Inferring competing endogenous RNA (ceRNA) interactions in cancer

Ziynet Nesibe Kesimoglu[1], Serdar Bozdag[1]

1 Marquette University, Department of Computer Science, Milwaukee, WI, USA

Abstract

To understand driving biological factors for cancer, regulatory circuity of genes needs to be discovered. Recently, a new gene regulation mechanism called competing endogenous RNA (ceRNA) interactions has been discovered. Certain genes targeted by common microRNAs (miRNAs) "compete" for these miRNAs, thereby regulate each other by making others free from miRNA regulation. Several computational tools have been published to infer ceRNA networks. In most existing tools, however, expression abundance, collective regulation, and groupwise effect of ceRNAs are not considered. In this study, we developed a computational pipeline named Crinet to infer cancer-associated ceRNA networks addressing critical drawbacks. Crinet considers lncRNAs, pseudogenes, and all mRNAs as potential ceRNAs and incorporates a network deconvolution method to exclude the amplifying effect of ceRNA pairs in terms of ceRNA regulation. We tested Crinet on breast cancer data in TCGA. Crinet inferred reproducible ceRNA interactions and groups, which were significantly enriched in cancer-related genes and processes. We validated our results using protein expression benchmarks. Crinet outperformed other tool predicting gene expression change in knockdown assays, and showed better performance for antibodies following miRNA transfection. Top high-degree genes in the inferred network included known suppressor/oncogene lncRNAs of breast cancer showing the importance of non-coding RNA's inclusion for ceRNA inference.

1 Introduction

MicroRNAs (miRNAs) are small RNA types that regulate gene expression by binding to them. Recently, a new regulatory layer related to miRNAs has been discovered [21]: certain RNAs targeted by common miRNAs "compete" for these miRNAs and thereby regulate each other indirectly by making the other RNA(s) free from miRNA regulation. Such indirect interactions between RNAs are called competing endogenous RNA (ceRNA) interactions having an important role in disease including cancer [13, 19, 28, 34].

There is a regulation multiplicity between miRNAs and RNAs, meaning that a miRNA could have multiple RNA targets, and an RNA could be targeted by multiple miRNAs. Given the enormous number of RNAs and difficulty of deciphering microRNA binding targets accurately, identifying ceRNA interactions experimentally is cost and labor prohibitive. Therefore, computational tools are crucial to infer ceRNA interactions in complex genomes like human. For the rest of the paper, "genes" refers to mRNAs, lncRNAs, and pseudogenes in our analysis.

Despite existing computational tools [4, 5, 8, 27], they have crucial drawbacks to address with more advanced methods. Most studies start with differentially expressed coding genes among matching normal and tumor samples to have a computationally manageable number of miRNA-target interactions, however, the inclusion of all genes that could be involved in ceRNA regulation including noncoding ones is important to understand the comprehensive regulation.

Current tools compute only pairwise ceRNA interactions or get ceRNA modules, however, inferring groupwise ceRNA interactions should be considered since several ceRNAs could work together to sequester miRNA(s) targeting key ceRNA(s). While considering ceRNA groups, collective measurements could be included to show the increasing preference of being a group rather than being single.

Also, in the existing tools, a miRNA could be assigned to mediate many interactions without considering whether miRNA expression abundance is enough or not. Furthermore, since ceRNAs positively regulate each other, two genes having many common ceRNA partners might be inferred just because of the amplifying effect of ceRNA partners in terms of ceRNA regulation, so excluding these indirect interactions is important to eliminate false-positive interactions from ceRNA network.

In this study, we developed a computational tool named Crinet (CeRna Interaction NETwork) to infer cancer-associated pairwise and groupwise ceRNA interactions genome-wide to address the aforementioned drawbacks. After integrating expression datasets and binding scores of miRNA-target interactions, we excluded if expression abundance of miRNA and target is not sufficient. To get ceRNA pairs under strong regulation jointly, we considered collective regulation for ceRNA inference. To cope with the amplifying effect of ceRNAs in terms of ceRNA regulation, we excluded indirect ceRNA pairs which are potentially inferred just because of the significant overlapping of common ceRNA partners. Considering the significant role in ceRNA regulation, we included noncoding (long noncoding RNAs (lncRNAs) and pseudogenes) in addition to all coding RNAs (mRNAs), even not differentially expressed. In addition to ceRNA interactions, we inferred ceRNA groups based on the premise that several ceRNAs could work together to sequester miRNA(s) with the high preference to be a group rather than being single. Getting the benefit of multiple biological datasets, this approach facilitates a better understanding of ceRNA regulatory mechanisms solving important drawbacks, which would shed light on the underlying complex regulatory circuitry in disease conditions.

Crinet was applied to breast cancer dataset inferring pairwise and groupwise ceRNA interactions. Since we built a ceRNA network relying on miRNA-target interactions, we evaluated Crinet-selected miRNA-target interactions with protein expression benchmarks having increasing performance following applied filtering. To validate the inferred ceRNA effect on expression change using an outside database, We used the LINCS [24] knockdown assays and observed that expression change in inferred ceRNAs was highly affected following knockdown of their ceRNA partners. Additionally, inferred ceRNAs were significantly associated with known cancer genes and significantly enriched with cancer-related processes. Moreover, ceRNA groups were associated with some biological processes suggesting that groupwise ceRNA interactions might involve in important processes because of their competition for common regulators.

