Supplementary information for: Human microRNAs co-silence in well-separated groups and have different essentialities

Gábor Boross [1,2], Katalin Orosz [1,2] and Illés J. Farkas [2,*]

[1] *Department of Biological Physics, Eötvös Loránd University,*

[2] *Statistical and Biological Physics Research Group and CellCom RET at the Hung. Acad. of Sci., Pázmány P. stny. 1A, H-1117 Budapest, Hungary*

[*] *To whom correspondence should be addressed:* fij@elte.hu

Abbreviations: miRNA, microRNA.

1. Comparing miRNA co-regulation scores and modules computed from TargetScan and PicTar data

Figure 1. of the main text shows that the relative overlap between the full miRNA-target gene lists of PicTar and TargetScan is only slightly above 10% (the maximum of the solid curve comparing the two databases). Such a low overlap between the two interaction lists cannot explain the observed good agreement between the strongest miRNA co-regulation scores and the lists of most and least essential miRNAs (see Fig. 4 of the main text). In Suppl. Fig. 1 we compare the miRNA co-regulation scores and miRNA co-regulation modules computed from TargetScan data to those computed from PicTar data. Fig. 4 and the high similarity of the modules obtained from the two data sources show that co-regulating modules of miRNAs efficiently extract high-quality information from the much less similar miRNA-target interaction lists. The distribution of co-regulation scores is clearly bimodal in both cases (panels a and c): a small group of high scores is separated from a large group of low scores. This separation is very strong in TargetScan, where there are almost no co-regulation scores between 0.5 and 0.95, and somewhat less pronounced in PicTar. When using PicTar data the presence of a small number of intermediate co-regulation scores (between 0.5 and 0.95) may explain the resulting lower number of co-regulation modules and also why several modules identified with TargetScan are merged into a single module (see on the left of panel d). In both cases co-regulating modules of miRNAs are well-separated: despite explicitly allowing overlaps, i.e., shared miRNAs, between them, we found only very few overlaps.

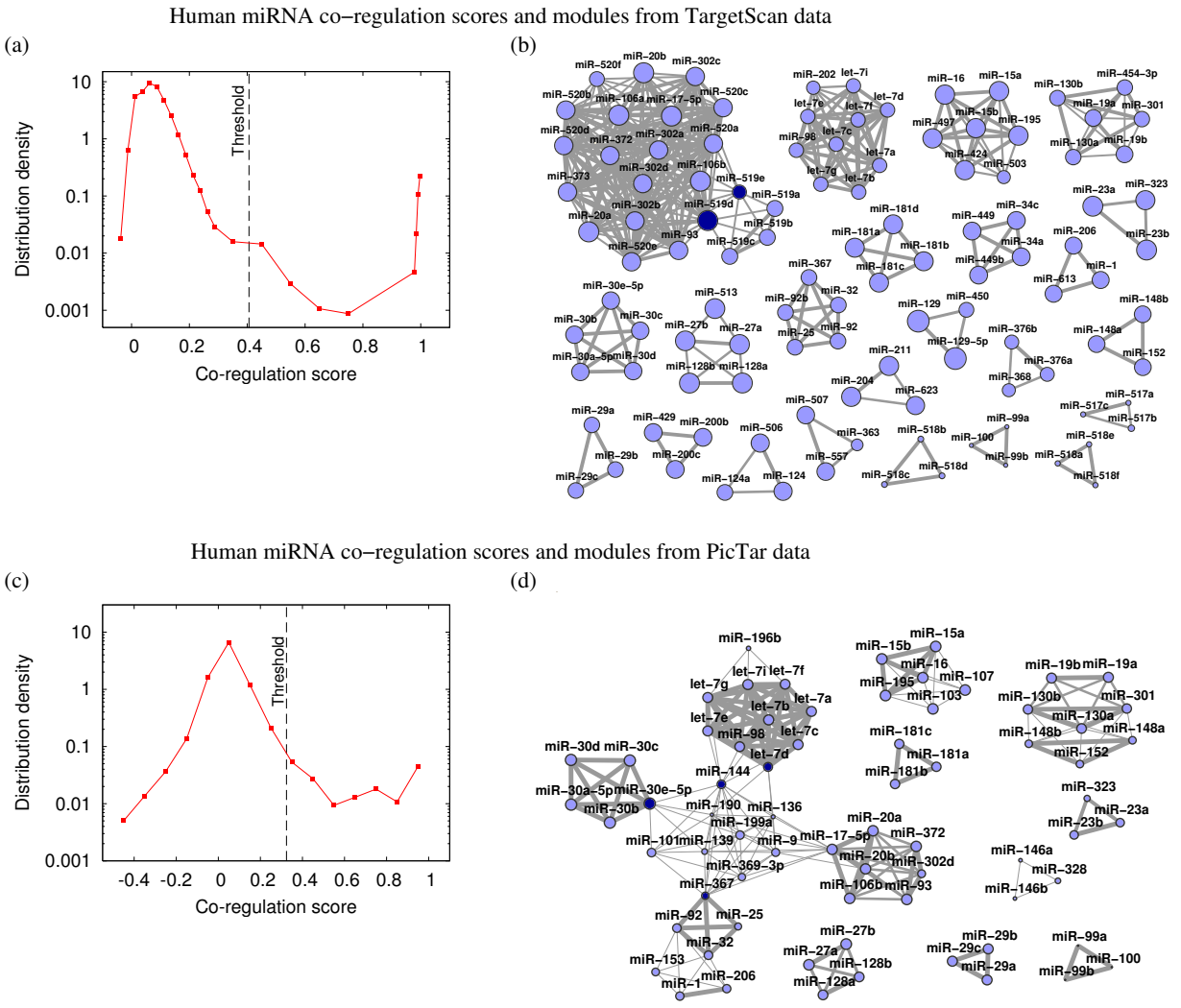
Almost all modules identified with PicTar were identified with TargetScan, too, with the same or highly similar participating miRNAs. Significant differences between the modules can be found only in the largest module. One reason for this can be that PicTar is less stringent with the prediction of low scoring miRNA - target pairs than TargetScan. Consequently, some of the co-regulation modules isolated entirely in the case of TargetScan may gain co-regulation links of intermediate strength (below 0.95, but above the threshold, W) when PicTar data are used.

