10614–10627 *Nucleic Acids Research, 2012, Vol. 40, No. 21* *Published online 12 September 2012*
doi:10.1093/nar/gks841

Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets

Noa Bossel Ben-Moshe[1], Roi Avraham[2], Merav Kedmi[2], Amit Zeisel[1], Assif Yitzhaky[1], Yosef Yarden[2] and Eytan Domany[1,*]

[1]Department of Physics of Complex Systems and [2]Department of Biological Regulation, Weizmann Institute of Science, Rehovot, 76100, Israel

Received March 20, 2012; Revised July 29, 2012; Accepted August 13, 2012

ABSTRACT

MicroRNAs (miRs) function primarily as post-transcriptional negative regulators of gene expression through binding to their mRNA targets. Reliable prediction of a miR's targets is a considerable bioinformatic challenge of great importance for inferring the miR's function. Sequence-based prediction algorithms have high false-positive rates, are not in agreement, and are not biological context specific. Here we introduce CoSMic (Context-Specific MicroRNA analysis), an algorithm that combines sequence-based prediction with miR and mRNA expression data. CoSMic differs from existing methods—it identifies miRs that play active roles in the specific biological system of interest and predicts with less false positives their functional targets. We applied CoSMic to search for miRs that regulate the migratory response of human mammary cells to epidermal growth factor (EGF) stimulation. Several such miRs, whose putative targets were significantly enriched by migration processes were identified. We tested three of these miRs experimentally, and showed that they indeed affected the migratory phenotype; we also tested three negative controls. In comparison to other algorithms CoSMic indeed filters out false positives and allows improved identification of context-specific targets. CoSMic can greatly facilitate miR research in general and, in particular, advance our understanding of individual miRs' function in a specific context.

INTRODUCTION

Background: microRNAs (miRs) are short, single-stranded non-coding RNA molecules which function as post-transcriptional negative regulators of gene expression. miRs act by recognizing complementary target sites in the 3′-UTR of their target genes, and consequently inducing transcript decay or translational arrest of their targets (1,2). Complementarity is mediated mainly by nucleotides 2–8 of the 5′-end of the miR, frequently referred to as the 'seed sequence' (3). Each miR can regulate hundreds of genes, and >30% of the mRNAs transcribed from human genes are predicted to be regulated by miRs (4). During the past 10 years the number of miRs that has been identified expanded enormously, and they were related to numerous biological processes, including development, cell-cycle control, differentiation and apoptosis (5).

One of the major difficulties in miR research is to unravel the function of a miR of interest and the pathways it regulates. Since there is no simple and widely used high-throughput experimental method for miR target identification, the amount of available information regarding miRs' function and their putative target genes is limited. A key factor for inferring the function of a miR is through its target genes. Therefore, several computational algorithms have been developed in the last few years in order to address this problem [such as PITA (6), TargetScan (4), miRanda (7) etc.]. These algorithms are based on a sequence similarity score, conservation and overall stability and accessibility of the miR–mRNA duplex. However, the current sequence-based available target prediction algorithms predict hundreds to few thousands of target genes for each miR, which makes it difficult to focus on a few likely targets of the miR of interest. Moreover, they are known to have high false-positive rates, and their predictions are not in agreement (8). A common procedure to overcome this problem is to intersect the results of several prediction algorithms in order to obtain a limited number of target genes for each miR, with less false-positive results. However, this procedure misses many *bona fide* targets, and hence although it has higher confidence it also has lower sensitivity (9–11). Although much effort was invested in improving sequence-based predictions [for most recent work see (12–19)], so far no significant progress has reached consensus.

*To whom correspondence should be addressed. Tel: +972 89343964; Fax: +972 89344109; Email: eytan.domany@weizmann.ac.il

An obvious problem with sequence-based methods is their generality. These algorithms are not taking into account biological context; for example, the top predicted targets of a certain miR might not be expressed at all in the specific tested model system. Thus, in spite of their high scoring by the sequence-based algorithm, they are not relevant to the specific model system (9). Our work was designed to address this issue, of context-dependent miR target prediction. It is fairly clear that in order to predict accurately the targets of a miR of interest with high sensitivity and specificity the sequence-based predictions have to be integrated with other kind of information. Since the problem is, on the one hand, unsolved, and on the other it is highly relevant and important, dozens of papers addressing the issue have been published in the last year. Several studies generated miR databases which contain sequence-based information along with lists of validated targets, expression data, signaling pathway resources and literature knowledge mining tools (20–28). A different approach was based on network analysis to identify signaling pathways associated with miRs (29,30).

Our approach is based on the belief that context-dependent functional targeting of a miR will be reflected in the expression data of its true mRNA targets (31,32). Therefore, we integrated another factor into miR target predictions: the correlation between the expression levels of the miR and the mRNAs. Here we propose an algorithm, Context Specific MicroRNA analysis (CoSMic), that combines experimental data from expression of mRNAs and miRs (measured in the same samples) with available sequence-based predictions. Combining these different kinds of information allows us to identify functional targets of miRs that play important roles in a specific experiment. As its output, CoSMic provides information about the statistical significance of the predictions, based on the enrichment of the high scoring sequence-based target genes (4,6,7) by the group of genes whose expression is highly correlated with the miR's (33). Hence CoSMic enables us to focus on the most significant candidate miRs for further investigation. Moreover, the number of predicted targets by the algorithm for each miR is only few tens, which is a reasonable number for further experimental validations and investigation. Last, we provide experimental evidence for the efficiency of CoSMic for finding functional miRs and their putative functional targets in a particular system of interest: induction of motility in an EGF-stimulated human mammary cell line. Our algorithm predicts the putative target genes of a miR more accurately and with less false positives than all other algorithms we tested, and allows the identification of functional context-specific target genes.

Brief review of recently developed related methods

Several methods combining sequence-based information with expression data have been developed in the past few years. Here we only list briefly the relevant methods (34–45)—see Supplementary Data for a detailed description of each. In 2006, Sood *et al.* developed a computational tool, miReduce (34), which correlates 3′-UTR motifs with changes in mRNA levels, to improve the sensitivity of target predictions. A few years later Dongen *et al.* introduced Sylamer (35), a method for detecting targets of a miR from expression data by assessing over or under representation of its seed region in the 3′-UTRs of a gene list, ranked by an expression-based criterion. Next, two other algorithms which integrated gene expression into their predictions were published—Sigterms (36) and CORNA (37). Both use the set of differentially expressed genes from a specific experiment, and perform enrichment analysis to determine whether this set of differentially expressed genes is enriched for targets of a particular miR (according to one of three sequence-based target prediction algorithms: TargetScan, PicTar or MiRanda). In 2010, Ulitsky *et al.* introduced FAME (38), a permutation-based statistical method that tests for over or under representation of miR targets in a set of co-expressed genes. All these algorithms (miReduce, Sylamer, Sigterms, CORNA and FAME) utilize only mRNA expression data and do not take into account miR expression. The potentially important association between the miR and mRNA expression levels is not used, and hence they lose key information which provides statistical evidence for a regulatory relationship between the miR and its putative mRNA targets. Moreover, Sigterms and CORNA use only the list of differentially expressed genes and disregard the level of change or profiles of gene expression.

The first algorithm that integrated both mRNA and miR expression data into the sequence-based prediction was GenMiR++ (39). GenMiR++ is a Bayesian model and learning algorithm, designed to explore functional miR targets. The algorithm outputs the posterior probabilities of whether a given miR putatively targets a given mRNA under the GenMiR++ model. Two other algorithms that also exploit the full-expression matrices of both the mRNA and miR expression data are MMIA (40) and MAGIA (41). In general, both algorithms intersect the group of predicted target genes of a specific miR (using one of the available sequence-based algorithms: TargetScan, PITA and PicTar) with the group of genes with inverse expression (MMIA) or anti-correlation (MAGIA) to the miR. Thus, both MMIA and MAGIA use sharp cutoffs for the statistical analyses and intersect the group of predicted target genes with the group of anti-correlated genes. Using this rigid approach might lose some putative targets, merely due to setting thresholds at some arbitrarily selected value. In addition, MMIA is suitable only for experiments with two conditions (e.g. control versus treatment), and therefore it is not applicable for datasets with more conditions (such as time-course experiments). During the last year several additional algorithms combining sequence-based target prediction with expression data were developed. Jayaswal *et al.* (2011) (42) proposed a two-step method for the identification of miRs–mRNAs relationships; the first step is the identification of miR and mRNA clusters and the second step is the estimation of association between the two types of clusters. Li *et al.* (2011) (43) suggested a computational approach to construct association networks between miRs and mRNAs, using partial least square (PLS) regression, without respect to any

sequence-based prediction information. Lu *et al.* (2011) (44) proposed a linear regression model to investigate one mRNA simultaneously regulated by multiple targeting miRs, with respect to their potential competition in binding sites. All these authors suggested approaches to improve target prediction, but they did not implement their methods and hence there is no tool readily available for the biologist to explore his own experimental data. Moreover, the results of these algorithms were not validated experimentally. Another algorithm, developed by Bang-Berthelsen *et al.* (2011) (45), is based on independent component analysis (ICA) that incorporates both seed matching and mRNA expression profiling. Bang-Berthelsen *et al.* do not consider miR expression data and hence, as CORNA and Sigterms, lose key information about the miR–mRNA regulation. In addition, no implementation is available for the biologist user.

