1380–1391 *Nucleic Acids Research, 2015, Vol. 43, No. 3* *Published online 27 January 2015*
doi: 10.1093/nar/gkv050

Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G

Rafal Gumienny and Mihaela Zavolan*

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50–70, 4056 Basel, Switzerland

Received September 22, 2014; Revised January 08, 2015; Accepted January 13, 2015

ABSTRACT

Small interfering RNA (siRNA)-mediated knock-down is a widely used experimental approach to characterizing gene function. Although siRNAs are designed to guide the cleavage of perfectly complementary mRNA targets, acting similarly to microRNAs (miRNAs), siRNAs down-regulate the expression of hundreds of genes to which they have only partial complementarity. Prediction of these siRNA 'off-targets' remains difficult, due to the incomplete understanding of siRNA/miRNA–target interactions. Combining a biophysical model of miRNA–target interaction with structure and sequence features of putative target sites we developed a suite of algorithms, MIRZA-G, for the prediction of miRNA targets and siRNA off-targets on a genome-wide scale. The MIRZA-G variant that uses evolutionary conservation performs better than currently available methods in predicting canonical miRNA target sites and in addition, it predicts non-canonical miRNA target sites with similarly high accuracy. Furthermore, MIRZA-G variants predict siRNA off-target sites with an accuracy unmatched by currently available programs. Thus, MIRZA-G may prove instrumental in the analysis of data resulting from large-scale siRNA screens.

INTRODUCTION

MicroRNAs (miRNAs) are ∼22 nucleotides long non-coding RNAs that guide Argonaute proteins to RNA targets. By silencing target expression (1), miRNAs take part in the regulation of many processes including cell differentiation and development (2). Aberrant miRNA expression has been implicated in many diseases, notably in carcinogenesis (3). The miRNA's 5′ end, particularly nucleotides 2–7 which are known as the 'seed' region (4,5), is thought to nucleate the miRNA–target interaction. Much experimental and computational work has established that perfect complementarity between the miRNA seed and the target site is important for the interaction (see (6) for a recent review). Target sites that satisfy this constraint are known as 'canonical' while those that do not as 'non-canonical'. High-throughput experimental studies point to a relatively high preponderance of non-canonical sites (7–10).

Exploiting the miRNA-dependent gene silencing pathway, exogenous small interfering RNAs (siRNAs) have been used as a tool to rapidly silence gene expression (11). Although an siRNA is designed to be perfectly complementary to its mRNA target, it rapidly became apparent that the transfection of the siRNA affects the expression of many other RNAs that are complementary to the siRNA seed region (12,13). These siRNA seed-dependent, 'off-target' interactions are frequently responsible for the observed phenotypes, and hamper the use of siRNAs for gene targeting. Nonetheless, large siRNA screens continue to be used to elucidate gene function, and therefore accurate prediction of siRNA off-targets has great practical importance.

One step in this direction has been made by approaches that uncover siRNA 'off-target' signatures from mRNA expression data (14,15). Prediction of siRNA off-targets has also been attempted (16) although stand-alone programs are not generally available. However, because siRNA off-target effects occur through the miRNA pathway, tools for miRNA target site prediction (17,18) can also be used to predict siRNA off-targets. An important limitation for this approach is that the strongest indicator of functionality of a putative miRNA target site, namely its evolutionary conservation (5), is unlikely to be relevant for the off-target sites of exogenous siRNAs. Yet it is precisely this feature that is exploited by the most accurate miRNA target prediction methods (19–21). Thus, the accuracy of siRNA off-target prediction is probably lower than the accuracy of miRNA target prediction, although such comparisons have not been carried out systematically. Interestingly, a tendency of active siRNA off-target sites to reside in transcript regions that are evolutionarily conserved has been noted (22).

*To whom correspondence should be addressed. Tel: +41 61 267 1577; Fax: +41 61 267 1584; Email: mihaela.zavolan@unibas.ch

The goal of our work was to develop a method that can predict canonical and non-canonical miRNA targets and siRNA off-targets with comparable accuracy. An important ingredient of our model is the miRNA–target interaction energy predicted by the MIRZA biophysical model that we previously inferred from Argonaute 2 crosslinking and immunoprecipitation (Ago2-CLIP) data (8). In addition to the MIRZA-predicted energy of interaction, the model includes features that we and others have shown to be predictive for functional miRNA target interactions, such as the nucleotide (nt) composition around putative target sites, their structural accessibility and location within 3′ untranslated regions (3′ UTRs) (21,23–25). We called the resulting miRNA target prediction method MIRZA-G (from MIRZA-Genome-wide). We illustrate the performance of the model on several large-scale data sets and demonstrate that MIRZA-G can help in the interpretation of large-scale siRNA screens.

MATERIALS AND METHODS

miRNA and siRNA transfection data

To train the model and evaluate its performance we made use of an extensive set of 26 experiments, carried out by seven different groups, in which the gene expression changes that were induced by the transfection of individual miRNAs were measured (26–32). A summary of the experimental data sets is given in Table 1. Data were processed as described previously (8) to obtain the log2 fold changes in gene expression levels upon transfection of individual miRNAs. The log2 fold changes for all used experiments can be found in Supplementary Table S1.

The gene expression changes induced by 12 different siRNA transfected individually were measured by Birmingham *et al.* (12) and processed by van Dongen *et al.* (14) to infer siRNA off-target signatures. We obtained the processed data from the supplementary material of this latter study.

Microarray-based measurements of gene expression changes that were induced by the transfection of individual siRNA were also carried out in the study of Jackson *et al.* (13). From the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/), we obtained the gene expression data as SOFT-formatted files (accession GSE5814). The data correspond to transfections of 10 distinct siRNAs (PIK3CB-6338, PIK3CB-6340, MAPK14–193, MAPK14-pos2-mismatch, MAPK14-pos3-mismatch, MAPK14-pos4-mismatch, MAPK14-pos5-mismatch, MAPK14-pos6-mismatch, MAPK14-pos7-mismatch and MAPK14-pos8-mismatch), with samples prepared 24 h after transfection. From this study, we also obtained the RefSeq annotations of the probes that were present on the microarray. Each probe was mapped to a RefSeq identifier and subsequently to Entrez Gene (http://www.ncbi.nlm.nih.gov/gene) identifier. If there were multiple probes per gene, the expression was averaged. For each gene, fold-changes were averaged over replicate experiments.

A more recent siRNA screen aiming to identify regulators of the TGF-β pathway (22) used a library of ∼21000 siRNAs that were designed to target approximately 6000 human genes that have been previously connected to cancers, including all known phosphatases, kinases and more generally, components of signal transduction pathways. The sequences of these siRNAs were obtained from the supplementary material of the paper. We scanned the set of 3′ UTRs (obtained as described in the section '3′ UTR Sequences') for matches to the seed regions of all siRNAs included in this screen, obtaining ∼50 million distinct matches. For each of these putative target sites, we calculated the associated features, as described below. Finally, we determined per-gene scores for all siRNAs as described in the section 'Computing Transcript/Gene Scores'.

miRNA and siRNA sequences

miRNA sequences were downloaded from miRBase (33) version 20. The sequences of siRNAs that were used in the experiments described above were obtained directly from the supplementary material of the studies that described the data (13–14,22). Some siRNA sequences were shorter than 21 nucleotides (nts). Because the MIRZA model assumes a small RNA sequence of 21 nts, we extended the sequences of these siRNAs to 21 nts with adenines which have been shown to be favorable for the functionality of the siRNA (34). For the miRNAs whose sequence in miRBase was shorter than 21 nts (a relatively uncommon situation), we extended to 21 nts based on the genomic locus of the miRNA. The correspondence between the names of the miRNAs that were used in the transfection experiments that we analyzed and those in the current version of miRBase is provided in Supplementary Table S2, together with the miRNA sequences.

