Article

Identification of novel multi-omics expression landscapes and meta-analysis of landscape-based competitive endogenous RNA networks in ALDH+ lung adenocarcinoma stem cells

Wei Yang[1,2,3,4], Yong Liang[5], Yuanyuan Zheng[1,2,3,4], Haitao Luo[1,2,3,4], Xiaofei Yang[1,2,3,4*] and Furong Li[1,2,3,4*]

1   Translational Medicine Collaborative Innovation Center, The Second Clinical Medical College(Shenzhen People's Hospital), Jinan University, Shenzhen 518020, China;

2   The First Affiliated Hospital(Shenzhen People's Hospital), Southern University of Science and Technology, Shenzhen 518020, China;

3   Guangdong Engineering Technology Research Center of Stem cell and Cell therapy, Shenzhen 518020, China;

4   Shenzhen key Laboratory of Stem Cell Research and Clinical Transformation, Shenzhen 518020, China

5   Shenzhen Xbiome Biotech Co., Ltd, Shenzhen 518057, China

*   Correspondence: <u>frli62@163.com</u>;<u>sophiayang81@hotmail.com</u>

Simple Summary: Non-coding RNAs(ncRNAs) landscape based on cancer stem cells (CSCs) provide novel predictive biomarkers for lung adenocarcinoma patients. Comprehensive and systematic in silico analysis on ncRNAs, including micro RNAs(miRNAs), long non-coding RNAs(lncRNAs), circular RNAs(circRNAs) uncover potential survival mechanisms on CSCs and cancer cells, respectively. The target of our research was to predict added novel ncRNAs, such as miRNAs, lncRNAs, circRNAs, and probe the interplay of those small molecules to excavate the potential signaling mechanism of ALDH+ H1975 lung adenocarcinoma stem cells(LSCs) and ALDH- H1975 lung adenocarcinoma tumor cells(LTCs). Our study integrated exhaustive ncRNAs and uncover their functional roles involved in LSCs and LTCs behaviors, which provided promising biomarkers for lung adenocarcinoma(LUAD) diagnosis and therapy.

Abstract: ALHD+ H1975 lung adenocarcinoma stem cells(LSCs), confirmed to be found in LUAD, is one rare subset within lung adenocarcinoma(LUAD) cancer cells. They could self-renew, drive tumor initiation, growth, metastasis, and recurrence, and are also considered to be the main cause of poor prognosis because of their intrinsic resistance to drug and chemotherapy, which prompts them to be a promising target for LUAD therapy. NcRNAs, including miRNAs, lncRNAs, circRNAs are thought to exert many significant regulatory functions in the development of LUAD. However, systematic and comprehensive molecular profiling on ncRNAs and messenger RNAs(mRNAs) between LSCs and ALDH- H1975 lung adenocarcinoma tumor cells(LTCs) is insufficient. Therefore, by using stringent bio-informatics pipelines, we identified a set of novel ncRNAs including miR-NAs, lncRNAs, circRNAs, and obtained differentially expressed landscapes on ncRNAs and mRNAs. Through further meta-analysis with those landscapes subsequently, we constructed novel competitive endogenous RNA(ceRNA) networks to excavate the potential molecular mechanisms that regulate the hallmarks of LSCs and LTCs. Herein, these results summarize novel ncRNAs and the fundamental roles of differentially expressed ncRNAs that have been implicated in LSCs and LTCs behaviors. It also provides a more comprehensive resource for the futural identification of diagnostic, therapeutic, and prognostic biomarkers in LUAD.

Keywords: novel ncRNAs; DE landscape; LSCs and LTCs; LncRNA; MiRNA; CircRNA; ceRNA networks; molecular mechanism

1. Introduction

Lung cancer is the most common respiratory tumor and the leading cause of cancer deaths worldwide, with a ~18% mortality rate and ~22.4% incident ratio of all cancer deaths[1,2]. Non-small cell lung cancer(NSCLC) is the most common form of lung cancer, accounting for ~90% of all patients with lung cancer, and adenocarcinoma of lung is its most common histological subtype[3], with a dismal prognosis and 15% of patients surviving 5 years after confirmation[4].

CSCs, a biologically particular subset with stem-like characteristics of most cancer cells, play critical roles in tumor heterogeneity, which are involved in tumor initiation, growth, progression, metastasis, recurrence, and drug resistance[5]. Increasingly research prove that the main leading cause of the poor prognosis for LUAD[6] is LSCs[7], which could initiate and sustain primary and metastatic cancer relapse and growth[8]. Further, previous evidence[9] and our last studies[10] shown that ALDH+ H1975 lung adenocarcinoma cells displayed strong CSC-like properties[11] and were acted as LSCs to influence the formation of a tumor niche and tumor initiation[6]. Our study shown that LSCs' capabilities of self-renewal and differentiation could enable to generate more tumor cells and promote tumor growth[10], metastasis, and recurrence. Therefore, LSCs are a promising objective in LUAD therapy[6]. Elucidation of the differences between LSCs and LTCs and relative molecular pathways that affect the characteristics of LSCs[12] is significant for identifying and evaluating potential diagnostic, therapeutic, and prognostic biomarkers[13].

LncRNAs, circRNAs, miRNAs are three subtypes of non-coding RNAs, which account for a huge proportion of the human transcriptome[14]. The diverse types of those ncRNAs are marked by sequence length, spatial architecture, and regulatory functions in various pathophysiological processes in cancer. Separately, miRNAs[15] are the most studied single-stranded ncRNAs with a length of ~20 nucleotides[16]. According to binding to specific 3' untranslated regions(UTR) of their target mRNAs or RNA, miRNAs could silence multiple targeted mRNAs expression and RNA simultaneously[17]. Regulatory modular of one-to-many of miRNAs affect multiple oncogenic and tumor repressive pathways[18,19], such as overexpressing has-miR-17-92 in lung cancer acting to enhance cell proliferation[20], has-let-7 in lung cancer effecting patients' survival[21], etc.. LncRNAs are over 200 nucleotides in length[22] with the ability to regulate gene expression rather than protein-encoding capability[23]. Extensive studies reply that lncRNAs may influence cancer progression via regulating relational mRNAs and miRNAs[24]. CircRNAs are a novel subtype of non-coding RNAs. Though circRNAs function likes lncRNAs[25], their uniquely circular structure makes them have higher stability than linear lncRNAs[26]. They are confirmed to play a vital role in various types of cancer[27]. Great advances show that a pile of mRNA transcripts is regulated or repressed by lncRNAs/circRNAs and miRNAs, individually or together[28]. It is a novel way of the regulatory network[29] according to the interaction between mRNAs, lncRNAs/circRNAs, and miRNAs, called ceRNA network. More importantly, lncRNAs/circRNAs, upregulate transcripts expression by binding miRNA sites, such as lncRNA HRCR acting as s sponge of miR-233 to prevent cardiac hypertrophy[30], lncRNA CHRF functioning as an endogenous sponge of miR-489 to limit miR-489 expression[31], lncRNA APF working as a sponge of miR-188-3p to prevent decomposition of ATG7[32].

Previous attempts focused on the discrepancy of lung adenocarcinoma cells and adjacent normal cells to explore the cause and treatment options of tumorigenesis, tumor metastasis, relapse, and drug resistance, the different characteristics and molecular mechanism between LSCs and LTCs, a new and crucial dimension in cancer therapy, is insufficient. Therefore, identification of biomarkers, a meta-analysis of diversity between LSCs and LTCs can help us to deeply reveal the molecular events associated with tumorigenesis, metastasis, relapse, and drug resistance in LUAD.

