Ye et al. *Zool. Res.* 2021, 42(6): 692−709
https://doi.org/10.24272/j.issn.2095-8137.2021.272

ZR | Zoological Research

Article Open Access

Comprehensive annotation of the Chinese tree shrew genome by large-scale RNA sequencing and long-read isoform sequencing

Mao-Sen Ye[1,2,#], Jin-Yan Zhang[1,2,#], Dan-Dan Yu[1,3], Min Xu[1,3], Ling Xu[1,3], Long-Bao Lv[3], Qi-Yun Zhu[4], Yu Fan[1,3], Yong-Gang Yao[1,2,3,*]

[1] *Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences & Yunnan Province, and KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan* 650204, *China*

[2] *Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, Yunnan* 650204, *China*

[3] *National Resource Center for Non-Human Primates, National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility), Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan* 650107, *China*

[4] *State Key Laboratory of Veterinary Etiological Biology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, Gansu* 730046, *China*

ABSTRACT

The Chinese tree shrew (*Tupaia belangeri chinensis*) is emerging as an important experimental animal in multiple fields of biomedical research. Comprehensive reference genome annotation for both mRNA and long non-coding RNA (lncRNA) is crucial for developing animal models using this species. In the current study, we collected a total of 234 high-quality RNA sequencing (RNA-seq) datasets and two long-read isoform sequencing (ISO-seq) datasets and improved the annotation of our previously assembled high-quality chromosome-level tree shrew genome. We obtained a total of 3 514 newly annotated coding genes and 50 576 lncRNA genes. We also characterized the tissue-specific expression patterns and alternative splicing patterns of mRNAs and lncRNAs and mapped the orthologous relationships among 11 mammalian species using the current annotated genome. We identified 144 tree shrew-specific gene families, including interleukin 6 (*IL6*) and STT3 oligosaccharyltransferase complex catalytic subunit B (*STT3B*), which underwent significant changes in size. Comparison of the overall expression patterns in tissues and pathways across four species (human, rhesus monkey, tree shrew, and mouse) indicated that tree shrews are more similar to primates than to mice at the tissue-transcriptome level. Notably, the newly annotated purine rich element binding protein A (*PURA*) gene and the *STT3B* gene family showed dysregulation upon viral infection. The updated version of the tree shrew genome annotation (KIZ version 3: TS_3.0) is available at http://www.

Received: 02 September 2021; Accepted: 24 September 2021; Online: 26 September 2021

Foundation items: This study was supported by the National Natural Science Foundation of China (U1902215 to Y.G.Y. and 31970542 to Y.F.), Chinese Academy of Sciences (Light of West China Program xbzg-zdsys-201909 to Y.G.Y.), and Yunnan Province (202001AS070023 and 2018FB046 to D.D.Y. and 202002AA100007 to Y.G.Y.)

[#]Authors contributed equally to this work

*Corresponding author, E-mail: yaoyg@mail.kiz.ac.cn

treeshrewdb.org and provides an essential reference for basic and biomedical studies using tree shrew animal models.

Keywords: Tree shrew; Genome annotation; Transcriptome; Gene family; Virus infection

INTRODUCTION

Suitable animal models are essential for expanding our knowledge regarding fundamental biological questions and for developing new drugs, vaccines, and therapeutics (McGonigle & Ruggeri, 2014; Robinson et al., 2019; Yao et al., 2015). An ideal animal model should possess many features, including high genetic similarity to humans, similar pathobiology and symptoms, efficacy with drug prediction and response, low cost, and low restriction (Bennett & Panicker, 2016; McGonigle & Ruggeri, 2014; Robinson et al., 2019; Yao et al., 2015). The Chinese tree shrew (*Tupaia belangeri chinensis*) is a small rat-sized (100–150 g) mammal with a short reproductive cycle (~6 weeks) (Yao, 2017; Zheng et al., 2014), and is widely distributed in Southeast Asia and South and Southwest China. In the past few decades, tree shrews have been widely used in a variety of biomedical studies, including research on viral infections (Amako et al., 2010; Li et al., 2018; Xu et al., 2007, 2020c), cancer (Ge et al., 2016; Lu et al., 2021), myopia (He et al., 2014; Levy et al., 2018; Phillips et al., 2000), visual cortex function (Fitzpatrick, 1996; Lee et al., 2016; Petry & Bickford, 2019), and neuroscience (Dimanico et al., 2021; Fan et al., 2018; Ni et al., 2018; Savier et al., 2021; Wei et al., 2017). Our research on tree shrews began with the genetic dissection of the Chinese tree shrew genome (Fan et al., 2013). We also aimed to promote the use of this animal in basic and biomedical research by continuing to update relevant genome information (Fan et al., 2014, 2019). Moreover, we developed two immortalized tree shrew cell lines for resource sharing (Gu et al., 2019b; Zhang et al., 2020b) and established the first genetic manipulation of tree shrews using spermatogonial stem cells to successfully generate transgenic offspring (Li et al., 2017). Compared with commonly used animal models such as rodents, tree shrews are phylogenetically closer to primates (Fan et al., 2013, 2019), and can more accurately mimic the physiological and pathological conditions of humans.

Accurate genome assembly and annotation are crucial for understanding tree shrew biology and for developing disease models using this animal. Indeed, creating an animal model of human disease using a tree shrew genome-based approach (Yao, 2017) is dependent on high-quality annotations of the tree shrew genome. Many attempts have been made to decipher the tree shrew genome in great detail and accuracy (Fan et al., 2013, 2019; Sanada et al., 2019). We successfully assembled the first high-quality genome of the Chinese tree shrew (KIZ version 1: TS_1.0) using high depth (~79X) short-read sequencing technology (Fan et al., 2013) and the first chromosome-level tree shrew genome (KIZ version 2: TS_2.0) using single-molecule real-time (SMRT) sequencing technology (Fan et al., 2019). The release of two versions of the tree shrew genome at www.treeshrewdb.org (Fan et al., 2014, 2019) has undoubtedly enhanced our knowledge on the usage of this species. Recently, Sanada et al. (2019) assembled a tree shrew genome using short reads for coding sequence (CDS) annotation. However, despite efforts to improve the annotation of the tree shrew genome, our understanding of the coding and non-coding genes of the tree shrew remains incomplete and unlikely to meet the growing needs of the research field.

RNA sequencing (RNA-seq) technology provides accurate and massive amounts of information regarding the direct transcription status of a genome (Cardoso-Moreira et al., 2019; Stark et al., 2019). The emergence of third-generation sequencing, which features long sequence reads that can cover the full-length of most transcripts (Gordon et al., 2016; Sharon et al., 2013), has greatly improved the accuracy of transcript structure annotation. Previous studies using next-generation sequencing (NGS) based on RNA-seq and long-read isoform sequencing (ISO-seq) have revealed the complexity and characteristics of eukaryotic transcriptomes (Chen et al., 2017). Using ortholog and *de novo* annotations (Garber et al., 2011; Yandell & Ence, 2012), transcriptome sequencing has been widely used to annotate the genomes of plants (Purugganan & Jackson, 2021; Wang et al., 2019), model animals (Ji et al., 2020; Nudelman et al., 2018; Zhang et al., 2020a), and livestock (Beiki et al., 2019; Foissac et al., 2019). In this study, we aimed to provide a more comprehensive tree shrew genome annotation using a wide range of transcriptome sequencing data. We collected high-quality RNA-seq datasets of tree shrews from publicly available sources (Supplementary Table S1), as well as two ISO-seq datasets and 139 RNA-seq datasets newly generated in this study. These transcriptome datasets included expression data of tree shrew cells and tissues under different conditions, including viral infection (Sanada et al., 2019; Yan et al., 2012), normal tissue (Fan et al., 2013; Han et al., 2020), and pathological tissue (Li et al., 2017; Lin et al., 2014; Tu et al., 2019; Wu et al., 2016b; Zhang et al., 2020b). Using a stringent pipeline, we obtained a total of 53 298 newly annotated coding transcripts and 115 562 newly annotated non-coding transcripts and produced a relatively complete and reliable tree shrew genome annotation (KIZ version 3: TS_3.0). Based on this comprehensive annotation, we further explored the spatial expression and alternative splicing patterns of the tree shrew transcripts and characterized the orthologous relationships among tree shrews and other species. We also compared expression similarity across species and provided a landscape of the innate immune response in tree shrews upon viral infection.

MATERIALS AND METHODS

Animals and tissue collection

Nine adult Chinese tree shrews were purchased from the Experimental Animal Center of the Kunming Institute of Zoology, Chinese Academy of Sciences. Animals were anesthetized with pentobarbital and intracardially perfused with phosphate-buffered saline (PBS). Eight tissues (small intestine, liver, heart, kidney, spleen, ovary, brain, and testis) from four animals were collected and snap-frozen in liquid

nitrogen. The remaining animals were used for the isolation of tree shrew primary renal cells (TSPRCs) for viral infection assays. All animal experiments were approved by the Institutional Review Board of the Kunming Institute of Zoology, Chinese Academy of Sciences.