3′ UTR sequences

A common stumbling block in comparing the accuracy of miRNA target prediction methods is that standalone versions of the software are not always available. Directly comparing the sets of predictions made by different methods is problematic because the set of transcripts/3′ UTRs that served as input for target prediction differed from study to study. Because TargetScan was the baseline algorithm with which we compared our results, we used human 3′ UTR sequences downloaded from TargetScan v6.2 (5) http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61 for our predictions.

Comparisons with other miRNA target prediction methods

MiRNA target predictions were obtained from the websites corresponding to each of the tools as follows: TargetScan: http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61, DIANA-microT: http://www.microrna.gr/webServer, MiRanda mirSVR: http://www.microrna.org/microrna/getDownloads.do. Version v3.0 of DIANA-microT (http://www.microrna.gr/microT) allows prediction of targets of individual small RNAs (miRNAs and siRNAs). Therefore, we used version v3.0 of the software to predict siRNA off-targets. We downloaded the predictions generated with DIANA-microT v5.0 (CDS)

Table 1. Summary of the experimental data sets that were used to train the model and evaluate its performance

Reference	Data source (Gene Expression Omnibus (GEO) accession / URL)	miRNAs in the data set
Dahiya *et al.* (28)	GSE10150	miR-200c, miR-98
Frankel *et al.* (29)	GSE31397	miR-101
Gennarino *et al.* (30)	GSE12100	miR-26b, miR-98
Hudson *et al.* (26)	GSE34893	miR-106b
Leivonen *et al.* (31)	GSE14847	miR-206, miR-18a, mir-193b, miR-302c
Linsley *et al.* (32)	GSE683	miR-103, miR-215, miR-17, miR-192, let-7c, miR-106b, miR-16, miR-20, miR-15a, miR-141, miR-200a
Selbach *et al.* (27)	http://psilac.mdc-berlin.de/download/	miR-155, let-7b, miR-30a, miR-1, miR-16

for the comparative analysis of mRNA and protein-level prediction of miRNA targets. To obtain a gene-level target score for methods that only score individual target sites (TargetScan and mirSVR), we summed up the scores of the target sites predicted in each individual gene.

Prediction of siRNA off-targets with DIANA-microT and TargetScan Context+

Of the miRNA target prediction tools that have been reported to have high accuracy, DIANA-microT and TargetScan Context+ are accessible and allow prediction of targets not only for miRNAs but also for siRNAs. Therefore, for TargetScan Context+ we downloaded scripts provided on the website (http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_61) and predicted target sites for all siRNAs from the Birmingham *et al.* (12) and Jackson *et al.* (13) studies. As for miRNAs, we obtained gene-level target scores by summing up the scores of individual sites within each gene. For DIANA-microT we used the available web server (http://diana.cslab.ece.ntua.gr/microT/) to obtain directly gene-level predictions of siRNA targets. Because some siRNAs yielded no predictions with DIANA-microT (one of the siRNA from Birmingham *et al.* (12) and five siRNAs from Jackson *et al.* (13)), in our comparisons of the performance of the methods we used only siRNAs for which all methods tested yielded predictions.

Putative binding sites

We focused our analysis and prediction on the following types of binding sites. First, we considered canonical sites in the sense used by TargetScan (5). Thus, we scanned the 3′ UTRs for miRNA seed matches (defined as exact match to the nucleotides 2–8 of mature miRNA or match to nucleotides 2–7 and followed by an adenine). Second, we sought to identify non-canonical sites that would interact strongly with miRNAs. We scanned the entire 3′ UTRs with MIRZA (current version at http://www.clipz.unibas.ch/index.php?r=tools/mirza/Submission/index) using a window of 50 nts, sliding by 30 nts at a time. Validated miRNA target sites in the literature do not surpass a length of 50 nucleotides and at the same time, it is relatively unlikely that such regions contain multiple sites because sites that are too close to each other presumably ‘interfere’ with each other (35). We then identified windows with a MIRZA target quality score of at least 50, a score threshold that we chose based on the distribution of MIRZA scores among Ago2-CLIP sites (see section ‘MIRZA Target Quality Score’ below). Then, we calculated the best miRNA-mRNA hybrid structure and inferred the region in the mRNA that would hybridize with the miRNA seed. We used this anchor region in the mRNA to define the full miRNA target site, comprising the miRNA seed match and the upstream 21 nts. For each of these sites, we computed the set of features described below. We applied the same procedure to the prediction of siRNA off-target sites.

Feature definition and computation

MIRZA target quality score. Computing the MIRZA target quality score, defined as in (8), was the first step in our transcriptome-wide prediction of miRNA/siRNA target sites. Because the target quality score depends on the length of the putative target site, we used windows of fixed length, 50 nts, in 3′ UTRs. To define a minimum target quality score, we reanalyzed the 2998 sites that were previously used by Khorshid *et al.* (8) to train the MIRZA model. For each site, we identified the miRNA that had the highest target quality score and then computed the highest-scoring hybrid structure between this miRNA and the CLIPed site. After classifying the sites into canonical/non-canonical, we determined the distributions of target quality score for these two categories of sites. We found as before, that the target quality scores were, on average, higher for canonical compared to non-canonical sites (316 compared to 15). The cumulative density function of the scores for the two types of sites showed that a score of 50 allows us to retain most (92%) of the canonical sites and a substantial proportion (18%) of the non-canonical sites, and we therefore chose 50 as a minimum target site quality score (Supplementary Figure S1).

Position of the target site in 3′ UTRs. We determined the distance to the closest 3′ UTR boundary as the minimum between the distance from the beginning of the seed-complementary region to the stop codon and to the poly-A tail.

Nucleotide content. The ‘Flanks G content’ and ‘Flanks U content’ features were defined as the proportion of G and U nts, respectively, within 50 nt upstream and 50 nt downstream of the miRNA seed-matching region.

Accessibility. The structural accessibility of the target site was defined as the probability that the 21 nucleotide long region (anchored on the right-hand side by the nucleotide matching the 5′-most nucleotide of the miRNA seed) is

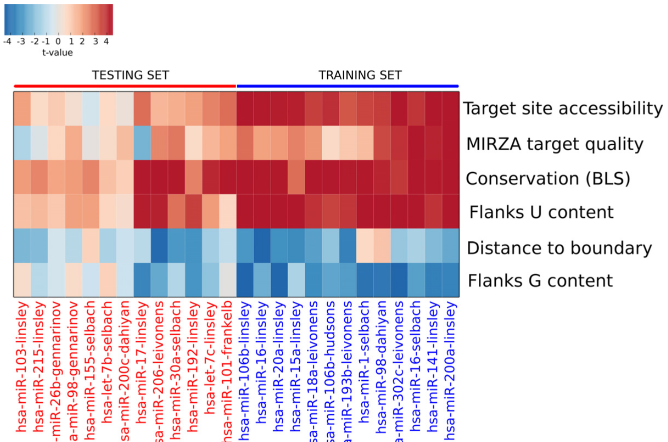

Figure 1. Value of *t*-statistic in comparing the mean values of features used in the model (rows) among functional and non-functional miRNA seed-complementary sites across 26 experiments (columns). The data from the experiments labeled in blue were used to train the model and those from experiments labeled in red were used in testing the model.

in single-stranded conformation, across all possible secondary structures. This probability was computed with CONTRAfold, a method for RNA secondary structure prediction that is based on conditional log-linear models (CLLMs) (36). CONTRAfold was applied to the region covering the miRNA seed match, and the 50 nucleotides upstream and 50 nucleotides downstream of the seed match. Computing the partition function over structures in which the target region was either constrained to be in single-stranded conformation or not (running CONTRAfold with –partition and –constraints flags, all other parameters left to default values) we could obtain the log-probability that the target site is in single-stranded conformation. We also carried out the entire model training and target prediction procedure using the energy necessary to open the secondary structure of the target region (computed with the RNAup program from the Vienna package (37), as described before in (23)) as a measure of target site accessibility. The results are comparable (Supplementary Figure S2, see section Evaluation of Model Performance for more details), although the top CONTRAfold-based predictions are slightly more down-regulated than the RNAup-based predictions. Thus, we used the CONTRAfold-based measure in the final model.