2. Materials and Methods

2.1 RNA isolation and miRNA sequencing

Two ALDH+ LSCs and two ALDH- LTCs cellular samples were collected from our previous study[10]. Sum up to ~3-5 x 106 cells were used for the isolation of RNA. Total RNAs of preparing cellular samples were extracted using TRIzolTM reagent (Invitrogen, Carlsbad, CA, USA) based on the manufacturer's protocols. At least 3μg RNA per sample were prepared as materials for the RNA samples. Sequencing libraries were generated using NEBNext® Multiplex Small RNA Library Prep Set for Illumina® (NEB, USA.) according to the manufacturer's recommendations. Index codes were added to attribute sequences to each sample. Briefly, NEB 3' sequence replication(SR) Adaptor was directly and specifically linked to the 3' end of miRNA, small interfering RNA(siRNA), and piwi-interacting(piRNA). After the 3' ligation reaction, the SR random primer(RT) Primer hybridized to the excess of 3' SR Adaptor, which transformed the single-stranded DNA adaptor into a double-stranded DNA molecule. This important step prevented adaptor--dimer formation. Based on that, dsDNAs were not substrates for ligation-mediated by T4 RNA Ligase 1 and therefore did not ligate to the 5' SR Adaptor in the subsequent ligation step. 5' ends adapter was ligated to 5' ends of miRNAs, siRNA, and piRNA. Then first-strand cDNA was synthesized using M-MuLV Reverse Transcriptase (RNase H-). PCR amplification was performed by using LongAmp® Taq 2X Master Mix, SR Primer for Illumina, and index(X) primer.

PCR products were purified on an 8% polyacrylamide gel (100V, 80 min). DNA fragments corresponding to 140~160bp(the length of small non-coding RNA plus the 3' and 5' adaptors) were recovered and dissolved in an 8μL elution buffer. At last, library quality was assessed on the Agilent Bioanalyzer 2100 system by using DNA High Sensitivity Chips. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq SR Cluster Kit v3-cBot-HS (Illumia) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq 2500/2000 platform and 50bp single-end reads were generated.

2.2 RNA raw data quality control and reads mapping statistics

Raw reads of miRNAs and lncRNAs/circRNAs/mRNAs from the Illumina Hiseq platform were processed to remove the reads with poly-N/adapter sequence, and low-quality reads. Raw reads of miRNAs were deeply processed to filter the reads containing poly A/T/G/G. Clean reads from each sample of miRNAs were mapped to the human reference genome(version GRCh38) by using Bowtie[33], without any mismatch. Clean reads of lncRNAs/circRNAs/mRNAs were aligned to the same human reference genome above mentioned by using HISAT2 with '—rna-standness RF' and other default parameters[34]. The mapping results were sorted, analyzed, and indexed by using ReSeqTools[35] and Samtools[36].

2.3 Identification of miRNAs

Based on those above aligned small RNA reads, miRBased20.0[37] was used as a reference, and miRDeep2/srna-tools-cli[38] were used to identify the potential miRNAs and depict the secondary structure. According to RepeatMasker[39], the Rfam database[40], mapped reads originating from protein-coding genes, repeat sequences, rRNA, tRNA, snRNA, and snoRNA[41] were removed. According to the characteristics of the hairpin structure of miRNA precursor, the Dicer cleavage site and the minimum information of the small RNA reads unannotated in the above steps, miREvo[42] and miRDeep2[38] were integrated to predict novel miRNAs. And base bias on the first position and all positions of all identified miRNAs was counted. In the alignment and annotation above mentioned, diversity RNA were counted with the following priority rule: known miRNA > rRNA > tRNA > snRNA > snoRNA > repeat > mRNA > novel miRNA, to make each unique small RNA matched to only one annotation. MiRNA expression levels were estimated by Transcripts Per Million(TPM)[43], which equation was normalized expression level:

$$TPM = (\text{Mapped Reads} * 1{,}000{,}000)/\text{Total Reads}$$

2.4 Identification and quantification of lncRNAs/circRNAs, mRNAs

Total mapped reads of each sample were assembled by StringTie[44] in a reference-based approach with an optional de novo assemble step, and a comprehensive annotation

file with full-length transcripts and potential novel ncRNAs was generated. Transcripts with low confidence, length <=200nt, repeat annotation, low-expression were removed. Quantification on known ncRNAs and mRNAs was performed by StringTie for each sample. All of the four tools, containing CNCI(Coding-Non-Coding-Index)[45], CPC(Coding Potential Calculator)[46], Pfam-scan[47], PhyloCSF(phylogenetic codon substitution frequency)[48] with default parameters were used to predict transcripts coding potential. Those with coding potential were filtered and the remaining were candidate sets of novel lncRNAs. Cufflinks were used to calculate Fragments Per Kilobase of exon model per Million mapped fragments (FPKM)[49] of both lncRNAs and coding genes in each sample.

FPKM=Total mapped exon Fragments/(mapped reads(millions)*exon Length(KB))

BAM files mentioned above were used to identify circRNAs. The detection and identification of circRNAs were based on our previous studies[10]. TMP was used to quantify circle RNAs.

2.5 Differential expression analysis

Following digital transcripts' expression quantification, differential expression analysis of lncRNAs and mRNAs between LSCs group and LTCs group was performed by using Cuffdiff[50], which was based on the negative binomial distribution model. Transcripts with a P-adjust <0.05 and |log2 foldchange|>=1.3 were assigned as differentially expressed. mRNAs with |log2 foldchange|>=1.3 and P-adjust <0.05 were defined as differential expression.

Differential expression analysis of miRNA and circRNAs between two conditions was performed using the DESeq R package[51]. The P-values were adjusted using the Benjamini & Hochberg method for controlling the false discovery rate[52]. Corrected P-value < 0.05 and |log2 foldchange| >=1.3 were set as the threshold for significantly differential expression.

2.6 Kyoto Encyclopedia of Genes and Genomes(KEGG) analysis and Gene Ontology(GO) analysis on DE ncRNAs

KEGG pathway analysis of DE mRNAs was carried by using DAVID 6.8[53]with a default parameter, to ascertain the potential functions of these genes that participated in cancer progression on LSCs and LTCs, separately. GO analysis was performed based on Metascape 3.5 websites[54], which could organize genes into hierarchical categories and uncover the gene regulatory network on the latest biological processes database. The P-value was adjusted by using the Benjamini & Hochberg method for controlling the false discovery and identifying the significant KEGG pathway and GO terms.

2.7 CeRNA network analysis and survival analysis

LncRNAs-mRNAs co-location networks were predicted and performed based on the parameters of upstream and downstream 100kb distance of differential expression lncRNAs.

Predicting the target gene of miRNAs was performed by miRanda[55], PITA[56], and RNAhybrid[57], separately. The overlap results on miRNAs and mRNAs pair in those three softwares were the last result. At the same time, Predicting the target miRNA of lncRNAs/circRNAs was performed by miRanda.

Produced the ceRNA networks was implemented by using Cytoscape 3.9.0[58] and identified the modules in ceRNA networks was performed by using the MCODE 2.0.0[59] Cytoscape 3.9.0 plug-in with default parameters. Survival analysis was performed by using UALCAN website[60] and lnCAR website[61] with default parameters.

