bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

High-Throughput Quantification of miRNA-3′-Untranslated-Region Regulatory Effects

Stephen Mastriano[1,2], Shaveta Kanoria[3], William Rennie[3], Chaochun Liu[3], Dan Li[1,2,4], Jijun Cheng[1,2], Ye Ding[3,5] and Jun Lu[1,2,5]

[1] Department of Genetics and Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA

[2] Yale Cancer Center and Center for RNA Science and Medicine, Yale University, New Haven, CT 06520, USA

[3] Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, NY 12201, USA

[4] Current address: Shanghai Eye and ENT Hospital, Fudan University, Shanghai, China

[5] Correspondence should be addressed to

J.L. (jun.lu@yale.edu), Phone: (203) 737-3426, Fax: (203) 785-4305

or Y.D. (ye.ding@health.ny.gov), Phone: (518) 486-1719

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

MicroRNAs (miRNAs) regulate gene expression post-transcriptionally, primarily through binding sites in 3′ untranslated regions (3′ UTRs). While computational and biochemical approaches have been developed to predict miRNA binding sites on target messenger RNAs, reliable and high-throughput assessment of the regulatory effects of miRNAs on full-length 3′ UTRs can still be challenging. Utilizing a miniaturized and high-throughput reporter assay, we present a 'pilot miRNA-targeting map', containing 4,994 successfully measured miRNA:3′ UTR regulatory outputs by pairwise assays between 461 miRNAs and eleven 3′ UTRs. This collection represents a large experimental miRNA:3′ UTR dataset to date on a single platform. The methodology can be generally applied to studies of miRNA-mediated regulation of critical genes. We found that seedless sites can lead to substantial downregulation. We utilized this dataset in the development of a quantitative total score for modeling the total regulatory effects by both seed and seedless sites on a full-length 3′ UTR. To assess the predictive value of the total score, we analyzed data from mRNA expression and proteomics studies. We found that the score can discriminate the potent miRNA inhibition from the weak inhibition and is thus useful for quantitative prediction of miRNA regulation. The score has been added to the STarMir program of the Sfold package now available via GitHub at https://github.com/Ding-RNA-Lab/Sfold.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

As a class of small non-coding RNAs, microRNAs (miRNAs) regulate diverse biological processes by post-transcriptionally modulating gene expression [1-3]. miRNAs bind to partially complementary sites on target messenger RNAs to downregulate target protein expression by destabilization of target mRNA and/or inhibition of protein translation. Experimental scientists working on miRNAs frequently face the question of which miRNAs regulate their genes of interest. This is often challenging because it is important to assess the level of miRNA-mediated regulation in the context of the whole gene, in addition to knowing that a miRNA can bind to a mRNA. miRNAs that can only weakly regulate targets are less interesting for further investigation.

Two approaches are often used to identify miRNAs that target genes of interest. Computational target predictions rely mainly on sequence information and known rules of miRNA:target recognition. It is widely perceived that computational algorithms have high levels of false-positive and false-negative rates [4,5] Additionally, these algorithms typically predict whether a miRNA can regulate a target gene, rather than the level of miRNA-mediated target regulation. The second approach is the biochemical purification of miRNAs and target binding sequences. For example, cross-linking and immunoprecipitation (CLIP)-based methodologies are powerful and capable of revealing miRNA-target binding events in cells [6-9]. However, it is impossible to deduce the functional output of these binding events. This is because CLIP methods do not inform the percentage of target messenger RNA occupied by a specific miRNA; it is also conceivable that binding by some miRNAs may not lead to effective functional repression of target protein production. As such, it is impossible for either approach to reliably assess the level of miRNA-mediated regulation. While one can estimate the level of miRNA-mediated regulation on predicted target genes via overexpressing or knocking out a miRNA followed by transcriptomic analyses, these transcriptomic level assays are prone to secondary and tertiary effects.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The high false rates of computational algorithms are likely due to an incomplete understanding of the miRNA-mediated target regulation. Some significant observations are summarized here. (1) Although miRNA binding sites have been experimentally identified in 3′ UTRs, coding sequences and beyond, it is believed that most functional regulatory output occurs through 3′ UTR binding sites [2]. (2) The 5′ region of the miRNA (also known as the "seed" region), particularly involving nucleotides 2-8, is critical for target recognition [2,10]. The Bartel laboratory has further classified seed-based interaction on target 3′ UTR into several categories. The presence of an "8mer" site is correlated with strong repression, "7mer" site with medium effect, whereas "6mer" site has weak to no effect [2,11]. (3) Other than seed-based interactions, seedless sites have also been observed to mediate miRNA binding on target 3′ UTRs [12-15], although the effect of such sites on target regulation is debated [8,14-16] and deemed very weak by some [17]. (4) The neighboring sequence content, the location of a miRNA binding site within 3′ UTR [6,18], and the structural accessibility of the target [19-22], can all influence miRNA binding and effectiveness.

In this study, we used a miniaturized cell-based 3′ UTR reporter assay to allow high-throughput quantification of miRNA-mediated regulation on 3′ UTR. It is important to note that our goal was to assay the effect from an abundant miRNA source in order to identify possible miRNA-mediated target regulation rather than physiological regulation. With this approach, we present an initial dataset (termed pilot miRNA-targeting map) that is comprised of 4,993 interactions for 11 3′ UTRs and 461 miRNAs, obtained on a single platform. We used this dataset to develop a quantitative score for modeling the total regulatory effects by both seed and seedless sites.

Results

Developing a high-throughput cell-based reporter assay to quantify miRNA:3′ UTR regulation

Luciferase reporter assay has long been a gold-standard to quantify miRNA-mediated regulation of 3′ UTR activity and to validate computational predicted miRNA:3′ UTR interactions. The basis of the assay

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

is that miRNA-mediated regulation directly via motifs within a 3′ UTR should be recapitulated on a non-related gene in a heterologous cellular background. It has been demonstrated that miRNA-mediated repression is not strongly influenced by cellular background [16]. Consistent with this notion, for a vast majority of reported cases, luciferase reporter assays reveal the functional consequences of direct miRNA-target interaction. Compared to measuring messenger RNA or proteins in cells which are prone to secondary and tertiary effects, the 3′ UTR reporter assay will allow lower false-positive reporting, because only a limited number of protein-coding genes are known to regulate 3′ UTRs and that miRNAs do not often regulate each other.

We have previously shown that we can use 384-well-based luciferase assay to quantify miRNA-mediated regulation of TET2 3′ UTRs [23]. However, whether this approach is reliable and expandable to many 3′ UTRs beyond TET2 has not been demonstrated, which led us to further optimize and test this approach (Fig. 1). Specifically, we used 293T cells for the assay, which is a cell line that has been widely used for assaying miRNA:3′ UTR interactions (e.g., [24,25]). In addition, 293T cells support high levels of gene expression in transient transfection assays, allowing enough signals for detecting luciferase activities. Indeed, using this 384-well format, we were able to capture the know regulation of miR-25 suppressing KLF4 levels [26] using a KLF4 3′ UTR reporter. This is a dual luciferase reporter with the KLF4 3′ UTR regulating *Renilla* luciferase activity, whereas firefly luciferase activity serves as an internal control. We co-transfected the KLF4 3′ UTR reporter with a miR-25 expression vector, into 6-well plate, 96-well plate and 384-well plate wells. In each case, we observed ~2-fold suppression of 3′ UTR activity (Fig. 2a), confirming literature findings. Importantly, the level of suppression observed in 384-well plates was similar to that observed in 6-well plates (Fig. 2a), indicating that the 384-well format can accurately capture miRNA-mediated 3′ UTR regulation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Next, we optimized two key parameters of the assay, namely the number of cells to plate in each 384-well plate well, and the time to read out luciferase activity after transfection. Varying cell densities were tested using the known targeting of miR-1 on KLF4 3′ UTR [18]. All densities generated similar levels of down regulation for the KLF4 3′ UTR reporter by miR-1 (Fig. 2b). However, plating 3,000 cells per well was deemed optimal because lower cell densities produced weaker luciferase activity readings and thus higher variation (Fig. 2b), whereas higher density leads to overcrowding of cells in wells. We then tested two time points for reading out luciferase activity at 24 or 48 hours post transfection. Again, using the known targeting of miR-1 on KLF4 or miR-125a on LIN28, we determined that the 48-hour time point produced better results (Fig. 2c). Of note, cells became over-confluent at day 3 post-transfection. These experiments established the optimal condition for our 384-well reporter assays.

We had three considerations regarding the use of 293T cells as an assaying cell type. First, can we achieve exogenous overexpression for miRNAs endogenously expressed in 293T cells? Based on miRNA expression in 293T cells, we selected miR-371, miR-125a, and miR-16, with low, medium and high endogenous expression, respectively. Quantitative RT-PCR analyses indicate that all three miRNAs can be overexpressed, with the exact fold overexpression dependent on the endogenous basal levels of the corresponding miRNAs (Fig. 2d). Of note, the levels of overexpression are underestimated, because transfection efficiency is below 100%. Second, can we observe repression by miRNAs with high endogenous expression? Our profiling data indicate that miR-25 is among the most abundant endogenous miRNAs in 293T cells. Nevertheless, we were able to observe repression by this miRNA on KLF4 3′ UTR (Fig. 2a). Third, can 3′ UTR regulatory relationships revealed in 293T cells be observed in other cell types? For this purpose, we tested the repression of miR-25 on the KLF4 3′ UTR in HeLa cells. Unlike 293T cells, 384-well format for HeLa cells yielded luciferase signals too weak to be quantifiable. We thus compared data from HeLa cells assayed in 96-well plates, which showed similar levels of repression as observed in 293T cells (Fig. 2e) and is consistent with existing data supporting that different cellular

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

backgrounds do not significantly impact miRNA-mediated gene regulation [16]. We will further examine this point in later sections.