ISO-seq for tree shrew tissues

Tissues from two adult Chinese tree shrews were used for ISO-seq (Supplementary Table S2). RNA extraction, library construction, and sequencing were performed by Annoroad Gene Technology (China). In brief, total RNA from each sample was isolated using a NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (Catalog # 7530; New England Biolabs Inc., USA) and processed following the manufacturer's protocols. RNA degradation and contamination were monitored by 1% agarose gels and RNA purity was checked using a NanoPhotometer® spectrophotometer (IMPLEN, USA). RNA integrity was assessed using a Qubit® RNA Assay Kit with a Qubit® 2.0 Fluorometer (Life Technologies, USA) and an RNA Nano 6000 Assay Kit with the Bioanalyzer 2100 system (Agilent Technologies, USA). Equal amounts of RNA from each tissue of the two tree shrews were pooled as one mixed RNA sample for ISO-seq. Two ISO-seq libraries (<4 kb and >4 kb) were prepared according to the Isoform Sequencing protocol (ISO-seq™) using a Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin™ Size Selection System (Sage Science, USA) protocol as described by PacBio (Menlo Park, USA). SMRT sequencing was performed on the Pacific Bioscience Sequel System using two SMRT cells.

The ISO-seq data were processed using IsoSeq v3.4.0 (https://github.com/PacificBiosciences/IsoSeq). Only sequence reads containing both 5' and 3' adaptors were retained to cover the entire transcript. We used LoRDEC (Salmela & Rivals, 2014) to correct errors in the SMRT reads by referring to the RNA-seq data. Subsequently, the corrected SMRT reads were aligned to the tree shrew reference genome TS_2.0 (Fan et al., 2019) using GMAP (Wu et al., 2016a) to locate the position of the predicted genes on the pseudochromosomes.

Compilation of publicly available tree shrew RNA-seq datasets

To ensure a robust and complete annotation of the tree shrew genome, we obtained all publicly available RNA-seq datasets of tree shrews from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), DNA Data Bank of Japan (DDBJ, https://www.ddbj.nig.ac.jp/index-e.html), and China National Center for Bioinformation (BIGD, https://bigd.big.ac.cn/). These transcriptome sequencing datasets were originally obtained by sequencing normal and pathological tissues or cells and represent a wide spectrum of biological and pathological conditions (Supplementary Table S1).

We used the following strategy for quality control (QC) of the RNA-seq data and filtered those data that did not meet requirements. Briefly, raw sequencing reads were processed by Trimmomatic (v0.38) (Bolger et al., 2014) to trim adaptor and low-quality sequences, with the parameters "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36". After read filtering, the quality of the clean reads was assessed by FastQC (https://sourceforge.net/projects/fastqc.mirror/). Datasets that passed QC (Q30>20) were aligned to the chromosome-level tree shrew genome (KIZ version 2: TS_2.0; https://www.treeshrewdb.org/download) using STAR (v2.6.0c) (Dobin et al., 2013). We discarded those datasets that failed to pass QC, i.e., mapping ratio to genome below 75%. After QC, a total of 91 publicly available RNA-seq datasets were retained for analysis (Supplementary Table S1).

RNA-seq of tree shrew tissues and cells with or without viral infection

To enhance the credibility of the genome annotation, we generated new RNA-seq data from tissues and/or cells of nine tree shrews (Supplementary Table S2) for further analysis with the publicly available RNA-seq data.

To explore the changes in gene expression in tree shrew cells in response to viral infection, we performed RNA-seq of virus-infected TSPRCs. We used the same procedure to isolate and culture TSPRCs and challenge cells with or without virus as described in our previous studies (Xu et al., 2016, 2020a; Yu et al., 2014). Briefly, TSPRCs were infected with a DNA virus (herpes simplex virus type 1, HSV-1; multiplicity of infection (MOI)=10) and RNA viruses (Sendai virus (SeV, 20 hemagglutinating units/mL), encephalomyocarditis virus (EMCV, MOI=2), and Newcastle disease virus (NDV, MOI=1)), respectively, for the indicated times before harvesting for RNA-seq (Supplementary Table S3). RNA-seq of tree shrew tissues and infected cells was performed by Annoroad Gene Technology (China). Approximately 1 μg of total RNA from each sample was used to construct the RNA-seq libraries (500–1 000 bp) with a NEBNext® Ultra™ RNA Library Prep Kit for Illumina®. The quality of each library was assessed using the Agilent Bioanalyzer 2100 system. Libraries were sequenced on the Illumina NovaSeq platform and 150 bp paired-end reads were generated. We followed the same QC procedures as described above for processing the publicly available RNA-seq data. We also included transcriptome datasets of lung tissues from influenza A virus (IAV)-infected tree shrews and of hepatitis C virus (HCV)-infected tree shrew primary hepatocytes in the current analyses (authors' unpublished data).

Evaluation of coding ability of transcripts and annotation of coding genes

We assembled RNA transcripts based on the RNA-seq reads from publicly available datasets (Supplementary Table S1) and the new data generated in this study (Supplementary Tables S2, S3) using StringTie (v2.1.1) in a reference-guided manner (-G) (Pertea et al., 2015). The assembled RNA-seq transcript models were merged with the SMRT models by StringTie merge (--merge option) to obtain a transcript model covering all transcriptome datasets. To ensure the credibility of the transcripts, we only retained transcripts with high-confidence expression levels (fragments per kilobase per million (FPKM)>0.5 in at least one sample) during the merging of the RNA-seq and SMRT transcripts. We used Gffcompare (https://ccb.jhu.edu/software/stringtie/gffcompare.shtml) (Pertea & Pertea, 2020) to compare the assembled transcript

models with the transcript models of the reference genome TS_2.0 (reference model). We treated those transcripts with no matches to the reference models (class code "=") as newly identified transcripts.

We evaluated the coding potential of the newly identified transcripts by incorporating the results predicted using the following approaches to achieve a more reliable annotation: (1) The Coding Potential Assessment Tool (CPAT) (Wang et al., 2013) was applied to evaluate the transcript coding ability using a logistic regression model. The hexamer frequency table was built using "make_hexamer_tab.py" script, and the logit model was built using "make_logitModel.py" script. (2) We used the Coding Potential Calculator 2 (CPC2) (Kang et al., 2017; Kong et al., 2007) to evaluate the coding ability of the transcripts employing a novel discriminative model based on four sequence-intrinsic features. (3) We used TransDecoder (https://github.com/TransDecoder/) to predict the high-confidence open reading frames (ORF) of each transcript. (4) Pfam (El-Gebali et al., 2019), which contains a comprehensive archive of protein domains, and UniProtKB/Swiss-Prot (Boutet et al., 2007), which contains a comprehensive archive of protein sequences from multiple species, were used to identify potentially translated ORFs. Except for CPAT, we ran all other programs with their default parameters and integrated the prediction results with the following procedures. First, we integrated the prediction results from both CPAT and CPC and only transcripts that met the coding cut-off of both approaches (CPAT, coding potential>0.4; CPC, designated as "coding") were subjected to further analyses. Second, the ORF of each transcript was predicted using TransDecoder (https://github.com/TransDecoder/) and only transcripts with at least one high-confidence ORF were retained. Third, we scanned the potentially translated ORFs against the Pfam (http://rfam.xfam.org/) (El-Gebali et al., 2019) and UniProtKB/Swiss-Prot databases (Boutet et al., 2007). Those transcripts with at least one predicted Pfam domain or high protein sequence identity (E-value>1e-5) with at least one known protein were defined as coding transcripts. We selected the longest transcript from each gene locus as the representative transcript of the gene. We BLASTed the representative transcripts against the UniProtKB/Swiss-Prot and UniProtKB/Trembl databases (Boutet et al., 2007) using blastall (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). The best hit gene name of each gene locus was designated as the name of the newly annotated gene. For multiple gene loci with the same best hit, we modified the gene name by adding "LI (like)+number" to avoid gene name redundancy. We used eggnog-mapper (Huerta-Cepas et al., 2017) to BLAST the translated ORFs for each transcript against the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa et al., 2017) and Gene Ontology (GO) database (http://geneontology.org/) for gene function annotation. The best KEGG pathway and GO term hits were designated for each gene. Fourth, lncRNAs were identified and annotated from the transcripts using the following criteria: (1) transcripts are >200 nucleotides (nt) long and meet the non-coding cut-offs of CPAT (coding potential<0.4) and CPC (designated as "noncoding"); (2) transcripts have a predicted ORF<100 nt; and (3) transcripts have a low similarity (E-value>1e-5) with the tRNA family in the Rfam database (El-Gebali et al., 2019) and UniProtKB/Swiss-Pro database (Boutet et al., 2007). We defined those transcripts showing inconsistent prediction of coding RNAs and lncRNAs based on the above approaches as biased transcripts.

After coding potential evaluation and gene annotation, we merged the TS_2.0 genome annotation file (Fan et al., 2019) with the newly annotated transcripts to generate the TS_3.0 genome annotation. We verified the accuracy of the TS_3.0 transcripts by comparing the annotated transcripts with those characterized by molecular cloning. In total, 30 transcripts reported in our previous studies (Gu et al., 2019a; Luo et al., 2018; Yao et al., 2019; Yu et al., 2014, 2016) (Supplementary Table S4) were selected for comparison using blastall (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). An E-value<1e-5 was used as a cutoff to define whether a transcript obtained by molecular cloning was included in the TS_3.0 genome annotation.

Benchmarking universal single-copy orthologs (BUSCO) analysis

We used BUSCO (Seppey et al., 2019) to evaluate the completeness of the TS_3.0 genome annotation using the *Mammalia* and *Eukaryota* BUSCO datasets (Simão et al., 2015; Waterhouse et al., 2018). To ensure that the sequences of each species originated from the genome annotation, we selected protein-coding genes according to the information provided in the gene transfer format (GTF) file for each species, and the longest transcript of each gene was selected for consideration. We compared the TS_3.0 annotation with the previous tree shrew genome annotations (Supplementary Table S6). We used Gffread (Pertea & Pertea, 2020) to extract the sequences from the reference genome according to the annotation information. BUSCO evaluation was run in transcriptome mode (-m tran).