We have made explicit comparisons of the predictive power of CoSMic with purely sequence-based predictions and with the five algorithms that use also expression data: GenMiR++, MAGIA, FAME, miReduce and Sylamer.

The added value provided by CoSMic over the other prediction algorithms that combine sequence-based information with expression data is summarized as follows:

First, CoSMic differs from the other methods in that it initially identifies miRs that play active roles in the specific biological system of interest, in addition to the identification of their functional and context-specific target genes. This feature is important when no prior knowledge is available about the miRs that play active and significant roles in the system of interest, and CoSMic may direct the biologist towards them.

Second, we provide experimental validation for CoSMic results, both for the identification of the significant miRs in a particular system (EGF-induced motility in a human breast cell line) and for their functional targets.

Third, the thresholds used by our algorithm are data driven; there is no sharp cutoff on the correlation or intersection of predicted target genes with correlated genes, instead we optimize a gene set enrichment procedure to get the group of correlated genes that are enriched in the sequence-based predictions. In addition, as opposed to the other prediction tools that combine sequence-based information with expression data, CoSMic takes into consideration in the enrichment analysis not only the identities of the genes identified as targets of a miR, but also their corresponding sequence-dependent scores.

Fourth, we implemented our algorithm as an easy-to-use stand alone software to allow biologists to apply it for analysis of their data.

Last, our algorithm considers not only negative correlations, but also positive correlations as an indicator for miR–mRNA direct regulation.

Thus, we offer CoSMic and the corresponding experimental design (Figure 1) as a global strategy for unveiling the functional significance of miRs in a given biological system. The CoSMic algorithm is freely available at http://www.weizmann.ac.il/complex/compphys/software/cosmic/ (27 August 2012, date last accessed).

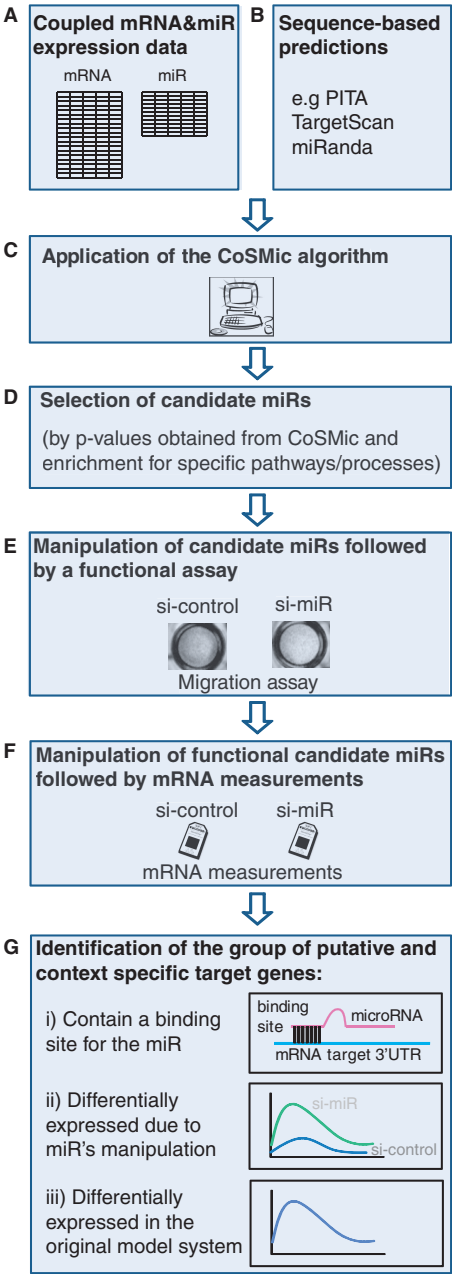

Figure 1. Flow chart of the experimental design. (**A**) Dataset of coupled mRNA and miR expression measurements from the same samples. Using predefined thresholds of expression across all samples and fold change, we filter genes and miRs that are expressed and changed in this specific model system. (**B**) Choosing one of the available sequence-based prediction algorithms (e.g. PITA, TargetScan or MiRanda). (**C**) Application of the CoSMic algorithm to the coupled mRNA and miR dataset. (**D**) Selection of candidate miRs based on the *P*-values obtained by the algorithm and on enrichment of the group of predicted target genes by specific pathways and processes that characterize the specific model system. (**E**) Silencing a few candidate miRs, and assessing their effect on the system, using a functional assay relevant to this specific model system. (**F**) Silencing the functional candidate miRs (these miRs that were found to affect the system in the previous step), and measuring the mRNA expression level after their perturbation. By this step we will be able to identify the group of putative target genes that mediate the effect on the system. (**G**) The group of putative and context-specific target genes should (i) contain a binding site for the miR, (ii) be differentially expressed due to the miR's manipulation and (iii) be differentially expressed in the original model system [i.e. the dataset in (A)].

MATERIALS AND METHODS

Expression data

CoSMic was applied on mRNA (46) and miR (47) expression data of human mammary (MCF10A) cells after EGF stimulation. Expression level threshold (EL) and fold change threshold (FC) were applied on the measured mRNA and miR expression data in order to detect significant changes above noise level, and to include in CoSMic only miRs and genes that were expressed and modulated due to EGF stimulation. For the mRNA data thresholds of $EL > 8$ and minimal $FC \geq 2$ were used (46), and for the miR data $EL > 10$ and minimal $FC \geq 1.4$. Changing the values of the miR thresholds by up to 50% does not significantly alter the results.

The CoSMic algorithm

The main idea of CoSMic is to integrate two kinds of gene-expression measurements—of mRNAs and miRs, with sequence-based miR target predictions. This procedure generates biological context-specific target predictions of improved reliability. The algorithm treats the miRs separately, one at a time, to predict the miR's putative target genes.The idea that underlies the algorithm is simple: identify as targets of a miR those genes which have high sequence-based scores and significant correlations of their expression levels with those of the miR. The algorithm defines this group of target genes and also calculates the corresponding *P*-value. In addition, it also identifies miRs that are functional in the specific biological context and assigns a false discovery rate (FDR)-corrected *P*-value to this identification.

As its first step, CoSMic extracts from the sequence-based prediction algorithm of choice (PITA, TargetScan or MiRanda) the predicted target genes of the miR, and sorts them by their sequence-based score (Supplementary Figure S1A). Next, it calculates the Spearman correlation coefficients and corresponding *P*-values between the expression levels of the miR and every mRNA across all samples, and sorts all mRNAs by their correlation to the miR (Supplementary Figure S1B, this is done separately for the positively and negatively correlated genes). Then CoSMic uses a strategy which is closely related (but not similar) to that of the gene set enrichment analysis (GSEA) (33) to calculate the enrichment of the top ranked sequence-based predicted target genes by the top ranked correlated genes (see 'Supplementary Methods' section for more information). The optimal set of high scoring and highly correlated genes is identified as the miRs targets. Last, a *P*-value is calculated for each miR using a random model, and then, using the FDR procedure, we correct for multiple testing and assign a *q*-value to each miR. This *q*-value represents the significance of the target prediction as well as the significance of this miR in the specific tested model system.

We used the CoSMic algorithm with Agilent microarrays for the miR expression measurements and Affymetrix exon arrays for the mRNA data, but any type of microarrays can be used, and also any other type of high-throughput expression data can be used by CoSMic [e.g. sequencing data and or qRT–polymerase chain reaction (PCR) expression data].

The CoSMic algorithm is freely available at http://www.weizmann.ac.il/complex/compphys/software/cosmic/ (27 August 2012, date last accessed).

Enrichment for migration gene ontology (GO)-terms

P-values for the enrichment of the group of predicted target genes by migration processes were calculated using a hyper-geometric test; all the expressed genes in MCF10A cells were used as background. *P*-values <0.05 were defined as significant.

Cell culture

MCF10A cells were grown in Dulbecco's modified Eagle's medium-F12 (DMEM-F12) supplemented with antibiotics, 10 mg/ml insulin, 0.1 mg/ml cholera toxin, 0.5 mg/ml hydrocortisone and heat-inactivated horse serum [5% (vol/vol) and 10 ng/ml EGF].

