3′-UTR shortening disrupts ceRNA crosstalk of housekeeping genes resulting in subtype-specific breast cancer development

Fan Zhenjiang[1], Soyeon Kim[2,3], Brenda Diergaarde[4,5], Hyun Jung Park[4†].

[1]Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

[2]Department of Pediatrics, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

[3]Division of Pulmonary Medicine, Children's hospital of Pittsburgh UPMC, Pittsburgh, Pennsylvania, United States

[4]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States

[5]Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania,

USA

[†] Corresponding author: hyp15@pitt.edu (H.J.P.)

ABSTRACT

Alternative polyadenylation (APA) is a post-transcriptional mechanism that regulates gene expression. In human cancer, shortening of the 3′-untranslated region (3′-UTR) through APA is widespread affecting thousands of genes[1]. We previously identified that 3′-UTR shortening (3′US) disrupts the competing-endogenous RNA (ceRNA) network (3′US-ceRNA effect) to promote breast cancer[2]. As different breast cancer subtypes are associated with different molecular mechanisms[3], we identified distinct 3′US profiles of different breast cancer subtypes in this work, calling for the characterization of subtype-specific 3′US-ceRNA effect in the ceRNA network. A quantitative challenge is that different sample sizes available for the different breast cancer subtypes can result in a systematic bias on size and topology of the constructed ceRNA networks. We addressed the bias by normalizing the networks in two-way, first between and second within the subtypes. Using the two-way network normalization, we built comparable ceRNA networks for estrogen receptor negative (ER-) and positive (ER+) subtype breast tumor samples of different size. Functional enrichment analyses associated subtype-specific 3′US-ceRNA effect with ER-'s aggressive phenotype[4] and unique growth mechanism[5]. Especially, for ER- specific growth mechanism, subtype-specific 3′US-ceRNA effect disrupts ceRNA crosstalk of housekeeping genes, which help maintain similar ceRNA network topology for ER- and ER+ normal samples. As ER- specific 3′US-ceRNA effect is associated with ER-'s pathological features, aggressive phenotype and unique growth mechanism, our study provides new insights into the interactive mechanism of 3′US and ceRNA for ER- specific cancer progression.

INTRODUCTION

Approximately, 70% of human genes contain multiple polyadenylation (polyA) sites in the 3′-untranslated region (3′-UTR)[6]. Through alternative polyadenylation (APA) during transcription, messenger RNAs (mRNA) from the same gene can have various lengths of 3′-UTR. Since 3′-UTR contains regulatory regions such as microRNA (miRNA) target sites, mRNAs with shortened or lengthened 3′-UTRs may effectively diversify transcriptomic dynamics in diverse pathological conditions such as cancer[7]. In human cancer, 3′-UTR lengthening (3′UL) was associated with cell senescence[8] with implication for tumor-associated processes such as cell cycle inhibitors, DNA damage markers, and tumor suppressors[9]–[12]. Widespread 3′-UTR shortening (3′US) was directly implicated for oncogene activation in cell line experiments[6]. Further, some 3′US genes demonstrate an additional prognostic power beyond common clinical and molecular covariates[1] and are associated with drug sensitivity[13]. These results suggest that APA events, both 3′-UTR shortening and lengthening, play important roles in tumor etiology and response to treatment.

3′-UTR is also implicated in competing-endogenous RNA crosstalk (ceRNA)[14] that co-regulate each other RNAs by competing to bind shared microRNAs (miRNA). When 3′-UTR shortening genes lose miRNA target sites in the 3′-UTR, the associated miRNAs bind to the 3′-UTR of the ceRNA partners, which would be competing for the binding of the miRNAs. As a result, 3′-UTR shortening disrupts ceRNA crosstalk (3′US-ceRNA effect). We recently reported that 3′US-ceRNA effect globally down-regulates tumor suppressors, promoting human cancer, including breast cancer[2].

Human cancers can be divided into different subtypes based on molecular and/or clinical features for more accurate treatment plans and prognosis. For example, breast cancer can be strictly classified into two major subtypes based on estrogen receptor (ER) status, a central component of the pathological evaluation of breast cancer[15]. Estrogen receptor negative (ER-) breast tumors have unique molecular dynamics compared to estrogen receptor positive (ER+) breast tumors including the unique growth mechanism. ER+ tumors can be treated with endocrine therapy, blocking ER activity or depleting estrogen levels, however, this therapeutic approach does not have efficacy in ER- breast tumors due to their difference in growth mechanism. In that sense, ER- breast tumors show worse prognosis than ER+ breast tumors[16] with more aggressive phenotype[4], [17].

To develop targeted therapies that effectively treat ER- breast cancer, its molecular dynamics needs to be understood comprehensively, especially for its aggressive phenotype and unique growth mechanism. With the profound tumorigenic effect of 3′US-ceRNA[2], we now hypothesize that 3′US-ceRNA effect specific to ER- breast tumor contributes to its unique molecular feature, aggressive phenotype and unique growth mechanism. To test this hypothesis, we compared ER- ceRNA networks with ER+ with regards to 3′US.

Global APA events differ between ER+ and ER-

To study the role of 3′-UTR shortening genes for subtype-specific breast tumor, we divided 97 breast tumor and the matched normal samples available in TCGA (see Methods) into 77 ER-positive (ER+) and 20 ER-negative (ER-) sample pairs. Using DaPars[1], we identified 3′UTR shortening (3′US, ΔPDUI < -0.2 and FDR < 0.05) and 3′UTR lengthening (3′UL, ΔPDUI > 0.2 and FDR < 0.05) genes. As there are totally 5,876 3′US and 5,379 3′UL genes, both 3′US and 3′UL events are widespread in both subtypes (**Fig. 1A**). In each normal/tumor sample, 3′US and 3′UL events occur equally highly on genes (**S. Fig. 1A, 1B**). To identify common mechanisms across tumor samples, we further identified recurrent 3′US and 3′UL genes (occurring in > 20% of the tumor samples[1], **S. Fig. 1C, 1D**). Although there are similar numbers of 3′US and 3′UL genes in terms of total number and in each sample pair (**S. Fig. 1A, 1B)**, more 3′US genes recur (occurring in > 20% of the sample pairs[1]) than 3′UL genes (**Fig. 1D, E** e.g. $P=5.0\times10^{-5}$ for ER+). The less recurrence of 3′UL genes partially explains why previous identifications, which focused on recurrent events[1], [13], observed less 3′UL genes than 3′US. Further analyses showed that 3′US and 3′UL play distinct roles for cancer. First, the recurrent 3′US and 3′UL genes show little overlap both for ER+ and ER- (1 and 13 genes in common for ER+ ($P=1.27e^{-6}$) and ER- ($P=3.97e^{-9}$), respectively, **Fig. 1D, 1E**). Second, the number of 3′UL events is not correlated with that of 3′US events across the tumor samples (P=0.35 for ER+ and P=0.61 for ER-, **Fig. 1B, 1C**). Third, IPA pathway analysis (**S. Fig. 1E**) shows that the recurrent 3′US and 3′UL genes are enriched for distinct sets of pathways in ER+ and ER- tumor samples. With our interest in the common mechanism of APA, we will focus on the function of recurrent 3′US genes.

Two-way Pairwise Normalization of ER+ and ER- ceRNA network

To study subtype-specific 3′US-ceRNA effect, we set out to compare the ceRNA networks from ER+ and ER- tumor and normal samples. While a network can be defined as a set of edges between genes, ceRNA networks can have edges between the genes that share a significant number of microRNA (miRNA) target sites and whose expression levels across samples are correlated (co-express)[2], [18]. For gene pairs that share a significant number of miRNA target sites (FDR < 0.05 based on hypergeometric test e.g. [2], [18]), co-expression cutoffs needs to be determined to build comparable ceRNA networks for ER+ and ER-. Using the common co-expression cutoff (e.g. Pearson's $\rho > 0.6$) would inflates the number of edges for ER- compared to ER+ (160,687 in ER- normal vs. 88,275 in ER+ normal, **Fig. 2A**). Our simulation study (using cutoff of Pearson's ρ 0.6) further confirmed that the network size difference (in terms of the number of edges) is attributable to the sample size difference. When smaller numbers of ER+ normal samples are used to construct the ceRNA network, the network size increase (**S. Fig. 2A**), consistent to the trend for ER- and ER+ network size (**Fig. 2A**). Further, the simulation shows that the actual size of ER- normal network (with 20 samples) falls in the non-outlier range of ER+ normal network size when 20 samples are used.

Although the network size difference should make a systematic bias in network comparison, it is not straightforward to address the network size difference. One might want to sample the number of ER+ normal samples to the same number of samples available for ER- (n=20) to remove the bias. Then, the ceRNA networks constructed from the subsample lose topological consistency within them (**Fig. 2B**), making it difficult to represent ER+ ceRNA dynamics. Also, if ER- ceRNA network is constructed using the expression correlation cutoff for the same statistical significance to ER+ (based on a permutation test, see Methods), it will drastically deflate the number of edges (**Fig. 2C**), making another systematic bias for comparison. To address this issue, we first construct the reference network from normal samples of larger size (ER+) using the common correlation cutoff (Pearson's $\rho > 0.6$). Based on the assumption that normal samples would have similar molecular dynamics between ER+ and ER-, we seek to find an expression correlation cutoff for ER- normal network that makes most topological similarity to the ER+ reference network. To estimate topological similarity, we employed normalized Laplacian Matrix Eigenvalue Distribution that discovers ensembles of Erdős–Rényi graphs better than other metrics such as Sequential Adjacency or Laplacian[19] (see Methods). While ER- normal network topology changes drastically by different correlation cutoff values (**S. Fig. 2B, 2C**), we found that cutoff of 0.68 makes ER- normal network most similar to the ER+ reference network (2,846 and 2,864 nodes for ER+ and ER-, respectively **Fig. 2D**). Correlation cutoff 0.68 is supported again when normal ER- network with the correlation cutoff makes the closest average clustering coefficient to the reference network, another measure of network similarity[20] (0.39 for the ER- network with cutoff 0.68 and 0.40 for the reference ER+ network, **S. Fig. 2D**). We further applied the subtype-specific cutoff (0.68 for ER- and 0.6 for ER+) to build tumor ceRNA networks (1,392 and 2,039 nodes for ER+ and ER-, respectively). Since this method normalizes ceRNA networks across different subtypes of normal samples (by identifying correlation cutoff for topological similarity) and within each subtype (between normal and tumor), we call it two-way network normalization.

