

## Supplementary Methods

### Manual curation of high-confidence experimentally verified ceRNAs

We performed a manual curation of experimentally verified ceRNA interactions from published articles. In this update, we retrieved published literature from the PubMed database by employing a combination of key words including '(lncrna OR lincrna OR long noncoding) AND (miRNA OR microrna OR mir) AND (sponge OR ceRNA OR competing OR compet OR competence OR mediate OR mediated OR interaction) OR (ceRNA) OR (miRNAs sponge) OR (miRNA decoy) OR (miRNA target lncRNA)' and found more than 15,000 relevant articles (before September 2021). These candidate articles were reviewed by at least two researchers. In this step, the researchers confirmed whether the regulatory relationship between each competing RNA member was verified via one of the high-confidence experimental methods, including qRT-PCR, western blot, luciferase reporter assay, RNAi, in vitro knockdown, RNA immunoprecipitation, or other reliable experiments. We filtered the ceRNA associations based on the following criteria: whether (i) RNAs within a ceRNA interaction act to dynamically regulate the expression of each other, (ii) RNAs within a ceRNA interaction share the same miRNA binding site, and (iii) ceRNA entries that were simultaneously confirmed by at least two researchers were retained. We retrieved more ceRNA types such as lncRNAs, coding mRNAs, pseudogenes, circular RNAs, intruded viral RNAs, and technically engineered RNAs. For each ceRNA interaction, detailed information including members of ceRNAs, species, tissues/cell lines, disease/phenotype, experimental methods, PubMed ID, article title, citation, and a functional description from the original studies were collected. Currently, LncACTdb 3.0 documents 5,667 high-confidence experimentally verified ceRNA interactions. The scope of

LncACTdb 3.0 has expanded to 24 species and 536 related diseases/phenotypes. The full list of experimentally verified ceRNA interactions can be downloaded from [http://bio-bigdata.hrbmu.edu.cn/LncACTdb/Lncact\\_down.jsp](http://bio-bigdata.hrbmu.edu.cn/LncACTdb/Lncact_down.jsp).

### **Manual curation of lncRNA biomarkers**

In LncACTdb 3.0, we performed manual curation of experimentally supported lncRNA biomarkers to provide new insights into tumor diagnosis and therapy. We used the following keyword combinations: 'circulating OR drug-resistant OR prognostic OR immune OR metastasis OR recurrence OR cell growth OR EMT OR apoptosis OR autophagy) AND (lncRNA)' to collect diagnostic and therapeutic biomarkers. A biomarker was selected depending on whether the lncRNA was associated with these processes by overexpression, RNA knockdown, or other functional experiments. Finally, a total of 10,084 experimentally supported lncRNA biomarkers associated with drug resistance, circulation, survival, immunity, metastasis, recurrence, cell growth, EMT, apoptosis, and autophagy were manually curated from the literature and integrated into the LncACTdb 3.0 database.

### **Curation of functional annotations and functional analysis**

LncACTdb 3.0 provides the Function tool to perform functional analysis of lncRNAs based on a 'guilt-by-association' strategy. The downstream mRNAs of lncRNA were used to perform functional enrichment analysis. LncACTdb 3.0 curated thousands of pathways and biological terms as functional backgrounds. For curating biological pathways and Gene Ontology (GO) annotation, a total of 1,329 pathways including KEGG (1), BioCarta (<https://cgap.nci.nih.gov/Pathways>), Reactome (2), PID (3), STKE (<http://stke.sciencemag.org/>), SIG (<http://www.signaling-gateway.org/>), and a total of 5,917

gene sets representing functional GO terms were collected from MSigDB (4). Gene sets of hallmark processes that promote tumor growth and metastasis (5), including the evasion of apoptosis, evasion of immune detection, genome instability and mutation, insensitivity to antigrowth signals, limitless replicative potential, reprogramming of energy metabolism, self-sufficiency in growth signals, sustained angiogenesis, tissue invasion and metastasis, and tumor-promoting inflammation, have been downloaded from our previously published article (6). To study the effects of ceRNA on tumor cell states, we downloaded the characteristic gene sets corresponding to the 14 functional states from CancerSEA, including angiogenesis, apoptosis, cell cycle, differentiation, DNA damage, DNA repair, EMT, hypoxia, inflammation, invasion, metastasis, proliferation, quiescence, and stemness (7). We performed hypergeometric tests to calculate the enrichment significance for target genes of lncRNAs based on different pathways and GO gene sets (Figure S3). The formula for calculating the statistically significant P-value for the enrichment of target genes of a particular lncRNA in a given pathway is as follows:

$$P = 1 - \sum_{t=0}^x \frac{\binom{S}{t} \binom{N-S}{M-t}}{\binom{N}{M}}$$

where N represents the total number of genes in the functional context (the total number of genes contained in all collected functional gene sets), S represents all genes contained in the functional pathway, M represents the number of all target genes of the lncRNA, and x indicates the number of intersections between target genes and pathway genes. Significantly enriched functions were defined using a threshold of  $P < 0.05$  and further illustrated as bar graphs based on  $-\log_{10}$  transformed P values. For individual function analysis, the R package GSEA was used to evaluate the functional state of each patient based on cancer hallmarks and cell state

gene sets (8).