3. Results

3.1. Identification of novel and known miRNAs and DE miRNAs in LSCs and LTCs

In our analysis, two LSC samples(Sample 1 of ALDH+ H1975 lung adenocarcinoma tumor cell(LSC1-m) and Sample 2 of ALDH+ H1975 lung adenocarcinoma tumor cell(LSC2-m)) and two LTC samples(Sample 1 of ALDH- H1975 lung adenocarcinoma tumor cell(LTC1-m) and Sample 2 of ALDH- H1975 lung adenocarcinoma tumor cell(LTC2-m)) were collected as described in previous study. MiRNAs raw data were obtained by

Illumina Hiseq 2500/2000 platform and 50bp single-end reads were generated. Sequencing reads of small RNA tags(18~30nt) extracted from raw reads were 16.62Mb(LTC1-m), 17.93Mb(LTC2-m), 15.58Mb(LSC1-m), 11.82Mb(LSC2-m), respectively. Correlation-ship between those four samples was shown in Figure 1b. MiRNAs expression landscape analysis on sequencing reads was performed to dissect ncRNAs distribution, predict novel miRNAs, known miRNAs, and DE miRNAs. Of those, at least 91.61% of the total small RNA tags were perfectly mapped to the Hg38.94 genome in four samples. Intriguingly, more than 55.74% of those small RNA tags in LTC2 were mapped to known miRNAs, and only 0.01% of those small RNA tags were predicted as novel miRNAs. On the contrary, the percentage of known miRNAs in LSC2 was only 32.5%, and the percentage of novel miRNAs in it was 0.02%. The rest of those small RNA tags were mapped to rRNA, tRNA, snoRNA, exon ,and so on(**Additional file:Table S1**), separately. The most abundant class in small RNA tags was known miRNAs, followed by other(Figure 1a). To sum up, 100 novel miRNAs, 1379 known miRNAs were obtained(Figure 1c). The base bias on the first position(based on 22nt) and all positions of identified miRNAs(18~30nt) were similar in novel miRNAs and known miRNAs(Figure 1d; Figure S1). The predicted novel miRNAs and known miRNAs(Figure 1e) were used for subsequent analysis. The overview of this research was shown in Figure 1 and Supplementary file: Figure S1.

Total 954 common miRNAs were obtained in two groups(LSCs contained LSC1-m/LSC2-m, and LTCs group consisted of LTC1-m/LTC2-m).208 specific miRNAs were expressed in the LSCs group, and 196 specific miRNAs were expressed in the LTCs group(Figure 1f). Of those, 91 up-regulated expressed miRNAs and 61 down-regulated expressed miRNAs were identified (Figure 1g).
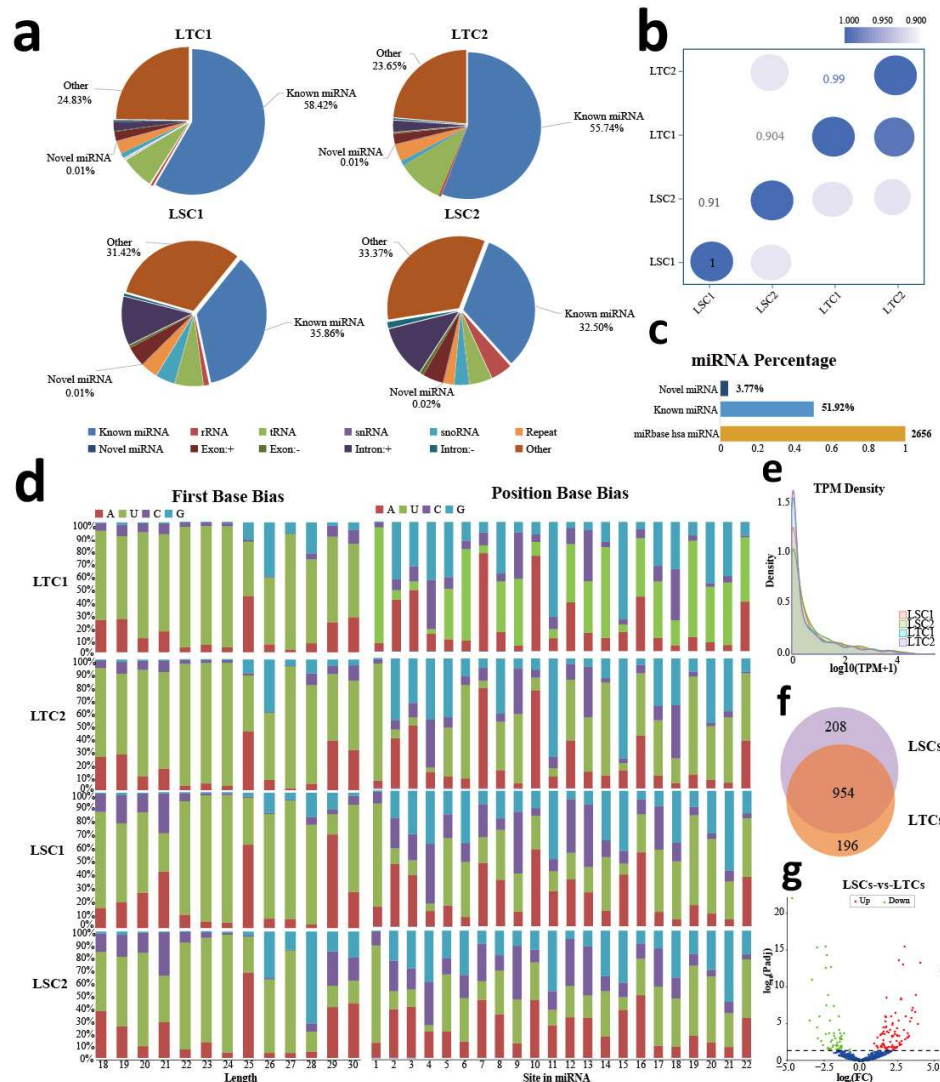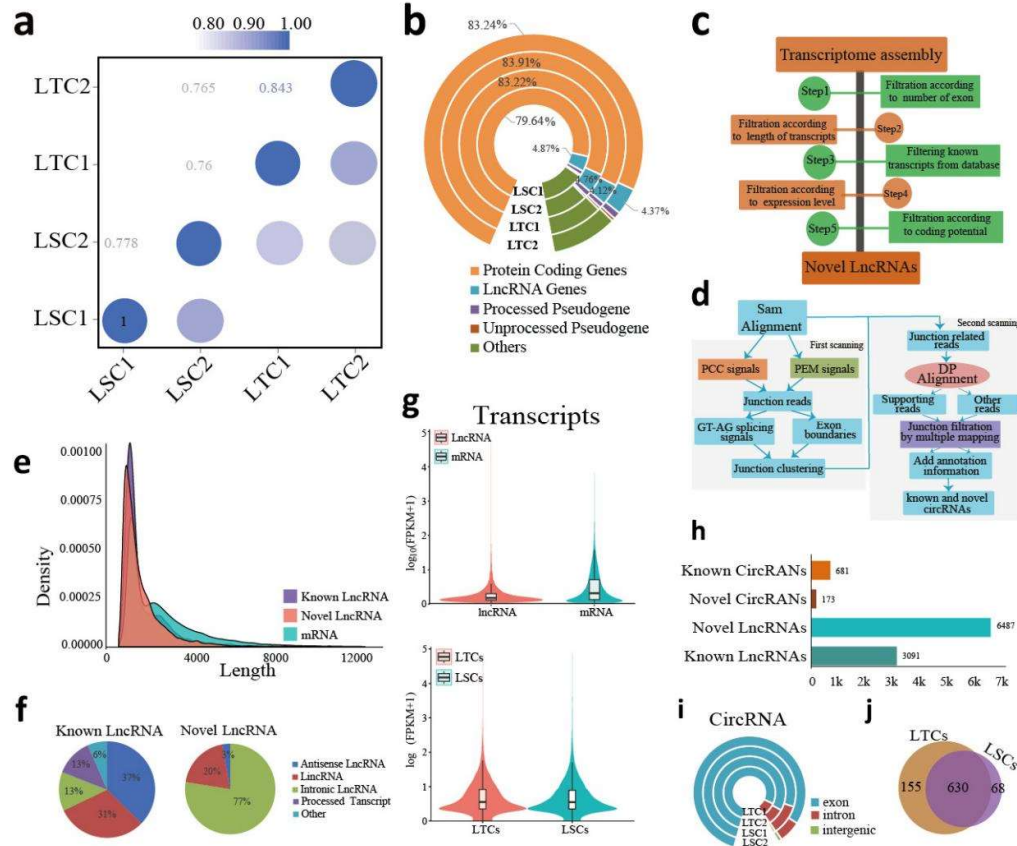
Figure 1. Distribution of novel and known miRNA and DE miRNAs. (a) small RNA(sRNA) category ratio;(b) Samples correlation-ship; (c)MiRNA percentage on miRbase known miRNAs database; (d)First base bias and position base bias on known miRNA; (e)TMP density of all miRNA; (f)Venn diagram about predicted miRNA in LSCs and LTCs; (g) volcano-plot of DE miRNAs

