Discovering Competing Endogenous RNA Interactions in Breast Cancer Molecular Subtypes

Gulden Olgun[1], Ozgur Sahin[2], and Oznur Tastan[1]

[1]Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
[2]Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, Ankara, 06800, Turkey.

Abstract

Motivation: Long non-coding RNAs(lncRNAs) can act as competing endogenous RNAs(ceRNAs); they indirectly regulate mRNAs expression levels by reducing the amount of microRNAs(miRNAs) available to target mRNAs. Previous work identified potential lncRNA-mediated ceRNA interactions in multiple cancer types including breast cancer. These ceRNA interactions have not been yet characterized for breast cancer subtypes.

Results:. To find lncRNA-mediated ceRNA interactions in molecular subtypes of breast cancer, we use partial correlation analysis and kernel independence tests on patient gene expression profiles and further refine the candidate interactions with miRNA target information. We find that although there are sponges common to multiple subtypes, there are also distinct ceRNA regulatory interactions specific to certain subtypes. Furthermore, we show that functional enrichment of mRNAs involved in ceRNA interactions proposed roles of different biological processes for different subtypes. Interestingly, spatially proximal ceRNA interaction analysis suggested a tight regulation of HOX genes by HOTAIR using miR-196a-1 and miR-196a-2. We also discover subtype specific ceRNA interactions with high prognostic potential. When grouped based on the expression patterns of these sponge interactions, patients differ significantly in their survival distributions while patients groups based on individual RNA expression profiles of the sponge participants, the groups do not yield a significant difference in survival.

Contact: oznur.tastan@cs.bilkent.edu.tr

1 Introduction

Advances in sequencing technologies have revealed that there are large number of RNAs that do not encode for proteins [11]. One class of non-coding RNAs(ncRNAs) are microRNAs(miRNAs) that repress gene expression by preferentially binding the complementary sequence of their target mRNAs and triggering translation repression or degradation[2]. MiRNAs play crucial roles in regulating gene expression programs in the normal cell and their aberrant expression contribute to pathogenesis in several diseases, including cancer. To date a large number of miRNAs have been shown to be associated with cancer progression, drug resistance or metastasis[26, 6, 47, 37, 32].

Another major class of non-coding RNAs are long non-coding RNAs(lncRNAs) that are longer than 200 nucleotides. Although a small number of lncRNAs are functionally characterized so far, accumulating evidence suggests that they are involved in regulation of diverse cellular and pathological processes[45, 36]. Recent work has provided evidence for an emerging role of lncRNAs; by acting as competitive endogenous

RNAs(ceRNA), lncRNAs can reduce the amount of miRNAs available for the target mRNA; in this way, they indirectly prevent the target gene repression[7, 14]. As cancer is characterized by aberrant expression of transcripts, dysregulation of these type of RNA-RNA interactions contributes to cancer[20].

lncRNA-associated ceRNA interactions have been investigated in gastric cancer[49], glioblastoma multiforme[8], pancreatic cancer[50], ovarian cancer[53] and in breast cancer[34]. However, subtype specific lncRNA-mediated ceRNA networks in breast cancer subtypes have not been characterized with functional and prognostic potential. Breast cancer is the second leading cause of cancer deaths among women[41] and its subtypes differ significantly in their molecular profiles and response to therapy. Because miRNAs exhibit different molecular activity patterns in breast cancer subtypes[23, 5], it is expected that there will be subtype specific lncRNA-mediated ceRNA interactions as well. Identifying these miRNA sponges can both shed light on the uncharacterized mechanisms of the breast cancer subtypes and potentially help in making better therapeutic decisions. In this work, we use an integrative approach to identify subtype specific lncRNA-miRNA-mRNA interactions in which lncRNAs compete for binding to shared miRNAs in breast cancer.

To identify subtype specific lncRNA-mediated ceRNA interactions, we develop and integrative methodology and systemically analyze lncRNA, miRNA and mRNA expression profiles of breast cancer patients made available through the Cancer Genome Atlas Project(TCGA)[33]. We identify statistically related lncRNA-miRNA-mRNA interactions through correlation and partial correlation analysis as in Paci et al.[34] and further refine these candidate interactions using a kernel-based conditional independence test (KCI)[52]. KCI, which does not assume any parametric form on the random variables tested, is used for the first time in finding regulatory interactions. The potential candidate interactions are further filtered in the light of available evidence regarding the miRNA-target interactions. We investigate the functional enrichment of mRNAs that participate in sponges, the genomic spatial organization and finally, through survival analysis of patients, we discover lncRNA-mediated ceRNA interactions with prognostic value.

2 Methods

2.1 Data Collection and Processing

2.1.1 lncRNA curation

As lncRNAs are not annotated in TCGA, we curated a list of lncRNAs using GENCODE v24 [17]. Based on Gencode24 annotation, 598 of the RNAs present in RNASeq expression data are designated as lncRNAs. To minimize erroneous annotations, we further checked each lncRNA's coding potential with alignment-free method Coding-Potential Assessment Tool(CPAT) [46] and alignment-based method Coding Potential Calculator(CPC) [22]. LncRNAs whose all transcripts are predicted to have high coding potential by both tools are eliminated. The number of lncRNAs that are predicted to have high coding potential by each tool are provided in Figure S1 (Supp. File1).

2.1.2 Expression data processing

Level 3 IlluminaHiSeq RNA-seq gene expression and miRNA expression data for human breast cancer were collected from The Cancer Genome Atlas [33] on August 9^{th} 2014 and patient survival data was obtained from the UCSC Cancer Genomics Browser on June 31^{st} 2016. Only the patient data that concurrently include mRNA, lncRNA and miRNA expression data were used. Patients were divided into subtypes based on information in TCGA defined by PAM50 method. The four subtypes used are Luminal A, Luminal B, Basal, HER2. The number of patients in each subtype is provided in Table S1 (Supp. File1).

In expression data, Reads Per Kilobase Million Reads(RPKM) values were used. To eliminate the genes and miRNAs with very low expression, we assumed that RKPM values below than 0.05 is missing and filtered out RNAs that are missing in more than 20 of the samples in each subtype. Expression values were added with a constant 0.25 to deal with the 0 gene expression values and are $\log 2$ transformed.

