Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supporting Information

Cross-cancer differential co-expression network reveals microRNA-regulated oncogenic functional modules

Chen-Ching Lin, Ramkrishna Mitra, Feixiong Cheng, Zhongming Zhao

Contents

S1 Differential co-expression between mRNAs in cancers

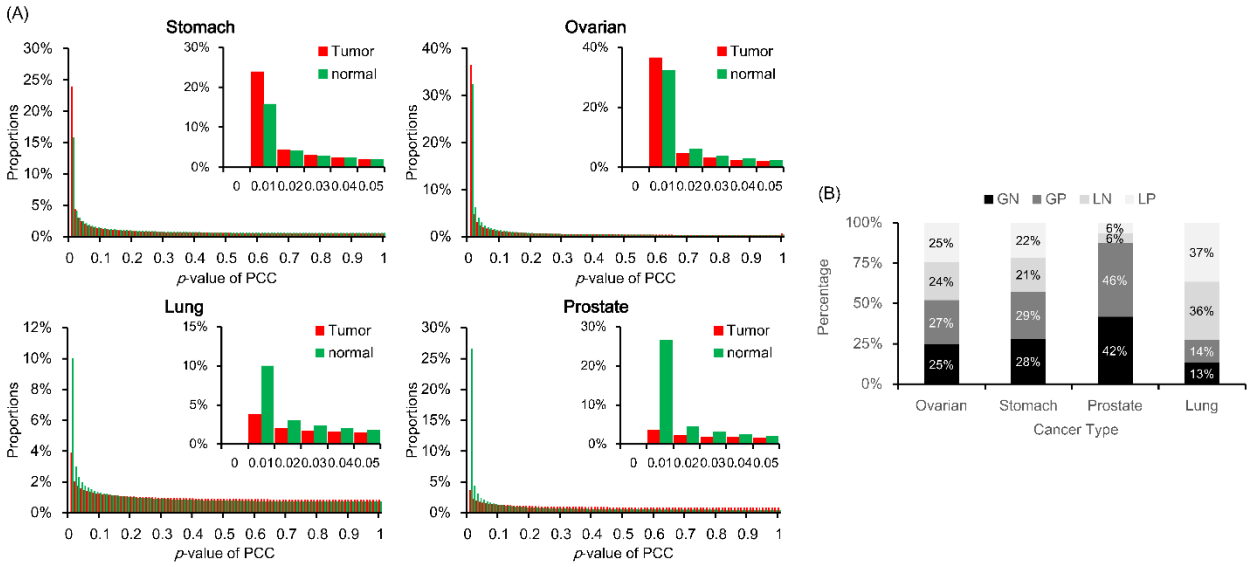


Figure S1: Differential co-expression among mRNAs.

(A) The distributions of PCC *P*-values in the four cancer and corresponding normal samples of mRNA expression profiles. (B) The distributions of mRNA differential co-expression in the four cancer types. The four differential co-expression categories are losing positive (LP), losing negative (LN), gaining positive (GP), and gaining negative (GN) co-expression, respectively.

S2 Differential co-expression between miRNAs in cancers

Table S1: The descriptive statistics of miRNA co-expression.

Total miRNAs: 850						
Sample type		Control	Cancers			
Cancer type			Lung	Ovarian	Prostate	Stomach
Expressed miRNAs		565	630	546	563	559
miRNA co-expression	Positive	37152	16234	14631	17912	7074
	Negative	35865	15000	11581	15715	6271
Differential miRNA co-expression	LP		30151	30858	30336	34119
	LN		31398	32576	30974	33944
	GP		9233	8337	11096	4041
	GN		10533	8292	10824	4350
Filtered differential miRNA co-expression ($\|\Delta Z\| \geq 1$)	LP		9014	9024	10608	11527
	LN		3025	3178	4217	4102
	GP		4171	3237	4115	2111
	GN		5143	3248	2607	1725
Ambiguous patterns	P ► N		218	168	205	121
	N ► P		110	120	163	91

Number of cancer types			1	2	3	4
Filtered differential miRNA co-expression ($\|\Delta Z\| \geq 1$)	LP		5624	6253	4633	2036
	LN		1877	2320	1759	682
	GP		11144	909	63	1
	GN		10134	961	52	5

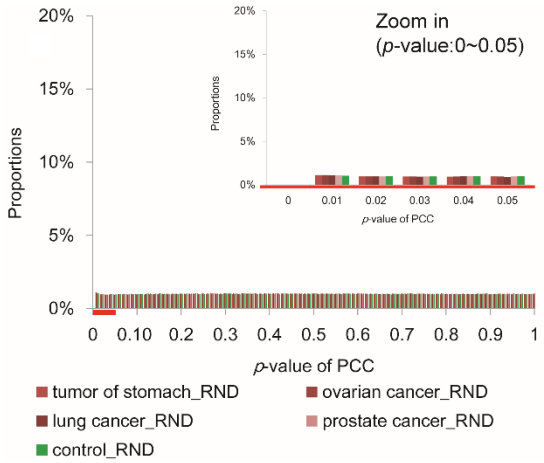


Figure S2: The distributions of PCC *P*-values of randomly permutated expression profiles

The distributions were derived from the random expression profiles. We randomly shuffled control and cancer samples 1,000 times to create the random expression profiles.