Taken together, we established experimental conditions to reliably quantify miRNA-mediated repression of 3′ UTRs using 384-well-based high-throughput reporter assays.

Generation of the pilot miRNA-targeting map

With a suitable tool in hand, we set out to quantify a large number of miRNA:3′ UTR pairs. We cloned eleven 3′ UTR reporters, representing both human (EIF2C2, DICER1, POU5F1, SOX2, KLF4, MYC, NANOG, LIN28, TP53 and MYB) and mouse (Bmi1) genes. This list of genes includes transcription factors, RNA binding proteins, epigenetic regulators, and cancer-related genes. For miRNAs, we used an expression library that we generated [23,27], containing 461 individual miRNA overexpression constructs. Each 3′ UTR reporter was assayed pairwise with each miRNA construct in quadruplicates (Fig. 1), generating a total of 31,249 data points (including replicates and controls), after filtering out wells with poor luciferase readings. We term this dataset the 'pilot miRNA-targeting map'.

To ensure comparability of data across different 384-well plates, we built two types of in-plate controls, also in quadruplicates. These controls included control 3′ UTR vector with each of the miRNAs assayed on the plate (including miRNA vector control), and each of the 3′ UTRs assayed on the plate (including 3′ UTR vector control) with a control miRNA vector. We utilized the two types of controls when normalizing the luciferase activity, represented by the *Renilla* versus firefly ratio, so that a normalized value of 1 represents no regulation, and numbers below 1 representing downregulation (see Methods for details). We noticed that it is important to normalize with both types of controls, as the full dataset becomes much tighter after normalization, resembling a Gaussian distribution (Fig. 3a), which is expected because most miRNA:3′ UTR pairs will not have regulatory effects. This concept is further supported by

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the fact that with each additional control, there was a stepwise shrinking of overall data variation in the dataset.

To evaluate the quality of our data, we next examined the variation of each assay. Among the 4,994 successfully measured miRNA:3′ UTR pairs, 79.2% showed a standard deviation ≤ 10% of the mean, and another 15.5% with a standard deviation between 10% and 15% of the mean (Fig. 3b). Taken together, the data above showed that we successfully generated the pilot miRNA-targeting map of overall good technical quality.

Among the 4,994 miRNA-3′ UTR pairs in the pilot miRNA targeting map, we identified 181 pairs with ≥25% down regulation (Fig. 3c). We set 25% downregulation as a main cutoff for two reasons. First, those repressive relationships of less than 25% will be less likely of interest to experimental scientists. Second, small levels of downregulation are more prone to experimental noise. Among the 181 downregulation pairs, at least 20 pairs (11%) were demonstrated previously, indicating that our assay captured known miRNA-mediated target regulation.

We compared our data to a biochemical binding dataset obtained from 293T cells using the CLASH method [8]. We chose CLASH data because it utilizes ligation to create miRNA-target chimeras which avoids the ambiguity of assigning miRNAs to target regions with other CLIP datasets. Due to the limited coverage of CLASH reads and possibly low expression of some associated genes in 293T cells, only seven miRNA:3′ UTR pairs overlapped between the CLASH data and our pilot miRNA-targeting map. Other than two pairs with a mild suppression, the other five pairs did not show downregulation in our data, supporting that not all CLASH sites result in functional suppression on 3′ UTRs. A much larger overlapping dataset would be needed to accurately estimate the percentage of functional CLASH sites.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Since we used 293T cells to generate the miRNA targeting map, we asked whether our data can reflect miRNA-target regulation in specific biological contexts other than 293T cells. To address this question, we examined eleven miRNA-3′ UTR pairs, including predicted pairs and those not predicted by any of three algorithms: TargetScan [11], PITA [21] and mirSVR [28] (Fig. 3d). These eleven pairs include ten miRNA targets expressed in human embryonic carcinoma cell line NCCIT. In addition, one pair involving MYB was examined in human hematopoietic cell line K562. We transduced corresponding miRNAs into relevant cell lines and measured endogenous protein expression by western blot. Specifically, the endogenous protein levels of LIN28, POU5F1, SOX2, and NANOG were assayed in NCCIT, and MYB in K562 cells. Careful quantification of endogenous protein levels after miRNA expression revealed a significant correlation between western data and luciferase reporter assay data (Fig. 3d; R=0.845, p value=0.001). Noticeably, our data also indicate that multiple unpredicted miRNA:3′ UTR relationships can be validated on endogenous protein level (Fig. 3d). Due to the lack of seed sites in these miRNA:3′ UTR pairs and the difficulty in selection among many putative seedless sites in these 3′ UTRs for mutational analysis, we cannot exclude the possibility that such non-predicted regulation are through indirect effects. Nevertheless, our luciferase data indicate that our luciferase data can reflect *bona fide* miRNA:3′ UTR regulation, either directly or indirectly. The data above demonstrate that the miRNA-targeting map can reflect miRNA:3′ UTR regulation even though the target is expressed in a tissue specific context. Our data also suggest that the false-positive rate of the miRNA-targeting map is low, given that many of the regulatory relationships can be observed on the levels of endogenous proteins.

Among all miRNA-3′ UTR pairs for which targets were downregulated, 60.1% of the pairs only have seedless sites in the 3′ UTR; for ≥10% downregulation, 47.1% of the pairs only have seedless sites in the 3′ UTR; for ≥20% downregulation, 36.8% of the pairs only have seedless sites in the 3′ UTR; for ≥25% downregulation, 27.4% of the pairs only have seedless sites in the 3′ UTR. These suggest that seedless sites can lead to substantial target downregulation in the miRNA-targeting map.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Development of quantitative score

We reasoned that we could use these data to develop computational scores for the quantitative prediction of miRNA-mediated regulation of full length 3′ UTRs. We used a linear weighted score for which the weighting coefficients were assigned to various site types (See Methods). To estimate the site type coefficients $\alpha_{site\text{-}type}$, we first used all available seed and seedless sites and performed the calculation as described in the Methods. However, the estimates did not meet the constraint $\alpha_{8mer} > \alpha_{7mer} > \alpha_{6mer} > \alpha_{seedless}$. We reasoned that the data might contain too much noise, particularly due to the inclusion of weak miRNA binding sites. The logistic probability [22] can be used to filter out weaker sites with a low logistic probability. Specifically, we applied logistic probability thresholds to both seed sites and seedless sites, filtering the data to obtain a representative set of combinations based on the corresponding thresholds. We considered thresholds ranging from 0.0 to 0.90, with an increment of 0.05, for the logistic probability values. Among a total of 324 threshold pairs, we found two pairs that yielded efficacy parameter estimates satisfying the above constraint. The first pair is 0.75 for seed sites and 0.55 for seedless sites. The second pair is 0.75 for seed sites and 0.6 for seedless sites. Since the two sets of estimates produced highly similar results and the 0.55 threshold for the seedless sites includes more data, we preferred the estimates for the first pair of thresholds. Given those thresholds, the estimates of α_{8mer}, α_{7mer}, α_{6mer} and $\alpha_{seedless}$ are 0.0198, 0.0133, 0.00658 and 0.00428, respectively, with the intermediate statistics in the calculation presented in Table 1.

With the estimates of the site type coefficients, we can compute the quantitative score for any miRNA:target pair as proposed in the Method. To assess the predictive value of the score, we analyzed high throughput mRNA and protein expression datasets from miR-1, miR-124 and miR-181a overexpression studies [29]. For all the down-regulated (mRNA/protein fold change less than or equal to 0) targets in the overexpression cell line, we compiled 8mers, 7mers, 6mers and seedless miRNA binding sites. For each miRNA target pair, in addition to $score_{total}$, the score for seed and seedless sites combined, we also computed $score_{seed}$, the component of the total score for seed sites, and $score_{seedless}$, the component of the total score for seedless sites ($score_{total} = score_{seed} + score_{seedless}$). A component score facilitates evaluation of contribution of seed or seedless sites. For example, scores for NM_000018 are shown in Table 2.

We computed the total and the two component scores for 2,406 targets down-regulated at the mRNA levels and 1,635 targets down-regulated at the protein level. To compare the scores between the subset

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with the highest and lowest level of downregulation, we applied various thresholds for the log fold change of the mRNA or the protein expression levels. For the set of targets with highest downregulation, the log fold change thresholds were -1.0, -0.5, -0.4, -0.3 (i.e., the set is defined by values below the threshold). For the lowest downregulated set of targets, the thresholds wereand -0.05, -0.10, and -0.20 (i.e., values between the threshold and 0).

We first analyzed mRNA expression data. For threshold pairs of -1.0 and -0.05, the average $score_{seed}$ for the set of targets with mRNA fold change values < -1.0 is 0.0205, which is significantly higher than the average of $score_{seed}$ of 0.0087 for the set of targets with mRNA fold change values between -0.05 and 0 (*p*-value of 1.3×10^{-7}). For the second threshold pairs of -0.5 and -0.05, the averages of $score_{seed}$ are 0.0131 and 0.0087, with the difference being significant (*p*-value of 2.1×10^{-5}). For the third threshold pairs of -0.4 and -0.10, the averages of $score_{seed}$ are 0.0117 and 0.008, with a p-value of 4.9×10^{-6} for the difference. For the last threshold pairs of -0.3 and -0.20, the averages of $score_{seed}$ of 0.0107, is significantly higher than 0.0082 (p-value of 9.9×10^{-6}). These data are summarized in Fig. 4a.