RNAs that do not vary across samples were filtered. We eliminated the genes with median absolute deviation (MAD) below than 0.5. MAD is calculated as follows:

$$\text{MAD } (r) = \text{median } (|r_i - \text{median}(\mathbf{r})|) \tag{1}$$

,where r_i denotes the RNA expression in sample i for RNA r and $\mathbf{r}$ denotes the vector that contains expression values for all samples for RNA r.

2.2 Statistical Analysis for Finding lncRNA Mediated ceRNA Interactions

To identify ceRNA interactions between lncRNA-miRNA-mRNA, we performed correlation analysis and kernel-based conditional independence test on expression data. Below, X random variable denotes a lncRNA, Y denotes an mRNA and finally Z denotes a miRNA.

2.2.1 Correlation and Partial Correlation Analysis

For a given ceRNA interaction, we expect expression values of the lncRNA and mRNA to be positively correlated and if this correlation relies on miRNA expression, the correlation between mRNA and lncRNA should weaken when miRNA expression is taken into account. To quantify this, first Spearman rank order correlation was calculated between lncRNA and mRNAs, which we denote with $\rho_{lncRNA,mRNA}$. Next, we calculated the Spearman partial rank order correlation between lncRNA and mRNA, this time controlling for miRNA expression, $\rho_{lncRNA,mRNA|miRNA}$, as follows:

$$\rho_{X,Y \mid Z} = \frac{\rho_{X,Y} - \rho_{X,Z}\ \rho_{Y,Z}}{\sqrt{1-\rho_{X,Z}^2}\sqrt{1-\rho_{Y,Z}^2}} \tag{2}$$

The difference between the correlation and the partial correlation for a miRNA measures the extend the miRNA Z is effective in the statistical correlation of X and Y, this value is calculated:

$$S_Z = \rho_{X,Y} - \rho_{X,Y \mid Z} \tag{3}$$

As we look for strongly positively correlated lncRNA and mRNA pairs, only those with correlation $\rho_{X,Y} > 0.5$ (p-value < 0.05) were considered. Among those, RNA triplets where S_Z is larger than a threshold value, t, were retained. We conducted our analysis at two different thresholds $t = 0.2$ and $t = 0.3$.

2.2.2 Kernel Based Conditional Independence Test

To find lncRNA interactions we also test directly for conditional independence. In a ceRNA interaction, if the interaction of a particular pair of lncRNA (X) and mRNA(Y) were through a shared miRNA (Z), we would expect that lncRNA and mRNA expressions to be conditionally independent given the miRNA expression level. Conditional independence is denoted by $X \perp\!\!\!\perp Y \mid Z$. Formally, X and Y are conditionally independent given Z if and only if the $\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z)$ (or equivalently $\mathbf{P}(Y \mid X, Z) = \mathbf{P}(Y \mid Z)$ or $\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\,\mathbf{P}(Y \mid Z)$). That is if X and Y are conditionally independent given Z, further knowing the values of X (or Y) does not provide any additional evidence about Y(or X). There are conditional independence tests available for continuous random variables[44, 18, 42, 52]. In our work we employ, kernel-based conditional independence (KCI) test proposed by Zhang et al.[52] as it does not make any distributional assumptions on the variables tested. Furthermore, KCI-test does not require explicit estimation of the joint or conditional probability densities and avoids discretization of the continuous random variables, both of which require large sample sizes for an accurate test performance. Below we describe the KCI-test briefly, details of which can be found in [52].

KCI-test defines a test statistic which is calculated from the kernel matrices associated with X, Y and Z random variables. A kernel function takes as its inputs vectors in the original space and returns the dot product of the input vectors in a trans-

formed feature space, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The feature transformation is denoted by $\Phi : \mathcal{X} \to \mathcal{H}$[39], $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle$. In this work we use the Gaussian kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_x^2})$, where $\sigma > 0$ is the kernel width. CI and KCI are based on kernel matrix of X, Y and Z, which are calculated by evaluating the kernel function for all pairs of samples, i.e. the (i,j)th entry of $\mathbf{K}_X$ is $k(\mathbf{x}_i, \mathbf{x}_j)$. The corresponding centralized kernel matrix is $\tilde{\mathbf{K}}_X \triangleq \mathbf{H}\mathbf{K}_X\mathbf{H}$ where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is a vector 1's. $\tilde{\mathbf{K}}_Y$ and $\tilde{\mathbf{K}}_Z$ are similarly calculated for Y and Z variables.

Given the i.i.d. samples $\mathbf{x} \triangleq (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} \triangleq (y_1, y_2, \ldots, y_n)$, the unconditional kernel test first calculates the centralized kernel matrices, $\tilde{\mathbf{K}}_X$ and $\tilde{\mathbf{K}}_Y$ from the samples $\mathbf{x}$ and $\mathbf{y}$ and then eigenvalues of the centralized matrices. The eigenvalue decompositions of centralized kernel matrices $\tilde{\mathbf{K}}_X$ and $\tilde{\mathbf{K}}_Y$ are $\tilde{\mathbf{K}}_X = \mathbf{V}_x\Lambda_x\mathbf{V}_x^T$ and $\tilde{\mathbf{K}}_Y = \mathbf{V}_y\Lambda_y\mathbf{V}_y^T$. Here Λ_x and Λ_y are the diagonal matrices containing the non-negative eigenvalues $\lambda_{\mathbf{x},i}$ and $\lambda_{\mathbf{y},i}$ in descending order, respectively. $\mathbf{V}_x$ and $\mathbf{V}_y$ matrices contain the corresponding eigenvectors. Zhang et al. [52] shows that under the null hypothesis that X and Y are independent, the following test statistic:

$$T_{UI} \triangleq \frac{1}{n} \operatorname{Tr}(\tilde{\mathbf{K}}_X\tilde{\mathbf{K}}_Y) \tag{4}$$

has the same asymptotic distribution ($n \to \infty$) as

$$\tilde{T}_{UI} \triangleq \frac{1}{n^2} \sum_{i,j=1}^{n} \lambda_{\mathbf{x},i}\lambda_{\mathbf{y},i}z_{i,j}^2, \tag{5}$$

