SUPPLEMENTAL MATERIAL

1. Supplemental Methods

1.1. Pseudo code of SiteTest

Algorithm: *z* = SiteTest(*GTmiRs*, *mRNA$_i$*). Calculate a binding score of GTmiRs that targets *mRNA$_i$*.

Input: *GTmiRs*: a set of miRNAs targeting GOI

mRNA$_i$: ith mRNA to be evaluated as a ceRNA candidate

Output: A score z for *GTmiRs* binding to *mRNA$_i$*

Steps:

1) *Evaluate following mRNA$_i$ features in 3' UTR*

1.1) Ratio of the number of miRNAs that target *mRNA$_i$* over the total number of miRNAs that target the GOI

$$z_1 = \frac{|\{miRNA_k : (miRNA_k \in GTmiRs) \ \& \ (miRNA_k \text{ targeting } mRNA_i)\}|}{|GTmiRs|}$$

Where |*| is the cardinality of the set GTmiRs.

1.2) *Density of MREs in the region*. Let MRE(*miRNA$_k$*) be all MREs i*miRNA$_k$*, and |MRE(*miRNA$_k$*)| the cardinally of MRE(*miRNA$_k$*). The density of the MREs of *miRNA$_k$*, *s$_k$*, is defined as,

$$s_k = \begin{cases} \frac{length(MRE(miRNA_k))}{length(3'UTR)} & |\text{MRE}(miRNA_k)| = 1 \\ \frac{|\text{MRE}(miRNA_k)| * length(MRE)}{D(\text{rightmost of MRE}(miRNA_k), \text{leftmost of MRE}(miRNA_k))} & \text{otherwise} \end{cases}$$

where D(*x, y*) is the distance between elements *x* and *y*, and the function length(•) is the length of the sequence element in base-pair unit. The score

for all *miRNA*$_k \in$ GTmiRs is $z_2 = \sum_{k=1}^{K} s_k$. The score prefers more MREs within the region they spans;

1.3) *Distribution of the sites*. The distribution of the binding sites of *miRNA*$_k$ is defined as,

$$m_k = \begin{cases} \delta & |\mathrm{MRE}(miRNA_k)| = 1 \\ \frac{\left(\text{rightmost MRE}(miRNA_k) - \text{leftmost MRE}(miRNA_k)\right)^2}{\sum_{k=1}^{K}(\text{distance between successive MREs of } miRNA_k)^2} & \text{otherwise} \end{cases}$$

where δ is a small quantity < 1, which is introduced to penalize the case of having single MRE per miRNA (we select δ = 0.1 in the actual implementation). The score for all *miRNA*$_k \in$ GTmiRs is $z_3 = \sum_{k=1}^{K} m_k$. The larger the value of z_3, the more evenly distributed the MREs. Thus, z_3 penalizes MREs that group together within a narrow region;

1.4) Ratio of the number of MREs for *mRNA*$_i \in$ GTmiRs over the total number of MREs for GOI.

$$z_4 = \frac{|\text{MREs targeting } mRNA_i|}{|\text{MREs targeting GOI}|}$$

2) Final prediction score is the multiplication of all 4 scores, or $z = z_1 \cdot z_2 \cdot z_3 \cdot z_4$.

1.2. TraceRNA Web Implementation

TraceRNA was developed under the Linux environment (2.6.32-279.11.1.el6.x86_64) and by using Server API Apache 2.0. TraceRNA application is composed of 3 software modules: an HTML module for user interface; a server-side module that contains scripts (PHP Version 5.3.3) to process the input from the user interface and interact with a MySQL database (client API 5.1.61); and an algorithm module that consists of a set of *R* scripts for statistical calculations. User interface was carefully designed to allow biologists to interact with TraceRNA without sophisticated bioinformatics skills and experiences. The ceRNA regulatory network is drawn by using Cytoscape Web (version 1.0.2). Users can also upload a gene expression dataset directly to TraceRNAs;

however, it may be limited by dataset size due to time and memory allocation to the web server. On the server side, a set of PHP programs were constructed to process the parameters inputted by users, create SQL queries to obtain required subsets of data, perform statistical analysis (*R* scripts), and visualize the results through web interface. All data are passed to *R* scripts for calculation of predicted *p*-values and FDR. Supplemental Fig. S1 provides a screenshot of the TraceRNA user-interface.

