarities and differences between recently proposed gene set methods for the integrated analysis of miRNA and mRNA expression data.

	Bossel Ben-Moshe *et al.*[4]	Engelmann *et al.*[6]	van Iterson
Method:	Pearson correlation	LARS (LASSO)	global test
Target predictions:	PITA, TargetScan, miRanda	microCosm, TargetScan, DIANA/microT and doRiNA (formely PicTar)	microCosm, PITA and TargetScan
Differential expression[a]:	fold-change filter	no	no
Model[b]:	-	mRNA $\sim$ miRNA, miRNA $\sim$ mRNA	miRNA $\sim$ mRNA, mRNA $\sim$ miRNA
Sets[c]:	mRNA	miRNA, mRNA	mRNA, miRNA
Overlap:	separate	partial	strict or partial
Association[d]:	both separate	both	both
Hypothesis test[e]:	no	no	yes
Ranking[f]:	yes	-	yes
Additional features[g]:	weights[h]	non-canonical	non-canonical, confounding, weights, permutations

- Indicates that this features is not meaning full for this method e.g. the method of Artmann uses a single target prediction tool so we can not speak about strict or partial overlap.

[a] Some methods first conduct a differential expression analysis on the miRNA and mRNA data, separately.

[b] Either the miRNA expression profiles are used as the responses in a linear model (miRNA $\sim$ mRNA) or the mRNA expression profiles (mRNA $\sim$ miRNA).

[c] All methods use *in silico* predicted targets to define gene sets, either miRNAs targeting the same gene or targets of the same miRNA.

[d] Some methods specifically test the association of up-regulated miRNAs with down-regulated mRNAs.

[e] Some methods require the computationally intensive bootstrapping in order to obtain p-values, others can directly calculate p-values using a theoretical null distribution.

[f] Our proposed method not only ranks the miRNA but also the targets.

[g] Some methods have specific additional features that others do not have, although these could be added, in principle, to the other methods as well.

[h] A GSEA-like approach is used were correlations and weights are combined.

References

[1] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[2] J. J. Goeman, J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957, 2005.

[3] J.J. Goeman, H.C. van Houwelingen, and L. Finos. Testing against a high-dimensional alternative in the generalized linear model: asymptotic type i error control. *Biometrika*, 98(2):381–390, 2011.

[4] N. Bossel Ben-Moshe, R. Avraham, M. Kedmi, A. Zeisel, A. Yitzhaky, Y. Yarden, and E. Domany. Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets. *Nucleic Acids Research*, 40(21):10614–10627, Nov 2012.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

[6] J. C. Engelmann and R. Spang. A Least Angle Regression Model for the Prediction of Canonical and Non-Canonical miRNA-mRNA Interactions. *PLoS ONE*, 7(7):e40634, 2012.

[7] Y. Lu, Y. Zhou, W. Qu, M. Deng, and C. Zhang. A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, 27(17):2406–2413, Sep 2011.

[8] E.J.G. Pitman. A note on normal correlation. *Biometrika*, 31(1):9–12, 1939.

[9] D. P. Bartel and C. Z. Chen. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics*, 5(5):396–400, May 2004.

[10] J.J Goeman, S.A. van dr Geer, and H.C. van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society Series B*, 68(3):477–493, 2006.

[11] A. Care, D. Catalucci, F. Felicetti, D. Bonci, A. Addario, P. Gallo, M. L. Bang, P. Segnalini, Y. Gu, N. D. Dalton, L. Elia, M. V. Latronico, M. Høydal, C. Autore, M. A. Russo, G. W. Dorn, O. Ellingsen, P. Ruiz-Lozano, K. L. Peterson, C. M. Croce, C. Peschle, and G. Condorelli. MicroRNA-133 controls cardiac hypertrophy. *Nature Medicine*, 13(5):613–618, May 2007.

[12] C. F. Wong and R. L. Tellam. Microrna-26a targets the histone methyltransferase Enhancer of Zeste homolog 2 during myogenesis. *Journal Biological Chemistry*, 283(15):9836–9843, 2008.