2. Comparing membership in miRNA co-regulation modules computed from miRBase and PITA data to those computed from TargetScan data

Comparing miRBase with TargetScan

With TargetScan data $n_1 = 111$ of the total $N_1 = 455$ miRNAs (nodes) are in the co-regulation modules, while miRBase lists $N_2 = 711$ nodes out of which $n_2 = 125$ are in modules. The intersection of the two full sets contains $M = 69$ miRNAs and the intersection of the two module node lists contains $m = 50$ miRNAs. To estimate the probability, P, of this event, we assume statistical independence in the selections and compare the number of ways for at least m module member nodes to be shared to the number of possibilities for a randomized control case. One

can pick the $s = m \ldots M$ nodes that are module members in both miRBase and TargetScan from the M shared nodes of miRBase and TargetScan in $\binom{M}{s}$ ways. The remaining $n_1 - s$ module member nodes of TargetScan that are not module members in miRBase can be selected from the remaining $N_1 - M$ nodes of TargetScan in $\binom{N_1-M}{n_1-s}$ ways. The $n_2 - M$ nodes that are module members with miRBase, but are not with TargetScan, can be picked from the remaining $N_2 - M$ nodes of miRBase in $\binom{N_2-M}{n_2-s}$ ways. Thus, the number of cases when miRBase and TargetScan share at least m module members is $\sum_{s=m}^{M} \binom{M}{s}\binom{N_1-M}{n_1-s}\binom{N_2-M}{n_2-s}$.



Supplementary Figure 1. Human miRNA - miRNA co-regulation scores and co-regulating miRNA modules computed from TargetScan and PicTar data. Subfigures **(a)** and **(b)** are reproduced from Fig. 2 of the main text. **(a, c)** The distribution of miRNA co-regulation scores is bimodal in both cases. Note that the vertical scale is logarithmic. A small group of strong (>0.95) co-regulation scores is clearly separated from the dominant group of weak co-regulation scores below the threshold, W. The optimal co-regulation score thresholds, $W = 0.406$ for TargetScan and $W = 0.324$ for PicTar, are marked in both cases. Lines connecting the data points are guides to the eye. **(b, d)** Co-regulating groups (modules) of miRNAs computed from TargetScan and PicTar data are highly similar. Modules with similar participating miRNAs are at the same positions on the two subfigures.