3′UTR shortening is associated with ER- tumors' aggressive metastatic phenotypes in ceRNA.

To identify ER- specific function of 3′US-ceRNA effect, we compared ER- and ER+ ceRNA networks after two-way normalization. Among 1,783 ceRNA partners of 521 3′US genes (3′US ceRNA partners) in normal ER- ceRNA network, 498 (27.9%) are only in ER- (ER- 3′US ceRNA partners), whereas 1,285 (72.1%) are also in ER+ as 3′US ceRNA partners (common 3′US ceRNA partners, **Fig. 3A**). We found that 118 IPA canonical pathways significantly ($P < 0.01$) enriched for the ER- 3′US ceRNA partners (**S. Table 2**) are linked with several aspects of ER- specific tumor phenotypes (**Fig. 3B**). The first aspect of the pathways are "cancer" pathways (pathways with keyword "cancer"). For example, "Molecular Mechanisms of Cancer" pathway ($P=10^{-5.25}$) is, according to IPA knowledgebase, a set of genes whose disruptions have been shown to drive tumor progression. Specific to breast cancer, the enrichment to "Breast Cancer Regulation by Stathmin1" ($P=10^{-3.92}$) pathway is interesting, since overexpression of Stathmin1 correlates with the loss of ER [21] and with aggressive phenotypes[22] of breast tumor. The second category of pathways underlies the aggressive metastasis of ER- tumors more directly. For example, 8 pathways were experimentally confirmed in association with breast tumor metastasis [23], and 5 of them are significantly enriched for ER- 3′US ceRNA partners with an exception of PI3K/AKT, whose enriched p-value is just below our cutoff ($P=10^{-1.95}$). Also, previous studies associated breast tumor malignancy and poor survival with abnormal control of Ephrin A (reviewed in [24]), which is enriched for ER- 3′US ceRNA partners ($P\text{-val}=10^{-5.05}$). Together, ER- specific 3′US ceRNA partners control pathways for cancer signaling and aggressive metastatic phenotypes in normal samples. However, in ER- tumors, 81.7% of 3′US ceRNA partners lost the ceRNA relationship (**S. Fig 3A)**, likely losing the normal control for the metastatic phenotypes. Further, as ER+ tumors also lost the 3′US ceRNA partners (95.4%, **S. Fig 3B**), ER- and ER+ share less 3′US ceRNA partners in tumor (35 of 416 (8.41%) 3′US ceRNA partners in ER- shared with ER+ (**Fig. 3C**)). Altogether, ER- specific loss of 3′US ceRNA partners can interrupt cancer signaling and aggressive metastasis pathways for ER- tumors.

Housekeeping genes keep normal ceRNA networks similar between ER- and ER+.
To identify the role of different gene classes for the ceRNA network dynamics, we first identified housekeeping (HK), tumor-associated (tumor suppressors or oncogenes, TA), and transcription factor (TF) genes in the ceRNA networks. Out of 3,804 HK[25], 932 TA[26], and 1,020 TF genes[27] curated in public databases (see Methods), the ceRNA networks consist of 3-fold more HK genes than TA or TF genes (**Fig. 4A** for normal and **S. Fig. 4A** for tumor). Expectedly, HK genes form 5 ~ 7 folds more edges than the other gene classes in both normal and tumor ceRNA networks for ER+ and ER- subtypes (**S. Fig. 4B**). Due to its active role in cell maintenance[25], HK genes are expected to maintain constant expression levels in most physiological conditions[25]. Consistent to the expectation, 958 HK genes on the ER- normal network express as highly as (**S. Fig. 4C**) and with a significantly less variation across the samples than ($P=1.72e^{-54}$, **Fig. 4B**) 1,906 non-HK genes. Together with the fact that the HK genes contain more miRNA binding sites than other genes in the 3′UTR (P=0.05, **Fig. 4C**), they would work as stable sponges for miRNAs[28]. ER- and ER+ normal networks share a very significant number of HK genes ($P=8.77e^{-771}$, **Fig. 4D**), leading us to hypothesize that HK ceRNA partners keep the normal ceRNA networks in similar topology between ER- and ER+. To test the hypothesis, we estimated the similarity between the sub-network of ER+ and ER- ceRNA networks consisting only of HK gene nodes and compared the similarity with those between sub-networks of ER+ and ER- consisting of non-HK ceRNA network nodes (sampled to the same number of HK genes). Since the subnetworks of HK genes are significantly more similar between the subtypes ($P < 0.01$), the results suggest a novel important role of HK genes to keep the normal ceRNA networks similar to each other.

3′US disrupts ceRNA crosstalk of housekeeping genes.
Further analyses suggest that 3′US indirectly disrupts the stable ceRNA crosstalk of HK genes. Out of 958 HK genes on the ER- normal network, 727 genes (75.8%) are connected to 3′US genes (3′US HK ceRNA partners), which is in the same scale as the other classes of genes that are known to be regulated by 3′US genes[2], [29] (61.8% from 317 TA genes and 90.2% from 271 TF genes). Additionally, such HK genes are connected to as many 3′US genes as the other classes of genes are (**Fig. 5A**). Compared to 231 HK genes on the ER- normal network that are not connected to 3′US genes, those that are ceRNA partners of 3′US genes (3′US HK ceRNA partners) are more highly connected in the network. The high connectivity of HK genes suggests their important roles for ER- normal ceRNA network (**Fig. 5B**). Previously, we showed that 3′US represses genes in tumor if they were the ceRNA partners in normal [2]. Hence, repression of 3′US HK ceRNA partners in tumor **(Fig. 5C)** signifies that they are indeed in ceRNA relationship with 3′US genes in normal. Simulation studies and cell line experiments have shown that ceRNA relationships propagate through the ceRNA network[18], [30]. Furthermore, ceRNA relationship changes, either loss or gain, between samples of conditions (e.g. tumor vs. normal) also could propagate[31]. Thus, when the ceRNA relationship of HK genes is disrupted in tumor due to 3′US, the disrupted ceRNAs should further disrupt the ceRNA relationship with their ceRNA partners (**S. Fig. 5A**). This indirect loss of ceRNA relationship due to 3′US disrupts

stable ceRNA crosstalk of HK genes and their role in the normal ceRNA network, as 727 3′US HK genes lost higher ratios of ceRNA partners in tumor (**Fig. 5D**).

CeRNAs of housekeeping genes in ER- tumor are associated with ER- specific growth.

With ceRNA networks reduced in tumor due to 3′US[2] and miRNA expression decrease[32], HK genes lost the ceRNA relationship in both ER+ and ER- (972 of 1,635 (59.4%) lost in ER- and 1,330 of 1,688 (78.8%) in ER+, **S. Fig. 5B**). As a result, HK ceRNA partners highly overlapping between ER+ and ER- normal samples (**Fig. 5A**) become specific to each tumor subtype (**Fig. 6B**). To assess if HK ceRNA partners specific to ER- tumor play important roles, we conducted functional enrichment analysis on the 505 and 144 HK ceRNA partners unique to ER- and ER+ tumor, respectively. While it is known that cell growth-related pathways and cell cycle-related pathways are differently regulated in the subtypes[33]–[35], our analysis shows that ceRNAs of HK genes specific to each subtype are enriched for cell growth- and cell cycle-related pathways, suggesting their role on subtype-specific molecular processes. First, we found that 505 HK ceRNA partners specific to ER- tumor are enriched for pathways associated to growth factor (with keyword "GF") (**S. Table 3**). Especially, EGF (P-val=$10^{-2.99}$) activates cell cycle progression in ER- tumors[36], and expression of VEGF (P-val=$10^{-2.42}$) is associated to ER- tumors[37]. Also, both EGF and VEGF are suspected to proliferate ER- tumors when estrogen cannot sustain them[37]. Second, cell cycle pathways are enriched for ER+ specific HK ceRNA partners, suggesting that ER-regulated cell cycle[38], [39] differentiates ER+ and ER- cancer partially at the ceRNA level. Especially, since regulation of cell cycle, G1- and S-phase and their transition ratio, is crucial for ER+ tumor's proliferation (reviewed in [40]), it is interesting that cell cycle regulation pathways for various phases (G1/S or G2/M) of various mediators (Estrogen or Cyclins) are enriched with 144 ER+ HK ceRNA partners. Third, considering that the enrichment analysis was for the disjoint sets of genes (505 unique to ER- and 144 unique to ER+), it is interesting that unique HK ceRNA partners of both subtypes are significantly enriched for some "cancer" pathways e.g. "Molecular Mechanisms of Cancer", because it shows that HK ceRNAs are involved in cancer mechanisms equally significantly but in a subtype-specific fashion.