2 Materials and Methods

Crinet is a computational tool to infer cancer-associated ceRNA network genome-wide with pairwise and groupwise interactions. Briefly, the first step is data preparation. The second step is computing miRNA-target interactions integrating with expression datasets and considering sufficient abundance. Starting with final miRNA-target

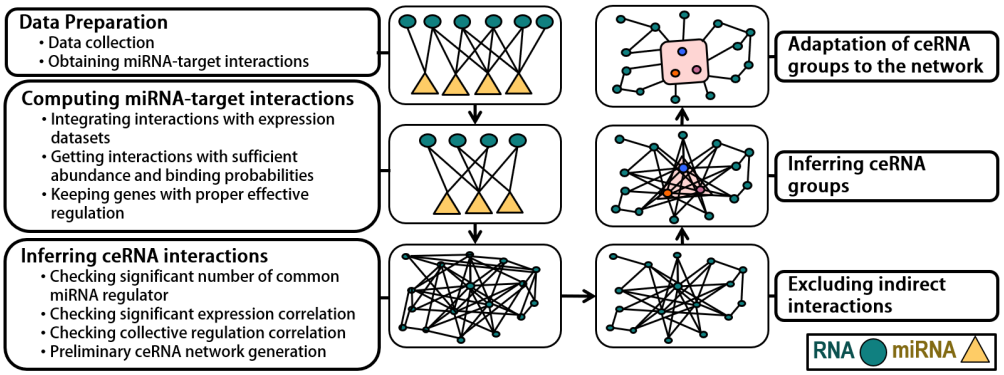

Figure 1. Crinet Pipeline

interactions, ceRNA interactions are inferred in the third step. At the last step, groupwise ceRNA interactions are inferred from the ceRNA network and adapted to the network. Each step of Crinet is explained as detailed:

2.1 Data preparation

As a starting point, Crinet needs miRNA-target interactions including binding scores, gene-centric copy number aberration (CNA), and expression datasets. If binding scores are not available, the same score for all interactions could be used.
Data collection: To collect datasets for Crinet, we used TCGAbiolinks R package [6] and obtained the datasets from the Cancer Genome Atlas (TCGA) project including gene expression, miRNA expression, and CNA for totally 1107 breast (BRCA) tumor samples. We preprocessed each datatype separately obtaining normalized expression values (gene expression as FPKM and miRNA expression as RPM) and filtered lowly expressed genes (if FPKM < 1 and RPM $= 0$ for 150 samples). To get gene-centric CNA from a segmented dataset for each RNA, we ran CNTools R package [31].
Obtaining miRNA-target interactions: To have publicly available putative miRNA-target interactions, we obtained all conserved and nonconserved mRNA-target interactions with weighted context++ scores from TargetScan [1]. Since we used lncRNAs and pseudogenes in addition to mRNAs as potential targets, we ran TargetScan separately for lncRNAs and pseudogenes, providing a constant dummy ORF length and supplying all transcript sequence as 3'UTR considering all the sequence is not translated. If multiple scores exist for a miRNA-target pair due to multiple transcripts, we used the mean score.

Since there is a corresponding score in TargetScan results even for weak miRNA-target interactions, we kept the top 40% of ranked interactions with respect to weighted context++ scores. To fit different distributions of scores for mRNAs, lncRNAs, and pseudogenes, we combined z-normalized scores from each data type and applied min-max normalization after having all the scores in the range of -1 and 1 (assigning low scores to -1, and high scores to 1). We used these normalized scores as a weight for each miRNA-target interaction assuming that these scores show the binding strength between miRNA and gene pairs.

2.2 Computing miRNA-target interactions

In this section, we computed putative miRNA-target interactions leveraging expression datasets and considering abundance sufficiency.

Integrating interactions with expression datasets (Correlation filtering): Since miRNAs are known to generally repress their target genes, we kept only miRNA-target pairs with significant anticorrelation among expression of miRNA and its targets. By default, we used the correlation coefficient threshold of -0.1, which was the median of the correlation coefficient. We also applied random sampling with replacement to compute correlation for 1000 times and required that the correlation threshold was satisfied for > 99% of the samplings.

Getting interactions with sufficient abundance and binding probabilities (Abundance filtering): In the miRNA-target interaction sets, a miRNA could be assigned to mediate thousands of ceRNA interactions without considering expression sufficiency, and similarly an RNA could be assigned for many miRNA regulators as potential target. To quantify the sufficiency for our putative interactions, we introduced an *Interaction Regulation* (IR) score formulated as:

$$IR_{ik} = \left(Exp(miRNA_k) \frac{Exp(RNA_i).score_{ik}}{\sum_{j \in miRNA_k's\ targets} Exp(RNA_j).score_{jk}} \right) * \left(Exp(RNA_i) \frac{Exp(miRNA_k).score_{ik}}{\sum_{j \in RNA_i's\ regulators} Exp(miRNA_j).score_{ij}} \right) \tag{1}$$

where IR_{ab} is the IR score of RNA_a and $miRNA_b$ through samples, $Exp(a)$ is the expression of a through samples, $score_{ab}$ is the normalized binding score for $miRNA_b$-RNA_a interaction. Basically, we calculated the sum of partial miRNA regulation on each RNA through samples by splitting the miRNA expression based on binding score proportions. Similarly splitting gene expression, we checked sufficiency of target abundance. Using this formula, we kept the final miRNA-target interactions having significantly high IR scores (log score > -4.89, which was third quartile of all IR scores through samples) for >80% of the samples. Supplementary Methods A.1 shows the confirmation of IR score as detailed.

Keeping genes with proper effective regulation: To exclude genes from analysis if they are not under strong miRNA regulation based on final miRNA-target interactions, we calculated correlation among gene expressions and their effective miRNA regulation. We called the sum of partial miRNA regulation on each RNA as *Effective Regulation* (ER) score formulated as:

$$ER(RNA_j) = \sum_{k \in RNA_j's\ regulators} \frac{Exp(miRNA_k) score_{jk}}{\sum_{i \in miRNA_k's\ targets} score_{ik}} \tag{2}$$

where $Exp(miR_a)$ is the expression of $miRNA_a$ and $score_{ab}$ is the normalized binding score for $miRNA_b$-RNA_a interaction. To keep genes with proper effective regulation by miRNAs, we filtered out genes without strong anticorrelation for gene expression and effective miRNA regulation, assuming that they are not having strong miRNA regulation for our specific dataset. By default, we kept genes having correlation, which was the correlation coefficient threshold of -0.01. We also applied random sampling with replacement to compute correlation for 1000 times and required that the correlation threshold was satisfied for > 99% of the samplings. We remained with our candidate genes to use for further analysis to generate all candidate ceRNA pairs.