3.2. Identification of novel and known lncRNAs/circRNAs and DE lncRNAs/circRNAs, mRNA in LSCs and LTCs

LncRNAs/circRNAs/mRNAs raw data were fetched from our previous studies. Those reads were sequenced by the Illumina Hiseq2500 platform and were 125bp pair-end reads. The correlation-ship between those four samples is shown in Figure 2a. We obtained approximately 79.3Mb pair-end reads, 93.45% of which could be matched to the Hg38.94 genome. Subsequently, by using HTseq software, we identified known transcripts distribution. Almost 79.64% of those reads were located in protein-coding genes and 4.12% were known lncRNAs(3,091 known lncRNAs)(Figure 2b; Figure 2h). Meanwhile, by using HISAT2 and StringTie, we assembled 188,945 new transcripts. According to those assembled transcripts and known transcripts, we then identified total 6487 novel lncRNAs by using a specific bioinformatics pipeline A(Figure 2c; Figure 2h; method 2.4). We found that the abundant size class of known lncRNAs was antisense lncRNAs(37%) and lincRNAs(31%), and that of novel lncRNAs was intronic lncRNAs(77%)(Figure 2f). Besides, from those short reads, we also identified 681 known circRNAs and 173 novel circRNAs(Figure 2h) by using in silico pipeline B(Figure 2d; method 2.4). Most of those

circRNAs were located in the exon(Figure 2i). Of those, 630 circRNAs were found(Figure 2j) at the same time in two groups. 155 circRNAs were only expressed in LTCs, and 68 circRNAs were only expressed in LSCs.

Figure 2. Distribution of novel and known lncRNAs/circRNAs and DE lncRNAs/circRNAs, mRNAs. (a) Samples correlation-ship;(b) Transcripts compartment;(c)In silico pipeline on the identification of novel lncRNAs; (d)Bioinformatics pipeline on the identification of circRNAs; (e)FPKM density of all transcripts; (f)Constitutions of known and novel lncRNAs ; (g) violin plot of Transcripts between lncRNA and mRNA, and group LTCs and group LSCs; (h)Count of lncRNA and circRNA; (i)Composition of circRNAs; (j)Venn about predicted circRNAs in LTCs and LSCs , respectively.

3.3. DE analysis and Annotation analysis

We used Cufflinks software to perform quantitative analysis on mRNAs/lncRNAs(Figure 2e; Figure 2g) and used Cuffdiff to perform DE analysis on lncRNAs and mRNAs, and finally obtained 16 up-regulated lncRNAs, 18 down-regulated lncRNAs, 378 up-regulated mRNAs, and 364 down-regulated mRNAs(Padj<0.05 and |log2(Fold Change)|>1.3). Meanwhile, We performed DE analysis on circRNAs by using DESeq2 software and eventually obtained 55 up-regulated circRNAs and 61 down-regulated circRNAs(Padj<0.05 and |log2(Fold Change)|>1.3).

We performed the Kyoto encyclopedia of genes and genomes(KEGG) analysis to excavate the potential molecular pathways that regulate the trait of LSCs and LTCs. The up-regulated gene sets and down-regulated gene sets were analyzed, separately. The results of KEGG analysis demonstrated the common pathways(Figure 3a) in LTCs and LSCs, which unveiled the commonality between LSCs and LTCs. KEGG analysis on up-regulated gene sets(Figure 3c) shown some pathways, such as Steroid biosynthesis, Fatty acid metabolism, Antigen processing and presentation, Fatty acid degradation, Inflammatory bowel disease(IBD), Biosynthesis of antibiotics, HTLV-I infection[62], Drug metabolism-other enzymes, which may be relative with immune evasion[63], poor prognosis[62] on

LUAD. The other pathways(Figure 3b), which were enriched by down-regulated gene sets: Cell cycle, Osteoclast differentiation, Progesterone-mediated oocyte maturation, Hippo signaling pathway, may be involved in cancer proliferation, cancer cell heterogeneity[64].