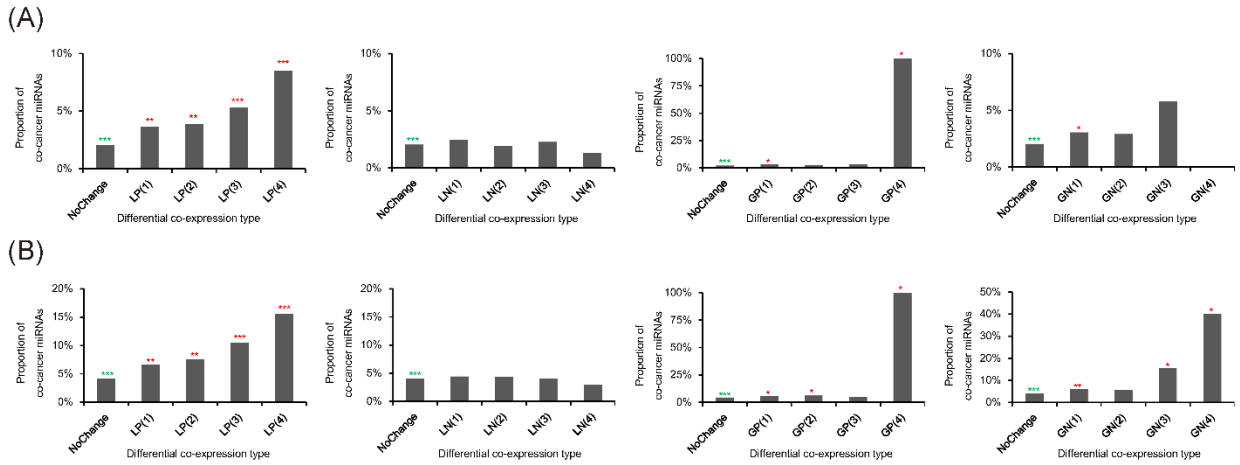


Figure S3: The co-cancer miRNAs in miRNA differential co-expression category

The co-cancer miRNA proportions of differential co-expressed miRNA pairs. The co-cancer miRNAs were defined as those paired miRNAs which were reported in at least one the same cancer type. The numbers in parentheses were the number of cancer types in which miRNA pairs losing their positive co-expression. The "NoChange" category contains non-differentially co-expressed miRNA pairs (PCC *P*-value > 0.01). The significance of each bar was tested by Fisher's exact test (***: $P \leq$ 1e-20, **: $P \leq$ 1e-10, *: $P \leq 0.05$; green asterisk: underrepresented, red asterisk: overrepresented). Both in (A) and (B), there is only one GP(4) and it is co-cancer, therefore the proportion of co-cancer GP(4) miRNAs is 100%. (A) Only the used four cancer types were considered; (B) all the cancer types in miRCnacer database were considered.

S3 Network-critical miRNAs play pivotal roles in cancer

Table S2: Enriched miRNA families in the cross-cancer differential co-expression network.

miRNA family	#members in network	#total members	*P*-value
mir-8	8	9	2.20E-03
mir-15	9	9	1.35E-04
mir-17	15	16	3.51E-06
mir-33	4	4	1.93E-02
mir-146	4	4	1.93E-02
mir-515	32	58	3.59E-03

Table S3: Network property comparisons between cancer miRNAs and non-cancer miRNAs.

miRNAs	Degree	CLV
Cancer miRNAs	11.11	2.42
Others	6.34	1.50
Wilcoxon *P*	< 0.001	< 0.001

This table shows the comparison between cancer-associated and the other miRNAs in respect to three network properties: degree and clique level (CLV). All the *P*-values were calculated by the Wilcoxon rank sum test.

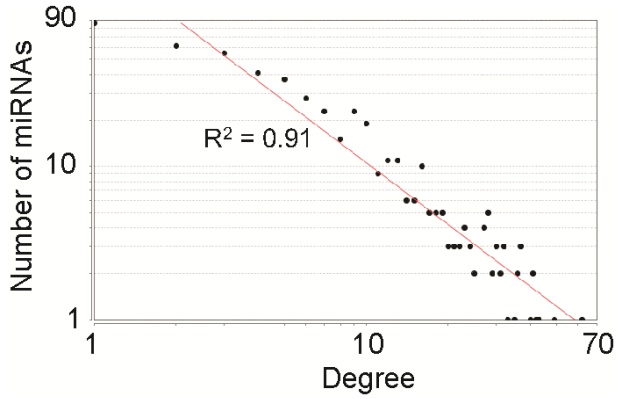


Figure S4: The degree distribution of the proposed network.

