PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Model-based analysis of competing-endogenous pathways (MACPath) in human cancers

Hyun Jung Park[1]*, Soyeon Kim[2], Wei Li[3,4]*

1 Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **2** Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas, United States of America, **3** Division of Biostatistics, Dan L Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, United States of America, **4** Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas, United States of America

* hj_park@pitt.edu (HJP); WL1@bcm.edu (WL)

Abstract

Competing endogenous RNA (ceRNA) has emerged as an important post-transcriptional mechanism that simultaneously alters expressions of thousands genes in cancers. However, only a few ceRNA genes have been studied for their functions to date. To understand the major biological functions of thousands ceRNA genes as a whole, we designed Model-based Analysis of Competing-endogenous Pathways (MACPath) to infer pathways co-regulated through ceRNA mechanism (cePathways). Our analysis on breast tumors suggested that NGF (nerve growth factor)-induced tumor cell proliferation might be associated with tumor-related growth factor pathways through ceRNA. MACPath also identified indirect cePathways, whose ceRNA relationship is mediated by mediating ceRNAs. Finally, MACPath identified mediating ceRNAs that connect the indirect cePathways based on efficient *integer linear programming* technique. Mediating ceRNAs are unexpectedly enriched in tumor suppressor genes, whose down-regulation is suspected to disrupt indirect cePathways, such as between DNA replication and WNT signaling pathways. Altogether, MACPath is the first computational method to comprehensively understand functions of thousands ceRNA genes, both direct and indirect, at the pathway level.

Author summary

Competing endogenous RNAs (ceRNA) are RNAs known to co-regulate each other's expression levels by competing for the binding of shared microRNAs. In human cancer, thousands genes are known to be in the ceRNA relationship. To understand the major biological functions of the ceRNA genes as a whole, we offer Model-based Analysis of Competing-endogenous Pathways (MACPath). Based on a network model representing ceRNA relationship changes in tumor, MACPath identifies biological pathways co-regulated through the ceRNA relationship (cePathways), since a biological pathway represents a group of genes coordinating to carry out the same biological function. In breast tumor

OPEN ACCESS

Citation: Park HJ, Kim S, Li W (2018) Model-based analysis of competing-endogenous pathways (MACPath) in human cancers. PLoS Comput Biol 14(3): e1006074. https://doi.org/10.1371/journal.pcbi.1006074

Editor: Edwin Wang, University of Calgary Cumming School of Medicine, CANADA

Received: September 5, 2017

Accepted: March 6, 2018

Published: March 22, 2018

Data Availability Statement: The open source MACPath program is freely available at https://github.com/thejustpark/MACPath.

Funding: This work was supported by the US National Institutes of Health (NIH) grants R01HG007538 and R01CA193466, Cancer Prevention Research Institute of Texas (CPRIT) grant RP150292 to WL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

data, MACPath shows that extensive ceRNA relationship changes in tumor are associated with previously under-characterized tumorigenic crosstalk e.g. between nerve growth factor and other tumor-related growth factor pathways. MACPath also identified indirect cePathways, whose ceRNA relationship is mediated by mediating ceRNAs. When usually thousands ceRNAs mediate indirect cePathways, MACPath employed efficient integer linear programming technique to characterize the mediation. In the breast tumor data, the characterization provides insights into how indirect cePathways and the mediating ceRNAs promote the tumorigenic process. To the best of our knowledge, MACPath is the first computational framework that can comprehensively identify cePathway relationship, both direct and indirect.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Competing endogenous RNAs (ceRNAs) are transcripts that share microRNA (miRNA) binding sites. Through competition for the binding of shared miRNAs, ceRNAs co-regulate each others' expression levels[1] (ceRNA relationship). Although its scope is not fully understood [2], ceRNA has been shown to play a prominent role in regulating gene expression in diverse physiological[3, 4] and pathological conditions, especially for human cancers (reviewed in [5]). Tumor-associated ceRNA function has been identified for a number of individual genes. For example, as *CNOT6L* and *VAPA* have ceRNA relationship with *PTEN*, they regulate *PTEN* itself and phenocopy its tumor suppressive properties[6].

Recently, several studies reported widespread ceRNA relationship (ceRNA network) in tumors involving thousands ceRNA genes[7, 8]. To identify the tumor-associated function of thousands ceRNA genes as a whole rather than arbitrarily chosen individual genes, we developed Model-based Analysis of Competing-endogenous Pathways (MACPath). MACPath groups ceRNA relationships by competing-endogenous pathways (cePathway), i.e. pairs of biological pathways enriched with ceRNA relationships in-between. Since a biological pathway represents a group of genes coordinating to carry out the same biological function, a cePathway relationship may represent a co-regulation between biological functions (crosstalk) through enriched ceRNA relationships. Furthermore, MACPath also identifies dysregulated cePathway relationships between cancer and normal cells. For example, cePathway relationships gained in tumors might be oncogenic, while those lost in tumors might be tumor-suppressive. Finally MACPath can identify dysregulation in indirect cePathway relationship, i.e. co-regulation between pathways mediated by (direct) ceRNAs[9, 10]. When applied to TCGA breast cancer data, MACPath discovered thousands of lost and gained (direct and indirect) cePathway relationships, which collectively provide novel biological insights into ceRNA regulations during tumorigenesis.

Results

Direct cePathways dysregulated in tumors

To identify direct cePathway relationships dysregulated in tumors, MACPath first builds a network model representing changes in ceRNA relationship between cancer and normal tissues (ceRNA relationship change network, **Def. 1**). Since co-expression between genes that share the same miRNA binding sites indicates the level of their ceRNA relationship[7, 10], their co-expression change (tumor vs. normal) can represent their ceRNA relationship change (Fig 1A). In the network model, MACPath identifies dysregulated cePathway relationship between pathway P and Q (Fig 1B, for example) in three steps. First, MACPath calculates the ceRNA relationship changes between all genes within pathway P and their ceRNA genes. Second, MACPath estimates if pathway Q is over-represented in the ceRNA relationship changes (ceRNA relationship change enrichment score ($CES_P(Q)$), Fig 1C). Third, MACPath estimates the statistical significance (empirical p-value) of the CES score by running a permutation test (**see** Materials and Methods). Pathways P and Q can gain cePathway relationship if they are significantly over-represented in the ceRNA relationship gains in tumors (Fig 1B, for example), or can lose cePathway relationship if over-represented in the ceRNA relationship losses.

In 97 breast tumor and their matched normal samples in TCGA, MACPath identified 1,019 lost and 1,698 gained cePathway relationships (FDR-corrected P value < 0.01 and CES > 0.6, **see** Materials and Methods) from all combinations ($1330^{*}1329/2 = 883{,}785$) of 1,330 mSigDB canonical pathways[11] (S1 Table). To validate the use of MACPath, we identified 377,446 and 36,320 ceRNAs in the normal and tumor data based on co-expression (>0.6 in Pearson's correlation coefficient) and miRNA binding site share (< 0.05 in B-H corrected P-value)[7]. Since ceRNAs between pathways were suggested to indicate crosstalk between pathways[7], checking the ceRNA number difference from normal to tumor would be another way to represent cePathway dysregulation. In particular, pathway pairs losing ceRNAs in-between may represent lost cePathways, and those gaining ceRNAs represent gained cePathways. However, while MACPath generally identified pathway pairs losing many ceRNAs as lost cePathways, some lost cePathways do not lose many ceRNAs (43.3% of the lost cePathways lost less than 10 ceRNAs, S1 Table**,** Fig 1D). For example, Reactome "lipid transport" lost cePathway relationship with PID "HIF1 TF" pathway with a strong over-representation in the ceRNA relationship loss in tumor (P-value = 0, Fig 1E). However, they lost only 2 ceRNAs in tumor. Further, the ceRNA number difference cannot effectively identify gained cePathways, since 10-fold decrease of ceRNAs in tumor (377,446 vs. 36,320) makes it very difficult for pathway pairs to gain ceRNAs (only 8 of the 1,698 gained cePathways (0.4%) gain ceRNAs). Additionally, the ceRNA number difference cannot control for false positives, since it does not estimate significance. Altogether, MACPath identifies statistically significant and biologically reasonable cePathway dysregulation, which is impossible with a naive, yet the only other available method.