For the control case, we drop the condition that there should be miRNAs that are inside modules with both

miRBase and TargetScan. This means that simultaneously n_1 nodes need to be selected from N_1 and n_2 nodes from N_2. The total number of ways to do this is $\binom{N_1}{n_1}\binom{N_2}{n_2}$. In summary, we estimate that the probability for the modules computed with miRBase and TargetScan to share at least the observed number of nodes is

$$P = \frac{\sum_{s=m}^{M} \binom{M}{s}\binom{N_1-M}{n_1-s}\binom{N_2-M}{n_2-s}}{\binom{N_1}{n_1}\binom{N_2}{n_2}}.$$

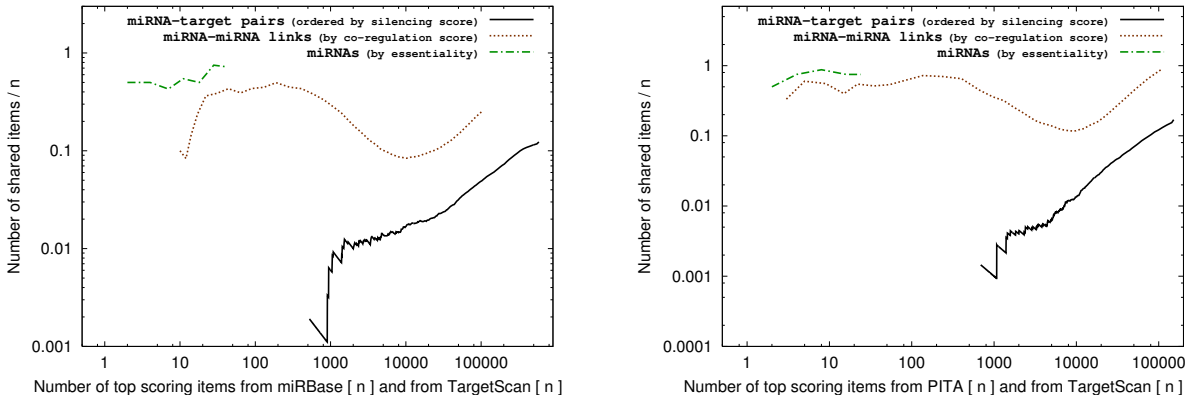
We apply the logarithms of the factorials and compute that in the randomized case the probability for at least $m = 50$ nodes (the same m nodes) to be contained by the co-regulation modules with both miRBase and TargetScan as a primary data source is $P = 1.54 \times 10^{-63}$.

Comparing PITA with TargetScan

PITA lists $N_2 = 470$ miRNAs (nodes) out of which $n_2 = 85$ are in modules. The intersection of the two full sets (PITA and TargetScan) contains $M = 70$ miRNAs and the intersection of the two module node lists contains $m = 55$ miRNAs. Similarly to the previous section, we estimate that in the randomized control case the probability for at least $m = 55$ nodes (the same nodes) to be contained by the co-regulation modules with both PITA and TargetScan is $P = 3.40 \times 10^{-76}$.

3. Similarity of miRNA co-regulation scores and predicted miRNA essentialities: Comparing miRBase with TargetScan and PITA with TargetScan

Similarly to Fig. 5 of the main text Suppl. Fig. 2 compares the miRNA co-regulation scores and essentiality levels computed from different data sources. We find again that co-regulation, scores, modules of miRNAs and miRNA essentiality levels efficiently extract high-quality information from the lists of miRNA-target pairs, which have, as of yet, a lower confidence.



Supplementary Figure 2. **Left panel.** Comparing miRBase and TargetScan through miRNA-target gene silencing scores (solid line, this curve is repeated from Fig. 2 of the main text), miRNA - miRNA co-regulation scores (dotted line) and miRNA essentialities (dash-dot). When replacing TargetScan as a data source by miRBase the list of top scoring miRNA co-regulation links and the list of the most essential miRNAs are clearly much more stable than the list of strongest (most efficiently silencing) miRNA-target gene pairs. **Right panel.** The same comparison for TargetScan and PITA.