The degree distribution of the cross-cancer miRNA differential co-expression network shows that this network is scale-free.

S4 Identification of the pan-cancer activated miRNA-regulated functional modules

To identify the pan-cancer activated miRNA-regulated functional modules, we applied a combinational procedure reported in two previous works[1,2]. This approach identified the modules according to three qualifiers: 1) miRNA regulation, 2) functional module, and 3) activation across multiple cancer types.

For the first part, three algorithms, TargetScan[3-5], miRanda[6], and MultiMiTar[7], were used to predict putative miRNA target genes. The predictions of TargetScan were directly downloaded from the website (http://www.targetscan.org/). For miRanda, we applied a stringent cut-off value (score $\geq$ 140) to obtain confident predictions of miRNA target genes. MultiMiTar was run with the default parameters suggested by the authors. To filter out possible false-positive predictions, only the putative miRNA targets predicted by at least two algorithms were used in further analysis.

Next, we combined predicted miRNA targets and their protein interaction partners to construct the miRNA-regulated networks. The protein-protein interactions (PPIs) were obtained from the Protein Interaction Network Analysis (PINA) v2[8,9]. Then, we performed functional enrichment analyses using Gene Ontology (GO)[10] annotations to uncover functional modules in the miRNA-regulated networks. Subnetworks in the miRNA-regulated networks with significantly overrepresented (Hypergeometric test, $P \leq 0.05$) GO terms were further considered as miRNA-regulated functional modules. The calculated *P*-values were adjusted by applying the Benjamini and Hochberg multiple testing procedures to control the false discovery rate (FDR)[11].

To assess the activities of identified miRNA-regulated functional modules in normal and cancer samples, we tested the enrichment of co-expressed protein-protein interactions (CePPIs) in each module. CePPIs were defined as PPIs that were formed by two proteins encoded by significantly and positively co-expressed genes (*P*-value of PCC $\leq$ 0.01). CePPIs were further considered as activated interactions in the corresponding condition. Accordingly, the miRNA-regulated functional modules with overrepresented CePPIs ($P \leq 0.05$, Fisher's exact test) were considered activated in the corresponding condition, e.g. cancer or normal. Finally, the miRNA-regulated functional modules that were activated in all four cancer types but not in normal samples were defined as pan-cancer activated miRNA-regulated functional modules.

However, based on this approach, we established seven cut-offs (from one to seven) for the number of K_7 miRNA. In addition, we considered three types of miRNA-regulated protein

interaction networks (PINs): "target gene only," "target gene plus the PPI partners," and "target gene plus the common PPI partners." Therefore, for each K_7, there were 21 possible combinations of its regulatory networks. To decide which combination would be used for further analyses, we mapped cancer-associated genes onto miRNA-regulated PINs to separately obtain the so-called precision and recall of cancer genes in each combination. Then, we used the F-measure, which is the harmonic mean of precision and recall, to assess the coverage capability of the cancer genes of the miRNA-regulated PINs. The highest F-measures for M_1 and M_2 were reached by target genes plus common PPI partners, using a cut-off of 4 miRNAs in K_7 (M_1: 0.18, M_2:0.17, Fig. S8). Therefore, we used this combination to identify the pan-cancer activated miRNA-regulated functional modules.

Table S4: The enrichment of co-expressed PPIs in two identified functional modules in seven TCGA cancer types.

M1	Description	BRCA	HNSC	KIRC	LUAD	OV	PRAD	STAD	COAD	THCA	UCEC	LUSC
GO:0048285	organelle fission	1.25E-05	5.69E-07	1.19E-07	6.88E-07	3.33E-07	2.65E-06	1.40E-05	1.99E-04	8.97E-05	2.82E-09	4.12E-06
GO:0000280	nuclear division	6.07E-11	9.26E-13	4.57E-10	8.13E-13	2.61E-16	1.02E-06	4.86E-15	1.96E-13	2.87E-03	1.19E-12	1.42E-12
GO:0007067	mitosis	1.25E-05	5.69E-07	1.19E-07	6.88E-07	3.33E-07	2.65E-06	1.40E-05	1.99E-04	8.97E-05	2.82E-09	4.12E-06
GO:0006260	DNA replication	2.04E-05	8.89E-07	2.24E-07	1.14E-06	4.78E-07	4.51E-06	2.12E-05	2.72E-04	1.39E-04	4.36E-09	6.28E-06