Dysregulated cePathways may underlie NGF-induced breast tumor cell proliferation

Many of the gained cePathway relationships involve nerve growth factor (NGF) signaling and tumor-related growth factor pathways, such as TGFβ[12] (transforming growth factor β) and EGFR[13] (epidermal growth factor receptor) (Fig 2A). An example is the gained cePathway relationship between REACTOME "NGF signaling via TRKA" and REACTOME "signaling by EGFR" (Fig 2B). NGF stimulates cell proliferation in tumor, but not in normal cells[14]. The mechanism may involve miRNAs, since NGF interacts with miRNAs for cell survival[15], which usually co-occurs with cell proliferation in cancer cells[16]. Therefore, we speculate that the mechanism underlying the NGF-induced cell proliferation might be due to the gained cePathway relationships between NGF and growth factor pathways. This is consistent with previous literature, in which crosstalk between NGF/TrkA and EGFR has been reported for cell activation[17], and NGF has been reported to up-regulate TGFβ in a cell differentiation experiment[18]. Furthermore, most of the cePathway gains in TCGA data are also found in an

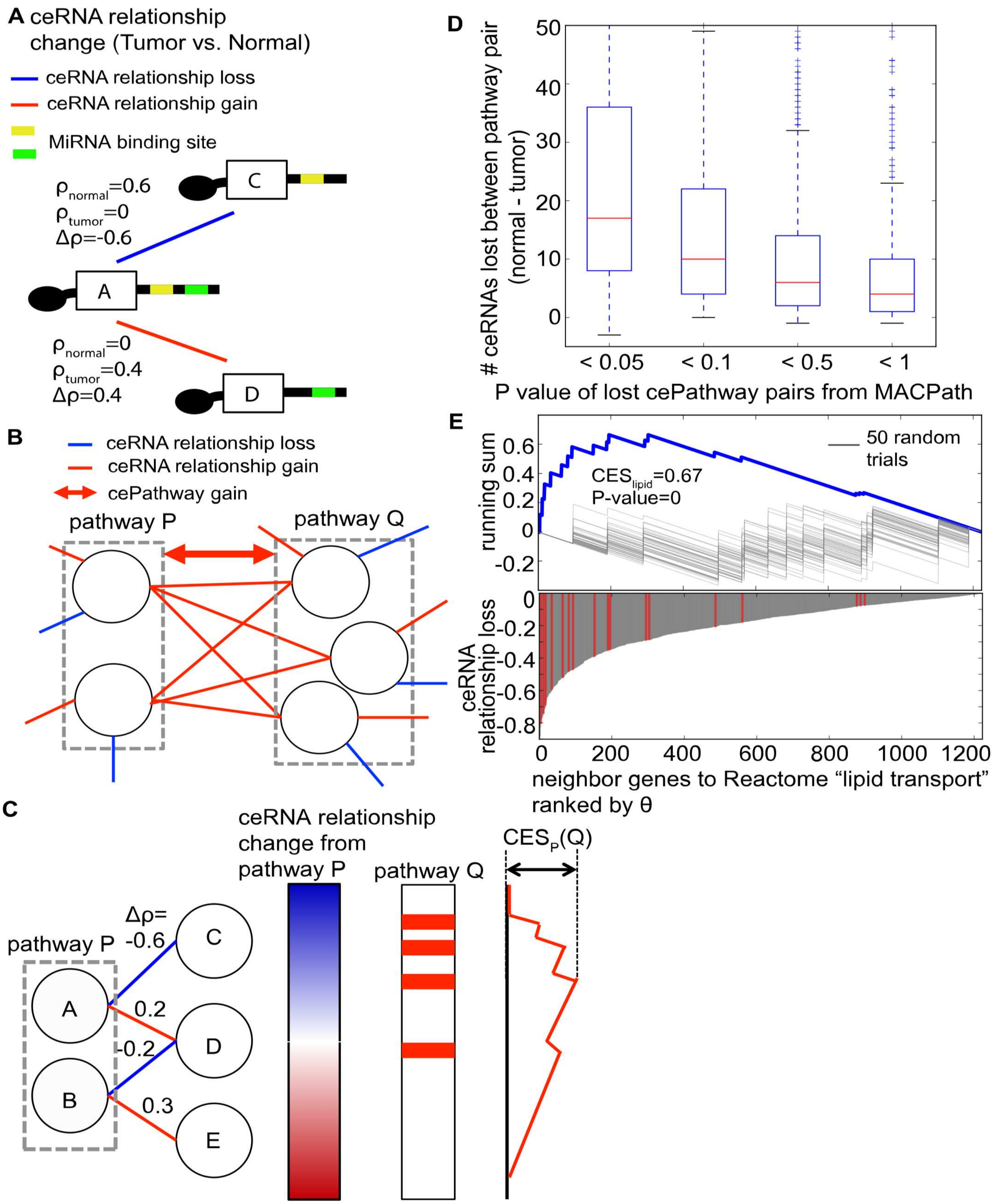
A ceRNA relationship change (Tumor vs. Normal)
ceRNA relationship loss
ceRNA relationship gain
MiRNA binding site
ρnormal=0.6
ρtumor=0
Δρ=-0.6
C
A
ρnormal=0
ρtumor=0.4
Δρ=0.4
D
B
ceRNA relationship loss
ceRNA relationship gain
cePathway gain
pathway P
pathway Q
C
ceRNA relationship change from pathway P
pathway Q
CESP(Q)
pathway P
Δρ= -0.6
0.2
-0.2
0.3
A
B
C
D
E
D
# ceRNAs lost between pathway pair (normal - tumor)
50
40
30
20
10
0
< 0.05
< 0.1
< 0.5
< 1
P value of lost cePathway pairs from MACPath
E
running sum
0.6
0.4
0.2
0
-0.2
50 random trials
CESlipid=0.67
P-value=0
ceRNA relationship loss
0
-0.2
-0.4
-0.6
-0.8
0
200
400
600
800
1000
1200
neighbor genes to Reactome “lipid transport” ranked by θ