2.3 Inferring ceRNA interactions

To infer ceRNA interactions, we generated all possible gene-gene combinations using candidate genes and filtered them based on the following criteria:

Checking significant number od miRNA regulator: We checked the number of common miRNA regulators for RNA pairs so that ceRNAs will have miRNAs to compete for. We kept RNA pairs having at least one common miRNA regulator with a significant overlap for common miRNAs (hypergeometric p-value < 0.01).
Checking significant expression correlation: Since the ceRNA pairs indirectly positively regulate each other and CNA considerably affects the expression values, we kept the candidate ceRNA pairs having significant partial correlation when excluding the CNA effect from gene expression. By default, we used the correlation coefficient threshold of 0.55 for partial correlation and 0.01 partial correlation p-value, which was the third quartile of the positive correlation coefficient (see Supplementary Figure 3).
Checking collective regulation correlation: If there exists ceRNA regulation among two RNAs then both RNAs compete for common miRNAs and those miRNAs affect both RNAs simultaneously. We called this regulation for ceRNA pairs *collective regulation*, and we kept a ceRNA pair if there exists a significant negative correlation between the sum of their ER and expression. By default, we used the correlation coefficient threshold of -0.01, which was the same as the RNA selection with proper effective regulation at the end of miRNA-target interaction computation.

We also applied random sampling with replacement to compute partial correlation and collective regulation correlation for 100 times and required that the correlation threshold was satisfied for $> 99\%$ of the samplings.
Preliminary ceRNA network generation: Using the remaining interactions, we generated our preliminary ceRNA network where each node is a ceRNA and each edge represents each existing ceRNA interaction.
Excluding indirect interactions from network: Since ceRNA pairs are known to positively regulate each other, having common ceRNA partners among any two RNAs will increase the correlation between their expression indirectly. If a ceRNA pair has many such indirect interactions then the regulation on this pair will be amplified. Thus, a ceRNA interaction will be inferred among two RNAs even these RNAs do not have direct ceRNA regulation. To exclude such indirect interactions from our network, we employed a network deconvolution algorithm [9] and kept the top one-third of high ranked interactions as our final ceRNA network.

2.4 Inferring ceRNA groups

In ceRNA regulation, each ceRNA pair compete for common miRNA(s) and act as a decoy to make the other RNA free from miRNA regulation. However, this competition could occur between more than two RNAs. Based on this premise, Crinet infers ceRNA groups besides ceRNA interactions.

To obtain ceRNA groups, we utilized one of the popular community detection algorithm Walktrap [18] on the weighted ceRNA network where weights were normalized partial correlation. For each iteration of the Walktrap algorithm, we kept a group if it satisfies all the group conditions (see below). Otherwise, we split the group until it satisfies all the group conditions or eliminated due to its small size. Our group requirements are listed as follows:
Common miRNA regulator: To have a common miRNA regulator to compete for, all the group members are required to have at least one common miRNA regulator.
Strong collective regulation: If the ceRNAs sequester common miRNAs collectively, they will be under the strong effect of miRNA regulation even when they are considered as a group. Therefore, we checked collective regulation for whole group
Strong regulation effect for a group than for a single ceRNA: CeRNAs in a group were expected to have a stronger miRNA regulation effect as a group than as individual ceRNAs. Thus, collective regulation was required to be more significant (i.e., reduced) for a group compared to single RNA's correlation among ER score and

expression showing that group of genes collaborate more than being single (satisfying 90% of the group members). Moreover, to guarantee that most of the genes in the group will be under strong regulation effect, we required to have an average decrease comparing the single to the group correlation among ER score and gene expression (average correlation difference > 0).

Adaptation of ceRNA groups to the network: To hold inference consistency of the ceRNA network, we expected the group to satisfy pairwise inference conditions among the group and each ceRNA partner of group members. We called ceRNA partners of any member in the group as neighbor RNA and kept the groups satisfying any two of three inference conditions among the group and most of the neighbor genes ($> 90\%$ of neighbors). Inference conditions are having at least one common miRNA regulator, having a strong Pearson correlation, and having strong collective regulation among all the group and neighbor.

With these requirements, we obtained 81 ceRNA groups after applying 1508 Walktrap runs. While keeping miRNA regulation at some level, we had stronger relations for the group than single without changing the inference consistency in the ceRNA network. After this step, we had our grouped ceRNA network with nodes as individual RNA or group of genes and edges between inferred nodes.

3 Results

We tested Crinet on breast (BRCA) tumor samples from the Cancer Genome Atlas (TCGA) project (see Section 2.1 for details). We computed miRNA-target interactions (Table 1) and used those interactions to infer ceRNA interactions (Table 2).

Table 1. Number of miRNA-target interactions for each filtering step in Crinet. ALL: All obtained miRNA-target interactions; ALL40: Top 40% of ALL interactions based on weighted context++ score; CORR: After correlation-based filtering on ALL40 interactions (<-0.1); CORR+ABUN: After abundance-based filtering on CORR interactions (>-4.89 for <%80 of samples)

Step	Interaction #	miRNA #	Gene #
ALL	8,193,904	888	28,129
ALL40	3,285,523	888	27,999
CORR	535,284	888	23,099
CORR+ABUN	165,937	888	21,261

To check the scale-free property and specificity of Crinet, we examined the inferred network (Supplementary Methods A.2 for details). Since biological networks generally exhibit scale-free property, we computed the inferred network's degree probability distribution function. Our inferred ceRNA network had a negative slope with high fitness ($R^2 = 0.93$), indicating that the inferred ceRNA network was scale-free (Supplementary Figure 2). To show the specificity of Crinet, we compared our inferred ceRNA pairs from different regulatory layers, namely protein-protein interactions (PPIs) and transcription factor (TF)-gene interactions. We collected 1,663,810 TF-target interactions from TRRUST v2 [10] database and the ENCODE Transcription Factor Targets dataset [7], and 1,847,774 PPIs from BIOGRID v3.5.186 [22]. Within all inferred ceRNA interactions, very few interactions were TF-gene interactions (0.46%) and PPIs (0.51%).

