 doi:10.20944/preprints202109.0337.v1

Identifying complex lncRNA/pseudogene-miRNA-mRNA crosstalk in hormone-dependent cancers

Dulari K. Jayarathna[1,2], Miguel E. Rentería[2,3], Emilie Sauret[4], Jyotsna Batra[1,3,5], and Neha S. Gandhi[1,5*]

[1]Centre for Genomics and Personalised Health, School of Chemistry and Physics, Queensland University of Technology, Brisbane, QLD 4000, Australia
[2]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia
[3]School of Biomedical Sciences, Queensland University of Technology, Brisbane, QLD 4059, Australia
[4]School of Mechanical, Medical & Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia
[5]Translational Research Institute, Brisbane, QLD 4102, Australia

***Correspondence:** Neha S. Gandhi (neha.gandhi@qut.edu.au)

Abstract

The discovery of microRNAs (miRNAs) has fundamentally transformed our understanding of gene regulation. The competing endogenous RNA (ceRNA) hypothesis postulates that not only messenger RNAs but also other RNA transcripts, such as long non-coding RNAs and pseudogenes, can act as natural miRNA sponges. These RNAs influence each other's expression levels by competing for the same pool of miRNAs through miRNA response elements on their target transcripts, thereby modulating gene expression and protein activity. In recent years, these ceRNA regulatory networks have gained considerable attention in cancer research. Several studies have identified cancer-specific ceRNA networks. Nevertheless, prior bioinformatic analyses have focused on long non-coding RNA-associated ceRNA networks. Here, we identify an extended-ceRNA network (including both long non-coding RNAs and pseudogenes) shared across a group of four hormone-dependent (HD) cancers, i.e., prostate, breast, colorectal, and endometrial cancers, using data from The Cancer Genome Atlas (TCGA). We performed a functional enrichment analysis for differentially expressed genes in the shared ceRNA network of HD cancers, followed by a survival analysis to determine their prognostic ability. We identified two long non-coding RNAs, nine genes, and seventy-four miRNAs in the shared ceRNA network across four HD cancers. Among them, two genes and forty-one miRNAs were associated with at least one HD cancer survival. This study is the first to investigate pseudogene associated ceRNAs across a group of related cancers and highlights the value of this approach to understanding shared molecular pathogenesis in a group of related diseases.

Keywords: hormone-dependent cancers, ceRNAs, lncRNAs, microRNAs, pseudogenes, multiple sensitivity correlation

1. Introduction

MicroRNAs (miRNAs) are endogenous non-coding RNAs consisting of 19-25 nucleotides in length. They regulate gene expression through degradation or inhibition of translation by binding to messenger RNA (mRNA) [1]. A single miRNA can target hundreds of genes. Therefore, miRNAs play a crucial post-transcriptional role in DNA-RNA-protein networks. In each cell, transcripts such as mRNAs, long non-coding RNAs (lncRNAs), and pseudogenes contain similar miRNA response elements (MREs) that can crosstalk via competition of binding to common miRNAs serving as miRNA sponges. In 2011, this phenomenon was described as "the competing endogenous RNA (ceRNA) hypothesis" [2]. As a major ceRNA component, lncRNA has a dual role in the nucleus and cytoplasm. Several studies suggest that lncRNAs directly interact with transcription factors as transcriptional co-activators in the nucleus, while others suggest that lncRNAs may

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

impair transcriptional complexes' assembly as the inhibitor of gene expression [3]. The pseudogenes are very similar to the coding genes, as they are produced by modifying and cutting off the coding transcripts in the transcription process. Both pseudogenes and cytoplasmic lncRNAs (not in the nucleus) act as regulators to affect their target genes [4]. These ceRNAs, lncRNAs and pseudogenes may influence cancer pathogenesis by regulating mRNA expression of crucial tumorigenic or tumour-suppressive genes and pathways [5].

Previous bioinformatics studies have identified ceRNA candidates as prognostic or predictive biomarkers for common cancer types such as colorectal, endometrial, prostate, and breast cancers [6-10]. Several web-based tools such as miRTissue$_{ce}$, LncACTdb 2.0, and lnCeDB have been developed, supporting the search for ceRNA interactions networks in multiple tissues [11-13]. A recent colorectal cancer ceRNA study identified a network of nine hub genes, thirteen lncRNAs, and twenty-nine candidate miRNAs integrating multiple genomic datasets [6]. The authors further revealed the *MFAP5*-miR-200b-3p-AC005154.6 axis as a potential biomarker of colorectal cancer. In 2019, bioinformatic analyses conducted by Wang, *et al.* [7] and Ouyang, *et al.* [8] revealed two endometrial cancer-associated ceRNAs, lncRNA LINC00958 (*DOLPP1*-miR-761-LINC00958) and lncRNA LINC00261 (*C2orf48*-LINC00261), respectively. Recent experimental studies have validated that these two lncRNAs act as critical regulators of endometrial cancer binding through multiple mRNA-miRNA axes [14,15]. A ceRNA network analysis of prostate cancer established a network consisting of four hub genes, homeobox B5 (*HOXB5*), glypican 2 (*GPC2*), pepsinogen A-5 (*PGA5*), and ameloblastin (*AMBN*) that are strongly associated with patient survival [9]. A comprehensive lncRNA-associated ceRNA analysis of breast cancer identified ninety-three lncRNAs, twenty-seven mRNAs, and nineteen miRNAs. In this dataset, fifteen lncRNAs were identified as prognostic biomarkers of breast cancer [10]. Studies described above suggest that existing ceRNA network analyses can be successfully applied for distinct cancer types to understand their biological mechanisms further. Identifying common ceRNAs networks across genetically related diseases such as hormone-dependent (HD) cancers will also significantly contribute to understanding the shared molecular pathogenesis.

This study identifies a shared ceRNA network across HD cancers, including prostate, breast, colorectal, and endometrial, which are among the world's highest cancer mortality and incident rates.

2. Materials and Methods

2.1. Patients and Samples

RNA expression data (RNA-seq and miRNA-seq) and clinical data for five HD cancers, prostate (PRAD), breast (BRCA), colon (COAD), rectal (READ), and endometrial (UCEC), were obtained from The Cancer Genome Atlas (TCGA). The PRAD, BRCA, COAD, READ, and UCEC consist of 499/52 (cases/controls), 1109/113, 480/41, 167/10 and 552/35, respectively. The HTSeq-counts RNA-seq data and isoform quantification data of miRNA-seq for the given five cancer types were downloaded to a local computing server from the GDC (genomics data commons) data portal [16].