Fig 1. Identifying direct cePathway relationship dysregulations in tumors. (a). An illustration of ceRNA relationship change network. RNA A shares a significant number of miRNA binding sites with RNA C and D, where ρ_{normal} and ρ_{tumor} represent their co-expression in normal and tumor samples, respectively. Between normal and tumor samples, co-expression decrease in tumors ($\Delta\rho = \rho_{tumor}\text{-}\rho_{normal}<0$) between A and C could represent their ceRNA relationship loss and co-expression increase ($\Delta\rho>0$) between A and D represent their ceRNA relationship gain. **(b)**. An example of a cePathway relationship gain between pathway P and Q in the ceRNA relationship change network. CePathway relationship loss would be identified in the same fashion.**(c).** An illustration of MACPath to identify cePathway relationship dysregulation between pathway P and Q. Genes showing ceRNA relationship changes from pathway P are sorted by the magnitude of the changes (θ_P from **Def. 2**, blue for loss and red for gain in the left bar). On the genes ranked by θ_P, a running sum is calculated for genes in pathway *Q* (in the right bar). The maximum variation of the running sum from 0 is the ceRNA relationship change enrichment score ($CES_P(Q)$). **(d).** The number difference of ceRNAs between pathway pairs (normal vs. tumor) divided by P value MACPath estimated for the pairs. **(e).** Upper panel shows behavior of running sum between Reactome "lipid transport" pathway and PID "HIF1 TF" pathway. Gray lines represent running sum of 50 permutation trials. Lower panel shows θ of neighbor genes to Reactome "lipid transport" pathways ranked by their θ. Red lines indicate where genes in PID "HIF1 TF" are placed.

https://doi.org/10.1371/journal.pcbi.1006074.g001

independent breast tumor data (GSE57297)[19] using MACPath with the same cutoff (**Fig 2A, see Materials and Methods**), suggesting the cePathway crosstalk between NGF and the growth factors is a common mechanism in breast cancers.

Indirect cePathways dysregulated in tumors

The ceRNA relationship change network that models direct ceRNA relationship (**Def. 1**) can further be used to model indirect ceRNA relationship. For genes x, y, and z, direct ceRNA relationship (x, y) can propagate to direct ceRNA relationship (y, z), making an indirect ceRNA relationship (x, z)[9, 10] (e.g. x, y, and z in normal condition in **Fig 3A**). Inspection on the same number (n = 5,000) of the ceRNA relationships showed that indirect ceRNA relationships are substantially co-expressed ($\rho>0.5$ for 53.8% of the relationships, **Fig 3B**), although less co-expressed than direct ceRNA relationships that are identified based on high co-expression. Additionally, indirect ceRNAs are more highly expressed than direct ceRNAs in the normal samples (P-value = $8.04e^{-44}$, **S2B Fig**). MACPath identifies indirect cePathway relationship based on indirect ceRNAs (x, z) that show the same direction of co-expression change as in (x, y) and (y, z) in the ceRNA network (**see Materials and Methods**). For TCGA breast cancer, MACPath identifies 509 lost and 84 gained indirect cePathway relationships (**S2 Table,** FDR-corrected P value <0.01 in ForwardStop[20]).

Mediating ceRNAs drive indirect cePathway relationships

Extensive ceRNA relationship in tumors[7, 8] implies that many genes would mediate indirect ceRNA relationships. Consistently, our network model identifies on average 840 mediating ceRNAs for each indirect cePathway relationship (**S3 Table**). For example, 1,132 genes are inferred to mediate the significant loss of indirect cePathway relationship between KEGG “DNA replication” and KEGG “WNT signaling” (P = 0.0005, **S2 Table**, **Fig 4A**). In normal cells, both WNT signaling and DNA replication were reported to crosstalk to the same genes, such as those in DNA damage pathways[22, 23], which would yield indirect crosstalk between DNA replication and WNT signaling. In addition, these two pathways were shown to suppress tumor in a miRNA dependent manner[24, 25], possibly through their indirect cePathway relationship.

One important question is to identify which of the 1,132 mediating ceRNAs would exert strong mediation effect for the loss of indirect cePathway. We address this problem by finding the minimum number of mediating genes covering all genes in the indirect cePathways. Finding minimally mediating genes is an NP-hard problem (**see Materials and Methods**), which requires trying all combinations of mediating genes. We formulated the problem such that it is amenable to efficient *integer linear programming* (ILP) solvers (**see Materials and Methods**) and found minimally mediating genes within a second for most indirect cePathway relationship dysregulation. In particular, MACPath employed GNU Linear Programming Kit (GLPK)

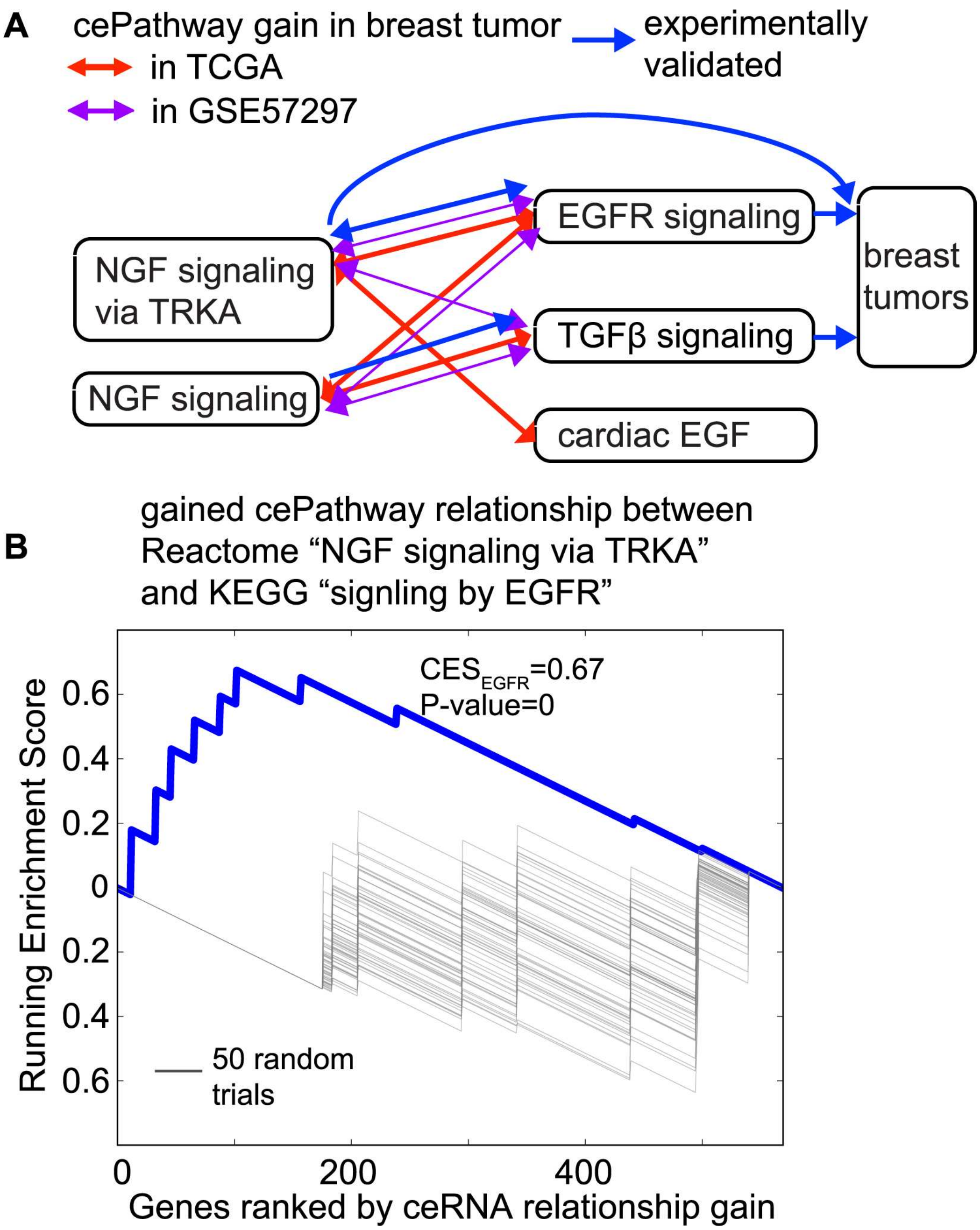

Fig 2. Dysregulated cePathways suggest NGF-induced cell proliferation. (a). Growth factor pathways that gained cePathway relationship in tumors with NGF signaling pathways in the TCGA data (red) and in GSE57297[19] (purple). **(b).** Behavior of running sum between Reactome “signaling by NGF” pathway and KEGG “TGFβ signaling” pathway. Gray lines represent running sum of 50 random trials (Actual p-values are estimated from 2,000 such random trials).