Branch length score. We quantified the selection pressure on putative target sites in terms of a 'branch length score' (38), defined as described below. The 3′ UTR sequences were aligned to the human genome (hg19) with GMAP (39). The pairwise alignments of the human genome (hg19) to the genomes of 41 other species were obtained from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html#human), and then anchored alignments (with the genomic region of the human 3′ UTRs serving as anchor) were constructed as described before (21). These alignments were used to assess the degree of evolutionary conservation of putative target sites.

The phylogenetic tree of 46 species (including *Homo sapiens*) was downloaded from the UCSC database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/) and the species for which pairwise alignments to human were not available were pruned. For each putative target site in human we carried out the following computation. Based on the alignment of the human 3′ UTRs with all the other species, we extracted the region that corresponded to the putative target site in the human 3′ UTR in all other species. Because the MIRZA target quality score depends on the length of the site, we either padded or trimmed the putative target sites in all of these species to precisely 50 nts. We then computed the target quality score of the putative target sites with the human miRNA, and we considered the target site to be conserved in a species when the target quality score was at least 50. Then, based on the evolutionary distances along the tree provided by UCSC, we computed the fraction of the total evolutionary distance in the phylogenetic tree along which the site was conserved. We called this measure branch length score. All manipulations of the phylogenetic tree were performed with DendroPy package (40). To assess the accuracy of this measure, we compared the estimates of selection pressure obtained in the manner described above with the posterior probabilities that individual putative target site are under evolutionary selective pressure, calculated with the ElMMo method (21). Because ElMMo only handles canonical sites, we did this comparison for seed-matching miRNA-complementary sites only. The two methods had comparable ability to distinguish between functional and non-functional sites (not shown).

Training of the generalized linear model

To train the model, we used only putative canonical sites of miRNAs in the test set (see below). Furthermore, to ensure that the impact of the miRNA can be attributed to specific sites, we analyzed only transcripts that contained a single putative canonical site for the transfected miRNA. For each experiment we extracted the 100 most downregulated and the 100 least-changing (whose log fold-change was closest to 0) transcripts with a single putative miRNA binding site in the 3′ UTR. These transcripts provided the 100 positive and the 100 negative target sites in the respective experiment.

For each site we then calculated the features described above: MIRZA target quality score, distance to the 3′ UTR boundary, G/U composition of flanking regions, structural accessibility, and branch length score. To assess the prediction power of these features, we carried out two-sample *t*-tests for the difference of the mean values of a given feature between the positive and negative target sites in each experiment.

Although the experiments show consistent differences between the positive and negative sites for all features that we used in our model, the significance of the difference differs to some extent between experiments (Figure 1). We used the subset of experiments in which the differences between the positive and the negative subsets of sites were most significant (labeled with blue in Figure 1) to train the model. The other subset of experiments (labeled with red) was used for testing the performance of the model.

We trained two generalized linear models (GLMs) with the logit link function (logistic regression) to classify the training data using the Statsmodels python library (41). The first model included the branch length score as a feature, the second model did not. The fitted parameters for both models can be found in Supplementary Table S3.

When predicting miRNA-binding sites transcriptome-wide we expect that the non-functional miRNA-complementary sites vastly outnumber those that are functional in gene repression, in contrast to our model training set-up, where we used an equal number of positive and negative sites. Thus, the value of the feature-dependent score at which a site has a 0.5 probability to be functional will likely be higher than the value inferred based on the training set. Formally, this would be equivalent to shifting the scores that we obtain from the linear predictor by a constant value ΔS, such that the probability of a site being bound changes from $p = \frac{e^S}{e^S+1}$ to $p\prime = \frac{e^{S+\Delta S}}{e^{S+\Delta S}+1}$. This leads to the transformation $p\prime = \frac{Kp}{Kp+1-p}$ where $K = e^{\Delta S}$. To determine an appropriate value for the constant K, we computed an overall measure of down-regulation of the predicted miRNA targets upon transfection. That is, for a given K, we computed the score T of an individual target M for a given miRNA as the expected number of bound sites in this target $T(M) = \sum_{s\in \mathrm{sites}(M)} p\prime(s)$, sorted the predicted targets of the miRNA from highest-to-lowest scoring, calculated the sum of fold-changes of top n targets for all values of n, and finally averaged these values over all miRNAs in the training set. To allow for the possibility that a minimum binding probability τ needs to be reached for a site to have a functional impact, we carried out the above calculations also allowing for different probability thresholds (Supplementary Figure S3). The optimized parameter values are $K = 0.24$ and $\tau = 0.12$. The resulting model was used to predict miRNA target sites and siRNA off-target sites across the entire set of 3′ UTRs.

Evaluation of model performance

Median fold changes. We compared the performance of various miRNA/siRNA target prediction methods as follows. For each miRNA, and for each method, we sorted all predicted target genes by their score, from highest to lowest. We determined the fold-change for each gene in each experiment and, when more than one experiment was available for a particular miRNA/siRNA, we computed the average fold-change in these experiments. Genes for which no expression estimates were available were filtered out. We then evaluated the median log fold-change of the targets predicted by a method $lm(n)$ as a function of the number n of top predicted targets. Lower median log fold-changes indicate a stronger down-regulation of the targets predicted by a given method upon miRNA/siRNA transfection. Finally, we calculated average median log fold-changes $< lm(n) >$ for all the miRNAs/siRNAs under consideration by averaging the functions $lm(n)$ over the considered miRNA/siRNA.

Estimating the number of functional targets. The number of functional targets predicted by each method for each miRNA was estimated as follows. For each miRNA transfection data set, we calculated the fraction f_{tot} of downregulated transcripts among all transcripts. This value is usually around 0.5. Then, considering the top n targets predicted by a given method for the transfected miRNA, we determined the fraction $f(n)$ of these predicted targets that are downregulated upon transfection. An $f(n)$ significantly larger than f_{tot}, indicates the presence of 'true' targets among the n predicted targets, as all of the true targets are expected to be downregulated. The total fraction $f(n)$ can be written as $f(n) = \rho(n) + f_{\mathrm{tot}}\ (1 - \rho(n))$, where $\rho(n)$ is the fraction of n predicted targets that are true targets. From this we can estimate the number of true, functional targets among to top n predicted by the method as $n_{\mathrm{func}}(n) = n \times \rho(n) = n\,(f(n) - f_{\mathrm{tot}})/(1 - f_{\mathrm{tot}})$. To summarize the data from all transfection experiments, we then determined the average number of functional targets over all considered experiments $<n_{\mathrm{func}}(\mathrm{n})>$. A similar approach was used previously in Khorshid *et al.* (8).

Analysis of the siRNA screen

siRNA-specific targeting score per gene. The score of a given siRNA for a given target gene was calculated as the sum of the scores of all unique target sites identified in the 3′ UTRs associated with the gene.