### **Collection and processing of high-throughput expression data**

We applied the following selection methods for data collection from TCGA and GEO databases:

(i) The dataset contained prognostic information, including patient alive/dead status and survival times. (ii) The dataset contained a sufficient number of samples (more than 20 samples) for survival analysis. (iii) For a GEO dataset, the normalized expression profile was provided by the GEO Accession page. (iv) For the TCGA dataset, RNA sequencing profiles from the Illumina HiSeq RNASeq V2 platform were collected. Finally, 62 high-throughput expression profiles containing 16,228 patients across 33 cancers were collected from the TCGA and GEO public datasets. We retrieved the already quantified gene expression values from the TCGA and GEO databases. In the GEO dataset, the same normalization method was performed for all samples according to the standard protocol of Affymetrix or Agilent. For example, the expression values of Affymetrix HG-U133\_Plus\_2.0 array were background adjusted using the robust multi-array analysis (RMA) method followed by quantile normalization as implemented in the R/Bioconductor package `frma`. Expression values of the Agilent Homo\_sapiens\_21.6K\_custom array were normalized using the locally weighted scatterplot smoothing (LOWESS) method. The already quantified expression profiles can be accessed from the Series Matrix File(s) at the GEO Accession page. For the TCGA dataset, RNA sequencing profiles from the Illumina HiSeq RNASeqV2 platform were collected. The expression values in different samples were normalized using the same method (normalized as FPKM values). The log<sub>2</sub> transformation was performed for all samples.

## Identification of sample-specific ceRNAs and construction of networks

Initially, candidate ceRNA pairs were collected from two databases: starBase v2.0 (9) and LncACTdb 2.0 (6). LncACTdb 2.0 and starBase v2.0 are comprehensive ceRNA databases with different characteristics. LncACTdb 2.0 identifies ceRNA interactions through miRNA-target prediction methods and expression correlation analysis. The experimentally verified ceRNA relations were also included in the LncACTdb 2.0. StarBase v2.0 identifies ceRNA interactions by analyzing a large set of Ago-binding sites determined from the Ago CLIP-Seq datasets. However, the experimentally verified ceRNAs were not collected using the starBase v2.0. Essentially, starBase v2.0 focuses on the overlap between Ago CLIP-Seq peaks and miRNA binding sites at the sequence level, while LncACTdb 2.0 focuses on the ceRNA correlations on expression level. Considering their different advantages, we trusted that these two databases will complement the pool of candidate ceRNAs. Thus, we collected ceRNAs from these two databases and used the union set as candidates in the LncACTdb 3.0. A total of 108,668 candidate ceRNA regulations were identified. To verify whether these ceRNA pairs were associated with each other in a specific sample, we used a published method for sample-specific network construction based on probability theory to identify ceRNA networks for each sample (Figure S6A) (10). We assumed that each ceRNA pair may have an association in some samples, but not in other samples, due to differences in sample types. We determined whether lncRNAs and mRNAs were related in a specific sample by testing the statistical independence of the candidate ceRNA expression values in the same sample. For a ceRNA pair of  $x(\text{mRNA})$  and  $y(\text{lncRNA})$  in sample  $q$ , we calculated the following statistic:

$$Q_{xy}^q = \frac{\sqrt{n-1} \cdot (n \cdot n_{xy}^q - n_x^q n_y^q)}{\sqrt{n_x^q n_y^q (n - n_x^q) (n - n_y^q)}}$$

where  $n$  is the total number of samples.  $n_x^q$  and  $n_y^q$  are predetermined integers. We set

$n_x^q = n_y^q = 0.1n$ . We draw the first two boxes near  $x_q$  and  $y_q$ , based on the predetermined  $n_x^q$  and  $n_y^q$ . The third box is simply the intersection of the previous two

boxes (Figure S6B). Thus, we can obtain the value of  $n_{xy}^q$  by counting the plots in the third

box. If  $x$  and  $y$  are independent of each other, this statistic follows a standard normal distribution, and the mean value and variance for the  $n$  samples are 0 and 1, respectively.

Therefore, we can determine the significance of the  $x$ ,  $y$  correlation with this statistic.

$edge_{xy}^q$  was set to 1 in the network of sample  $q$  with a false discovery rate (FDR) of  $< 0.05$ .