https://doi.org/10.1371/journal.pcbi.1006074.g002

solver, which returns a different set of minimally mediating genes for a different permutation of the same input values. Using the solver on multiple (10,000) permutations on the input values for the lost indirect cePathway relationship KEGG "DNA replication" and KEGG "WNT signaling", MACPath yielded 615 minimally mediating genes, which are surprisingly enriched for tumor suppressors, but not oncogenes (**Fig 4B**). Tumor suppressors that are also minimally

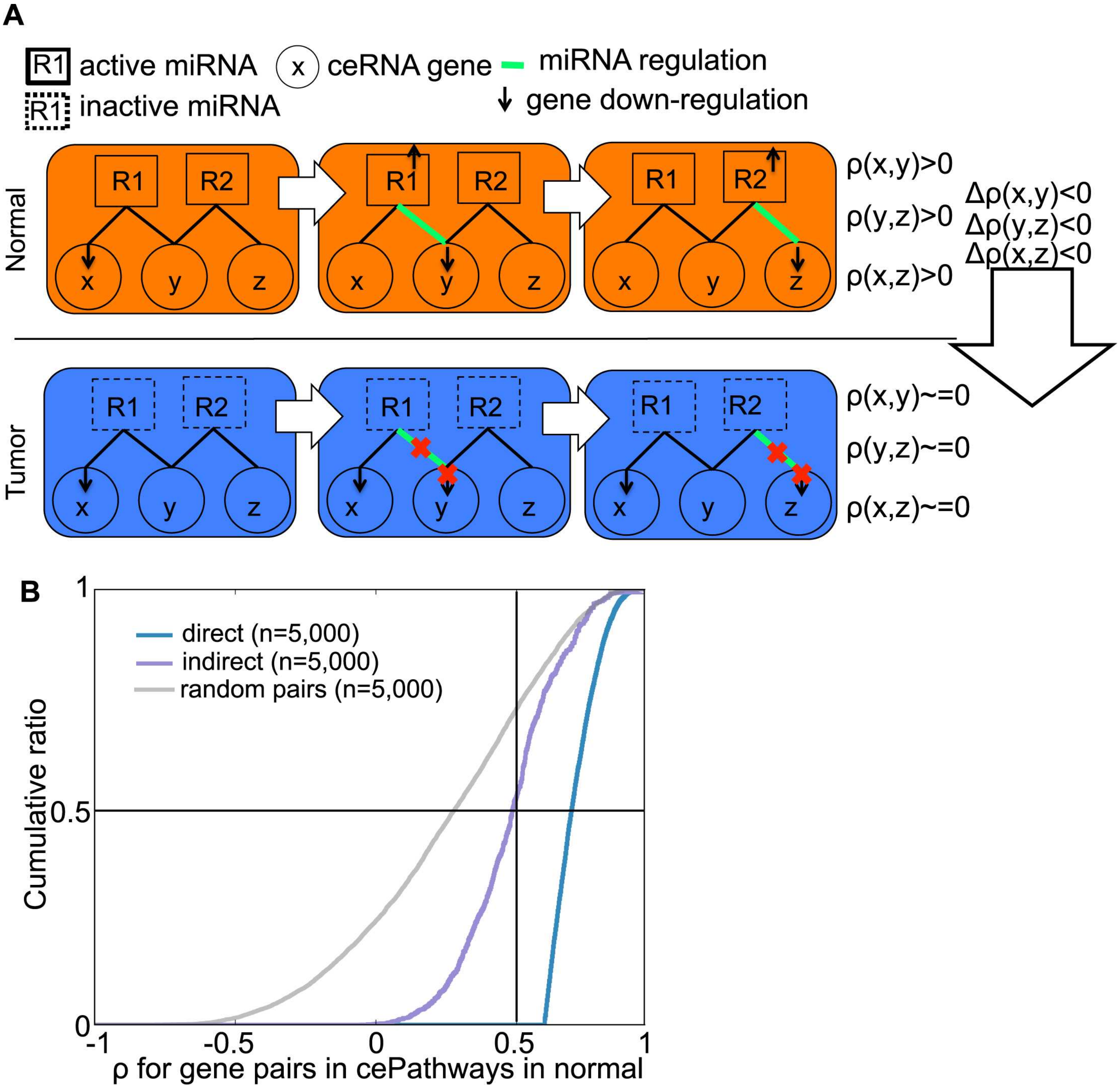

Fig 3. Identifying gained/lost indirect cePathway relationships (a). In normal when regulating miRNAs (R1 and R2) are active (in abundance relative to other competing molecules[21]), down-regulation of x would direct more copies of R1 to repress y. In turn, repressed y would direct R2 to repress z. Overall, ceRNA relationship on (*x*, *y*) and (*y*, *z*) would yield indirect ceRNA relationship on (*x*, *z*). In tumor when (*x*, *y*) lost ceRNA relationship possibly in association to inactivation of the regulating miRNAs, down-regulation of x would not repress y and z, representing the loss of not only direct ceRNA relationship in (*x*, *y*) and (*y*, *z*) but also indirect ceRNA relationship in (*x*, *z*). From normal and tumor condition, (*x*, *z*) would lose co-expression in tumors (Δρ<0) altogether with (*y*, *z*) and (*x*, *z*). **(b).** Accumulative distribution of co-expression (ρ) in direct and indirect ceRNA relationships (found in direct and indirect cePathway relationships) in normal samples, compared to the same number of random pairs (n = 5,000).