DISCUSSION

To investigate the role of 3′US-ceRNA effect [2] for estrogen receptor negative (ER-) breast tumors vs. ER+, we constructed ceRNA networks for ER+ and ER- subtype comparable to each other by addressing the bias owing to the different number of samples (72 for ER+ and 20 for ER-). Comparison of the networks suggests that 3′US disrupts the ceRNA network for ER- tumors' aggressive phenotypes. Further, we revealed the role of 3′US-ceRNA effect on housekeeping (HK) genes. Although HK genes highly and stably express in diverse biological contexts[41], our understanding of their roles is limited, especially with regards to ceRNA. For the first time, we found their role in keeping normal ceRNA networks similar between the subtypes.

Further analysis shows that 3′US indirectly disrupts ceRNA crosstalk of HK genes for ER- specific growth mechanism. Indirect ceRNA crosstalk propagates ceRNA effects through the ceRNA networks, demonstrated in simulation studies[18], cell line experiments[30], and

TCGA breast cancer data[31]. In this paper, we identified an indirect ceRNA effect of 3′US-ceRNA for the first time, where 3′US-ceRNA indirectly disrupts ceRNA crosstalk of HK genes in tumor.

Identifying ER-'s aggressive metastasis and unique growth pathways in ceRNA networks also indicates a clinical potential regarding miRNA therapeutics. For example, "Breast Cancer Regulation by Stathmin1" ($P=10^{-3.92}$) pathway, whose expression is associated with ER- tumors[21] and with the aggressive phenotypes[22], are disrupted by 3′US-ceRNA effect directly (**Fig. 3B**) and indirectly (**Fig. 6C**) through HK genes. Since 3′US-ceRNA effect is mediated by miRNAs[2], [31], treating ER- tumors with microRNAs involved in the effects is expected to mitigate ER-'s aggressive phenotype.

In network analysis, the network of interest is often compared to the reference network. However, if the networks are constructed from different numbers of samples, the comparison will be confused due to the sample size difference. Based on a biological assumption that normal samples would share a similar size of molecular interactions, we determined the subtype-specific cutoff value for normal ceRNA networks and apply the cutoff value to construct tumor ceRNA network (two-way pairwise normalization). As the resulting ceRNA networks facilitate novel discoveries on the subtype-specific 3′US-ceRNA effect, we expect that the two-way pairwise normalization method can further help normalize biological networks constructed with the different number of samples if the matched normal samples are available.

METHODS

TCGA breast tumor RNA-seq data and identification of breast cancer subtypes.

Quantified gene expression files (RNASeqV1) for primary breast tumors (TCGA sample code 01) and their matching solid normal samples (TCGA sample code 11) were downloaded from the TCGA Data Portal[42]. We used 97 breast tumor samples that have matched normal tissues, which were further categorized into 77 estrogen receptor positive (ER+) and 20 estrogen receptor negative (ER-). For ER+ and ER-, we collected both normal (ER+ normal and ER- normal) and tumor (ER+ tumor and ER- tumor) samples. A total of 10,868 expressed RefSeq genes (fragments per kilobase of transcript per million mapped reads (FPM) ≥ 1 in > 80% of all samples) were selected for downstream analyses.

Selection of miRNA target sites

Predicted miRNA-target sites were obtained from TargetScanHuman version 6.2[43]. Only those with a preferentially conserved targeting score (Pct) more than 0 were used[1]. Experimentally validated miRNA- target sites were obtained from TarBase version 5.0[44], miRecords version 4[45] and miRTarBase version 4.5[46]. The target sites found in indirect studies such as microarray experiments and high-throughput proteomics measurements were filtered out [47]. Another source is the microRNA target atlas composed of public AGO-CLIP data[48] with significant target sites (q-value < 0.05). The predicted and validated target site information was then combined to use in this study.

Statistical significance of Pearson correlation coefficient

The implementation of the Pearson r function is provided by a python package, SciPy, and available at https://scipy.org/, which returns the calculated correlation coefficient and a 2-tailed p-value for testing non-correlation. The Pearson correlation coefficient measures the linear relationship between two variables (e.g. gene X and gene Y) and when the two covariates follow binormal distribution, we can assume that their Pearson's correlation follows student t distribution. The p-value is calculated by three steps: 1) calculating the value of the Pearson's correlation t, 2) defining the degree of freedom df (N-2, where N is the sample size), 3) getting the probability of having t or more extreme than t from a Student's t-distribution with the degrees of freedom df.

Housekeeping, transcription factor and tumor-associated genes

Housekeeping genes are required for the maintenance of basic cellular functions that are essential for the existence of a cell, regardless of its specific role in the tissue or organism. Generally, housekeeping (HK) genes are expected to be expressed at relatively constant rates in most non-pathological situations[41]. We used 3,804 HK genes defined in RNA-Seq data for 16 normal human tissue types: adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells[25].

Transcription factors (TFs) play an important role in the gene regulatory network. We used 2,020 TF genes defined in TFcheckpoint database[27], in which TF information is collected from 9 different resources followed by manual inspections for sequence-specific DNA-binding RNA polymerase II TF.

The tumor-suppressor genes and oncogenes were defined by the TUSON algorithm from genome sequencing of > 8,200 tumor/normal pairs[26], in particular residue-specific activating mutations for oncogenes and discrete inactivating mutations for tumor-suppressor genes. TUSON computationally analyzes patterns of mutation in tumors and predicts the likelihood that any individual gene functions as a tumor- suppressor gene or oncogene. We used 466 oncogenes and 466 tumor suppressor genes at the top 500 in each prediction (after subtracting 34 genes in common).

Building subtype ceRNA networks

For each of the breast cancer data (ER+ normal, ER+ tumor, ER- normal, and ER- tumor) that we defined above, we constructed a ceRNA network based on microRNA (miRNA) target site share and expression correlation[2], [18]. The same miRNA target site information was determined regardless of the subtypes, resulting into a miRNA target site share network (based on FDR > 0.05 in hypergeometric test with miRNA target site information). And given the same miRNA target site share network, the expression correlation information for each subtype will select ceRNA network edges for each subtype.

We first constructed the ER+ normal reference ceRNA network by applying a traditional correlation cutoff (>=0.6) on the miRNA target site share network. Then, to identify ER- normal ceRNA network comparable to ER+ normal reference ceRNA network, we applied different correlation cutoff values (0 to 1 with a step size of 0.01) on the miRNA target site share network for ER- normal samples, and select the correlation cutoff values that makes ER- normal ceRNA network most similar to ER+ normal reference ceRNA network. To estimate topological similarity, we employed normalized Laplacian Matrix Eigenvalue Distribution that discovers ensembles of Erdős–Rényi graphs better than other metrics such as Sequential Adjacency or Laplacian[19]. After identifying the ER+ normal reference network and the corresponding ER- normal network, we used the same cutoffs (0.6 for ER+ subtypes and 0.68 for ER- subtypes) to construct the ER+ tumor network and the ER- tumor network, respectively.

Estimating topological similarity

To identify the structural equivalence between two networks, we employed spectral analysis not only to identify the structural similarities, but also to track down the underlying dynamic behavior changes between them. Spectral clustering on networks uses the eigenvalues of several matrices, such as adjacency matrix, the Laplacian matrix, the normalized Laplacian matrix. In this research, we used the normalized Laplacian matrix since it involves both the degree matrix and adjacency matrix, where the degree matrix can identify the node related equivalence of networks and the adjacency matrix can capture the structural equivalence of networks. Another very important reason of using the normalized Laplacian eigenvalue matrix is that it is more sensitive to small changes because it considers more information[17].

For network G, the normalized Laplacian of G is the matrix:

$$N = D^{-1/2} - LD^{-1/2} \tag{1}$$

where L is the Laplacian matrix of G and D is the degree matrix. The Laplacian matrix L is defined as: $L = D - A$, where A is the adjacency matrix of G.

In N, each of its entry elements is given by:

$$N_{i,j} = \begin{cases} 1, & if\ i = j\ and\ \mathrm{degree}(v_i) \neq 0 \\ -\dfrac{1}{\sqrt{\mathrm{degree}(v_i)\,\mathrm{degree}(v_j)}}, & if\ i \neq j\ and\ v_i\ is\ adjacent\ to\ v_j \\ 0, & otherwise \end{cases} \tag{2}$$

where degree (vertex v) is the function that return the degree of the vertex v.

To assess how close two eigenvalue distributions of network G_1 and G_2 are, we used the Kolmogorov–Smirnov test (KS test), which is defined as:

$$K_{n,m} = \sup_x |dist_{1,n}(x) - dist_{2,m}(x)| \tag{4}$$

where $dist_{1,n} and\ dist_{2,m}$ are the empirical distribution functions of the first and the second eigenvalue distribution respectively, and $\sup_x$ is the supremum of the set of distances.

By using the normalized Laplacian Matrix and KS test, ER+ normal reference network G_{ref}^{ER+} is compared with a ER- normal subnetwork with a particular correlation cutoff i G_i^{ER-} in the following three steps:

1) Compute the normalized Laplacian metrics N_{ref}^{ER+} and N_i^{ER-} from G_{ref}^{ER+} and G_i^{ER-} respectively.
2) Compute the eigenvalues E_{ref}^{ER+} and E_i^{ER-} from N_{ref}^{ER+} and N_i^{ER-} respectively.
3) Compute the KS statistic between E_{ref}^{ER+} and E_i^{ER-} .

The third step test the null hypothesis that eigenvalues E_{ref}^{ER+} and E_i^{ER-} are drawn from the same continuous distribution. If the two-tailed p-value returned by the KS test is high, then we cannot reject the hypothesis that G_{ref}^{ER+} and G_i^{ER-} are the same network. In another word, the higher the p-value is, the more similar G_{ref}^{ER+} and G_i^{ER-} .