KEGG pathway analysis. For the 100 siRNAs with the strongest effect in the screen (22), we obtained seed-MIRZA-G (see Table 2) off-target predictions. Then, for each gene that was predicted to be targeted by at least one of the 100 siRNAs, we calculated the average prediction score over all of these 100 siRNAs. Additionally, we determined the number of siRNAs (from the 100 with the highest score in the screen) that were predicted to target each individual gene. We sorted genes based on the number of targeting siRNAs and extracted the top 1000 for further analysis. We performed the same analysis considering all siRNAs in the libraries, not only the 100 that were found active in the screen. KEGG pathways analysis was performed using DAVID (42,43). As background we used the human genes whose 3′ UTRs we used for target site prediction.

Estimating the impact of an siRNA on individual pathways. For a given siRNA, we averaged the seed-MIRZA-G scores over all genes in a pathway of interest. siRNAs that did not target any gene in the pathway of interest were not considered in this analysis. We then examined the relationship between the z-score of an siRNA in the screen and the average seed-MIRZA-G scores over genes in the pathway of interest.

RESULTS

Features of miRNA binding sites that are active in mRNA degradation

In line with previous studies (19,23,25), we sought to combine in our model a small number of sequence and structure features that are known to affect the efficacy of miRNA binding sites in mRNA degradation. These features were as follows

Table 2. Four alternative MIRZA-G models (see Materials and Methods for additional details)

Model name	Features	Target site type
seed-MIRZA-G	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary	canonical
seed-MIRZA-G-C	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary, evolutionary conservation	canonical
MIRZA-G	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary	canonical and non-canonical
MIRZA-G-C	MIRZA target quality score, structure accessibility, nucleotide composition of flanks, distance to boundary, evolutionary conservation	canonical and non-canonical

- MIRZA quality score of the target site—reflects the free energy of binding between a miRNA and a target site and has been shown to enable identification of non-canonical binding sites that are effective in mRNA degradation (8).
- Accessibility of the target site—defined as the probability that the target site (defined as 7 nucleotide seed match plus 14 nucleotides upstream) is in single-stranded conformation within the mRNA (23,44).
- Nucleotide composition of regions flanking the miRNA binding site—effective miRNA binding sites have been shown to reside in G-poor and U-rich sequence environments (23).
- Evolutionary conservation—this feature has been repeatedly shown to be highly informative for functional miRNA binding sites (5,21), capturing probably a variety of distinct factors that have not been characterized yet.
- Distance to the boundary—functional miRNA binding sites tend to be located at the beginning and at the end of 3′ UTRs (21,24–25) and this seems to be the case for siRNA target sites as well (data not shown).

The computation of these features is described in the Methods. To demonstrate that these features are informative for the prediction of functional miRNA target sites we used a set of 26 experimental data sets consisting of mRNA expression measurements before and after the transfection of individual miRNAs, that were obtained by seven different laboratories. From each experiment, we determined the 100 most downregulated (positive, effective sites) and the 100 least-changing (negative, ineffective sites) transcripts that had in the 3′ UTR a single canonical match to the transfected miRNA. We then computed the features of the corresponding sites as described in the Methods section, and we evaluated the significance of the difference between the means of each feature's values in the positive and negative sets with the t-test. The results, shown in Figure 1, indicate that the features that we selected indeed distinguish the positive from the negative sites consistently, across the entire set of experiments.

In particular, the feature with the most consistent predictive power is the branch length score, that reflects the evolutionary conservation of miRNA–target interaction. We used this measure of selection pressure rather than the ElMMo score that we developed previously developed (21) because although the two measures have comparable predictive power (not shown), the branch length score can be more readily be computed for non-canonical sites compared to the ElMMo score, that was designed specifically for miRNA seed matches.

Also consistent with previous results (23), the sequence composition of the flanking regions is highly predictive for their responding in miRNA transfection experiments, to an extent comparable with the branch length score. Among the features that describe structural accessibility (accessibility of the seed-complementary region, target site, extended target site), the accessibility of the target site (probability that a 21 nucleotides long target site anchored on the right-hand side by the match to the miRNA seed region is in single stranded conformation) has the most consistent performance across data sets (not shown). The accessibility of an RNA fragment for interaction with cognate factors can be defined in various ways. For example, the RNAup program from the Vienna package (45) calculates the energy that is necessary to generate a single-stranded conformation for the RNA sequence of interest, whereas the CONTRAfold program (36) computes the probability that the RNA sequence is in single-stranded conformation in the ensemble of all possible structures that it can assume. Because the CONTRAfold-based model appears to have slightly better performance than the RNAup-based model in predicting transcript down-regulation (Supplementary Figure S2), we used the CONTRAfold-based accessibility in our generalized linear model.

As shown in Figure 1, the experimental data sets appear to separate into two clusters that differ in the t-values of the differences between the feature values of positive and negative sites. To train our model we decided to use the set of experiments that gave the most significant t-values in the t-tests comparing feature values among the positive and negative sites (labeled in blue in Figure 1). The remaining set of experiments (labeled in red in Figure 1) were used for testing.

Performance of the model in predicting the response of mRNAs to miRNA transfection

We used the features defined above and the 'training set' of miRNA transfection experiments to construct a generalized linear model to predict positive sites—that confer downregulation to the host mRNA upon transfection of the cognate miRNA—and negative sites—that do not confer increased decay rate to the host mRNA—as described in the section 'Training of the Generalized Linear Model'.

We used the 'test set' of miRNA transfection experiments (Figure 1) and a procedure that we described before (8) to evaluate the performance of our model. Briefly, we sorted the putative targets of a miRNA in the order of the scores assigned to them by a given prediction method and then we traversed the list of targets from top to bottom, computing, at each target rank x, the median fold change of all top x targets in response to miRNA transfection. Although many miRNA target prediction methods have been proposed, the benchmarking studies that are available (8,46) consistently identify a few methods that yield consistently good results. We included these methods here and further refer the reader to the above-mentioned benchmarking studies for additional comparisons. One of the most widely used miRNA target prediction methods is TargetScan which consistently shows close-to-best performance (8,46). We therefore used TargetScan as the base-line for our assessment of algorithms' performance. TargetScan has two variants, one that relies on the evolutionary conservation of the putative target sites (TargetScan PCT) (20) and one that uses information about the context in which the target site resides (TargetScan Context+) (17,25). We used both of these variants in the initial testing of our model's performance. We further included DIANA-microT (18), which has also been reported to have high accuracy (46) and miRanda-mirSVR (19), which has been proposed for the prediction of both canonical and non-canonical sites.

We constructed and compared the accuracy of two types of MIRZA-based models: one that uses the branch length scores of sites in training and prediction and one that does not. Furthermore, we considered predicting only canonical targets or targets that possibly contained non-canonical sites. In the first case, we scanned the 3′ UTRs for canonical miRNA seed matches, while in the latter case we scanned the 3′ UTRs for 50 nts-long putative binding regions whose target quality score for a given miRNA was at least 50 (as described in the Materials and Methods). These models are summarized in Table 2, and the performance evaluations in Figure 2A.

We found that models that take into account evolutionary conservation perform distinctly better than those that do not (Figure 2A). When considering evolutionary conservation, the targets predicted by the model that only considers canonical sites (seed-MIRZA-G-C) undergo the strongest down-regulation in response to miRNA transfection, followed by targets predicted by DIANA-microT, TargetScan PCT, our model that also considers non-canonical sites (MIRZA-G-C) and finally those predicted by miRanda-mirSVR. Among models that do not consider evolutionary conservation, our model that only takes into account canonical sites (seed-MIRZA-G) has by far the best performance followed by our model that includes non-canonical sites (MIRZA-G), and TargetScan Context+. The top targets of MIRZA-G respond stronger to miRNA transfection compared to those of TargetScan Context+, but for targets with mid-range scores, the relative magnitude of the response is reversed. The results are comparable when we assess the performance of the models in predicting protein-level changes (measured in (27)) in response to miRNA perturbations (Supplementary Figure S4).