Similarly, for each of the threshold pairs above, the average $score_{seedless}$ for the set of highest downregulation is significantly higher than that for the corresponding set of targets with lowest downregulation (Fig. 4b). E.g., for the threshold pair of -0.5 and -0.05, the difference between 0.2147 and 0.1497 has a *p*-value of 7.32×10^{-9}. The same statistical trends were also observed for $score_{total}$ (Fig. 4c). E.g., for threshold pair of -0.5 and -0.05, the difference between 0.2278 and 0.1584 has a *p*-value of 3.4×10^{-9}. For the protein expression data, we also found that the average for each of the three scores is significantly higher for the set of targets with the highest levels of downregulation (*p*-value < 1.0×10^{-15} for all three scores; Fig. 5).

The scores have been implemented in the STarMir program [30] of the Sfold package. In addition to web service (http://sfold.wadsworth.org/cgi-bin/starmirtest2.pl), STarMir is now freely available through Github at https://github.com/Ding-RNA-Lab/Sfold.

Discussion

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In this study, we present a pilot miRNA-targeting map based on a 384-well-plate high-throughput 3′ UTR reporter assay to quantify miRNA mediated regulation. This method can be particularly useful for experimental scientists to quickly identify miRNAs that regulate genes of interest, and to focus on those with stronger regulatory relationships. As an example, we have used this method to identify previously unappreciated oncogenic miRNAs by focusing on miRNAs that strongly regulate TET2, an epigenetic enzyme important in hematopoiesis [23]. It is important to note that, according to our data, miRNA regulation of a 3′ UTR does not necessarily imply physiological regulation which is dependent on the endogenous miRNA expression levels in specific cell types.

The pilot miRNA-targeting map contains quantitative regulatory outcomes from 461 miRNAs interacting with 11 3′ UTRs. This dataset differs from measuring genome-wide RNA or protein perturbations when a single miRNA is overexpressed or knocked out. It is conceivable that significant levels of secondary and tertiary effects on gene expression, even at early time points after miRNA perturbation, will increase the noise in the effort to decipher miRNA-targeting relationships. In contrast, by directly measuring 3′ UTR activity, we can avoid much of the secondary and tertiary effects, because there are relatively limited mechanisms that can influence 3′ UTR activity *in vitro*. Furthermore, when miRNAs are overexpressed, we rarely observe significant changes in endogenous miRNA expression. Our data is also different from measuring miRNA-target binding using biochemical approaches such as CLIP-based methods [6-9]. Such biochemical methods, although highly useful to inform in vivo miRNA binding on targets, do not necessarily indicate functional miRNA-based regulation. We observed that not all CLASH-identified miRNA:3′ UTR pairs have functional suppressive effect on 3′ UTR reporters, despite that the small overlap between CLASH data and the pilot miRNA targeting map prevents us from deriving statistical estimate of the percentage of functional CLASH sites. Consistent with this notion, even though miRNA binding sites were frequently observed in coding sequences [6,7], it is believed that the majority of these interactions do not lead to strong target regulation [2,11]. This notion gained additional support from a

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

recent study using PAR-CLIP on human embryonic stem cells, where only ~21% of detected target gene RNA expression could be altered when the corresponding miRNA is perturbed [31] , although some level of ambiguity of computational assignment of miRNA:target relationship with the PAR-CLIP method complicates the interpretation of these data.

Several lines of evidence support the quality of the data in our pilot miRNA-targeting map. First, the variation for the majority of the reporter assays is small. Second, known repressive miRNA:3′ UTR relationships have been captured. Third, both reporter assays in HeLa cells and western validation experiments in relevant cell types indicate that we can capture effects of miRNAs on genes not expressed in 293T cells. Fourth, known rules for miRNA target recognition were supported by our data. The Western blotting data also support that the false-positive rate of our data is low. We noticed that a reported suppressive effect of miR-145 on POU5F1 and SOX2 [32] cannot be observed in our pilot miRNA-targeting map. Neither were we able to observe the miR-145-based POU5F1 and SOX2 protein repression in NCCIT cells despite a strong overexpression of miR-145. Note that we did not use human embryonic stem cells as used in the original report [32]. This is because miR-145 can cause embryonic stem cell differentiation [32], and it will be difficult to distinguish a direct effect of miR-145 on POU5F1 and SOX2 versus an indirect effect via miR-145-induced embryonic stem cell differentiation. Although we do not fully understand why our data are different, we did notice that a CLIP study in human embryonic stem cells did not identify miR-145 binding on POU5F1 or SOX2 mRNA [31].

In the computational work, we utilized the high-throughput reporter data in the development of a new quantitative total score for modeling the total regulatory effects by both seed and seedless sites, as well as component scores for assessing individual contributions by seed sites and seedless sites. We adopted two assumptions in our modeling efforts: the linear additivity of contributions from multiple sites, and the hierarchical order in efficacy by different site types. During the model development process, we derived efficacy parameter estimates for modeling the variation in the effectiveness of different site types. To

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

assess the predictive value of the total score, we analyzed high-throughput functional data from mRNA expression and proteomics studies. We found that, each of the scores can discriminate the potent miRNA inhibition from the weak inhibition and is thus useful for quantitative prediction of the levels of downregulation. In particular, the inclusion of the contribution of seedless sites is important for modeling and prediction.

Methods

Cell culture

293T cells and HeLa were cultured in DMEM (Invitrogen) with 10% FBS (Invitrogen) and 1% Penicillin-Streptomycin-Glutamine (Invitrogen). NCCIT and K562 cells were cultured in RPMI (Invitrogen) with 10% FBS and 1% Penicillin-Streptomycin-Glutamine. All cells were cultured at 37°C with 5% CO_2.

MiRNA overexpression retrovirus was produced as previously described (Lu et al. 2008) by co-transfecting 293T cells with miRNA constructs and retroviral gag, pol, and VSVG protein expression constructs. Viral infection was carried out using spin-infection by centrifuging cells, retrovirus, and 8 µg/ml polybrene (Millipore) in a retronectin (Clontech)-coated well of a 6-well plate. Cells were selected with 5 µg/ml puromycin.

Constructs

The miRNA expression library used in this study comprises 461 single miRNAs or miRNA clusters, each derived from human genomic DNA through cloning. MiRNA overexpression constructs and their cloning were previously described [23,27]. Briefly, genomic fragments containing miRNA hairpin and flanking sequences were cloned into a retroviral expression construct pMIRWAY-puro [25,27]. The control construct for miRNA is the vector backbone without any miRNA insert. These constructs were also used to produce virus to examine their effects on endogenous proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3′ UTR dual-luciferase reporter constructs were cloned from human genomic DNA (except for Bmi1, which was cloned from mouse genomic DNA) into Promega's psiCHECK2 vector through PCR. The specific enzyme sites, along with the start and end sequences of each 3′ UTR are listed below. For the 3′ UTR control, the unaltered Promega psiCHECK2 was utilized.

POU5F1 (XhoI & NotI) ggtgcctgcccttctaggaa…acagtagatagacacactta; SOX2 (XhoI & NotI) gggccggacagcgaactgga…ttgaaatatggacactgaaa; NANOG (XhoI & NotI) agatgagtgaaactgatatt…gtaaatatacagcttaaaca; KLF4 (XhoI & NotI) atcccagacagtggatatga…aacctataatattttatctg; LIN28 (XhoI & NotI) gccacaatgggtgggggcta…tattggtacgcaaactgtca; MYC (XhoI & NotI) ggaaaagtaaggaaaacgat…aaatatatcattgagccaaa; TP53 (PmeI & NotI) acattctccacttcttgttc…acctgtgtgtctgaggggtg; EIF2C2 (XhoI & NotI) catgttttagtgtttagcga…tatatatctggtttgcagta; DICER1 (XhoI & NotI) aaccgctttttaaaattcaa…ataaagttatcgtctgttca; MYB (XhoI & NotI) acatttccagaaaagcatta…gatgtgcccttattttacct; Bmi1 (PmeI & NotI) actgttaaggaaaagatttt…taataaaacaacagaaagat

High-throughput luciferase reporter assay

293T cells were trypsinized and resuspended at a concentration of 150,000 cells/ml in culture media. We used robotics in cell seeding and transfection, as detailed below, but such steps can also be carried out manually. The CyBi-Well automatic liquid handler (CyBio) was used to seed 20 µl of the cell dilution (3,000 cells) into each well of white, clear-bottom, 384-well plates (Corning 3707). The plates were then spun at 453 *g* for 30 seconds and cultured overnight. For experiments evaluating the density of cells, we varied the concentration of the cell suspension before seeding.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The following day, transfection mixtures were prepared in 96-well PCR plates. We included two types of controls on each plate, including control miRNA assayed with each of the 3′ UTR reporters (including a control reporter), and a control reporter assayed with each of the miRNAs on the plate (including a control miRNA construct). DNA mixture for each pair of 3′ UTR reporter and miRNA construct was prepared in each well to contain 27 ng of 3′ UTR reporter and 243 ng of miRNA construct in 5.4 µl total volume). For each well, 0.81 µl FuGENE 6 (Promega E2692) and 11.79 µl serum-free DMEM were mixed as per the manufacturer's instructions. We prepare the DMEM/FuGENE6 master mixture in a microcentrifuge tube for all the wells on a 96-well DNA plate, and then use multichannel pipet to add 12.6 µl of the mixture into each DNA mixture well. The plate was spun at 453 *g* for 30 seconds and incubated at room temperature for 15 minutes. Using the automatic liquid handler, each transfection mixture was then added to the 293T cells seeded in the 384-well plates at 4 µl per well in quadruplicate, with each well receiving a transfection of 6 ng of 3′ UTR reporter, 54 ng of a miRNA construct, and 0.18 µl FuGENE 6 in 2.61 µl serum-free DMEM. Plates were then cultured for 48 hours or a specified amount of time.