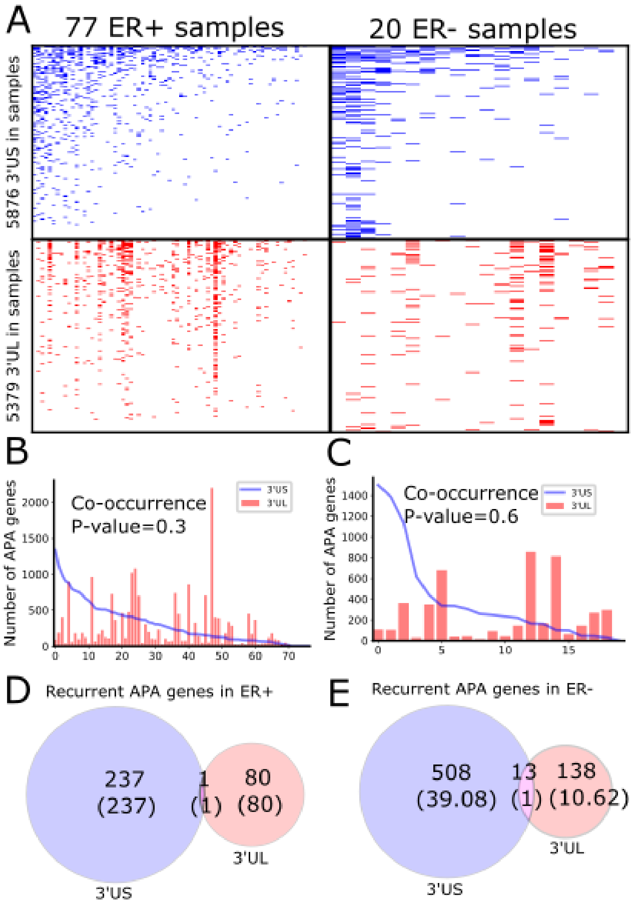


Figure 1. Global APA events distinct for ER+ and ER-. (A). Heatmaps showing genes with 3′US (first row) or 3′UL (second row) in ER+ samples (left column) or ER- samples (right column). The number of APA genes (3′US in line and 3′UL in red bar) in ER+ (B) and ER- (C). Samples are aligned in the same order in Fig. A. Overlap of recurring (>20% in tumor samples) 3′US and 3′UL genes in ER+ (D) and ER- (E).

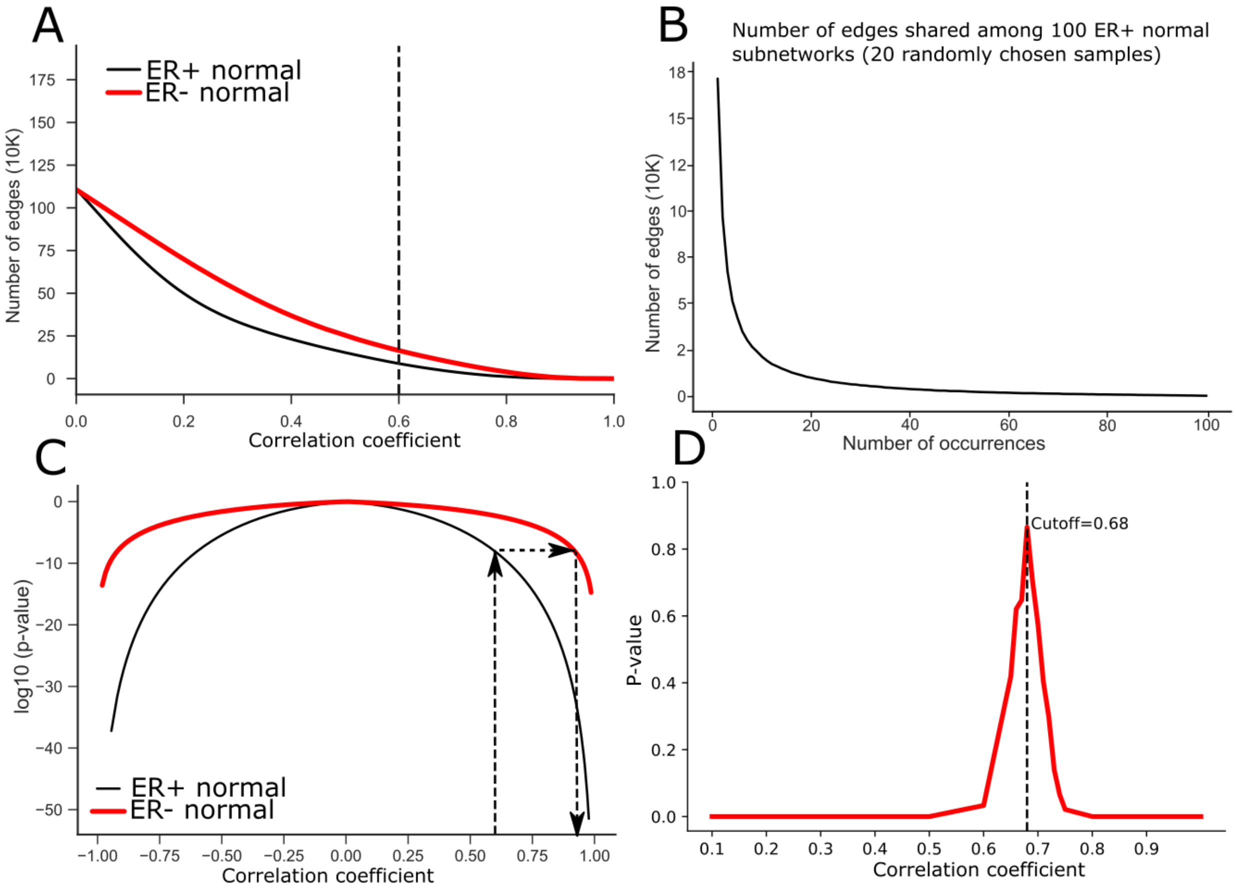


Figure 2. Two-way Pairwise Normalization of ER+ and ER- ceRNA network. (A) Number of edges in the ceRNA networks by the correlation coefficient cutoff (black and red line for ER+ and ER- networks, respectively). (B) The number of edges shared among 100 ER+ subnetworks from normal samples, where each of them was built by using 20 randomly chosen samples. (C) Statistical significance (p-value) achievable by using different correlation coefficient cutoff values for ER+ (black) and ER- (red) samples. Statistical significance for a correlation coefficient cutoff value is described in Methods. To achieve the same statistical significance of the traditional cutoff value (0.6) from ER+ to ER-, the cutoff value would inflate to 0.89, resulting in drastically a deflated number of edges. (D). Topological similarity (y-axis) between ER+ and ER- normal ceRNA networks by the cutoff value for ER- (x-axis). The bigger the p-value is, the more similar the two networks are (see Methods)[19]. The ER- normal network with the cutoff of 0.68 looks most similar to the ER+ normal reference network.

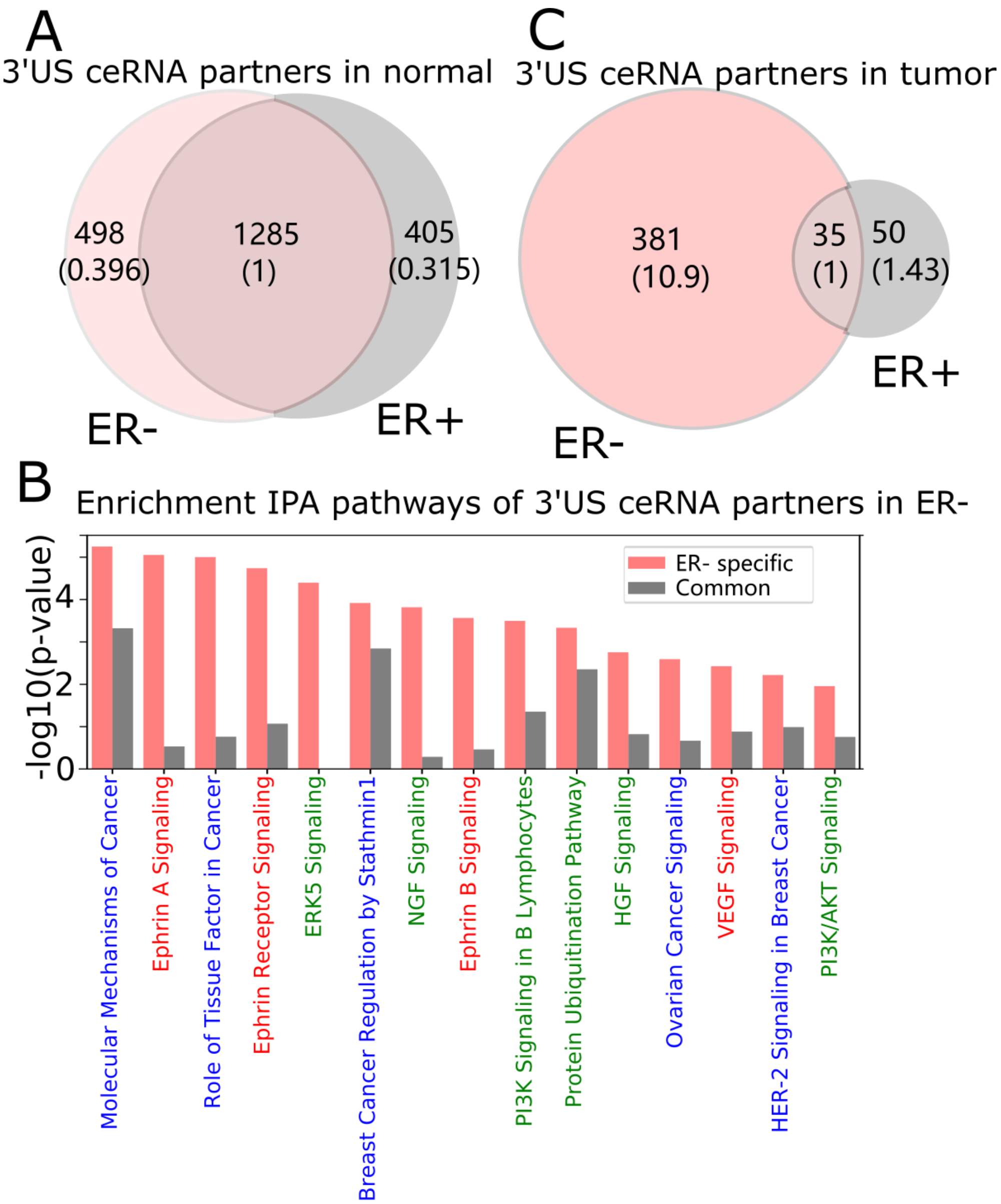


Figure 3. 3′UTR shortening is associated to ER-’s aggressive phenotypes in ceRNA. (A) Intersection of 3′US ceRNA partners between ER- and ER+ normal ceRNA networks. (B) IPA canonical pathways significantly (P < 0.01) enriched for the ER- 3′US ceRNAs. Pathways are colorcoded by keyword, “Cancer” in blue, “Signaling” in red and those associated with aggressive phenotypes[23] in green. (C) Intersection of 3′US ceRNA partners between ER- and ER+ tumor ceRNA networks.