To check the reproducibility based on different datasets, robustness to different hyperparameters, and importance of each individual step in ceRNA inference, we examined Crinet more detailed (see Supplementary Methods A.3 and A.4 for details). To check the reproducibility of Crinet based on different datasets, we ran the whole

Table 2. Number of remained ceRNA pairs for each filtering step. ALL: All candidate pairs after keeping proper genes with effective regulation; Step1: pairs with at least one common miRNA regulator with a significant overlap for common miRNAs; Step2: pairs after filtered based on partial correlation among gene expressions excluding copy number aberration effect; Step3: pairs after filtered based on collective regulation; Step4: pairs after applying random sampling for Step2 and Step3; Step5: pairs after applying the network deconvolution method to exclude indirect ceRNA pairs.

Step	Interaction #	Gene #
ALL	212,994,480	20,640
Step1	16,082,000	20,640
Step2	247,885	11,910
Step3	209,220	11,726
Step4	52,858	7,263
Step5	17,443	4,494

pipeline multiple times with different datasets and had high overlapping among different runs for interactions and unique ceRNAs (Supplementary Table 4). When we checked the distribution of consistently overlapping ceRNAs, the degree distribution of these ceRNAs was much higher as compared to the overall degree distribution (p-value $< 2.10^{-16}$) suggesting that consistently inferred ceRNAs were the important ceRNAs highly involving in our inferred ceRNA network. Moreover, to examine the importance of each step, we skipped major steps in ceRNA inference. Skipping individual steps made a substantial difference in inferred results (Supplementary Table 1) and the ceRNA interaction obtained from the modified network was not good at predicting gene expression change of its ceRNA partner or potentially inferred more false ceRNA interactions. Besides, to show that Crinet was robust to hyperparameter selection, we modified hyperparameters in ceRNA inference and we had a high number of overlapping interactions, still keeping half of the interactions with highly-overlapping ceRNAs (Supplementary Table 6). This analysis suggests that even each hyperparameter had a significant effect on the results, they consistently kept similar ceRNA interactions with a high number of overlapped ceRNAs. Furthermore, we showed that the selection of ranking scores in the incorporated network deconvolution algorithm [9] did not change results much, and the evaluation of the obtained network was around the same accuracy even with a wide range for ranking selection (Supplementary Table 5). These analyses showed that the network deconvolution step is important to eliminate indirect ceRNA interactions, but the selection of ranking did not significantly change Crinet results.

3.1 miRNA-target interaction filtering showed increasing performance on protein expression benchmarks

Since we built a ceRNA network relying on miRNA-target interactions, proper selection of these interactions is important; therefore, we evaluated each filtering step of miRNA-target interactions with protein expression benchmarks.

Transfection analysis: We used RPPA output for MDA-MB-231 breast cancer cell line to assess our miRNA-target interactions [2]. Data is available from The Cancer Proteome Atlas (TCPA; https://www.tcpaportal.org/tcpa/download.html) under accession number TCPA00000001. We used 104 unique antibodies, their fold-change for 141 transfected miRNAs, and mock controls from transfection data. We checked each miRNA-target interaction measuring expression fold reduction of each antibody of targets for each targeting miRNA and also for an average of all targeting miRNAs relative to average mock transfections. Table 3 showed the prediction accuracy of each filtering step having ratios greater than 1, confirming the preferential down-regulation

Table 3. Evaluation of each filtering step of computed miRNA-target interactions with miRNA transfection data. Interaction phase shows the expression fold reduction of each antibody of target for its transfected miRNA regulator relative to mock transfection. Gene phase shows average expression fold reduction of each antibody of target for all transfected miRNAs regulators relative to mock transfection. Down-regulated over up-regulated numbers along with the ratio is shown (ratio is expected to be more than 1 to have down-regulation preference. Higher is better). ALL: All obtained miRNA-target interactions; ALL40: Top 40% of ALL interactions; CORR: After expression correlation filtering on ALL40 interactions; CORR+ABUN: After abundance sufficiency filtering on CORR interactions.

Interaction Set	Interaction # in Set	Interaction Phase	Gene Phase
ALL	8,193,904	6,248/4,932=1.3	90/74=1.2
ALL40	3,285,523	613/403=1.5	97/59=1.6
CORR	535,284	183/112=1.6	71/37=1.9
CORR+ABUN	165,937	179/111=1.6	68/39=1.7

of predicted miRNA targets. As compared to the starting interaction set, all other sets showed higher preference and the ratio increased through each step of pipeline showing the positive effect of filtering for each independent interaction. We had 179 down-regulated antibodies of targets, while 111 of them up-regulated following transfection of each miRNA regulator. For proteins ERCC1, BAK1, CTNNA1, PXN, MSH2, XIAP, MAPK3, EEF2K, CAV1, IGFBP2, PRKAA1, PCNA, CASP9, IGF1R, SMAD4, COL6A1, PIK3R1, CHEK1, EIF4E, PTK2, CDK1, SMAD1, BCL2L1, BCL2, LCK, DIABLO, NF2, and EIF4EBP1 including the important genes in breast cancer, the results were consistently down-regulated by their predicted miRNA regulators. As the negative control, we used non-inferred interactions having a random preference for all the filtering, even slightly up-regulation preference (Supplementary Table 2).