Knockdown of selected microRNAs

MCF10A cells (1×10^5) were plated in six-wells, and allowed to adhere. The next day, cells were transfected using Oligofectamine transfection reagent (Invitrogen), with siRNA oligonucleotides directed at hsa-miR-20a, hsa-miR-671-5p and hsa-miR-212 and a non-targeting control (purchased from Dharmacon).

Transwell cell migration assay

Cells were plated in the upper compartment of a Transwell tray (Corning, Corning, NY, USA) and allowed to migrate through an intervening nitrocellulose membrane for 24 h at 37°C. The membrane was then removed and fixed in paraformaldehyde (3%), followed by cell permeabilization in Triton X-100 and staining with methyl violet. Cells growing on the upper side of the membrane were scraped using a cotton swab, and cells growing on the bottom side of the membrane were photographed, and then disintegrated in 10% acetic acid for quantification.

RNA isolation and microarray hybridization

Total RNA was extracted from biological duplicates at four time-points following EGF stimulation. RNA was isolated using Qiagen's microRNA isolation kit (Valencia, CA, USA) and was hybridized to Affymetrix GeneChip Human Gene 1.0 ST arrays. Microarray data are deposited in Gene Expression Omnibus (GSE33538).

Microarray data analysis

Affymetrix Expression Console was used, followed by normalization of all arrays together using a Lowess multi-array algorithm and signal-dependent noise estimation (48). The data were thresholded at 5 (log scale) and duplicate samples were averaged.

Real-time quantitative PCR

mRNA analysis

Complementary DNA (cDNA) was generated by the use of SSII reverse transcriptase (Invitrogen, Carlsbad, CA,

USA). Real-time quantitative PCR (qPCR) analysis was performed using SYBR Green I (Finnzymes, Invitrogen) as a fluorescent dye. Primers were designed using UniversalProbeLibrary, and β2 microglobulin (B2M) served for normalization.

miR analysis

cDNA was generated by the use of the Qiagen miScript kit, according to the manufacture instructions. Real time qPCR analysis was performed using the Qiagen miScript kit, with specific primers to miR-212, miR-671-5p and miR-20a. U6 small RNA served for normalization.

Luciferase reporter assay

For 3′-UTR reporter assays of *EDN1*, cells were transfected with reporter plasmid encoding a wild-type *EDN1* 3′-UTR and a pGL3-CMV containing *Firefly* luciferase (Promega, Madison, WI, USA). Cells were co-transfected with miRNA mimic oligomucleotides (Qiagen, Valencia, CA, USA) of miR-671-5p or a control. Forty-eight hours after transfection with the reporter plasmid, cells were harvested and *Firefly* and *Renilla* luciferase activities were measured using the Promega dual-luciferase assay system.

RESULTS

The CoSMic algorithm

The main idea of CoSMic is to combine gene expression measurements of mRNAs and miRs from the same samples, with sequence-based miR target predictions, to improve the reliability of the predictions. The algorithm works on each miR separately, and searches for a group of mRNAs with correlated expression to the miR, which also have high sequence-based scores. Assuming that this correlation implies regulation of the gene by the miR, we repeat the process twice, separately for genes with positive and negative correlations to the miR. Negative correlation between a miR and its targets represents a classical repression relationship in which the miR is upregulated to inhibit its target gene, or vice versa, downregulated to allow the upregulation of its target gene. Positive correlation between a miR and its target gene can be due to transcriptional coregulation of the miR and its target, with the miR fine-tuning its target's expression (49).

Thus, CoSMic provides for each miR a group of genes that were identified as its predicted targets in a particular experiment or condition, and a corresponding *P*-value for the significance of this prediction (related to the miR and its group of predicted targets). Using the FDR procedure to correct for multiple testing, a set of miRs with significant *P*-values is declared. These miRs are predicted to be relevant in the specific model system, and to regulate the expression levels of their target genes, as identified by the CoSMic algorithm. For more details about CoSMic see the 'Materials and Methods' section.

Application of CoSMic to real data

We tested CoSMic algorithm with coupled mRNA and miR time-course expression measurements performed on mammary epithelial (MCF10A) cells after EGF stimulation (46,47). The mRNA and miR expression measurements were done in two independent experiments, but since the system is highly reproducible, it is equivalent to measuring the mRNA and miR simultaneously. Applying the algorithm to the coupled dataset, we used 677 genes and 138 miRs that were differentially expressed in a time-dependent manner (exceeding predefined thresholds, see 'Materials and Methods' section). This step ensures that all mRNAs and miRs inserted into the algorithm are expressed and modulated, and hence are potentially functional in this specific model system (the expression data used here can be found in the 'Supplementary Methods' section). Using CoSMic with PITA predictions, we found 50 miRs which were identified as significant (FDR 20%) in this experiment; the results of using CoSMic with other algorithms, such as TargetScan (4) and miRanda (7) are described later in 'Comparing the results of CoSMic when used in conjunction with other sequence-based predictors' section. The number of target genes, as identified by the algorithm for each miR, was few tens; much less than the number of target genes predicted when using the sequence-based predictions alone (see in Table 1 the target groups associated with the top 10 significant miRs obtained by CoSMic and Supplementary Table S1 for all 70 miRs that passed FDR 20%). 37 out of the 70 significant miRs had negative correlation with their target mRNAs, 29 had positive correlation with their target genes and four miRs had both positively and negatively correlated target genes (miR-320a, miR-31, miR-671-5p and miR-20a).

Selection of candidate miRs

Since the major phenotypic response of MCF10A cells to EGF stimulation is migration, we searched among the significant miRs obtained from the algorithm, those miRs whose predicted target genes were enriched by migration processes. Specifically, for each significant miR, we calculated the enrichment of its set of target genes

Table 1. Top 10 miRs identified as significant by the CoSMic algorithm (used with PITA)

miR name	Correlation	*q*-value	No. of targets by CoSMic	No. of targets by PITA
hsa-miR-212	+	0.007	12	2749
hsa-miR-623	+	0.007	24	4849
hsa-miR-769-5p	−	0.007	13	3582
hsa-miR-500	−	0.007	23	2923
hsa-miR-320a	+	0.007	35	4901
hsa-miR-424	−	0.009	17	2901
hsa-miR-98	−	0.009	34	3514
hsa-miR-601	+	0.010	40	2810
hsa-miR-31	+	0.019	18	4906
hsa-miR-22	+	0.019	28	2316

For each miR we indicate whether it was identified as significant by the positively (+) or negatively (−) correlated targets, the corresponding *q*-value (see 'Materials and Methods' section), number of predicted target genes obtained by the algorithm, and the number of target genes obtained by the sequence-based predictions alone (PITA algorithm, targets with negative scores).

(as defined by CoSMic) in each one of the 33 migration GO-terms. We counted for each significant miR, the number of migration GO-terms, that its group of target genes was significantly enriched in (P-value < 0.05, Figure 2). The target genes of 35 miRs, out of the 70 significant ones, were enriched for at least one migration process. These 35 miRs are our candidates to regulate migration in MCF10A cells in response to EGF stimulation.

Effect of manipulation of the candidate miRs on the migration phenotype of the cells

After identifying our candidate miRs to regulate migration in MCF10A cells, we examined experimentally their actual effect on the migration phenotype of the cells by silencing few of them and performing a migration assay. Three candidate miRs were selected for this purpose (Figure 2, indicated by red): (i) miR-212 which had the lowest q-value and its target genes were significantly enriched by five migration gene ontology (GO)-terms, (ii) miR-20a, whose target genes were significantly enriched by seven migration GO-terms, different from those of miR-212 and (iii) miR-671-5p which was found to be significant by the algorithm also when using TargetScan predictions instead of PITA predictions (see 'Comparing the results of CoSMic when used in conjunction with other sequence-based predictors' section).

In addition we selected a miR that was not identified as significant by CoSMic (miR-27b; FDR $> 20\%$) and two miRs (let-7e and miR-20b, indicated by green in Figure 2), that were significant (FDR $< 10\%$) according to CoSMic, but their target genes were not enriched for migration GO terms. We tested experimentally whether these 'true negative' control miRs indeed did not affect the migratory response of the cells.

