

Supplemental Information

Genome-wide Dissection of MicroRNA Functions and Cotargeting Networks Using Gene Set Signatures

John S. Tsang, Margaret S. Ebert, and Alexander van Oudenaarden

Supplemental Figures

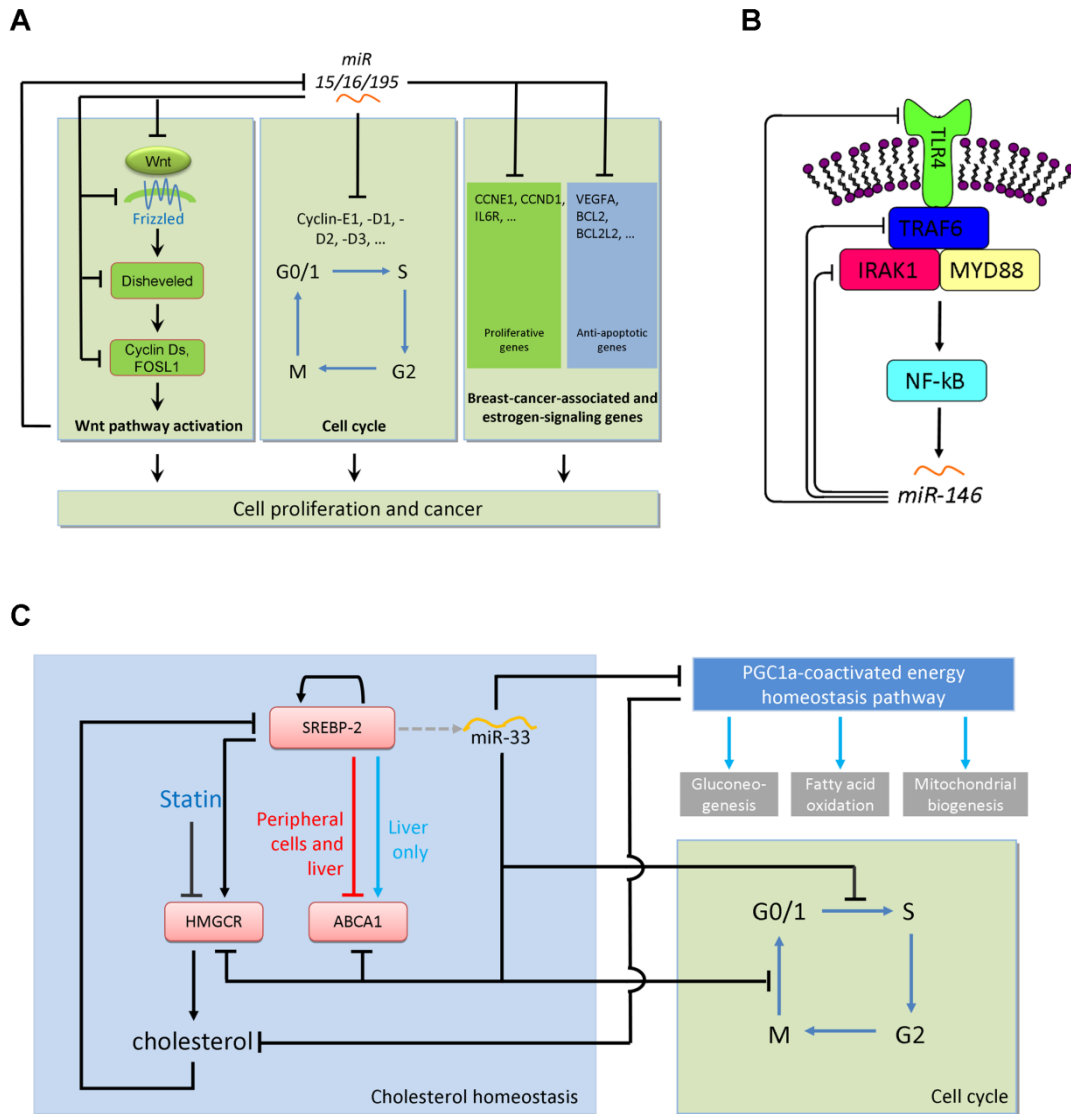


Figure S1 Network diagrams of selected mirBridge predictions discussed in the main text (related to Figure 2). Aside from the miRNA targeting links, the networks are compiled based on the literature. (A) [mirBridge](#) predicts that *miR-15/16/195* could regulate several intricately linked pathways that control cell proliferation and cancer, suggesting that a general function of the *miR-15/16/195* family is to control proliferation and/or growth. Several putative targets have multiple high-quality seed-matched sites (Table S1a). (B) [mirBridge](#) indicates that *miR-146* functions in NF- κ B, IL4 and TOLL pathways where *miR-146* mediates several negative feedback loops to upstream signaling factors. (C) [mirBridge](#) indicates that *miR-33* functions in cholesterol homeostasis. *miR-33a* is probably co-expressed with SREBP2 because it is embedded in an intron of SREBP2. *miR-33* also putatively regulates the cell cycle network and the PGC1a pathway, forming a double-negative (i.e. positive) feedback to cholesterol.