Protein expression analysis: To evaluate predicted miRNA-target interactions in [4], the authors focused on the protein ESR1 to validate inferred miRNA-target interactions, showing that ESR1 protein expression in TCGA breast cancer tumors (profiled by RPPA using the antibody ER.alpha.R.V_GBL.9014870) is strongly correlated with the expression of predicted miRNA regulators. Ranking samples based on miRNA expression, the top 10% and bottom 10% samples were compared with protein expression. Similarly, we generated a heatmap showing protein expression for top and bottom 10% samples for our selected miRNAs regulating ESR1. As shown in Figure 2, we had nine miRNAs regulating ESR1 in our selected miRNA-target interactions, while Cupid had six predicted and six known miRNA regulators. Among our selected miRNA regulators, five of them were the same with Cupid's regulators: four known and one predicted. We quantified the anticorrelation among miRNA expression and protein expression by measuring fold-change of mean protein expression for top 10% with respect to the bottom 10% samples. Our results indicated that the expression of Crinet-selected miRNA regulators for ESR1 had high anticorrelation with protein expression. While Cupid had some low fold-change such as hsa-mir-381, our results had high fold-change consistently. If we consider only predicted (not known) regulators, Crinet-selected regulators were still higher than the Cupid-predicted ones. In Figure 2, we also showed nine regulators eliminated by Crinet after expression correlation and sufficient abundance filtering of miRNA-target selection.These Crinet-eliminated interactions did not show strong anticorrelation among miRNA and protein expression with respect to fold-change for the majority of miRNAs, showing the strength of our filtering approaches. As a negative control, we added the average of 100 random miRNA regulators which were not selected as ESR1's regulator by Cupid and

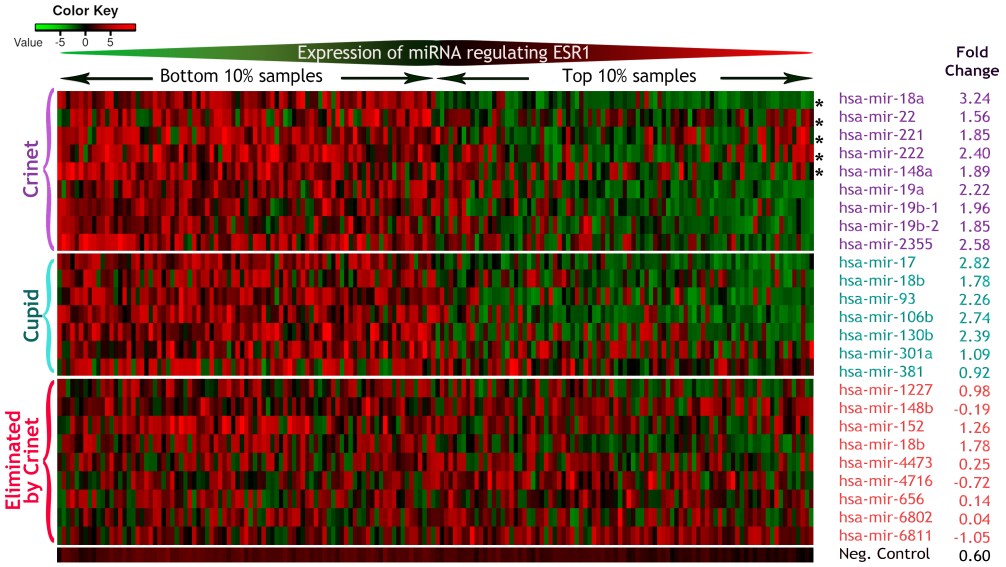

Figure 2. Heatmap showing protein expression of ESR1 for top and bottom 10% ranked samples with respect to miRNA expression. Protein expression is shown for top and bottom 10% samples ranked with respect to miRNA expression regulating ESR1 by Cupid-selected, Crinet-selected, Crinet-eliminated, and negative control along with the mean difference of log fold-change of protein expression for bottom 10% with respect to top 10% samples. Each row is independently ranked by miRNA expression. (* known miRNA regulator selected by Cupid, ** predicted miRNA regulator selected by Cupid overlapping with Crinet).

Crinet, and they did not show any preference through all samples.

Protein expression anticorrelation analysis: Using protein expression dataset (TCPA), we analyzed the anticorrelation among expression of target's protein and miRNA regulators for each applicable miRNA-target interaction and also for each target with average miRNA regulators. In the protein expression dataset, there is more than one protein for some targets, while there is no protein for some others. In the starting miRNA-target interaction set, we had 201 protein-gene pairs between unique 154 genes and 175 proteins to evaluate while we had 161 protein-gene pairs between unique 124 genes and 143 proteins to evaluate in final miRNA-target interactions.

While the anticorrelation preference slightly decreases after selecting top 40% miRNA-target interactions, our filtering steps substantially increased the ratios showing the anticorrelation preference in our selected interactions 4. Also considering average miRNAs per target, we had slightly anticorrelation preference for top 40% interactions; however, our filtering steps did a better job increasing the anticorrelation preference more. As the negative control, non-inferred interactions showed the random preference for anticorrelation (See Supplementary Table 3).

3.2 Inferred ceRNA interactions were able to predict gene expression change

To assess our accuracy for ceRNA inference, we used the Library of Integrated Network-based Cellular Signature (LINCS) [24] L1000 shRNA-mediated gene knockdown experiment in breast cancer cell line and checked whether ceRNA interactions can predict the effects of RNAi-mediated gene silencing perturbations in MCF7 cells. In [5], the authors also used LINCS shRNA transfection data. Since Crinet

Table 4. Analysis with miRNA-protein expression anticorrelation. Interaction phase shows the anticorrelated expressions of miRNA-protein target pairs with respect to not anticorrelated ones. Gene phase shows the anticorrelated expression of target with average of all miRNA regulators with respect to not-anticorrelated ones. (ratio is expected to be more than 1 to have anticorrelation preference. Higher is better). ALL: All obtained miRNA-target interactions; ALL40: Top 40% of ALL interactions; CORR: After expression correlation filtering on ALL40 interactions; CORR+ABUN: After abundance sufficiency filtering on CORR interactions.

Interaction Step	Interaction # in Set	Interaction Phase	Gene Phase
ALL	8,193,904	38,505/33,243=1.2	95/106=0.9
ALL40	3,285,523	2,720/2,688=1.0	107/91=1.2
CORR	535,284	751/370=2.0	115/50=2.3
CORR+ABUN	165,937	719/352=2.0	114/47=2.4

uses a high number of genes, it was not computationally feasible to run many tools with our dataset. However, Hermes [5] runs any given ceRNA pair independently, therefore we ran Hermes just for the genes in knockdown assays using the same expression datasets and Crinet-selected miRNA-target interactions, and compared with Crinet using LINCS database. The database is a rich resource having an expression change of almost 1000 genes as a response to a silenced gene. When a gene is silenced then its ceRNA partner will be affected since more miRNA regulators will be available to suppress the ceRNA partner. Thus, given a ceRNA pair, expression level should be lower in response to the silenced ceRNA partners in comparison to the genes that are not ceRNA partners. Therefore, we calculated the ratio of gene expression fold-change following the knockdown of its ceRNA partners to following the genes that are not its ceRNA partners for each perturbagen ceRNA. A perturbagen ceRNA's fold-change ratio was expected to be smaller than 1, and lower ratios indicate a better prediction of gene expression change for ceRNA interactions. To quantify the ratios, we accepted the accuracy of a ceRNA network as the percentage of ceRNAs whose ratios were smaller than 1 with respect to all ceRNAs.