Alternative splicing prediction, differential gene expression, and dimension reduction analyses

We used Kallisto (Bray et al., 2016) to quantify the expression level of each transcript. Briefly, the clean reads obtained from each RNA-seq dataset were mapped to the transcriptome constructed using the TS_3.0 annotation. We used SUPPA2 (Trincado et al., 2018) to predict alternative splicing events, including skipped exons (SE), alternative 5' splice sites (A5), alternative 3' splice sites (A3), mutually exclusive exons (MXE), and retained introns (RI). The splicing level of each gene in each RNA-seq dataset was quantified using the Percent Spliced-In (PSI) index. The PSI values of each gene were calculated based on the transcript model and expression level (transcripts per million, TPM) of all transcripts for the gene (Trincado et al., 2018).

To characterize tissue-specific gene/transcript expression levels and alternative splicing events, we identified specifically

expressed genes by applying the Wilcoxon rank-sum test and Dunn test in the R (http://www.R-project.org/) package Seurat (Butler et al., 2018). *P*-values were adjusted (P_{adjust}) by the Benjamini-Hochberg (BH) method. Genes/transcripts with a P_{adjust}<0.05 were defined as differentially/specifically expressed in a given condition. Using Seurat (Butler et al., 2018), we performed uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) for each tree shrew tissue based on TPM of mRNA and lncRNA at the gene and transcript levels, respectively. Those genes/transcripts expressed in all samples of a given tissue and with a gene/transcript |log2 fold-change|>0.5 when compared with other tissues were regarded as tissue-specific genes/ transcripts.

We used the R package DESeq2 (Love et al., 2014) to identify differentially expressed genes (DEGs) under virus infection conditions. The P_{adjust} values were calculated using the BH method, as described above. Genes were identified as dysregulated upon viral infection if P_{adjust}<0.05 and $|\log_2$ fold-change|>1 were met. KEGG and GO enrichment analyses were performed using the R package ClusterProfiler (Yu et al., 2012), with *P*-values adjusted by the BH method. A pathway with a P_{adjust}<0.05 was defined as significantly enriched.

Gene family analyses

We obtained protein sequences of multiple mammals from the Ensembl database (https://asia.ensembl.org/index.html), including *Homo sapiens* (GRCh38.p13), *Pan troglodytes* (Pan_tro_3.0), *Gorilla gorilla gorilla* (gorGor4), *Macaca mulatta* (Mmul_10), *Rattus norvegicus* (Rnor_6.0), *Mus musculus* (GRCm39), *Sus scrofa* (Sscrofa11.1), *Bos taurus* (ARS-UCD1.2), *Canis lupus familiaris* (CamFam3.1), and *Oryctolagus cuniculus* (OryCun2.0) (Supplementary Table S5). The orthologous relationships among species were calculated using OrthoFinder (Emms & Kelly, 2019). We used CAFÉ (De Bie et al., 2006; Mendes et al., 2020) to detect gene family size changes, including expansion and contraction, based on the orthogroups and phylogenetic tree constructed by OrthoFinder (Emms & Kelly, 2019). The phylogenetic tree was constructed using all protein-coding genes of the genomes.

Two gene families showed expansion in this study, i.e., STT3 oligosaccharyltransferase complex catalytic subunit B (*STT3B*) and subunit A (*STT3A*) and the interleukin 6 (*IL6*) gene family, which were featured for their potential roles in viral infection. We constructed gene trees of these gene families using the maximum-likelihood (ML) method (K2+G model) with 1 000 bootstraps. Trees were based on protein sequence alignment and constructed using MEGA (Kumar et al., 2018).

Tissue expression pattern and pathway gene similarity across species

We retrieved expression data from five tissues (liver, brain, kidney, testis, and heart) of mice (https://www.ebi.ac.uk/arrayexpress/E-MTAB-6798), rhesus macaques (https://www.ebi.ac.uk/arrayexpress/E-MTAB-6813), and humans (https://www.ebi.ac.uk/arrayexpress/E-MTAB-6814) from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) (Cardoso-Moreira et al., 2019). Principal component analysis of gene expression levels (TPM of each gene) for each tissue was constructed using the R package FactoMineR based on 11 059 one-to-one orthologous genes and 3 291 highly variable genes (defined by a coefficient of variation>0.1 TPM). Pathway gene information was retrieved from the KEGG database (Kanehisa et al., 2017). Protein sequence identities were calculated by BLASTing the tree shrew genes against human homologs and mouse genes against human homologs, respectively. Comparisons of protein sequence identity between mice and humans and between tree shrews and humans were performed using the Wilcoxon rank-sum test adopted in the R package. Here, *P*<0.05 was considered statistically significant.

RESULTS

Identification of high-quality tree shrew transcripts

To refine and annotate the chromosome-level tree shrew reference genome TS_2.0 (Fan et al., 2019), we adopted a stringent pipeline to integrate the related RNA-seq and SMRT datasets (Figure 1A). We collected and curated tree shrew transcriptome data across different tissues and cells with different viral infections. A total of 234 tree shrew transcriptome datasets were used in this study (Supplementary Tables S1–S3). These datasets covered a wide range of biological and pathological conditions, including cells infected with different viruses (Figure 1B) and normal and pathological tissue expression (Figure 1C). We used these datasets of diverse conditions to ensure that we captured the diversity and quantity of the transcripts, especially those with low abundancy in normal conditions. After QC, we discarded four samples with a mapping ratio below 75% from further analysis. The remaining datasets had a mean mapping ratio of 93.26% (Figures 1C, D), reflecting relatively high completeness of the reference genome and high quality of the RNA-seq datasets used. Based on the transcriptome datasets, we assembled a transcript model with FPKM>0.5 for each sample using StringTie (Pertea et al., 2015) and obtained 423 965 transcripts for the tree shrew (hereafter referred as RNA-seq transcripts).

To verify the accuracy of the tree shrew RNA-seq transcripts, we constructed two ISO-seq full-length transcriptome libraries based on pooled RNA samples from eight tissues of two tree shrews. After QC and error correction, we obtained 36 381 non-redundant non-chimeric full-length transcripts located at 12 366 loci (hereafter referred as ISO-seq transcripts). The mean length of the ISO-seq transcripts was 2 371 nt and the longest ISO-seq transcript (death effector domain containing [DEDD] gene) was 9 417 nt. A total of 10 968 transcripts were matched between the RNA-seq and ISO-seq transcripts, accounting for 30.15% of the ISO-seq transcripts. Nearly all ISO-seq transcripts were captured by the RNA-seq transcripts, and only 0.9% (1 317/146 347) of exons and 2.6% (319/12 366) of loci captured by ISO-seq were missed by the RNA-seq transcripts. We combined the RNA-seq and ISO-seq transcripts and obtained a total of 403 792 transcripts located in 98 142 loci in the tree shrew genome. The compiled tree shrew transcripts were deposited in the Tree shrew Database (http://www.treeshrewdb.