2.2. Differential Expression Analysis of Hormone-Dependent Cancers Data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

At the data pre-processing stage, we removed TCGA samples with duplicated sample IDs. Then metastatic samples were eliminated as we compared primary tumour and adjacent normal samples using the differential expression analysis. The raw counts expression data were normalised by the TMM (trimmed mean of M values) method implemented in the edgeR R package [17]. The normalised data were transformed into a standard scale using the voom method implemented in the limma (linear modelling for microarrays) R package [18]. Low-expressed genes (log counts per million<1 in more than 50% of the samples) were removed by default. Ignoring low-expressed genes increases the total count of differentially expressed genes after multiple testing correction and improves sensitivity and precision. Genes or miRNAs that were differentially expressed between tumour and normal tissue were identified by applying "lmFit" followed by "eBayes (empirical Bayes)", in-built functions in the limma R package [18]. We fitted a linear model for each gene using the "lmFit" function. Then eBayes moderation was applied, borrowing information across all the genes to obtain more precise estimates of gene-wise variability. Expression differences were assessed by linear modelling results, log fold-change (logFC) and false discovery rate (FDR) adjusted p-values. |logFC| >1 and FDR < 0.01 were considered thresholds to identify statistically significant mRNAs, lncRNAs, pseudogenes and miRNAs. Differentially expressed lncRNAs, pseudogenes, and mRNAs were separately recorded for ceRNA network analysis. Figure 1 depicts the workflow of this study.

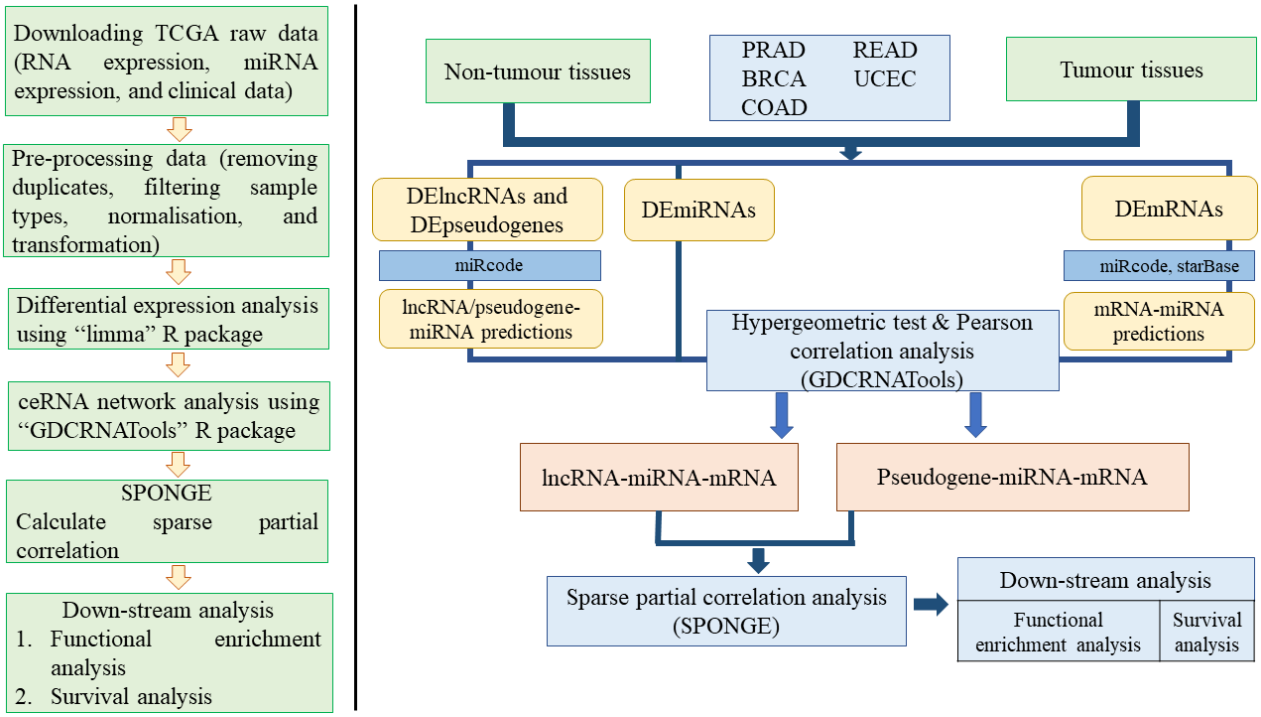

Figure 1. Methodological workflow of the study. RNA-seq and miRNA-seq data extracted from The Cancer Genome Atlas (TCGA) were pre-processed, normalised, and transformed into a standard scale. Differential expression analysis was conducted to identify differentially expressed long non-coding RNAs (lncRNAs), pseudogenes, messenger RNAs (mRNAs), and microRNAs (miRNAs). The competing endogenous RNA (ceRNA) network analysis followed three steps, i. identifying lncRNA/pseudogene-mRNA sharing the significant number of miRNAs, ii. calculating Pearson correlation between lncRNA/pseudogene and mRNAs, and iii. calculating multiple sensitivity correlation considering a set of miRNAs targeted by given lncRNA/pseudogene-mRNA pair. The first two steps were conducted using the GDCRNATools R/Bioconductor package [19]. Step iii, sparse partial correlation analysis was executed using the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

SPONGE (Sparse Partial correlation ON Gene Expression) R/Bioconductor package [20]. The three-step analysis filtered out statistically significant ceRNAs for individual HD cancers. Then only shared ceRNA components across all four HD cancers were involved in the downstream analysis.

2.3. Competing Endogenous RNA Network Analysis

Initially, we constructed both lncRNA-based and pseudogene-based ceRNA networks for individual HD cancers. Then we identified shared ceRNA associations across four HD cancers. Two downstream analyses, functional enrichment and survival analyses, were conducted for shared genes, lncRNAs, pseudogenes, and miRNAs across HD cancers ceRNA networks.

2.4. Long non-coding RNA/pseudogene-microRNA-mRNA Networks

We followed three steps to identify ceRNA interactions, i. detecting lncRNA/pseudogene-mRNA pairs that share a significant number of miRNAs, ii. selecting positively correlated lncRNA/pseudogene-mRNA pairs, and iii. jointly estimating the significance of multiple miRNAs in lncRNA/pseudogene-mRNA pairs. The miRNA-mRNA, miRNA-lncRNA, and miRNA-pseudogene interactions required for steps i and iii were obtained from two databases, miRcode, and starBase [21,22]. The miRcode database facilitates mRNA-miRNA, lncRNA-miRNA and pseudogene-miRNA target predictions using a broad searchable map that contains 10,419 lncRNAs and 12,549 pseudogenes. The starBase includes miRNA-mRNA interactions predicted by probing 108 CLIP-seq datasets. As described above, a similar three-step approach has been previously followed by the miRTissue$_{ce}$, a ceRNA-ceRNA web application tool [11]. In the first step, we used the hypergeometric test to identify lncRNA/pseudogene-mRNA pairs with a significant number of shared miRNAs. The hypergeometric test associated p-value can be computed using the following equation, equation 1:

$$p = 1 - \sum_{k=0}^{m} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (1)$$