As gene expression in the MCF7 dataset was measured in two different time points (96h and 144h), we analyzed each time point separately and combined them as the overall percentage of accuracy. Hermes accepts two different filtering parameters: the significance of the size of the common miRNA program and the significance of conditional regulation. We executed first run of Hermes with 10^{-2} for common miRNA significance and 10^{-3} for conditional regulation significance, for second run 10^{-3} and 10^{-4}, for third run 10^{-4} and 10^{-5}, for fourth run 10^{-5} and 10^{-6}, and for fifth run 10^{-6} and 10^{-7} respectively, till having the similar number of knockdown assessment with Crinet results. The accuracy of this assessment is shown in Table 5.

Based on these results, Crinet outperformed Hermes for each timepoint and overall accuracy. Even we eliminated input interaction set based on our filtering, Hermes behaved not much selective in terms of the number of ceRNA interactions by inferring many interactions as significant. Considering binding scores, including collective regulation, and excluding the amplifying effect of ceRNAs in terms of regulation in Crinet seem powerful to predict gene expression change in ceRNA partners in the LINCS database.

Table 5. Analysis to check the accuracy of inferred ceRNA interactions using LINCS-L1000 shRNA-mediated gene knockdown experiment in breast cancer cell line. The values show the prediction accuracy of inferred network for Crinet and different Hermes runs along with the number of downstream ceRNAs whose ratios are smaller than 1 over all with the percentages.

	96h Timepoint	144h Timepoint	Overall
Crinet	48/77=62%	44/77=57%	92/154=60%
Hermes 1.run	412/897=46%	441/913=48%	853/1810=47%
Hermes 2.run	199/466=43%	217/481=45%	416/947=44%
Hermes 3.run	109/269=41%	121/273=44%	230/542=42%
Hermes 4.run	65/143=45%	68/144=47%	133/287=46%
Hermes 5.run	33/69=48%	39/71=55%	72/140=51%

3.3 Inferred ceRNAs were significantly associated with known cancer genes and significantly enriched with cancer-related processes

To analyze the biological significance of the inferred ceRNA network, we applied enrichment analysis for the inferred ceRNAs. We used ClusterProfiler R package [30] for all enrichment analysis The inferred ceRNAs were significantly enriched in 398 GO terms from biological process ontology and 39 KEGG pathways. When we did the same analysis for non-inferred genes, they were significantly enriched just for 40 GO terms and 3 KEGG pathways. To associate enriched terms to broader categories, we analyzed GO Slim terms (Supplementary Table 8).

To check that inferred ceRNAs were associated with cancer genes, we collected 3078 known cancer genes obtained from Cancer Gene Census in COSMIC v91 [23], Bushman's cancer gene list v3 [14], human oncogenes from ONGene [16], Network of Cancer Genes 6.0 [20], and LncRNADisease database from the Cui Lab [3]. In our inferred ceRNAs, we had significant overlap (hypergeometric test p-value 9.10^{-105}) between known cancer genes and inferred ceRNAs having 789 out of 3078 known cancer genes in our inferred network. When we repeated the same analysis for non-inferred genes, they did not show significant p-value (almost 1). We also had 54 breast cancer-related genes from available datasets (LncRNADisease database and Network of Cancer Genes 6.0 had disease information), and we had 14 out of 54 breast cancer-related genes as inferred in our network. These results indicated that inferred ceRNAs were significantly associated with known cancer genes.

We compared our inferred ceRNA pairs with lncACTdb v2.0 [25] database consisting of functionally activated lncRNA-miRNA-gene triples matching 3918 lncRNA-gene pairs with our datasets. Even we did not have any matching ceRNA pairs, we had 3 ceRNA interactions at the database before we applied network deconvolution to our network. These three interactions were between lncRNA NR2F1-AS1 and three genes namely: CORO2B, ZEB1, and RECK. LncRNA NR2F1-AS1 had 43 targets before network deconvolution filtering, but remained with just three of them, most probably due to its low gene expression (expression distribution through all samples had a median of 0.5 even with a maximum of 6 as FPKM). If we didn't filter out indirect interactions, we will have these interactions as inferred, however, these interactions could be inferred because of the high overlapping ceRNA partners having a high indirect effect.

Moreover, we analyzed the top genes with a high degree in our network. The degree distribution of our network had a median of 3 with a maximum of 152 and a minimum of 1. We got the top 81 ceRNAs having at least 50 degrees in our network as high-degreed genes. 45 out of 81 were applicable for the enrichment and these genes were significantly enriched in 58 GO terms and 8 KEGG pathways. Also, these high-degree genes were

among known cancer genes with a hypergeometric p-value of 0.0009. 15 out of 81 genes were involved in known cancer genes, while 3 of them were breast cancer-related.

Furthermore, we obtained 19 known and important lncRNA oncogene and suppressors in breast cancer obtained from [32]. We compared these lncRNAs with our inferred network, and six of them were inferred in our ceRNA network (hypergeometric test p-value = 5.10^{-06} among known lncRNAs and inferred lncRNAs). These genes were namely MAGI2-AS3, MALAT1, MIR100HG, ITGB2-AS1, MEG3, and XIST having very high degrees for some in our inferred network (89, 51, 26, 20, 17, 1 degree, respectively).