Here $z_{i,j}$ are i.i.d. standard Gausian variables, thus $z_{i,j}^2$ are i.i.d χ_1^2 - distributed. The unconditional independence test procedure involves calculating T_{UI} according to Eq (4). Empirical null distribution of $\tilde{T}_{UI}$ is simulated by drawing i.i.d random samples for $z_{i,j}^2$ variable from $\tilde{\chi}^2$. Finally, the p-value is calculated by locating T_{UI} in the emprical null distribution. We use this

The kernel conditional independence test also makes use of the centralized kernel matrices. Under the null hypothesis that X and Y are conditionally independent given Z, the following test statistic is calculated:

$$\tilde{T}_{CI} \triangleq \frac{1}{n} \operatorname{Tr}(\tilde{\mathbf{K}}_{\ddot{X}\,|\,Z}\tilde{\mathbf{K}}_{Y\,|\,Z}), \tag{6}$$

where $\ddot{X} \triangleq (X, Z)$ and $\mathbf{K}_{\ddot{X}}$ is the centralized kernel matrix for $\ddot{X}$. As Zhang et al. [52] reports has the same asymptotic distribution as

$$\tilde{T}_{CI} \triangleq \frac{1}{n} \sum_{k=1}^{n^2} \mathring{\lambda}_k \cdot z_k^2 \tag{7}$$

The details of the definition of $\mathring{\lambda}_k$ and z_k^2 can be found in [52]. The procedure involves calculating the emprical p-value based on the test statistic as defined in Eq (4) and simulating the null distribution based on Eq (7).

Using the unconditional kernel independence test, we first test the null hypothesis that a lncRNA and mRNA pair is independent against the alternative hypothesis they are dependent. For those pairs where the null is rejected at significance level 0.01 are considered. For each of the lncRNA-mRNA pair, we test their conditional independence with all miRNAs. A lncRNA and mRNA pair is tested for conditionally independence given an miRNA using KCI. Those pairs that are found to be independent at significance level 0.01 are considered as potential lncRNA-mediated ceRNAs.

2.3 Filtering ceRNAs Based on miRNA-Target Interactions

To identify interactions that are biologically meaningful, we filtered the potential ceRNA interactions that were not supported by miRNA target information. The miRNA-mRNA and miRNA-lncRNA interactions are retrieved from multiple databases as listed in Table 1. The candidate sponges are retained if both mRNA and lncRNA have support for being targeted by the miRNA of the sponge.

Table 1: Computationally and experimentally validated miRNA-target databases used for mRNA and lncRNA. Plus signs denote databases that are used for the miRNA interactions of the RNA type. 'P' denotes predicted target information while 'E' denotes experimentally supported target information.

miRNA-Target Databases	mRNA	lncRNA	P/C	Reference
TargetScan	+	+	P	[1]
mirCode	+	+	P	[19]
mirSVR	+	+	P	[3]
PITA	+	+	P	[21]
RNA22	+	+	P	[30]
lnCeDB		+	P	[10]
mirWalk		+	P	[12]
mirTarBase	+	+	E	[9]
Diana		+	E	[35]
LncBase				

2.4 Identifying ceRNAs with Prognostic Value

To evaluate the ceRNA interactions in terms of their prognostic potential, we analyzed the survival of the patients based on the expression patterns of each sponge interaction. In a sponge interaction, we expect the lncRNA and mRNA to be regulated in the same direction and miRNA to be in the opposite direction. For each ceRNA found in a subtype, the patients are divided into two groups based on the regulation patterns of the RNAs that participate in the ceRNA. For the up-down-up pattern, the first group comprises patients whose sponge lncRNA and mRNA are upregulated and miRNA is downregulated; the second group includes all patients that do not fit in this pattern. Similarly, we divide the patients based on the down-up-down pattern: if both lncRNA and mRNA are downregulated whereas miRNA is unregulated such patients constitute one group and the rest of the patients constitute the second group.

For each of the subtype specific ceRNAs identified, the patient groups are tested whether their survival rates significantly differ from each other using log-rank test[16] (p-value < 0.05). We further excluded ceRNAs if any of the RNAs can by itself divide the patients into groups that differ in terms of survival distributions significantly. In each subtype, the patients are divided as upregulated and downregulated for each of the RNA participating in the ceRNA interaction separately. If at least one of the molecules leads to groups with significant survival difference (log-rank test, (p-value < 0.05)), we disregard this ceRNA from the list of prognostic ceRNAs. This last step ensures that the prognostic difference is due to the interactions between the RNAs and that it does not stem from expression of the single RNA's expression patterns.

The identified ceRNA interactions are further divided based on the following f score that reflects the interaction's prognostic value with respect to the individual RNA's prognostic value:

$$f_{xyz} = -\log \frac{p_{xyz}}{\texttt{min}(p_x, p_y, p_z)} \tag{8}$$

Here, p_{xyz} is the p-value attained in testing whether patient survivals differ based on the log-rank test whereas p_x, p_y, p_z indicate the p-values obtained by testing patient survival distribution differences due to lncRNA, mRNA, and miRNA expression patterns, respectively.

In the above analysis, RNAs that have expression levels above (or below) a certain threshold value are considered upregulated (or downregulated). This threshold value is selected among the candidate cut-off values of expression as the one that results in the lowest p-value in the log-rank test when patients are grouped based on this cut-off. The candidate cut-off values were the 10^{th} and 90^{th} percentiles, mean, median, the lower and upper quartiles of the expression values of the patients in each subtype.

2.5 Pathway and GO Enrichment Analysis

We conducted pathway and GO enrichment of mRNAs that participate in subtype specific sponges. Enrichment tests are conducted with clusterProfiler [51] with Bonferroni multiple hypothesis test correction.