We manipulated the cells by silencing the three candidate miRs and the three control miRs (each miR separately), and assessed, using a Transwell migration assay, the effect of these perturbations on the migration phenotype of MCF10A cells. We compared the migration of these cells to the migration of an empty si-oligo control (si-control). As can be seen from Figure 3A, the three candidate miRs were indeed functional and affected the migration phenotype of the cells; silencing of miR-20a reduced cell migration ($P = 0.0005$), whereas the silencing of miR-671-5p and miR-212 enhanced the migration phenotype of the cells ($P = 0.012$ for miR-671-5p). As expected, the three 'true negative' control miRs did not affect the migration phenotype (Figure 3B). It is important to mention that we performed the enrichment analysis for migration GO terms in order to narrow down the list of significant miRs obtained by CoSMic to the most promising candidates. Nevertheless, we cannot exclude the possibility that a significant miR, whose targets were not found to be enriched for migration GO terms, does regulate migration (since enrichment for migration depends also on the validity and completeness of the GO). Still, we do believe that such a miR is less likely to regulate migration relative to the significant miRs whose target genes were enriched for migration.

Estimate the validity of CoSMic

After showing that CoSMic indeed identified correctly functional miRs in the system, we need to demonstrate that these miRs mediated the effect on the cell through the targets that were defined by the algorithm. For this purpose we experimentally found the putative and context-specific target genes of miR-20a and miR-671-5p (see next section); using these as our 'ground truth' we compared the performance of CoSMic to the sequence-based predictions alone (see 'Comparison of CoSMic predictions and the sequence-based predictions' section) and to algorithms combining sequence-based predictions with expression data (see 'Comparison of CoSMic performance relative to other algorithms combining sequence-based predictions with expression data' section). Moreover we performed a Luciferase assay to prove direct interactions between miR-671-5p and one of its putative target genes (see 'Experimental validation of CoSMic results' section).

Identification of the putative and context-specific target genes of miR-20a and miR-671-5p, which mediated the effect on migration

To find experimentally the group of putative and context-specific target genes of the two functional miRs, miR-20a and miR-671-5p, we silenced each miR in MCF10A cells and measured the mRNA expression levels after EGF stimulation (using microarrays). In addition, we measured the mRNA levels of MCF10A cells transfected with si-oligo control (si-control) after EGF stimulation. We identified target genes which mediated the effect of a miR on migration, as the group of genes which fulfill the following three criteria: (i) contain a binding site for the miR; (ii) are differentially expressed due to the miR's manipulation; the logic behind this criterion is that silencing of a miR should allow the upregulation of its target genes, and hence the target gene should be expressed differently between si-control and si-miR. This statement is relevant also for the positively correlated targets since by silencing the miR we are no longer in the regime in which the miR and its target are transcriptionally coregulated; and (iii) are differentially expressed due to EGF stimulation (i.e. in the original experiment); we are looking for targets that mediate the migration process of MCF10A cells in response to EGF stimulation, and therefore these targets should be modulated due to EGF in order to mediate the migration process. Using these criteria we found 15 target genes of miR-671-5p that mediate migration, and 28 target genes of miR-20a (Figure 4 and Supplementary Table S2); we refer to these as 'confirmed targets'. We measured by qRT–PCR the expression levels of several of these confirmed target genes after EGF stimulation, in order to validate the microarray results (Figure 5). As can be seen these target genes were indeed upregulated due to the miR's silencing.

Comparison of CoSMic predictions and the sequence-based predictions

After identifying the group of confirmed target genes of miR-671-5p and miR-20a, which mediate the migration phenotype, we can evaluate whether CoSMic improves the identification of confirmed and context-specific

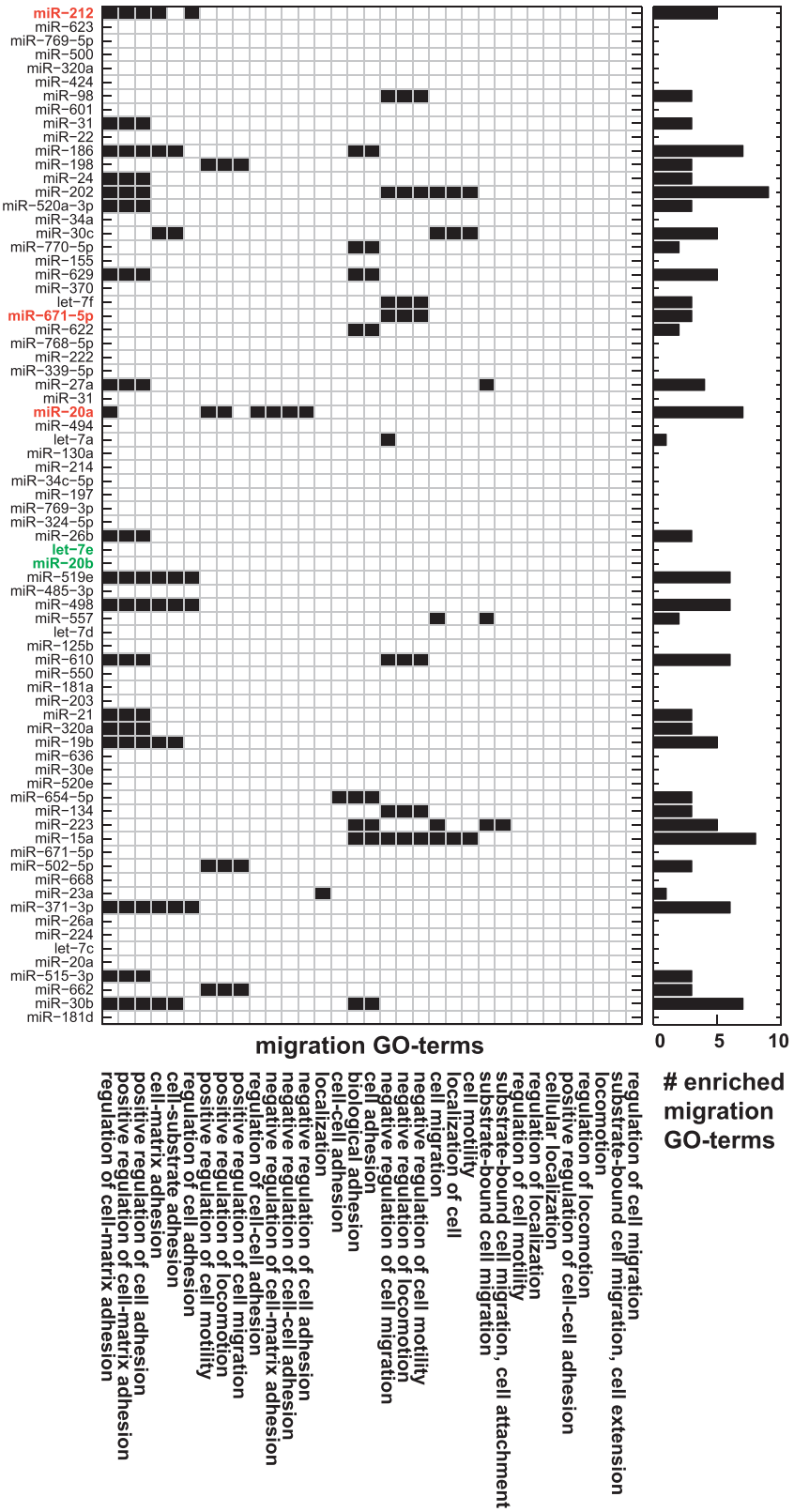

Figure 2. Enrichment of the groups of predicted target genes by migration processes. For each miR that was identified as significant by CoSMic algorithm (FDR 20%, 70 miRs), we calculated the enrichment of its predicted target genes (as identified by CoSMic) by migration processes. There are 33 GO-terms related to migration, for each one of them we calculated the enrichment of the target genes of the significant miR by a hyper-geometric test. Left panel: the *y*-axis lists the significant miRs obtained by the algorithm ordered by their *q*-values; the *x*-axis lists the 33 migration GO-terms; black squares represents significant enrichment (P-values < 0.05). Right panel: the *x*-axis presents the number of migration GO-terms that were significantly enriched by the predicted target genes of each indicated miR (*y*-axis). miRs indicated by red are our selected candidates for further investigation and the 'true negative' control miRs (see text for details) are marked in green.