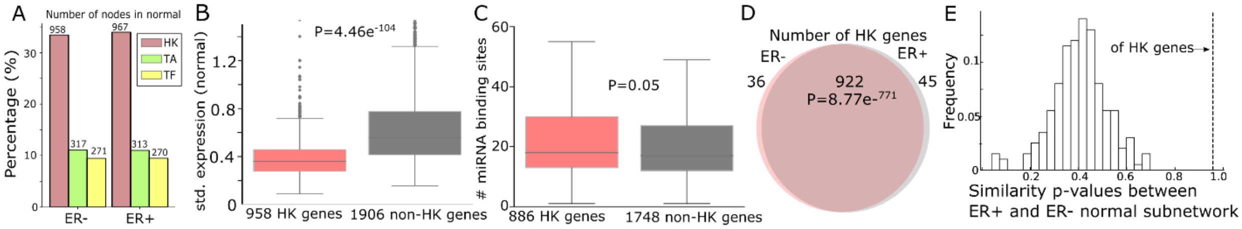


Figure 4. Housekeeping genes make consistent ceRNA networks between ER- and ER+ normal samples. (A) The number (and the percentage to the total number of nodes in the networks) of housekeeping (HK), tumor-associated (TA), or transcription factor (TF) genes in ER+ and ER- normal ceRNA networks. (B) Standard deviation of gene expressions of 958 HK genes and 1,906 non-HK genes across ER- normal samples. (C) Number of miRNA binding sites on the 3′UTR of 886 HK and 1,748 non-HK genes (those that have miRNA binding site information). (D) Number of HK genes shared by ER- and ER+ normal ceRNA networks (with those in common). (E) Distribution of the similarity p-values between subnetworks sampled with 922 HK genes from ER+ and ER- normal networks (horizontal dotted line) and 200 subnetworks sampled with 1,990 non-HK genes to the same number of HK genes (bar). The higher the p-value is, the more similar the networks are[19].

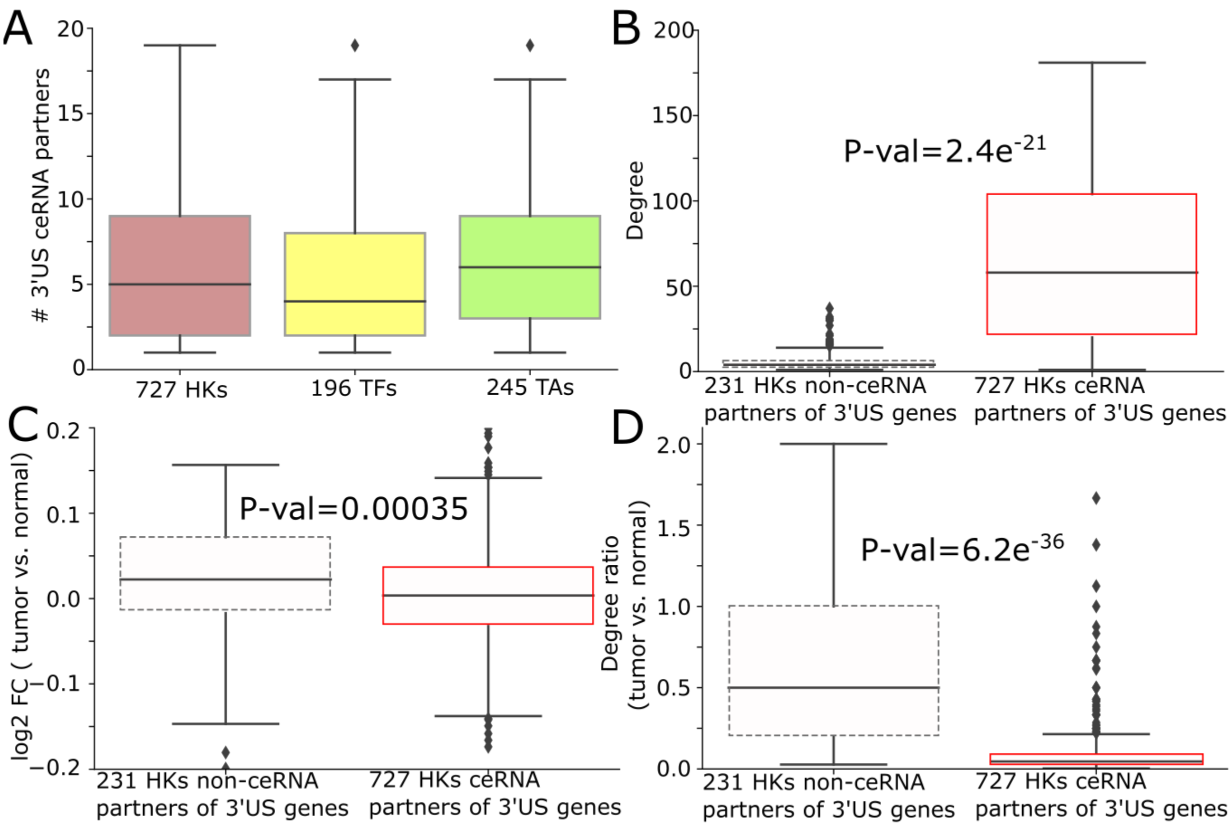


Figure 5. 3′US disrupts ceRNA relationship of HK genes. (A) The number of 3′UTR shortening genes connected to housekeeping (HK), transcription factor (TF), and tumor-associated (TA) genes. Degree (# neighbors in ERN normal ceRNA network) (B), log2 fold chance (tumor vs. normal) (C), degree ratio (tumor vs. normal) (D) of 727 and 231 HK genes that are ceRNA partners of 3′US genes or not, respectively. Since degree ratio in (D) represents the ratio of the number of neighbors retained in tumor, low degree values of 727 3′US HK ceRNA partners represents their higher loss of ceRNA neighbors in tumor.