M2	Description	BRCA	HNSC	KIRC	LUAD	OV	PRAD	STAD	COAD	THCA	UCEC	LUSC
GO:0048285	organelle fission	3.08E-06	8.44E-09	8.67E-09	3.98E-10	3.80E-09	3.34E-06	1.63E-07	1.08E-04	3.36E-06	1.11E-10	3.23E-07
GO:0000280	nuclear division	7.38E-11	2.24E-14	2.44E-09	6.32E-10	3.55E-13	6.55E-06	2.51E-13	3.19E-13	4.42E-02	9.44E-12	2.66E-12
GO:0007067	mitosis	3.08E-06	8.44E-09	8.67E-09	3.98E-10	3.80E-09	3.34E-06	1.63E-07	1.08E-04	3.36E-06	1.11E-10	3.23E-07
GO:0006260	DNA replication	3.08E-06	8.44E-09	8.67E-09	3.98E-10	3.80E-09	3.34E-06	1.63E-07	1.08E-04	3.36E-06	1.11E-10	3.23E-07

These tables show the enrichment of co-expressed PPIs in the two identified functional modules. All the *P*-values were two-tailed and calculated by Fisher's exact test.

Table S5: Enriched GO functions in mRNA co-expression network in the four cancer types

Cancer type	Enriched Functions
Stomach	regulation of biological process
	regulation of cellular process
	antigen processing and presentation of peptide antigen via MHC class I
	antigen processing and presentation of exogenous peptide antigen via MHC class I
	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
Ovarian	antigen processing and presentation
	antigen processing and presentation of exogenous antigen
	regulation of cellular process
	antigen processing and presentation of peptide antigen
	antigen processing and presentation of peptide antigen via MHC class I
	antigen processing and presentation of exogenous peptide antigen
	antigen processing and presentation of exogenous peptide antigen via MHC class I
	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
Lung	DNA replication initiation
Prostate	regulation of cellular process
	metabolic process
	biological regulation
	regulation of biological process
	organic substance metabolic process
	single-organism cellular process

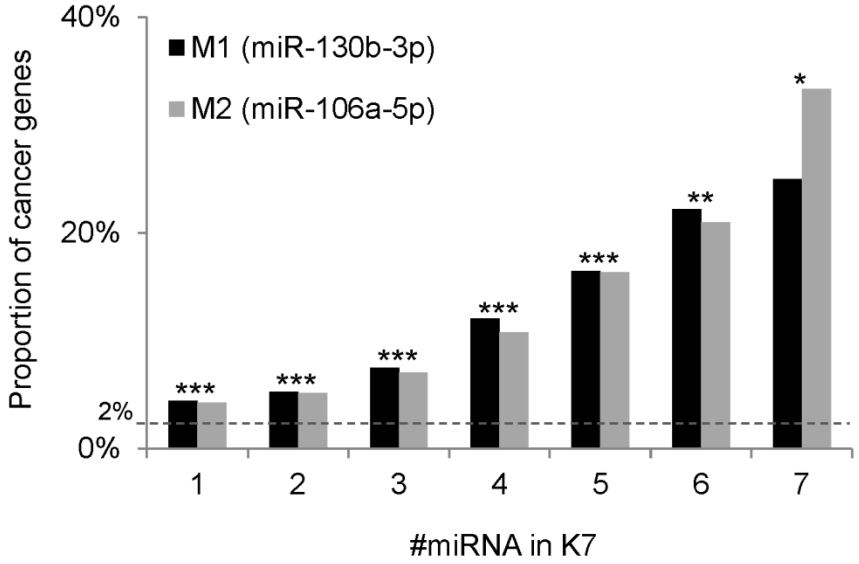
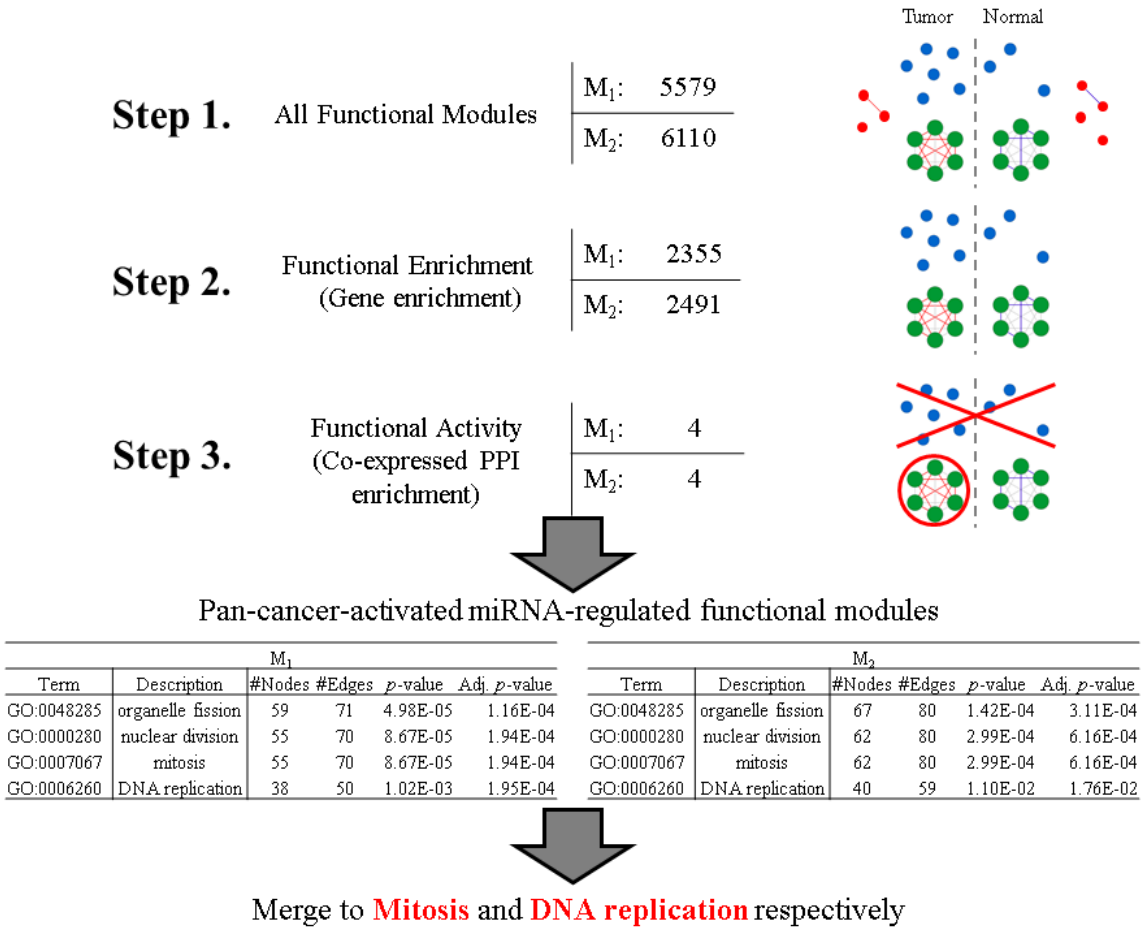