In deciding enriched pathways and GO terms, a p-value cutoff of 0.05 and FDR cutoff of 1×10^{-4} are used. In both pathway and GO enrichment analysis the background genes were the union of mRNAs that remained after MAD filtering step (Step B in Figure 1(A)). For pathway enrichment analysis, different pathway data sources were downloaded from Baderlab GeneSets Collection[28]. List of all pathways that were employed in this analysis is provided in Table S2 (Supp. File1). Redundant pathways are eliminated when different sources are combined. Additionally, a pathway enrichment analysis is conducted with KEGG pathways (downloaded on February 28^{th} 2017).

2.6 Clustering mRNAs

If mRNAs are highly correlated among each other, we often find that correlated mRNAs participate in ceRNA interactions with the lncRNA and miRNA pair. We consider the mRNAs that participate in a ceRNA interaction with the same pair of lncRNA and miRNA. If all mRNAs are strongly correlated among each other, where all the pairwise correlations are above than 0.7, all mRNAs are assigned into the same cluster. Otherwise, we apply Ward hierarchical clustering method to find groups of correlated mRNAs[48]. We determine optimal number of clusters with Mojena's stopping rule[31] using Milligan and Cooper's[29] correction.

3 Results and Discussion

3.1 Overview of Discovered ceRNA Interactions

In order to discover subtype specific breast cancer ceRNA interactions, we employ the methodology summarized in Figure 1(A) and identify ceRNAs specific to four molecular subtypes of breast cancer: Luminal A, Luminal B, HER2 and Basal. The number of candidate ceRNA interactions that remain after each main step when in the partial correlation analysis step S value threshold $t = 0.2$ is employed, is provided in Figure 1(B) (see Figure S2(A) in Supp. File1 for $t = 0.3$). The total number of ceRNA interactions found in all subtypes is 11.614. Figure 1(C)

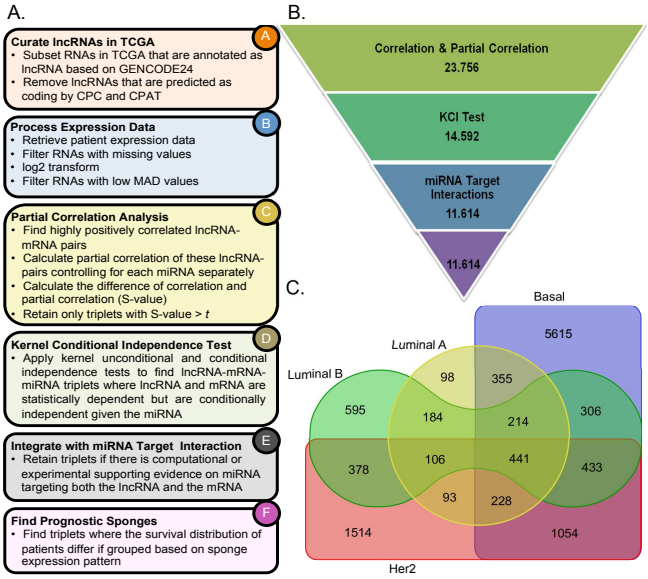

Figure 1: A)Overview of the methodology, each box represents a step in the methodology. Steps B-F are conducted for each breast cancer subtype separately. B)Number of ceRNAs remained after each main filtering step when $t = 0.2$ (Step C in Figure 1A). C)Venn diagram of ceRNA interactions discovered in each of the breast cancer molecular subtype.

shows the Venn diagram of number of ceRNA interactions discovered for the four subtypes (see Figure S2(B) in Supp. File1 for $t = 0.3$).

Although there are sponges that are detected in multiple subtypes, there are also a large number of sponges that are only specific to a single subtype (Table S3 and Figures S3A and S3B in Supp. File1). The list of sponges identified in each subtype, their partial correlation analysis, KCI-test results and target information are provided in Supp. File 2.

We analyze the specificity of the individual RNAs that participate in each of the subtypes. Figures 2A and 2B display the number of sponges per lncRNA and miRNA for $t = 0.2$ (Figure S3C in Supp. File1 for $t = 0.3$). Some lncRNAs and miRNAs participate in sponges of all the subtypes (Table 2); i.e. KIAA0125 participates in a large number of sponges across the four subtypes. KIAA0125 has been re-

ported to act as an oncogene in bladder cancer related to cell migration and invasion[27]; however, no functional relevance to breast cancer has been reported to date. HOTAIR, which is one of the lncRNAs that has been associated with metastasis[15], is found to participate in sponges of all the subtypes except HER2. Similarly, miRNAs hsa-miR-142, hsa-miR-150, and hsa-miR-155 participate in ceRNA interactions of all subtypes.

There are also RNAs that take part in sponges of exclusively in a single subtype (Table S4 in Supp. File1). For example, the lncRNA C17orf44 is specific to HER2 (Figure 2(A)) while hsa-miR-342 is only found in Basal ceRNA interactions (Figure 2(B)). Similarly, some RNAs are only regulated in single subtype (see Supp. File 3 for all the mRNAs in the interactions and for only the prognostic mRNAs see Supp. File 4). These subtype specific RNAs are of great value for understanding the dsyregulated cellular mechanisms in each subtype.

The lncRNA-mRNA networks for each subtype where each node denotes a lncRNA or an mRNA while an edge represents an interaction through a shared miRNA is shown in Figure S4 (Supp. File1) and Supp. File 8. The number of nodes and edges are provided in Table S5 (Supp. File1). In Luminal A, lncRNA LOC100188949 regulates majority of the sponge interactions, while C21orf34 also form a smaller connected component of its own. In Luminal B, KIAA0125 is at the center of the many interactions while a few other lncRNAs among them are HOTAIR and C21orf34 mediates a small number of interactions. Basal and HER2 include a large number of interactions. In Basal, among others HCP5, MIR155HG, MIAT are the hubs of the network. In HER2 KIAA0125, LOC100188949 and LOC100233209 are the top 3 largest hubs.