2. Comparison of PTEN ceRNAs predictions by TraceRNA and those in (Sumazin et al. 2011)

We compared the predictions of *PTEN* ceRNAs in GBM by TraceRNA with those reported by (Sumazin et al. 2011). Thirteen *PTEN* ceRNAs (*ABHD13, CCDC6, CTBP2, DCLK1, DKK1, HIAT1, HIF1A, KLF6, LRCH1, NRAS, RB1, TAF5* and *TNKS2*) were identified in (Sumazin et al. 2011) and six negative predictions (*POFUT1, DDX24, SLC46A3, EXTL3, PIK3R2* and *EHMT2*) were also verified. For TraceRNA, at *p*-value < 0.05, its predictions included five of these thirteen genes (*CCDC6, TNKS2, CTBP2, HIF1A* and *RB1*) and none of the six negative predictions by (Sumazin et al. 2011). To investigate potential reasons for TraceRNA missing the other 8 genes, we examined the Pearson correlation between the expressions of *PTEN* and each of the 8 genes and found that all the Pearson correlations are low (Supplemental Table S2). As a result, these 8 genes are unlikely to be predicted as ceRNAs by TraceRNA because ceRNAs need to have high expression correlation with GOI. The fact that the expression correlations are low in this study is because only 262 GBM samples were available from TCGA 2008 deposit were used in (Sumazin et al. 2011), whereas we used 450 GBM

samples from TCGA 2012 deposit. The difference in expression and hence the predicted ceRNAs suggest that ceRNAs are context-dependent.

[Sumazin, et al 2011] Sumazin P, Yang X, Chiu HS, Chung WJ, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J et al. 2011. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 147(2): 370-381.

3. Supplemental Tables

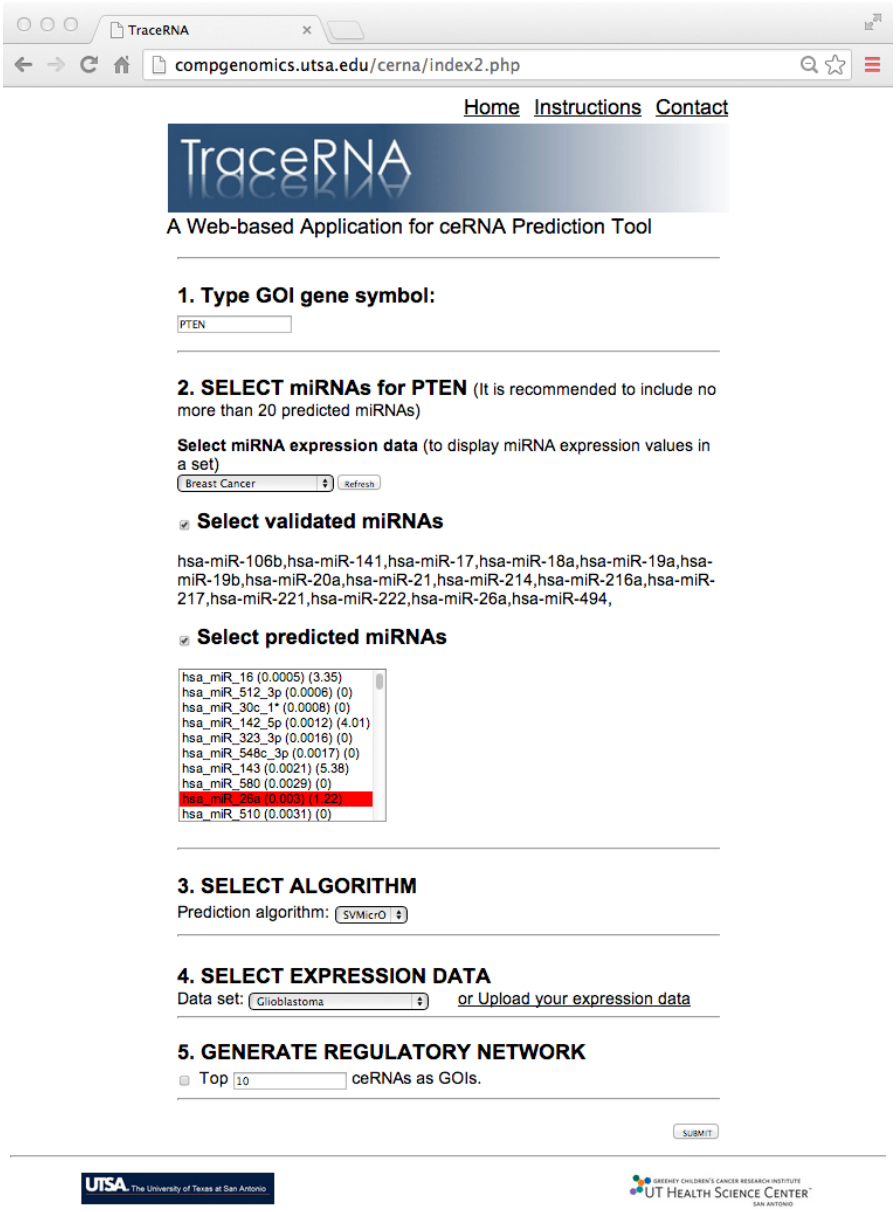
Supplemental Table S1. The fitted Gamma parameters for three algorithms