Dual luciferase assays were performed using the Dual-Glo Luciferase Assay System kit (Promega E2940) as per the manufacturer's instructions. To avoid signal variation between plates, we planned the experiments so that each plate experienced the same amount of time between adding luciferase reagent and plate reading. Luminescence levels were read out on a Wallac Victor^2 plate reader (Perkin Elmer).

Data preprocessing and normalization

Raw luciferase reporter data were preprocessed and normalized using custom Matlab codes. First, we calculated the ratio of *Renilla* luciferase activity (reflect 3′ UTR activity) versus firefly luciferase activity (control luciferase in the dual luciferase reporter), which is referred to as RvF ratios. Wells with weak readings were removed before the next steps of processing. Weak wells were defined as those with weak

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

control firefly luciferase activity with a raw reading below 10,000. RvF ratios were then normalized based on the following formula. Normalized Luciferase Activity = (RvF_{Well} / mean($RvF_{UTR\text{-}CtrlMiR}$))/(mean($RvF_{CtrlUTR\text{-}miR}$)/mean($RvF_{CtrlUTR\text{-}CtrlMiR}$)). RvF_{Well} represents the RvF ratio of an assay, with mean (RvF) referring to the mean readings from all replicates of the corresponding control wells on the same plate as the RvF_{Well} assay. UTR-CtrlMiR stands for test 3′ UTR reporter assayed with a control miRNA construct. CtrlUTR-miR stands for a control 3′ UTR reporter assayed with a test miRNA. CtrlUTR-CtrlMiR stands for a control 3′ UTR reporter assayed with a control miRNA construct. After normalization, the mean of control assays becomes 1.

To identify miRNA-3′ UTR interactions that resulted in ≥25% downregulation, we examined the mean luciferase activity after normalization. Similarly, mean luciferase activities were applied to other cutoffs used in this study.

For each miRNA, we assigned major and minor strands based on miRbase data [33] and in-house next-generation sequencing reads. We analyzed predictions with both major and minor strands, as well as only with major strands.

To compare to CLASH dataset, CLASH data were downloaded from the supplement of Helwak *et al.* [8]. Only 3′ UTR binding sites were used to compare to miRNA-targeting map data.

Western blot analysis

Western blot analyses were performed as described previously [27]. For NCCIT cells, collections were conducted on ice by scrapping in cold PBS. Cells were lysed using RIPA buffer supplemented with protease inhibitors (Roche 11873580001) per manufacturer's instructions on ice for 20 minutes. Samples were then sonicated and spun at 16,100 *g* for 30 minutes. The supernatant was transferred to a new

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

microcentrifuge tube, and protein was quantified by Bradford Assay (BioRad). Samples were boiled at 95°C in SDS loading buffer for five minutes, and were analyzed using 10% polyacrylamide gels. For K562 cells, cells counts were determined, and cells were then directly boiled in SDS loading buffer, and analyzed using 10% polyacrylamide gels. Samples were loaded based on equal cell counts. We obtained more quantitative data with this method in K562 cells. Antibodies used were Oct-4A (C30A3C1), Sox2 (D6D9), Nanog (D73G4), and Lin28A (A117) from Cell Signaling Technology, and Myb (clone 1-1) from Millipore.

To quantify the western data, we exclusively considered non-saturated exposures. In addition, quantification was performed on several non-saturated exposures with different exposure times and band intensities. To ensure data accuracy, each exposure was quantified using ImageJ software, and the relative protein amounts were calculated relative to loading control. Similar results were obtained with different exposures.

RNA expression analysis

The miRNA expression profiles of 293T cells were determined using a plate-capture-based labeling and a bead-based detection system, as described before [34,35]. Data normalization was performed as described [35].

The expression levels of mature miRNAs in other experiments were determined by quantitative RT-PCR. Total RNA was collected using TRIzol Reagent (Invitrogen) as per manufacturer's instructions. cDNA was prepared using the miScript II RT Kit (Qiagen), and qPCR was conducted with the miScript SYBR Green PCR Kit (Qiagen) as per manufacturer's instructions. The Qiagen miRNA specific primers used during qPCR reaction were MS00003528 (miR-145), MS00004060 (miR-371), MS00003423 (miR-125a), and MS00006517 (miR-16). Relative gene expression was calculated using the ddCt method and human RNU6B for normalization.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Statistical analysis

Student's t-test was used to evaluate the significance of luciferase reporter data and expression data.

STarMir prediction of miRNA binding sites

We analyzed the data from the 3′ UTR study. Rvf ratio, a measure of the level of miRNA-mediated 3′ UTR regulation relative to a baseline control, is computed from the readouts of the reporter assay. An Rvf ratio of around one indicates a lack of regulation. For ratios below one, a lower ratio indicates a greater level of down-regulation.

We first performed the prediction of miRNA binding sites using the STarMir program [30] for all the miRNA:3′ UTR pairs in the luciferase dataset. Similarly, for data from high throughput mRNA expression and proteomics data [29], we performed prediction of miRNA binding sites for 2406 targets down-regulated at the mRNA levels and 1,635 targets down-regulated at the protein level. This yielded a list of various miRNA site types, including 8mer, 7mer, 6mer (regular or offset) [2], and seedless sites for each miRNA:target pair.

Quantitative modeling

Site-type specific efficacy parameters

Usually, an 8mer seed site is more effective than a 7mer site, which is more effective than a 6mer or a seedless site [2]. To quantify such functional variation solely due to various site types (independent of site/context features), we introduce parameter $\alpha_{\text{site-type}}$ as the average efficacy of a specific site-type for target down-regulation, where site-type can be 8mer, 7mer, 6mer, or seedless, and the following constraint is imposed: $\alpha_{8mer} > \alpha_{7mer} > \alpha_{6mer} > \alpha_{seedless}$.

A range of logistic probability thresholds was considered from 0 to 1.0 with an increment of 0.05, with the hope that we can get estimates under the above constraint. From the luciferase data, we assembled 471 3′ UTR: miRNA pairs with RvF below 1.0, a threshold considered adequate for ensuring down-regulation. This down-regulation dataset can be used to estimate the efficacy parameters. Specifically, we divide the dataset into four subsets. Set 1 includes 3′ UTR: miRNA pairs for which only seedless miRNA binding sites are present (i.e., no 6mer, 7mer, or 8mer sites); set 2 is for 6mer containing 3′ UTR:miRNA

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

pairs, with or without seedless sites (no 7mer, 8mer sites); set 3 is for 3′ UTR:miRNA pairs with 7mers, with or without 6mer or seedless sites (no 8 mer sites); set 4 is for 8mer sites possessing 3′ UTR:miRNA pairs, with or without other site types. Note that a lower RvF ratio shows a higher level of down-regulation, thus a measure of regulatory efficacy can be given by as (1-RvF). For $n1_{seedless}$, the total seedless sites in set 1, and total efficacy 1 as the sum of (1-RvF) for all 3′UTR: miRNA pairs in set 1, $\alpha_{seedless}$, the average efficacy per seedless site, can be estimated from the equation, $\alpha_{seedless} * n1_{seedless}$ = total efficacy 1, i.e., $\alpha_{seedless}$ = total efficacy 1/ $n1_{seedless}$. Similarly, $\alpha_{6mer} * n2_{6mer} + \alpha_{seedless} * n2_{seedless}$ = total efficacy 2, where $n2_{6mer}$ is the total number of 6mer sites, $n2_{seedless}$ is the total number of seedless sites, and total efficacy 2 is the sum of efficacy for this set. With the estimate available for $\alpha_{seedless}$, an estimate for α_{6mer} can be computed from the linear equation, i.e., α_{6mer} = (total efficacy 2 – $\alpha_{seedless} * n2_{seedless}$)/ $n2_{6mer}$. Next, $\alpha_{7mer} * n3_{7mer} + \alpha_{6mer} * n3_{6mer} + \alpha_{seedless} * n3_{seedless}$ = total efficacy 3, where for set 3, $n3_{7mer}$ is the total number of 7mer sites, $n3_{6mer}$ is the total number of 6mer sites, $n3_{seedless}$ is the total number of seedless sites, and total efficacy 3 is the sum of efficacy for this set. With estimates available for $\alpha_{seedless}$ and α_{6mer}, an estimate for α_{7mer} is calculated by (total efficacy 3 – $\alpha_{6mer} * n3_{6mer} - \alpha_{seedless} * n3_{seedless}$)/ $n3_{7mer}$. Finally, $\alpha_{8mer} * n4_{8mer} + \alpha_{7mer} * n4_{7mer} + \alpha_{6mer} * n4_{6mer} + \alpha_{seedless} * n4_{seedless}$ = total efficacy 4, where for set 4, $n4_{8mer}$, $n4_{7mer}$ and $n4_{6mer}$ are total numbers of 8mer, 7mer and 6-mer sites, respectively, and total efficacy 4 is the sum of efficacy for this set. With estimates available for $\alpha_{seedless}$, α_{6mer} and α_{7mer}, an estimate for α_{8mer} is given by (total efficacy 4 – $\alpha_{7mer} * n4_{7mer} - \alpha_{6mer} * n4_{6mer} - \alpha_{seedless} * n4_{seedless}$)/ $n4_{8mer}$.