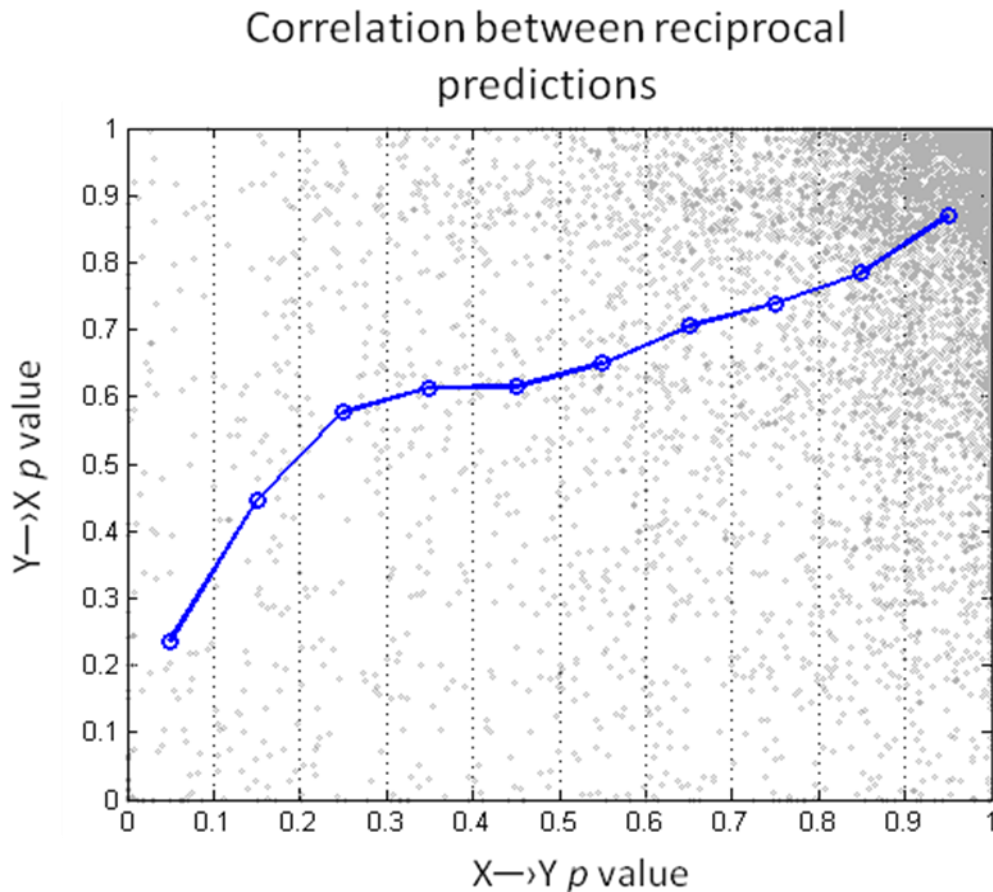


Figure S2 (related to Figure 3) Correlation between reciprocal co-targeting predictions For each miRNA-family pair (X,Y), the lowest [mirBridge](#) p values of X→Y and Y→X are plotted against each other. The entries were partitioned into 10 bins by the X→Y p values and the average Y→X p value was computed and plotted against the

average $X \rightarrow Y$ p value of each bin (resulting in the blue line). The reciprocal p values are significantly correlated (Spearman correlation = 0.42, $p=0$). It is important to note that while many miRNA families are reciprocal co-targeting pairs ($X \leftrightarrow Y$), it is biologically plausible that $X \rightarrow Y$ need not imply $Y \rightarrow X$. For instance, Y may function in more diverse contexts than X , yet co-targeting may be functionally important only in the contexts where X functions. A likely example, albeit on the more extreme end, involves the *miR-99/100* and *miR-125/351* families with 80 and 1362 predicted targets, respectively. The PTS of *miR-99/100* has a large number of seed-matched sites for *miR-125/351* with a significant fraction of those being conserved and/or having high context scores, yielding a q -value of 0.03. In contrast, the reciprocal q -value is 0.92 because the larger *miR-125/351* PTS only contains a small number of sites for *miR-99/100*, and an insignificant fraction of those are conserved and/or have high context scores, suggesting that most of *miR-125/351*'s functional contexts are not shared with *miR-99/100*. A similar example involves the *miR-17* and *-18* families where the latter has a smaller PTS. Individual cases aside, PTS-size difference is not a major contributing factor: the size-difference distribution between PTSs for miRNA-family pairs having both $X \rightarrow Y$ and $Y \rightarrow X$ q -values of less than 0.2 do not significantly deviate from those pairs with a significant p -value in only one direction ($p=0.24$ Kolmogorov-Smirnov Test).

Supplemental Experimental Procedures

The **mirBridge** algorithm

Inputs:

1. A set **M** of miRNA seed-matched motifs. The motifs can be partitioned into two classes: **m2-8** and **m1-7-A-anchor**
2. A gene set **G** with n genes and their 3' UTRs
3. The context score of all seed-matches (from **M**) in the 3' UTRs in **G**
4. A context score threshold (t)

Processing:

1. For each motif m in the class **m2-8**, determine the following statistics in **G**:
 - a. The number of seed matches (N) (for OC)
 - b. The number of genes (T) in **G** with at least one seed-matched site
 - c. The number of conserved seed matches (K) (for CE)
 - d. The number of seed-matches (H) with context score in the top t -percentile (for CTX)
2. Build the gene neighborhood:
 - a. For each gene g in **G**, build an ordered array A of gene neighbors by sorting all 3' UTR x in the genome by the normalized Euclidean distance between the 3' UTRs

of g and x using length, GC content, and general conservation. $A[1]$ is the closest, $A[2]$ the next closest, and so on.

3. Build the putative target neighborhoods for each miRNA motif:
 - a. For each motif m from Step 1, form ordered array A_m (as in Step 2) for each g in G by removing entries in the corresponding A that do not contain the motif m in its 3' UTR (i.e. not a putative target of the miRNA)
4. Compute the null distributions for N (the number of seed-matches)
 - a. Determine the bandwidth parameter σ
 - i. For each α from a list of possible σ 's
 1. For each g in G
 - a. draw a random number x from the Gaussian density with mean 0 and variance α^2
 - b. round x to the nearest integer and take its absolute value
 - c. use x to index to g 's neighbor list to draw a gene/3' UTR; i.e. $A[x]$
 2. Compute the Kolmogorov-Smirnov p value between the drawn random gene set and G for each of length, GC-content, and general conservation
 3. Repeat the above for 100 (or more) times
 4. Take the average p value for each of length, GC content, and general conservation over the 100 iterations
 - ii. Pick the largest α such that the lowest of the three average p values must be greater than a given threshold (currently set to 0.67)
 - b. Repeat 10,000 times (or more)
 - i. Using the σ from Step a, draw a random gene set R as in Step 4-a-i-1
 - ii. For each seed-matched motif in Step 1, compute N as in Step 1 for the random gene set to obtain the null distribution for N
5. Compute the null distributions for K (the number of conserved sites) and H (the number of high-context-scoring sites) conditional on T
 - a. For each motif from Step 1
 - i. Identify the putative targets in G (i.e. genes in G with at least one site)
 - ii. Determining the bandwidth parameter as in Step 4a except: 1) use the putative target neighborhood for the motif (from Step 3); 2) only use the putative targets as members of G (i.e. ignore/remove genes without sites)
 - iii. Using the procedure in Step 4-a-i-1, generate random putative target sets by replacing each of the putative targets in G with a randomly sampled putative target from the putative target neighborhood array (A_m) for the motif and gene (from Step 3). Note that each random target set would have exactly T genes with at least one motif site
 - iv. For each random target set, compute K and H (note that by design each random target set has exactly T putative targets)
 - v. Repeat 10,000 times (or more) to obtain the null K and H distributions conditional on T

6. Compute the p value for N by using the null distribution from Step 4: count the percentage of random gene sets that have an equal or higher N . Similarly compute the p values for K and H by using the null distributions from Step 5: count the percentage of random putative target sets that have an equal or higher corresponding statistics.

7. FDR analysis: computing the q values across all **m2-8** motifs:

- a. Use $\lambda = 0.5$ to estimate the proportion of null features, π_0 , by counting the number of p values that are greater than 0.5; and divide this by $n(1 - \lambda)$.
- b. For each motif with p value p_i
 - i. Use $\lambda = 0.5$, and estimate the proportion of null features by counting the

$$\widehat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, 2, \dots, n\}}{n(1 - \lambda)}$$

- ii. Compute q_i as $q_i = \frac{\pi_0 \cdot n \cdot t}{\#\{p_j < p_i; j = 1, 2, \dots, n\}}$

- c. For each q_i , set q_i to q_j where q_j is the minimum of all q values for which the corresponding p values are greater than p_i .