Figure S5: The proportion of cancer-associated genes in K_7 miRNA target gene sets.

Genes targeted by more miRNAs in M1 or M2 tend to be reported as cancer-associated genes by the Cancer Gene Census. The dashed line shows the proportion of cancer-associated genes which are targeted by at least one human miRNA. The significance of each bar was tested by Fisher's exact test (***: $P \leq$ 1e-05, **: $P \leq$ 1e-03, *: $P \leq$ 1e-01).



M_1

Term	Description	#Nodes	#Edges	*p*-value	Adj. *p*-value
GO:0048285	organelle fission	59	71	4.98E-05	1.16E-04
GO:0000280	nuclear division	55	70	8.67E-05	1.94E-04
GO:0007067	mitosis	55	70	8.67E-05	1.94E-04
GO:0006260	DNA replication	38	50	1.02E-03	1.95E-04

M_2

Term	Description	#Nodes	#Edges	*p*-value	Adj. *p*-value
GO:0048285	organelle fission	67	80	1.42E-04	3.11E-04
GO:0000280	nuclear division	62	80	2.99E-04	6.16E-04
GO:0007067	mitosis	62	80	2.99E-04	6.16E-04
GO:0006260	DNA replication	40	59	1.10E-02	1.76E-02

Figure S6: The number of modules during the identification of the pan-cancer-activated miRNA-regulated functional modules.

The number of modules during the identification of the pan-cancer activated miRNA-regulated functional modules. Genes are denoted as nodes and assigned with the same color when they share the same GO terms. In the step 1, genes the pan-cancer-activated miRNAs-regulated network were annotated by GO terms and grouped as functional modules according to their sharing terms. Next, in the step 2, only the functional modules significantly enriched with number of genes (blue and green module) were kept. Finally, in the step 3, the remaining functional modules which are overrepresented co-expressed PPIs in tumor but not in normal were denoted as candidates, i.e. green module in tumor. Among the candidate functional modules, because both mitosis and nuclear division are subsets of organelle fission, we further merged these three modules into one. This union module was termed mitosis. Additionally, M_1- and M_2-regulated mitosis modules were merged as a union mitosis module, because they shared about 70% of their module members. Analogously, M_1- and M_2-regulated DNA replication modules were also merged as a union DNA

replication, due to their 77% overlapped member genes. The overlapping proportions were calculated by using the Jaccard index. Finally, we only left two pan-cancer activated miRNA-regulated functional modules, mitosis and DNA replication.