A quantitative score

We aim to develop a score for quantitative prediction of the level of target down-regulation by miRNAs. The average efficacy parameters address functional variation among different site types. On the other hand, for sites of the same type, e.g., two 8mer seed sites, there can be significant differences in effectiveness due to other site/context features such as structural accessibility. This variation can be addressed by the logistic probability of individual site [22]. With the estimates of the efficacy parameters and logistic probabilities, for a target with *n* predicted sites of various types, we propose to compute a predictive score as a weighted sum of site probabilities: $\sum_{site\ i} Pr_i\ \alpha_{site\text{-}type}(i)$, where i= 1, …, n, and Pr_i and $\alpha_{site\text{-}type}(i)$ are the logistic probability and the site-type specific efficacy for the ith site on the target, respectively. This score not only addresses cooperative regulatory effects of seed and seedless sites, but also takes into account functional differences among different site types as well as differences due to other site features.

Assessment of quantitative score

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To assess the predictive value of the score, we analyze high throughput data from proteomics and microarray studies [29]. We compared the scores of the most down-regulated targets and the least down-regulated ones. Statistical significances were computed for the comparison.

Acknowledgments

This work was supported in part by NIH grants 1R01GM099811, 1R01GM138856 (to Y.D. and J.L.) and 1R01CA149109 (to J.L.), NSF grant DBI-0650991 (to Y.D.), as well as a Connecticut Stem Cell Award to J.L. Additionally, S.M. was partially funded by the NIGMS/NIH-sponsored Predoctoral Training Grant in Genetics T32 GM007499. We thank Dianqing Wu and his lab for technical help.

Disclosure declaration

The authors declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure Legends

Figure 1. Schematic of a high-throughput 3′ UTR reporter assay and the generation of the pilot miRNA-targeting map

A schematic for the high-throughput 3′ UTR luciferase assay was used to generate the pilot miRNA Target Map, in which 461 miRNA expression constructs were assayed pairwise with eleven 3′ UTR dual-luciferase reporters in 293T cells.

Figure 2. **Assay parameter optimization for the high-throughput 3′ UTR reporter assay**

(a) Down regulation of the Klf4 3′ UTR luciferase reporter by miR-25 in 6-, 96-, and 384-well plates was determined. 293T cells were transfected with a control miRNA construct (Ctrl) or miR-25. Normalized luciferase activities are plotted with one (red line) indicating no change. Error bars represent standard deviations. N = 4. **(b)** The seeding density of 293T cells was evaluated using Klf4 3′ UTR reporter assayed with miR-1. The numbers of 293T cells seeded per well in 384-well plates are indicated. Error bars represent standard deviation. N = 4. **(c)** Luciferase activity reading at 24 and 48 hours post-transfection was evaluated for miR-1-mediated suppression of KLF4 3′ UTR and miR-125a-mediated suppression of LIN28 3′ UTR. Normalized luciferase activities are shown. Error bars represent standard deviation. N = 4. **(d)** MiRNA constructs for miR-371, miR-125a, and miR-16, representing low, medium and high endogenous levels, were evaluated for overexpression in 293T cells 48 hours post transfection. A control miRNA construct (Ctrl) was used for comparison. MiRNA expression was determined by qRT-PCR. Representative data are shown. Error bars represent standard deviation. N = 3. **(e)** The regulation of miR-25 on KLF4 3′ UTR was evaluated in 293T and HeLa cells in comparison with a control miRNA construct. Error bars represent standard deviation. For HeLa, N=3; For 293T, N=4. For all sub-figures, *$p<0.01$.

Figure 3. The pilot miRNA-targeting map

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(a) Data normalization strongly reduced data variation. Raw and normalized *Renilla* vs firefly luciferase activities were plotted in a histogram. Note the shrinking of variation and near normal distribution of the normalized data. **(b)** Most assays in the pilot miRNA-targeting map have low levels of assay variation. A histogram is shown for standard deviation versus mean for the 4,994 miRNA:3′ UTR pairs assayed. Red dashed lines indicate 0.1 and 0.15 of standard deviation versus mean values. **(c)** A heatmap is shown to reflect the mean values of the 181 miRNA:3′ UTR pairs that resulted in ≥25% downregulation, with rows representing 3′ UTRs and columns representing miRNAs. Data were filtered to retain only those miRNAs that resulted in ≥25% downregulation for any of the assayed 3′ UTRs and were organized by hierarchical clustering afterward. Deeper blue reflects stronger suppression. MiRNA:3′ UTR pairs that did not result in ≥25% downregulation is set to 1 and represented by white color. **(d)** Eleven miRNA:3′ UTR pairs were evaluated on the level of endogenous protein. Corresponding miRNAs were transduced into NCCIT cells or K562 cells, and the corresponding endogenous protein changes were determined by western blot analysis. The mean luciferase fold changes from the pilot miRNA-targeting map were plotted against protein level fold change. Western data reflect averages of two independent biological replicates. Circles indicate analysis in NCCIT cells; squares indicate K562 cells. The color red indicates interactions predicted by TargetScan, PITA and mirSVR; blue indicates those interactions that were not predicted.

Figure 4. Validation of score by mRNA expression data

For each of the seed component score **(a)**, the seedless component score **(b)** and the total score **(c)**, the genes with the highest down-regulation in mRNA levels have significantly higher scores than those with lowest levels of down-regulation.

Figure 5. Validation of score by protein expression data

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For each of the seed component score **(a)**, the seedless component score **(b)** and the total score **(c)**, the genes with highest down-regulation in protein levels has a significantly higher score than those with lowest levels of down-regulation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 1: Calculation of site-type specific efficacy

No. of 8mer sites	No. of 7mer sites	No. of 6mer sites	No. of seedless sites	RvF mean	No. of 3'UTR:miRNA pairs	Efficacy
28	0	10	699	20.39	24	0.01980
0	19	6	591	14.18	17	0.01327
0	0	76	1389	49.55	56	0.00658
0	0	0	7279	342.83	374	0.00428

Table 2: An example of score calculation for NM_000018

Target	miRNA	No. of 8mer sites	No. of 7mer sites	No. of 6mer sites	No. of seedless sites	$Score_{seed}$	$Score_{seedless}$	$Score_{total}$
NM_000018	hsa-miR-1	0	0	2	22	0.00444	0.04372	0.04816
NM_000018	hsa-miR-124	0	1	1	131	0.00587	0.26711	0.27298

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116 (2):281-297. doi:10.1016/s0092-8674(04)00045-5
2. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136 (2):215-233. doi:10.1016/j.cell.2009.01.002
3. Farazi TA, Spitzer JI, Morozov P, Tuschl T (2011) miRNAs in human cancer. J Pathol 223 (2):102-115. doi:10.1002/path.2806
4. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics 25 (23):3049-3055. doi:10.1093/bioinformatics/btp565
5. Rajewsky N (2006) microRNA target predictions in animals. Nat Genet 38 Suppl:S8-13. doi:10.1038/ng1798
6. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460 (7254):479-486. doi:10.1038/nature08170
7. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141 (1):129-141. doi:10.1016/j.cell.2010.03.009
8. Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell 153 (3):654-665. doi:10.1016/j.cell.2013.03.043
9. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods 8 (7):559-564. doi:10.1038/nmeth.1608
10. Lai EC (2002) Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet 30 (4):363-364. doi:10.1038/ng865
11. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 27 (1):91-105. doi:10.1016/j.molcel.2007.06.017
12. Vella MC, Choi EY, Lin SY, Reinert K, Slack FJ (2004) The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. Genes Dev 18 (2):132-137. doi:10.1101/gad.1165404
13. Didiano D, Hobert O (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol 13 (9):849-851. doi:10.1038/nsmb1138
14. Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, O'Day E, Chowdhury D, Dykxhoorn DM, Tsai P, Hofmann O, Becker KG, Gorospe M, Hide W, Lieberman J (2009) miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. Mol Cell 35 (5):610-625. doi:10.1016/j.molcel.2009.08.020