8. Construct the composite test statistics (CE-CTX and OC-CE-CTX) and compute the corresponding p and q values (modified inverse-normal method):

- a. For each p value from the basic statistics (i.e. p_{ce}, p_{oc}, p_{ctx}), compute $t_{ce} = \Phi^{-1}(1 - p_{ce})$ where Φ^{-1} is the inverse of the standard normal cumulative distribution function (i.e. normal with mean 0 and std 1). Similarly compute t_{oc} and t_{ctx} .

- b. Construct the composite statistics for each motif in G:

$$t_{ce_ctx} = w_{ce} t_{ce} + w_{ctx} t_{ctx}$$

$$t_{oc_ce_ctx} = w_{ce} t_{ce} + w_{ctx} t_{ctx} + w_{oc} t_{oc}$$

where $w_{ce} + w_{ctx} = 1$ and $w'_{ce} + w'_{ctx} + w'_{oc} = 1$.

The w 's can be adjusted to assign different weights to basic statistics (currently $w_{ce} = 0.5, w_{ctx} = 0.5; w'_{ce} = 0.4, w'_{ctx} = 0.35, w'_{oc} = 0.25$).

- c. To compute p values, obtain the null distributions of t_{ce_ctx} and $t_{oc_ce_ctx}$ by the following method:

- i. Compute the covariance between each pair of t_{ce}, t_{oc}, t_{ctx} (by using values across all motifs. In cases where multiple gene sets are considered, values from all motif-gene-set combinations can be used)

- ii. Compute the variance of t_{ce_ctx} and $t_{oc_ce_ctx}$ by using the formula:

$$\text{var}(at_1 + bt_2 + ct_3) = a^2 \text{var}(t_1) + b^2 \text{var}(t_2) + c^2 \text{var}(t_3) + 2ab \cdot \text{cov}(t_1, t_2) + 2ac \cdot \text{cov}(t_1, t_3) + 2bc \cdot \text{cov}(t_2, t_3)$$

- iii. Compute the means of t_{ce_ctx} and $t_{oc_ce_ctx}$ by using the formula:

$$\text{mean}(at_1 + bt_2 + ct_3) = a \cdot \text{mean}(t_1) + b \cdot \text{mean}(t_2) + c \cdot \text{mean}(t_3)$$

- iv. The null distributions of t_{ce_ctx} and $t_{oc_ce_ctx}$ are normal distributions with the mean and variances computed above.
 - v. The p values of the observed statistics for each motif can be computed from the null distributions.
- d. Compute the q values for the composite p values as in Step 7.
9. Repeat steps 1-8 for **m1-7-A-anchor** motifs

Output:

For each input motif, the q value of each test is provided.

Note:

If multiple gene sets are being tested simultaneously, the FDR procedure (Step 7) can be adjusted to include p values from all motif-gene-set combinations. Similarly for Step 8c the t 's from all motif-gene-set combinations can be used to estimate the covariances and to compute the q values (Step 8d).

The mirBridge null model

The discussion below focuses on defining the appropriate null models for the test statistics used in [mirBridge](#) (i.e. CE, CTX, OC). As discussed, the null model of the CE and CTX tests is based on randomizing putative target sets while that of OC is based on randomizing the entire gene set (Fig. 1 in main text). The main task is, however, that of generating a random set of genes that has similar properties as a particular gene set (i.e. for [mirBridge](#) the gene set can be a putative target set or the input gene set itself). Thus the following discussion revolves around “gene sets,” but it should be understood that it equally applies to “putative target sets.”

The simplest null model is to generate size-matched uniformly sampled random gene sets. However, as discussed in the main text, this can be an inappropriate null model because other factors, such as general (or non-specific) motif conservation rate, may lead to systematic biases. Below these key factors are empirically analyzed to show that they can indeed introduce systematic biases. The analysis of 3' UTR length is omitted because it is obvious that it is correlated with motif occurrences.

General evolutionary rate For a given 3' UTR, the general (or non-specific) conservation rate is defined as the number of conserved 7-mers (because seed matches are 7-mers) divided by the total number of 7-mers (i.e. 3' UTR length – 6). By counting only the occurrences of a particular motif type, a similar definition is used for the conservation rate of a motif. To investigate whether non-specific conservation rate can affect the CE statistic, the general conservation rate and conservation rate of each seed-

matched motif were computed for all 3' UTRs. The Spearman correlation¹ between the general and specific conservation rates for each motif was computed across all human 3' UTRs, resulting in 314 correlation coefficients (one for each of the Targetscan seed motifs of conserved miRNAs) (Fig. S3). 309 out of 314 of the motifs exhibit significant correlations ($p < 0.01$). To ensure that the correlation is not primarily due to unusually short 3' UTRs, the correlations were recomputed using only 3' UTRs that are longer than 1000 bp; the same result holds (Fig. S3). The significant correlations persist when the correlation between general conservation rate and the occurrence count of each motif were computed (309/314 have $p < 0.01$) even though the absolute correlation coefficients are lower (Fig. S4).

This analysis strongly indicates that non-specific conservation rate is a strong predictor for the conservation rate of specific motifs. Therefore, an effective null model has to take the general conservation level of a gene set into account. For instance, genes in many biological gene sets, such as the human PIP3 signaling pathway in cardiac myocytes, have significantly higher general conservation levels than the rest of the genome (Fig. S5).

GC content A key property used to compute the context score is the GC content around the seed match: higher GC contents can lead to more stable local secondary structures that block miRNA-RISC access (Grimson et al., 2007). This implies that the overall GC content of the 3' UTR can have an effect on the context score. To investigate this possibility, the percentage of bases that are either G or C was computed for each 3' UTR. The Spearman correlation between the percent-GC and the context score for each type of seed match was computed across all 3' UTRs, resulting in 314 correlation coefficients (Fig. S6). 304 out of 314 motifs exhibit significant negative correlation at $p < 0.01$, indicating that the overall GC content of the 3' UTR is a strong predictor of the context score.

Correlation between different factors Significant pair-wise correlation exists between length (L), GC-content (GC), and general conservation rate (C) across human 3' UTRs, indicating that accounting for systematic biases introduced by any one of the factors alone can over- or under-compensate others (table below). An effective null model needs to consider all factors simultaneously (see below).