https://doi.org/10.1371/journal.pcbi.1006074.g003

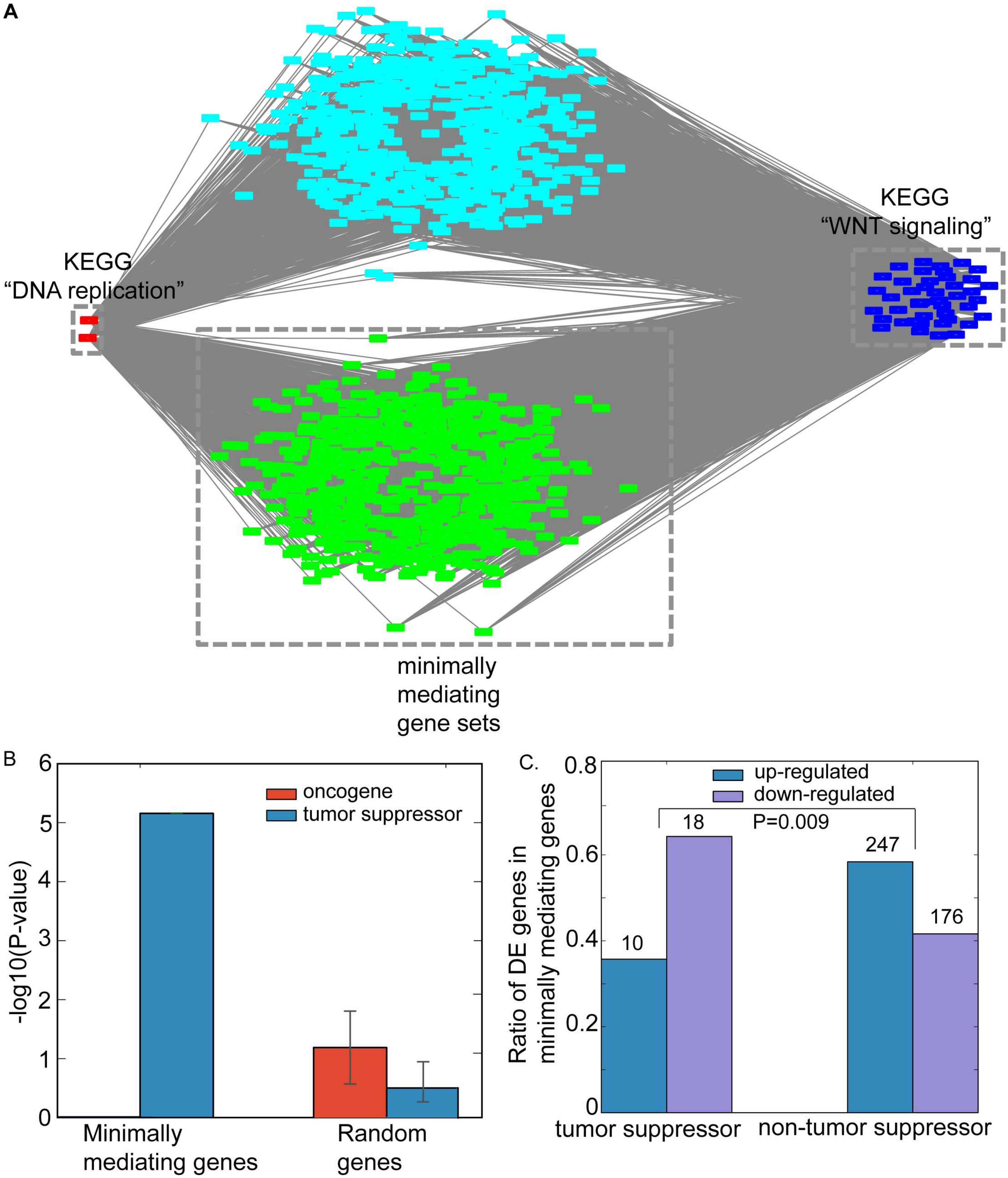
A
KEGG
"DNA replication"
KEGG
"WNT signaling"
minimally
mediating
gene sets
B
-log10(P-value)
0
1
2
3
4
5
6
oncogene
tumor suppressor
Minimally
mediating genes
Random
genes
C.
Ratio of DE genes in
minimally mediating genes
0
0.2
0.4
0.6
0.8
up-regulated
down-regulated
P=0.009
10
18
247
176
tumor suppressor
non-tumor suppressor

Fig 4. The lost indirect cePathway relationship between KEGG "DNA replication" and KEGG "WNT signaling". (a). Red nodes are ceRNA genes of KEGG "DNA replication" and blue nodes are those of KEGG "WNT signaling". 1,132 genes (green and light blue nodes) mediate the ceRNA loss between the pathways, where green nodes are 615 minimally mediating gene set. Visualization is taken from Cytoscape[27]. **(b).** Enrichment of the minimally mediating genes to oncogenes (red) and tumor suppressors (blue) compared to the same number of random genes. Error bar represents standard variation values from 100 random trials. **(c).** Ratio of differentially expressed tumor suppressors in the set of the mediating genes compared to other genes in the set. EdgeR estimates differential expression with FDR-corrected P value < 0.05.

https://doi.org/10.1371/journal.pcbi.1006074.g004

mediating genes are more likely to be down-regulated than other non-tumor-suppressor minimally mediating genes (P-value = 0.009, **Fig 4C**). While tumor suppressor down-regulation is a hallmark of cancer[26], one of its tumorigenic effects may be to disrupt indirect cePathway relationships, for example between KEGG "DNA replication" and KEGG "WNT signaling".

YAP, TAZ, and vitamin D receptor promote or inhibit lung adenocarcinoma possibly through ceRNA

To ascertain its general applicability, we ran MACPath on 58 TCGA lung adenocarcinoma (LUAD) and their matched normal samples, identifying direct cePathway dysregulation (13,247 gains and 7,121 losses in tumor) (**S3 Fig**, **S4 Table**). Further analyses on the dysregulation provide new insights into some tumor-associated crosstalks in lung adenocarcinoma (AC). Reactome "YAP1 and TAZ stimulated gene expression" that gained the most number (339) of cePathway relationships is an example. Overexpression of YAP and TAZ promotes growth signals in cancer types including lung AC [28]. Their expressions were found to be regulated by miRNAs in diverse cancer types [29], but their post-transcriptional tumorigenic mechanisms remain elusive. Since Reactome "YAP1 and TAZ stimulated gene expression" gains cePathway relationships with several tumor-associated growth factor pathways (e.g. EGFR, TGFβ and IGF), YAP and TAZ would promote the growth factor pathways possibly through ceRNA relationships. Further, cePathway dysregulations suggest their potential roles as prognostic biomarkers. As an example, vitamin D receptor pathways gained many cePathway relationships (5^{th} and 11^{th} highest in the number of gains). Vitamin D receptor is known to regulate cell cycle [30], whose high expression correlating with longer survival through antiproliferative effects and cell cycle arrest [31]. Since its function as cell cycle regulator involve miRNAs in diverse types of cancers [30], vitamin D receptor pathways gaining cePathway relationships to 23 cell cycle checkpoint pathways may enhance the antiproliferative and cell cycle arrest mechanisms at the ceRNA level.

Discussion

CeRNA plays an important role in the pathogenesis of cancers by forming extensive ceRNA network (ceRNET)[32] involving thousands ceRNA genes[21, 33]. For the ceRNETs to produce appropriate tumor-associated responses, the ceRNA genes are expected to coordinate to function in the level of biological pathways. Therefore, identifying ceRNA relationship between biological pathways (cePathway) would be a useful tool to understand tumor-associated ceRNA functions as a whole. For this purpose, we developed a computational method MACPath that identifies pathway pairs enriched in a differential network that encodes the ceRNA relationship changes (gain/loss) in tumor (cePathway relationship gain/loss).

A few computational methods have been proposed to identify pathway pairs enriched in a differential network[34, 35], which can be adapted to identify direct cePathway relationship dysregulation. However, they are not directly applicable for indirect cePathway relationships, since they do not consider indirect relationships. To the best of our knowledge, MACPath is the first computational method that can comprehensively identify both direct and indirect cePathway relationship dysregulation in a unified computational platform.