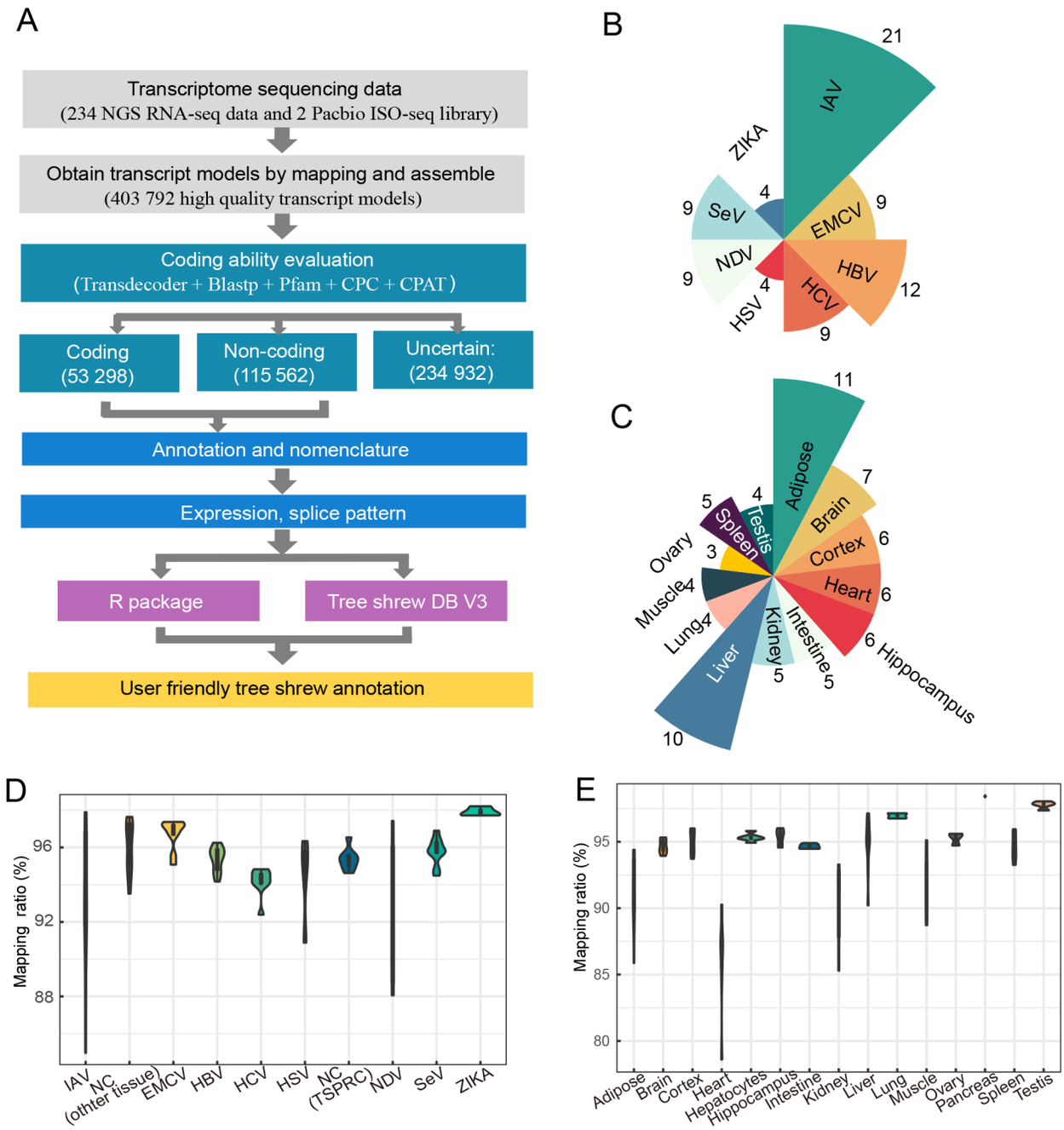

Figure 1 Reference-guided transcriptome assembly of tree shrew TS_3.0 genome annotation

A: Integrative pipeline for tree shrew genome annotation using publicly available and newly generated transcriptome datasets. B,C: Number of RNA-seq datasets of virus-infected tissues/cells (B) and normal tissues (C) analyzed in this study. D, E: Mapping ratio of RNA-seq data of virus-infected tissues/cells (D) and normal tissues (E) relative to reference tree shrew genome TS_2.0 (Fan et al., 2019). Sample information in (D) is listed in Supplementary Table S1.

org/download.html).

Expanded list of tree shrew coding and long non-coding transcripts

Both lncRNAs and mRNAs share similar biogenesis pathways and are involved in multiple biological processes (Jiang et al., 2019; Quinn & Chang, 2016), but exert their functions in different manners (Dahariya et al., 2019; Melé et al., 2017). We categorized transcripts into high-confidence coding transcripts (mRNAs), lncRNA transcripts, and biased transcripts. Among the 403 792 newly obtained transcripts, 115 562 (>200 nt) were located in 56 401 loci and were predicted to be lncRNA transcripts. Among these lncRNA transcripts, 50 576 were antisense, 60 118 were intergenic, and 4 868 were bidirectional. We predicted 53 298 coding

transcripts located in 19 242 loci. In addition, 234 932 transcripts located in 93 426 loci had inconsistent prediction regarding the characteristics of coding and lncRNA transcripts.

Overall, the expression levels of tree shrew lncRNAs were significantly lower than the expression levels of mRNAs at the gene and transcript levels (Figure 2A), consistent with previous reports on the human transcriptome (Iyer et al., 2015; Jiang et al., 2019). The exon number per lncRNA transcript was also significantly lower than that of mRNA in the tree shrew (Figure 2B), as observed in humans (Jiang et al., 2019). Moreover, the length of lncRNA was significantly shorter than that of mRNA in the tree shrew (Figure 2C). It has been reported that lncRNAs can regulate mRNA expression in a cis-regulatory manner (Jiang et al., 2019; Ørom et al., 2010; Ponjavic et al., 2009). Here, we calculated the expression correlation (Pearson correlation coefficient) between 10 000 mRNAs and their closest lncRNAs in the same genomic region across all RNA-seq datasets and compared the expression correlation between 10 000 randomly selected pairs of mRNAs and lncRNAs. Results showed that the expression correlation between closely located mRNA-lncRNA pairs was significantly stronger (Wilcoxon tests, P<2.2e-16) than that between randomly chosen pairs (Figure 2D). This provides additional evidence for the good accuracy of the lncRNA and mRNA annotations in the tree shrew genome.

We further compared 30 transcripts obtained by molecular cloning in our previous studies (Gu et al., 2019a; Luo et al., 2018; Yao et al., 2019; Yu et al., 2014, 2016) (Supplementary Table S4) with those predicted in the TS_3.0 genome annotation. All 30 transcripts showed very good alignment with the currently annotated transcripts (blastall, E-value<1e-5), and 47 additional transcripts were identified in these gene loci according to the TS_3.0 genome annotation, suggesting high accuracy and completeness of the transcript annotation. For instance, we observed all tree shrew *TLR* gene family members (Yu et al., 2016) and six *IL7* transcripts (Yu et al., 2014) reported in our previous studies in TS_3.0. The four *IL7* transcripts showed a complete sequence match with the corresponding transcripts in TS_3.0 (Supplementary Figure S1). Among the five alternative splicing events in the tree shrew transcriptome, SE was the most common type of alternative splicing event in the TS_3.0 transcripts (Figure 2E). This pattern is consistent with that of humans and mice (Figure 2E).

Compared to the TS_2.0 tree shrew genome (Table 1), we found 6 126 coding transcripts (including 207 single-exon transcripts) located in 3 514 loci, none of which had been previously annotated and thus represented newly annotated genes. We profiled the gene expression patterns of these newly annotated genes across all RNA-seq datasets and found that the expression levels of the genes were significantly lower than those of the previously annotated ones (Figure 2F). The low abundancies of these newly annotated genes may be the reason for missing annotation in our previous studies (Fan et al., 2013, 2019). The newly annotated genes were enriched (BH adjusted, P_{adjust}<0.05) in immune-related KEGG pathways, such as “Pattern recognition receptors” and “Inflammatory bowel disease (ko05321)” (Figure 2H), partly due to the bias of the RNA-seq datasets of tree shrew cells with viral infection. Combined with the previously annotated genes (Fan et al., 2019), 27 082 coding genes were finally annotated in the tree shrew genome.

We further compared the gene, transcript, and lncRNA numbers between TS_3.0 and the previously reported versions of tree shrew genome annotation and found remarkable improvement (Table 1). Evaluation of gene completeness of the TS_3.0 annotation relative to the TS_2.0 annotation by BUSCO (Figure 2G; Supplementary Table S6) showed that the ratio of complete BUSCOs increased from 92.16% to 98.04% for *Eukaryota* BUSCOs (255 genes) and from 81.52% to 92.80% for *Mammalia* BUSCOs (9 224 genes). Compared with the NCBI TupChi_1.0 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000334495.1) and TupaiaBase (Sanada et al., 2019), the current TS_3.0 annotation showed better completeness and better quality (Table 1).

Tissue expression and alternative splicing profiles of TS_3.0

To characterize the tissue-specific expression and alternative splicing patterns of each gene, we analyzed the transcriptome datasets from each tissue using UMAP and calculated the expression correlations. Both mRNAs and lncRNAs showed clear tissue-specificity in the context of UMAP (Figure 3A). The correlation matrix showed that the testis had the most unique tissue expression pattern compared with other tissues (Figure 3B). Notably, the testis possessed the most unique mRNAs and lncRNAs at the gene and transcript level (Figure 3C), consistent with the patterns reported in humans (Djureinovic et al., 2014). Many of the tree shrew testis-specific genes were involved in spermatogenesis (Supplementary Figure S2), as also reported in rats (Ji et al., 2020).

We also characterized the specificity and intensity of RNA alternative splicing. Alternative splicing intensity of genes differed significantly across the 13 tree shrew tissues under study (Kruskal-Wallis rank-sum test, P<2.2e-16) (Figure 3D). Brain-related tissues (including the brain, hippocampus, and cortex) showed the highest PSI, whereas heart tissue had the lowest PSI. Furthermore, UMAP using PSI showed that the alternative splicing pattern for each gene also presented tissue specificity (Figure 3E). Collectively, we found that alternative splicing intensity and gene specificity were meticulously regulated across different tree shrew tissues (Figure 3E), which may account for the different functions of the respective tissues and organs.