For network construction in a specific sample, we retained pairs that met the  $FDR < 0.05$ . For any pair of candidate ceRNAs (an lncRNA and an mRNA), we first screened all samples in which the ceRNAs were not statistically independent in expression, and then calculated whether there was a positive relationship between the lncRNA and mRNA in these samples at a broader level. We performed Pearson correlation tests for the expression of lncRNAs and mRNAs in these samples. If the correlation coefficient was positive and the  $p$ -value was less than 0.05, the ceRNA relationships in these samples were retained; otherwise, they were excluded.

The algorithm requires the inclusion of both mRNA and lncRNA expressions in the expression dataset, and is more efficient when the number of samples is greater than 100. This method is not sensitive to normalization methods for gene expression matrices. Any data type such as FPKM, TPM, or count can be used as the algorithm input. In RNA-seq data, the statistic may

approach zero due to experimental errors, which is meaningless in biological terms and may produce errors in the data analysis. Hence, we treat the zeros as follows (10).

(1) If we cannot distinguish whether the zeros result from zero expression or the experimental errors,  $edge_{xy}^q$  is set to 0 when  $x_q = 0$  or  $y_q = 0$  without the consideration of the statistic.

(2) If we know that the zeros resulting from the zero expression,  $edge_{xy}^q$  is determined by the statistic.

### **Identification of cancer-specific ceRNAs**

Based on the above pipeline, we purified ceRNA pairs in a cancer-specific manner. For a candidate lncRNA-miRNA-mRNA relation in a cancer, if the miRNA targeting both RNAs of the pair was expressed in more than 50% of samples, the ceRNA-cancer relationship was retained. The specificity of the ceRNAs was characterized quantitatively by calculating the specificity score using a previously described method (11-13). For a ceRNA, the specificity score was calculated as:

$$Specificity\ score = \frac{\sum_{i=1}^N (1 - x_i)}{N - 1}$$

where  $N$  is the number of cancers and  $x_i$  is the percentage of samples in which the ceRNA can be found in a cancer. The value of  $x_i$  was normalized to the maximum percentage value of cancers. For example, the specificity score for a ceRNA with a percentage profile of '0 0 0.2 0 0 0 0.2 0 0 0.8 0' was calculated to be 0.9545. The range of specificity scores was between 0 and 1, whereas a perfect specific pattern was scored as 1. According to an earlier study (11), a specificity score < 0.15 indicates a housekeeping gene. We used a specificity score > 0.5 as the threshold for purified cancer-specific ceRNAs.

## Survival analysis

LncACTdb 3.0 performs COX survival analysis and displays Kaplan–Meier survival curves for ceRNAs, lncRNAs, miRNAs, and mRNAs. Clinical follow-up information of patients from TCGA and GEO was collected. A univariate Cox regression model was built to evaluate the association between survival and the expression level of each lncRNA/miRNA/mRNA member in a ceRNA interaction. A risk score formula was developed to evaluate the association between survival time, living status, and expression, which considers both the strength and positive/negative association between each gene and the probability of survival. The integrated risk score for each patient was calculated based on the linear combination of ceRNA expression values weighted by the Cox regression coefficients:

$$Risk\ score = \sum_{i=1}^n r_i Exp(i)$$

where  $r_i$  is the COX regression coefficients for lncRNA, miRNA, and mRNA,  $n$  is the number of genes, and  $Exp(i)$  is the expression value of gene in patient  $i$ . We used the median and mean risk scores as cut-offs to classify patients into different risk groups.

## Network construction and illustration

In LncACTdb 3.0, the network tool constructs and illustrates the ceRNA network in the following ways:

(i) **Gene-centric:** For an lncRNA or mRNA, LncACTdb 3.0, provides a comprehensive view of all possible associated ceRNA relationships. A network consisting of this lncRNA or mRNA and its associated competing neighbors was constructed and illustrated using the Java script plugin ECharts (V4.0). The ceRNA network scale can be reset by selecting different steps of



the neighbors. In the one-step-neighbours scale, the top 20 competing mRNA partners (ordered by activity score) of the lncRNA were illustrated. In the two-step- and three-step-neighbors scale, this network expands with another 20 and 40 competing lncRNAs and mRNAs, respectively.

(ii) **Patient-centric:** For a cancer sample, LncACTdb 3.0, provides a patient-specific ceRNA network. In this section, users can input a sample name to obtain a patient-specific ceRNA network. In the network, ceRNA interactions were determined to be specifically active in a sample by testing the statistical independence of the candidate ceRNA expression values. When users move the cursor over a node in the network, all the edges and nodes connected to it are highlighted. Different network layouts, such as the circular and force layouts, can be used to illustrate the network.

(iii) **User-designed:** For user-selected data, LncACTdb 3.0 provides a user-designed ceRNA network in which lncRNAs, miRNAs, mRNAs, and diseases can be determined by users. This network was constructed based on the experimentally validated ceRNA relationships in LncACTdb 3.0. The network degree of ceRNAs is listed in the data table. Users can start a new search for experimentally validated ceRNA relations by clicking a specified gene in the network.