4. Essentiality of each miRNA computed with TargetScan and PicTar data

In the table below we list the essentiality score and the rank of that score for each miRNA with TargetScan and PicTar data. A dash indicates that the essentiality score is not available. The names of miRNAs that are among the 10 most essential in both cases are printed in boldface with red color. Similarly, the names of miRNAs that are among the 10 least essential in both cases are printed in italics with blue color.

miRNA	Predicted essentiality	
	with TargetScan (rank) score	or PicTar data (rank) score
hsa-miR-130a	(1) 1.080	(1) 1.083
hsa-miR-195	(2) 1.040	(2) 1.030
hsa-miR-196b	-	(3) 1.029
hsa-miR-144	-	(4) 1.013
hsa-miR-15b	(3) 0.993	(12) 0.911
hsa-miR-30d	(4) 0.970	(5) 1.001
hsa-miR-136	-	(6) 0.976
hsa-miR-30a-5p	(6) 0.944	(8) 0.946
hsa-let-7g	(7) 0.938	(7) 0.967
hsa-miR-9	-	(9) 0.931
hsa-miR-103	-	(11) 0.921
hsa-miR-30e-5p	(5) 0.960	(13) 0.897
hsa-miR-30c	(8) 0.917	(10) 0.922
hsa-miR-25	(9) 0.891	(15) 0.886
hsa-miR-32	(10) 0.884	(16) 0.876
hsa-let-7c	(11) 0.856	(14) 0.889
hsa-miR-99b	(12) 0.849	(17) 0.856
hsa-miR-27b	(13) 0.828	(22) 0.833
hsa-let-7i	(14) 0.825	(18) 0.855
hsa-let-7f	(15) 0.825	(19) 0.852
hsa-miR-101	-	(24) 0.823
hsa-miR-99a	(16) 0.820	(26) 0.820
hsa-miR-100	(17) 0.814	(23) 0.825
hsa-miR-27a	(18) 0.812	(27) 0.816
hsa-let-7e	(19) 0.808	(20) 0.842
hsa-miR-15a	(20) 0.797	(25) 0.821
hsa-miR-23a	(21) 0.786	(30) 0.791
hsa-let-7d	(22) 0.785	(21) 0.841
hsa-miR-23b	(23) 0.780	(32) 0.784
hsa-let-7b	(24) 0.776	(29) 0.815
hsa-miR-16	(25) 0.757	(31) 0.785
hsa-miR-93	(26) 0.751	(33) 0.744
hsa-miR-106b	(27) 0.748	(34) 0.743
hsa-miR-19a	(28) 0.734	(35) 0.733
hsa-miR-19b	(29) 0.729	(36) 0.728
hsa-let-7a	(30) 0.716	(37) 0.721
hsa-miR-20a	(31) 0.683	(38) 0.678
hsa-miR-29a	(32) 0.662	(39) 0.670
hsa-miR-17-5p	(33) 0.656	(28) 0.815
hsa-miR-29c	(34) 0.649	(40) 0.654
hsa-miR-181b	(35) 0.618	(42) 0.617
hsa-miR-29b	(36) 0.615	(41) 0.618
hsa-miR-181a	(37) 0.603	(43) 0.599

5. miRNA - target interactions

We first compared human miRNA - human target interactions from one manually curated and four computational data sources. Then, for our analyses we used data from TargetScan and as a control PicTar (see, e.g., Suppl. Fig. 1). The five data sources were TarBase (as provided in a filtered form under "known targets" by miRBase in June 2008) [1], miRBase (version 5) [2], PicTar (vertebrates: "conservation in mammals", Dec. 2007) [3], PITA (top: "3-15", Nov. 2007) [4] and TargetScan v4.1 (conserved and non-conserved sites) [5]. TarBase provides a manually collected list of experimentally verified interactions, while the four computational data sets (i) provide a score for each predicted miRNA - target (transcript, protein or gene) link quantifying the efficiency of silencing and (ii) apply a lower cutoff score (a threshold) below which they discard all links.