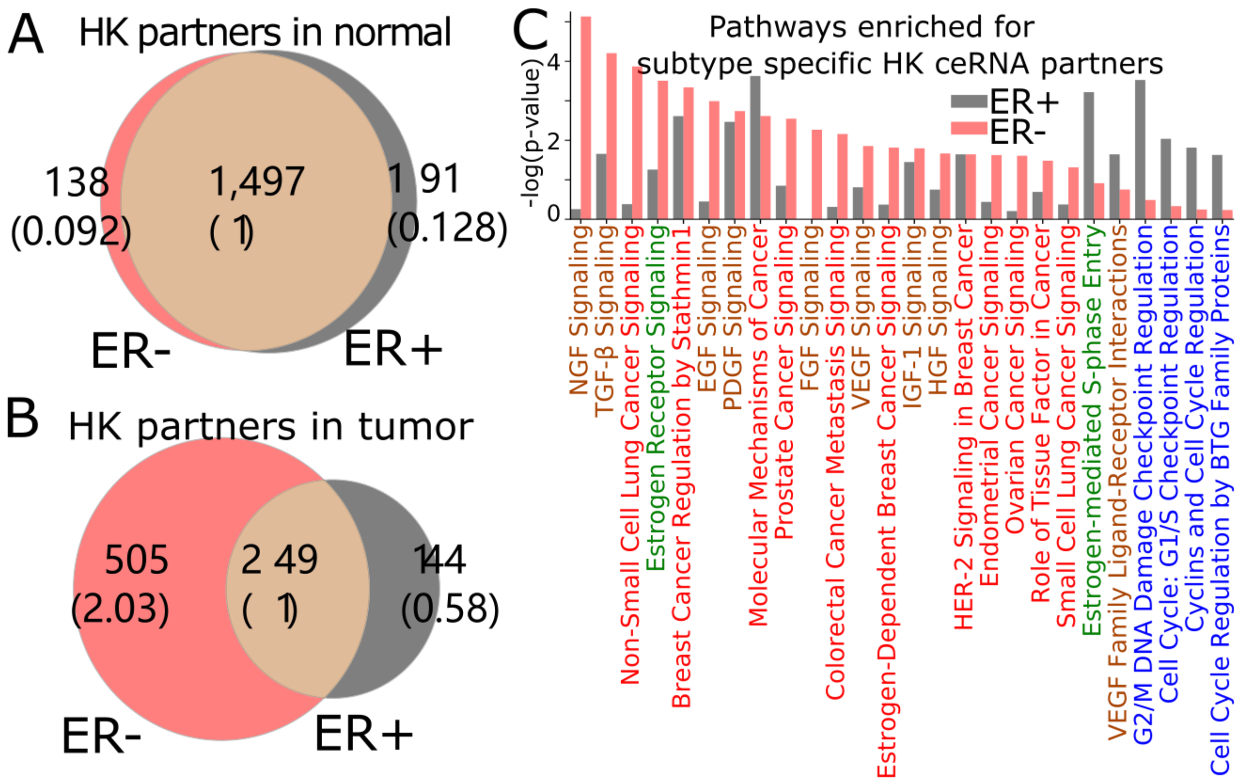
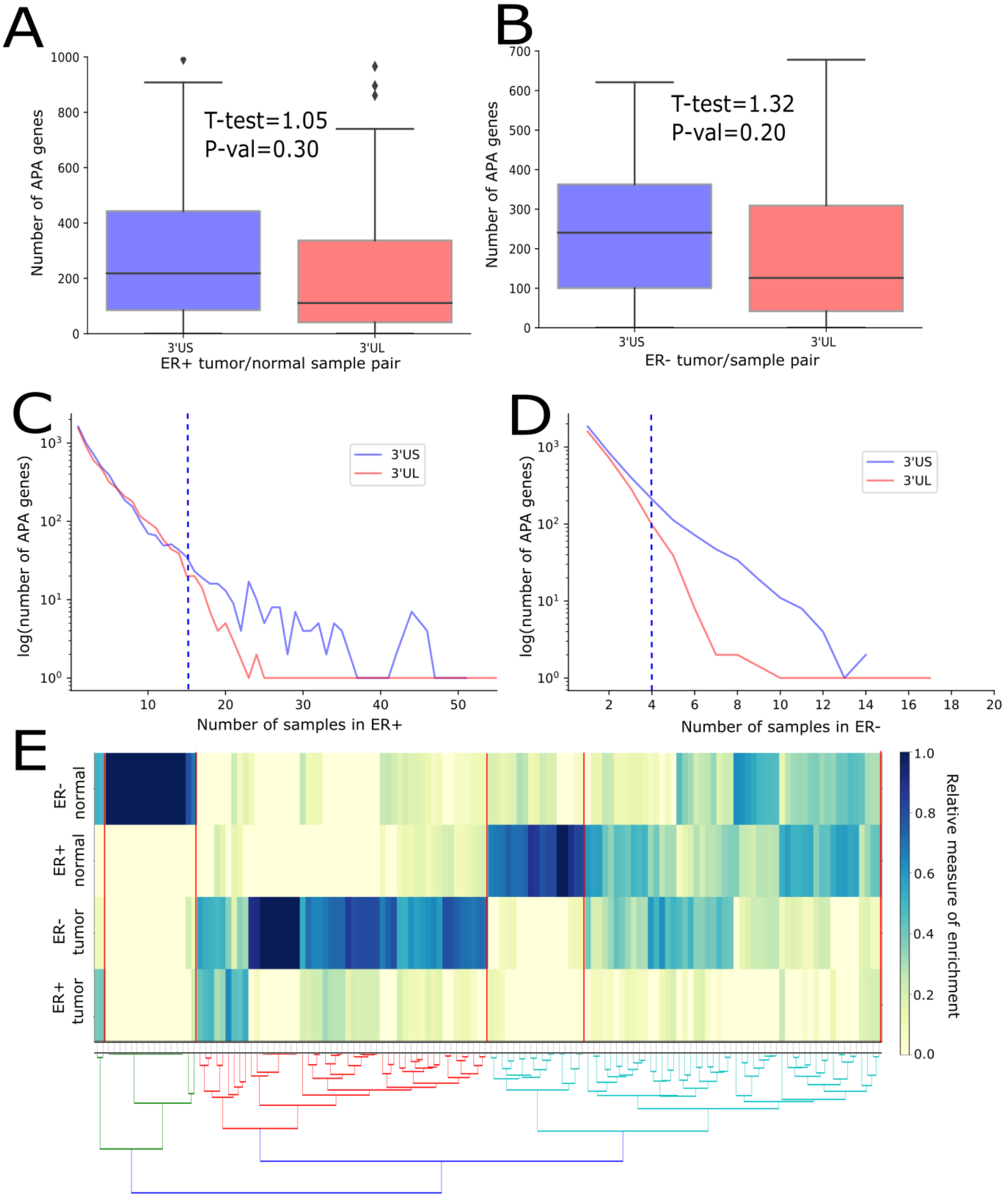
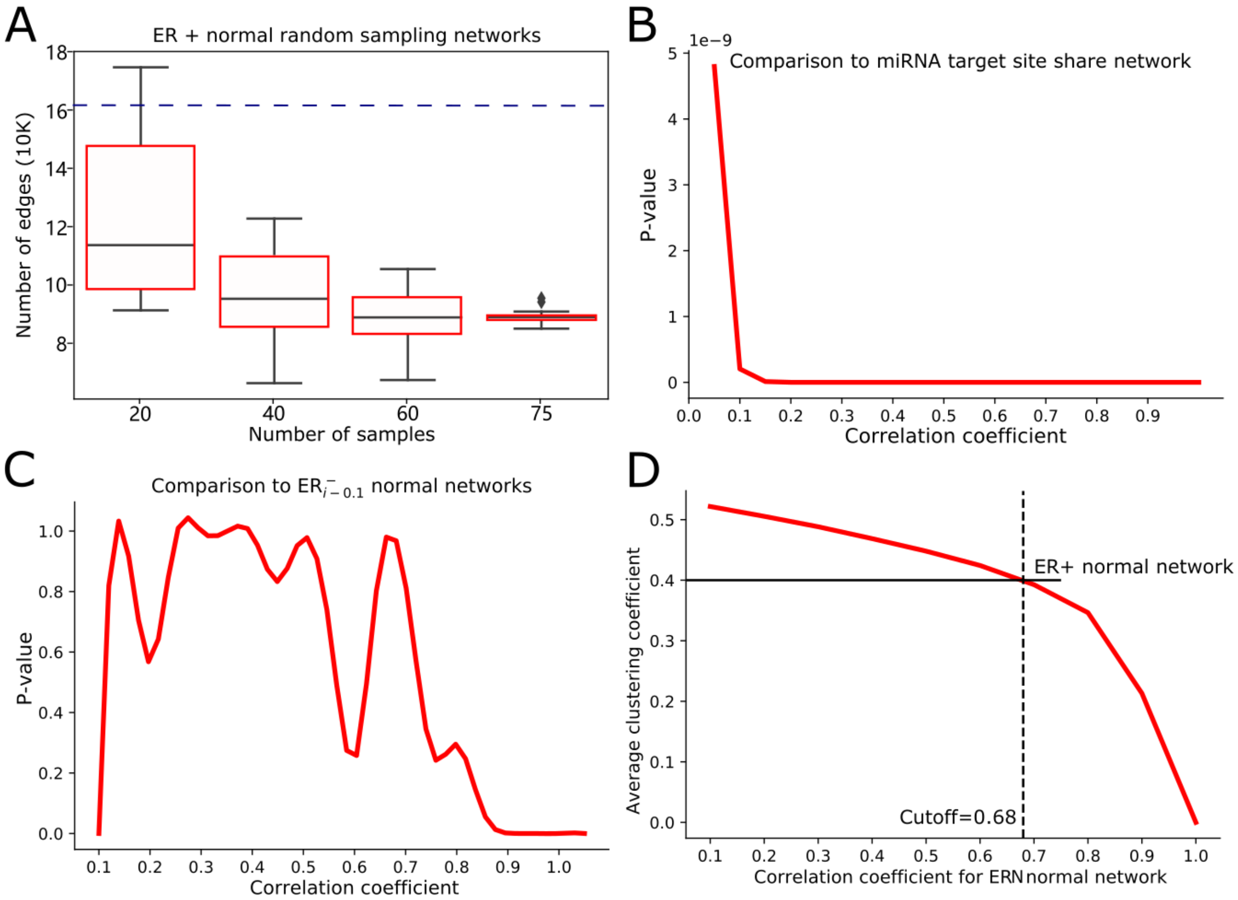


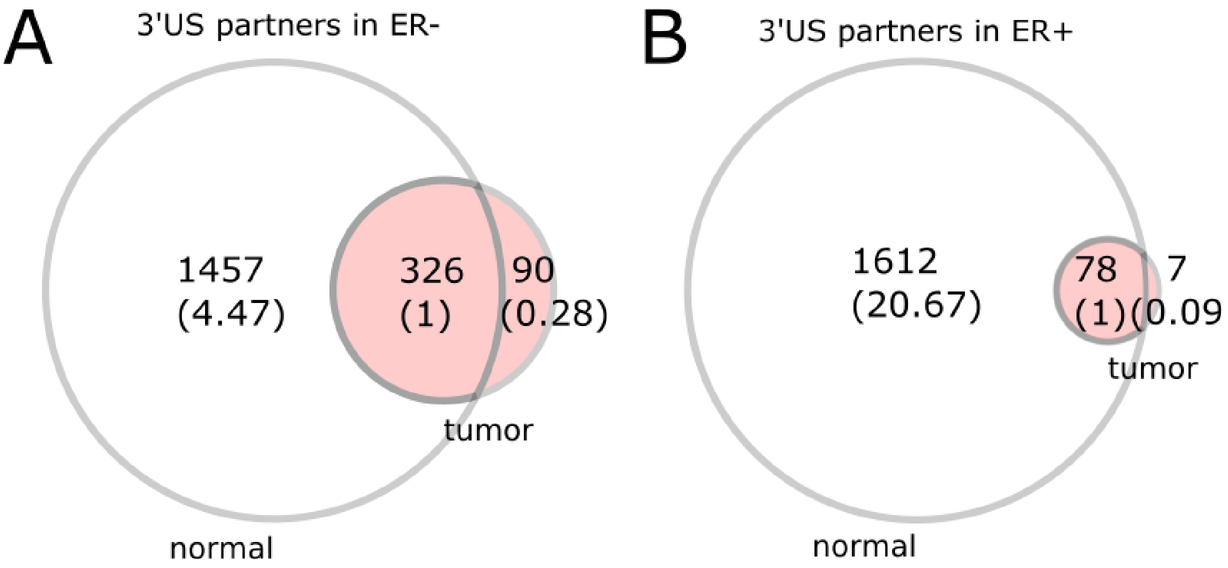
Figure 6. 3′US disrupts ceRNA relationship of HK genes for ER- specific growth. (A) Number of HK ceRNA partners unique and common to ER- and ER+ normal (left) and tumor (right) ceRNA networks. The numbers in parentheses are normalized to the number of genes shared between tumor and normal. (B) Degree in ER- tumor ceRNA network of 727 and 231 HK genes that are ceRNA partners of 3′US genes or not, respectively (C) IPA canonical pathways significantly (P < 0.01) enriched for ER+ and ER-specific HK ceRNA partners. Pathways are color-coded by keyword, “Cancer” in red, “GF” in brown, “Estrogen” in green, and “Cell Cycle” in blue.