Where m is the number of shared miRNAs, N is the total number of available miRNAs, n is the number of miRNAs targeting the lncRNA/pseudogene, K is the number of miRNAs targeting the mRNA. MiRNAs are known as negative regulators of gene expression. If a lncRNA/pseudogene occupies the majority of miRNAs, only a small proportion is available to bind to the target mRNA, increasing the mRNA's expression level. Based on this phenomenon, the lncRNA/pseudogene-mRNA pair should be positively correlated. As the second step, we applied the Pearson correlation analysis to extract positively correlated lncRNA/pseudogene-mRNA pairs from all possible lncRNA/pseudogene-mRNA interactions. Both hypergeometric test and Pearson correlation analysis were carried out using the GDCRNATools R/Bioconductor package [19]. In GDCRNATools, the regulation contribution towards a ceRNA interaction has been quantified using the sensitivity correlation (*scor*) [23]. The *scor* value does not account for a combinatorial effect of multiple miRNAs. Subsequently, strong ceRNAs mediated by multiple moderate miRNA regulators cannot be detected. Therefore, we utilised an extension of *scor*, the multiple sensitivity correlation (*mscor*) method, which has been implemented in SPONGE (Sparse Partial correlation ON Gene Expression) R/Bioconductor package [20]. The derived formula to calculate *mscor* is given in equation 2:

$$\mathrm{mscor}(g_1, g_2, M) = \mathrm{cor}(g_1, g_2) - \mathrm{pcor}(g_1, g_2 | M) \quad (2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

Where $M=m_1,m_2\ldots,m_i$ and i is the number of shared miRNAs between g_1 and g_2 genes. The *cor*() term defines the Pearson correlation between g_1 and g_2 genes expression profiles, and *pcor*() is the partial correlation that estimates how two variables are correlated when they are controlled by additional variables. Furthermore, the SPONGE method defines a null distribution which allows estimating an empirical p-value for *mscor*.

We filtered ceRNA interactions returned by three user-defined significant thresholds, i. in hypergeometric test, FDR adjusted p-value < 0.05, ii. Pearson correlation coefficient between ceRNA pairs > 0.4, and iii. the adjusted p-value of *mscor* in the SPONGE method < 0.05. This study defined colorectal cancer (COLCA) ceRNA network by integrating unique ceRNA associations of colon and rectal cancers ceRNA networks. The resulted lncRNA-mRNA-miRNA and pseudogene-mRNA-miRNA combinations in each HD cancer-specific ceRNA network were integrated into a single variable. The format of the derived categorical variable is "<lncRNA/pseudogene gene ensemble ID>_<gene ensemble ID>_<miRNA name>". After that, we constructed the one-way table to identify shared lncRNA/pseudogene-mRNA-miRNA associations across all four HD cancer types. Suppose one-way frequency equals 4 for a given lncRNA/pseudogene-miRNA-mRNA pair. In that case, a ceRNA association is classified as "the shared ceRNA network of HD cancers". We used the Cytoscape software to visualize the shared ceRNA network of HD cancers [24]. Two downstream analyses were conducted for genes and miRNAs included in the shared ceRNA network of HD cancers.

2.5. Functional Enrichment Analysis

The Functional enrichment analysis was performed for genes in the shared ceRNA network of HD cancers. The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses were conducted using the R/Bioconductor clusterProfiler R package [25].

2.6. Survival Analysis

We performed survival analysis using the Kaplan-Meier (K-M) survival curves, implemented in the survival R package [26] to explore the role of genes in the shared ceRNA network. For each gene/miRNA, the tumour samples were divided into two groups (low-expressed and high-expressed) according to the median gene/miRNA expression value. The log-rank test (Mantel-Haenszel test) was used as the statistical method for the Kaplan-Meier curves. The log-rank test statistic has a chi-square (χ^2) distribution with one degree of freedom. Therefore, significant genes and miRNAs were chosen under the χ^2 test statistic p-value < 0.05. We used miRCancer [27], a literature-curated database for miRNAs experimental studies in cancers, to check descriptions of prognostic miRNAs.

3. Results

After the quality-control process, we retrieved 495/52 (cases/controls), 1091/113, 456/41, 166/10, and 543/35 samples from PRAD, BRCA, COAD, READ, and UCEC. In the differential expression analysis, we used 15509, 15244, 14771, 14866, and 15197 genes from PRAD, BRCA, COAD, READ, and UCEC after removing that low expressed.

3.1. Differential Expression Analysis Results

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

The differential expression analysis between tumour and healthy samples was conducted using the limma R package. The count of differentially expressed (up/down) lncRNAs, pseudogenes, mRNAs and miRNAs are given in Table 1.

Table 1. Count of differentially expressed (up/down) lncRNAs, pseudogenes, mRNAs, and miRNAs in each hormone-dependent (HD) cancer (corrected for multiple testing).

Cancer	lncRNA		Pseudogene		mRNA		miRNA	
	Up	Down	Up	Down	Up	Down	Up	Down
BRCA	61	106	17	28	1125	1642	71	87
COAD	137	72	44	31	1200	1778	186	153
PRAD	139	49	28	18	434	1079	34	27
READ	181	53	52	18	1169	1790	165	114
UCEC	116	137	43	43	1584	2000	142	103

3.2. Shared Competing Endogenous RNA Networks across Hormone-Dependent Cancers

First, we identified significant lncRNA-mRNA-miRNA and pseudogene-mRNA-miRNA networks for each HD cancer. The number of lncRNAs/pseudogenes, mRNAs, and miRNAs in ceRNA networks of individual HD cancers are reported in Table S1. We used both GDCRNATools and SPONGE R packages in ceRNA network analysis. Table 2 contains all shared ceRNA associations found from the GDCRNATools approach (steps i and ii). In Table 2, common ceRNAs from both GDCRNATools and SPONGE partial correlation analysis (step iii) are labelled by an asterisk (*).

Table 2. Shared lncRNA and pseudogene-associated ceRNA associations among hormone-dependent (HD) cancers.

lncRNA/pseudogene	mRNA (number of shared miRNAs among lncRNA and mRNA)	
MBNL1-AS1 (lncRNA)	DnaJ Heat Shock Protein Family (Hsp40) Member B4 (*DNAJB4*) (6)	
MAGI2-AS3* (lncRNA)	*DNAJB4**(12)	Cofilin2* (*CFL2**) (36)
	Fibroblast Growth Factor 2* (*FGF2**) (8)	Phospholipid Scramblase 4* (*PLSCR4**) (19)
	Myosin Light Chain Kinase* (*MYLK**) (15)	Endothelin Receptor Type B (*EDNRB*) (12)
	Junctophilin 2 (*JPH2*) (6)	Tensin 1* (*TNS1**) (18)
MIR100HG* (lncRNA)	FERM Domain Containing Kindlin 2* (*FERMT2**) (21)	*FGF2* (9)
	DIX Domain Containing 1* (*DIXDC1**) (17)	Sushi Repeat Containing Protein X-Linked* (*SRPX**) (7)
	R-Spondin 3 (*RSPO3*) (8)	*JPH2* (6)
	DNAJB4 (13)	
MEIS3P1 (pseudogene)	*TNS1* (16)	KN Motif And Ankyrin Repeat Domains 2 (*KANK2*) (14)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

TUBAP5 (pseudogene)	MYB Proto-Oncogene Like 2 (*MYBL2*) (40)	

The shared lncRNAs and mRNAs using both methods, GDCRNATools and SPONGE have been labelled with the "*" sign.