Variable Pair	Spearman correlation	Simulated P value
L-C	0.185	0
L-GC	-0.085	0
C-GC	-0.125	0

Additional factors So far three gene set properties (length, GC content and general conservation) that can introduce systematic biases have been discussed. A key

¹ A non-parametric correlation measure is used because the normality assumption does not hold

question is whether additional factors need to be considered. In other words, are other factors largely conditionally independent² of the test statistics given L , GC , and C ? This is a difficult question to answer empirically because there are a large number of possible factors. For instance, can the occurrence rates of certain k-mers ($k=2, 3, 4\dots$) affect the context score and/or evolutionary rate of certain seed-matched motifs? The frequency of a given k-mer can affect the frequency of motifs containing subsequences that are correlated in frequency to the k-mer. However, aside from OC, our test statistics are conditional on N , so factors that affect motif frequencies are unlikely to have a significant effect (as discussed in the main text, OC is only used in the composite score but is not used alone as an indication of functional targeting). A related concern is that the evolutionary rate of a subset of the motifs may be dependent upon the frequency of some k-mers, but such dependencies should be largely captured by the general conservation rate measure, especially if the number of affected motifs is relatively large. In fact, one would not want to miss the signal if the differential rate is specific to a small set of motifs, because such signals can reflect constraints imposed by miRNA-mediated regulation. L , GC , and C are likely the most direct gene-set properties that affect the test statistics. The p value distributions from the analysis of a large number of biological gene sets (using OC-CE-CTX) indicate that a null model that accounts for these three factors is effective (i.e. the distribution is quite uniform). In addition, our formulation of the null model and our method to compute the null distribution do not preclude the incorporation of additional factors (see below). In fact, in principle any combination of factors can be incorporated.

Defining the null model

The above analysis indicates that an effective null model can be defined based on *comparable* random gene sets, i.e. ones that have similar L , GC and C distributions as the given gene set (\mathbf{G}). Formally, given a statistic S (e.g. $K|N$) and a gene set \mathbf{G} , whose genes have a joint empirical (L, GC, C) distribution D (i.e. $L, GC, C|\mathbf{G} \sim D$), the goal is to obtain the distribution of $S|D$. By conditioning on D , this model formally requires that the random gene sets have similar properties as \mathbf{G} . Note how this definition allows the incorporation of additional factors by conditioning on a joint distribution. The p values of the observed statistics of \mathbf{G} can be computed from the $S|D$ distribution.

The advantage of this model is that the joint empirical (L, GC, C) distribution of \mathbf{G} is taken into account, but the computation of the null distributions can be challenging. A simpler alternative is to only condition on a summary statistic of the empirical distribution, such as the mean or median, to account for overall trends. However, this is problematic if the higher moments of the empirical distribution are also significantly

² A random variable X is conditionally independent of Y given Z if $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$. In other words, all correlation between X and Y is through Z ; once Z is fixed, X and Y are no longer correlated.

different from the genome-wide distribution. Below a novel sampling scheme is introduced to compute the null distribution of any gene-set based statistic given the (L, GC, C) distribution of \mathbf{G} .

Computing the null distributions

Given \mathbf{G} with n genes (or putative targets), a direct way to compute the null distribution is to generate random gene sets by sampling n gene from the genome according to the empirical distribution D . One approach to accomplish this is to repeatedly draw a sample from D (i.e. a (l, gc, c) triple) and pick a gene whose length, GC content, and general conservation is closest to the drawn sample. This sampling procedure requires that a parametric form be fitted to the empirical (L, GC, C) distribution; the joint density can also be obtained by techniques such as kernel-based estimation (Duda et al., 2001). We opted to pursue the latter because it is non-parametric and purely data driven, and can thus avoid potential biases introduced by parametric models; it also allows the easy incorporation of additional conditioning factors because different parametric models are likely needed for different combinations of factors.

A kernel-based estimator fits a given empirical density by a set of parameterized functions called *kernels*. The density function is the sum of kernel functions defined over the domain of the random variable(s). Formally, let $f_i(\mathbf{x}|\bar{\theta}_i)$ be the i^{th} kernel with parameter vector $\bar{\theta}_i$; the estimated density is $f(\mathbf{x}) = \sum_{i=1}^{n_k} f_i(\mathbf{x}|\bar{\theta}_i)$ where \mathbf{x} can be a vector and n_k is the total number of kernels.

A simple example of a kernel-based density estimation procedure is the construction of histograms from data (Fig. S7). The kernels in this case are constant functions in a defined interval. Each kernel is parameterized by two parameters: location and height. For instance, a one-dimensional kernel has the form:

$$f(x) = \begin{cases} h & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

, where $[a, b]$ specifies the location and h specifies the height (or probability mass) in $[a, b]$. The location of the kernels is determined by the center of each bin and the height reflects the number of data points that fall within the bin (Fig. S7). The location parameter in a multidimensional kernel specifies a hypercube. The size or volume (also called the bandwidth) of the location parameter (e.g. $|b-a|$ in the 1-d case) is a key that determines the performance of the estimator. Ideally the bandwidth should always be small if sufficient data are available; because if the bandwidth were too large each data point would exert bias on the density of the nearby points. However, in practice, data can be limiting and hence the bandwidth parameter needs to be optimized so that the maximum amount of information can be extracted from the data with minimum bias (Turlach, 1993).

A common approach is to use one kernel per data point and then infer the bandwidth parameter, either individually for each kernel or one for all kernels. Gaussian

kernels are often used because they have a tractable analytical form and nicely model the intuitive notion that the density influence of a data point should gradually diminish as one moves away from the data point (rather than abruptly going to 0 if a constant function is used). For instance, given n one-dimensional data points d_i , the estimated density is $f(x) = \frac{1}{n} \sum_{k=1}^n g(x|d_i, \sigma_i)$, where $g(\cdot | \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2 (Fig. S8). Sampling from such kernel-based densities is straightforward: one can randomly pick one of the kernels and sample according to the kernel density.

Gene-neighborhood sampling

Multidimensional Gaussian kernels (i.e. in L - GC - C space), one per gene in the input gene set \mathbf{G} , can be used to obtain the empirical (L, GC, C) distribution of \mathbf{G} . The following algorithm can be used to generate a random gene set:

- For each gene g in \mathbf{G} ,
1. Sample a (l, gc, c) triple from the Gaussian kernel of g
 2. Find the gene in the genome whose 3' UTR length, GC content, and general conservation is the closest to (l, gc, c) .

To evaluate “closeness” in the second step, a distance metric is needed in the L - GC - C space. The Euclidean distance can be used after normalizing each dimension by their mean and standard deviation³ to ensure that the variables with larger absolute magnitudes do *not* dominate the distance measure (e.g. 3' UTR length).

A verbatim implementation of this algorithm can be inefficient because locating the closest gene for any given (l, gc, c) takes time proportional to the number of genes in the genome. However, note that for each g , the above algorithm is equivalent to sampling from genes that are close to g in the L - GC - C space (i.e. the neighbors of g), so by indexing the neighbors using their normalized Euclidean distance to g , the look-up step for the closest gene can be made more efficient:

1. For every gene in the genome, sort all genes in the genome in the order of normalized Euclidean distance to g ; index them by the distance.
2. For each gene g in \mathbf{G}
 - a. sample a (l, gc, c) triple from the Gaussian kernel of g
 - b. determine the distance d between (l, gc, c) and g
 - c. use d to look up the index to obtain the closest gene

³ For the (l, gc, c) triple associated with each gene, the normalized length, gc-content, and general conservation level is $(\frac{l - \langle l \rangle}{\sigma_l}, \frac{gc - \langle gc \rangle}{\sigma_{gc}}, \frac{c - \langle c \rangle}{\sigma_c})$, where $\langle \cdot \rangle$ and σ are the mean and standard deviation of the respective variables.