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

15. Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB, Leslie CS, Rudensky AY (2012) Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol Cell 48 (5):760-770. doi:10.1016/j.molcel.2012.10.002
16. Nam JW, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, Yildirim MA, Rodriguez A, Bartel DP (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. Mol Cell 53 (6):1031-1043. doi:10.1016/j.molcel.2014.02.013
17. Shin C, Nam JW, Farh KK, Chiang HR, Shkumatava A, Bartel DP (2010) Expanding the microRNA targeting code: functional sites with centered pairing. Mol Cell 38 (6):789-802. doi:10.1016/j.molcel.2010.06.005
18. Xie C, Huang H, Sun X, Guo Y, Hamblin M, Ritchie RP, Garcia-Barrio MT, Zhang J, Chen YE (2011) MicroRNA-1 regulates smooth muscle cell differentiation by repressing Kruppel-like factor 4. Stem Cells Dev 20 (2):205-210. doi:10.1089/scd.2010.0283
19. Robins H, Li Y, Padgett RW (2005) Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A 102 (11):4006-4009. doi:10.1073/pnas.0500775102
20. Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y (2007) Potent effect of target structure on microRNA function. Nat Struct Mol Biol 14 (4):287-294. doi:10.1038/nsmb1226
21. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39 (10):1278-1284. doi:10.1038/ng2135
22. Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, Ding Y (2013) CLIP-based prediction of mammalian microRNA binding sites. Nucleic Acids Res 41 (14):e138. doi:10.1093/nar/gkt435
23. Cheng J, Guo S, Chen S, Mastriano SJ, Liu C, D'Alessio AC, Hysolli E, Guo Y, Yao H, Megyola CM, Li D, Liu J, Pan W, Roden CA, Zhou XL, Heydari K, Chen J, Park IH, Ding Y, Zhang Y, Lu J (2013) An extensive network of TET2-targeting MicroRNAs regulates malignant hematopoiesis. Cell Rep 5 (2):471-481. doi:10.1016/j.celrep.2013.08.050
24. Song SJ, Poliseno L, Song MS, Ala U, Webster K, Ng C, Beringer G, Brikbak NJ, Yuan X, Cantley LC, Richardson AL, Pandolfi PP (2013) MicroRNA-antagonism regulates breast cancer stemness and metastasis via TET-family-dependent chromatin remodeling. Cell 154 (2):311-324. doi:10.1016/j.cell.2013.06.026
25. Guo S, Bai H, Megyola CM, Halene S, Krause DS, Scadden DT, Lu J (2012) Complex oncogene dependence in microRNA-125a-induced myeloproliferative neoplasms. Proc Natl Acad Sci U S A 109 (41):16636-16641. doi:10.1073/pnas.1213196109
26. Kuhn AR, Schlauch K, Lao R, Halayko AJ, Gerthoffer WT, Singer CA (2010) MicroRNA expression in human airway smooth muscle cells: role of miR-25 in regulation of airway smooth muscle phenotype. Am J Respir Cell Mol Biol 42 (4):506-513. doi:10.1165/rcmb.2009-0123OC
27. Adams BD, Guo S, Bai H, Guo Y, Megyola CM, Cheng J, Heydari K, Xiao C, Reddy EP, Lu J (2012) An in vivo functional screen uncovers miR-150-mediated regulation of hematopoietic injury response. Cell Rep 2 (4):1048-1060. doi:10.1016/j.celrep.2012.09.014
28. Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11 (8):R90. doi:10.1186/gb-2010-11-8-r90
29. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP (2008) The impact of microRNAs on protein output. Nature 455 (7209):64-71. doi:10.1038/nature07242
30. Rennie W, Liu C, Carmack CS, Wolenc A, Kanoria S, Lu J, Long D, Ding Y (2014) STarMir: a web server for prediction of microRNA binding sites. Nucleic Acids Res 42 (Web Server issue):W114-118. doi:10.1093/nar/gku376

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.05.626985; this version posted December 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

31. Lipchina I, Elkabetz Y, Hafner M, Sheridan R, Mihailovic A, Tuschl T, Sander C, Studer L, Betel D (2011) Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev 25 (20):2173-2186. doi:10.1101/gad.17221311
32. Xu N, Papagiannakopoulos T, Pan G, Thomson JA, Kosik KS (2009) MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. Cell 137 (4):647-658. doi:10.1016/j.cell.2009.02.038
33. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39 (Database issue):D152-157. doi:10.1093/nar/gkq1027
34. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR (2005) MicroRNA expression profiles classify human cancers. Nature 435 (7043):834-838. doi:10.1038/nature03702
35. Lu J, Guo S, Ebert BL, Zhang H, Peng X, Bosco J, Pretz J, Schlanger R, Wang JY, Mak RH, Dombkowski DM, Preffer FI, Scadden DT, Golub TR (2008) MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. Dev Cell 14 (6):843-853. doi:10.1016/j.devcel.2008.03.012

Figure 1

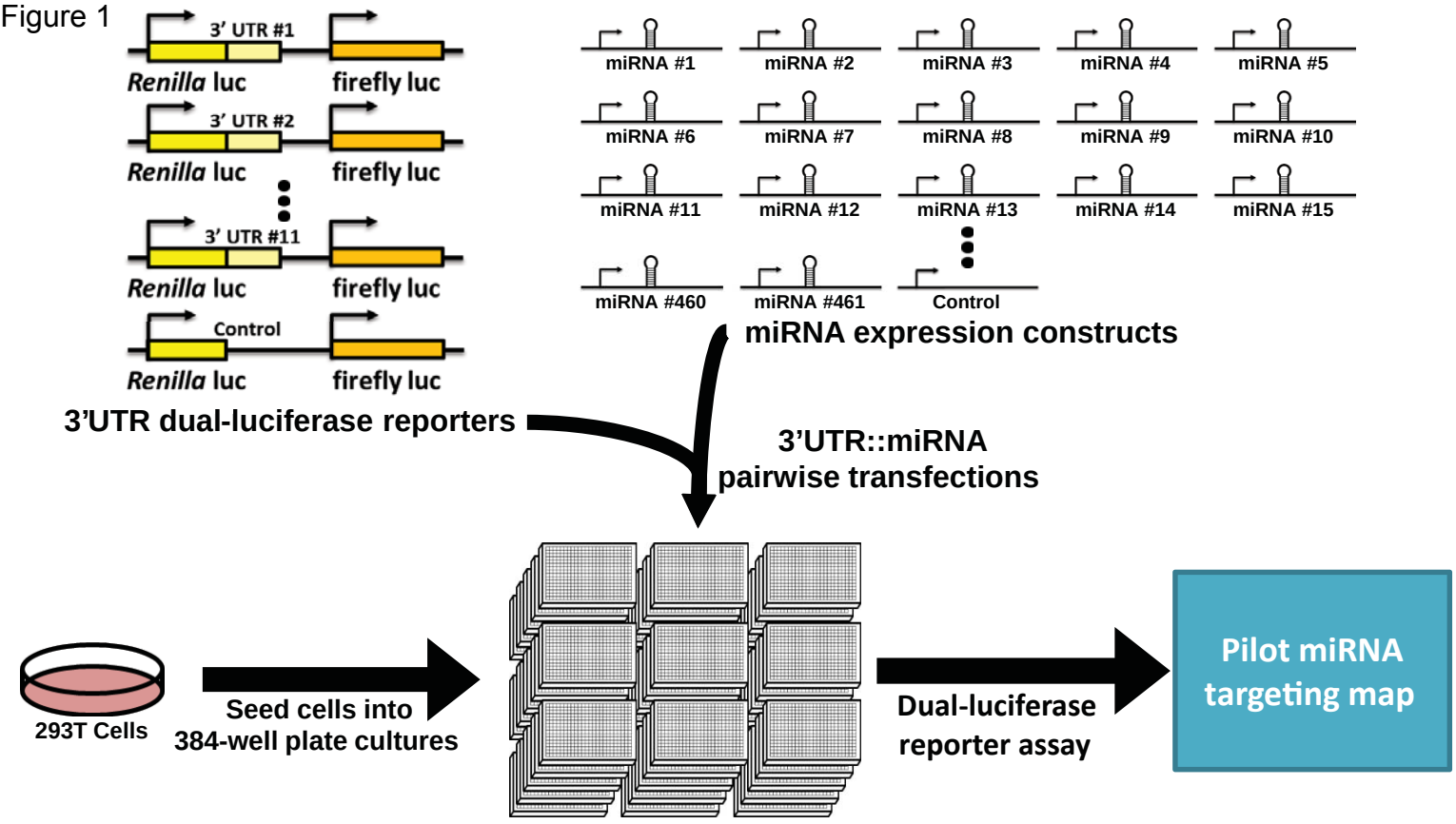

3' UTR #1
Renilla luc
firefly luc
3' UTR #2
Renilla luc
firefly luc
3' UTR #11
Renilla luc
firefly luc
Control
Renilla luc
firefly luc
3'UTR dual-luciferase reporters
miRNA #1
miRNA #2
miRNA #3
miRNA #4
miRNA #5
miRNA #6
miRNA #7
miRNA #8
miRNA #9
miRNA #10
miRNA #11
miRNA #12
miRNA #13
miRNA #14
miRNA #15
miRNA #460
miRNA #461
Control
miRNA expression constructs
3'UTR::miRNA
pairwise transfections
293T Cells
Seed cells into
384-well plate cultures
Dual-luciferase
reporter assay
Pilot miRNA
targeting map