According to Table 2, integrative analysis of GDCRNATools and SPONGE packages (GDCRNATools+SPONGE) resulted in two lncRNAs, nine mRNAs and seventy-four miRNAs. The list of lncRNA/pseudogene-mRNA-miRNA associations from GDCRNATools and GDCRNATools+SPONGE methods are available in Table S2 and Table S3, respectively. We conducted two downstream analyses for genes and miRNAs in the shared ceRNA network of HD cancers described in Table 2. Figure 2 illustrates a graphical representation of the shared ceRNA networks of HD cancers found in our study which was prepared using the Cytoscape software [24].

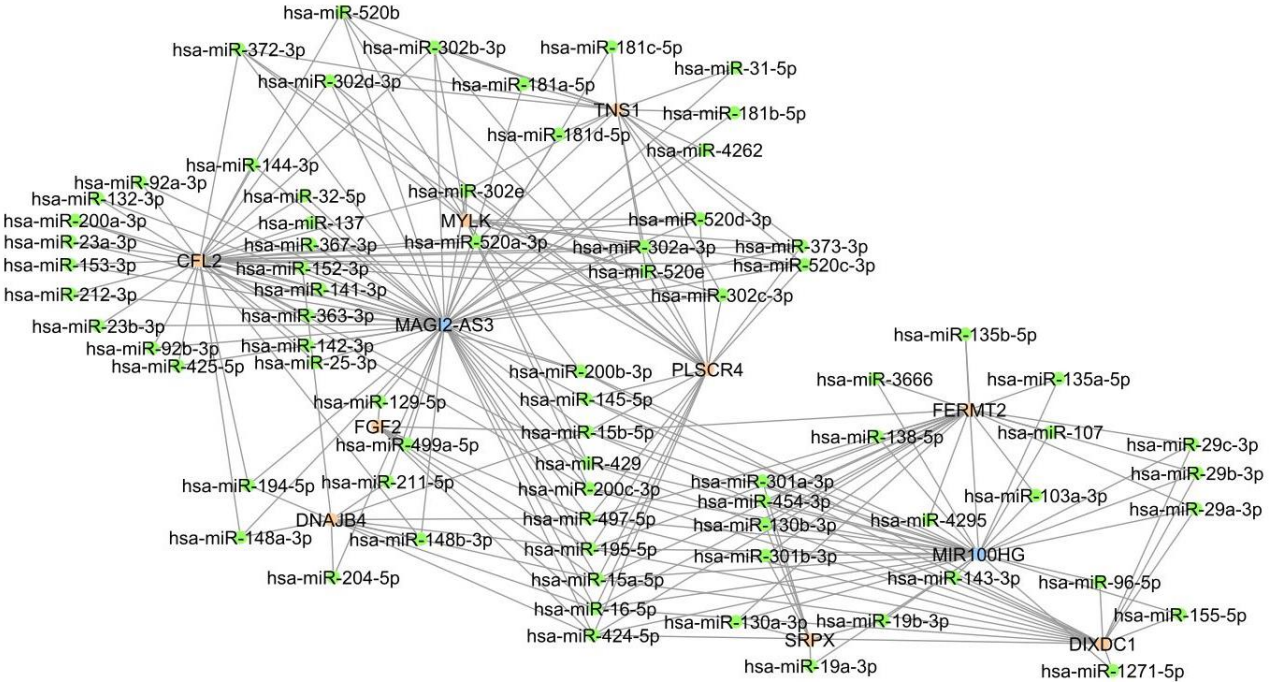

Figure 2. The shared competing endogenous RNA network across four hormone-dependent cancers, breast, prostate, colorectal, and endometrial, was constructed by the Cytoscape tool [24]. The blue, orange and green-colored circles represent long non-coding RNAs, mRNAs, and microRNAs, respectively.

According to Figure 2, the majority of miRNAs binds with the MAGI2-AS3 associated ceRNA network.

3.3. Functional Enrichment Analysis

Functional enrichment analysis was performed on the nine mRNAs obtained from the shared ceRNA network of HD cancers. The GO-cellular components (CC) of enrichment were mainly I band, stress fiber, contractile actin filament bundle, actin filament bundle, actomyosin, focal adhesion, and cell-substrate junction. Five out of nine genes (*CFL2*, *MYLK*, *TNS1*, *FERMT2*, and *DIXDC1*) were enriched in the actin-binding component in the GO-molecular functions (MF) pathway. KEGG pathway analysis showed that 3 out of 9 mRNAs (*FGF2*, *CFL2*, and *MYLK*) are involved in the regulation of actin cytoskeleton pathway. Results of enrichment analysis are illustrated in Figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

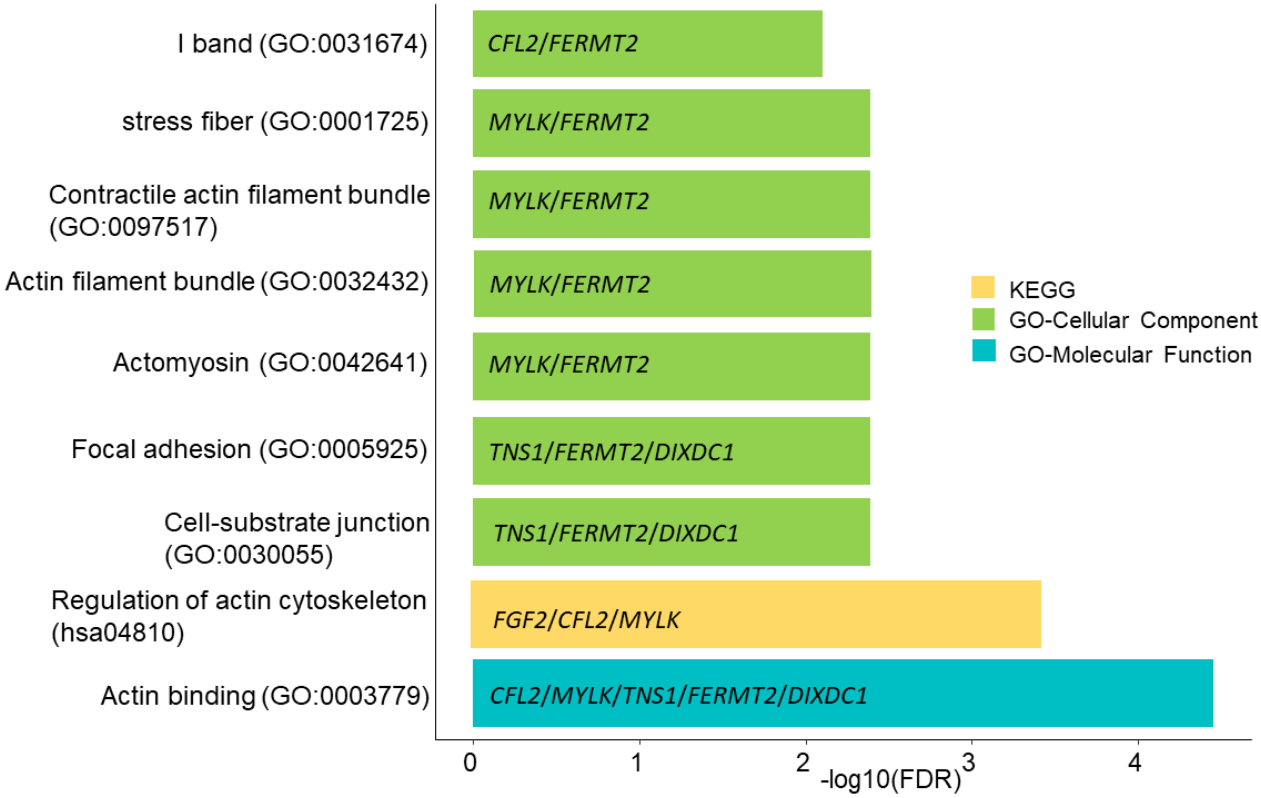

Figure 3. Functional enrichment analysis of nine genes in the shared competing endogenous RNA network of hormone-dependent cancers. There were seven, one, and one statistically significant component(s) in GO-CC, GO-MF, and KEGG pathways, respectively. Six out of nine genes, *CFL2*, *MYLK*, *TNS1*, *FERMT2*, *DIXDC1*, and *FGF2* are associated with actin-related pathways.