In PicTar and PITA target transcripts are identified by RefSeq mRNA IDs, miRBase contains Ensembl transcript

IDs, while TargetScan and TarBase contain gene/protein names. We mapped all target names to Ensembl gene IDs and in each of the four computationally predicted lists we calculated a single unified interaction score for each miRNA - target gene pair. Consider now one of the four computationally predicted interaction lists. A miRNA, M, may bind at multiple sites to the same transcript, T_1, with the interaction scores $w^{(1)}_{M,T_1}$, $w^{(2)}_{M,T_1}$, etc. In addition, M may silence several transcript variants (T_1, T_2, T_3, ...) produced from the same gene, G. To obtain a single unified (silencing) interaction score, $w_{M,G}$, for the interacting pair $M-G$, one needs to merge into $w_{M,G}$ all interaction scores between M and (the transcript variants produced from) G. The list of all scores between M and G may look like this: $w^{(1)}_{M,T_1}, w^{(2)}_{M,T_1}, w_{M,T_2}, w^{(1)}_{M,T_3}, w^{(2)}_{M,T_3}, w^{(3)}_{M,T_3}, \ldots$. We denote by w_i one of the interaction scores from this list. In the four computationally predicted data sets we computed the unified interaction score, $w_{M,G}$, between a miRNA and a target gene with different rules. Following closely the thermodynamical and biological concepts behind each prediction algorithm and additional advice received from the authors/maintainers of the databases we applied these rules: $w_{M,G} = -\max_i(\log w_i)$ for miRBase, $w_{M,G} = \sum_i \log w_i$ for PicTar, $w_{M,G} = -\log(\sum_i \mathrm{e}^{-w_i})$ for PITA and $w_{M,G} = \max(0, -\sum_i w_i)$ for TargetScan. We selected the sign of $w_{M,G}$ everywhere with the goal to make a larger score represent a higher silencing efficiency.

We used the following raw scores (w_i): orthology P value (P_{og}) from miRBase, "PicTar score" from PicTar, difference of free energy change ($\Delta\Delta G$) from PITA and the "context score" of both conserved and non-conserved sites from TargetScan. We mapped target transcript and protein names to gene names with our Perl scripts processing Ensembl files and with the online conversion tool Synergizer [6]. To locate the current miRNA names for the interactions listed in TarBase we used the sequence of each regulator as provided by TarBase and looked up the full name in miRBase by the sequence. We note that all five sources contain miRNA names with the suffixes *-3p* and *-5p*, but only miRBase and PITA contain *miR** names. We discuss only human miRNAs and omit the *hsa-* prefix from each name.

6. The co-regulation network of miRNAs and its modules

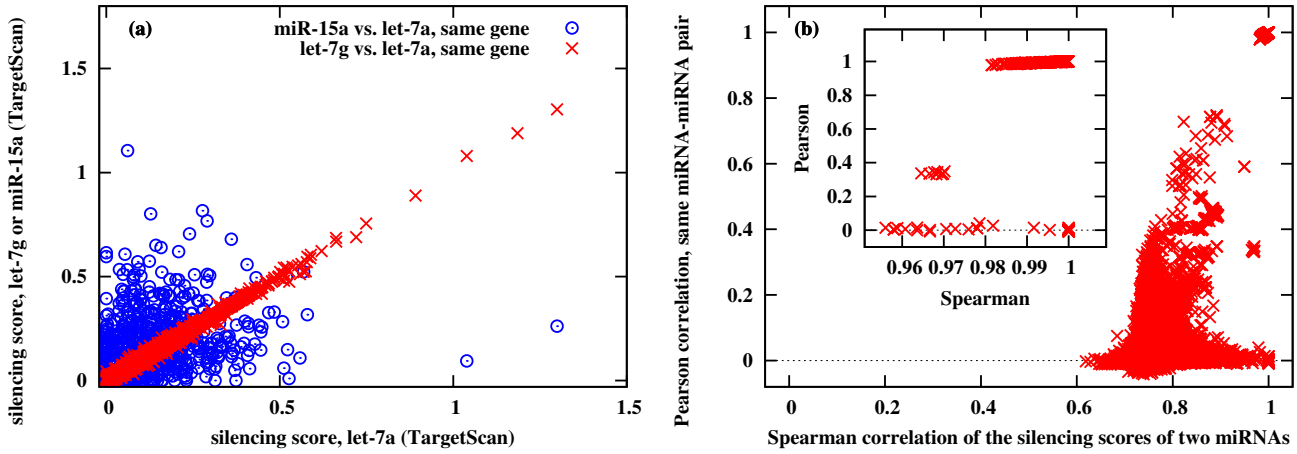
In the co-regulation network of miRNAs two miRNAs (nodes) are connected, if they share at least one target gene and the weight (score) of each link is computed as the similarity of the regulation patterns of the two miRNAs. To compute the score of each link we first listed for any miRNA, M, its unified interaction score with all genes in the genome, $\vec{v}_M = (w_{M,G_1}, w_{M,G_2}, \ldots)$, based on TargetScan data. We set the silencing score to zero for any non-interacting regulator-target pair and computed the co-regulation score (link weight) of two miRNAs (network nodes) as the correlation of their $\vec{v}_M$ vectors.

