Appendix S1

Supplementary Materials and methods

RNA quantification and qualification. The liver tissue samples (100 mg) collected from biliary atresia or choledochal cyst patients were stored in a tube with RNALater stabilization solution (Invitrogen, USA) in liquid nitrogen. Total RNA was extracted from liver tissues using TRIzol (Invitrogen, Carlsbad, CA, USA) following the manufacturer's protocol. Total RNA was evaluated mainly by three ways: A Nanodrop ND-2000 spectrophotometer (Thermo Scientific, Wilmington, DE) was applied to preliminary quantify and qualify the RNA concentration; Qubit was used to specifically calculate the concentration of RNA; 1.5% agarose gel electrophoresis was used to detect the degradation or contamination of the RNA; Agilent 2100 Bioanalyzer was used to evaluate he integrity of RNA (RIN value).

Library preparation for RNA-seq. The input standard for cDNA library generation were: (1) RNA concentration ≥200 ng/μl; (2) total RNA >5 μg per sample; and (3) OD 260/280 value between 1.8 and 2.2.

After the RNA samples were qualified, rRNA was removed by using Ribo-zero kit (EpiCentre, Madison, WI), and then the RNA was fragmented under high temperature and metal ions. The first-stranded cDNA strand was synthesized by using ribosomal-depleted RNA as a template with random hexameric primers, while the second-stranded cDNA strand was synthesized with by adding buffer, dNTPs (dUTP, dATP, dGTP, and dCTP) and enzymes. The double-stranded cDNA was purified by using VAHTS™ DNA Clean Beads (Vazyme, Nanjing, China). The final strand-specific cDNA library was constructed through a series of experiments such as end repair, tailing, sorting, and digestion of cDNA containing U by using UDG enzyme, and PCR enrichment. After the construction of the library, Qubit 3.0 was applied to preliminarily determined the concentration of cDNA. Agilent 2100 Bioanalyzer was used to evaluate the library quality. And then ABI Step One Plus Real-Time PCR system was used to specifically calculate the concentration of the library. Finally, Illumina HiSeq (Illumina, San Diego, CA) sequencing was performed after pooling different libraries according to the requirements of effective concentration and target data volume.

Sequencing and quality control and mapping of clean reads. The raw image data obtained by high-throughput sequencing is converted into sequence data by CASAVA base calling, i.e. raw data or raw reads, were as FASTQ format, followed by subsequent evaluation of the quality control of FASTQ data. Clean reads were obtained by removing those contained adapter, poly-N (N>5%, i.e. base cannot be identified), or low quality reads (reads with a quality score <10 accounts for >50% of the total reads) .

Differential expression analysis of mRNAs and lncRNAs. HISTA2 was applied for sequence alignment analysis using reference sequences. The reference sequences of the corresponding species were downloaded from the database Ensemble GRCh38.p7 (ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens). Based on the alignment results of sequencing data, combined with the annotation file of the reference sequence, the reads to different locations of the reference sequence, including exon, intronic and intergenic regions, were statistically analyzed. Expression of mRNAs and lncRNAs was normalized and outputted with StringTie version 1.3.3b (http://ccb.jhu.edu/software/stringtie/). Cufflinks was then applied for the differential expression analysis. |log2FC|>1 and P-value <0.05 were used as the cut-off criteria. Volcano plots and hierarchical clustering were drawn to visualize the overall distribution of differential transcripts.

Bioinformatic RNA-seq analysis. The functional enrichment analysis was performed for each cluster of genes by using the DAVID (1) (https://david.ncifcrf.gov). Gene Ontology database (2) (GO; http://www.geneontology.org/), and Kyoto Encyclopedia of Genes and Genomes (3) (KEGG; https://www.kegg.jp/) with a P-value <0.05 as cutoff criterion. Top 100 up- and down-regulated DEmRNAs-protein interaction network analysis was performed based on STRING protein interaction database (http://string-db.org/) and the R language package STRINGdb. The protein-coding mRNAs (100-kb upstream and downstream) adjacent to lncRNAs were selected as their target mRNAs and DElncRNA-DEmRNA nearby-targeted network was obtained. The DElncRNA-DEmRNA pairs with absolute values of PCC >0.99 and P<0.01 were selected, and DElncRNA-DEmRNA co-expression network was constructed. Cytoscape software version 3.5.0 was applied to visualize above networks. The expression of screened DEmRNAs obtained from our RNA sequencing were verified using the GSE46960 dataset. The schematic representation of the bioinformatics pipeline used is shown in Fig. S1.

References

1. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA: DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 4: P3, 2003.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 25: 25-29, 2000.
3. Kanehisa M and Goto S: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30, 2000.

Figure S1. Processing and database construction for RNA-seq.