3.4. Survival Analysis

We performed Kaplan-Meier (K-M) survival analysis for genes and miRNAs in the shared ceRNA network of HD cancers. The genes and miRNAs lists were applied to individual HD cancer survival analysis. We filtered out genes and miRNAs that were significant from at least one survival analysis. We found two genes and forty-one miRNAs in the shared ceRNA network are significant in at least one HD cancer. Two mRNAs out of nine, *SRPX* and *DNAJB4* were significant in COAD and UCEC survival analysis. Figure 4 illustrates K-M curves for the two prognostic genes in COAD and UCEC.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

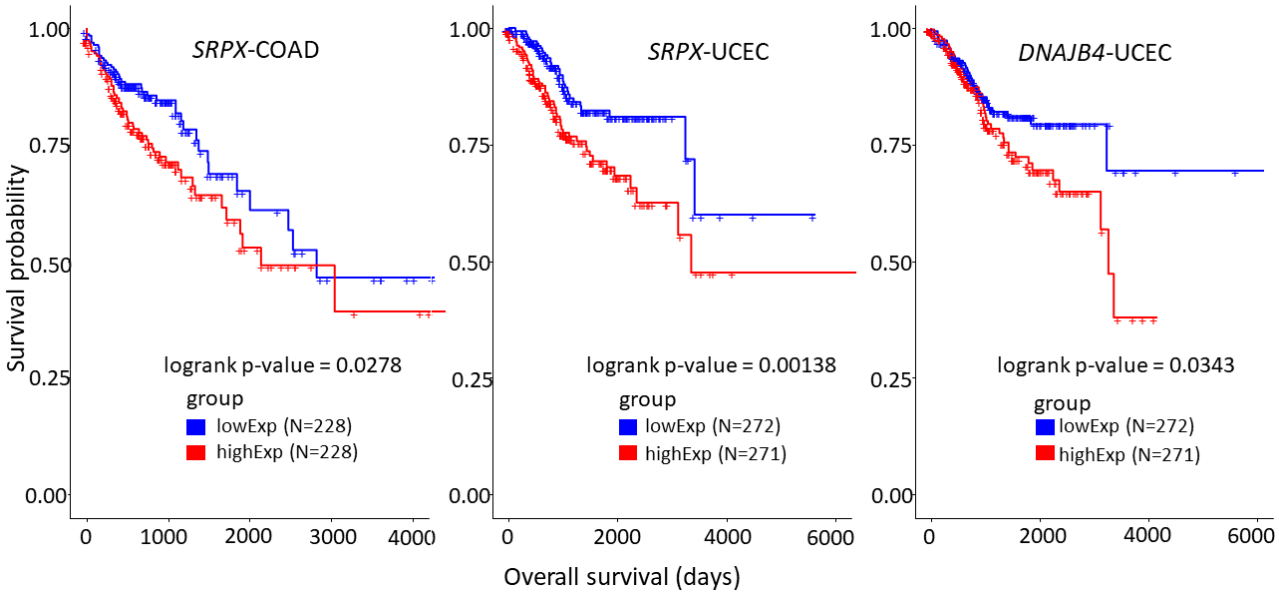

Figure 4. Kaplan Meier (K-M) survival analysis plots for prognostic mRNAs in the shared ceRNA networks of hormone-dependent cancers. The low-expressed *SRPX* gene exhibits prognostic ability in both colon and endometrial cancer. The low-expressed *DNAJB4* shows better survival in endometrial cancer.

We performed a K-M survival analysis for the list of miRNAs (seventy-four) in the shared ceRNA network of HD cancers. We selected miRNAs that are significant in at least one HD cancer survival analysis. We found that forty-one out of seventy-four miRNAs exhibit predictive ability in at least one HD cancer. Table 3 shows the significant miRNAs with hazard ratios and log-rank test p-values. We used the miRCancer web-based tool [27] to explore miRNA-HD cancer associations of listed significant miRNAs from the survival analysis.

Table 3. Log-rank test results (survival analysis) for miRNAs in the shared ceRNA network of hormone-dependent (HD) cancers.

Cancer	**miRNA**	**Hazard ratio**	**p-value**	**Cancer**	**miRNA**	**Hazard ratio**	**p-value**
BRCA	hsa-miR-16-5p[1,2,3]	0.672	0.0136	**UCEC**	hsa-miR-142-3p[1,2,3,4]	0.5634	0.0078
BRCA	hsa-miR-181c-5p[1,2,3,4]	0.6578	0.0114	**UCEC**	hsa-miR-148a-3p[1,2,3,4]	0.55	0.0055
BRCA	hsa-miR-195-5p[1,2,3,4]	0.6859	0.0212	**UCEC**	hsa-miR-152-3p[1,2,3,4]	1.6863	0.0156
BRCA	hsa-miR-200c-3p[1,2,3,4]	0.7097	0.04	**UCEC**	hsa-miR-212-3p[1,2,3]	1.7536	0.0096
BRCA	hsa-miR-204-5p[1,2,3,4]	0.6294	0.0052	**UCEC**	hsa-miR-25-3p[2,3,4]	1.5573	0.0365
BRCA	hsa-miR-29a-3p[1,2,3,4]	0.7168	0.0429	**UCEC**	hsa-miR-301a-3p[2,3,4]	1.8982	0.0032
BRCA	hsa-miR-29c-3p[1,3,4]	0.6313	0.0061	**UCEC**	hsa-miR-301b-3p[1,2,4]	1.6064	0.0277
BRCA	hsa-miR-301b-3p[1,2,4]	1.3884	0.0478	**UCEC**	hsa-miR-302a-3p [1,2,3,4]	0.5663	0.0071
BRCA	hsa-miR-31-5p[1,2,3]	0.5542	0.0003	**UCEC**	hsa-miR-302b-3p[1,2,3,4]	0.5608	0.0061
BRCA	hsa-miR-363-3p[2,3,4]	0.6961	0.0279	**UCEC**	hsa-miR-302c-3p[1,2,3,4]	0.5498	0.0049
BRCA	hsa-miR-372-3p[1,2,3]	1.409	0.0392	**UCEC**	hsa-miR-302d-3p[1,2,3,4]	0.5531	0.0053
COAD	hsa-miR-1271-5p[1,2,3,4]	1.6083	0.0166	**UCEC**	hsa-miR-302e[1,2,3,4]	0.4897	0.0008
COAD	hsa-miR-130a-3p[1,2,3,4]	1.8346	0.0021	**UCEC**	hsa-miR-367-3p[1,2,3,4]	0.5204	0.0021
COAD	hsa-miR-145-5p[1,2,3]	1.5823	0.0214	**UCEC**	hsa-miR-425-5p[1,2,3]	1.6045	0.0301
COAD	hsa-miR-181b-5p[1,2,4]	1.5294	0.0326	**UCEC**	hsa-miR-4262[1,2,3,4]	0.4897	0.0008