To find modules in the co-regulation network of miRNAs, we discarded co-regulation links with weights below a fixed threshold, W. With the Clique Percolation Method, [7], implemented by CFinder [8], we simultaneously selected the optimal link weight threshold, W, and computed the modules of the network. The Clique Percolation Method (CPM) finds groups of nodes in the network that are connected more densely inside the group than between groups and at the same time it selects an optimal link weight threshold, W. Note that the CPM explicitly allows for overlaps between the identified network modules. In the CPM the requested within-module link density is controlled by the clique size parameter, k. When requesting a high density of links (high k) inside modules, one will obtain a few small, but very densely internally linked modules. In this case many nodes of the network might not be contained by the modules, i.e., coverage will be low. On the other hand, requesting a low density of within-module links (low k) will lead to large and biologically meaningless modules. In other words, the precision of the modules will be low in this case.

A sign of low coverage is the absence of large and medium-sized modules, while low precision is accompanied by a few of huge and many small modules. Similarly to the clique size parameter, k, the link weight threshold, W, can adjust the sizes of the resulting modules. At a low W the modules will contain many links of the network and

have a high coverage combined with a low precision. To find the optimal (k, W) parameter pair, a tradeoff between the coverage and precision of the modules, we followed the technique from Ref. [7]. We identified the richest (most informative) module structure by the (k, W) pair for which small, medium sized and large modules are all present. From the commonly used distribution functions many, e.g., the exponential and normal distributions, decay quickly and would allow almost exclusively small modules. Thus, we selected the (k, W) pair for which the module size distribution is closest to a power-law. To achieve this first we scanned a wide range of (k, W) pairs and for each parameter set we computed the modules of the network with the CPM. Then for each (k, W) parameter pair we calculated the sizes of the modules, i.e., the numbers of nodes (miRNAs) in each module: $s_1 \geq s_2 \geq \ldots$. Finally, we selected the (k, W) pair (i) providing the largest module sizes such that (ii) s_1/s_2 just reached 2 from above indicating the transition point between high coverage and high precision. With TargetScan (PicTar) data the optimal parameter pair was $k = 3$ and $W = 0.406$ ($k = 3$ and $W = 0.324$).

7. Computing the similarities of the vectors of silencing scores with the Pearson (covariance) and Spearman (rank) correlations