Note that in this algorithm the sampling from L-GC-C space essentially reduces down to sampling from the distance space, i.e. each (l, gc, c) triple sampled was converted to d , which is the critical parameter for locating which gene to pick. Hence a one-dimensional kernel in distance space can be defined for each gene in \mathbf{G} to replace the three-dimensional L-GC-C kernel. The distance-space sampling can be further simplified to distance-rank-space sampling:

1. For every gene u in the genome, assign ranks to all genes in the genome based on their normalized Euclidean distance to u (e.g. the closest gene has rank 1, next has rank 2, and so on).
2. For each gene g in \mathbf{G}
 - a. sample a rank from the Gaussian kernel of g (draw a sample from the Gaussian, take the absolute value and round to the nearest integer).
 - b. return the gene with the sampled rank

Note that the rank is gene-dependent and can correspond to different actual distance units across genes. A rank-based kernel, such as the one used above, is desirable if one wants to ensure that every gene has an equal-size sampling neighborhood (i.e. with the same number of genes). This makes intuitive sense in that if a gene in G resides in a sparse neighborhood in the L-GC-C space, its effect on the mass of the estimated density in L-GC-C space around the neighborhood should be broader. This is equivalent to scaling the kernel bandwidth in distance space by the gene density around the gene (i.e. genes with rare L-GC-C attributes have a kernel with larger bandwidth).

The parameter remaining to be specified is the bandwidth of the kernels (i.e. the σ of Gaussians). If σ is too large, the L-GC-C distribution of the random gene sets would be significantly different from \mathbf{G} ; whereas a small σ can lead to bias as illustrated in Fig. S8. In practice σ is largely a function of the size of \mathbf{G} . To determine a reasonable σ , we use the algorithm above to draw random gene sets using different σ and compare the L, GC and C distributions of each random set to the respective L, GC and C distributions of \mathbf{G} . For each σ , a large number (>100) of random gene sets are used so that an average deviation based on the Kolmogorov-Smirnov Test can be computed. The largest σ that does *not* result in an average deviation greater than a pre-specified threshold⁴ from the L-GC-C distributions of \mathbf{G} can be used as a good bandwidth estimate. An example can be found in Fig. S9.

Compiling high-quality putative target sets

To compile high-quality putative target (HPT) sets for co-targeting analysis (and also for examining HPTs within gene sets), we aim to include Targetscan predictions that either

⁴ Currently set to 0.67, which was determined based on a simulation experiment

have at least one perfectly conserved seed match and/or predictions with at least one seed-matched site that has a high context score. To infer a good context score cutoff, we examined the context score distributions of conserved and non-conserved seed-matched sites (Fig. S10). Below (above) a context score of ~68, non-conserved (conserved) sites are enriched. This suggests that a context score of 68 is a good cutoff to use for inferring high quality non-conserved sites if we make the plausible assumption that conserved sites are enriched with true positives. Thus we defined high-quality targets as ones having at least one conserved seed-matched site and/or ones having at least one seed-matched site with a context score greater than 68.

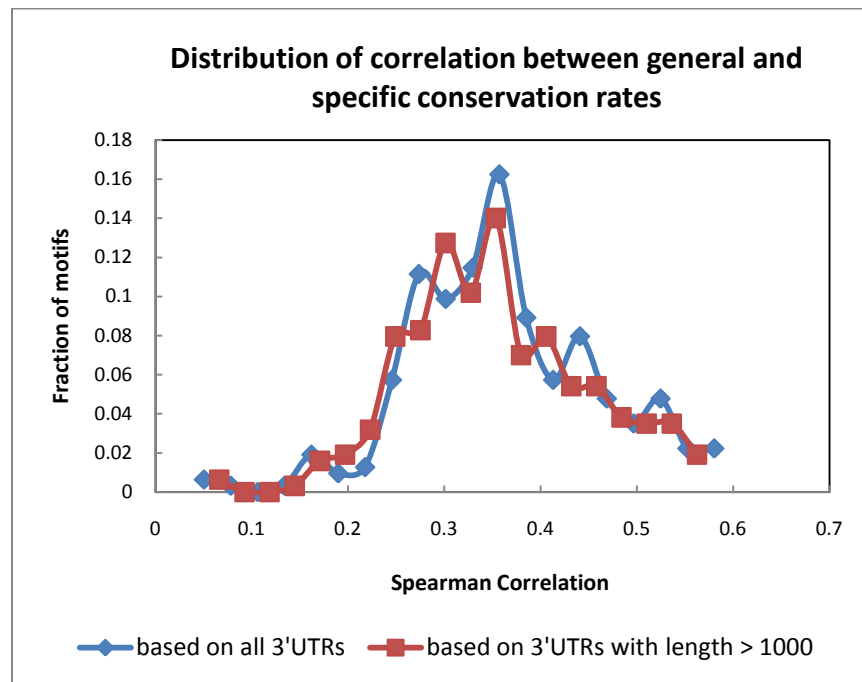


Figure S3 General conservation level is predictive of the conservation level of individual motifs. The distribution of correlations between general and specific conservation rates across 314 seed-matched motifs (i.e. one correlation value for each motif) is shown. The specific conservation rate was computed based on individual motifs whereas the general conservation rate was computed across all 7-mers. All but 5 of the correlations have P values less than or equal to 0.01. The results are similar if only 3' UTRs that are at least 1000 nt long were used.

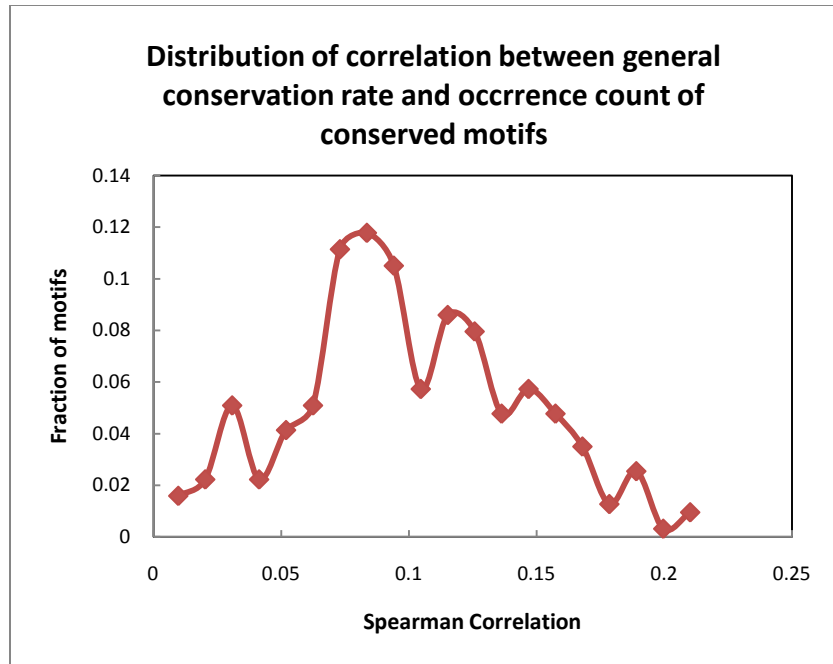


Figure S4 Similar to Fig. S3, but the correlation was computed based on general conservation rate and the occurrence count of individual motifs. All but 5 of the correlations have P values less than or equal to 0.01.