We find that ceRNA interactions often contain the same lncRNA-miRNA pair but for those interactions mRNAs vary. As an example, HER2 subtype specific C14orf72-hsa-miR-150 lncRNA-miRNA pair interacts with 45 different mRNAs, the same is not true for lncRNA-mRNA pairs. Number of ceRNA interactions per lncRNA-miRNA pairs are provided in Figure S5 (Supp. File1). We also analyze the data by clustering mRNAs that participate in a sponge with the same lncRNA-miRNA pair based on mRNA expression correlation. The view of the identified sponges in terms of these clusters are provided in Supp. File 5.

Table 2: List of lncRNAs & miRNAs that are found to participate in sponges of all four subtypes.

miRNA	lncRNA
hsa-miR-142	LOC100188949
hsa-miR-196a-1	C5orf58
hsa-miR-127	LOC100233209
hsa-miR-155	HCP5
hsa-miR-150	KIAA0125
hsa-miR-196a-2	C21orf34
hsa-miR-125b-2	MIR155HG

3.2 Spatially Proximal ceRNAs Interactions

Although regulatory interactions can take place between molecules encoded in different chromosomes, spatial proximity often hints a tight regulatory coordination. To characterize sponge interactions, we examine the genomic locations of the RNAs that participate in the genome. The sponge interactions for which the participating RNAs are within 100KB distance of each other are identified. The most striking case is the set of sponge interactions that take place between HOTAIR, hsa-miR-196a miRNAs and HOXC genes (Figure 3(A)). These sponge interactions are identified in all subtypes except HER2. Both HOTAIR, has-miR-196a-1, and hsa-miR-196a-2 are all spatially proximal to the HOX gene clusters on chromosome 12 (Figure 3(B)). HOX genes are highly conserved transcription factors that take master regulatory roles in numerous cellular process including development, apoptosis, receptor signaling, differentiation, motility and angiogenesis. Their aberrant expression is reported in multiple cancer types [4]. HOXA are reported to have altered expression in breast and ovarian cancers; other HOX genes are also associated with other tumor types, including colon, lung, and prostate cancer. The other lncRNA partner of this sponge interaction is HOTAIR. Upregulation

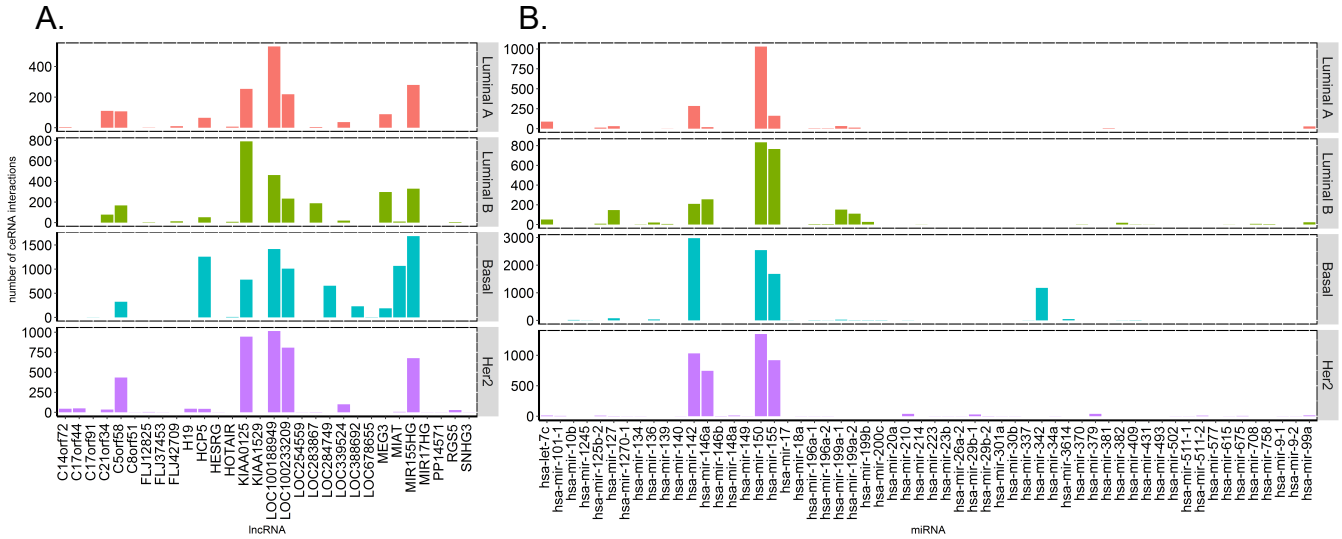

Figure 2: Number of ceRNA interactions discovered that A) lncRNAs and B) miRNAs take part in each breast cancer subtype ($t = 0.2$).

of HOTAIR is associated with metastatic progression and low survival rates in breast, colon and liver cancer patients [14, 20, 27, 49, 8, 53, 41, 43, 25]. The complete list of sponge interactions whose members exhibit such spatial proximity at least between two RNAs in the sponge are provided in Supp. File 6.

3.3 Functional Enrichment Analysis of mRNAs in ceRNAs

To understand the patterns of pathways related to identified sponges, we conducted pathway enrichment of mRNAs that participate in the sponges separately. The top enriched pathways are found to be common across subtypes (see S6-S7 Figures, Supp. File1) and these pathways are mostly related to the immune system and signaling pathways, which are known to critical for breast cancer[13]. Interestingly, interferon alpha/beta signaling pathway is among the top pathways for Basal subtype (p-value 7.20×10^{-23}) while it is not found enriched in other subtypes (p-value cut-off 0.05 and FDR cutoff 1×10^{-4}).

The overlap between the enriched pathways in different subtypes are shown on a Venn diagram (Figure S8 in Supp. File1). The list of pathways that are found enriched only in a single subtype are listed in Table S7 in Supp. File1 with p-value cut-off 0.05 and FDR cutoff 1×10^{-4}. Interestingly, PI3K pathway is found to be enriched specifically in Luminal A. This is interesting as the most frequently mutated gene in Luminal A is PIK3CA (45% of the patients in TCGA) and there are PIK3CA mutations that are specific to this subtype[33]. Complement cascade induces cell proliferation which causes carcinogenesis including invasion, cell death and metastasis[38], which are Basal subtype characteristics. We detected C2, C3, C3AR1, C4A, C7 complement genes in Basal ceRNA interactions. Consequently, complement cascade pathway may significant for Basal subtype. Integrin signaling widely studied in breast cancer literature since integrins incorporate breast cancer progression[24]. Moreover, integrin have role in cell migration and tissue invasion. Thus, they drive tumor cell to metastasis[24]. HER2 subtype specific enriched pathways contain integrin signaling pathways (TableS7 in Supp. File1).