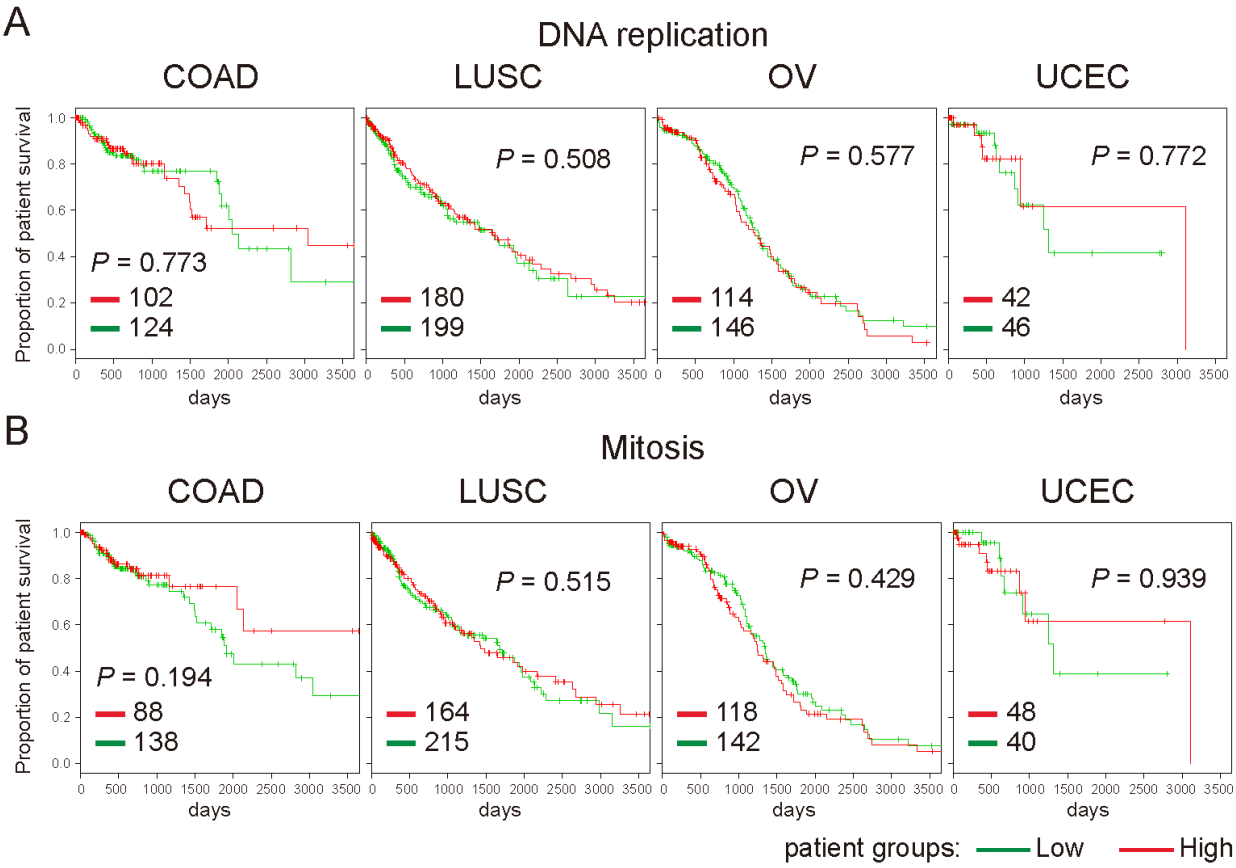


Figure S7: Kaplan-Meier survival curves for identified functional modules in COAD, LUSC, OV, and UCEC samples.

Kaplan-Meier 10-year survival curves for two pan-cancer activated functional modules in four cancer types: COAD, LUSC, OV, and UCEC. Patients were grouped into lowly and highly expressed groups, based on the average expression levels of genes in the pan-cancer activated functional modules. The *P*-values were derived from the Cox's regression model with age as an explanatory variable.

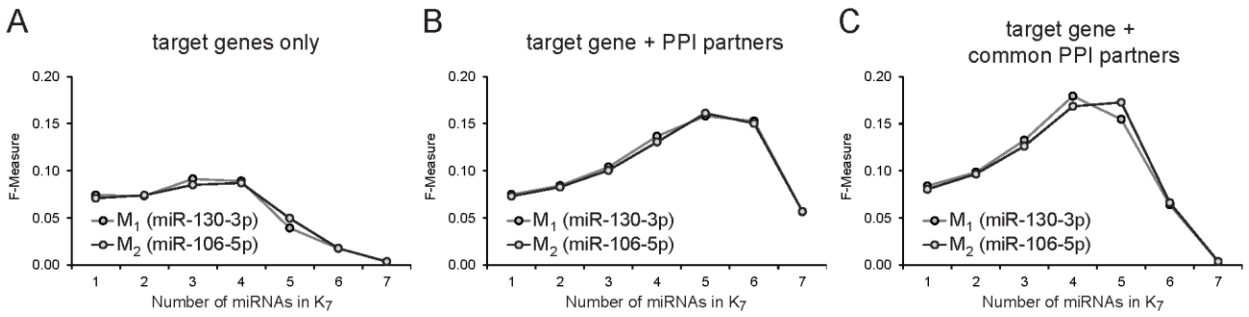


Figure S8: The F-Measure for all combinations of miRNA-regulated PINs.