Algorithm	Parameters of fitted Gamma distribution	
	α	β
SVMicrO	0.71861	0.35867
BCMicrO	0.32997	13.00204
SiteTest	0.30132	14.00241

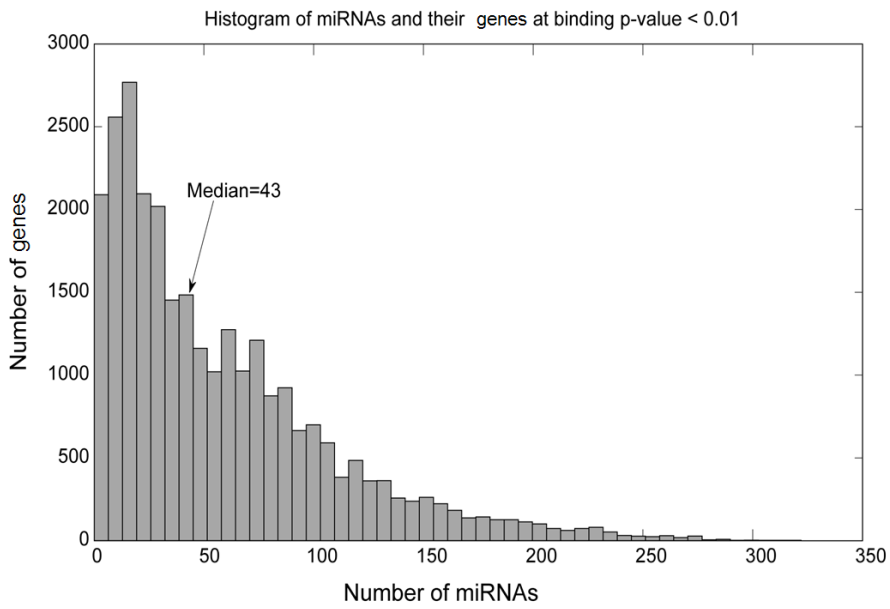
Supplemental Table S2. Correlation of PTEN and 8 predicted ceRNAs in (Sumazin et al. 2011)

Gene	Correlation
ABHD13	0.3032
DCLK1	0.2215
DKK1	0.1900
HIAT1	0.1125
KLF6	0.4137
LRCH1	0.1498
NRAS	0.2192
TAF5	0.3675

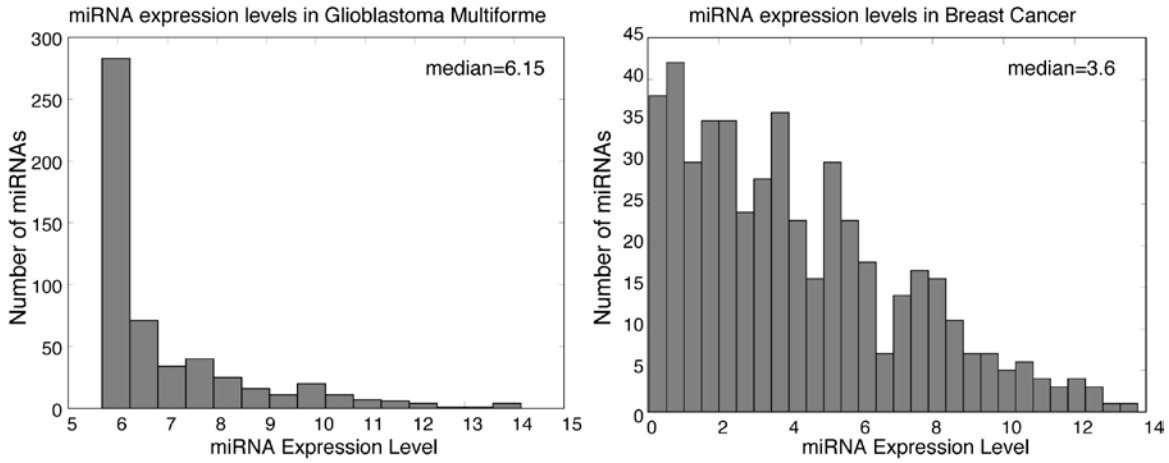
4. Supplemental Figures.