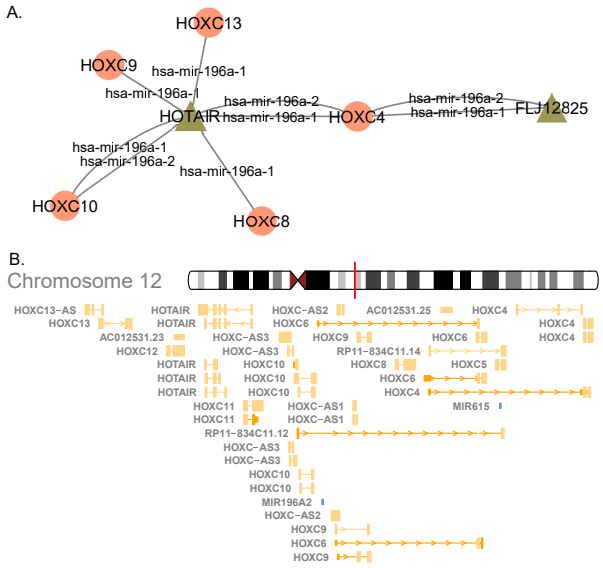

Figure 3: A) Network of sponge interactions between HOTAIR, hsa-miR-196a miRNAs and HOXC genes. The circles denote lncRNAs and triangles denote mRNAs. An edge exists between a lncRNA and an mRNA if there is a sponge interaction between them; the edge label indicates the miRNA that regulates the interaction. B) The genomic locations of the sponge interactions on chromosome 12.

3.4 Prognostic Sponge Interactions

To identify ceRNA interactions with prognostic value, for each of the identified sponge we checked whether the sponge expression pattern divides the patients into groups that differ in their survival probability. To further verify that this difference is due to the interaction and not due to an individual RNA molecule that participates in the sponge, we filter them further. We only consider interactions where there is significance difference in survival when patients are grouped based on ceRNA expression pattern while there is no significance due to a grouping based on a single RNA molecules' expression pattern. These prognostic ceRNA interactions are ranked based on f-score (details in 2.4) and are provided as a list in Supp. File 7 and the network of interactions are shown in Figure 10 in Supp. File1. KM plots for the two examples are shown in Figure 4. The distribution of lncRNAs that participate in the prognostic sponges are provided in Figure S9 in Supp File1 and the network of interaction among prognostic sponges are provided in Figure 5 and the Supp. Cytoscape file in Supp. File 9.

4 Conclusion

In this study, we focus on a specific type of interaction between lncRNAs, mRNAs and miRNAs. lncRNAs are known to reduce the relative amount of available miRNAs to mRNAs by binding to shared miRNAs. This way, they serve as sponges for miRNA by preventing the repression of the target mRNAs. We characterize these lncRNA-mediated sponge interactions in each breast cancer molecular subtypes. Our analysis method integrates the statistical analysis of gene expression profiles of lncRNA, mRNA and miRNAs of patients and reveals that there are distinct interactions specific to each subtype that are also supported by miRNA target information. We also find that breast cancer patients who possess certain expression patterns pertaining to certain sponge RNAs have different survival distributions.

References

[1] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel. Predicting effective microrna target sites in mammalian mrnas. *elife*, 4:e05005, 2015.

[2] D. P. Bartel. Micrornas: target recognition and regulatory functions. *cell*, 136(2):215–233, 2009.

[3] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie. Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8):R90, 2010.

[4] S. Bhatlekar, J. Z. Fields, and B. M. Boman. Hox genes and their role in the development of human cancers. *Journal of molecular medicine*, 92(8):811–823, 2014.

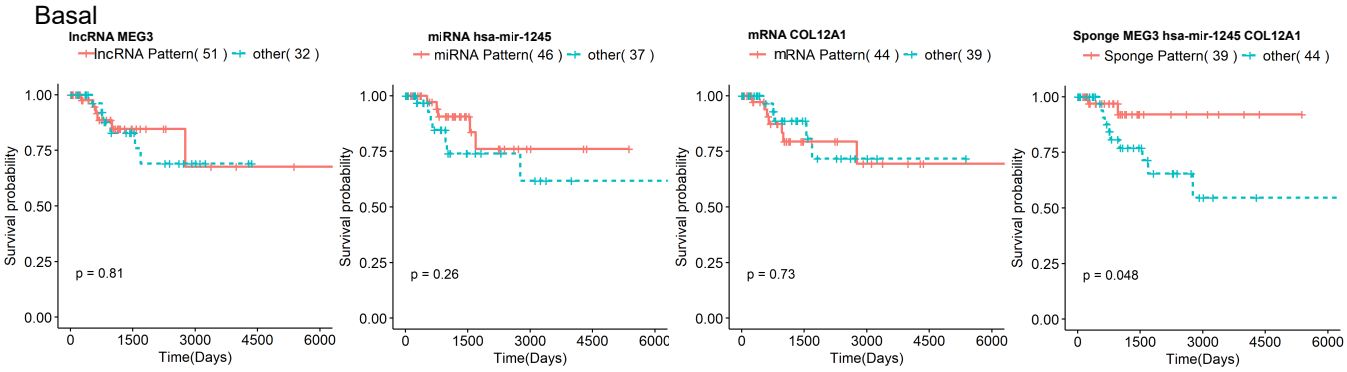

Figure 4: Kaplan Meir survival plots when patients are divided based on individual expression patterns of the RNAs (the first three plots in each panel) and when patients are divided based on the sponge expression pattern (4th plot) for MEG3, hsa-miR-1245, COL12A1 sponge

[5] C. Blenkiron, L. D. Goldstein, N. P. Thorne, I. Spiteri, S.-F. Chin, M. J. Dunning, N. L. Barbosa-Morais, A. E. Teschendorff, A. R. Green, I. O. Ellis, et al. Microrna expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome biology*, 8(10):R214, 2007.