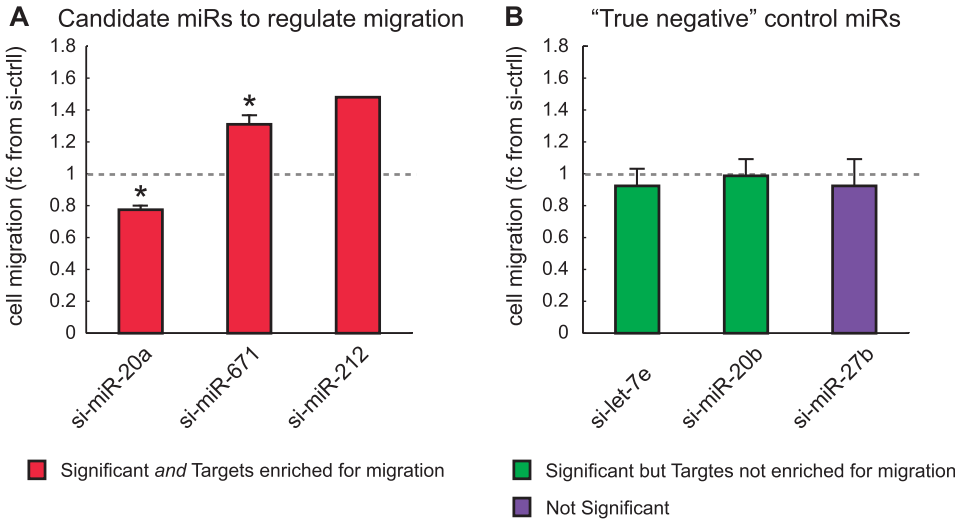

Figure 3. Quantification of Transwell migration assays of MCF10A cells treated with si-oligos for the indicated miRs, or controls. MCF10A cells were transfected with indicated si-oligos or controls, and migration was assessed using a Transwell migration assay; Quantification was made 24 h after stimulation with EGF. (**A**) The three candidate miRs that were found as significant by CoSMic *and* their target genes were significantly enriched for migration processes. (**B**) 'True negative' control miRs, 2 of which (let-7e and miR-20b) were identified as significant by CoSMic but their target genes were not enriched for migration processes, and miR-27b which was not found to be significant by the algorithm. Data is presented as mean fold change from si-control ± SD ($n = 3$ or 4 replicates for each miR); *P*-values were calculated relative to the si-control by paired *t*-test, significant results ($P < 0.05$) are indicated by asterisk. For si-miR-212, due to technical problems, only one repeat is presented.

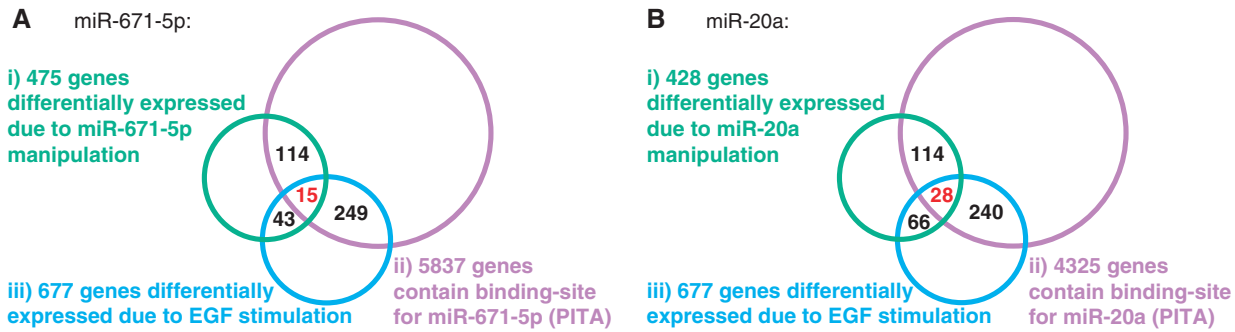

Figure 4. Identification of 'confirmed target genes' of miR-671-5p and miR-20a. (**A–B**) Venn Diagrams showing the group of confirmed context-specific target genes of miR-671-5p and miR-20a, respectively. We identified as confirmed target genes that mediate the effect of a miR on migration as those genes which contain a binding site of the miR, are differentially expressed due to the miR's manipulation and due to EGF stimulation.

targets relative to the naïve sequence-based algorithm (we describe first the results for PITA, and treat other algorithms in 'Comparing the results of CoSMic when used in conjunction with other sequence-based predictors' section). Practically, we assessed whether CoSMic performs better than simply taking, for each miR, the top x predictions of PITA, where x corresponds to the number of targets identified by CoSMic. Our measure for the performance of each algorithm is the positive predictive value (ppv), which is also referred to as purity. The ppv is the fraction of predicted target genes that were indeed confirmed; it is calculated by the number of true positives (i.e. the number of confirmed targets identified by the algorithm) divided by the total number of targets predicted by the algorithm. As can be seen from Tables 2 and 3, CoSMic has better results than using only the sequence-based predictions. CoSMic identified 69 target genes for miR-671-5p (33 were negatively correlated and 36 were positively correlated with the miR). Among these 69 targets, 5 (three negatively and two positively correlated genes) were from the group of 15 confirmed and context-specific targets (33% sensitivity, ppv = 0.0725), to be compared with 0 among the top 69 predictions of PITA. To be even more conservative, we doubled the top list from PITA (i.e. instead of taking the top 69 targets, we took the top 138 targets); the results remained the same—none of the confirmed targets were found by PITA (0%). Regarding miR-20a, CoSMic found 13 out of its 28 confirmed targets (46% sensitivity, ppv = 0.113) among its 115 identified target genes (32 negatively correlated and 83 positively correlated), whereas PITA's top 115 predicted targets contained only 2 of the 28 confirmed targets (7% sensitivity, ppv = 0.0174). Doubling the list to 230 top PITA targets did not add any new confirmed target gene. Thus, CoSMic algorithm indeed filters out many false-positive results,

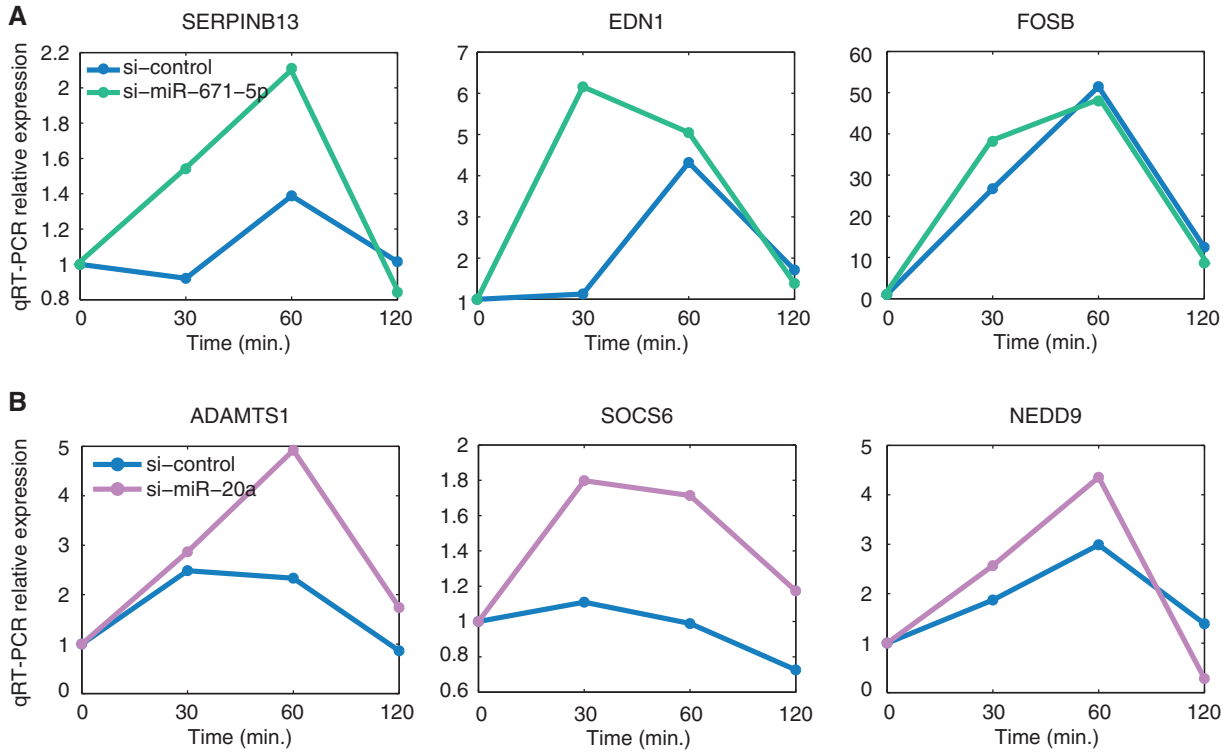

Figure 5. qRT–PCR expression levels of several putative target genes of miR-671-5p (**A**) and miR-20a (**B**). (**A**) *SERPINB13*, *EDN1* and *FOSB* expression levels by qRT–PCR in si-control (blue) and si-671-5p (green). (**B**) *ADAMTS1*, *SOCS6* and *NEDD9* expression levels by qRT–PCR in si-control (blue) and si-20a (magenta).

and allows discovery of 'real' and context-specific target genes.