Figure 2

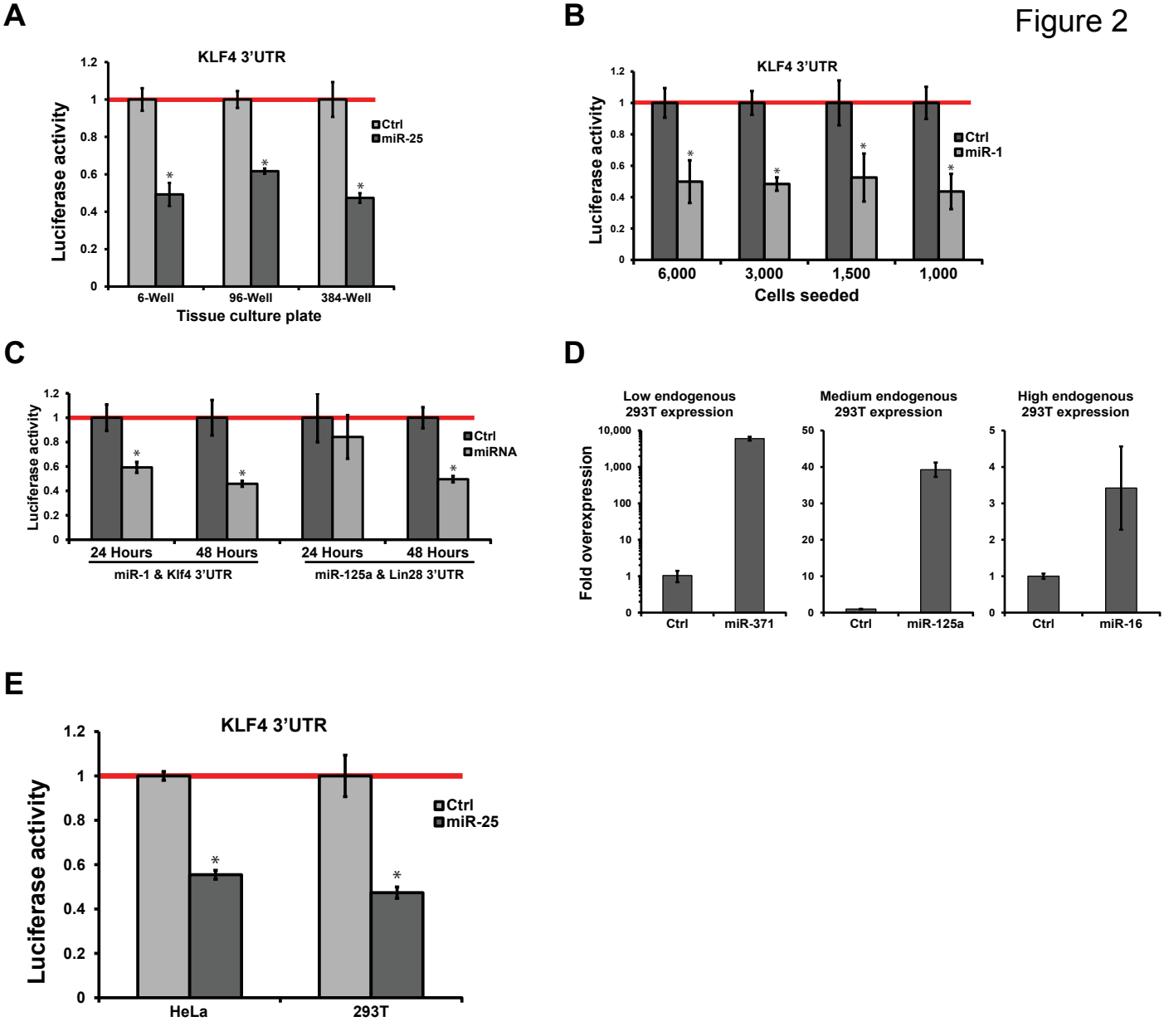

A
KLF4 3'UTR
Luciferase activity
1.2
1
0.8
0.6
0.4
0.2
0
Ctrl
miR-25
6-Well
96-Well
384-Well
Tissue culture plate
B
KLF4 3'UTR
Luciferase activity
Ctrl
miR-1
6,000
3,000
1,500
1,000
Cells seeded
C
Luciferase activity
Ctrl
miRNA
24 Hours
48 Hours
24 Hours
48 Hours
miR-1 & Klf4 3'UTR
miR-125a & Lin28 3'UTR
D
Low endogenous 293T expression
Medium endogenous 293T expression
High endogenous 293T expression
Fold overexpression
10,000
1,000
100
10
1
0
50
40
30
20
10
5
4
3
2
Ctrl
miR-371
Ctrl
miR-125a
Ctrl
miR-16
E
KLF4 3'UTR
Luciferase activity
Ctrl
miR-25
HeLa
293T

Figure 3

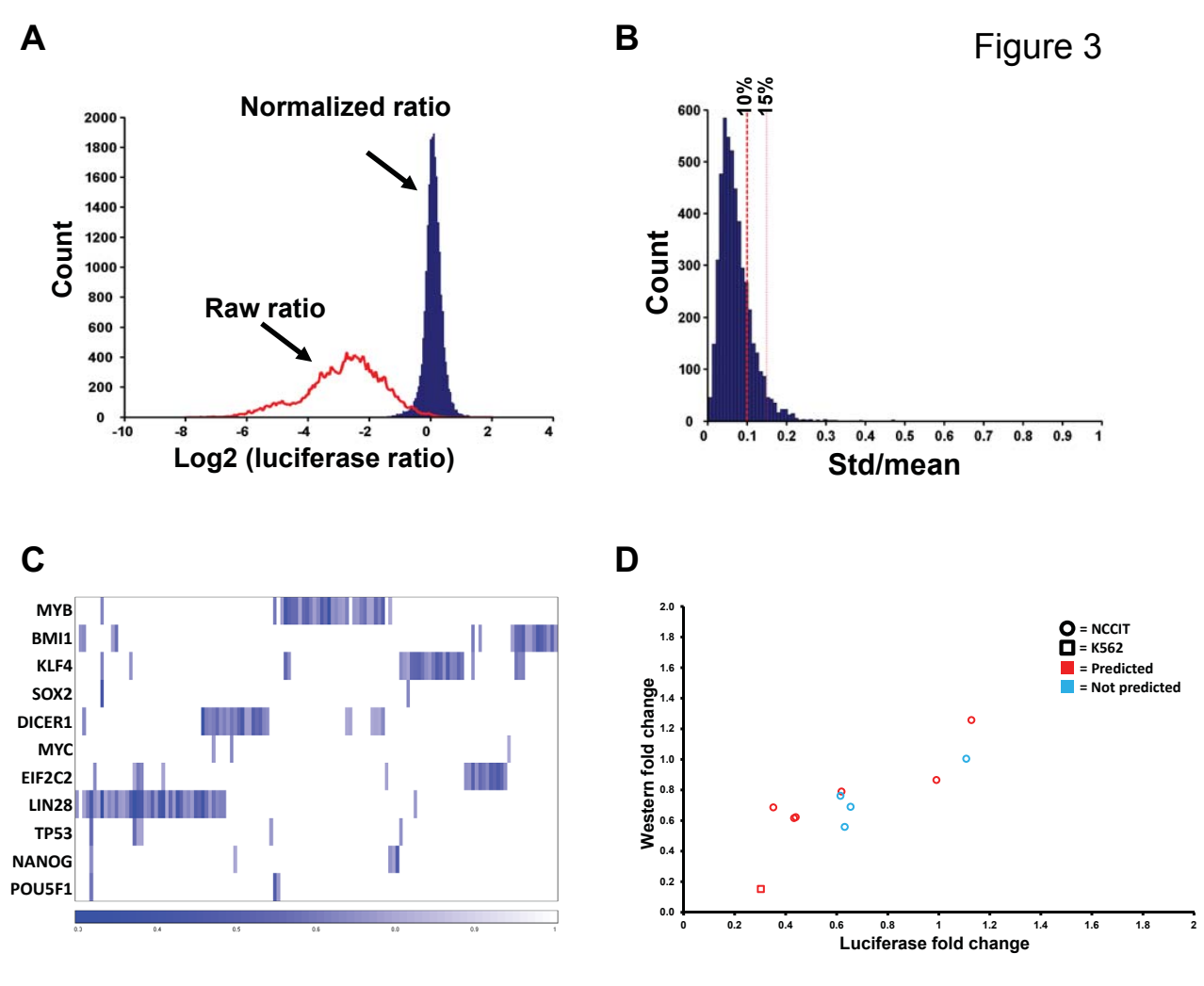
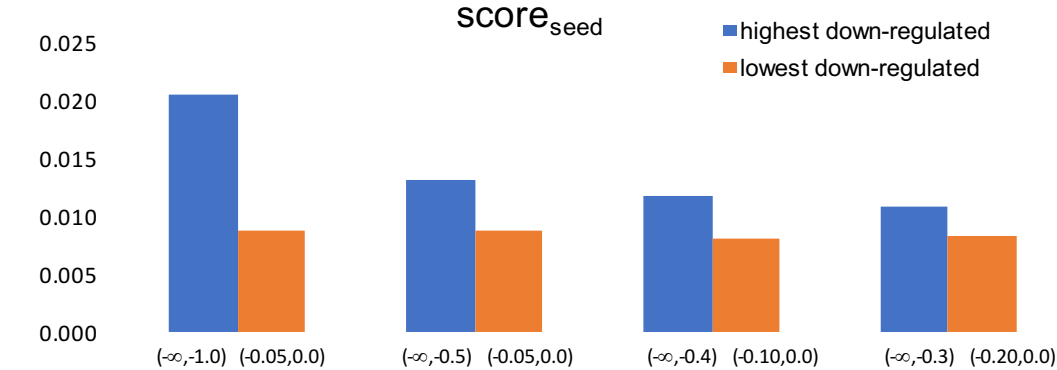
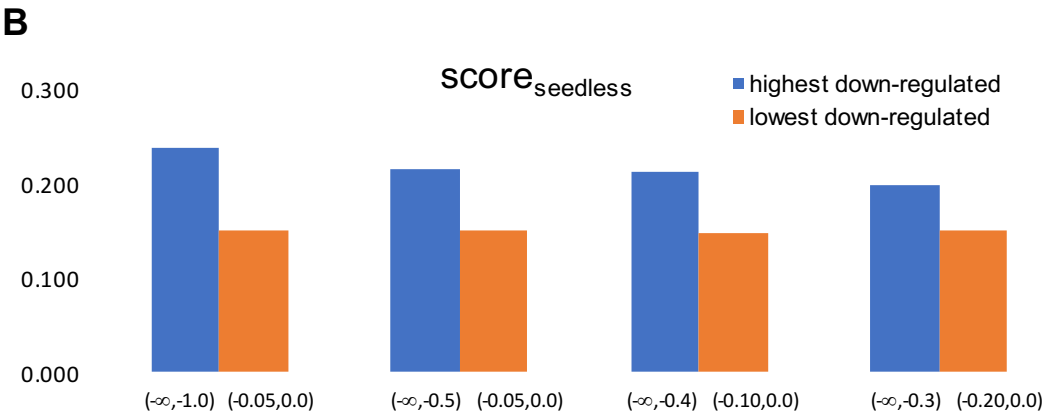