Orthologous relationships of tree shrew genes with other mammals

We identified a total of 272 814 genes in the 11 mammals under study (Supplementary Table S5). Among them, 95.1% (259 313 genes) could be assigned to an orthogroup (i.e., a set of genes from multiple species descended from a single gene from the last common ancestor of that set of species) (Emms & Kelly, 2019). In total, 25 249 tree shrew genes could be assigned to 12 485 orthogroups. Among these orthogroups, 191 were paralogs and appeared to be tree shrew specific. We identified 17 299 orthogroups (including

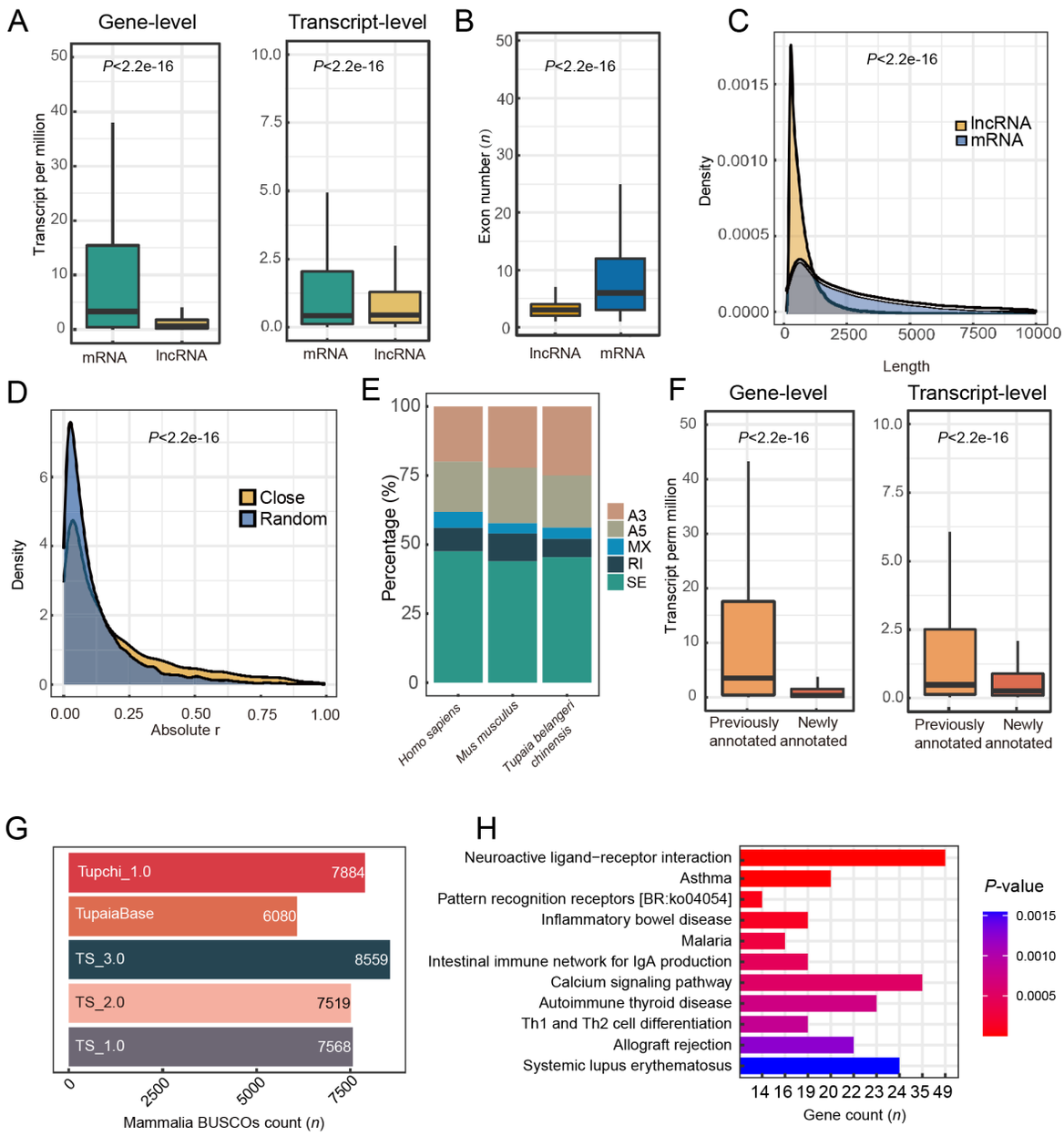

Figure 2 Characteristics of tree shrew TS_3.0 transcripts

A: Expression level of mRNAs is greater than that of lncRNAs at gene and transcript levels. B: mRNA transcripts have a higher number of exons than lncRNA transcripts. C: Average length of mRNA transcripts is longer than that of lncRNA transcripts. Density plot was drawn based on kernel density and statistical analysis was performed with Wilcoxon rank-sum test. D: Tree shrew lncRNAs exert cis-regulatory function on expression of proximal mRNAs. Close, mRNA and lncRNA pairs neighboring each other on pseudochromosome of tree shrew. Random, randomly selected mRNA and lncRNA pairs distant from each other in the genome. E: Percentages of alternative splicing events in reference genome annotations of tree shrew (TS_3.0), mouse (GRCm39), and human (GRCh38.p13). SE, skipped exon; A5, alternative 5' splice site; A3, alternative 3' splice site; MXE, mutually exclusive exons; RI, retained intron. F: Expression level of newly annotated coding genes is lower than that of previously annotated genes at gene and transcript levels. G: BUSCO evaluation of different tree shrew genome annotations showing that current version (TS_3.0) is superior. H: Pathway enrichment analysis of newly annotated genes showing enrichment in 11 pathways (P_{adjust}<0.05). Values in A, B and F are presented as a boxplot, and statistical analyses were performed by Wilcoxon rank-sum test.

14 549 one-to-one orthologs) shared between humans and tree shrews (Supplementary Table S7), which is a substantial improvement compared with the 12 840 one-to-one orthologs in TS_2.0 (Fan et al., 2019). Based on the current comprehensive orthologous relationships among species, we constructed a phylogenetic tree using the STAG algorithm (Emms & Kelly, 2019). We confirmed that the tree shrew is phylogenetically closer to primates than to rodents

Table 1 Comparisons of five tree shrew genome annotations

Parameters	TupChi_1.0 (NCBI)	TupaiaBase	TS_1.0	TS_2.0	TS_3.0
Coding genes					
Total number of coding genes	23 527	19 230	22 121	23 568	27 082
Transcript per coding gene	1.59	1	1	1	2.17
Annotated coding genes	23 537	12 612	20 225	20 811	25 127
Average mRNA length	48 104		33 712	40 114	41 239
Average CDS length	1 682	1 419	1 404	1 527	1 684
Average exon number	8.34	7.68	7.54	8.86	9.32
Average exon length	229	185	186	172	181
Average intron length	6 003	3 411	4 937	4 907	4 863
Complete BUSCOs (*Eukaryota* 255 genes)	216(84.7%)	195(76.5%)	221 (86.7%)	235(92.2%)	250(98.0%)
Complete BUSCOs (*Mammalia* 9 224 genes)	7 884(85.5%)	6 080(65.9%)	7 568(82.0%)	7 519(81.5%)	8 559(92.8%)
Non-coding genes					
Total number of lncRNA genes	3 718	–	–	–	56 401
Transcripts per lncRNA gene	5 179	–	–	–	2.05
Average lncRNA length	914	–	–	–	823
Average exon number	3.54	–	–	–	3.06
Average intron length	17 614	–	–	–	4 658

Tree shrew genome annotations TS_1.0 (Fan et al., 2013), TS_2.0 (Fan et al., 2019), and TS_3.0 (this study) were established in our studies. Tupchi_1.0, NCBI tree shrew annotation (https://www.ncbi.nlm.nih.gov/assembly/GCF_000334495.1/). TupaiaBase was reported by (Sanada et al., 2019). BUSCO: Benchmarking with Universal Single-Copy Orthologs. –: Not available.

(Figure 4A), as described in our previous study based on 2 117 single-copy one-to-one orthologs (Fan et al., 2013).

Gene gain and loss are important evolutionary processes that allow organisms to adapt to their environment (Page, 1998). Here, we analyzed changes in gene family size across 11 mammal species, including tree shrews (Supplementary Table S8), to validate and refine the previously characterized gene expansion and contraction events. We identified 120 gene families showing rapid expansion and 22 gene families showing rapid contraction (Figure 4A). The gene family exhibiting the greatest expansion was long-interspersed element-1 (LINE-1) retrotransposable element ORF (*LIRE1*), with 180 *LIRE1* genes identified in the TS_3.0 genome annotation. The gene family exhibiting the greatest contraction was immunoglobulin heavy variable 3–35 (Supplementary Table S9). The guanylate binding protein (*GBP*) gene family was found to have rapidly contracted, consistent with the findings of our previous study (Gu et al., 2019a). We found that *IL6* (Supplementary Figure S3) and *STT3B* (Supplementary Figure S4) were significantly expanded, with 34 of 39 *STT3B* gene family members being newly annotated and four of 13 *IL6* gene family members being refined in the TS_3.0 genome annotation, respectively. The *IL6* family contains *IL6*, cardiotrophin like cytokine factor 1 (*CLCF1*), cardiotrophin 1 (*CTF1*), ciliary neurotrophic factor (*CNTF*), interleukin 11 (*IL11*), interleukin 27 (*IL27*), LIF interleukin 6 family cytokine (*LIF*), and oncostatin M (*OSM*) (Rose-John, 2018). The tree shrew contained all these *IL6* family members, and 12 of the 13 *IL6* copies in tree shrews were not detected in other species. We found that each member of the *IL6* family was grouped into a single clade in the ML tree of the *IL6* family genes, confirming the close relationship of the expanded copies of each *IL6* gene family member (Figure 4B). *IL6LI7* appeared to be the ancestral gene of the tree shrew *IL6* gene copies. The tree shrew *IL6* family members were all located on pseudochromosome 6 (Figure 4C), with consistent exon numbers for each family member (Supplementary Figure S5). This suggests that the *IL6* gene copies were most likely generated from tandem duplication and segmental duplication. We constructed an ML tree for the tree shrew *STT3B* gene family members, together with those of the other mammals, and the *STT3A* paralog. The *STT3A* and *STT3B* copies showed a gene-specific clustering pattern (Figure 4D). In the clade for the expanded *STT3B* copies from the tree shrew, *STT3BLI27* diverged first and appeared to be the ancestral gene of the tree shrew *STT3B* family. Intriguingly, all 39 copies of *STT3B* in the tree shrew were distributed on 15 pseudochromosomes and one unplaced contig (Figure 4E). Of note, the tree shrew *STT3BLI27* had 16 exons, while the other copies of *STT3B* contained no more than four exons (Supplementary Figure S6), suggesting that expansion of the tree shrew *STT3B* was most likely caused by retrotransposon activity.