### **BLAST method**

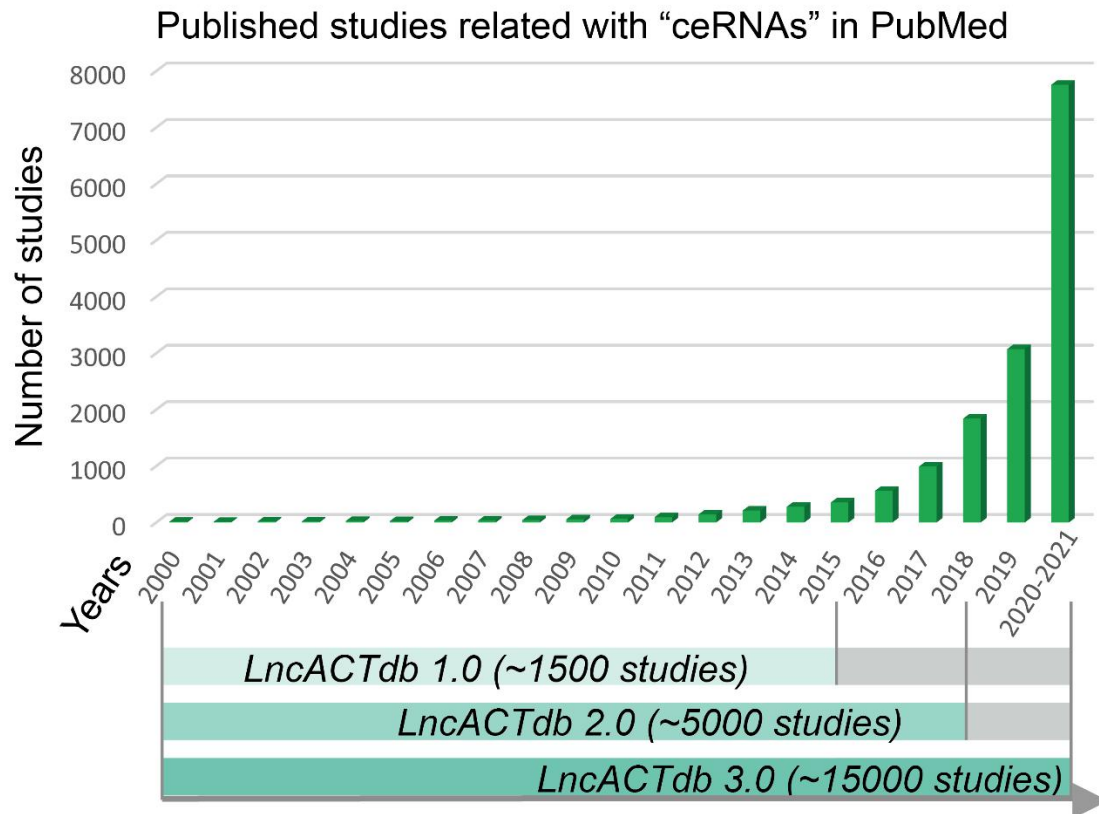
As described in LncACTdb 2.0, the current database also allows users to query datasets via the input of custom sequences. The Basic Local Alignment Search Tool (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>) was integrated into LncACTdb 3.0, to compare the

input sequence with sequences in our database and calculate the statistical significance of the matches. The parameters for running BLAST were set as `-perc_identity 80 -outfmt '7 qacc sacc evaluate length pident'`. To filter the results, we used '>80% identity' as the threshold to identify highly similar sequences.

## REFERENCES











1. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, **40**, D109-114.
2. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic acids research*, **42**, D472-477.
3. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K.H. (2009) PID: the Pathway Interaction Database. *Nucleic acids research*, **37**, D674-679.
4. Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems*, **1**, 417-425.
5. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646-674.
6. Wang, P., Li, X., Gao, Y., Guo, Q., Wang, Y., Fang, Y., Ma, X., Zhi, H., Zhou, D., Shen, W. *et al.* (2019) LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. *Nucleic Acids Res*, **47**, D121-D127.
7. Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., Xu, L., Luo, T., Yan, H., Long, Z. *et al.* (2019) CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res*, **47**, D900-D908.
8. Hänzelmann, S., Castelo, R. and Guinney, J. (2013) GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, **14**, 7.
9. Li, J.H., Liu, S., Zhou, H., Qu, L.H. and Yang, J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*, **42**, D92-97.
10. Dai, H., Li, L., Zeng, T. and Chen, L. (2019) Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res*, **47**, e62.
11. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650-659.
12. Wang, P., Ning, S., Zhang, Y., Li, R., Ye, J., Zhao, Z., Zhi, H., Wang, T., Guo, Z. and Li, X. (2015) Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res*, **43**, 3478-3489.
13. Guo, Q., Wang, J., Gao, Y., Li, X., Hao, Y., Ning, S. and Wang, P. (2020) Dynamic TF-lncRNA Regulatory Networks Revealed Prognostic Signatures in the Development of Ovarian Cancer. *Front Bioeng Biotechnol*, **8**, 460.