MAGI2-AS3 had 89 degrees being highly active for ceRNA regulation in our network, and it was known as suppressor involving in cell growth [29]. We had 51 degrees for MALAT1, which contributes significantly to cancer initiation and progression in breast cancer [12]. Some other lncRNAs had also important functionality: MIR100HG as an oncogene involving in proliferation [26], ITGB2-AS1 as an oncogene involving migration and invasion [15], and MEG3 as a suppressor involving in proliferation and EMT [33]. These suggest that we had known lncRNA oncogenes and suppressors as actively involved in the inferred ceRNA network showing the lncRNAs in inferred ceRNA network could have an important role in cancer-related processes, and the inclusion of lncRNAs to the ceRNA inference pipeline was important to find out them.

3.4 Inferred ceRNA groups included known cancer-related genes and were enriched in cancer-related processes

Crinet inferred 81 ceRNA groups from the ceRNA network after 1508 Walktrap runs. Seventy-seven of these groups had three of fewer elements. We further analyzed the remaining four groups (i.e., groups 1, 2, 3, and 4 in Supplementary Figure 4). To evaluate the inferred ceRNA groups, we performed enrichment of ceRNA groups in cancer-related genes and biological processes.

CeRNAs in group 1, 3 and 4 had a significant overlap with cancer-related genes (hypergeometric p-value < 0.0006). Group 1 genes were enriched with 82 GO terms from biological process ontology and 5 KEGG pathways, while Group 3 were enriched with 35 GO terms and Group 4 in 17 GO terms. Moreover, Group 1 was highly enriched with the processes related to ancestor categories including “cell cycle”, “response to stress”, “DNA metabolic process”, “chromosome segregation”, and “cell division”. Even group 2 was limited in terms of genes applicable for enrichment, all the remaining groups (groups 2, 3, and 4) were consistently enriched with ancestor category “immune system process”. Also, group 3 had some ancestor category including “cell death”, “cell adhesion”, and “cell motility”, while group 4 had “response to stress” category additionally (Supplementary Table 7 for details).

CeRNA groups were significantly overlapped with known cancer-related genes and significantly enriched in biological processes suggesting that ceRNA groups could have important roles in cancer including the immune system and cell repair.

4 Discussion

In this study, we developed a tool named Crinet to infer pairwise and groupwise ceRNA interactions and applied it to breast tumor samples from TCGA. Leveraging multiple types of biological datasets and considering expression abundance among miRNA and their targets, we inferred a ceRNA network including 52,858 ceRNA interactions between unique 7263 ceRNAs. To exclude the amplifying effect of ceRNA regulation, we employed a network deconvolution algorithm [] on the network and excluded indirect interactions, remaining with 17,443 ceRNA interactions among 4494 unique ceRNAs.

Since we built a ceRNA network relying on the miRNA-target interaction set, computing the miRNA-target interactions properly is important for ceRNA inference. Unlike the existing tools, we filtered the miRNA-target interactions considering abundance sufficiency and binding scores so that a miRNA will not mediate thousands of targets with a very limited expression level. We introduced a *Interaction Regulation* (IR) score for each miRNA-target interaction, and confirmed that the top 25% miRNAs with the highest number of targets were eliminated following IR-score filtering, while the new top 25% miRNAs with the highest number of targets had higher expression level (Supplementary Methods A.1 and Supplementary Figure 1). Since ceRNAs positively regulate each other, expecting positive correlation among ceRNAs is a common approach [21]. While some studies use additional metrics in addition to simply calculating positive correlation [8, 17], we measured the partial correlation excluding the CNA since expression values are highly affected by copy number amplification and deletion. Different from the studies starting with differentially expressed genes among tumor and normal samples, we included all long noncoding RNAs (lncRNAs) and pseudogenes in addition to all coding RNAs (mRNAs) considering the significant role of all these genes in ceRNA inference. Since we considered binding probabilities in our pipeline, we started also with a large set of miRNA-target interactions. These cost more computational power while generating ceRNA interactions based on common miRNA regulators, however, it was important to include all possible genes especially for the measurements like our collective regulation and *effective regulation* to split expression among targets and regulators.

Crinet employs a network deconvolution algorithm to exclude indirect ceRNA interactions from the network. CeRNAs positively regulate each other competing for common miRNA regulators, therefore, ceRNA partners have an indirect effect on each other. If two genes have no direct ceRNA interaction but have many common ceRNA partners then they will have amplified effect of correlation, therefore these indirect interactions might be inferred in the ceRNA network. Thus, excluding these indirect interactions is important, and network deconvolution could be a good solution to this problem. In [9], authors find a systematic method for inferring the direct dependencies in a network removing the effects of transitive interactions by indirect effects. As a novel part, we used a network deconvolution method and eliminated ceRNA interactions potentially inferred by the amplifying effect of ceRNA pairs. Since [9] does not remove indirect edges but gives a ranking score, we selected some threshold to apply on our ceRNA network, showing the selection of this threshold is not changing the accuracy of inferred ceRNA network much (Supplementary Table 5).

Based on the premise that more than two genes can work together to compete for common miRNAs, Crinet inferred groupwise ceRNA interactions using a community detection algorithm. We defined our ceRNA groups as a group of ceRNAs working together to sequester some common miRNA regulators. To hold the inference consistency in the ceRNA network, we adapted our groups into the network checking all the inference conditions between groups and neighbor genes. In that way, we did not find only a group of closely related genes in our network, but we had a ceRNA network in which the groups and individual ceRNAs had ceRNA regulation showing the comprehensive regulation in the network.

When the expression profiling of other data types such as circular RNAs is available from TCGA, including them into the pipeline could be worthy of further investigation to improve ceRNA inference [11], especially for comprehensively measuring metrics like our collective regulation.

In conclusion, we developed Crinet to infer cancer-associated ceRNA networks addressing critical drawbacks in existing computational tools. Getting the benefit of multiple biological datasets, considering expression abundance among miRNA and their

targets, differentiating direct and indirect ceRNA interactions, and inclusion of noncoding RNAs (namely lncRNAs and pseudogenes), our tool comprehensively investigates ceRNA regulation. A better understanding of this new regulatory layer with novel insights would shed light on the underlying complex regulatory circuitry in disease conditions like cancer.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM133657. The datasets used in the analysis were generated by the TCGA Research Network: `https://www.cancer.gov/tcga`.