For each method, we also estimated the number of functional targets, comparing the proportion of predicted targets that are downregulated with the proportion of all genes that are downregulated in the transfection experiment. The relative performance of the methods, shown in Figure 2B, shows a pattern similar to that shown in Figure 2A.

Prediction of siRNA off-target effects

Small interfering RNAs (siRNAs) have become a very important tool for studying gene function. Many studies have employed siRNAs or short hairpin RNAs (shRNAs) to screen for genes that are relevant to specific phenotypes (47–49). It is not trivial to interpret the outcomes of these screens, due to a large extent to the so-called 'off-target' effects that the siRNAs have because they act through the miRNA effector pathway. SiRNAs being exogenous molecules, the feature that is most informative in the prediction of functional miRNA target sites, namely their strong evolutionary conservation, is unlikely to be informative. Thus, accurate prediction of siRNA off-target effects has remained challenging. As a the main aim of our study was to improve the prediction of siRNA off-target effects, we next tested our models on siRNA transfection data sets.

The first siRNA transfection data set that we used covered 12 distinct siRNAs (12) and previously used in the development of the Sylamer tool for the detection (though not prediction) of siRNA off-target effects (14). Figure 3A shows that our models clearly outperform TargetScan Context+ and DIANA-microT in the prediction of off-target effects of these siRNAs, whether we consider only canonical or both canonical and non-canonical sites. Interestingly, when we take into account the evolutionary conservation of the siRNA-complementary sites, we observe a somewhat stronger downregulation of the predicted mRNA targets, consistent with prior observations (22). This does not appear to be the result of siRNAs acting on the target sites of miRNAs with the same seed sequence, because we obtain similar results when we use only siRNAs that do not share six or more contiguous seed nucleotides with any of the known miRNAs (Supplementary Figure S5A and B). The results obtained for each individual siRNA in this set are given in Supplementary Figure S6A–I. In Figure 3C and D we show two examples, one corresponding to an siRNA that was inferred (14) to have strong off-target effects (siRNA-C52), and the other to an siRNA with small off-target signature (siRNA-C3). In contrast to the targets predicted by TargetScan Context+ and DIANA-microT, the top targets that are predicted by our models consistently show stronger down-regulation compared to targets with lower prediction scores.

We further analyzed the data set obtained in one of the first studies that showed that the siRNA off-target effects are mediated by the siRNA seed, similarly to miRNAs (13). This study measured the transcriptome-wide response induced by mutants of an siRNA that was designed to target the MAP kinase. As shown in Supplementary Figure S7A–G and summarized in Figure 3B, in 6 of the 7 siRNA transfections the highest-scoring predictions of our models show a stronger down-regulation compared to TargetScan Context+ or DIANA-microT-predicted targets.

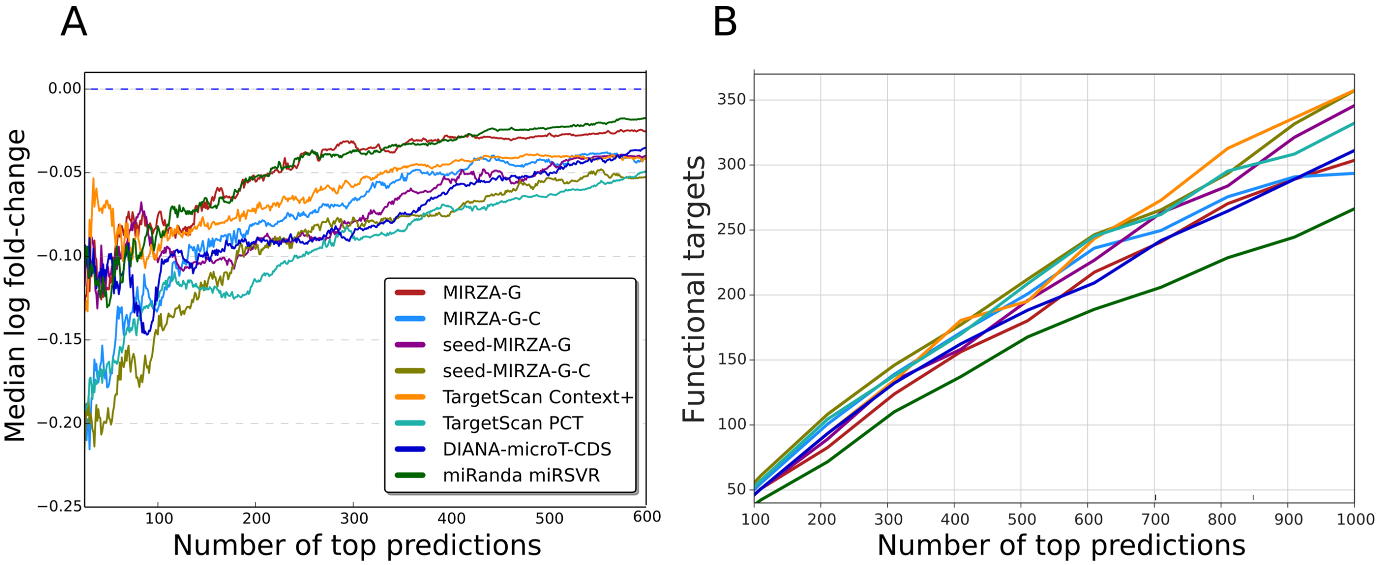

Figure 2. Comparative evaluation of various models. (**A**) Models' performance in predicting mRNA down-regulation following miRNA transfection. The expectation is that a model performs well when its top predicted targets undergo the strongest downregulation after miRNA transfection. (**B**) Estimated number of functional targets for different methods as the function of the number of top predictions. Variants of the MIRZA-G model are described in Table 2. The other tested models are TargetScan Context+, TargetScan PCT, DIANA-microT-CDS and miRanda-mirSVR (the most conservative predictions). See text for additional details on these methods.

Analysis of siRNA screening results with MIRZA-G

SiRNAs have been used in many high-throughput screens to identify key regulators or components of various biological processes. Most of these studies do not specifically investigate the off-target effects. However, a recent study found that of the ∼20 000 siRNAs that were designed, in an 'unbiased' manner, to target the coding sequence (CDS) of 6000 distinct genes (phosphatases, kinases, signal transducers and cell-surface receptors) previously implicated in cancer, a large proportion had off-target effects on the TGF-β pathway (22). We sought to determine whether the results of the screen could be interpreted in light of MIRZA-G's prediction of off-target effects.

From the screening results we identified the 100 siRNAs with the strongest phenotypic readout of TGF-β pathway inhibition, which was the translocation of a GFP-SMAD2 reporter to the nucleus. For each gene in our 3′ UTR set we calculated the average MIRZA-G targeting score over all of these siRNAs as described in Methods. We repeated this procedure using predictions from all MIRZA-G variants, as well as from TargetScan and DIANA-microT. We found that TGFBR2 is the gene with the highest seed-MIRZA-G-C and MIRZA-G-C average score for the siRNAs that were most active in the screen (Supplementary Tables S4 and S5), consistent with previous results (22). It is also a top target (3rd and 2nd, respectively) in the MIRZA-G and seed-MIRZA-G predictions. In contrast, the rank of TGFBR2 based on the TargetScan and DIANA-microT predictions is 13 and 43, respectively (Supplementary Table S5). We further used the 1000 genes with the highest average score to determine whether specific KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) (50) are targeted by the active siRNAs. In this test again, the TGF-β pathway is most enriched among the prediction of the MIRZA-G variants compared to the other methods (Supplementary Table S6). These are in fact the pathways that should be targeted through on-target effects, guided by the perfect complementarity between the siRNAs and the coding regions of the mRNAs. Interestingly, these pathways are also predicted to be targeted through off-target effects, the reason being that all of these pathways contain TGF-β. These results are consistent with the phenotypic readout of the screen as well as with our predictions (Supplementary Table S7). Figure 4A shows a sketch of the TGF-β pathway with the genes predicted to be targeted by the active siRNAs labeled with a red.