Comparison of CoSMic performance relative to other algorithms combining sequence-based predictions with expression data

First, we compared CoSMic results with other algorithms that integrate both mRNA and miR expression data with sequence-based predictions (as is done by CoSMic). We selected those algorithms that were implemented as a tool and are available to the biologist to analyze his own experimental data (see Supplementary Data for more details). Thus, we compared CoSMic results with GenMiR++ and MAGIA algorithms. As can be seen from Tables 2 and 3, CoSMic has better ppv's than any of the other tested algorithms, both for miR-671-5p and miR-20a. Using MAGIA, with either Spearman or Pearson correlation, none of the confirmed target genes of miR-671-5p and miR-20a were identified. Using GenMiR++, nine confirmed target genes of miR-671-5p were identified among its 198 predicted targets, but the ppv (0.045) was lower than that of CoSMic (0.0725). We compared CoSMic's results also to the top 69 predictions of GenMiR++ (the same number of target genes as identified by CoSMic for this miR). Among the top 69 predictions of GenMiR++ there were only three confirmed target genes of miR-671-5p, to be compared with five putative target genes identified by CoSMic. Regarding miR-20a, GenMiR++ identified 89 genes as its targets; among them 6 were confirmed target genes, representing a ppv of 0.0674, relative to 0.113 of CoSMic. When comparing CoSMic results with the top 115 prediction of GenMiR++, GeneMiR++ identified six confirmed targets, whereas CoSMic identified 13 confirmed target genes among its 115 predicted targets. Thus, predictions obtained by CoSMic for miR20a have higher ppv's than both MAGIA and GenMiR++ algorithms. Furthermore, it should be noted that GenMiR++ was not designed to identify the functional miRs that play an active role in the specific model system of interest.

Next, we applied FAME, miReduce and Sylamer (three algorithms that consider mRNA expression data along with the sequence-based information, but not miR expression) on our mRNA expression data. miReduce did not find any motif of size 6–8 nt corresponding to a miR, whose presence or absence was significantly (P-value < 0.05) correlated with the fold changes of the differentially expressed genes (see results in Supplementary Tables S3 and S4 and 'Supplementary Results' section). Sylamer found three depleted and one enriched motif (with Bonferroni corrected P-value < 0.01), corresponding to miR-151-3p (6-mer), miR-218 (7-mer), miR-643 (8-mer) and miR-328 (8-mer), respectively (Supplementary Figure S2). miR-218 and miR-643 are not expressed in MCF10A cells according to the miR array data, and hence are not functional; Regarding the other two miRs, miR-151-3p obtained q-values of 0.34 (0.7) for the negatively (positively) correlated targets by

Table 2. Comparison of CoSMic results for miR-671-5p (using PITA) with other algorithms

Algorithm	No. of predicted targets	No. of overlap with 15 confirmed targets	Positive predictive value
CoSMic (FDR = 20%)	69	5	5/69 (0.0725)
PITA (top 69)	69	0	0/69 (0)
PITA (top 138)	138	0	0/138 (0)
MAGIA-spearman (FDR = 30%)	2	0	0/2 (0)
MAGIA-pearson (FDR = 30%))	0	0	0/0 (0)
GenMiR++ (75^{th} percentile)	198	9	9/198 (0.0455)
GenMiR++ (top 69)	69	3	3/69 (0.0435)

Table 3. Comparison of CoSMic results for miR-20a (using PITA) with other algorithms

Algorithm	No. of predicted targets	No. of overlap with 28 confirmed targets	Positive predictive value
CoSMic (FDR = 20%)	115	13	13/115 (0.113)
PITA (top 115)	115	2	2/115 (0.0174)
PITA (top 230)	230	2	2/230 (0.0087)
MAGIA-spearman (FDR = 30%)	1	0	0/1 (0)
MAGIA-pearson (FDR = 30%)	0	0	0/0 (0)
GenMiR++ (75^{th} percentile)	89	6	6/89 (0.0674)
GenMiR++ (top 115)	115	6	6/115 (0.0522)

CoSMic, and miR-328 obtained q-values of 0.2 (0.46); the overlap between the targets predicted by the two algorithms for these miRs was not significant (Supplementary Table S5). Thus, Sylamer did not identify miR-671-5p or miR-20a as potential regulators of the differentially expressed genes in MCF10A cells in response to EGF stimulation. Regarding the miRs that were predicted as regulators, we cannot assess the performance of this algorithm since we do not know the confirmed targets of these miRs. Moreover, two of these predicted miRs were not expressed in our dataset, and the number of target genes predicted by Sylamer for each miR was > 100 (see 'Supplementary Results' section). FAME predicted miR-26ab/1297 to regulate eight of the differentially expressed genes (with corrected P-value of 0.001). miR-26a was also found as significant by CoSMic algorithm (q-value = 0.05), with a group of 19 target genes. Five targets were shared between FAME and CoSMic predictions (P-value = 3.6×10^{-9}), moreover, four out of the five shared targets were the top four targets predicted by CoSMic (see Supplementary Tables S6–S8 and 'Supplementary Results' section). Thus, there is high agreement between FAME and CoSMic predictions regarding this miR, which give validation to the predictions. Nevertheless, CoSMic identified also other miRs (some of which were experimentally validated), which FAME missed (see the full results in the 'Supplementary Results' section).

Experimental validation of CoSMic results

EDN1 was identified by CoSMic as a target gene of miR-671-5p. Using a Luciferase assay we proved the direct interaction between miR-671-5p and *EDN1* (Figure 6). In this experiment the 3′-UTR of *EDN1* is placed downstream to the Luciferase gene, and the effect of over-expressing miR-671-5p on Luciferase expression was measured directly. Interestingly, *EDN1* was found to be commonly overexpressed in a broad range of human tumors. In esophageal squamous cell carcinoma its upregulation is associated with cellular migration, tumor cell metastasis and invasion (50); in breast cancer increased expression of *EDN1* enhanced tumor cell invasion (51). These findings are compatible with our results that upon silencing of miR-671-5p, *EDN1* was upregulated, which, in turn, enhanced the migration phenotype of MCF10A cells.

Comparing the results of CoSMic when used in conjunction with other sequence-based predictors

Our algorithm can integrate the mRNA and miR expression data with sequence-based predictions from any one of the available prediction algorithms (PITA, TargetScan, MiRanda, etc.). Until now we explored CoSMic's results when using the sequence-based predictions of PITA. Next, we examined the effect of implementing other sequence-based prediction algorithms, e.g. TargetScan and miRanda, on the results of CoSMic. We evaluated the differences between the results obtained using PITA, TargetScan and MiRanda at two levels: first, the significant miRs obtained by the algorithm with each sequence-based prediction algorithm; second, the target genes predicted by our algorithm for each miR, using the different sequence-based predictions. The predictions of CoSMic used with TargetScan or miRanda can be found in Supplementary Table S9. Briefly stated, using TargetScan the algorithm identified 48 miRs as significant (at FDR of 20%), 18 of which (~40%) were shared with the miRs identified as significant by CoSMic with PITA's predictions. Using miRanda's predictions, 10 miRs were identified as significant (at FDR of 20%), among them 2 were shared with PITA's significant miRs and 2 others with TargetScan's significant miRs. Thus, there is higher agreement between the significant miRs obtained by CoSMic when used with PITA and TargetScan predictions, relative to using miRanda predictions.

Next, we evaluated whether CoSMic algorithm predicts the same target genes for these shared significant miRs. We compared the number of shared targets obtained when CoSMic was used with PITA and TargetScan, to the number of shared targets between the top predictions of the same two algorithms. As before, the number of top sequence-based predicted targets used was the same as the number of targets identified by CoSMic for each miR; to be even more conservative, we repeated the same comparison using twice the number of targets identified by CoSMic (i.e. doubled the list of prediction from PITA and from TargetScan). As can be seen from Figure 7, the number of shared targets derived by using CoSMic with PITA and with TargetScan is much larger than

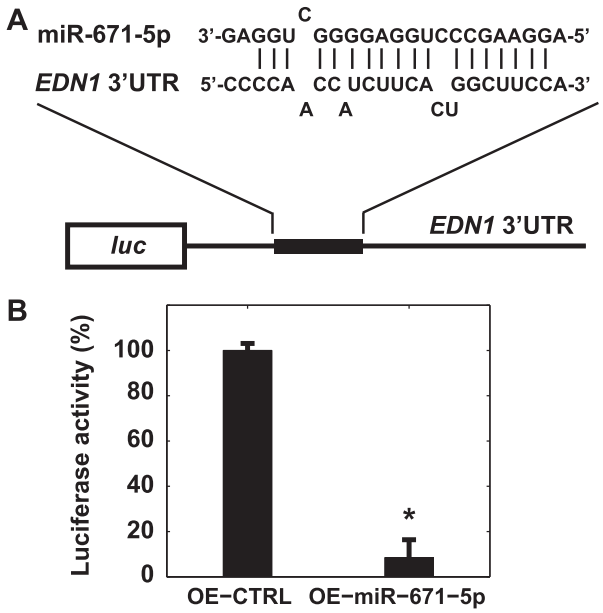

Figure 6. Evidence that the 3′-UTR of *EDN1* is targeted by miR-671-5p. (**A**) Schematic diagram of the Luciferase assay–reporter construct with wild-type 3′-UTR of *EDN1*, containing a binding-site for miR-671-5p. (**B**) HeLa cells were cotransfected with the luciferase reporter construct and either mimic-miR-671-5p (OE-miR-671-5p) or mimic-control (OE-CTRL). Luciferase activities were measured and normalized to the level of control Renilla luciferase. Mean and SD values of duplicates of six repeats are presented; *P*-value = 7.83×10^{-13}, indicated by asterisk as significant.

what one gets by intersecting top PITA and TargetScan predictions alone (or by intersecting top PITA and TargetScan predictions of twice the number of CoSMic targets tested). Moreover, the shared targets by top PITA and TargetScan alone are not expressed or changed in the original experiment, and hence, these targets have low potential to be relevant in this model system (for the results of all 22 shared miRs see Supplementary Table S10). Thus, CoSMic indeed filters out correctly false-positive predictions; thereby increasing the agreement between targets lists obtained using different sequence-based prediction algorithms.