To further dissect the potential evolutionary roles of the tree shrew gene family size changes, we conducted enrichment analysis using the canonical genes of each rapidly changing gene family. Results showed that gene families that have undergone rapid size change were enriched in the “immune response to tumor cell”, “regulation of cytokine production”, and “regulation of DNA metabolic process” pathways (Figure 4F).

Expression similarity across different species

To study the mRNA expression patterns of tissues and related pathways across humans, rhesus monkeys, mice, and

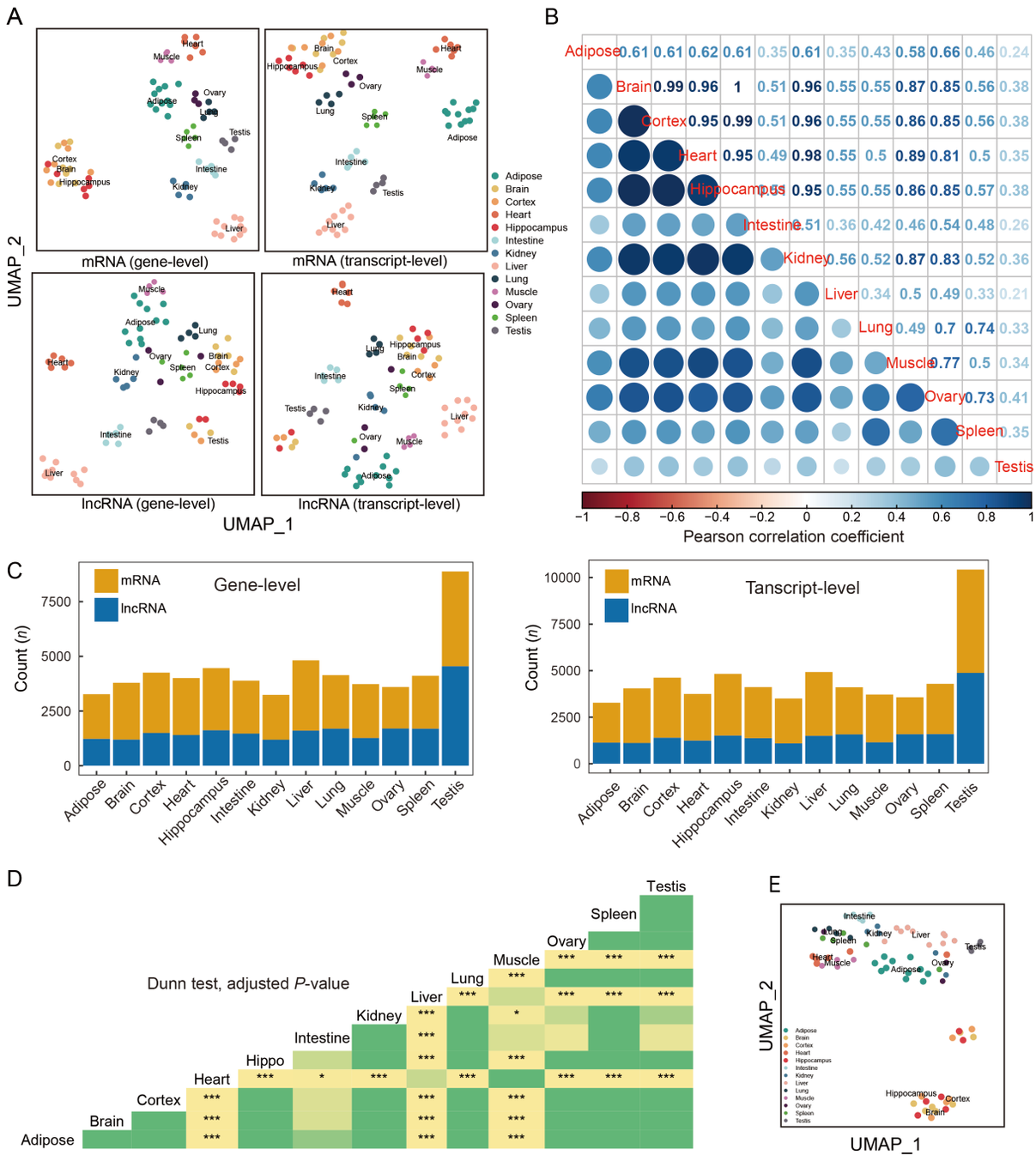

Figure 3 Tissue expression and alternative splicing profiles of tree shrew TS_3.0 transcripts

A: Tissue expression profiles of mRNAs and lncRNAs annotated in TS_3.0 at gene level (*left* panel) and transcript level (*right* panel). UMAP_1, UMAP dimension 1; UMAP_2, UMAP dimension 2. Detailed information on RNA-seq datasets of 13 tree shrew tissues is listed in Supplementary Tables S1, S2. B: Expression correlation matrix among different tree shrew tissues based on expression levels of mRNAs and lncRNAs. C: Tissue-specific expression patterns of mRNAs and lncRNAs at gene level (*left*) and transcript level (*right*). D: Comparison of PSI across 13 tree shrew tissues. *P*-value was calculated based on Dunn test and adjusted by Benjamini-Hochberg method. *: P_{adjust}<0.05; ***: P_{adjust}<0.0005. Different colors indicate different P_{adjust} values in triangle map. E: UMAP was constructed based on PSI of each gene in 13 tree shrew tissues.

Chinese tree shrews, we retrieved tissue RNA-seq data of mice, monkeys, and humans (Cardoso-Moreira et al., 2019), and compared their clustering patterns via principal component (PC) analysis. The species clustering patterns based on expression data from brain, liver, testis, kidney, and heart tissues showed distant divergence of mice from primates and tree shrews in the second PC, whereas humans, monkeys, and tree shrews were mainly separated by the first PC (Figure 5A). However, these clustering patterns should be considered with caution as the first and second PCs only

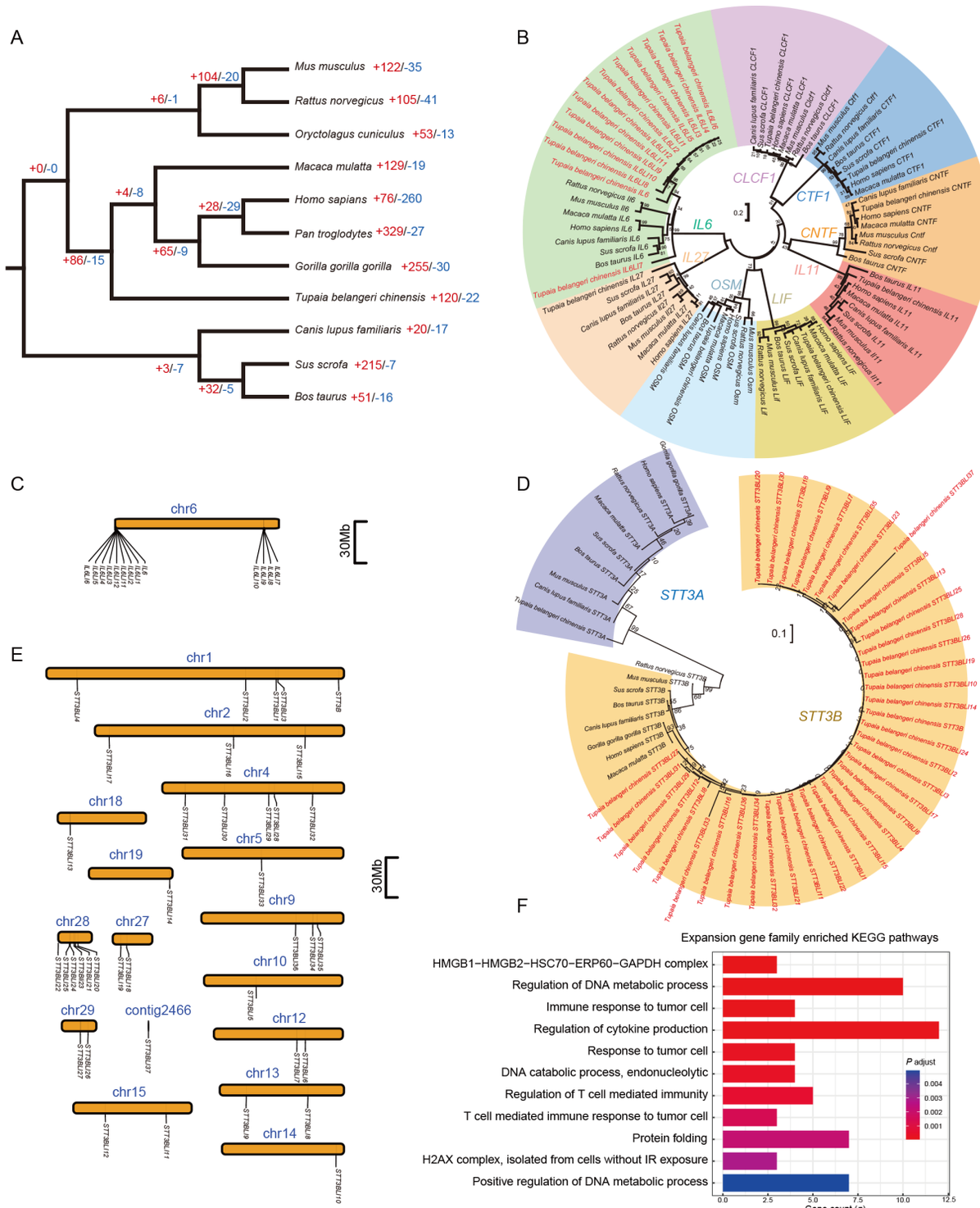

Figure 4 Orthologous relationships and gene family size changes among different species