We further found a significant anti-correlation between the *z*-score, that quantifies the magnitude of the cellular response to an siRNA in the screen, and the score that our model gives to the interaction of the siRNA with TGFBR2 (Figure 4B). This anti-correlation is weaker to absent when we include more genes of the TGF-β pathway (TGFBR1, SMAD2 and SMAD4, Supplementary Figure S8) to compute an average score of interaction of the siRNA with TGF-β pathway components.

These results suggest that the gene that most responsible for the observed phenotype is TGFBR2. Although TGF-β has two main receptors, TGFBR1 and TGFBR2, it has been remarked that these two receptors do not appear to be similarly targeted (22). Indeed, we found that more of the top 100 most active siRNAs are predicted to target TGFBR2 (Figure 4C) and with higher MIRZA-G off-target scores compared to TGFBR1.

In the above analysis, we started from siRNAs that were identified in the screen to be effective in modulating the response to TGF-β. However, a question of high relevance in an experimental setting is whether relevant off-targets could be predicted *a priori*. To address this question, we computed, for each human gene, an average seed-MIRZA-

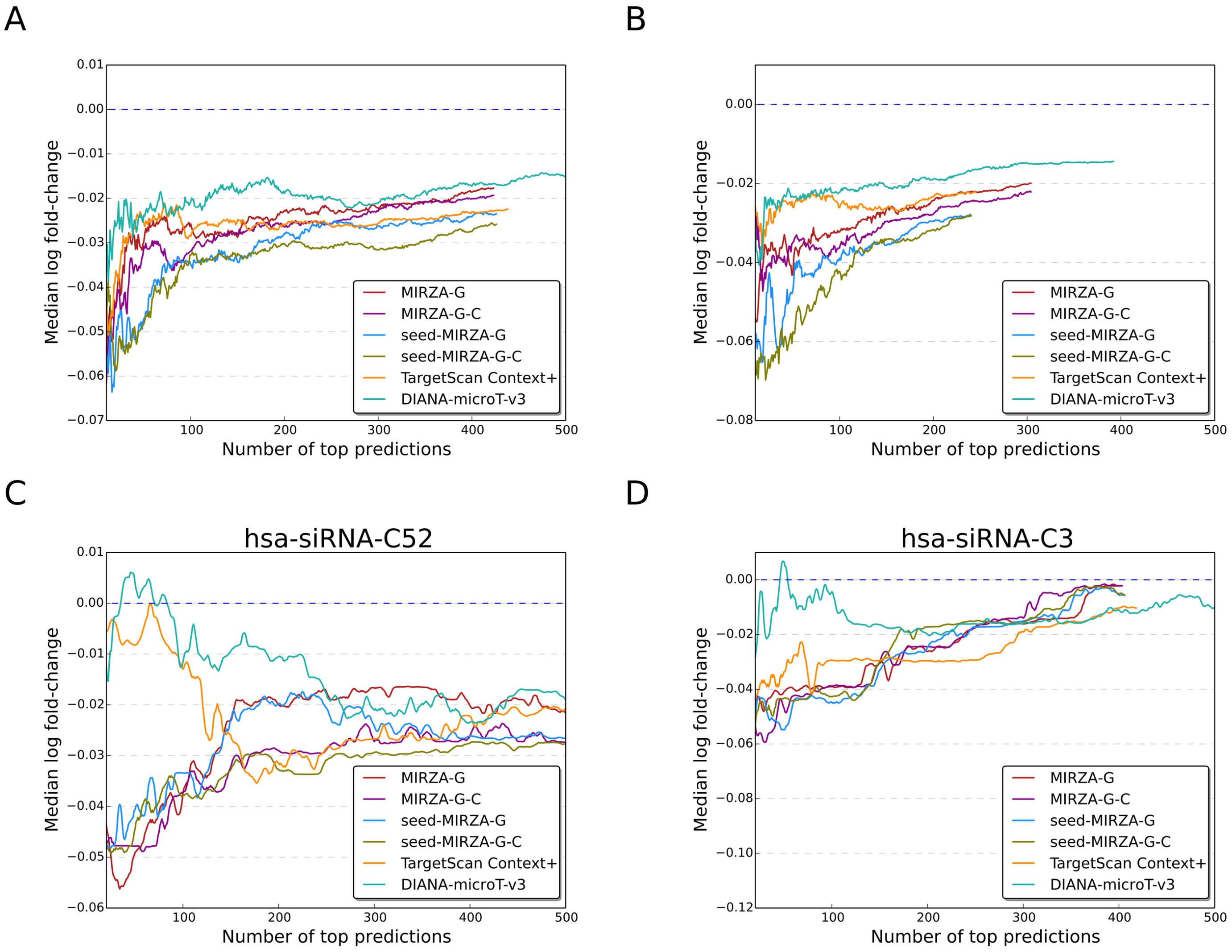

Figure 3. Relationship between the prediction scores obtained with different target prediction methods and the extent of down-regulation of target mRNAs upon siRNA transfections. (**A**) Average over the siRNAs in the data set of Birmingham *et al.* (12). (**B**) Average over the siRNAs from Jackson *et al.* (13). (**C**) Data from an individual siRNA identified by van Dongen *et al.* (14) to have prominent off-target effects. (**D**) Data from an individual siRNA identified by van Dongen *et al.* (14) to have modest off-target effects. See also Table 2 and the text for details on the methods.

G targeting score across all the siRNAs of this library (Supplementary Table S8). We then determined the enrichment of KEGG pathways among the top 1000 genes with the highest average score (Supplementary Table S9). Taking all human genes as the background set, the TGF-β pathway shows the 12th most significant enrichment. Other pathways that are even more enriched than TGF-β and would thus be expected to confound screening studies are the MAPK, neurotrophin, insulin, mTOR and ErbB pathways. Relevant for siRNA screening could be that the siRNAs in this library are also predicted to affect endocytosis.