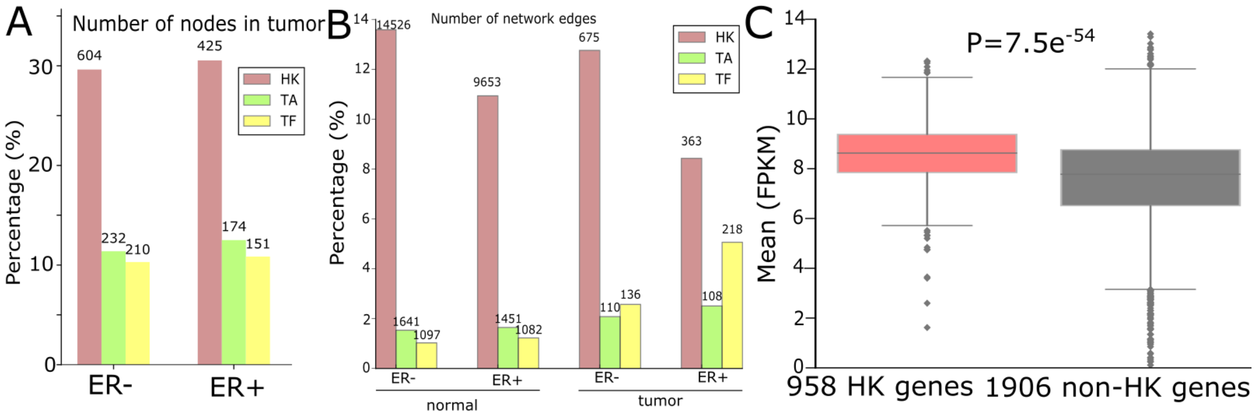
S. Figure 1. Boxplot showing the number of 3′US and 3′UL genes in each sample of ER+ (A) and ER- (B). T-test is t-test statistic value. Number of APA genes (y-axis blue in log10 for 3′US and red for 3′UL) common to ER+(C) and ER- (D) samples (x-axis) for ER+. The vertical dotted blue line marks the 20% threshold for recurrent events (E). IPA pathways enriched for 3′UL and 3′US genes in ER- and ER+. Colors represent enrichment of each pathway (column) for each class of genes (The higher the enrichment is, the higher the associated term is enriched). The red lines cut the pathways into 5 clusters, where each cluster is enriched in a set of genes.



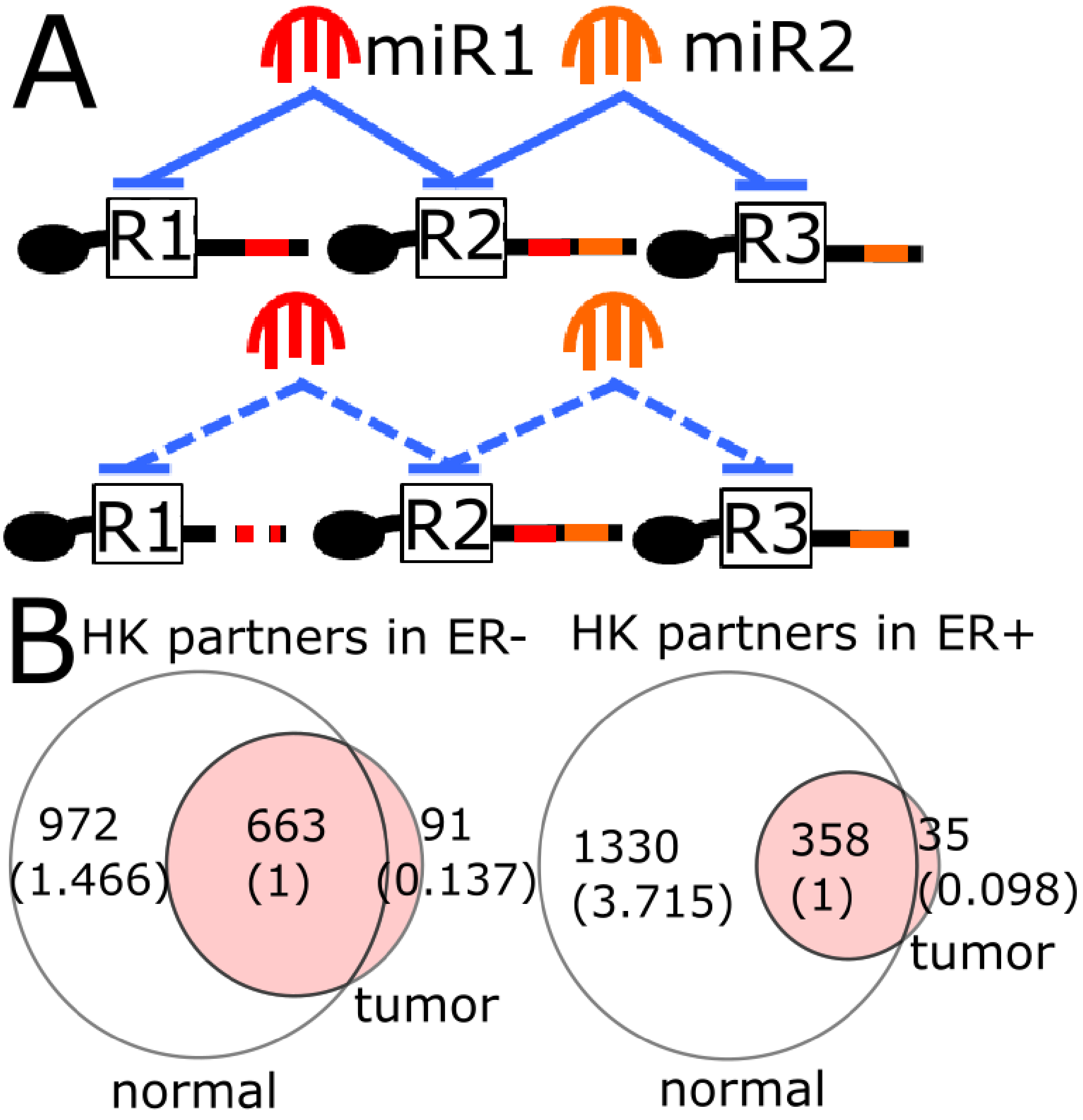
S. Figure 2. (A) Number of edges of networks from a subset of ER+ normal samples sampled in different size. Blue dotted line represents the number of edges of ER- normal network whose sample size is 20 (160,687) (B) Comparison of the miRNA target site share network to ER- normal networks with different correlation cutoff values (see Methods). As illustrated, ER- normal network topology changes drastically by different cutoff values in reference to the miRNA target site share network. (C) Comparison of ER- normal networks with previous correlation cutoff values in the stepwise increase (see Methods). (D) Comparing ER- normal ceRNA networks of different correlation cutoff values with the ER+ normal reference network in the average clustering coefficient. The average clustering coefficient for the ER+ normal reference network is 0.40 (indicated by the horizontal black line), which is quite close to the average clustering coefficient for the ER- normal network with a cutoff of 0.68 (indicated by the vertical dashed line).



S. Figure 3. Number of 3′US ceRNA partners in normal and tumor ceRNA networks found in (A) ER- and (B) ER+. The numbers in parentheses are normalized to the number of genes shared between tumor and normal.



S. Figure 4. (A) Number (and percentage to the total number of nodes in tumor networks) of HK genes and other important classes of genes in ER+ and ER- normal ceRNA networks. (B) The number (and the percentage to the total number of edges) of housekeeping (HK), transcription factor (TF), and 245 tumor-associated genes (TF) edges in ER- and ER+ for normal and tumor. (C) Average gene expression values of 1,003 HK genes and 1,990 non-HK genes in ER+ and ER- normal networks.



S. Figure 5. (A) Without 3'US (illustrated on top), R1 and R2 would compete for miR1 (red), forming ceRNA crosstalk and R2 and R3 would compete for miR2 (orange), forming ceRNA crosstalk. With 3'US on R1 (illustrated below), R1 would lose its ceRNA crosstalk with R2. Through indirect ceRNA effect that propagate the relationship loss, R2 would lose its ceRNA crosstalk with R3. (B) Numbers of HK ceRNA partners in ER- normal and in ER- tumor with overlap in common (The numbers in parentheses are normalized to the number of genes shared between tumor and normal). Numbers of HK ceRNA partners in ER+ normal and in ER- tumor (The numbers in parentheses are normalized to the number of genes shared between tumor and normal).