A: Phylogenetic tree of 11 mammals using orthogroups. Numbers on tree branches refer to numbers of gene family expansion (+red) and contraction (-blue), respectively. B: Maximum-likelihood (ML) trees of *IL6* gene family. Coding sequence of the longest transcript for each gene in each species was used to construct ML tree. Values on tree branches refer to support of 1 000 bootstraps. The tree shrew *IL6* gene family had 13 copies, labeled in red in the tree. C: Locations of 13 tree shrew *IL6* gene copies on pseudochromosome 6 (chr6). D: ML trees of *STT3B* gene family and *STT3A*. Tree shrew *STT3B* gene family had 39 copies. E: Locations of 39 tree shrew *STT3B* gene copies on 15 pseudochromosomes and one unplaced contig. Pseudochromosomes and unplaced contig were defined in reference tree shrew genome TS_2.0 (Fan et al., 2019). F: Pathway enrichment of genes from tree shrew-specific gene families with significant expansion.

contributed to a proportion of expression variance.

We further determined the gene identity of tree shrews to humans using pathway analysis. For each human KEGG pathway, we compared protein sequence identities between mice and humans and between tree shrews and humans. Genes in 13 pathways showed greater protein sequence identity between tree shrews and humans than between mice and humans (Figure 5B). These 13 pathways included neuro-related pathways such as "Axon guidance", "Parkinson disease", and "Alzheimer disease". Furthermore, proteins belonging to the "pathway in cancer" also showed higher identity between tree shrews and humans than between mice and humans (Figure 5B), suggesting that the tree shrew could be used to create valid cancer and neurodegenerative animal models. We also profiled the expression patterns of five pathways related to the brain (Figure 5C) and found that mice had a more distant clustering pattern than tree shrews with primates. Collectively, these results suggest that tree shrews are more genetically similar to primates than to mice at the transcriptomic level.

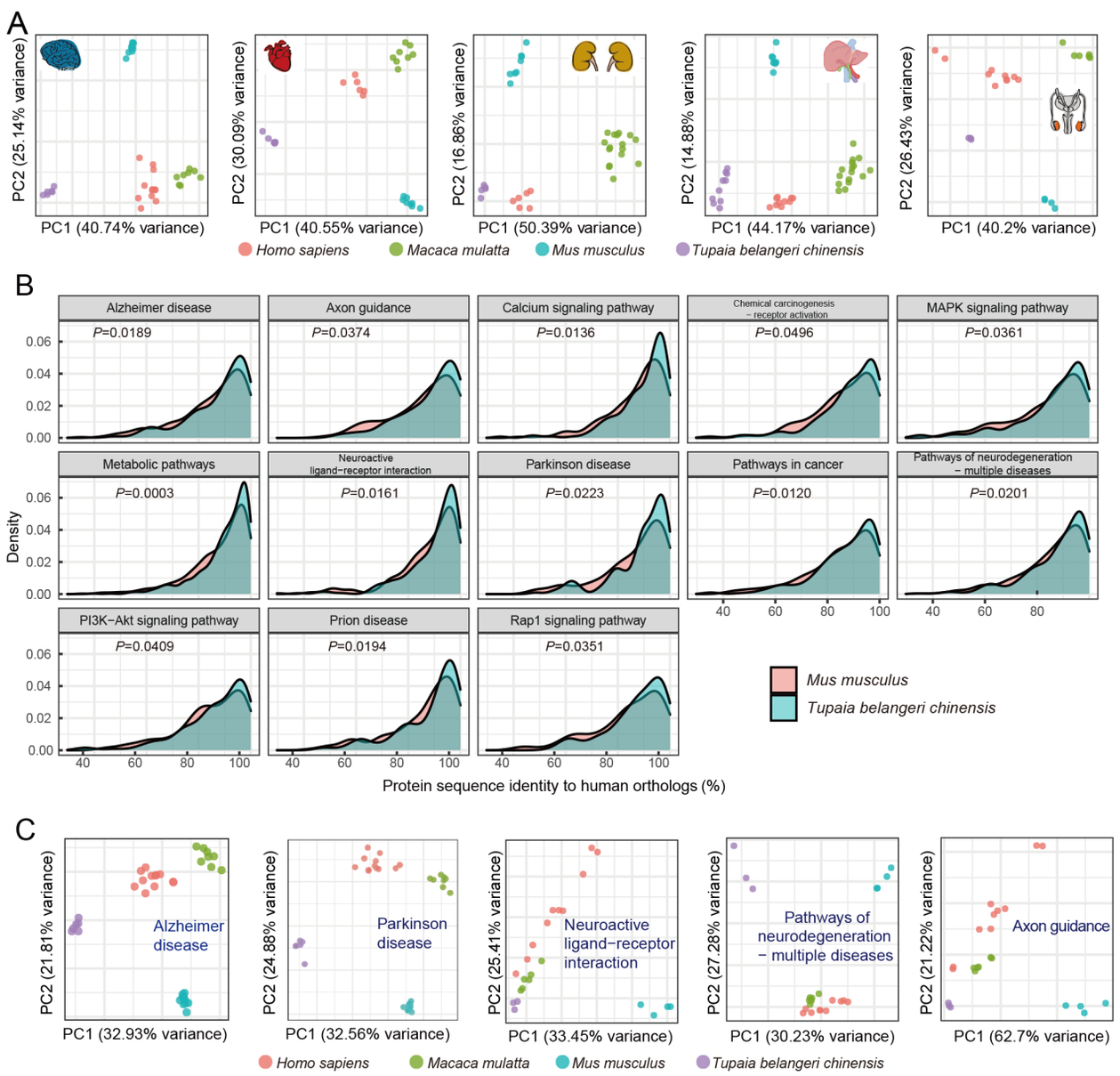

Figure 5 Expression similarities among different species

A: Tissue expression similarities among humans, rhesus monkeys, tree shrews, and mice. Expression patterns in five tree shrew tissues more closely resembled that of primates than that of mice. B: Comparisons of protein sequence identity of genes in KEGG pathways between tree shrews and humans and between mice and humans. C: Expression patterns of genes in brain-related pathways in tree shrews, rhesus monkeys, humans, and mice. Brain-related pathways included "Alzheimer's disease", "Parkinson disease", "Neuroactive ligand-receptor interaction", "Pathways of neurodegeneration-multiple diseases", and "Axon guidance".

Changes in newly identified genes upon viral infection

To profile the transcriptome patterns of host immune responses to viral infection using the TS_3.0 genome annotation, we focused on the differential expression of the newly identified genes upon viral infection. Results showed that the TS_3.0 genome annotation had better accuracy and resolution for gene identification in tree shrew cells with or without viral infection. Some of the newly annotated genes were significantly dysregulated upon infection with HBV (67 genes), SeV (99 genes), NDV (48 genes), EMCV (seven genes), and ZIKA (178 genes) (Figure 6A). Of the 3 779 DEGs identified from the RNA-seq datasets by comparing infected and uninfected cells, only purine rich element binding protein A (*PURA*) and interferon induced protein 35 (*IFI35*) were significantly dysregulated in cells with all virus infections. Both genes are reported to have a pro-viral effect in other species (Das et al., 2014; Gounder et al., 2018; Krachmarov et al., 1996). As *PURA* is a newly annotated gene in TS_3.0, more studies should be carried out to characterize its role in viral infections in tree shrews.

Among the expanded gene family members, 19 of the 39 copies of the *STT3B* gene family in the tree shrew were up-regulated upon SeV infection. The oligosaccharyltransferase complex is known to be an essential host factor for dengue virus (DENV) replication (Lin et al., 2017). Considering that the 19 *STT3B* gene copies were not located on the same pseudochromosome, we speculated that they were up-regulated by the same transcription regulation system. The