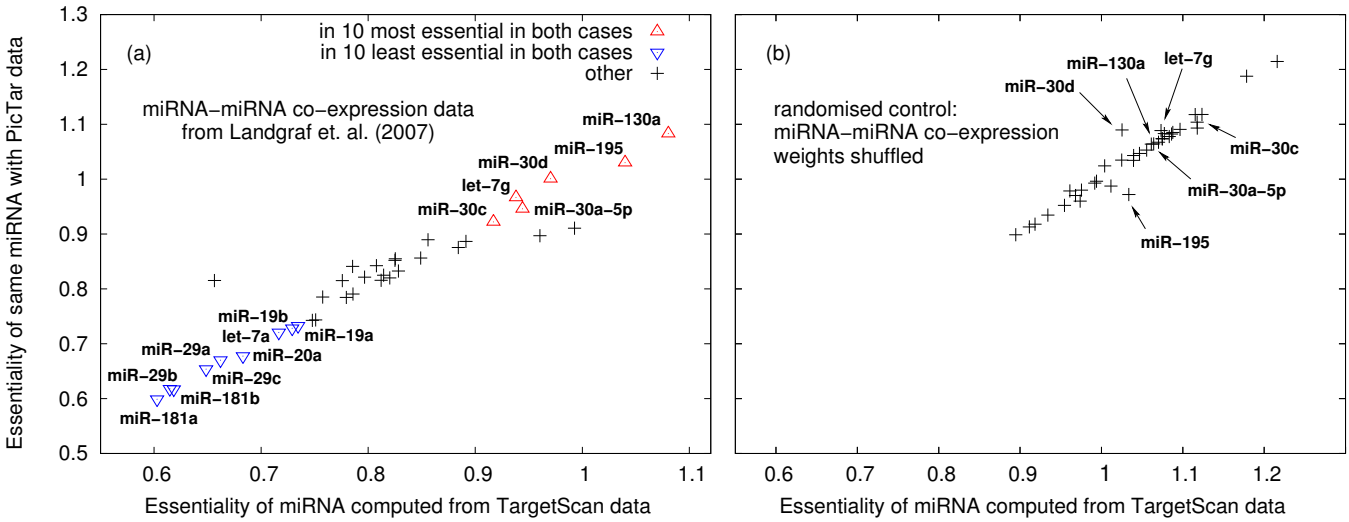


Supplementary Figure 3. Comparing how the Pearson (covariance) and Spearman (rank) correlation measures quantify the similarity between the vectors of the silencing scores of two miRNAs. **(a)** The horizontal coordinate of each point shows the silencing score between *let-7a* and a gene, while the vertical coordinate of the same point shows the silencing score between *let-7g* (red crosses) or *miR-15a* (blue circles) and the same gene. Compared to the normal distribution the "silencing vectors" of all three miRNAs, i.e., the lists of horizontal the vertical coordinates, contain outliers. Nevertheless, the Pearson correlation can still efficiently distinguish between the two types of relationships: the Pearson correlation for the pair *let-7g* - *let-7a* is 0.999, while for the pair *miR-15a* - *let-7a* it is 0.0787. The Spearman rank correlation for the same two pairs is 0.999 and 0.795, respectively. To see which of these two differences is significant, we have compared all pairs with both correlation measures. **(b)** The Spearman rank correlation and the Pearson correlation between the silencing score vectors of any two miRNAs. A high ($S > 0.98$) Spearman correlation and a high ($P > 0.8$) Pearson correlation are usually equivalent. There are 224 points, i.e., miRNA-miRNA pairs in the top right corner (high S, high P) of the figure and only 26 points, in the bottom right corner (high S, $P < 0.2$). The first group (high S, high P), where the two correlation measures agree, is clearly separated from the "cloud" of other points, thus, it is statistically significant. The second group (high S, low P), where the two correlation measures disagree, is much smaller and seems to be continuously connected to the "bulk" cloud of points, thus, it is statistically much less significant.

We have tested whether the Pearson (covariance) and Spearman (rank) correlation measures provide significantly

different results when comparing the vectors of silencing scores of miRNA-miRNA pairs. In panel (a) of Suppl. Fig. 3 we have plotted the silencing scores of selected miRNAs on the same genes, while in panel (b) we show for each miRNA-miRNA pair the Pearson vs. the Spearman correlation of their silencing scores. We find that, with a few exceptions, the most strongly correlated miRNA-miRNA pairs are identical when computed with the two correlation measures, thus, for the particular data set neither of the two measures is more appropriate than the other.