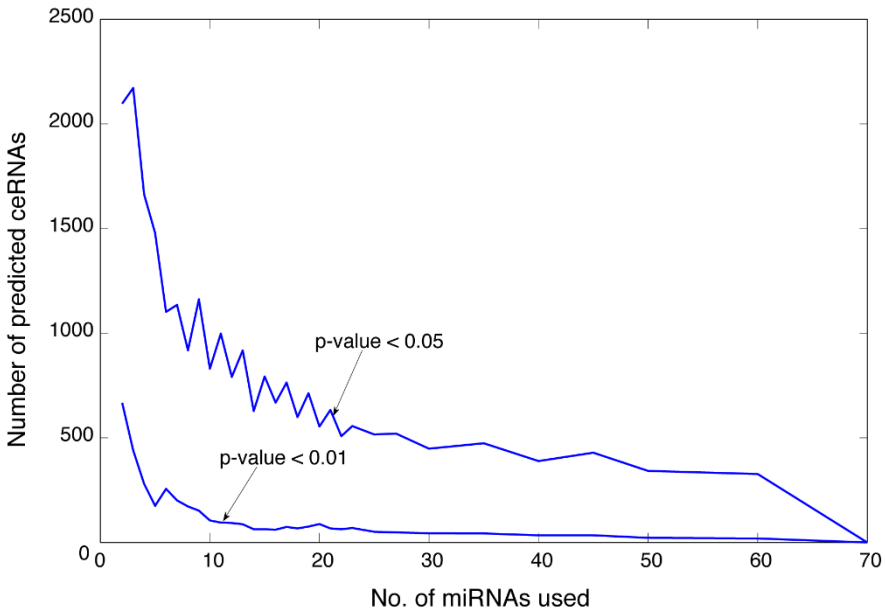
Supplemental Figure S1. TraceRNA user interface. 1) The gene of interest (GOI) is the initial input from the user; 2) After selecting the gene, a user needs to select the validated miRNAs and/or predicted miRNAs from the list. The predicted miRNAs is a list of predictions of binding of miRNAs to the GOI (in this figure, PTEN). The list is ordered by the sequence prediction *p*-value (first value in the parentheses, smaller p-value is more significant). User can also choose a miRNA expression dataset (in the figure, breast cancer) as a reference for selecting ceRNAs. miRNA expression level is provided as second value within the parentheses; 3) The user selects one of the three possible algorithms for ceRNAs predictions: SVMicrO-based, BCMicrO-based or SiteTest; 4) The user selects one expression data or upload a new set of expression data; and 5) The user select this option to generate a regulatory network as a list of genes and a graphic representation.



Supplemental Figure S2. Number of mRNAs (y-axis) targeted by *n* miRNAs (x-axis) with binding *p*-value < 0.01. As indicated in the figure, 50% of genes targeted by more than 43 miRNAs with binding *p*–value < 0.01, or in other words, we most likely need to select no more than 43 miRNAs for each GOI with binding *p*-value < 0.01. Other selection criterion will further limit the choice: by using miRNA expression (Figure S3) or by using specificity (Figure S4).



Supplemental Figure S3. miRNA expression histograms from TCGA glioblastoma multiforme dataset and breast cancer dataset. The median expression levels are 6.15 and 3.6 for glioblastoma and breast cancer, respectively.



Supplemental Figure S4. Number of predicted ceRNAs (with *p*-value < 0.05 or 0.01) when vary the number of miRNAs targeting each gene. The figure is generated by examining how many mRNAs will be targeted by *k* miRNAs with predicting *p*-value less than specified (Eq. 2). For very few miRNAs, the average number of predicted ceRNAs is quite large, indicating non-specific prediction. For about 20 or 10 miRNAs for *p*-value < 0.05 or 0.01, respectively, the number of predicted ceRNAs is stabilized, potentially indicating capturing ceRNA candidates.