DISCUSSION

Computational prediction of miRNA targets has progressed at a fast pace after the discovery of miRNAs, aiming to facilitate functional characterization of the thousands of miRNA genes that emerged from next-generation sequencing-based studies. Many methods are now available (46). However, a tendency to converge on a small number of determinants has been apparent, even for tools that have been in use for almost a decade. Although increasingly large numbers of non-canonical miRNA binding sites have been reported in the recent years, it is clear that many miRNA target sites are perfectly complementary to miRNA seed regions and that the degree of evolutionary conservation of the miRNA-seed complementary region is a strong predictor of target site functionality. In our study, we took advantage of a biophysical model (http://www.clipz.unibas.ch/index.php?r=tools/sub/mirza) of miRNA–target interaction that is able to identify not only canonical but also non-canonical interactions that are effective in mRNA destabilization from CLIP data (8) to predict such sites genome-wide. On its own, the biophysical model can be used to identify the miRNAs that guided the interaction of the Argonaute protein with CLIP-identified sites. However, for an accurate prediction of miRNA as

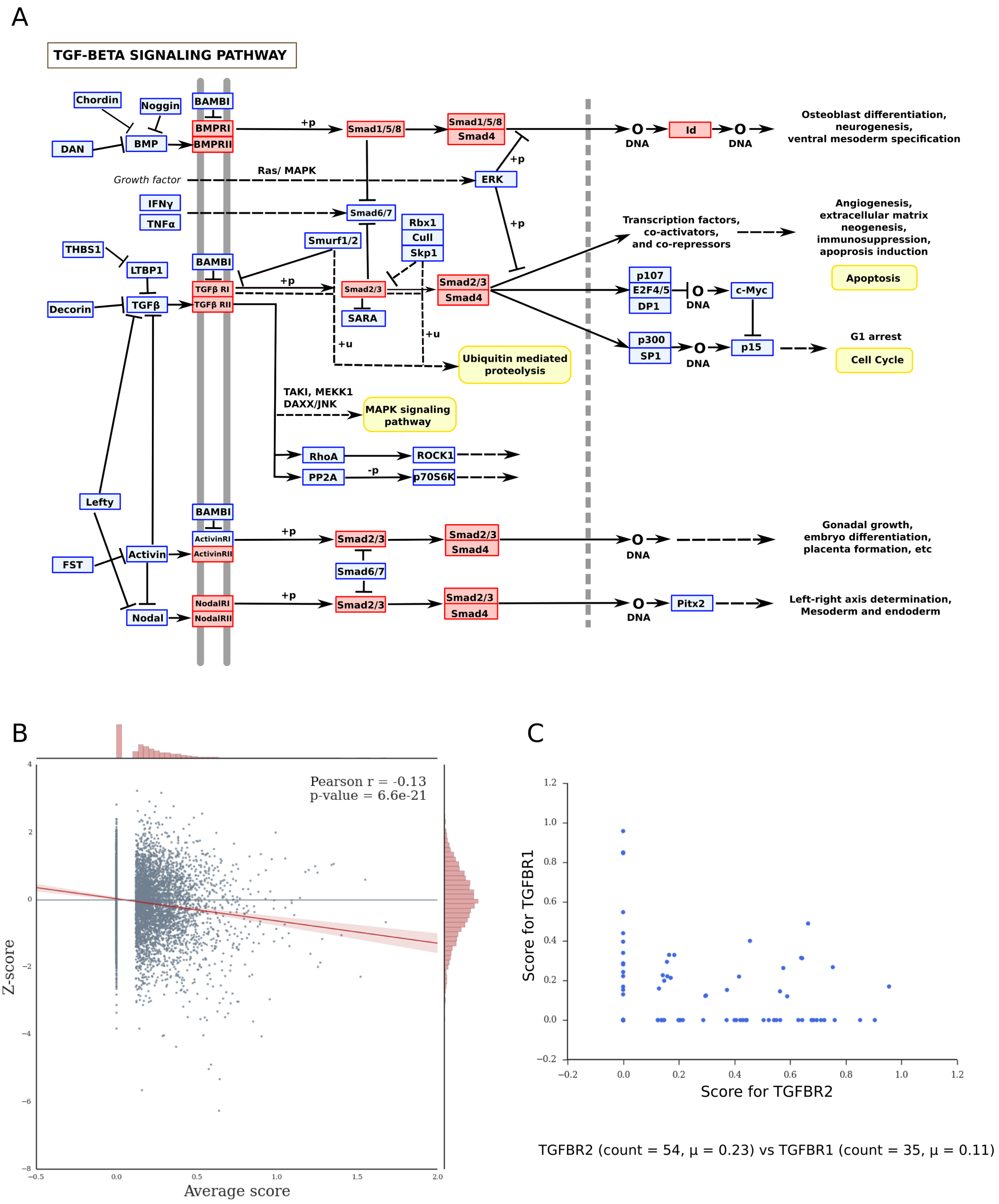

Figure 4. SiRNA off-targets in the TGF-β pathway. (**A**) Schema of the TGF-β pathway drawn based on the figure provided by the DAVID server (42,43). Genes predicted to be off-targets of the top 100 siRNAs with the strongest effect in the screen are marked with red boxes. (**B**) Correlation between the *z*-score of an siRNA in the screen (*y*-axis) and the score that our model assigns to the interaction of the siRNA with TGFBR2 (*x*-axis). (**C**) Scatter plot of the predicted activities of the top 100 most active siRNAs on TGFBR1 and TGFBR2.

well as siRNA binding sites at a genome-wide scale, features beyond the energy of the small RNA–target site interaction need to be taken into consideration. This was the motivation for developing MIRZA-G. We have shown that MIRZA-G improves to some extent the genome-wide prediction of miRNA targets and substantially the prediction of siRNA off-targets. The software is accessible at http://www.clipz.unibas.ch/index.php?r=tools/sub/mirza_g. The pipeline was implemented with the ruffus framework (51).

Our analysis indicates that the features that were previously found to characterize effective miRNA target sites, whether they are located in the 3′ UTRs or coding regions (52), are also informative for predicting siRNA off-target sites, as has been argued before (22). Overall, this is not unexpected because that siRNAs and miRNAs use the same effector pathway. What may be surprising is that taking evolutionary conservation into account improves the prediction of siRNA target sites (compare the results of seed-MIRZA-G-C with those of seed-MIRZA-G in Figure 3A and B). This is consistent with the results of a previous study which found that conserved siRNA seed matches are more likely to be effective than non-conserved seed matches (22). Although a trivial explanation could be that some siRNAs share the seed sequence with endogenous miRNAs, excluding these siRNAs from the analysis does not completely eliminate the signal (Supplementary Figure S6A and B). A possible explanation is that the conservation of a 3′ UTR region, indicative of its relevance for some biological process, is correlated with other properties, such as its structural accessibility and nucleotide composition, that support targeting by siRNAs or miRNAs. The same reasoning may explain why functional miRNA-complementary sites preferentially emerge at the beginning and end of long 3′ UTRs (21).

Although much work has been invested in computational miRNA target prediction, there remains substantial room for improvement. This may come from improved estimates of the rates of interaction between miRNAs and targets, from the inclusion of context-dependent effects such as 3′ UTR isoforms (53), modulation of miRNA–target interactions by RNA-binding proteins (54) and others. Computational modeling of the miRNA-induced effects in systems in which measurements of relevant rate constants and abundances of relevant molecular species are available, will provide further insights into this mode of regulation (55). Predictions generated by models such as MIRZA-G can provide essential entry points into such studies. Specifically in the analysis of siRNA screens, an avenue that has not been explored yet, is to use siRNA off-target predictions in conjunction with the measured phenotypic effects to infer the contribution of individual genes to the measured phenotype. This approach has been successfully used in the identification of transcription factors and miRNAs that have an important contribution to the pattern of mRNA expression in individual cell types (56). It would be interesting to apply this methodology to a large number of siRNA screens to further unravel the contributions of individual molecular pathways to phenotypes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We are grateful to Jean Hausser for sharing the pipeline for calculating miRNA target site features and to Erik van Nimwegen for discussions on the miRNA target prediction model. We also thank members of the Zavolan group for their comments on the manuscript.

FUNDING

Marie Curie Initial Training Network, RNPnet (project #289007), from the European Commission [to R.G.]. Funding for open access charge: ITN-RNPnet.

Conflict of interest statement. None declared.