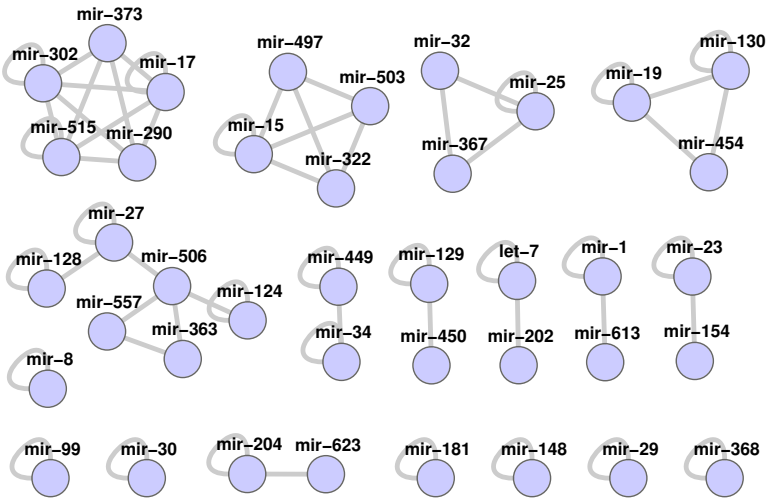
8. Predicted essentialities of miRNAs after shuffling the miRNA-miRNA co-expression weights.



Supplementary Figure 4. Changes in the predicted miRNA essentialities after the co-expression weights of miRNA-miRNA pairs (from Supplementary Table 20 of Ref. [9]) have been randomly permuted. Only module member miRNAs are shown. **(a)** This panel is a copy of Figure 4 from the main text. **(b)** The same predicted essentiality values after keeping the list of co-expressed miRNA-miRNA pairs constant and shuffling the co-expression weights. Observe that differences between the predicted essentialities of miRNAs are strongly reduced after randomisation. In particular, low predicted essentialities disappear. We have marked the originally predicted 6 most essential miRNAs after randomisation. Note that these miRNAs have lost their top ranks and are randomly mixed with the other miRNAs.

As an additional test of robustness, we have tested whether the predicted miRNA essentiality levels are preserved after the randomisation of miRNA expression data. By keeping the list of co-expressed miRNA-miRNA pairs constant, we have randomly permuted the co-expression weights among these pairs. We have used this shuffled list of miRNA-miRNA co-expression weights to produce Suppl. Fig. 4b. Panel (a) of this figure is a copy of Figure 4 from the main text and a comparison between panels (a) and (b) shows that differences between the predicted essentialities of miRNAs are strongly reduced after randomisation. In particular, low predicted essentialities disappear. As the randomisation step destroys expression correlations between co-regulating miRNAs, and thus, increases the number of miRNAs not co-expressed with their co-regulating partners, this is indeed the expected behaviour of the null control.

9. The co-regulation modules merging miRNAs from the same miRBase-defined miRNA family.



Supplementary Figure 5. Co-regulation modules of miRNA families. While in figure Fig. 2.c of the main text each miRNA is displayed separately, this figure shows the same modules by merging all miRNAs from the same family. The lists of miRNA family members were taken from miRBase.

To test how miRNA families are represented in the co-regulation modules shown in Fig. 2c of the main text, in Suppl. Fig. 5 we have merged miRNAs from the same family (as defined by miRBase). After collapsing miRNA families, we find well-separated modules, similarly to the original case. We add here two technical comments. At the time of writing the human miRNA *miR-368* contained by TargetScan is not present in miRBase, but there is a miRNA family called *mir-368*, thus, we assigned *miR-368* to the family *mir-368*. Similarly, the human miRNA *miR-623* is not assigned to a family in miRBase, therefore, we assigned it a separate family, *mir-623*.

---

[1] Sethupathy P, Corda B, Hatzigeorgiou AG (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12:192-197.

[2] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucl Acids Res*, 36:D154-D158.

[3] Lall S, *et al.* (2006) A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. *Curr Biol*, 16:460-471.

[4] Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet*, 39:1278-1284.

[5] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of mammalian microRNA targets. *Cell*, 115:787-798.

[6] Berriz GF, Roth FP (2008) The Synergizer service for translating gene, protein, and other biological identifiers. *Bioinformatics*, 24:2272-2273.

[7] Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814-818.

[8] Adamcsek B, Palla G, Farkas I J, Derényi I, Vicsek T (2006) CFinder: Locating cliques and overlapping modules in biological

networks. *Bioinformatics* 22:1021-1023.

[9] Landgraf P, *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**:1401-1414.