[6] G. A. Calin and C. M. Croce. Microrna signatures in human cancers. *Nature reviews. Cancer*, 6(11):857, 2006.

[7] M. Cesana, D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, and I. Bozzoni. A long noncoding rna controls muscle differentiation by functioning as a competing endogenous rna. *Cell*, 147(2):358–369, 2011.

[8] Y.-C. Chiu, T.-H. Hsiao, Y. Chen, and E. Y. Chuang. Parameter optimization for constructing competing endogenous rna regulatory network in glioblastoma multiforme and other cancers. *BMC genomics*, 16(4):S1, 2015.

[9] C.-H. Chou, N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu, et al. mirtarbase 2016: updates to the experimentally validated mirna-target interactions database. *Nucleic acids research*, 44(D1):D239–D247, 2015.

[10] S. Das, S. Ghosal, R. Sen, and J. Chakrabarti. lncedb: database of human long noncoding rna acting as competing endogenous rna. *PloS one*, 9(6):e98965, 2014.

[11] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. M. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101, 2012.

[12] H. Dweep, C. Sticht, P. Pandey, and N. Gretz. mirwalk–database: prediction of possible mirna binding sites by walking the genes of three genomes. *Journal of biomedical informatics*, 44(5):839–847, 2011.

[13] P. Eroles, A. Bosch, J. A. Pérez-Fidalgo, and A. Lluch. Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer treatment reviews*, 38(6):698–707, 2012.

[14] J. M. Franco-Zorrilla, A. Valli, M. Todesco, I. Mateos, M. I. Puga, I. Rubio-Somoza, A. Leyva, D. Weigel, J. A. García, and J. Paz-Ares. Target mimicry provides a new mechanism for regulation of microrna activity. *Nature genetics*, 39(8):1033, 2007.

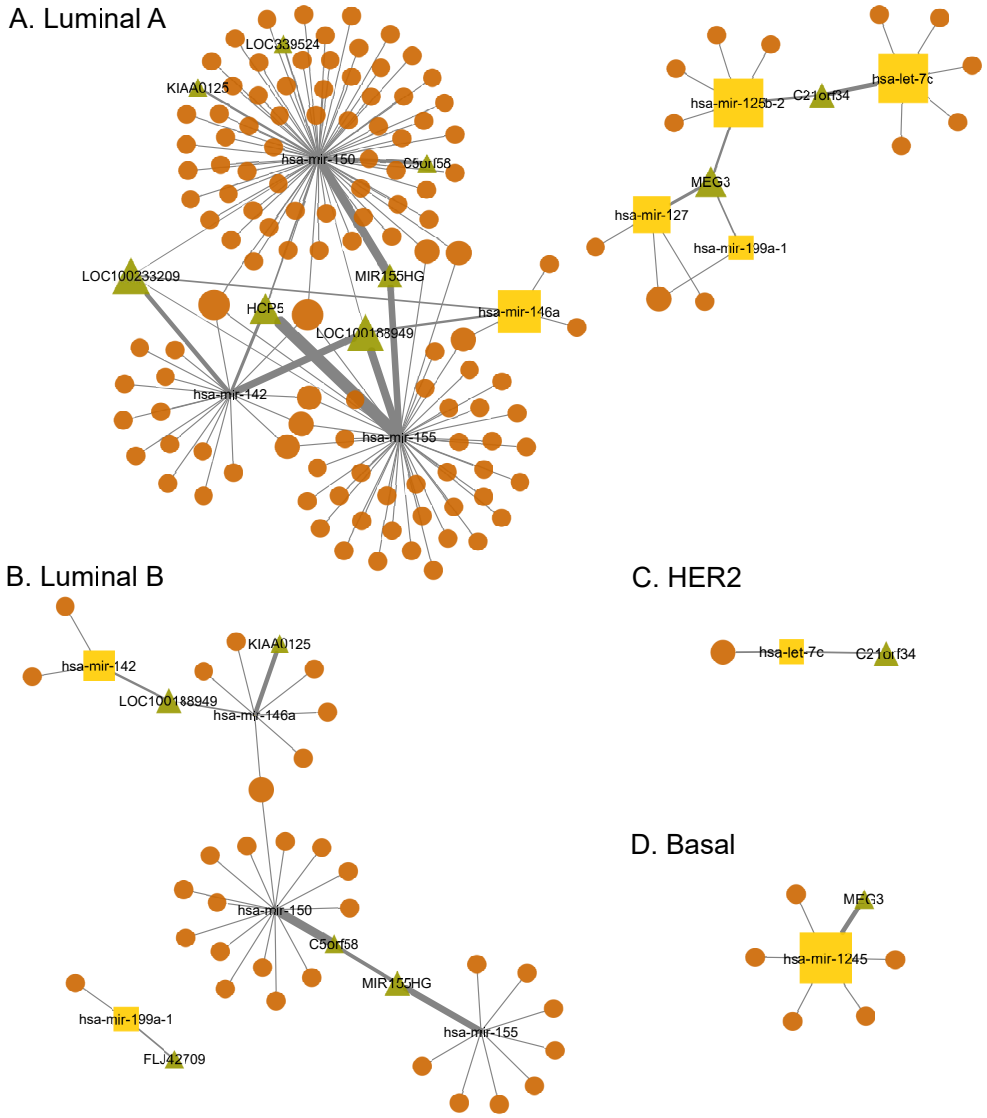

Figure 5: lncRNA-miRNA-mRNA network for all breast cancer subtypes. lncRNAs are represented with green triangle symbol, mRNAs are represented with orange ellipse symbol and miRNAs are with yellow rectangle. Each node size is scaled by its degree, the number of edges incident to the nodes and edge width is scaled by number of occurrence of node pair. Network was constructed using the Cytoscape(v3.4.0)[40].