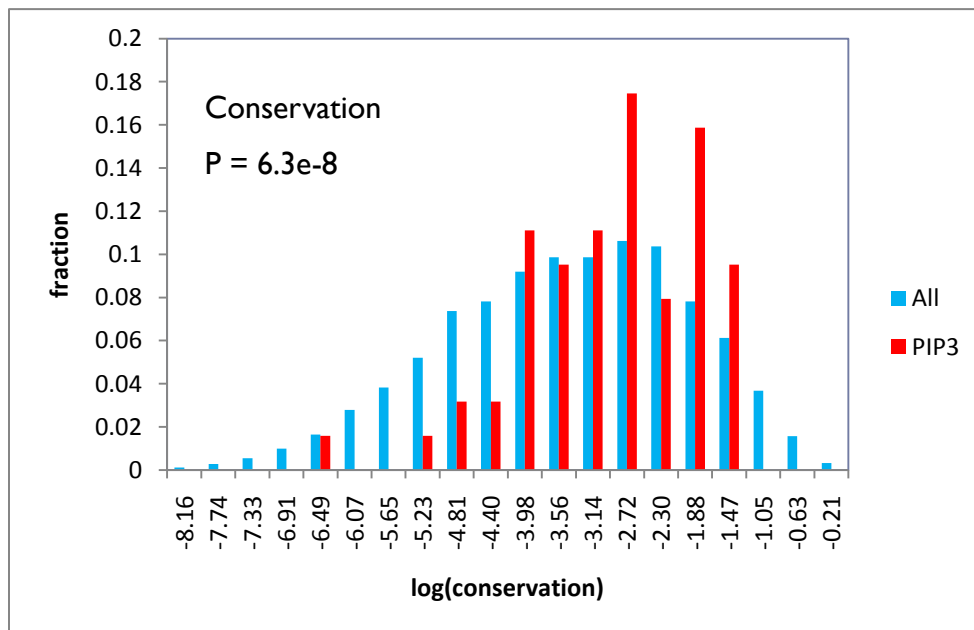


Figure S5 The general conservation rate distribution of genes in the PIP3 gene set vs. that of all genes in the genome. PIP3 genes in general have higher background conservation levels. The two distributions are significantly different (Kolmogorov-Smirnov Test; the *p* value is as shown).

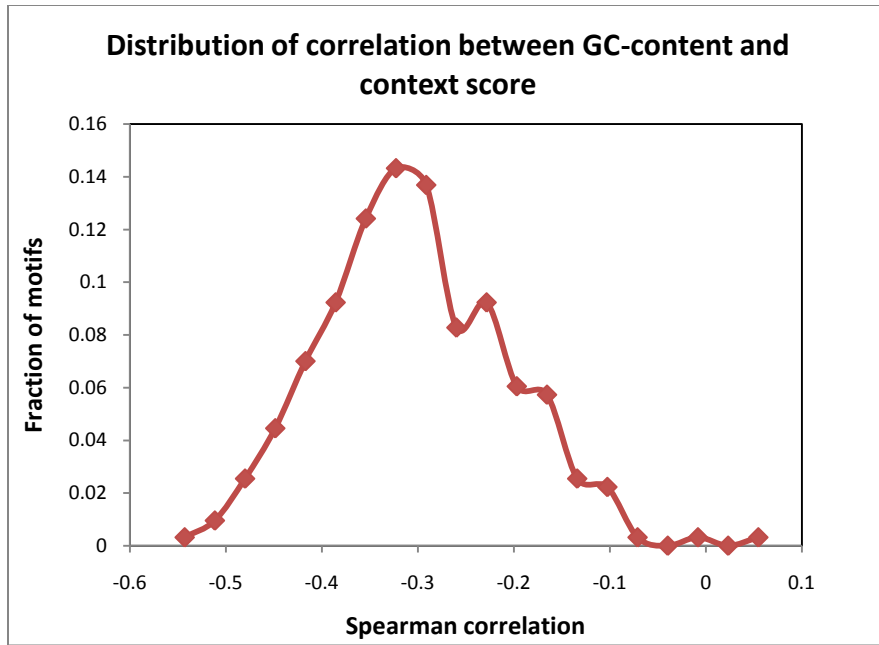


Figure S6 GC content of a 3' UTR is negatively correlated with the context score.

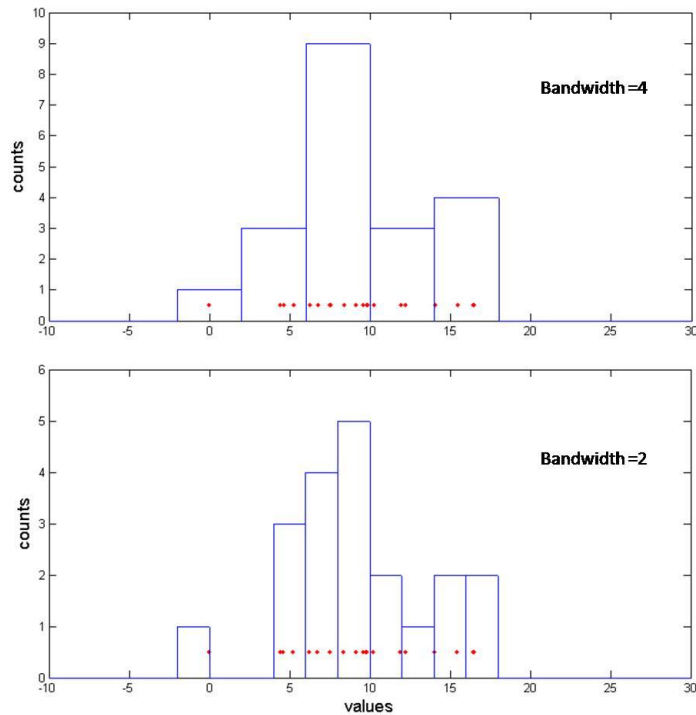


Figure S7 Histograms are examples of kernel-based density estimators. The kernels are constant functions with a fixed value within a defined neighborhood and zero everywhere else. The red dots are samples, which were drawn from a normal distribution with mean=10 and standard deviation=5. The top example uses kernel functions of width=4. The histogram was constructed by sliding a window of size 4 starting from -10 and

counts the number of samples that fall within the window. The bottom example estimates the density by using kernels of width=2.