We calculated the proportion of cancer-associated genes in the miRNA-regulated PINs to obtain so-called precision and recall of cancer-associated genes in each combination separately. Then, we used the F-measure, which is the harmonic mean of precision and recall, to assess the coverage capability of cancer-associated genes in the miRNA-regulated PINs. The highest F-measure for M_1 and M_2 can be reached by the target gene plus common PPI partners on the cut-off of 4 miRNAs in K_7 (M_1: 0.18, M_2:0.17).

S5 Influence of sample size to the distributions of Pearson correlation coefficients and *P*-values.

To discuss the influence of sample size to construct the co-expression network, we randomly generated data with sample size from 3, 4, 5, 6, 7, 8, 9, 10, 13, 15, 23, 32, 70, 100, 200, 300, 400, 500, 1000, and 2000. The first 8 and last 7 sizes were selected to see the influence when the sample size is smaller and larger than the dataset we used in the study, respectively. The result was depicted in Figure S9 and S10. Obviously, the distributions of Pearson correlation coefficient (PCC) were affected by sample size, while the distributions of PCC *P*-value didn't. That is, the size of co-expression network was influenced by the sample size if we used PCC as cut-off to construct the network. For example, if we used $|PCC| > 0.6$ as the cut-off to construct the co-expression network, we would get different number of co-expression with different sample size (Table S6). Moreover, if the size of control sample is 70 and the sizes of case sample are 13, 15, 23, and 32, we would only get the differential co-expression of losing positive and negative ones. However, the number of co-expression was stable if we applied PCC *P*-value as the cut-off (Table S6). Therefore, the bias of using PCC as cut-off could be reduced, even removed. That is the reason we used PCC *P*-value instead of PCC value to obtain significantly correlated miRNA pairs. We added these above observations in the Supporting Information to discuss the influence of sample size to construct the co-expression network.

Table S6: Number of selected co-expression under varied sample size with fixed cut-off of PCC or *P*-value

Sample Size	3	4	5	6	7	8	9	10	13	15
$\|PCC\| \geq 0.6$	55085	39820	28705	21265	15869	12008	9226	7131	3313	2100
P-value < 0.01	992	1042	1085	1101	1072	1128	1089	1078	1067	1124
Sample Size	23	32	70	100	200	300	400	500	1000	2000
$\|PCC\| \geq 0.6$	295	37	0	0	0	0	0	0	0	0
P-value < 0.01	1047	993	1022	1022	999	1003	1001	1000	1015	952

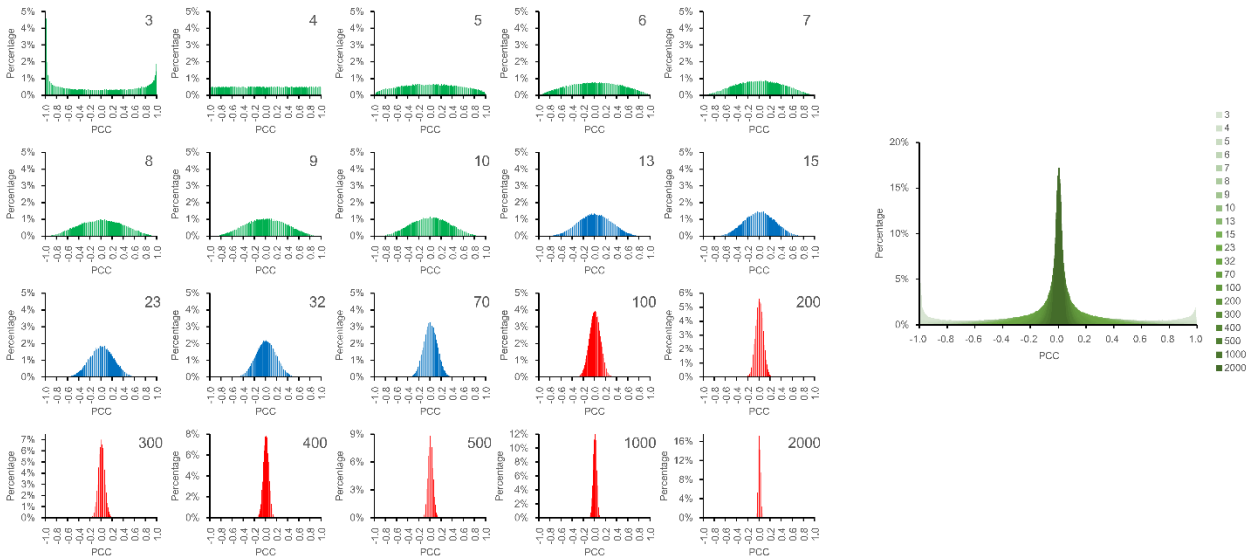


Figure S9: The distributions of Pearson correlation coefficient (PCC) from 100000 randomized pairs under different sample size.