## Supplementary Figures

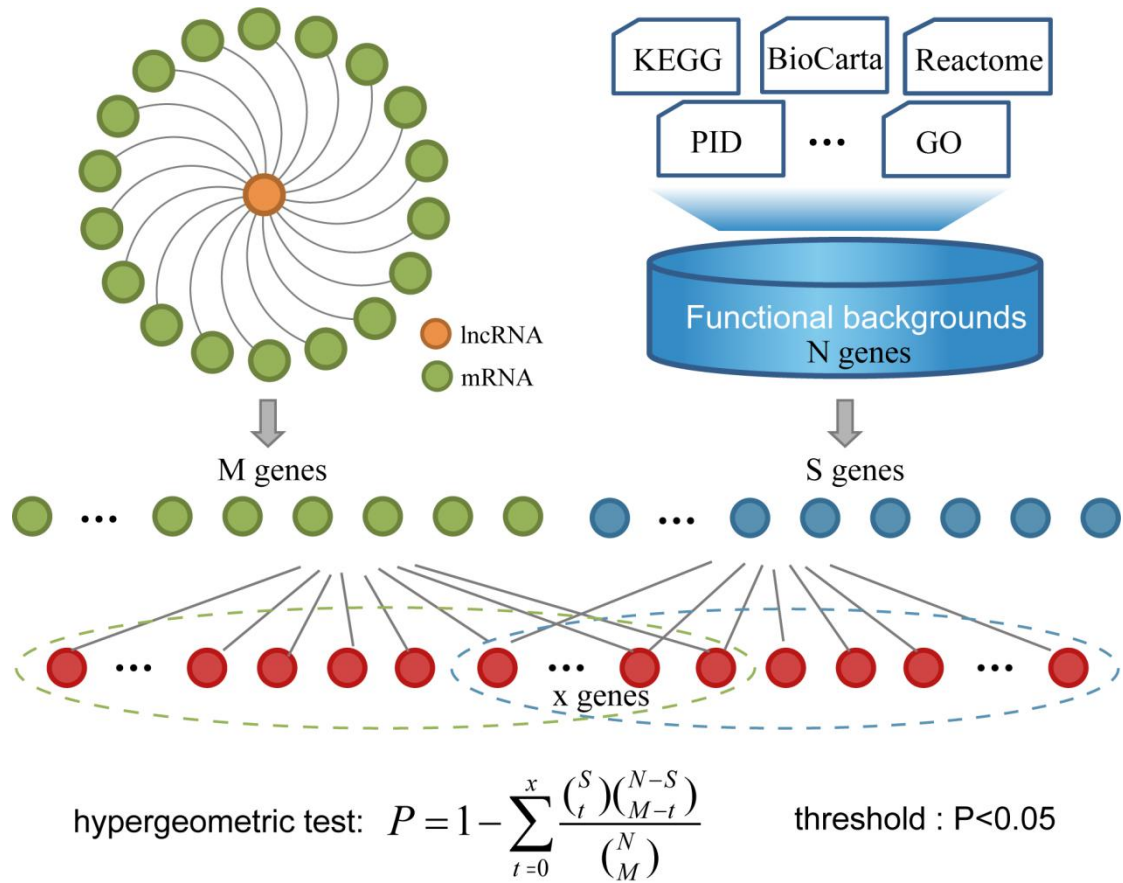


**Figure S1.** The development of LncACTdb database and collection of ceRNA related studies in recent years.

### The biomarker statistics in LncACTdb 3.0

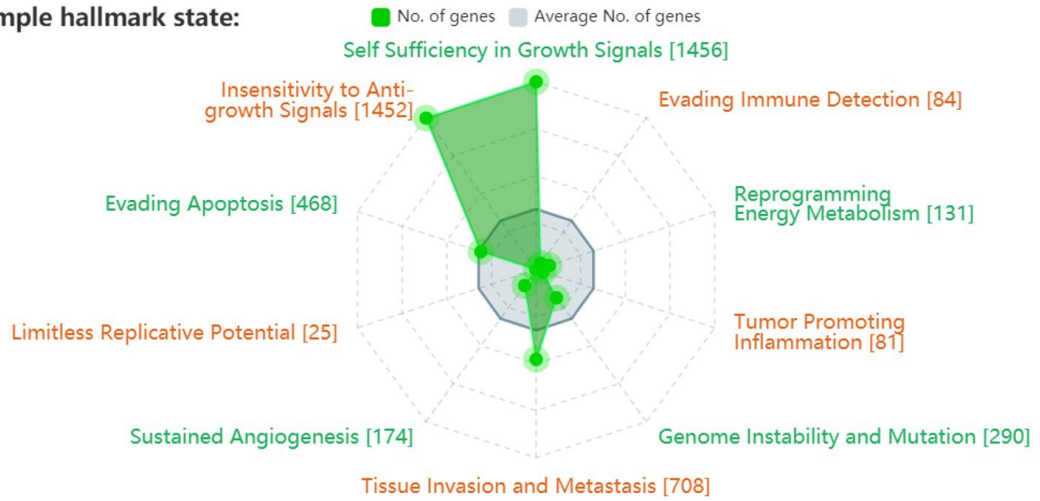
 Biomarker: drug	Counts: 458	<a href="#">Detail</a>	 Biomarker: circulating	Counts: 1012	<a href="#">Detail</a>
 Biomarker: survival	Counts: 4223	<a href="#">Detail</a>	 Biomarker: immune	Counts: 104	<a href="#">Detail</a>
 Biomarker: metastasis	Counts: 3150	<a href="#">Detail</a>	 Biomarker: recurrence	Counts: 294	<a href="#">Detail</a>
 Biomarker: cellgrowth	Counts: 2479	<a href="#">Detail</a>	 Biomarker: emt	Counts: 1050	<a href="#">Detail</a>
 Biomarker: apoptosis	Counts: 2617	<a href="#">Detail</a>	 Biomarker: autophagy	Counts: 118	<a href="#">Detail</a>