One of the significant shared miRs obtained by using both PITA and TargetScan was miR-671-5p, which was shown above (see 'Identification of the putative and context-specific target genes of miR-20a and miR-671-5p, which mediated the effect on migration' section) to affect migration in MCF10 cells through regulating 15 target genes. Hence we could use the previously described experimental results to compare the performance of CoSMic with TargetScan to the list produced by TargetScan alone, as was done above for PITA (see 'Comparison of CoSMic predictions and the sequence-based predictions' section and Table 2). As can be seen from Table 4, again the results obtained using CoSMic were better than using only TargetScan sequence-based predictions. Our algorithm identified 17 target genes negatively correlated with miR-671-5p and 20 targets positively correlated with the miR expression. Among these 37

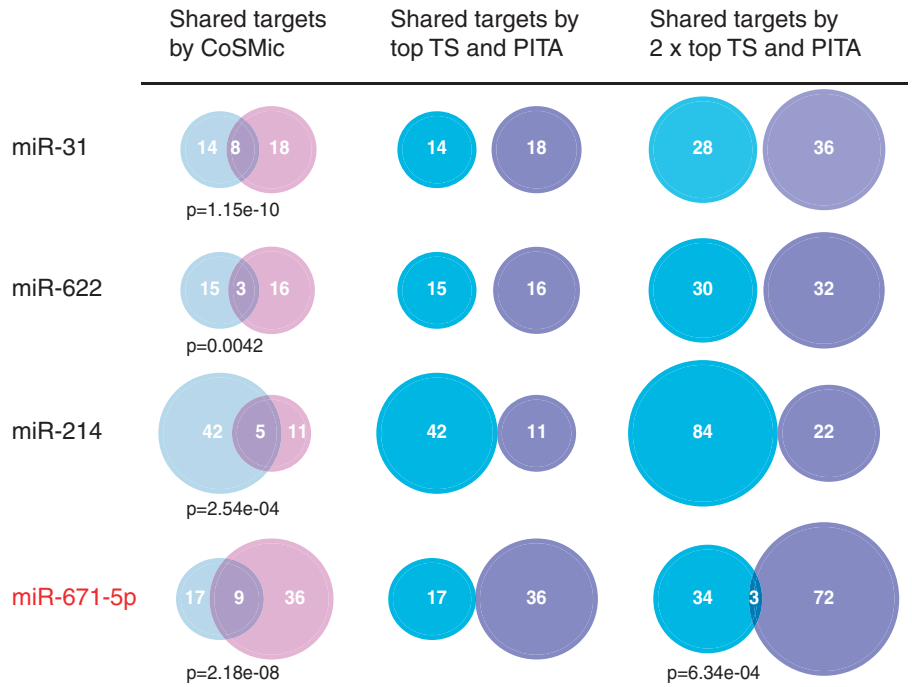

Figure 7. Number of shared targets obtained by CoSMic and by intersecting top PITA and TargetScan (TS) predictions alone. We presented 4 out of the 18 shared significant miRs obtained when CoSMic is used with PITA and with TS predictions. For each miR the numbers of targets identified as significant by CoSMic, and the number of shared targets, are indicated in column 1 (light blue circles for using CoSMic with TS predictions and pink circles for using CoSMic with PITA predictions). The intersection of top × and top 2× predictions of TS and PITA, obtained using sequence information alone, are presented in columns 2 and 3, respectively (blue circles for TS and purple circles for PITA). For both PITA and TS we used × that corresponds to the number of significant targets predicted by CoSMic, when used in conjunction with the corresponding sequence-based method. The *P*-values for the number of shared targets were calculated using the hyper-geometric test.

Table 4. Comparison of CoSMic results for miR-671-5p (using TargetScan) with TargetScan predictions alone

Algorithm	No. of predicted targets	No. of overlap with 15 confirmed targets	True positive rate
CoSMic	37	4	4/37 (0.108)
TargetScan (top 37)	37	0	0/37 (0)
TargetScan (top 74)	74	0	0/74 (0)

targets, 4 were from the group of 15 putative and context-specific target genes of miR-671-5p (26%). However, none of the putative and context-specific target genes were either among TargetScan's top 37 predictions, or TargetScan's top 74 predictions. This serves as additional evidence demonstrating that CoSMic filters out correctly false-positive results, and significantly improves identification of 'real' and context-specific target genes.

DISCUSSION

The challenge of predicting the function and target genes of a miR of interest is not yet solved. Currently there is no simple and widely used high-throughput experimental method for it, and the available computational sequence-based prediction algorithms suffer from high rate of false-positive results. We introduced the CoSMic algorithm which deals with this problem by combining mRNA and miR expression data with the sequence-based predictions. The rationale behind our algorithm is that functional targeting of a miR should be reflected in its target mRNA expression level, and therefore we integrated to the predictions the correlation between the miR and its target gene. Thus, CoSMic searches for a group of genes that are correlated with the miR expression and are enriched in its sequence-based predictions. By integrating the expression measurements to the predictions, CoSMic assures that the identified predicted target genes are expressed in the specific model system, and hence are context specific to this system of interest; an aspect which is missing from the sequence-based predictions alone. Moreover, CoSMic provides information about the statistical significance of the predictions, which enables the researcher to focus on the most relevant candidate miRs in the specific model system. Furthermore, the number of target genes predicted by CoSMic algorithm for each miR is only few tens, much less than by the sequence-based predictions alone, which allows the researcher to focus on a few context-specific target genes of the miR of interest for further investigation.

We tested the CoSMic algorithm on experimental data of coupled mRNA and miR expression measurements performed on mammary epithelial cells (MCF10A) after EGF stimulation. We showed experimentally that CoSMic indeed identified correctly functional miRs in this system using a migration assay, since the major phenotypic response of MCF10A cells to EGF stimulation is migration. Moreover, we demonstrated that CoSMic filters out many false-positive targets, relative to the sequence-based predictions alone, and has higher ppvs than the other tested algorithms that combine expression data with sequence-based predictions. Last we showed that using CoSMic algorithm we increase the overlap between the predictions of different sequence-based algorithms, relative to their agreement alone; additional evidence that CoSMic filters out correctly false-positive predictions.

We propose the CoSMic algorithm and experimental design as a global strategy for unveiling the functional significance of miRs in biological systems. The algorithm allows the possibility to start with a biological system with no prior knowledge about the functional miRs in this system, and through several steps, identifying functional miRs in the system of interest, and their context-specific target genes.

Since CoSMic algorithm searches target genes that are both positively and negatively correlated with the miR expression, it's opens additional question for further investigation: whether the direction of the correlation between the miR and its target gene imply about the mechanism of regulation by the miR; does negative correlation implies about classical repression or degradation of the mRNA target, and positive correlation about fine-tuning or inhibition of the target at the level of translation.

We believe that improvement of miR target prediction will greatly facilitate miR research in general and, specifically, the understanding of individual miRs' function, pathways they are regulate and mechanism of regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. Supplementary Introduction, Supplementary Methods, Supplementary Results, Supplementary Figures 1 and 2, Supplementary Tables 1–10 and Supplementary References [52,53].

ACKNOWLEDGEMENTS

The authors thank Prof. Gideon Rechavi and Jasmine Jacob-Hirsch for performing the microarray hybridization of the si-experiment. E.D. is the incumbent of the Henry J. Leir Professorial Chair and Y.Y. is the incumbent of the Harold and Zelda Goldenberg Professorial Chair.