REFERENCES

[1] Z. Xia *et al.*, "Dynamic Analyses of Alternative Polyadenylation from RNA- Seq Reveal Landscape of 3 ' UTR Usage Across 7 Tumor Types," *Nat. Commun.*, pp. 1–38, 2014.

[2] H. J. Park *et al.*, "3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk," *Nat. Genet.*, 2018.

[3] T. C. G. A. Network, "Comprehensive molecular portraits of human breast tumours.," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[4] M. Sheikh, G. M, P. P, F. JA, and R. H, "Why are estrogen-receptor-negative breast cancers more aggressive t," *Invasion Metastasis*, vol. 14, no. 1–6, pp. 329–36, 1994.

[5] S. T. Pearce and V. C. Jordan, "The biological role of estrogen receptors and in cancer," *Crit. Rev. Oncol. Hematol.*, vol. 50, pp. 3–22, 2004.

[6] C. Mayr and D. P. Bartel, "Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.," *Cell*, vol. 138, no. 4, pp. 673–84, Aug. 2009.

[7] C. P. Masamha and E. J. Wagner, "The contribution of alternative polyadenylation to the cancer phenotype," *Carcinogenesis*, vol. 39, no. 1, pp. 2–10, 2018.

[8] M. Chen *et al.*, "3 ′ UTR lengthening as a novel mechanism in regulating cellular senescence," pp. 285–294, 2018.

[9] G. P. Dimri *et al.*, "A biomarker that identifies senescent human cells in culture and in aging skin in vivo.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 20, pp. 9363–7, 1995.

[10] R. A. Busuttil, M. Rubio, M. E. T. Dollé, J. Campisi, and J. Vijg, "Oxygen accelerates the accumulation of mutations during the senescence and immortalization of murine cells in culture.," *Aging Cell*, vol. 2, no. 6, pp. 287–294, 2003.

[11] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging," *Cell*, vol. 153, no. 6, 2013.

[12] D. Muñoz-Espín and M. Serrano, "Cellular senescence: From physiology to pathology," *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 7, pp. 482–496, 2014.

[13] Y. Xiang *et al.*, "Comprehensive Characterization of Alternative Polyadenylation in Human Cancer," vol. 110, no. November 2017, pp. 1–11, 2018.

[14] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?," *Cell*, vol. 146, no. 3, pp. 353–8, Aug. 2011.

[15] M. E. H. Hammond *et al.*, "American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version).," *Arch. Pathol. Lab. Med.*, vol. 134, no. 7, pp. e48-72, Jul. 2010.

[16] S. Tsutsui, S. Ohno, S. Murakami, Y. Hachitanda, and S. Oda, "Prognostic value of epidermal growth factor receptor (EGFR) and its relationship to the estrogen receptor

status in 1029 patients with breast cancer," *Breast Cancer Res. Treat.*, vol. 71, no. 1, pp. 67–75, 2002.

[17] A. Pergamenschikov *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2002.

[18] U. Ala *et al.*, "Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 18, pp. 7154–9, Apr. 2013.

[19] R. Gera *et al.*, "Identifying network structure similarity using spectral graph theory," *Appl. Netw. Sci.*, vol. 3, no. 1, p. 2, 2018.

[20] C. C. Friedel and R. Zimmer, "Inferring topology from clustering coefficients in protein-protein interaction networks," *BMC Bioinformatics*, vol. 15, pp. 1–15, 2006.

[21] P. Curmi *et al.*, "Overexpression of stathmin in breast carcinomas points out to highly proliferative tumours," *Br. J. Cancer*, vol. 82, no. 1, pp. 142–150, 2000.

[22] S. Obayashi *et al.*, "Stathmin1 expression is associated with aggressive phenotypes and cancer stem cell marker expression in breast cancer patients," *Int. J. Oncol.*, vol. 51, no. 3, pp. 781–790, 2017.

[23] K. Krishnan *et al.*, "miR-139-5p is a regulator of metastatic pathways in breast cancer.," *RNA*, vol. 19, no. 12, pp. 1767–80, Dec. 2013.

[24] D. Vaught, D. M. Brantley-Sieders, and J. Chen, "Eph receptors in breast cancer: Roles in tumor promotion and tumor suppression," *Breast Cancer Research*. 2008.

[25] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes, revisited.," *Trends Genet.*, vol. 29, no. 10, pp. 569–74, Oct. 2013.

[26] T. Davoli *et al.*, "Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome.," *Cell*, vol. 155, no. 4, pp. 948–962, Oct. 2013.

[27] K. Chawla, S. Tripathi, L. Thommesen, A. Lægreid, and M. Kuiper, "TFcheckpoint: A curated compendium of specific DNA-binding RNA polymerase II transcription factors," *Bioinformatics*, vol. 29, no. 19, pp. 2519–2520, 2013.

[28] Y. Tay, J. Rinn, and P. P. Pandolfi, "The multilayered complexity of ceRNA crosstalk and competition," *Nature*, vol. 505, no. 7483, pp. 344–352, Jan. 2014.

[29] P. Sumazin *et al.*, "An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma.," *Cell*, vol. 147, no. 2, pp. 370–81, Oct. 2011.

[30] M. Nitzan, A. Steiman-Shimony, Y. Altuvia, O. Biham, and H. Margalit, "Interactions between distant ceRNAs in regulatory networks," *Biophys. J.*, vol. 106, no. 10, pp. 2254–2266, 2014.

[31] H. J. Park, S. Kim, and W. Li, "Model-based analysis of competing- endogenous pathways (MACPath) in human cancers," *PLoS Comput. Biol.*, vol. 22, no. 14, 2018.

[32] J. Lu *et al.*, “MicroRNA expression profiles classify human cancers.,” *Nature*, vol. 435, no. 7043, pp. 834–8, Jun. 2005.

[33] G. Hong *et al.*, “Genes Dysregulated to Different Extent or Oppositely in Estrogen Receptor-Positive and Estrogen Receptor-Negative Breast Cancers,” *PLoS One*, vol. 8, no. 7, p. e70017, 2013.

[34] M. C. Alles *et al.*, “Meta-Analysis and Gene Set Enrichment Relative to ER Status Reveal Elevated Activity of MYC and E2F in the ‘Basal’ Breast Cancer Subgroup,” *PLoS One*, vol. 4, no. 3, p. e4710, 2009.

[35] M. C. Abba *et al.*, “Gene expression signature of estrogen receptor α status in breast cancer,” *BMC Genomics*, vol. 6, pp. 1–13, 2005.

[36] D. K. Biswas, a. P. Cruz, E. Gansberger, and a. B. Pardee, “Epidermal growth factor-induced nuclear factor kappa B activation: A major pathway of cell-cycle progression in estrogen-receptor negative breast cancer cells,” *Proc. Natl. Acad. Sci.*, vol. 97, no. 15, pp. 8542–8547, Jul. 2000.

[37] D. Fuckar *et al.*, “VEGF expression is associated with negative estrogen receptor status in patients with breast cancer,” *Int. J. Surg. Pathol.*, vol. 14, no. 1, pp. 49–55, 2006.

[38] S. Javanmoghadam, Z. Weihua, K. K. Hunt, and K. Keyomarsi, “Estrogen receptor alpha is cell cycle-regulated and regulates the cell cycle in a ligand-dependent fashion,” *Cell Cycle*, vol. 15, no. 12, pp. 1579–1590, 2016.

[39] S. Paruthiyil, H. Parmar, V. Kerekatte, G. R. Cunha, G. L. Firestone, and D. C. Leitman, “Estrogen Receptor □ Inhibits Human Breast Cancer Cell Proliferation and Tumor Formation by Causing a G 2 Cell Cycle Arrest,” pp. 423–428, 2004.

[40] D. C. Henley, J. S. Foster, J. Wimalasena, P. Seth, and A. Bukovsky, “Multifaceted Regulation of Cell Cycle Progression by Estrogen: Regulation of Cdk Inhibitors and Cdc25A Independent of Cyclin D1-Cdk4 Function,” *Mol. Cell. Biol.*, vol. 21, no. 3, pp. 794–810, 2002.

[41] E. Eisenberg and E. Levanon, “Human housekeeping genes are compact,” *TRENDS Genet.*, vol. 19, no. 7, pp. 362–365, 2003.

[42] M. Goldman *et al.*, “The UCSC Cancer Genomics Browser: update 2013.,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D949-54, Jan. 2013.

[43] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.,” *Cell*, vol. 120, no. 1, pp. 15–20, Jan. 2005.

[44] G. L. Papadopoulos, M. Reczko, V. a Simossis, P. Sethupathy, and A. G. Hatzigeorgiou, “The database of experimentally supported targets: a functional update of TarBase.,” *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D155-8, Jan. 2009.

[45] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, “miRecords: An integrated resource for microRNA-target interactions,” *Nucleic Acids Res.*, vol. 37, no. November 2008, pp. 105–110, 2009.

[46] S.-D. Hsu *et al.*, "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D78-85, Jan. 2014.

[47] H. Dvinge *et al.*, "The shaping and functional consequences of the microRNA landscape in breast cancer.," *Nature*, vol. 497, no. 7449, pp. 378–82, May 2013.

[48] M. P. Hamilton *et al.*, "Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif.," *Nat. Commun.*, vol. 4, p. 2730, Jan. 2013.

35. Park HJ, Kim S, Li W. Model-based analysis of competing-endogenous pathways (MACPath) in human cancers. Wang E, editor. *PLOS Comput Biol*. 2018;14: e1006074. doi:10.1371/journal.pcbi.1006074