Materials and methods

MiRNA binding site database

Predicted miRNA binding sites were obtained from TargetScanHuman version 6.2[36]. For accurate analyses, those with preferentially conserved targeting score (Pct) more than 0 were used. Experimentally validated miRNA binding sites were obtained from TarBase version 5.0 [37], miRecords version 4[38], and miRTarBase version 4.5[39]. The binding sites found in indirect studies such as microarray experiments and high-throughput proteomics measurements were filtered out[40]. Another source is microRNA target atlas composed of public AGO-CLIP data[41]. As suggested, we employed only significant binding sites (q-value $<$ 0.05). The predicted and validated binding site information was combined for use[40]. Note that all miRNAs are converted to miRNA family ID[36], and representative miRNA names for the families (appeared first in the ID) are presented.

TCGA breast tumor RNA-seq data

Quantified gene expression files (RNASeqV1) for 97 primary breast tumors (TCGA sample code 01) and their matching normal samples (TCGA sample code 11) were downloaded from TCGA Data Portal[42]. Tumor samples are composed of various subtypes. 10,868 expressed RefSeq genes (FPKM $\geq$ 1 in $>$ 80% of all samples) were selected for downstream analyses. To better quantify gene expression in the presence of 3′ UTR-shortening, we only counted reads in the coding regions, approximated by taking maximum CDS region for each exon across different CDS annotations of multiple transcripts. Exon and CDS annotation for TCGA data were downloaded from Sage Bionetworks' Synapse database system.

CeRNA relationship change network

Given gene expression levels in normal and tumor samples, MACPath first builds ceRNA relationship change network in the following.

Definition 1. CeRNA relationship change network G is an undirected graph (V, E), where V is a set of genes and $E \subset V \times V$ represents gene pairs with significant miRNA binding site overlap (**S1A Fig**).

Since, for a gene pair with significant miRNA binding site overlap $(x,y) \in E$, co-expression increase/decrease in tumors would represent ceRNA relationship gain ($\Delta\rho_{tumor-normal}(x,y) > 0$) or loss ($\Delta\rho_{tumor-normal}(x,y) < 0$), G would encode ceRNA relationship change by holding the co-expression change value for $(x,y) \in E$. In building G for the breast tumor data (**S1B and S1C Fig**), we used only expressed mRNAs (FPKM>1) and miRNAs with moderate expression (FPKM between 0.01 and 100), since these are known to facilitate active ceRNA relationship [10]. After removing genes with less than 6 such miRNA binding sites, gene pairs with significant miRNA binding site overlap ($<$ 0.05 in B-H corrected p-value) were selected, where significance is estimated from hypergeometric test.

Identifying direct cePathway relationships gained/lost in tumors

Given the ceRNA relationship change network (**Def. 1**), MACPath assesses if pathway P and Q would gain or lose the cePathway relationship in tumors in the three steps.

Step 1. Estimating ceRNA changes. ceRNA relationship change between P and its neighbor gene y is measured.

Definition 2. $\theta_P(y) = \sum_{x\in P} \Delta\rho(x,y)$, if $(x,y) \in E$, where $\Delta\rho(x,y) = \rho_{tumor}(x,y) - \rho_{normal}(x,y)$.

$\theta_P(y)$ represents gain or loss of co-expression (and thus ceRNA relationship) in tumors between y and *P*; negative values represent loss and positive values represent gain. $\theta_P(y)$ is calculated for all genes connected to *P* in *G* and they are ranked by θ_P (**Fig 1C**).

Step 2. Calculating ceRNA relationship enrichment score. If pathway *Q* is not enriched in the neighbor genes, MACPath calls that they are not in a cePathway relationship. If pathway *Q* is enriched, we walk down the ranked list produced in Step 1, as a running sum is increased when we encounter a gene in *Q* and decreased when otherwise (**Fig 1C**). From the running sum, ceRNA relationship enrichment score ($CES_P(Q)$) is the maximum deviation from zero. To avoid yielding high scores for *Q* that is clustered near the middle of the ranked list, the amount of increase is set proportional to θ_P. However, the use of weighted increase (θ_P) can cause an asymmetric distribution of CES scores, complicating the interpretation of the scores [11]. Therefore, we followed Subramanian *et al.*[11] and considered genes with the negative and positive θ_P values separately for lost and gained cePathway relationships, respectively.

Step 3. Estimating significance of CES. From 2,000 permutations on the gene labels in the ranked list, p-value is the fraction of the permutations where the calculated CES_p values are greater than or equal to the observed CES_P.

For the TCGA breast tumor data, we selected significantly lost or gained cePathway relationships based on the CES values corresponding to top 5% of the empirical distribution (>0.6) and the multiple-testing corrected significance (B-H P value < 0.01).

GEO breast tumor microarray data

To validate our findings from the TCGA breast tumor data using an independent data set, we downloaded gene expression profiles on 32 human breast tissues of multiple cancer subtypes (including 19 Luminal A, 3 Luminal B, 3 Triple Negative) and 7 (non-matched) normal controls in microarray data (GSE57297) [19]. Although TCGA RNA-Seq data contain matched normal samples, this data do not. We used the normalized signal intensity of genes provided in the series matrix file.

Identifying indirect cePathway relationships gained/lost in tumors

To identify indirect cePathway relationship dysregulation in tumors, we define $\kappa_P(z)$, indirect ceRNA relationship change between pathway P and its indirect ceRNA partner *z* through mediating genes in the ceRNA relationship change network (**Def. 1**).

Definition 3. $\kappa_P(z) = \sum_{x\in P}\Delta\rho(x,z)$, if

i) $(x,y) \in E$ and $(y,z) \in E$ for some y,

ii) $\Delta\rho(x,y)$, $\Delta\rho(y,z)$, and $\Delta\rho(x,z)$ are in the same direction.

Since indirect ceRNA change between *x* and *z* would be propagated from ceRNA (x, y) and (y,z), the second condition filters out invalid indirect ceRNA changes. On the estimated indirect ceRNA changes using $\kappa_P(z)$, MACPath identifies indirect cePathway dysregulations by following **Step2** and **3** defined for direct cePathway relationship dysregulation.

Controlling false discovery rate in an adaptive manner

Between direct and indirect cePathways, it would be more difficult to identify true signals for indirect cePathways, because it involves another dimension of data, mediating RNAs. To identify true signals for indirect cePathways in a more sophisticated way, we tested for FDR control in an adaptive way[43]. Among adaptive methods showing higher statistical power, we selected ForwardStop[20] with $\alpha = 0.01$ to be stringent.

Pathway annotation in the analyses

We annotate pathways based on keyword search in the following way: A(_B) pathways are identified with keyword "A" (and "B") in the corresponding table. For example, cancer pathways are with keyword "CANCER". After collecting pathways with the keywords, manual inspection was conducted to filter out irrelevant pathways. Manual inspection indicated that the keyword-based identification mostly identified only relevant pathways.

Minimum number of mediating genes for indirect cePathway relationship dysregulation

Assuming that all mediating genes cover genes both in P and Q, this problem is identical to Minimum Set Cover Problem (MSCP)[44]. For multiple sets covering items in a universe, MSCP finds the minimum number of sets that covers the whole universe. With each mediating gene taken as a set covering genes in P and Q, seeking the minimum number of sets covering P and Q is identical to MSCP. Hence, this problem is also NP-hard. We formulate the problem into *integer linear programming*, in which the objective function and the constraints are given as follows.