A
Normalized ratio
Raw ratio
Count
Log2 (luciferase ratio)
B
10%
15%
Count
Std/mean
C
MYB
BMI1
KLF4
SOX2
DICER1
MYC
EIF2C2
LIN28
TP53
NANOG
POU5F1
D
= NCCIT
= K562
= Predicted
= Not predicted
Western fold change
Luciferase fold change

Figure 4


A
score seed
highest down-regulated
lowest down-regulated
0.025
0.020
0.015
0.010
0.005
0.000
(-∞,-1.0) (-0.05,0.0)
(-∞,-0.5) (-0.05,0.0)
(-∞,-0.4) (-0.10,0.0)
(-∞,-0.3) (-0.20,0.0)


B
score seedless
highest down-regulated
lowest down-regulated
0.300
0.200
0.100
0.000
(-∞,-1.0) (-0.05,0.0)
(-∞,-0.5) (-0.05,0.0)
(-∞,-0.4) (-0.10,0.0)
(-∞,-0.3) (-0.20,0.0)

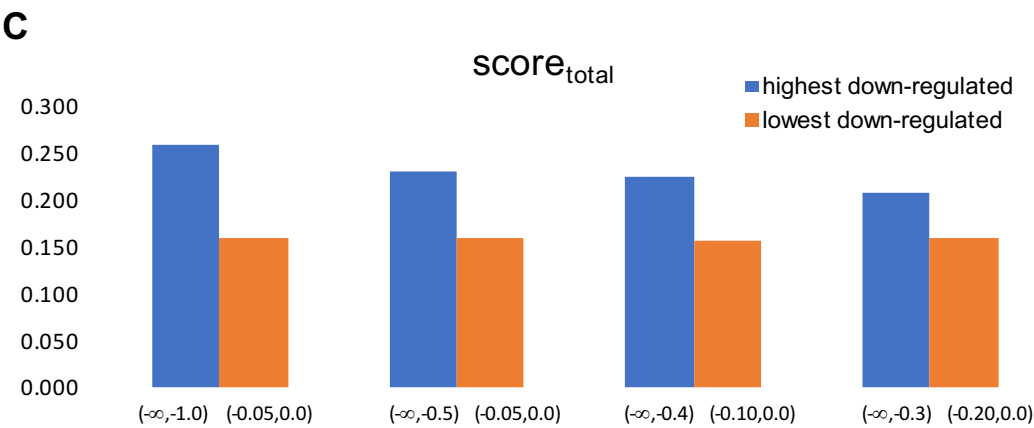
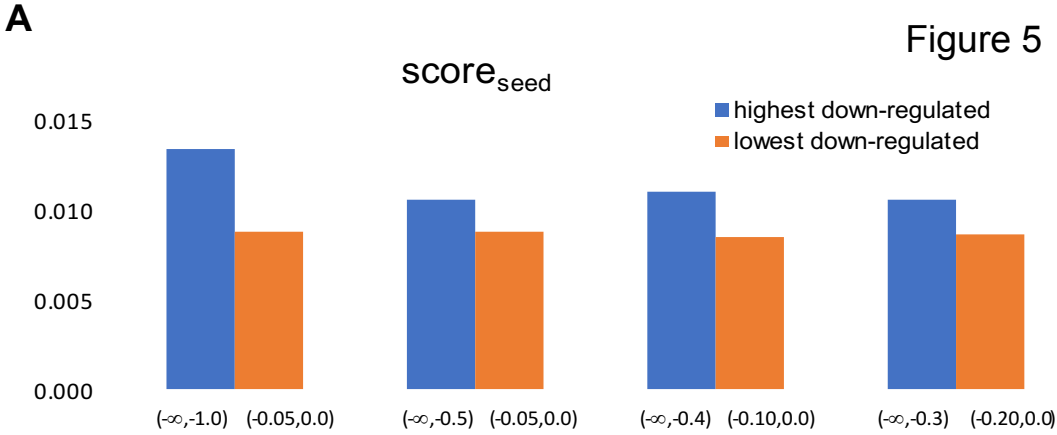
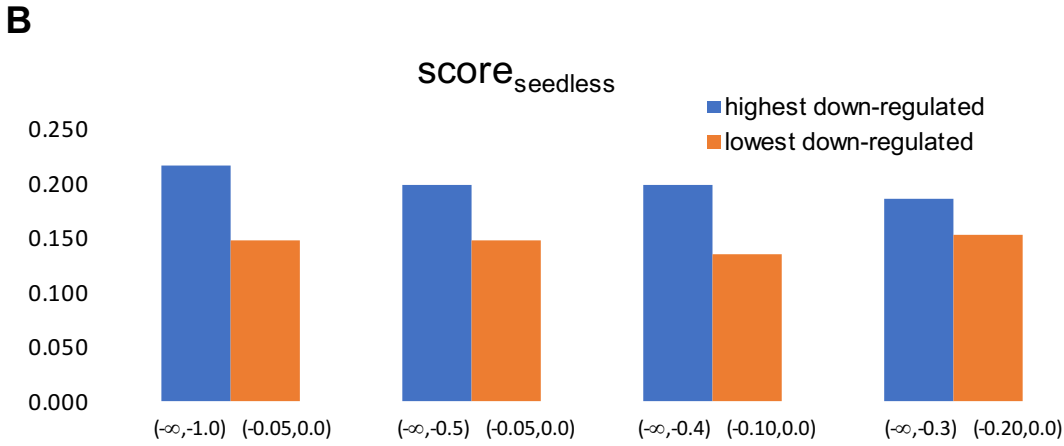
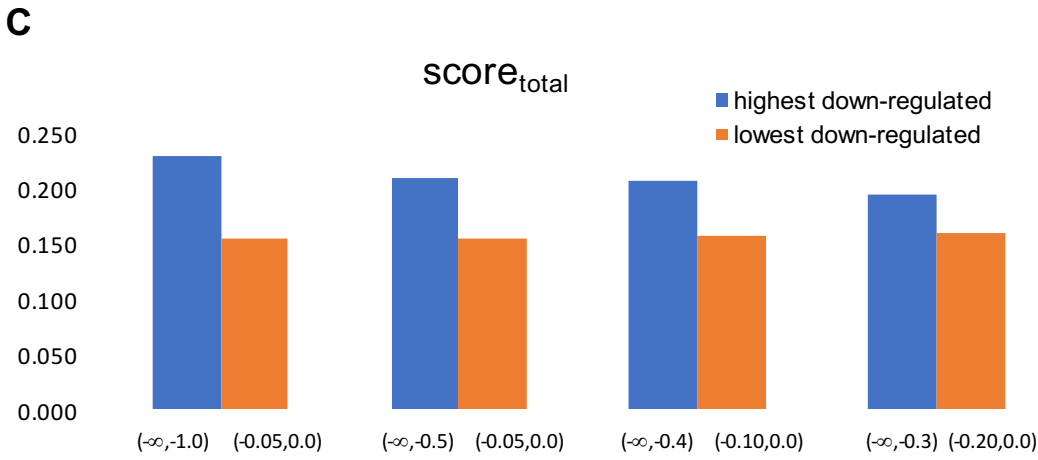

C
score total
highest down-regulated
lowest down-regulated
0.300
0.250
0.200
0.150
0.100
0.050
0.000
(-∞,-1.0) (-0.05,0.0)
(-∞,-0.5) (-0.05,0.0)
(-∞,-0.4) (-0.10,0.0)
(-∞,-0.3) (-0.20,0.0)

Figure 5


A
score seed
highest down-regulated
lowest down-regulated
0.015
0.010
0.005
0.000
(-∞,-1.0)
(-0.05,0.0)
(-∞,-0.5)
(-0.05,0.0)
(-∞,-0.4)
(-0.10,0.0)
(-∞,-0.3)
(-0.20,0.0)


B
score seedless
highest down-regulated
lowest down-regulated
0.250
0.200
0.150
0.100
0.050
0.000
(-∞,-1.0)
(-0.05,0.0)
(-∞,-0.5)
(-0.05,0.0)
(-∞,-0.4)
(-0.10,0.0)
(-∞,-0.3)
(-0.20,0.0)


C
score total
highest down-regulated
lowest down-regulated
0.250
0.200
0.150
0.100
0.050
0.000
(-∞,-1.0)
(-0.05,0.0)
(-∞,-0.5)
(-0.05,0.0)
(-∞,-0.4)
(-0.10,0.0)
(-∞,-0.3)
(-0.20,0.0)