**Figure S2.** Detail information of diagnostic and therapeutic biomarkers in LncACTdb 3.0.

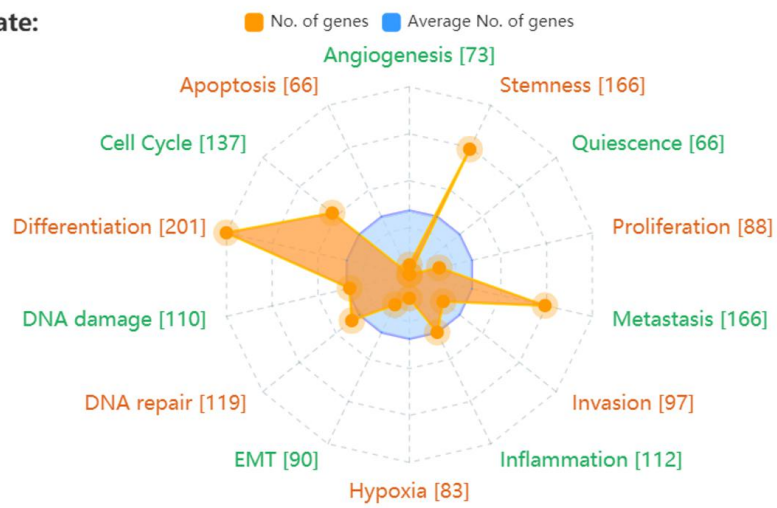


**Figure S3.** Schematic diagram of hypergeometric test principle.

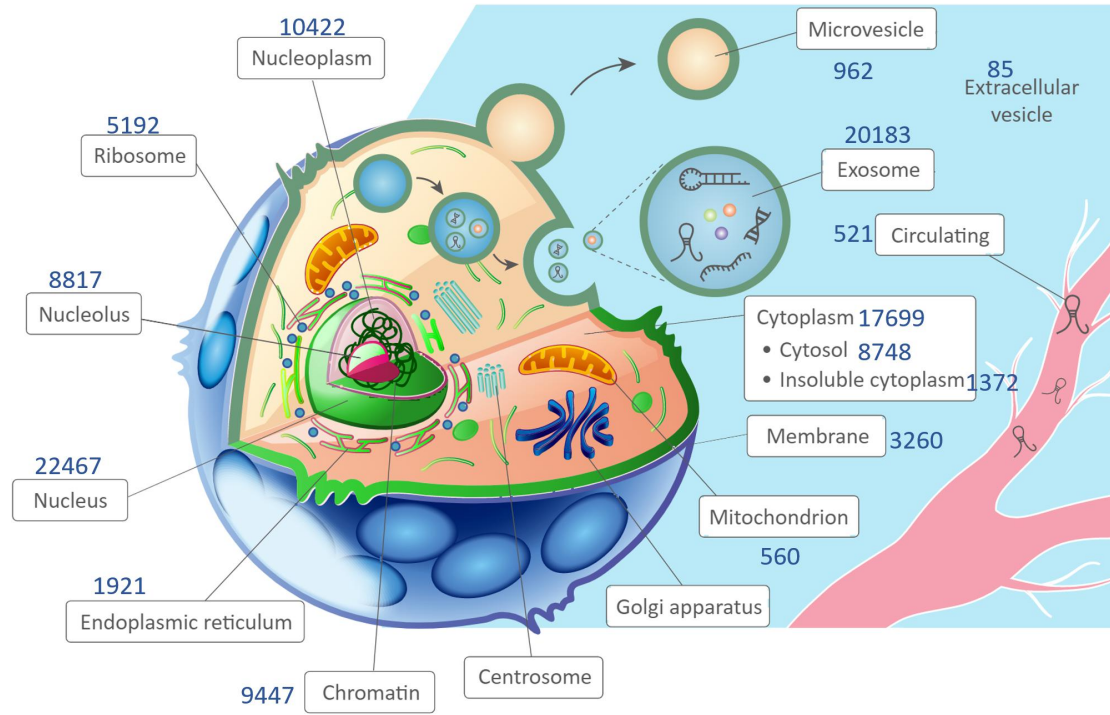
**A Sample hallmark state:**



**B Tumor cell state:**



**Figure S4.** The number of genes for 10 classic cancer hallmark processes (A) and 14 functional states of tumors cells (B) in LncACTdb 3.0.

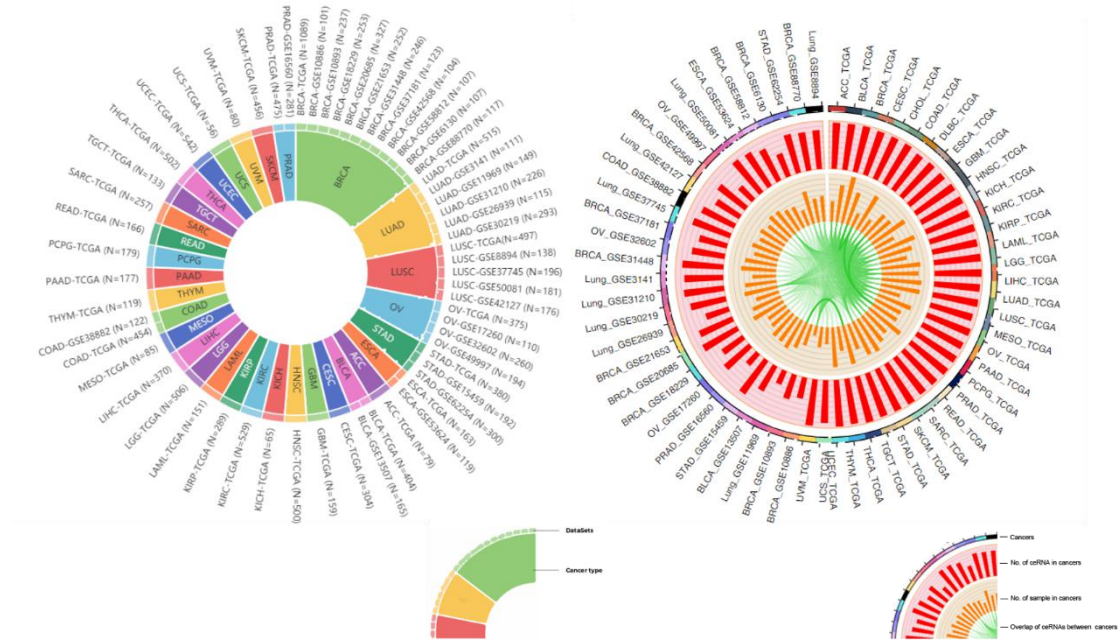


**Figure S5.** Data distribution of sub-cellular and extracellular vesicle locations for lncRNAs, miRNAs, and mRNAs in LncACTdb 3.0.





boxes represent the neighbourhood of  $x_q$  and  $y_q$  respectively. The intersection of the two boxes is the dark grey box, which represents the neighbourhood of  $(x_q, y_q)$ . The number of plots in the light, medium and dark grey boxes is  $n_x^q$ ,  $n_y^q$  and  $n_{xy}^q$  respectively. The statistic is designated as  $Q_{xy}^q$ . If  $x$  and  $y$  are independent of each other, the statistic follows standard normal distribution. If the statistic  $Q_{xy}^q$  is significantly larger, there is an edge between  $x$  and  $y$  in sample  $q$ , otherwise there is no edge.



**Figure S7.** Detail information of 62 high-throughput expression profile data containing 16,228 patients across 33 cancers from TCGA and Gene GEO public datasets in LncACTdb 3.0. The bar graph indicates the number of ceRNAs and patients in different datasets. The links indicate the ceRNA overlap (Jaccard index) between different cancers.

**LncACTdb 3.0** Home Search Browser Tools Statistic Download Submit & Contact Us Help

**Welcome to LncACTdb 3.0**  
A comprehensive database of experimentally supported ceRNA interactions and personalized networks contributing to precision medicine

Quick Search  
Search Keywords... **SEARCH**

Hot words: NEAT1 TP53 hsa-let-7a-5p MALAT1

**LncACTdb 3.0**

Analyze Tools

**Search**

We provide a variety of search methods for more flexible and convenient use of LncACTdb 3.0.