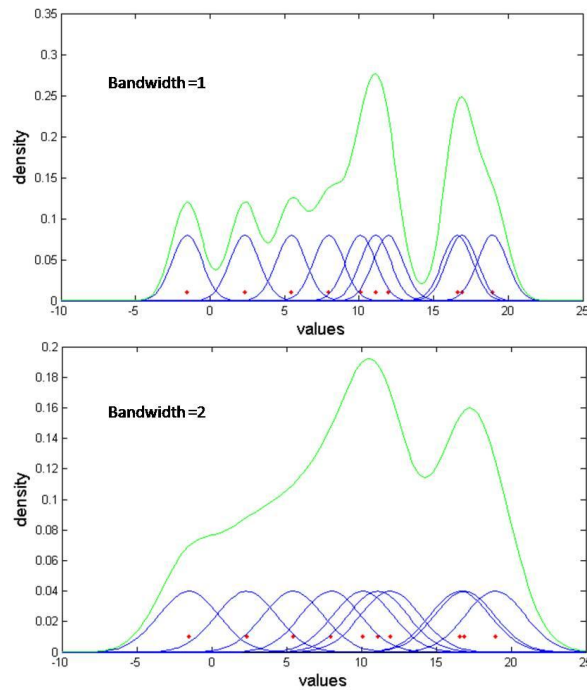


Figure S8 Density estimation by using Gaussian kernels. The red dots are samples, which were drawn from a normal distribution with mean=10 and standard deviation=5. The estimated density is the sum of normal densities with means set to the values of individual samples; the standard deviation is specified by the bandwidth parameter. The blue densities are the individual kernels and the green density is the sum. Note when the number of samples and the bandwidth are both small, there are lots of local bumps in the resulting density (top plot). A larger bandwidth avoids such biases and results in a smoother estimate (bottom plot).

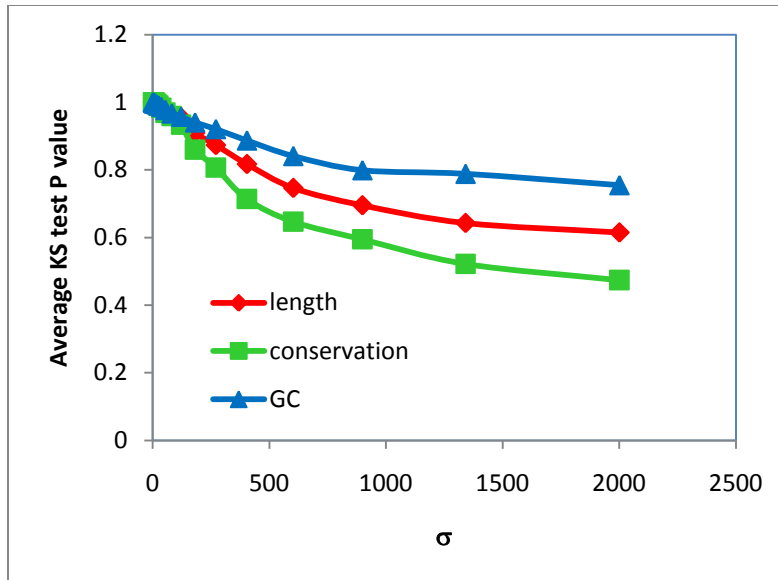


Figure S9 The input gene set is the PIP3 signaling pathway in cardiac myocytes. For each bandwidth parameter σ , 100 random gene sets were generated using the algorithm described in the text. The length, general conservation rate, and GC content distributions of each of the random gene sets were compared to those of the input gene set by the Kolmogorov-Smirnov (KS) test. The average KS test p value across the 100 random gene sets is plotted. Note that as expected, the higher the bandwidth, the lower the p value. [mirBridge](#) uses the largest bandwidth so that the lowest of the three average p values is higher than a predetermined threshold.

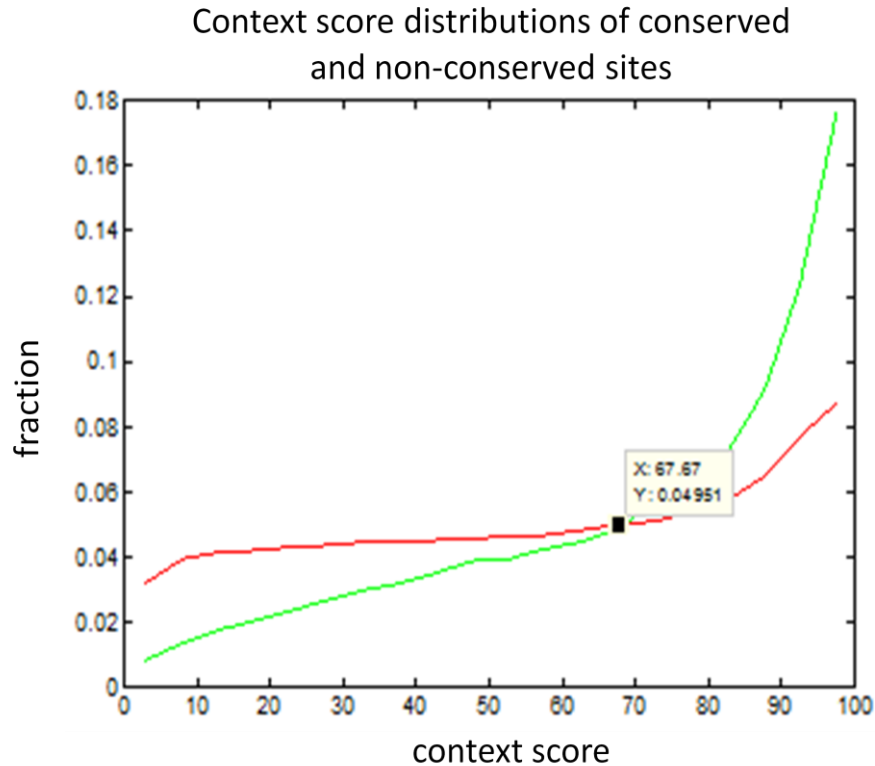


Figure S10 Context score distributions of conserved (green) and non-conserved (red) sites.

The connection between CE and prior tests that use evolutionary conservation

The CE test is fundamentally different from a couple of seemingly similar tests (Lewis et al., 2005; Stark et al., 2005): CE evaluates the degree of gene set-specific conservation of the miRNA target sequence above that of the *same* sequence in comparable random gene sets, whereas the earlier tests evaluate whether the conservation level of the target sequence is significantly above that of *random* sequences in the same gene set. miRNA target sequences are typically significantly more conserved than random sequences across all genes and gene categories (Stark et al., 2005; Xie et al., 2005). Thus, merely having higher conservation than random motifs in the same gene set may not be sufficiently specific to establish functional linkage between a miRNA and a gene set; the type of conservation enrichment detected by the CE test is more appropriate.