With screening criteria of P-value < 0.01, we then further conducted GO analysis on up-regulated DE mRNAs and down-regulated DE mRNAs, separately. We then conducted hierarchical clustering on GO analysis. Nine clusters of GO terms based on down-regulated gene sets, twelve clusters of GO terms based on up-regulated gene sets were found(Figure 3d: Figure 3e). Nine clusters included cell fate commitment, positive regulation of programmed cell death, intrinsic apoptotic signaling pathway, cell junction organization, actin cytoskeleton organization, regulation of cell-substrate adhesion[65], positive regulation of cell migration, regulation of I-kappaB kinase/NF-kappaB signaling[66], which may play a vital role in cell migration, cell apoptosis, cell communication in most previous studies[67]. Clusters of GO term based on up-regulated gene sets were enriched in mitotic cell cycle, DNA replication initiation, telomere maintenance, regulation of cell morphogenesis, homeostasis of number of cells, cell morphogenesis involved in differentiation, cell junction organization, positive regulation of alpha-beta T cell activation, steroid biosynthetic process, which may be linked to cancer growth, cancer proliferation, metastasis, immune evasion, etc.
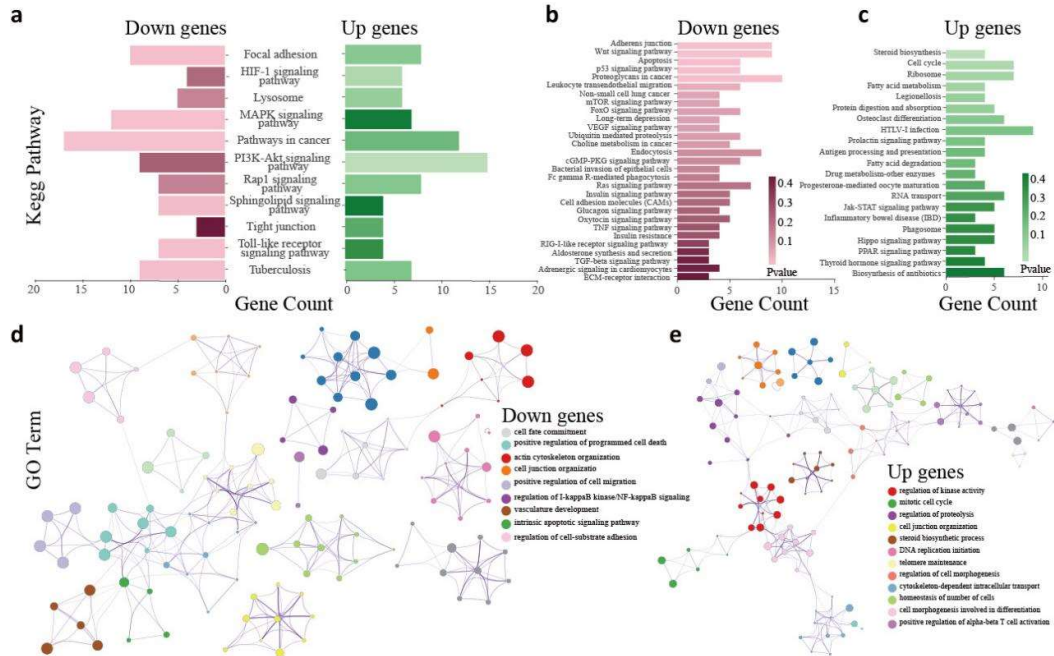

Figure 3. KEGG pathway analysis and GO cluster analysis. (a) Common pathway found in both up-regulated gene sets and down-regulated gene sets;(b)Specific Kegg pathway in down-regulated gene sets ; (c) Specific KEGG pathway in up-regulated gene sets; (d)clusters of GO terms in down-regulated gene sets; (e) clusters of GO terms in down-regulated gene sets.

3.4. Target prediction and Integrative analysis of competitive endogenous RNA(ceRNA) networks

Of the 742 DE mRNAs, 152 DE miRNAs, we predicted the target mRNAs of miRNAs by using miRana, PITA, RNAhybrid, separately. The overlap results in that three software were our final results. Of the 32 DE lncRNAs/115 DE circRNAs, 152 DE miRNAs, we researched the target lncRNAs/circRNAs of miRNAs by using miRanda. We further searched the target mRNAs of lncRNAs by scanning coding genes 10k/100k upstream and downstream of lncRNAs, and obtained 12 lncRNA-mRNA pairs. Finally, we built ceRNA networks, which consisted of 90 DE mRNAs, 143 DE miRNAs, 128 lncRNA/circRNAs. Interestingly, according to produce the ceRNA networks by using Cytoscape, we found

that the network could be divided into two subpopulations(Figure 4). Those two sub-networks matched the ceRNA theory. That is, one sub-network was made up of down-regulated miRNAs, up-regulated mRNAs, and lncRNA/circRNAs, the other one consisted of up-regulated miRNAs, down-regulated mRNAs, and lncRNA/circRNAs.
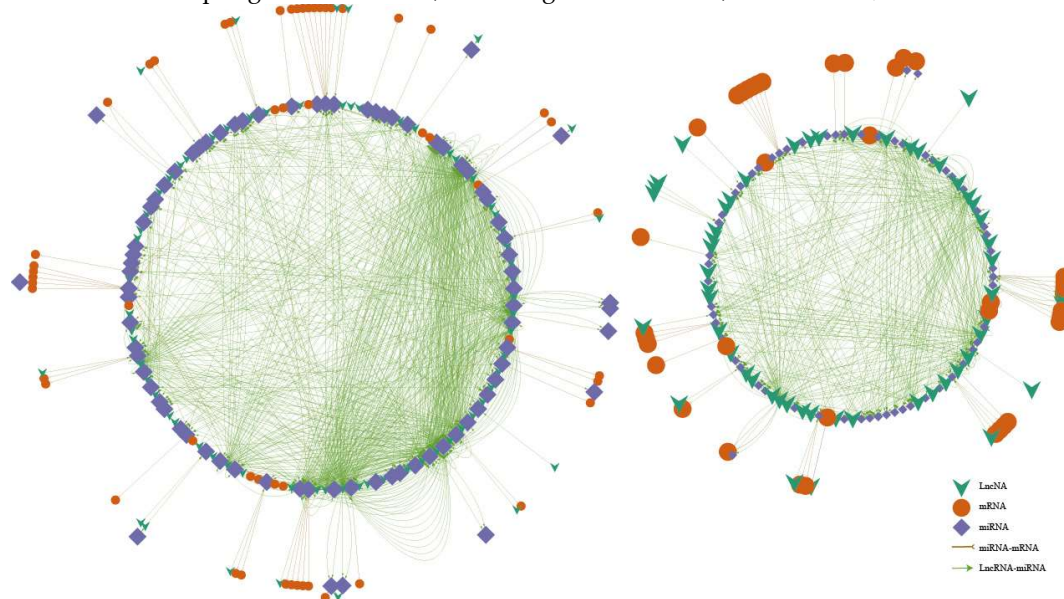

Figure 4. CeRNA networks

To reveal different pathways according to DE mRNAs in ceRNA network between LSCs and LTs, we performed GO analysis. GO analysis on 33 up-regulated gene sets and 54 down-regulated gene sets was respectively analyzed. The results were shown in Figure 5a and Figure 5b.The GO terms of down-regulated gene sets clustered in cell junction, localization of cell, cell motility, cell migration, locomotion, intracellular signal transduction, endocytosis, extracellular matrix may be linked to cell communication, tumor microenvironment, exosome, and cancer metastasis[68], which were similar with that of upregulated gene sets enriched in cell adhesion, positive regulation of signaling transduction, vesicle-mediated transport, lysosome, regulation of secretion, cell migration, locomotion, localization of cell, cell motility etc..That is, communication between cells, such as LSCs-LSCs, LTCs-LTCs, and LSCs-LTCs, may depend on exosome secreted by neighboring cells, which ultimately leads to the migration and proliferation of cancer cells. Those GO terms of up-regulated gene sets, such as fatty acid binding, positive regulation of defense response, Ras protein signal transduction, defense response, leukocyte aggregation, innate immune response, immune response-regulating signaling pathway, cell surface receptor signaling pathway, probably affected the immune resistance of LSCs, and eventually leaded to dismal prognosis. The rest GO terms on up-regulated gene sets were cell growth, cell differentiation, cell junction assembly, cell morphogenesis, cell proliferation, which were related to the stemness of LSCs[69].