The distribution of PCC varies from sample size to sample size: the range of distributions became narrower and narrower as the sample size increased. Consequently, the size of co-expression network was influenced by the sample size if we used PCC as cut-off to construct the network. The numbers on the upper right corner denote sample size. Green charts are from the distributions with sample size smaller than the data we used in this study; blue ones are with sample size equal to our dataset; red ones are with sample size larger than we used. We merged these 20 distributions as the chart on the right (darker green denotes larger sample size).

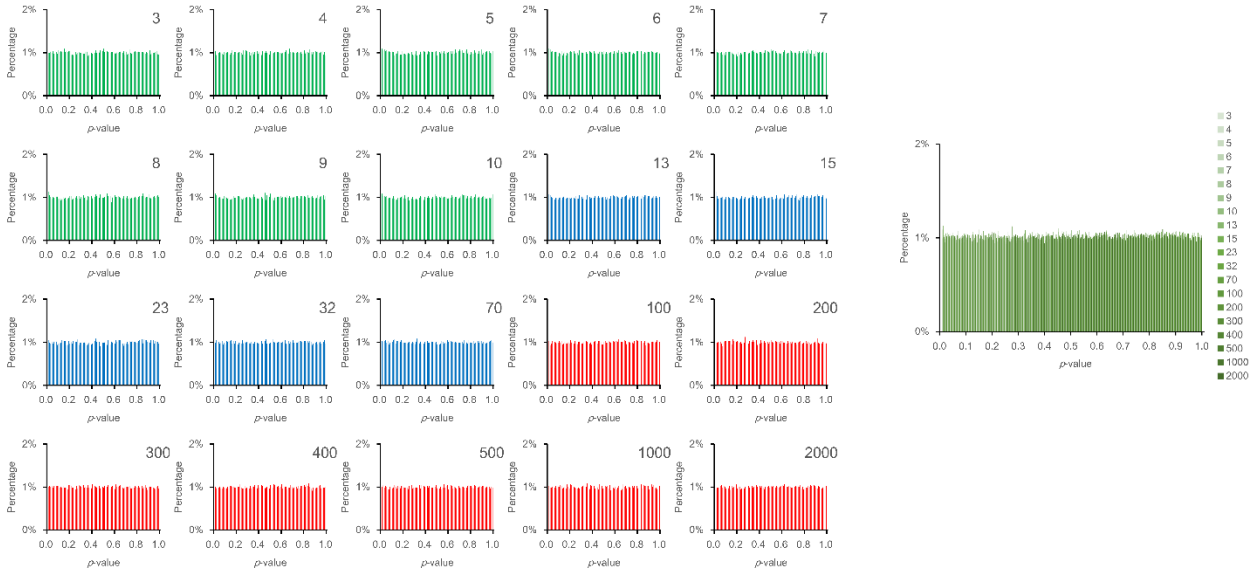


Figure S10: The distributions of PCC *P*-value from 100000 randomized pairs under different sample size.

Obviously, the distributions of PCC *P*-value were not affected by the sample size: they are all uniform. Consequently, the size of co-expression network was not influenced by the sample size if we used PCC *P*-value as cut-off to construct the network. The numbers on the upper right corner denote sample size. Green charts are from the distributions with sample size smaller than the data we used in this study; blue ones are with sample size equal to our dataset; red ones are with sample size larger than we used. We merged these 20 distributions as the chart on the right (darker green denotes larger sample size).

References

1 Lin CC, Hsiang JT, Wu CY, Oyang YJ, Juan HF, Huang HC. Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC systems biology* 2010;4: 138.

2 Tseng CW, Lin CC, Chen CN, Huang HC, Juan HF. Integrative network analysis reveals active microRNAs and their functions in gastric cancer. *BMC systems biology* 2011;5: 99.

3 Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;120(1): 15-20.

4 Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* 2007;17(12): 1850-64.

5 Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* 2011;469(7328): 97-101.

6 Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. *Genome Biol* 2003;5(1): R1.

7 Mitra R, Bandyopadhyay S. MultiMiTar: a novel multi objective optimization based miRNA-target prediction method. *PLoS One* 2011;6(9): e24583.

8 Cowley MJ, Pinese M, Kassahn KS, et al. PINA v2.0: mining interactome modules. *Nucleic acids research* 2012;40(Database issue): D862-5.

9 Wu J, Vallenius T, Ovaska K, Westermarck J, Makela TP, Hautaniemi S. Integrated network analysis platform for protein-protein interactions. *Nature methods* 2009;6(1): 75-7.

10 Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000;25(1): 25-9.

11 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57(1): 289-300.