**Browser**

Browser interface provides a list of diseases, lncRNAs, mRNAs and biomarkers.

**Function Enrichment**

A tool to identify dysregulated functions of lncRNA-associated ceRNA based on Gene Ontology and biological pathways.

**Cell location**

Cell Location tool can find out the sub-cellular locations for a ceRNA. (Only Symbol or Ensembl ID of LncRNA/mRNA is required)

**Survival**

A tool to identify dysregulated functions of lncRNA-associated ceRNA based on Gene Ontology and biological pathways. In the new version we offer the forestplot

**LncACTdb 3.0**

24h visits: 0 12 15 18 21 0 3 6 9

**Analysis tools**

- » LncACT-Function
- » LncACT-Cell Location
- » LncACT-Survival
- » LncACT-Hallmark
- » LncACT-BLAST
- » LncACT-Network (Sample)
- » LncACT-Network (Pie)

**Tools & Other Page**

- » LncACT-Network (Exp)
- » LncACT-State
- » LncACT-Tree
- » LncACTdb 3.0 - Browser
- » LncACTdb 3.0 - Statistic
- » LncACTdb 3.0 - Download
- » LncACTdb 3.0 - Help

**Contact us**

» Xia Li: lixia@hrbmu.edu.cn

» Phone & Fax: +86-451-86615922

» Contact us

» Address: 194 Xuefu Road, Harbin 150081, CHINA

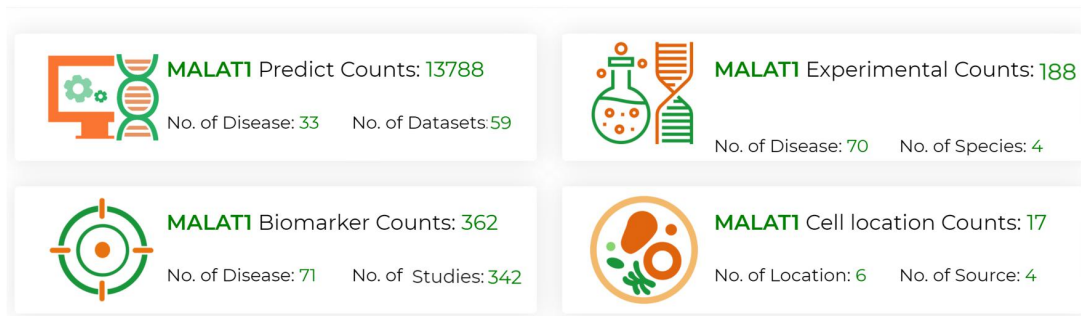
**Quick Links**

- » LncACTdb2.0
- » LncCell
- » LncCeVar
- » Lnc2Cancer 3.0
- » lincSNP 3.0
- » Lnc2Meth
- » MSDD

**Figure S8.** A screenshot of LncACTdb 3.0 Home page.

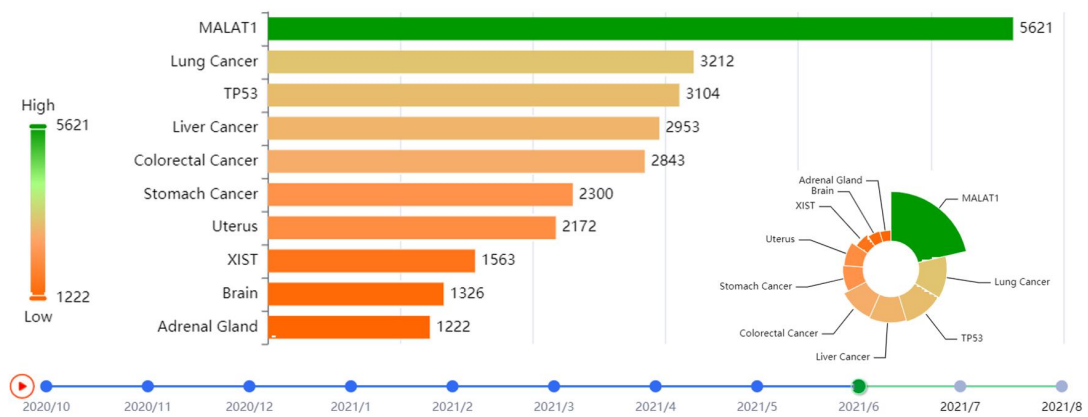
A

 The quick search result in LncACTdb 3.0



B

 LncACT-Hotpoint



**Figure S9.** A quick search engine is available for users to directly investigate data or perform analyses. (A) A quick search result of MALAT1 in LncACTdb 3.0. (B) A screenshot of LncACTdb 3.0 illustrating the most visited items in recent months.