We then identified the hub modules in ceRNA networks by using the Cytoscape plug-in: MCODE with default parameters. We obtained four main modules, which were shown in Figure 5c.
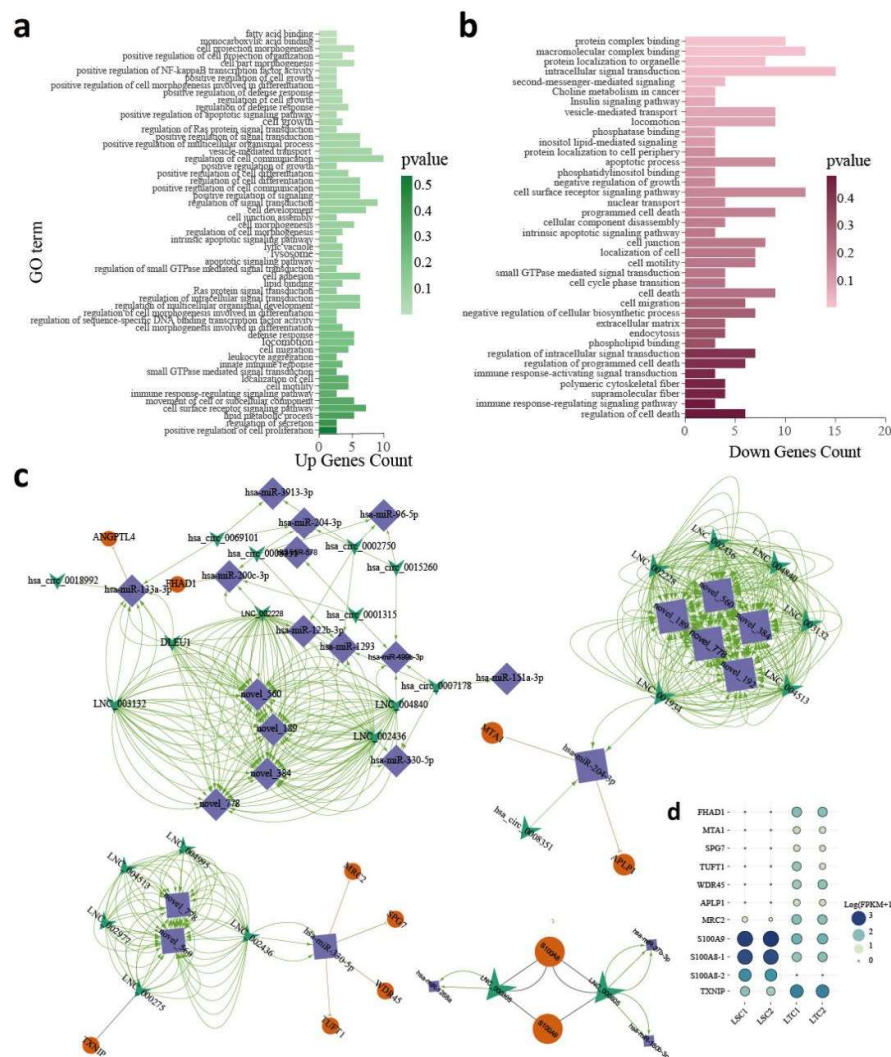
Figure 5. GO analysis and MCODE analysis. (a) GO analysis on up-regulated gene sets on ceRNA networks;(b) GO analysis on down-regulated gene sets on ceRNA networks.; (c)Four main sub-networks; (d)heatmap on 11 transcripts in four sub-networks.

We presented the clustering heatmap about all 11 mRNAs in four modules (Figure 5d). Among those genes, we found that two S100A8 transcript and one S100A9 transcript were significantly up-regulated, and they were regulated by two same novel lncRNAs(LNC_000305/LNC_000605). Those three mRNAs were validated to be associated with several regulatory functions, such as regulation of defense response, regulation of NF-kappaB TF activity, toll-like receptor signaling pathway.In addition, S100A9 was validated to be correlated with a poor prognosis in LUAD[10]. According to survival analysis from UALCAN website and lnCAR website, we demonstrated the survival curve of one novel lncRNAs and one mRNA, LNC_00236 and TUFT1, which were belonged to one module.(Figure 5c;Figure 6A;Figure 6b;Figure 6c).Those three curves shown that the expression levels of TUFT1 and LNC_00236 were associated with the overall survivability of LUAD patients. Plus, LNC_00236 was a high expression in 40 LUAD patients. It could be a potential prognostic biomarker. That is, the other novel lncRNAs could be a potential resource to research better prognostic biomarkers in the future (Figure 6d).
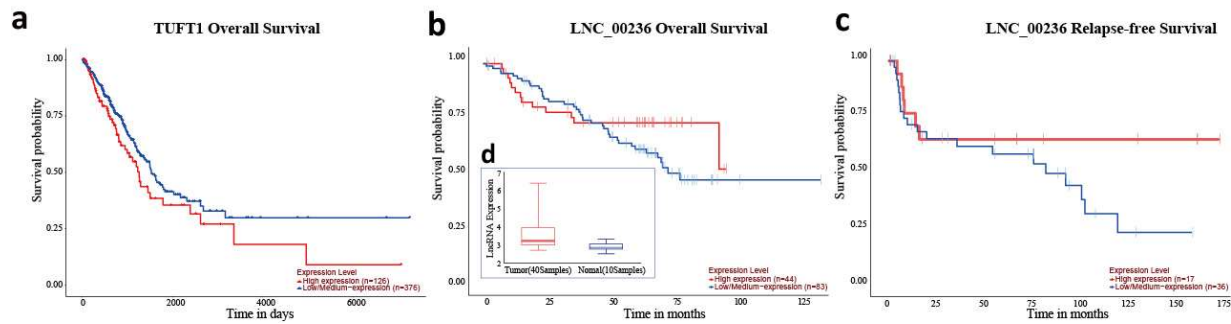
Figure 6. Survival analysis (a)Survival analysis on TUFT1; (b)Survival analysis on LNC_00236; (c) Relapse-free Survival analysis on LNC_00236; (d)LNC_00236 expression atlas in tumor and normal tissue.

4. Conclusions

To sum up, our study identified novel miRNAs, lncRNAs, and circRNAs, which were the basic resource to reveal the important molecular mechanism and hallmarks between LSCs and LTCs during tumor progression, recurrence, metastasis and poor prognosis. In the following, our study provided the detailed dissection of the constituents and functional properties of differentially expressed miRNAs, mRNAs, lncRNAs, and circRNAs between ALDH+ H1975 LSCs and ALDH- H1975 LTCs. Based on these data, an interaction network of the differentially expressed miRNAs, mRNAs, and lncRNAs/circRNAs was constructed to investigate a potential novel regulatory mechanism. The findings of this meta-analysis provided exhaustive and systematic information into underlying potential mechanisms and hallmarks between LSCs and LTCs, respectively. It also would aid further experimental validation which exhibits important roles in the prognosis analysis and therapeutic biomarkers in lung adenocarcinoma.

Appendix A

ncRNAs:Non-Coding RNAs
CSCs: Cancer Stem Cells
miRNAs:Micro RNAs
lncRNAs:Long Non-Coding RNAs
circRNAs:Circular RNAs

mRNAs:Messenger RNAs
LSCs:Lung Adenocarcinoma ALDH+ H1975 Stem Cells
LTCs:Lung Adenocarcinoma ALDH- H1975 Tumor Cells
LUAD:Lung Adenocarcinoma
ceRNA:Competitive Endogenous RNA
NSCLC:Non-Small Cell Lung Cancer
UTR:Untranslated Region
TMP:Transcripts Per Million
FPKM:Fragments Per Kilobase of exon model per Million mapped fragments
PCC:Pair Chiastic Clipping
PEM:Pair End Mapping
DP:Dynamic Programming Alignment
KEGG:Kyoto Encyclopedia of Gene and Genome
GO:Gene Ontology
Padj:Adjusted P-value

Appendix B(ORCID identifiers)
0000-0003-2138-5563 (Wei Yang)
0000-0002-3702-3113(Yong Liang)
0000-0003-3671-7786 (Haitao Luo)
0000-0001-6255-5181 (Xiaofei Yang)
0000-0002-0606-8861 (Furong Li)

Uncategorized References

1.   Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **2021**, *71*, 209-249.