Sensitivity and specificity of the OC-CE-CTX test: Alternative test scores and comparisons

Other combinations of the three basic tests (CE, CTX and OC) are possible. For instance, by combining the CE and CTX tests one can form the “CE-CTX” score, which can lead to miRNA-gene set predictions solely from known functional targeting signals (i.e.

conservation and favorable 3' UTR sequence context). Comparing the performance of different tests is difficult because true positives (i.e. known miRNA functions), especially in the context of pathways, are lacking. Below we discuss several analyses that suggest OC-CE-CTX has the best sensitivity and specificity among tests that use the three basic scores. Specifically, we will compare two basic tests (CE and CTX) and the CE-CTX composite test to the OC-CE-CTX test using the pathway and module gene sets. While other tests are possible, e.g., OC-CTX and OC-CE, their utility is clearly bested by the OC-CE-CTX test (and the OC test alone is insufficient to suggest functional targeting as discussed in the main text).

At a global FDR cutoff of 0.2 (across gene-set and seed-motif combinations), the CE, CTX, CE-CTX, and OC-CE-CTX tests predict 7, 1, 37 and 215 miRNA-gene-set associations, respectively, for the pathway gene sets; and 4, 2, 23, and 186 respective predictions for the module gene sets. The CE and CTX predictions are all in the CE-CTX and OC-CE-CTX lists, indicating that, as expected, the composite tests are more sensitive. Below we focus on comparing the CE-CTX and OC-CE-CTX pathway prediction results.

The CE-CTX pathway predictions are largely in the OC-CE-CTX set, except four pathways with higher (close to 0.2) CE-CTX q values (in the case of modules, only one prediction is in CE-CTX exclusively; we only focus on the pathway results in the discussion below as the module results share the same trend). However, the relative ranking of some individual predictions (based on the q values) are different across the OC-CE-CTX and CE-CTX lists. For example, predictions ranked near the top of the CE-CTX list but having a low OC score are ranked lower in the OC-CE-CTX predicted list. The *miR-1*-PIP3 association is such an example, where it has a higher rank (9/37 versus 72/215) and a more significant q value (0.065 versus 0.098) in the CE-CTX list because the number of putative *miR-1* binding sites is not unusually high ($p=0.38$) in the PIP3 gene set (even though the proportion of conserved and high-context-scoring sites are unusually high—the basis of significant CE and CTX scores). The fact that the OC-CE-CTX test only excludes a few CE-CTX predictions with higher q values is encouraging as this suggests that OC-CE-CTX achieves higher sensitivity (i.e., significantly larger number of predictions) without sacrificing specificity (that is, OC-CE-CTX selectively excludes only the less-confident predictions in the CE-CTX list; see below).

To infer whether the additional predictions made by OC-CE-CTX are enriched for true positives, we compare the CE-CTX p -value distribution of miRNA-pathway pairs that are exclusively predicted by OC-CE-CTX to that of miRNA-pathway pairs *not* predicted by OC-CE-CTX. If the use of OC signals by OC-CE-CTX largely results in false positives, we expect the two distributions to be statistically indistinguishable (they would also have comparable median p values). However, the two distributions are drastically different ($p < 2.3 \times 10^{-155}$, Kolmogorov-Smirnov Test) and the median p values are 0.009 and 0.5 respectively (their difference is highly significant: $p < 4.4 \times 10^{-132}$, Mann-Whitney Test). Reassuringly, the latter distribution is essentially uniform,

as is expected for p values randomly drawn from the null. Furthermore, if we compute the CE-CTX q values by using only those miRNA-pathway pairs predicted by OC-CE-CTX exclusively, all CE-CTX pairs would have a q value smaller than 0.2. This suggests that these pairs had insignificant CE-CTX q values (>0.2) only because the CE-CTX test has insufficient statistical power when many miRNAs and gene sets are tested simultaneously.

In stark contrast to the analysis of pathway and module gene sets, CE alone gives a much larger number of miRNA-miRNA co-targeting predictions at a FDR cutoff of 0.2 than both CE-CTX and OC-CE-CTX (3053, 85, and 221 distinct miRNA-family pairs predicted by CE, CE-CTX, and OC-CE-CTX tests, respectively). A majority ($>75\%$) of the CE predictions that overlap with those of OC-CE-CTX have small CE q values (<0.1), while more than 90% of non-overlapping pairs have CE q values larger than 0.1. This strongly suggests that a large percentage of the non-overlapping predictions are false positives, where OC-CE-CTX excludes them because they are not simultaneously supported by other tests (CTX and/or OC). This apparent lack of specificity of CE compared to the composite tests indicates that the non-specific conservation biases in these predicted target sets are extremely strong; only by combining multiple tests that use different aspects of functional targeting can we enrich for true positives. Our method for correcting for non-specific conservation bias has already helped significantly as CE gives a significantly smaller number of predictions than gene-set overlap analysis using Fisher's Exact Test at the same FDR cutoff (see main text). Similar to the results in pathway analysis, CE-CTX and OC-CE-CTX results are largely overlapping (70 out of 85 CE-CTX predictions are in the OC-CE-CTX list). Taken together, our analyses strongly suggest that the OC-CE-CTX test has significantly better sensitivity and specificity than other tests.

Gene sets used for *miR-218* analysis

Glutamate set

Slc1A1
Slc1A2
Slc1A3
Slc1A6
Slc1A7
Slc17A6
Slc17A7
Slc17A8
Grm1
Grm2
Grm3
Grm4
Grm5

Grm6
Grm7
Grm8
Grik1
Grik2
Grik3
Grik4
Gria1
Gria2
Gria3
Gria4
Grin1
Grin2A
Grin2B
Grin2C
Grin2D
Grin3A
Grin3B
GrinA
GrinL1A
Grid1
Grid2
Homer1
Homer2
Homer3
GLS
GAD
GLUL

GABA set

SLC6A1
SLC6A11
SLC6A13
SLC32A1
GABRA1
GABRA2
GABRA3
GABRA4
GABRA5
GABRA6
GABRB1
GABRB2
GABRB3

GABRD
GABRE
GABRG1
GABRG2
GABRG3
GABRP
GABRQ
GABRR1
GABRR2
GAD
ABAT
ALDH5A1

Dopamine set

DRD2
DRD3
DRD4
DRD5
DBH
DDC
COMT
MAOA
SLC6A3
TYR
TH
PAH
SLC29A4

Serotonin

HTR1A
HTR1B
HTR1D
HTR1E
HTR1F
HTR2A
HTR2C
HTR3A
HTR4
HTR6
HTOR
SLC6A4
HTR5A
5HTT

HTR7
HTR2B
HTR3B
HTR5A
HTR3E
HTR3D
HTR5B
TPH
MAOA
SLC29A4

Adrenaline epinephrine set

ADRA1A
ADRA1B
ADRA1D
ADRA2A
ADRA2B
ADRA2C
ADRB1
ADRB2
ADRB3
ADRBK1
ADRBK2
COMT
PNMT
TH
DBH

Synaptic vesicle formation set

BSN
RAPGEF4
RIMS1
RIMS2
PCLO
UNC13A
ERC2
SV2A
SV2B
NAPA
STXBP1
SYT1
CLPX1
CLPX2
NSF

Supplemental References

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*, 2nd edn (New York: Wiley).

Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: A review, Paper presented at: Discussion Paper 9317 (Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium: Institut de Statistique).

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338-345.