FUNDING

Funding for open access charge: The Leir Charitable Foundation; the Israel Science Foundation (ISF); the MD Moross Institute for Cancer Research; the Dr Miriam and Sheldon Adelson Medical Research Foundation; the German Research Foundation (DIP) and the National Cancer Institute [CA72981].

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
4. 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.
5. Hwang,H.W. and Mendell,J.T. (2006) MicroRNAs in cell proliferation, cell death, and tumorigenesis. *Br. J. Cancer*, **94**, 776–780.
6. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
7. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.
8. Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38(Suppl.)**, S8–S13.
9. Farazi,T.A., Spitzer,J.I., Morozov,P. and Tuschl,T. (2011) miRNAs in human cancer. *J. Pathol.*, **223**, 102–115.
10. Sethupathy,P., Megraw,M. and Hatzigeorgiou,A.G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods*, **3**, 881–886.
11. Iorio,M.V., Ferracin,M., Liu,C.G., Veronese,A., Spizzo,R., Sabbioni,S., Magri,E., Pedriali,M., Fabbri,M., Campiglio,M. *et al.* (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.*, **65**, 7065–7070.
12. Sturm,M., Hackenberg,M., Langenberger,D. and Frishman,D. (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, **11**, 292.
13. Chandra,V., Girijadevi,R., Nair,A.S., Pillai,S.S. and Pillai,R.M. (2010) MTar: a computational microRNA target prediction architecture for human transcriptome. *BMC Bioinformatics*, **11(Suppl. 1)**, S2.
14. Liu,H., Yue,D., Chen,Y., Gao,S.J. and Huang,Y. (2010) Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, **11**, 476.
15. Marin,R.M. and Vanicek,J. (2011) Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res.*, **39**, 19–29.
16. Lekprasert,P., Mayhew,M. and Ohler,U. (2011) Assessing the utility of thermodynamic features for microRNA target prediction under relaxed seed and no conservation requirements. *PLoS One*, **6**, e20622.
17. Ogul,H., Umu,S.U., Tuncel,Y.Y. and Akkaya,M.S. (2011) A probabilistic approach to microRNA-target binding. *Biochem. Biophys. Res. Commun.*, **413**, 111–115.
18. Ragan,C., Zuker,M. and Ragan,M.A. (2011) Quantitative prediction of miRNA-mRNA interaction based on equilibrium concentrations. *PLoS Comput. Biol.*, **7**, e1001090.
19. Corrada,D., Viti,F., Merelli,I., Battaglia,C. and Milanesi,L. (2011) myMIR: a genome-wide microRNA targets identification and annotation tool. *Brief. Bioinform.*, **12**, 588–600.
20. Cho,S., Jun,Y., Lee,S., Choi,H.S., Jung,S., Jang,Y., Park,C., Kim,S. and Kim,W. (2011) miRGator v2.0: an integrated system for functional investigation of microRNAs. *Nucleic Acids Res.*, **39**, D158–D162.
21. Naeem,H., Kuffner,R., Csaba,G. and Zimmer,R. (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, **11**, 135.
22. Gamazon,E.R., Im,H.K., Duan,S., Lussier,Y.A., Cox,N.J., Dolan,M.E. and Zhang,W. (2010) Exprtarget: an integrative approach to predicting human microRNA targets. *PLoS One*, **5**, e13534.
23. Kowarsch,A., Preusse,M., Marr,C. and Theis,F.J. (2011) miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA*, **17**, 809–819.
24. Dweep,H., Sticht,C., Pandey,P. and Gretz,N. (2011) miRWalk—database: prediction of possible miRNA binding sites by 'walking' the genes of three genomes. *J. Biomed. Inform.*, **44**, 839–847.
25. Mestdagh,P., Lefever,S., Pattyn,F., Ridzon,D., Fredlund,E., Fieuw,A., Ongenaert,M., Vermeulen,J., De Paepe,A., Wong,L. *et al.* (2011) The microRNA body map: dissecting microRNA function through integrative genomics. *Nucleic Acids Res.*, **39**, e136.
26. Hsu,J.B., Chiu,C.M., Hsu,S.D., Huang,W.Y., Chien,C.H., Lee,T.Y. and Huang,H.D. (2011) miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics*, **12**, 300.
27. Huang,J., Townsend,C., Dou,D., Liu,H. and Tan,M. (2011) OMIT: a domain-specific knowledge base for microRNA target prediction. *Pharm Res.*, **28**, 3101–3104.
28. Ritchie,W., Flamant,S. and Rasko,J.E. (2010) mimiRNA: a microRNA expression profiler and classification resource designed to identify functional correlations between microRNAs and their targets. *Bioinformatics*, **26**, 223–227.
29. Shirdel,E.A., Xie,W., Mak,T.W. and Jurisica,I. (2011) NAViGaTing the micronome–using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs. *PLoS One*, **6**, e17429.
30. Satoh,J. and Tabunoki,H. (2011) Comprehensive analysis of human microRNA target networks. *BioData Min.*, **4**, 17.
31. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.
32. Israel,A., Sharan,R., Ruppin,E. and Galun,E. (2009) Increased microRNA activity in human cancers. *PLoS One*, **4**, e6045.
33. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
34. 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. USA*, **103**, 2746–2751.
35. 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.
36. Creighton,C.J., Nagaraja,A.K., Hanash,S.M., Matzuk,M.M. and Gunaratne,P.H. (2008) A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA*, **14**, 2290–2296.
37. Wu,X. and Watson,M. (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*, **25**, 832–833.
38. Ulitsky,I., Laurent,L.C. and Shamir,R. (2010) Towards computational prediction of microRNA function and activity. *Nucleic Acids Res.*, **38**, e160.
39. Huang,J.C., Morris,Q.D. and Frey,B.J. (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J. Comput. Biol.*, **14**, 550–563.
40. Nam,S., Li,M., Choi,K., Balch,C., Kim,S. and Nephew,K.P. (2009) MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.*, **37**, W356–W362.
41. Sales,G., Coppe,A., Bisognin,A., Biasiolo,M., Bortoluzzi,S. and Romualdi,C. (2010) MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Res.*, **38**, W352–W359.
42. Jayaswal,V., Lutherborrow,M., Ma,D.D. and Yang,Y.H. (2011) Identification of microRNA-mRNA modules using microarray data. *BMC Genomics*, **12**, 138.
43. Li,X., Gill,R., Cooper,N.G., Yoo,J.K. and Datta,S. (2011) Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. *BMC Med. Genomics*, **4**, 44.
44. 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.
45. Bang-Berthelsen,C.H., Pedersen,L., Floyel,T., Hagedorn,P.H., Gylvin,T. and Pociot,F. (2011) Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics*, **12**, 97.
46. Zeisel,A., Kostler,W.J., Molotski,N., Tsai,J.M., Krauthgamer,R., Jacob-Hirsch,J., Rechavi,G., Soen,Y., Jung,S., Yarden,Y. *et al.* (2011) Coupled pre-mRNA and mRNA dynamics unveil

operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.*, **7**, 529.
47. Avraham,R., Sas-Chen,A., Manor,O., Steinfeld,I., Shalgi,R., Tarcic,G., Bossel,N., Zeisel,A., Amit,I., Zwang,Y. *et al.* (2010) EGF decreases the abundance of microRNAs that restrain oncogenic transcription factors. *Sci. Signal*, **3**, ra43.
48. Zeisel,A., Amir,A., Kostler,W.J. and Domany,E. (2010) Intensity dependent estimation of noise in microarrays improves detection of differentially expressed genes. *BMC Bioinformatics*, **11**, 400.
49. Li,Z., Huang,H., Chen,P., He,M., Li,Y., Arnovitz,S., Jiang,X., He,C., Hyjek,E., Zhang,J. *et al.* (2012) miR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia. *Nat. Commun.*, **3**, 688.
50. Chattopadhyay,I., Singh,A., Phukan,R., Purkayastha,J., Kataki,A., Mahanta,J., Saxena,S. and Kapur,S. (2010) Genome-wide analysis of chromosomal alterations in patients with esophageal squamous cell carcinoma exposed to tobacco and betel quid from high-risk area in India. *Mutat. Res.*, **696**, 130–138.
51. Wiesmann,F., Veeck,J., Galm,O., Hartmann,A., Esteller,M., Knuchel,R. and Dahl,E. (2009) Frequent loss of endothelin-3 (EDN3) expression due to epigenetic inactivation in human breast cancer. *Breast Cancer Res.*, **11**, R34.
52. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.Roy. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.
53. Huang,J.C., Babak,T., Corson,T.W., Chua,G., Khan,S., Gallie,B.L., Hughes,T.R., Blencowe,B.J., Frey,B.J. and Morris,Q.D. (2007) Using expression profiling data to identify human microRNA targets. *Nat. Methods*, **4**, 1045–1049.