2.   Ferlay, J.; Colombet, M.; Soerjomataram, I.; Parkin, D.M.; Piñeros, M.; Znaor, A.; Bray, F. Cancer statistics for the year 2020: An overview. *International Journal of Cancer* **2021**.

3.   Tang, X.; Cheng, L.; Li, G.; Yan, Y.-M.; Su, F.; Huang, D.-L.; Zhang, S.; Liu, Z.; Qian, M.; Li, J. A small-molecule compound D6 overcomes EGFR-T790M-mediated resistance in non-small cell lung cancer. *Communications biology* **2021**, *4*, 1-13.

4.   de Bruin, E.C.; McGranahan, N.; Mitter, R.; Salm, M.; Wedge, D.C.; Yates, L.; Jamal-Hanjani, M.; Shafi, S.; Murugaesu, N.; Rowan, A.J. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **2014**, *346*, 251-256.

5.   Nassar, D.; Blanpain, C. Cancer Stem Cells: Basic Concepts and Therapeutic Implications. *Annu Rev Pathol* **2016**, *11*, 47-76, doi:10.1146/annurev-pathol-012615-044438.

6.   Morrison, R.; Schleicher, S.M.; Sun, Y.; Niermann, K.J.; Kim, S.; Spratt, D.E.; Chung, C.H.; Lu, B. Targeting the mechanisms of resistance to chemotherapy and radiotherapy with the cancer stem cell hypothesis. *J Oncol* **2011**, *2011*, 941876, doi:10.1155/2011/941876.

7.   Chen, W.-J.; Ho, C.-C.; Chang, Y.-L.; Chen, H.-Y.; Lin, C.-A.; Ling, T.-Y.; Yu, S.-L.; Yuan, S.-S.; Chen, Y.-J.L.; Lin, C.-Y. Cancer-associated fibroblasts regulate the plasticity of lung cancer stemness via paracrine signalling. *Nature communications* **2014**, *5*, 1-17.

8.   Mitra, S.; Stemke-Hale, K.; Mills, G.B.; Claerhout, S. Interactions between tumor cells and microenvironment in breast cancer: a new opportunity for targeted therapy. *Cancer science* **2012**, *103*, 400-407.

9.   Zhang, W.C.; Shyh-Chang, N.; Yang, H.; Rai, A.; Umashankar, S.; Ma, S.; Soh, B.S.; Sun, L.L.; Tai, B.C.; Nga, M.E. Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell* **2012**, *148*, 259-272.

10.   Luo, H.-T.; Zheng, Y.-Y.; Tang, J.; Shao, L.-J.; Mao, Y.-H.; Yang, W.; Yang, X.-F.; Li, Y.; Tian, R.-J.; Li, F.-R. Dissecting the multi-omics atlas of the exosomes released by human lung adenocarcinoma stem-like cells. *NPJ Genomic Medicine* **2021**, *6*, 1-14.

11.   Ginestier, C.; Hur, M.H.; Charafe-Jauffret, E.; Monville, F.; Dutcher, J.; Brown, M.; Jacquemier, J.; Viens, P.; Kleer, C.G.; Liu, S. ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell stem cell* **2007**, *1*, 555-567.

12.   Reya, T.; Morrison, S.J.; Clarke, M.F.; Weissman, I.L. Stem cells, cancer, and cancer stem cells. *nature* **2001**, *414*, 105-111.

13.   Eramo, A.; Haas, T.L.; De Maria, R. Lung cancer stem cells: tools and targets to fight lung cancer. *Oncogene* **2010**, *29*, 4625-4635, doi:10.1038/onc.2010.207.

14.   He, X.; Qi, Y.; Zhang, X.; Liu, X.; Li, X.; Li, S.; Wu, Y.; Zhang, Q. Current landscape of tumor-derived exosomal ncRNAs in glioma progression, detection, and drug resistance. *Cell death & disease* **2021**, *12*, 1-11.

15.   Dalmay, T.; Edwards, D. MicroRNAs and the hallmarks of cancer. *Oncogene* **2006**, *25*, 6170-6175.

16.   Correia de Sousa, M.; Gjorgjieva, M.; Dolicka, D.; Sobolewski, C.; Foti, M. Deciphering miRNAs' Action through miRNA Editing. *Int J Mol Sci* **2019**, *20*, doi:10.3390/ijms20246249.

17.   Esquela-Kerscher, A.; Slack, F.J. Oncomirs—microRNAs with a role in cancer. *Nature reviews cancer* **2006**, *6*, 259-269.

18.   Tai, M.C.; Yanagisawa, K.; Nakatochi, M.; Hotta, N.; Hosono, Y.; Kawaguchi, K.; Naito, M.; Taniguchi, H.; Wakai, K.; Yokoi, K. Blood-borne miRNA profile-based diagnostic classifier for lung adenocarcinoma. *Scientific reports* **2016**, *6*, 1-9.

19.   Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell* **2004**, *116*, 281-297.

20.   Hayashita, Y.; Osada, H.; Tatematsu, Y.; Yamada, H.; Yanagisawa, K.; Tomida, S.; Yatabe, Y.; Kawahara, K.; Sekido, Y.; Takahashi, T. A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer research* **2005**, *65*, 9628-9632.

21.   Yanaihara, N.; Caplen, N.; Bowman, E.; Seike, M.; Kumamoto, K.; Yi, M.; Stephens, R.M.; Okamoto, A.; Yokota, J.; Tanaka, T. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer cell* **2006**, *9*, 189-198.

22.   Kopp, F.; Mendell, J.T. Functional classification and experimental dissection of long noncoding RNAs. *Cell* **2018**, *172*, 393-407.

23.   Nie, J.-H.; Li, T.-X.; Zhang, X.-Q.; Liu, J. Roles of non-coding RNAs in normal human brain development, brain tumor, and neuropsychiatric disorders. *Non-coding RNA* **2019**, *5*, 36.

24.   Deng, X.; Xiong, W.; Jiang, X.; Zhang, S.; Li, Z.; Zhou, Y.; Xiang, B.; Zhou, M.; Li, X.; Li, G. LncRNA LINC00472 regulates cell stiffness and inhibits the migration and invasion of lung adenocarcinoma by binding to YBX1. *Cell death & disease* **2020**, *11*, 1-13.

25.   Wu, P.; Mo, Y.; Peng, M.; Tang, T.; Zhong, Y.; Deng, X.; Xiong, F.; Guo, C.; Wu, X.; Li, Y. Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Molecular cancer* **2020**, *19*, 1-14.

26.   Li, X.; Yang, L.; Chen, L.-L. The biogenesis, functions, and challenges of circular RNAs. *Molecular cell* **2018**, *71*, 428-442.

27.   Wang, Y.; Liu, J.; Ma, J.; Sun, T.; Zhou, Q.; Wang, W.; Wang, G.; Wu, P.; Wang, H.; Jiang, L. Exosomal circRNAs: biogenesis, effect and application in human diseases. *Molecular cancer* **2019**, *18*, 1-10.