[15] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M.-C. Tsai, T. Hung, P. Argani, J. L. Rinn, et al. Long noncoding rna hotair reprograms chromatin state to promote cancer metastasis. *nature*, 464(7291):1071, 2010.

[16] D. P. Harrington and T. R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[17] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[18] T.-M. Huang et al. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091, 2010.

[19] A. Jeggari, D. S. Marks, and E. Larsson. mircode: a map of putative microrna target sites in the long non-coding transcriptome. *Bioinformatics*, 28(15):2062–2063, 2012.

[20] F. A. Karreth and P. P. Pandolfi. cerna crosstalk in cancer: when ce-bling rivalries go awry. *Cancer discovery*, 3(10):1113–1121, 2013.

[21] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microrna target recognition. *Nature genetics*, 39(10):1278, 2007.

[22] L. Kong, Y. Zhang, Z.-Q. Ye, X.-Q. Liu, S.-Q. Zhao, L. Wei, and G. Gao. Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, 35(suppl_2):W345–W349, 2007.

[23] S. Kurozumi, Y. Yamaguchi, M. Kurosumi, M. Ohira, H. Matsumoto, and J. Horiguchi. Recent trends in microrna research into breast cancer with particular focus on the associations between micrornas and intrinsic subtypes. *Journal of human genetics*, 2016.

[24] A. W. Lambert, S. Ozturk, and S. Thiagalingam. Integrin signaling in mammary epithelial cells and breast cancer. *ISRN oncology*, 2012, 2012.

[25] J. Li, J. Wang, Y. Zhong, R. Guo, D. Chu, H. Qiu, and Z. Yuan. Hotair: a key regulator in gynecologic cancers. *Cancer cell international*, 17(1):65, 2017.

[26] S. Lin and R. I. Gregory. Microrna biogenesis pathways in cancer. *Nature reviews cancer*, 15(6):321–333, 2015.

[27] W. Lv, L. Wang, J. Lu, J. Mu, Y. Liu, and P. Dong. Long noncoding rna kiaa0125 potentiates cell migration and invasion in gallbladder cancer. *BioMed research international*, 2015, 2015.

[28] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*, 5(11):e13984, 2010.

[29] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[30] K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos. A pattern-based method for the identification of microrna binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217, 2006.

[31] R. Mojena. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, 20(4):359–363, 1977.

[32] M. Mutlu, Ö. Saatci, S. A. Ansari, E. Yurdusev, H. Shehwana, Ö. Konu, U. Raza, and Ö. Şahin. mir-564 acts as a dual inhibitor of pi3k and mapk signaling networks and inhibits proliferation and invasion in breast cancer. *Scientific reports*, 6:32541, 2016.

[33] C. G. A. Network et al. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61, 2012.

[34] P. Paci, T. Colombo, and L. Farina. Computational analysis identifies a sponge interaction network between long non-coding rnas and messenger rnas in human breast cancer. *BMC systems biology*, 8(1):83, 2014.

[35] M. D. Paraskevopoulou, I. S. Vlachos, D. Karagkouni, G. Georgakilas, I. Kanellos, T. Vergoulis, K. Zagganas, P. Tsanakas, E. Floros, T. Dalamagas, et al. Diana-lncbase v2: indexing microrna targets on non-coding transcripts. *Nucleic acids research*, 44(D1):D231–D238, 2016.

[36] J. R. Prensner and A. M. Chinnaiyan. The emergence of lncrnas in cancer biology. *Cancer discovery*, 1(5):391–407, 2011.

[37] U. Raza, Ö. Saatci, S. Uhlmann, S. A. Ansari, E. Eyüpoğlu, E. Yurdusev, M. Mutlu, P. G. Ersan, M. K. Altundağ, J. D. Zhang, et al. The mir-644a/ctbp1/p53 axis suppresses drug resistance by simultaneous inhibition of cell survival and epithelial-mesenchymal transition in breast cancer. *Oncotarget*, 7(31):49859, 2016.

[38] M. J. Rutkowski, M. E. Sughrue, A. J. Kane, S. A. Mills, and A. T. Parsa. Cancer and the complement cascade. *Molecular Cancer Research*, 8(11):1453–1465, 2010.

[39] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[40] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[41] R. Siegel, J. Ma, Z. Zou, and A. Jemal. Cancer statistics, 2014. *CA: a cancer journal for clinicians*, 64(1):9–29, 2014.

[42] K. Song. Testing conditional independence via rosenblatt transforms. 2007.

[43] K. P. Sørensen, M. Thomassen, Q. Tan, M. Bak, S. Cold, M. Burton, M. J. Larsen, and T. A. Kruse. Long non-coding rna hotair is an independent prognostic marker of metastasis in estrogen receptor-positive primary breast cancer. *Breast cancer research and treatment*, 142(3):529–536, 2013.

[44] L. Su and H. White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

[45] K. C. Wang and H. Y. Chang. Molecular mechanisms of long noncoding rnas. *Molecular cell*, 43(6):904–914, 2011.

[46] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li. Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6):e74–e74, 2013.

[47] L. Wang and J. Wang. Microrna-mediated breast cancer metastasis: from primary site to distant organs. *Oncogene*, 31(20):2499, 2012.

[48] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[49] T. Xia, Q. Liao, X. Jiang, Y. Shao, B. Xiao, Y. Xi, and J. Guo. Long noncoding rna associated-competing endogenous rnas in gastric cancer. *Scientific reports*, 4, 2014.

[50] S. Ye, L. Yang, X. Zhao, W. Song, W. Wang, and S. Zheng. Bioinformatics method to predict two regulation mechanism: Tf–mirna–mrna and lncrna–mirna–mrna in pancreatic cancer. *Cell biochemistry and biophysics*, 70(3):1849–1858, 2014.

[51] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287, 2012.

[52] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[53] M. Zhou, X. Wang, H. Shi, L. Cheng, Z. Wang, H. Zhao, L. Yang, and J. Sun. Characterization of long non-coding rna-associated cerna network to reveal potential prognostic lncrna biomarkers in human ovarian cancer. *Oncotarget*, 7(11):12598, 2016.