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 September 2021 doi:10.20944/preprints202109.0337.v1

COAD	hsa-miR-32-5p[1,2,3,4]	1.5932	0.0213	**UCEC**	hsa-miR-497-5p[1,2,3]	0.5285	0.0037
COAD	hsa-miR-497-5p[1,2,3]	1.5895	0.0206	**UCEC**	hsa-miR-520b[1,2,3]	0.5896	0.0129
COAD	hsa-miR-96-5p[1,2,3]	1.4888	0.0474	**UCEC**	hsa-miR-520c-3p[1,2,3]	0.5791	0.0099
PRAD	hsa-miR-19a-3p[1,2,3]	6.9585	0.026	**UCEC**	hsa-miR-520d-3p[2,3]	0.5043	0.0012
PRAD	hsa-miR-29b-3p[1,2,3]	0.2343	0.0434	**UCEC**	hsa-miR-520e[2,3]	0.626	0.0273
READ	hsa-miR-155-5p[1,2,3,4]	0.4544	0.0426				

The superscripted 1, 2, 3, and 4 implies that the given miRNA has been experimentally validated in prostate, breast, colorectal, and endometrial cancers.

As stated in Table 3, hsa-miR-301b-3p acts as a prognostic candidate in BRCA and UCEC, whereas hsa-miR-497-5p is significant in both COAD and UCEC survival analyses. According to results, eleven, seven, one, two, and twenty miRNAs were obtained from the BRCA, COAD, READ, PRAD, and UCEC survival analyses, respectively.

4. Discussion

Previous genome-wide and transcriptome-wide analyses have reported the existence of a shared genetic aetiology of HD cancers [28]. As ceRNAs have a critical role in gene and molecular pathways, identifying a shared ceRNA network of HD cancers will contribute to understanding the shared genetic aetiology of HD cancers. Here we investigated the availability of a shared ceRNA network of four common HD cancers. Previous HD cancer-related ceRNA analyses have focused on lncRNA-associated networks [6-10]. We extended the scope of ceRNA research to include pseudogene-associated cross- HD cancer ceRNA networks.

We utilised two ceRNA analysis R packages, GDCRNATools and SPONGE, to improve the predictive power of ceRNA analyses [19,20]. Prior HD cancer-associated ceRNA analyses have used the sensitivity correlation method defined in the GDCRNATools. The sensitivity correlation cannot account for the presence of multiple miRNAs for a given ceRNA pair. To address this limitation, we used the sparse partial correlation method implemented in the SPONGE R/Bioconductor package. We aggregated HD cancer-specific lncRNA/pseudogene-miRNA-mRNA associations (significant from both GDCRNATools and SPONGE) to evaluate the shared lncRNA/pseudogene-miRNA-mRNA triplets across four HD cancer types. We identified two lncRNAs, nine mRNAs, and seventy-four miRNAs common across lncRNA-mRNA-miRNA networks in HD cancers. None of the pseudogene-related shared ceRNA associations selected from GDCRNATools was significant from the SPONGE method. Previous cancer studies have extensively described two lncRNAs in the shared ceRNA network, MAGI2-AS3, and MIR100HG [29-35]. Du, *et al.* [30] have shown that MAGI2-AS3 upregulation inhibits BRCA metastatic progression by decreasing miR-374a and enhancing *PTEN* expression. Ren, *et al.* [31] have revealed that MAGI2-AS3 promotes COLCA progression through regulating the miR-3163-*TMEM106B* axis. Moreover, the MAGI2-AS3 promoter was hypermethylated in several cancers such as COAD, READ and UCEC [32].

The lncRNA MIR100HG, host gene for miR-100, let-7a-2 and miR-125b cluster, has been previously reported to have a role in gastric cancer, COLCA, and BRCA [33-35]. Li, *et al.* [34] demonstrated that MIR100HG overexpression cause COLCA progression and is a poor prognosis in COLCA patients. It also promotes triple-negative BRCA cells' migration, invasion, and proliferation by sponging the miR-5590-3p-*OTX1* axis [35]. Our study reveals the ceRNA role of MIR100HG in PRAD and UCEC; and MAGI2-AS3 in PRAD for

 doi:10.20944/preprints202109.0337.v1

the first time. Wet-lab experiments are required to understand the molecular mechanism of MAGI2-AS3 and MIR100HG in these cancers.

We found nine mRNAs in the shared ceRNA network of HD cancers in which six and three mRNAs were paired with MAGI2-AS3 and MIR100HG, respectively. Eight out of nine mRNA-lncRNA axes were identified for the first time in cancer-related ceRNAs. These ceRNA pairs are likely to be involved in cancer pathways as they were significant across four cancer types. The MAGI2-AS3/miR-31-5p/*TNS1* axis identified in our study has been shown to regulate migration and invasion ability in bladder cancer cell lines [36].

We conducted two downstream analyses, survival analysis and functional enrichment analysis to identify important mRNAs and miRNAs in the shared ceRNA network. Two out of nine mRNAs, *SRPX* (UCEC and COAD) and DNAJB4 (UCEC) were found as prognostic markers in at least one HD cancer from the survival analysis. The *SRPX* gene acts as a tumour-suppressor gene, which is down-regulated in several malignancies, including PRAD, COLCA and neuroendocrine (cells that release hormones into the blood in response to stimulation of the nervous system) cancers [37]. All these malignancies are biologically related to hormones. Therefore, the role of the *SRPX* gene in hormone-related cancers should be further investigated. Currently, *SRPX* is being examined as a potential cancer drug under patent number US 9,290,744 B2 [38].

The *DNAJB4* gene (also known as *HLJ1*) belongs to the DNJ family heat shock proteins (HSPs) and is regarded as a tumour-suppressor in COAD, BRCA, lung, and gastric cancers [39]. HSPs have been reported as biomarkers and potential drug targets of cancers for decades. A recent integrative analysis of multi-omics data uncovered the distinct impact of several HSPs (including *DNAJB4*) members in BRCA progression [40]. Our study discovered that *DNAJB4* could act as a prognostic marker in UCEC. Moreover, GDCRNATools analysis (only hypergeometric test and Pearson correlation analysis) identified *DNAJB4* can be paired with all three lncRNAs in the shared network, MBNL1-AS1, MAGI2-AS3, and MIR100HG. Therefore, *in-vivo*/*vitro* experiments are required to evaluate its tumour-suppressive/oncogenic role in HD cancers.