28.   Wang, J.; Zhao, X.; Wang, Y.; Ren, F.; Sun, D.; Yan, Y.; Kong, X.; Bu, J.; Liu, M.; Xu, S. circRNA-002178 act as a ceRNA to promote PDL1/PD1 expression in lung adenocarcinoma. *Cell death & disease* **2020**, *11*, 1-11.

29.   Chen, W.; Li, W.; Liu, Z.; Ma, G.; Deng, Y.; Li, X.; Wang, Z. Comprehensive analysis of competitive endogenous RNA associated with immune infiltration in lung adenocarcinoma. *Scientific Reports* **2021**, *11*, 1-14.

30.   Wang, K.; Long, B.; Liu, F.; Wang, J.-X.; Liu, C.-Y.; Zhao, B.; Zhou, L.-Y.; Sun, T.; Wang, M.; Yu, T. A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223. *European heart journal* **2016**, *37*, 2602-2611.

31.   Wang, K.; Liu, F.; Zhou, L.-Y.; Long, B.; Yuan, S.-M.; Wang, Y.; Liu, C.-Y.; Sun, T.; Zhang, X.-J.; Li, P.-F. The long noncoding RNA CHRF regulates cardiac hypertrophy by targeting miR-489. *Circulation research* **2014**, *114*, 1377-1388.

32.   Wang, K.; Liu, C.-Y.; Zhou, L.-Y.; Wang, J.-X.; Wang, M.; Zhao, B.; Zhao, W.-K.; Xu, S.-J.; Fan, L.-H.; Zhang, X.-J. APF lncRNA regulates autophagy and myocardial infarction by targeting miR-188-3p. *Nature communications* **2015**, *6*, 1-11.

33.   Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **2009**, *10*, 1-10.

34.   Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **2019**, *37*, 907-915.

35.   He, W.; Zhao, S.; Liu, X.; Dong, S.; Lv, J.; Liu, D.; Wang, J.; Meng, Z. ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis. *Genet Mol Res* **2013**, *12*, 6275-6283.

36.   Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.

37.   Kozomara, A.; Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **2014**, *42*, D68-D73.

38.   Friedländer, M.R.; Mackowiak, S.D.; Li, N.; Chen, W.; Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research* **2012**, *40*, 37-52.

39.   Saha, S.; Bridges, S.; Magbanua, Z.V.; Peterson, D.G. Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biology* **2008**, *1*, 85-96.

40.   Kalvari, I.; Nawrocki, E.P.; Ontiveros-Palacios, N.; Argasinska, J.; Lamkiewicz, K.; Marz, M.; Griffiths-Jones, S.; Toffano-Nioche, C.; Gautheret, D.; Weinberg, Z. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **2021**, *49*, D192-D200.

41.   Saha, S.; Bridges, S.; Magbanua, Z.V.; Peterson, D.G. Empirical comparison of ab initio repeat finding programs. *Nucleic acids research* **2008**, *36*, 2284-2294.

42.   Wen, M.; Shen, Y.; Shi, S.; Tang, T. miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. *BMC bioinformatics* **2012**, *13*, 1-10.

43.   Wagner, G.P.; Kin, K.; Lynch, V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences* **2012**, *131*, 281-285.

44.   Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* **2016**, *11*, 1650-1667.

45.   Sun, L.; Luo, H.; Bu, D.; Zhao, G.; Yu, K.; Zhang, C.; Liu, Y.; Chen, R.; Zhao, Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research* **2013**, *41*, e166-e166.

46.   Kang, Y.-J.; Yang, D.-C.; Kong, L.; Hou, M.; Meng, Y.-Q.; Wei, L.; Gao, G. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research* **2017**, *45*, W12-W16.

47. Punta, M.; Coggill, P.C.; Eberhardt, R.Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J. The Pfam protein families database. *Nucleic acids research* **2012**, *40*, D290-D301.

48. Lin, M.F.; Jungreis, I.; Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **2011**, *27*, i275-i282.

49. Love, M.I.; Hogenesch, J.B.; Irizarry, R.A. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology* **2016**, *34*, 1287-1291.

50. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; Van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **2010**, *28*, 511-515.

51. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Nature Precedings* **2010**, 1-1.

52. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **2003**, *31*, 2013-2035.

53. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **2009**, *37*, 1-13.

54. Zhou, Y.; Zhou, B.; Pache, L.; Chang, M.; Khodabakhshi, A.H.; Tanaseichuk, O.; Benner, C.; Chanda, S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications* **2019**, *10*, 1-10.

55. Enright, A.; John, B.; Gaul, U.; Tuschl, T.; Sander, C.; Marks, D. MicroRNA targets in Drosophila. *Genome biology* **2003**, *4*, 1-27.

56. Kertesz, M.; Iovino, N.; Unnerstall, U.; Gaul, U.; Segal, E. The role of site accessibility in microRNA target recognition. *Nature genetics* **2007**, *39*, 1278-1284.

57. Krüger, J.; Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research* **2006**, *34*, W451-W454.

58. 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 research* **2003**, *13*, 2498-2504.

59. Bader, G.D.; Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics* **2003**, *4*, 1-27.

60. Chandrashekar, D.S.; Bashel, B.; Balasubramanya, S.A.H.; Creighton, C.J.; Ponce-Rodriguez, I.; Chakravarthi, B.V.; Varambally, S. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* **2017**, *19*, 649-658.

61. Zheng, Y.; Xu, Q.; Liu, M.; Hu, H.; Xie, Y.; Zuo, Z.; Ren, J. lnCAR: a comprehensive resource for lncRNAs from cancer arrays. *Cancer research* **2019**, *79*, 2076-2083.

62. Manns, A.; Wilks, R.J.; Murphy, E.L.; Haynes, G.; Figueroa, J.P.; Barnett, M.; Hanchard, B.; Blattner, W.A. A prospective study of transmission by transfusion of HTLV-I and risk factors associated with seroconversion. *International journal of cancer* **1992**, *51*, 886-891.

63. Mahata, B.; Pramanik, J.; van der Weyden, L.; Polanski, K.; Kar, G.; Riedel, A.; Chen, X.; Fonseca, N.A.; Kundu, K.; Campos, L.S. Tumors induce de novo steroid biosynthesis in T cells to evade immunity. *Nature communications* **2020**, *11*, 1-15.

64. Huang, J.; Wu, S.; Barrera, J.; Matthews, K.; Pan, D. The Hippo signaling pathway coordinately regulates cell proliferation and apoptosis by inactivating Yorkie, the Drosophila Homolog of YAP. *Cell* **2005**, *122*, 421-434.

65. Knights, A.J.; Funnell, A.P.; Crossley, M.; Pearson, R.C. Holding tight: cell junctions and cancer spread. *Trends in cancer research* **2012**, *8*, 61.

66. Staudt, L.M. Oncogenic activation of NF-kappaB. *Cold Spring Harb Perspect Biol* **2010**, *2*, a000109, doi:10.1101/cshperspect.a000109.

67. Adli, M.; Baldwin, A.S. IKK-i/IKKε controls constitutive, cancer cell-associated NF-κB activity via regulation of Ser-536 p65/RelA phosphorylation. *Journal of Biological Chemistry* **2006**, *281*, 26976-26984.

68. El-Sabban, M.; Pauli, B. Role of cell adhesion molecules in endothelial cell/tumor cell communication and tumor cell extravasation. In *Progress in Cell Research*; Elsevier: 1995; Volume 4, pp. 87-92.

69. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *cell* **2011**, *144*, 646-674.