We define a binary variable as follows:

B_{pm}, $\forall p \in (P \cup Q)$, $\forall m \in V - (P \cup Q)$. B_{pm} will take value 1 if $(p,m) \in E$. $B_{pm} = 1$ represents the case where gene $p \in (P \cup Q)$ is connected to m.

Then, the ILP program is:

Minimize |M|

Subject to $B_{pm} = 1, \forall p \in (P \cup Q), m \in M$.

To solve based on the ILP program, we used GNU Linear Programming Kit (GLPK) solver (version 5.1.3). Although the GLPK solver does not return a comprehensive set of solutions, permutation on the input values returns multiple solutions of the same size. Based on the multiple solutions, we conducted functional analysis (**Fig 3B and 3C**).

Tumor suppressors and oncogenes

The tumor suppressors and oncogenes used in this study were defined by the TUSON algorithm from genome sequencing of >8,200 tumor-normal pairs[45], namely residue-specific activating mutations for oncogenes and discrete inactivating mutations for tumor suppressors. TUSON is a computational method that analyzes patterns of mutation in tumors and predicts the likelihood that any individual gene functions as a tumor suppressor or oncogene. We ranked genes by their TUSON prediction P values from the most to the least significant and used the top 500 genes (P-value < 0.01) as the reference tumor suppressors or oncogenes. After removing 30 genes in common, 470 tumor suppressors and oncogenes were used for the enrichment analysis. Please note that there were very few breast tumor-specific tumor suppressors and oncogenes (36 and 3 with breast_q-value ≤ 0.5, respectively) and 90% of them were found in the top 500 pan-cancer predictions.

Supporting information

S1 Fig. CeRNA relationship change network for the TCGA breast tumor data. (a). Illustration of the pipeline to build the ceRNA relationship change network (**Def. 1**) for the TCGA

breast tumor data. **(b).** Connectivity (number of neighbor genes) of genes in the network against the number of miRNA binding sites residing in their 3′UTR. **(c).** Distribution of $\Delta\rho_{tumor-normal}$ values in the network. 63.8% of edges lose co-expression ($\Delta\rho < 0$) in tumors, marked by gray vertical line. Red line indicates $\Delta\rho$ value for the case of $r = 0.01$.
(TIF)

S2 Fig. Indirect cePathway relationships. (a). An illustration of invalid indirect ceRNA change. In tumor samples (column), gene x, y, z are expressed (white for lowly expressed and black for highly expressed). If they were not correlated at all in the matched normal samples, co-expression change (tumor vs. normal) of (x,y) and of (y,z) would be positive with their correlation gain in the half of the tumor samples. However, co-expression of (x,z) would not be gained in tumor (thus, not functional), because (x,y) and (y,z) were correlated in the other half of the tumor samples (red rectangles). **(b).** Average expression levels of genes across normal samples (y-axis) belonging to direct ceRNAs and equally correlated ($\rho > 0.6$) indirect ceRNAs (t-statistic -14.1 and P-value 8.04e^{-44}).
(TIF)

S3 Fig. Number of cePathway partners gained or lost in TCGA lung adenocarcinoma tumor samples vs. matched normal.
(TIF)

S1 Table. Information of direct cePathways gained or lost in the TCGA breast cancer samples (vs. matched normal samples).
(XLSX)

S2 Table. Information of indirect cePathways gained or lost in the TCGA breast cancer cells (vs. matched normal cells).
(XLSX)

S3 Table. 615 minimally mediating ceRNAs and 517 other mediating ceRNAs connecting the lost indirect cePathway relationship KEGG "DNA replication" and KEGG "WNT signaling".
(XLSX)

S4 Table. Information of direct cePathways gained or lost in the TCGA lung adenocarcinoma samples (vs. matched normal samples).
(XLSX)

Acknowledgments

We thank Benjamin Rodriguez in our laboratory for his valuable suggestions.

Author Contributions

Conceptualization: Hyun Jung Park, Wei Li.

Formal analysis: Hyun Jung Park, Soyeon Kim.

Funding acquisition: Wei Li.

Investigation: Hyun Jung Park, Soyeon Kim, Wei Li.

Methodology: Hyun Jung Park, Soyeon Kim.

Software: Hyun Jung Park.

Supervision: Wei Li.

Writing – original draft: Hyun Jung Park, Soyeon Kim.

Writing – review & editing: Wei Li.

References

1. Fabian MR, Sonenberg N. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. Nature Structural & Molecular Biology. 2012; 19:586–93. https://doi.org/10.1038/nsmb.2296 PMID: 22664986.
2. Broderick Ja, Zamore PD. Competitive endogenous RNAs cannot alter microRNA function in vivo. Molecular cell. 2014; 54:711–3. https://doi.org/10.1016/j.molcel.2014.05.023 PMID: 24905003.
3. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, et al. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell. 2011; 147:358–69. https://doi.org/10.1016/j.cell.2011.09.028 PMID: 22000014.
4. Wang Y, Xu Z, Jiang J, Xu C, Kang J, Xiao L, et al. Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal. Developmental Cell. 2013; 25:69–80. https://doi.org/10.1016/j.devcel.2013.03.002 PMID: 23541921.
5. Wang Y, Hou J, He D, Sun M, Zhang P, Yu Y, et al. The Emerging Function and Mechanism of ceRNAs in Cancer. Trends in Genetics. 2016; 32:211–24. https://doi.org/10.1016/j.tig.2016.02.001 PMID: 26922301.
6. Tay Y, Kats L, Salmena L, Weiss D, Tan SM, Ala U, et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell. 2011; 147:344–57. https://doi.org/10.1016/j.cell.2011.09.029 PMID: 22000013.
7. Sumazin P, Yang X, Chiu H-S, Chung W-J, Iyer A, Llobet-Navas D, et al. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. Cell. 2011; 147:370–81. https://doi.org/10.1016/j.cell.2011.09.041 PMID: 22000015.
8. Xu J, Li Y, Lu J, Pan T, Ding N, Wang Z, et al. The mRNA related ceRNA-ceRNA landscape and significance across 20 major cancer types. Nucleic Acids Research. 2015; 43:8169–82. https://doi.org/10.1093/nar/gkv853 PMID: 26304537.
9. Nitzan M, Steiman-Shimony A, Altuvia Y, Biham O, Margalit H. Interactions between distant ceRNAs in regulatory networks. Biophysical Journal. 2014; 106:2254–66. https://doi.org/10.1016/j.bpj.2014.03.040 PMID: 24853754.
10. Ala U, Karreth Fa, Bosia C, Pagnani A, Taulli R, Léopold V, et al. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:7154–9. https://doi.org/10.1073/pnas.1222509110 PMID: 23536298.
11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. 2005. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517
12. Gorsch SM, Memoli VA, Stukel TA, Gold LI, Arrick BA. Immunohistochemical staining for transforming growth factor beta 1 associates with disease progression in human breast cancer. Cancer Res. 1992; 52:6949–52. PMID: 1458485.
13. Masuda H, Zhang D, Bartholomeusz C, Doihara H, Hortobagyi GN, Ueno NT. Role of epidermal growth factor receptor in breast cancer. Breast Cancer Research and Treatment. 2012; 136:331–45. https://doi.org/10.1007/s10549-012-2289-9 PMID: 23073759.
14. Descamps S, Lebourhis X, Delehedde M, Boilly B, Hondermarck H. Nerve growth factor is mitogenic for cancerous but not normal human breast epithelial cells. The Journal of biological chemistry. 1998; 273:16659–62. https://doi.org/10.1074/jbc.273.27.16659 PMID: 9642218.
15. Terasawa K, Ichimura A, Sato F, Shimizu K, Tsujimoto G. Sustained activation of ERK1/2 by NGF induces microRNA-221 and 222 in PC12 cells. FEBS Journal. 2009; 276:3269–76. https://doi.org/10.1111/j.1742-4658.2009.07041.x PMID: 19438724.
16. Aktipis CA, Boddy AM, Gatenby RA, Brown JS, Maley CC. Life history trade-offs in cancer evolution. Nature reviews Cancer. 2013; 13:883–92. https://doi.org/10.1038/nrc3606 PMID: 24213474.
17. El Zein N, D'Hondt S, Sariban E. Crosstalks between the receptors tyrosine kinase EGFR and TrkA and the GPCR, FPR, in human monocytes are essential for receptors-mediated cell activation. Cellular Signalling. 2010; 22:1437–47. https://doi.org/10.1016/j.cellsig.2010.05.012 PMID: 20566383.
18. Chung J, Kubota H, Ozaki YI, Uda S, Kuroda S. Timing-dependent actions of NGF required for cell differentiation. PLoS ONE. 2010;5. https://doi.org/10.1371/journal.pone.0009011 PMID: 20126402.