We conducted a separate survival analysis for miRNAs in the shared ceRNA network. Interestingly, our miRNA survival analysis revealed that ~55% (41/74) of miRNAs in the shared ceRNA network of HD cancers are associated with disease survival in at least one HD cancer type. These forty-one miRNAs have been experimentally validated for their functional role in at least three HD cancers (out of four types of interest), providing confidence to our computational results [27]. Among these forty-one miRNAs, twenty-eight have been differentially expressed in HD cancers. We found multiple prognostic miRNAs from the same miRNA family, two from miR-181 (in COAD and BRCA), three from miR-29 (in BRCA and PRAD), three from miR-301 (in BRCA and UCEC), five from miR-302 (in UCEC), and four from miR-520 (in UCEC). The four members of the miR-520 family were prognostic in UCEC are required to be determined through experiments. We found both miR-302 and miR-367 as prognostic markers from the shared ceRNA network. The miR-302/367 cluster has been previously identified in PRAD, BRCA, COLCA, and UCEC associated pathways supporting our findings [41].

According to functional enrichment analysis, six out of nine mRNAs were associated with actin-related GO and KEGG pathways. The actin dynamics and actin-specific molecular signalling have shown potential clinical significance on non-genomic steroid hormone actions on tumour cells [42]. All these facts supported by literature have improved the significance of our study's shared ceRNA network of HD cancers.

 doi:10.20944/preprints202109.0337.v1

A limitation of this study is that we selected both experimentally validated and predicted miRNA-target interactions only from two databases, miRcode and starBase to include a substantial set of miRNA-mRNA/lncRNA/pseudogene interactions. We did not include circular RNAs (circRNAs) for the ceRNA network analysis as their expression levels are not available in TCGA. Nevertheless, our findings have important biological implications of HD cancers.

Herein, we identified a shared ceRNA network that can be facilitated to understand the shared genetic aetiology of HD cancers. The shared ceRNA network consists of two lncRNAs, nine mRNAs, and seventy-four miRNAs that have shown links with individual HD malignancies from the literature. Our study lays the groundwork for future research into the understanding role of these mRNAs, miRNAs, and lncRNAs in the shared genetic susceptibility of HD cancers. Future directions could lead to a supervised machine learning approach to understand molecular effects on ceRNA networks of HD cancers.

5. Conclusions

We conducted the first extensive computational study that compares ceRNA networks (both lncRNA and pseudogene) in a group of related cancers, HD cancers. The shared ceRNA network comprises two lncRNAs, nine mRNAs, and seventy-four miRNAs, and part of them were described for the first time in certain HD cancers. A global view of the functional ceRNA networks of large sample sets encompassing multiple tumour types may help identify potential unexpected targets that apply to a cancer subset, such as HD cancers. Moreover, identifying novel risk-associated lncRNAs, pseudogenes, miRNAs, and mRNAs across a group of related cancers will significantly contribute to understanding their shared disease biology. Further experimental investigations should be conducted to understand the tumour-suppressive/oncogenic/cancer-driven role of identified ceRNAs in HD cancers.

Supplementary Materials: Table S1: Number of statistically significant long non-coding RNA/pseudogene-mRNA-microRNA triplets in each cancer type (the ceRNA associations from GDCRNATools and both GDCRNATools and SPONGE-Sparse Partial correlation ON Gene Expression have been tabulated separately). Table S2: A detailed list of long non-coding RNA/pseudogene-mRNA-microRNAs identified from the GDCRNATools analysis. Table S3: A detailed list of long non-coding RNA-mRNA-microRNAs identified from both GDCRNATools and SPONGE (Sparse Partial correlation ON Gene Expression).

Data availability: Publicly available TCGA RNA-seq and miRNA-seq expression data were downloaded through the GDC Data Portal (https://portal.gdc.cancer.gov/repository). All statistical analyses and graph preparations were performed using the R statistical software freely available at https://cran.r-project.org/. The ceRNA network graphs were drawn using the Cytoscape software freely available at https://cytoscape.org/download.html.

Acknowledgements: The datasets reported in this work originated from the TCGA Research Network: https://www.cancer.gov/tcga. The computational resources and services used in this work were provided by the eResearch Office, Queensland University of Technology, Brisbane, Australia.

Author Contribution: Conceptualization: D.K.J, N.S.G., and M.E.R.; methodology: D.K.J.; formal analysis: D.K.J., writing-original draft: D.K.J.; writing-review and editing: D.K.J., N.S.G., M.E.R., J.B., and E.S.; supervision: N.S.G., M.E.R., J.B., and E.S. All authors read and approved the final manuscript.

Funding: D.K.J. acknowledges QUTPRA and QUT HDR tuition fee sponsorship. M.E.R. thanks the support of the NHMRC and Australian Research Council (GNT1102821). N.S.G. and J.B.

acknowledge funding support from the Advance Queensland Industry Research Fellowship. J.B. acknowledges support from the NHMRC Career Development Fellowship and a Cancer Council Queensland grant.

Institutional Review Board Statement: Not applicable

Institutional Consent Statement: Not applicable

Conflict of interest: The authors declare no conflict of interest.

References

1. Bhaskaran, M.; Mohan, M. MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. *Vet. Pathol.* **2014**, *51*, 759-774.
2. Salmena, L.; Poliseno, L.; Tay, Y.; Kats, L.; Pandolfi, P.P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **2011**, *146*, 353-358.
3. Lin, W.; Liu, H.; Tang, Y.; Wei, Y.; Wei, W.; Zhang, L.; Chen, J. The development and controversy of competitive endogenous RNA hypothesis in non-coding genes. *Mol. Cell. Biochem.* **2021**, *476*, 109-123.
4. Wei, Y.; Chang, Z.; Wu, C.; Zhu, Y.; Li, K.; Xu, Y. Identification of potential cancer-related pseudogenes in lung adenocarcinoma based on ceRNA hypothesis. *Oncotarget* **2017**, *8*, 59036-59047.
5. Wang, Y.; Hou, J.; He, D.; Sun, M.; Zhang, P.; Yu, Y.; Chen, Y. The Emerging Function and Mechanism of ceRNAs in Cancer. *Trends. Genet.* **2016**, *32*, 211-224.
6. Guo, L.; Yang, G.; Kang, Y.; Li, S.; Duan, R.; Shen, L.; Jiang, W.; Qian, B.; Yin, Z.; Liang, T. Construction and Analysis of a ceRNA Network Reveals Potential Prognostic Markers in Colorectal Cancer. *Front. Genet.* **2020**, *11*, 418.
7. Wang, Y.; Huang, T.; Sun, X.; Wang, Y. Identification of a potential prognostic lncRNA-miRNA-mRNA signature in endometrial cancer based on the competing endogenous RNA network. *J. Cell. Biochem.* **2019**, *120*, 18845-18853.
8. Ouyang, D.; Li, R.; Li, Y.; Zhu, X. Construction of a Competitive Endogenous RNA Network in Uterine Corpus Endometrial Carcinoma. *Med. Sci. Monit.* **2019**, *25*, 7998-8010.
9. Xu, N.; Wu, Y.P.; Yin, H.B.; Xue, X.Y.; Gou, X. Molecular network-based identification of competing endogenous RNAs and mRNA signatures that predict survival in prostate cancer. *J. Transl. Med.* **2018**, *16*, 274.
10. Tuersong, T.; Li, L.; Abulaiti, Z.; Feng, S. Comprehensive analysis of the aberrantly expressed lncRNA-associated ceRNA network in breast cancer. *Mol. Med. Rep.* **2019**, *19*, 4697-4710.
11. Fiannaca, A.; Paglia, L.; Rosa, M.; Rizzo, R.; Urso, A. miRTissue (ce): extending miRTissue web service with the analysis of ceRNA-ceRNA interactions. *BMC Bioinformatics* **2020**, *21*, 199.
12. Wang, P.; Li, X.; Gao, Y.; Guo, Q.; Wang, Y.; Fang, Y.; Ma, X.; Zhi, H.; Zhou, D.; Shen, W.; et al. LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. *Nucleic Acids Res.* **2019**, *47*, D121-d127.
13. Das, S.; Ghosal, S.; Sen, R.; Chakrabarti, J. lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. *PLoS One* **2014**, *9*, e98965.
14. Jiang, Y.; Qiao, Z.; Jiang, J.; Zhang, J. LINC00958 promotes endometrial cancer cell proliferation and metastasis by regulating the miR-145-3p/TCF4 axis. *J. Gene. Med.* **2021**, *23*, e3345.
15. Fang, Q.; Sang, L.; Du, S. Long noncoding RNA LINC00261 regulates endometrial carcinoma progression by modulating miRNA/FOXO1 expression. *Cell. Biochem. Funct.* **2018**, *36*, 323-330.

16. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375*, 1109-1112.
17. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139-140.
18. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47.
19. Li, R.; Qu, H.; Wang, S.; Wei, J.; Zhang, L.; Ma, R.; Lu, J.; Zhu, J.; Zhong, W.D.; Jia, Z. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics* **2018**, *34*, 2515-2517.
20. List, M.; Dehghani Amirabad, A.; Kostka, D.; Schulz, M.H. Large-scale inference of competing endogenous RNA networks with sparse partial correlation. *Bioinformatics* **2019**, *35*, i596-i604.
21. Jeggari, A.; Marks, D.S.; Larsson, E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* **2012**, *28*, 2062-2063.
22. Li, J.H.; Liu, S.; Zhou, H.; Qu, L.H.; Yang, J.H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **2014**, *42*, D92-97.
23. Paci, P.; Colombo, T.; Farina, L. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst. Biol.* **2014**, *8*, 83.
24. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498-2504.
25. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* **2012**, *16*, 284-287.
26. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model*; Springer, New York, NY: 2000.
27. Xie, B.; Ding, Q.; Han, H.; Wu, D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* **2013**, *29*, 638-644.
28. Kar, S.P.; Beesley, J.; Amin Al Olama, A.; Michailidou, K.; Tyrer, J.; Kote-Jarai, Z.; Lawrenson, K.; Lindstrom, S.; Ramus, S.J.; Thompson, D.J.; et al. Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types. *Cancer Discov.* **2016**, *6*, 1052-1067.
29. Kai-Xin, L.; Cheng, C.; Rui, L.; Zheng-Wei, S.; Wen-Wen, T.; Peng, X. Roles of lncRNA MAGI2-AS3 in human cancers. *Biomed. Pharmacother.* **2021**, *141*, 111812.
30. Du, S.; Hu, W.; Zhao, Y.; Zhou, H.; Wen, W.; Xu, M.; Zhao, P.; Liu, K. Long non-coding RNA MAGI2-AS3 inhibits breast cancer cell migration and invasion via sponging microRNA-374a. *Cancer Biomark.* **2019**, *24*, 269-277.
31. Ren, H.; Li, Z.; Tang, Z.; Li, J.; Lang, X. Long noncoding MAGI2-AS3 promotes colorectal cancer progression through regulating miR-3163/TMEM106B axis. *J. Cell. Physiol.* **2020**, *235*, 4824-4833.
32. Xiong, Y.; Wei, Y.; Gu, Y.; Zhang, S.; Lyu, J.; Zhang, B.; Chen, C.; Zhu, J.; Wang, Y.; Liu, H.; et al. DiseaseMeth version 2.0: a major expansion and update of the human disease methylation database. *Nucleic Acids Res.* **2017**, *45*, D888-d895.
33. Li, J.; Xu, Q.; Wang, W.; Sun, S. MIR100HG: a credible prognostic biomarker and an oncogenic lncRNA in gastric cancer. *Biosci. Rep.* **2019**, *39*.
34. Li, W.; Yuan, F.; Zhang, X.; Chen, W.; Tang, X.; Lu, L. Elevated MIR100HG promotes colorectal cancer metastasis and is associated with poor prognosis. *Oncol. Lett.* **2019**, *18*, 6483-6490.

35. Chen, F.Y.; Zhou, Z.Y.; Zhang, K.J.; Pang, J.; Wang, S.M. Long non-coding RNA MIR100HG promotes the migration, invasion and proliferation of triple-negative breast cancer cells by targeting the miR-5590-3p/OTX1 axis. *Cancer Cell Int.* **2020**, *20*, 508.
36. Tang, C.; Cai, Y.; Jiang, H.; Lv, Z.; Yang, C.; Xu, H.; Li, Z.; Li, Y. LncRNA MAGI2-AS3 inhibits bladder cancer progression by targeting the miR-31-5p/TNS1 axis. *Aging (Albany NY)* **2020**, *12*, 25547-25563.
37. Werner, R.J.; Schultz, B.M.; Huhn, J.M.; Jelinek, J.; Madzo, J.; Engel, N. Sex chromosomes drive gene expression and regulatory dimorphisms in mouse embryonic stem cells. *Biol. Sex. Differ.* **2017**, *8*, 28.
38. Green, M.; Gao, G.; Santra, M.K.; Bhatnagar, S. SRPX for treatment of cancer. US 9,290,744 B2, 2016.
39. Acun, T.; Doberstein, N.; Habermann, J.K.; Gemoll, T.; Thorns, C.; Oztas, E.; Ried, T. HLJ1 (DNAJB4) Gene Is a Novel Biomarker Candidate in Breast Cancer. *Omics* **2017**, *21*, 257-265.
40. Buttacavoli, M.; Di Cara, G.; D'Amico, C.; Geraci, F.; Pucci-Minafra, I.; Feo, S.; Cancemi, P. Prognostic and Functional Significant of Heat Shock Proteins (HSPs) in Breast Cancer Unveiled by Multi-Omics Approaches. *Biology (Basel)* **2021**, *10*.
41. Liu, J.; Wang, Y.; Ji, P.; Jin, X. Application of the microRNA-302/367 cluster in cancer therapy. *Cancer Sci.* **2020**, *111*, 1065-1075.
42. Stournaras, C.; Gravanis, A.; Margioris, A.N.; Lang, F. The actin cytoskeleton in rapid steroid hormone actions. *Cytoskeleton (Hoboken)* **2014**, *71*, 285-293.