REFERENCES

1. Huntzinger,E. and Izaurralde,E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev. Genet.*, **12**, 99–110.
2. Shivdasani,R.A. (2006) MicroRNAs: regulators of gene expression and cell differentiation. *Blood*, **108**, 3646–3653.
3. Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.
4. Rajewsky,N. and Socci,N.D. (2004) Computational identification of microRNA targets. *Dev. Biol.*, **267**, 529–535.
5. 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.
6. Pasquinelli,A.E. (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat. Rev. Genet.*, **13**, 271–282.
7. Chi,S.W., Hannon,G.J. and Darnell,R.B. (2012) An alternative mode of microRNA target recognition. *Nat. Struct. Mol. Biol.*, **19**, 321–327.
8. Khorshid,M., Hausser,J., Zavolan,M. and van Nimwegen,E. (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods*, **10**, 253–255.
9. Helwak,A., Kudla,G., Dudnakova,T. and Tollervey,D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.
10. Grosswendt,S., Filipchyk,A., Manzano,M., Klironomos,F., Schilling,M., Herzog,M., Gottwein,E. and Rajewsky,N. (2014) Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol. Cell*, **54**, 1042–1054.
11. Elbashir,S.M., Harborth,J., Weber,K. and Tuschl,T. (2002) Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, **26**, 199–213.
12. Birmingham,A., Anderson,E.M., Reynolds,A., Ilsley-Tyree,D., Leake,D., Fedorov,Y., Baskerville,S., Maksimova,E., Robinson,K., Karpilow,J. *et al.* (2006) 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat. Methods*, **3**, 199–204.
13. Jackson,A.L., Burchard,J., Schelter,J., Chau,B.N., Cleary,M., Lim,L. and Linsley,P.S. (2006) Widespread siRNA 'off-target' transcript silencing mediated by seed region sequence complementarity. *RNA*, **12**, 1179–1187.
14. 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.
15. Yilmazel,B., Hu,Y., Sigoillot,F., Smith,J.A., Shamu,C.E., Perrimon,N. and Mohr,S.E. (2014) Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis. *BMC Bioinformatics*, **15**, 192–197.
16. Das,S., Ghosal,S., Chakrabarti,J. and Kozak,K. (2013) SeedSeq: off-target transcriptome database. *Biomed. Res. Int.*, **2013** , 905429–905437.

17. Garcia,D.M., Baek,D., Shin,C., Bell,G.W., Grimson,A. and Bartel,D.P. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.
18. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Vlachos,I.S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.
19. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
20. Friedman,R.C., Farh,K.K.-H., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
21. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69–90.
22. Schultz,N., Marenstein,D.R., De Angelis,D.A., Wang,W.-Q., Nelander,S., Jacobsen,A., Marks,D.S., Massagué,J. and Sander,C. (2011) Off-target effects dominate a large-scale RNAi screen for modulators of the TGF-β pathway and reveal microRNA regulation of TGFBR2. *Silence*, **2**, 3.
23. Hausser,J., Landthaler,M., Jaskiewicz,L., Gaidatzis,D. and Zavolan,M. (2009) Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res.*, **19**, 2009–2020.
24. Majoros,W.H. and Ohler,U. (2007) Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics*, **8**, 152–160.
25. Grimson,A., Farh,K.K.-H., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
26. Hudson,R.S., Yi,M., Esposito,D., Glynn,S.A., Starks,A.M., Yang,Y., Schetter,A.J., Watkins,S.K., Hurwitz,A.A., Dorsey,T.H. *et al.* (2013) MicroRNA-106b-25 cluster expression is associated with early disease recurrence and targets caspase-7 and focal adhesion in human prostate cancer. *Oncogene*, **32**, 41394147.
27. Selbach,M., Schwanhäusser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.
28. Dahiya,N., Sherman-Baust,C.A., Wang,T.-L., Davidson,B., Shih,I.-M., Zhang,Y., Wood,W. 3rd, Becker,K.G. and Morin,P.J. (2008) MicroRNA expression and identification of putative miRNA targets in ovarian cancer. *PLoS One*, **3**, e2436.
29. Frankel,L.B., Wen,J., Lees,M., Høyer-Hansen,M., Farkas,T., Krogh,A., Jäättelä,M. and Lund,A.H. (2011) microRNA-101 is a potent inhibitor of autophagy. *EMBO J.*, **30**, 4628–4641.
30. Gennarino,V.A., Sardiello,M., Avellino,R., Meola,N., Maselli,V., Anand,S., Cutillo,L., Ballabio,A. and Banfi,S. (2008) MicroRNA target prediction by expression analysis of host genes. *Genome Res.*, **19**, 481–490.
31. Leivonen,S.-K., Mäkelä,R., Ostling,P., Kohonen,P., Haapa-Paananen,S., Kleivi,K., Enerly,E., Aakula,A., Hellström,K., Sahlberg,N. *et al.* (2009) Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*, **28**, 3926–3936.
32. Linsley,P.S., Schelter,J., Burchard,J., Kibukawa,M., Martin,M.M., Bartz,S.R., Johnson,J.M., Cummins,J.M., Raymond,C.K., Dai,H. *et al.* (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell. Biol.*, **27**, 2240–2252.
33. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
34. Elbashir,S.M., Martinez,J., Patkaniowska,A., Lendeckel,W. and Tuschl,T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J.*, **20**, 6877–6888.
35. Saetrom,P., Heale,B.S.E., Snøve,O. Jr, Aagaard,L., Alluin,J. and Rossi,J.J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.*, **35**, 2333–2342.
36. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
37. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
38. Kheradpour,P., Stark,A., Roy,S. and Kellis,M. (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res.*, **17**, 1919–1931.
39. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
40. Sukumaran,J. and Holder,M.T. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
41. Seabold,S. and Perktold,J. (2010) Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference* , 57–61.
42. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
43. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
44. 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.
45. Mückstein,U., Tafer,H., Hackermüller,J., Bernhart,S.H., Stadler,P.F. and Hofacker,I.L. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.
46. Alexiou,P., Maragkakis,M., Papadopoulos,G.L., Reczko,M. and Hatzigeorgiou,A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.
47. Sharma,S., Quintana,A., Findlay,G.M., Mettlen,M., Baust,B., Jain,M., Nilsson,R., Rao,A. and Hogan,P.G. (2013) An siRNA screen for NFAT activation identifies septins as coordinators of store-operated Ca2+ entry. *Nature*, **499**, 238–242.
48. Zhou,H., Xu,M., Huang,Q., Gates,A.T., Zhang,X.D., Castle,J.C., Stec,E., Ferrer,M., Strulovici,B., Hazuda,D.J. *et al.* (2008) Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, **4**, 495–504.
49. Moreau,D., Kumar,P., Wang,S.C., Chaumet,A., Chew,S.Y., Chevalley,H. and Bard,F. (2011) Genome-wide RNAi screens identify genes required for Ricin and PE intoxications. *Dev. Cell*, **21**, 231–244.
50. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
51. Goodstadt,L. (2010) Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics (Oxford, England)*, **26**, 2778–2779.
52. Hausser,J., Syed,A.P., Bilen,B. and Zavolan,M. (2013) Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.*, **23**, 604–615.
53. Nam,J.-W., Rissland,O.S., Koppstein,D., Abreu-Goodger,C., Jan,C.H., Agarwal,V., Yildirim,M.A., Rodriguez,A. and Bartel,D.P. (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell*, **53**, 1031–1043.
54. Bhattacharyya,S.N., Habermacher,R., Martine,U., Closs,E.I. and Filipowicz,W. (2006) Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, **125**, 1111–1124.
55. Hausser,J. and Zavolan,M. (2014) Identification and consequences of miRNA-target interactions - beyond repression of gene expression. *Nat. Rev. Genet.*, **15**, 599–612.
56. Balwierz,P.J., Pachkov,M., Arnold,P., Gruber,A.J., Zavolan,M. and van Nimwegen,E. (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.*, **24**, 869–884.