19. Fu J, Allen W, Xia A, Ma Z, Qi X. Identification of biomarkers in breast cancer by gene expression profiling using human tissues. Genomics Data. 2014; 2:299–301. https://doi.org/10.1016/j.gdata.2014.09.004 PMID: 25396118.

20. G'Sell MG, Wager S, Chouldechova A, Tibshirani R. Sequential Selection Procedures and False Discovery Rate Control. arXiv preprint. 2013. https://doi.org/10.1111/rssb.12122

21. Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. Nature Reviews Genetics. 2016; 17:272–83. https://doi.org/10.1038/nrg.2016.20 PMID: 27040487

22. Karimaian A, Majidinia M, Bannazadeh Baghi H, Yousefi B. The crosstalk between Wnt/β-catenin signaling pathway with DNA damage response and oxidative stress: Implications in cancer therapy. DNA Repair. 2017; 51:14–9. https://doi.org/10.1016/j.dnarep.2017.01.003 PMID: 28108274

23. Willis N, Rhind N. Regulation of DNA replication by the S-phase DNA damage checkpoint. Cell division. 2009; 4:13. https://doi.org/10.1186/1747-1028-4-13 PMID: 19575778.

24. Song JL, Nigam P, Tektas SS, Selva E. MicroRNA regulation of Wnt signaling pathways in development and disease. Cellular Signalling. 2015; 27:1380–91. https://doi.org/10.1016/j.cellsig.2015.03.018 PMID: 25843779.

25. Zhang X, Xiao D, Wang Z, Zou Y, Huang L, Lin W, et al. MicroRNA-26a/b Regulate DNA Replication Licensing, Tumorigenesis, and Prognosis by Targeting CDC6 in Lung Cancer. Molecular cancer research: MCR. 2014:1535–46. https://doi.org/10.1158/1541-7786.MCR-13-0641 PMID: 25100863.

26. Hanahan D, Weinberg Ra. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–74. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230.

27. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13:2498–504. https://doi.org/10.1101/gr.1239303 PMID: 14597658.

28. Zanconato F, Cordenonsi M, Piccolo S. YAP/TAZ at the Roots of Cancer. Cancer Cell. 2016; 29:783–803. https://doi.org/10.1016/j.ccell.2016.05.005 PMID: 27300434.

29. Zhou X, Lei QY. Regulation of TAZ in cancer. Protein and Cell. 2016; 7:548–61. https://doi.org/10.1007/s13238-016-0288-z PMID: 27412635.

30. Fleet JC, Desmet M, Johnson R, Li Y. Vitamin D and Cancer: A review of molecular mechanisms James. Biochem Journal. 2012; 441:61–76. https://doi.org/10.1042/BJ20110744.Vitamin

31. Kim SH, Chen G, King AN, Jeon CK, Christensen PJ, Zhao L, et al. Characterization of vitamin D receptor (VDR) in lung adenocarcinoma. Lung Cancer. 2012; 77:265–71. https://doi.org/10.1016/j.lungcan.2012.04.010 PMID: 22564539.

32. Cheng D-l, Xiang Y-y, Ji L-j, Lu X-j. Competing endogenous RNA interplay in cancer: mechanism, methodology, and perspectives. 2015. https://doi.org/10.1007/s13277-015-3093-z PMID: 25604144

33. Jens M, Rajewsky N. Competition between target sites of regulators shapes post-transcriptional gene regulation. Nature Reviews Genetics. 2014; 16:113–26. https://doi.org/10.1038/nrg3853 PMID: 25488579

34. Hung J-H, Whitfield TW, Yang T-H, Hu Z, Weng Z, DeLisi C. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome biology. 2010; 11:R23. https://doi.org/10.1186/gb-2010-11-2-r23 PMID: 20187943.

35. Han J, Shi X, Zhang Y, Xu Y, Jiang Y, Zhang C, et al. ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. Scientific Reports. 2015; 5:13044. https://doi.org/10.1038/srep13044 PMID: 26267116.

36. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120:15–20. https://doi.org/10.1016/j.cell.2004.12.035 PMID: 15652477.

37. Papadopoulos GL, Reczko M, Simossis Va, Sethupathy P, Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of TarBase. Nucleic acids research. 2009; 37:D155–8. https://doi.org/10.1093/nar/gkn809 PMID: 18957447.

38. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic acids research. 2009; 37:D105–10. https://doi.org/10.1093/nar/gkn851 PMID: 18996891.

39. Hsu S-D, Tseng Y-T, Shrestha S, Lin Y-L, Khaleel A, Chou C-H, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic acids research. 2014; 42:D78–85. https://doi.org/10.1093/nar/gkt1266 PMID: 24304892.

40. Dvinge H, Git A, Gräf S, Salmon-Divon M, Curtis C, Sottoriva A, et al. The shaping and functional consequences of the microRNA landscape in breast cancer. Nature. 2013; 497:378–82. https://doi.org/10.1038/nature12108 PMID: 23644459.

41. Hamilton MP, Rajapakshe K, Hartig SM, Reva B, McLellan MD, Kandoth C, et al. Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. Nature communications. 2013; 4:2730. https://doi.org/10.1038/ncomms3730 PMID: 24220575.

42. Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M, et al. The UCSC Cancer Genomics Browser: update 2013. Nucleic acids research. 2013; 41:D949–54. https://doi.org/10.1093/nar/gks1008 PMID: 23109555.

43. Li A, Barber RF. Accumulation tests for FDR control in ordered hypothesis testing. arXiv:statME/150507352. 2015. PubMed PMID: arXiv:stat.ME/1505.07352.

44. Gens G, Levner E. Complexity of approximation algorithms for combinatorial problems: a survey. ACM SIGACT News. 1980; 12:52–65.

45. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. Cell. 2013; 155:948–62. https://doi.org/10.1016/j.cell.2013.10.011 PMID: 24183448.