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. R. Akbani, P. K. S. Ng, H. M. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.-Y. Yang, K. Yoshihara, J. Li, et al. A pan-cancer proteomic perspective on the cancer genome atlas. *Nature communications*, 5(1):1–15, 2014.
3. G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui. Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic acids research*, 41(D1):D983–D986, 2012.
4. H.-S. Chiu, D. Llobet-Navas, X. Yang, W.-J. Chung, A. Ambesi-Impiombato, A. Iyer, H. R. Kim, E. G. Seviour, Z. Luo, V. Sehgal, et al. Cupid: simultaneous reconstruction of microrna-target and cerna networks. *Genome research*, 25(2):257–267, 2015.
5. H.-S. Chiu, M. R. Martínez, M. Bansal, A. Subramanian, T. R. Golub, X. Yang, P. Sumazin, and A. Califano. High-throughput validation of cerna regulatory networks. *BMC genomics*, 18(1):418, 2017.
6. A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.
7. E. P. Consortium et al. A user's guide to the encyclopedia of dna elements (encode). *PLoS biology*, 9(4), 2011.
8. D. Do and S. Bozdag. Cancerin: A computational pipeline to infer cancer-associated cerna interaction networks. *PLoS computational biology*, 14(7):e1006318, 2018.
9. S. Feizi, D. Marbach, M. Médard, and M. Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726, 2013.
10. H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

11. M. Huang, Z. Zhong, M. Lv, J. Shu, Q. Tian, and J. Chen. Comprehensive analysis of differentially expressed profiles of lncrnas and circrnas with associated co-expression and cerna networks in bladder carcinoma. *Oncotarget*, 7(30):47186, 2016.

12. J. Kim, H.-L. Piao, B.-J. Kim, F. Yao, Z. Han, Y. Wang, Z. Xiao, A. N. Siverly, S. E. Lawhon, B. N. Ton, et al. Long noncoding rna malat1 suppresses breast cancer metastasis. *Nature genetics*, 50(12):1705–1715, 2018.

13. M. S. Kumar, E. Armenteros-Monterroso, P. East, P. Chakravorty, N. Matthews, M. M. Winslow, and J. Downward. Hmga2 functions as a competing endogenous rna to promote lung cancer progression. *Nature*, 505(7482):212–217, 2014.

14. B. Lab. Bushman lab: Cancer gene list (version 4). Available from: `http://www.bushmanlab.org/links/genelists`, 05 2018. Accessed: 2020-06-06.

15. M. Liu, L. Gou, J. Xia, Q. Wan, Y. Jiang, S. Sun, M. Tang, T. He, and Y. Zhang. Lncrna itgb2-as1 could promote the migration and invasion of breast cancer cells through up-regulating itgb2. *International journal of molecular sciences*, 19(7):1866, 2018.

16. Y. Liu, J. Sun, and M. Zhao. Ongene: A literature-based database for human oncogenes. *J Genet Genomics*, 44(2):119–121, 2017.

17. 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.

18. P. Pons and M. Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

19. X. Qi, D.-H. Zhang, N. Wu, J.-H. Xiao, X. Wang, and W. Ma. cerna in cancer: possible functions and clinical implications. *Journal of Medical Genetics*, 52(10):710–718, 2015.

20. D. Repana, J. Nulsen, L. Dressler, M. Bortolomeazzi, S. Kuppili Venkata, A. Tourna, A. Yakovleva, T. Palmieri, and F. Ciccarelli. The network of cancer genes (ncg): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biology*, 20, 01 2019.

21. L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi. A cerna hypothesis: the rosetta stone of a hidden rna language? *Cell*, 146(3):353–358, 2011.

22. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.

23. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, and S. A. Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018.

24. t. text and t. text. text. *text*, 1(1):129, 2016.

25. P. Wang, X. Li, Y. Gao, Q. Guo, Y. Wang, Y. Fang, X. Ma, H. Zhi, D. Zhou, W. Shen, et al. Lncactdb 2.0: an updated database of experimentally supported cerna interactions curated from low-and high-throughput experiments. *Nucleic acids research*, 47(D1):D121–D127, 2019.

26. S. Wang, H. Ke, H. Zhang, Y. Ma, L. Ao, L. Zou, Q. Yang, H. Zhu, J. Nie, C. Wu, et al. Lncrna mir100hg promotes cell proliferation in triple-negative breast cancer through triplex formation with p27 loci. *Cell death & disease*, 9(8):1–11, 2018.

27. X. Wen, L. Gao, and Y. Hu. Lacemodule: Identification of competing endogenous rna modules by integrating dynamic correlation. *Frontiers in genetics*, 11:235, 2020.

28. J. Yang, T. Li, C. Gao, X. Lv, K. Liu, H. Song, Y. Xing, and T. Xi. Foxo1 3 utr functions as a cerna in repressing the metastases of breast cancer cells via regulating mirna activity. *FEBS letters*, 588(17):3218–3224, 2014.

29. Y. Yang, H. Yang, M. Xu, H. Zhang, M. Sun, P. Mu, T. Dong, S. Du, and K. Liu. Long non-coding rna (lncrna) magi2-as3 inhibits breast cancer cell growth by targeting the fas/fasl signalling pathway. *Human cell*, 31(3):232–241, 2018.

30. 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.

31. J. Zhang. Cntools: Convert segment data into a region by sample matrix to allow for other high level computational analyses. *R package (Version 1.6. 0.)*, 2016.

32. T. Zhang, H. Hu, G. Yan, T. Wu, S. Liu, W. Chen, Y. Ning, and Z. Lu. Long non-coding rna and breast cancer. *Technology in cancer research & treatment*, 18:1533033819843889, 2019.

33. W. Zhang, S. Shi, J. Jiang, X. Li, H. Lu, and F. Ren. Lncrna meg3 inhibits cell epithelial-mesenchymal transition by sponging mir-421 targeting e-cadherin in breast cancer. *Biomedicine & Pharmacotherapy*, 91:312–319, 2017.

34. 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.