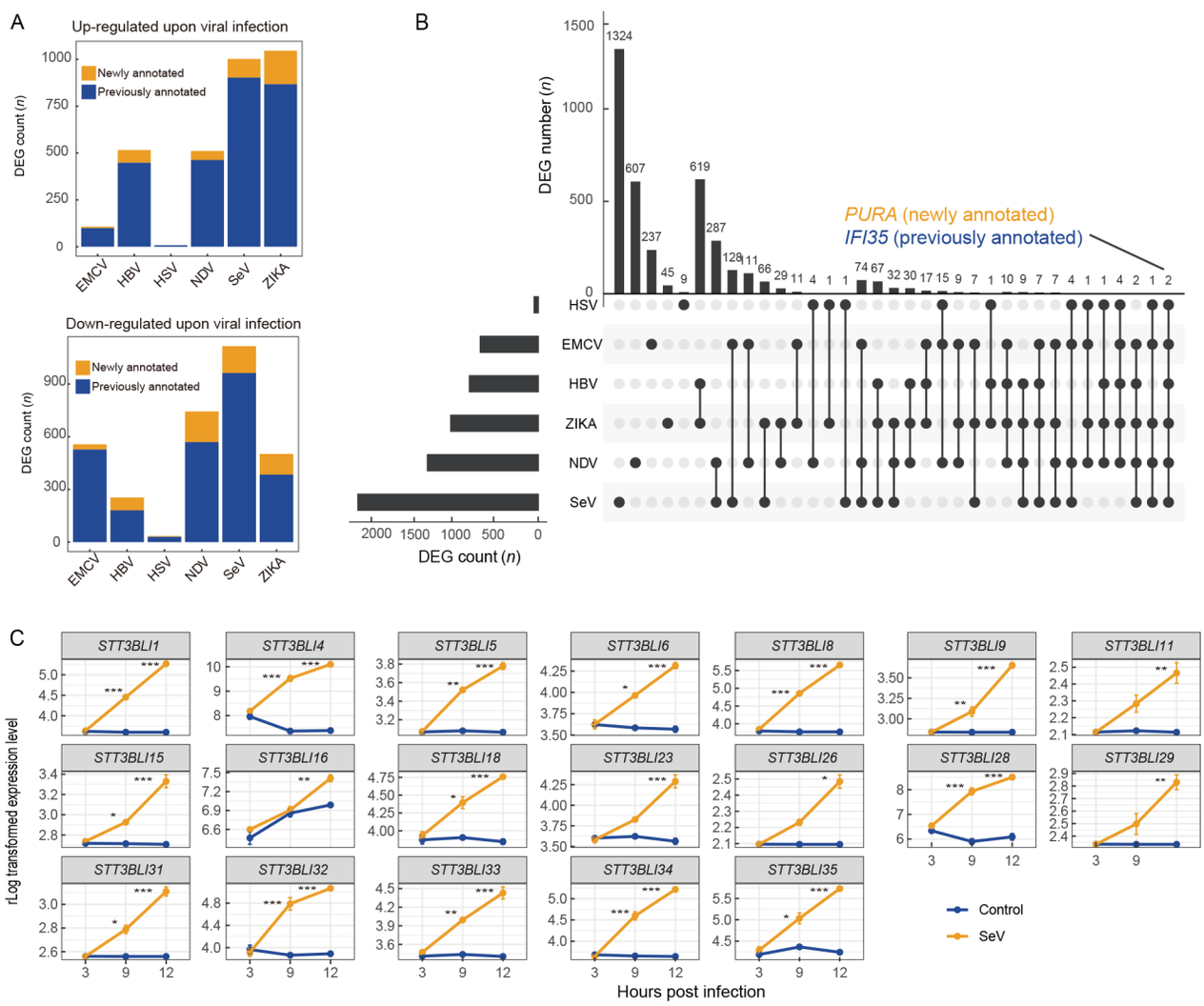

Figure 6 Changes in expression of genes in virus-infected tree shrew tissues and cells

A: Changes in newly annotated genes upon viral infection. HBV, hepatitis B virus; NDV, Newcastle disease virus; EMCV, encephalomyocarditis virus; HSV-1, herpes simplex virus type 1; SeV Sendai virus; ZIKA, Zika virus. Genes were identified as differentially expressed genes (DEGs) upon virus infection if P_{adjust}<0.05 and |log2 fold-change|>1. B: Plot of DEGs upon virus infection. *PURA* and *IFI35* were dysregulated in cells and tissues infected with different viruses. Horizontal bar on left represents number of DEGs in each RNA-seq dataset. Dots and lines represent subsets of DEGs. Vertical histogram represents number of DEGs in each subset. C: Changes in expression of 19 gene copies of *STT3B* gene family upon SeV infection. Results are mean±standard deviation (*SD*). *: P_{adjust}<0.05; **: P_{adjust}<0.005; ***, P_{adjust}<0.0005. P_{adjust} values were calculated using DESeq2.

ancestral gene of the *STT3B* gene copies, *STT3BLI27*, was not dysregulated upon viral infection, suggesting that the 19 *STT3B* gene copies may have acquired the association with pro-viral function at a later stage of gene family expansion. However, further studies are required to confirm this speculation.

DISCUSSION

Comprehensive tree shrew genome annotation is crucial for developing animal models and for studying basic scientific questions (Yao, 2017). In this study, we annotated the Chinese tree shrew genome by integrating diverse RNA-seq datasets and newly generated ISO-seq datasets. We obtained a total of 27 082 coding genes (including 3 514 previously unannotated coding genes in TS_2.0 (Fan et al., 2019)) and 56 401 lncRNAs. Evaluation of the completeness of multiple tree shrew genome annotations using BUSCO (Seppey et al., 2019; Simão et al., 2015; Waterhouse et al., 2018) indicated that the current TS_3.0 annotation showed remarkable improvement in terms of completeness, which was achieved by incorporating diverse RNA-seq datasets that covered a wide range of biological and pathological conditions. The newly updated tree shrew TS_3.0 genome annotation can be downloaded from the Tree shrew Database (http://www.treeshrewdb.org/download.html).

Compared with the previous tree shrew genome annotation (Table 1), TS_3.0 provides a complete list of lncRNAs, which could help in the interpretation of the roles of lncRNAs in tree shrew biology and disease. The lack of lncRNA conservation among species is a considerable obstacle for functional annotation (Iyer et al., 2015). Here, we compared multiple characteristics between tree shrew mRNAs and lncRNAs and found smaller average exon number and shorter transcript length in lncRNAs than in mRNAs. We confirmed that the identified tree shrew lncRNAs may exert a cis-regulatory role on mRNA expression (Figure 3D). The overall characteristics of the tree shrew lncRNAs versus mRNAs resembled that of human lncRNAs versus mRNAs reported in previous studies (Iyer et al., 2015; Jiang et al., 2019; Ørom et al., 2010; Ponjavic et al., 2009).

Alternative splicing plays a key role in transcript processing and biological functions (Baralle & Giudice, 2017; Ule & Blencowe, 2019). By generating multiple transcripts from a particular gene, alternative splicing events can dramatically increase the diversity and complexity of the transcriptome, and can impact mRNA stability, localization, and translation (Baralle & Giudice, 2017; Ule & Blencowe, 2019). We previously showed that alternative splicing events in STING have played an important role in the innate immunity response of tree shrews against DNA and RNA viral infections (Xu et al., 2020a). Our updated annotation of tree shrew transcripts, especially from SMRT reads, provided an accurate model to characterize alternative splicing events in tree shrews. Using the TS_3.0 transcripts, we quantified and characterized the alternative splicing events and found a tissue-specific pattern of splicing intensity. The high occurrence of alternative splicing events in the tree shrew brain-related tissues is consistent with that found in humans (Rodriguez et al., 2020), suggesting that alternative splicing constitutes a straightforward strategy for enacting diverse functions such as tissue formation (Baralle & Giudice, 2017). It would be worth performing functional characterization of important genes that exhibit alternative splicing in different tree shrew tissues and/or in response to viral infection in the TS_3.0 genome annotation, as exemplified by the elegant functional assay for the STING isoform described in our recent study (Xu et al., 2020a).

The updated tree shrew TS_3.0 genome annotation could also provide insightful information regarding cross-species comparisons to initiate genome-based methods for creating animal models of human disease (Yao, 2017). We systematically characterized the orthologous relationships among experimental animals, including mice, monkeys, and tree shrews, using the newly updated tree shrew genome annotation. Orthologous comparison confirmed the closer relationship between primates and tree shrews than between primates and mice (Figure 4A), suggesting that tree shrews would be better model animals for biomedical research. We also compared the tissue expression patterns and related genes in particular pathways across four species, which again showed that tree shrews are closer to primates than to mice at the transcriptomic level (Figure 5C).

Gene expansion and contraction play key roles in environment adaptation (Yim et al., 2014). We re-appraised gene expansion and contraction events using the TS_3.0 transcripts and confirmed the gene families highlighted in our previous study (Fan et al., 2013). Among the 144 gene families that experienced size changes in the tree shrew, the *IL6* and *STT3B* families may have particular biological implications. Notably, IL6 is thought to be actively involved in the cytokine storms observed in COVID-19 patients (Mehta et al., 2020; Vabret et al., 2020; Zhou et al., 2020) and therapy with the IL-6-receptor antagonist tocilizumab is considered a promising treatment for COVID-19 patients (Fu et al., 2020; Jones & Hunter, 2021). In SARS-CoV-2-infected tree shrews (Xu et al., 2020c; Zhao et al., 2020), different individuals demonstrated different susceptibility to SARS-CoV-2 and showed different viral loads after infection, though none of the infected tree shrews showed severe symptoms. Whether the expanded *IL6* gene family played a role in this process is an interesting and important question. Cloning all 13 *IL6* copies and characterizing the respective roles of each gene copy could help clarify why this gene family underwent expansion in the tree shrew. Among the 39 gene copies of the *STT3B* gene family, 19 were up-regulated upon SeV infection, whereas the other copies, including ancestral *STT3BLI27*, showed no such effect. The STT3B protein is a part of the oligosaccharyltransferase complex in humans (Lu et al., 2019), and is reported to play a pro-viral role in Dengue virus and HSV-1 infections (Lin et al., 2017; Lu et al., 2019). Expansion of the *STT3B* gene family may indicate a new immune response mechanism for tree shrews to counteract or facilitate these viral infections. However, more studies are required to characterize the function of the tree shrew *STT3B* gene family and to confirm the above speculation.

An important update of the TS_3.0 genome annotation was the inclusion of newly generated RNA-seq data from tree shrew cells and tissues challenged with different viruses. The

inclusion of these datasets offers the chance to identify genes that are up-regulated or down-regulated upon viral infection for further study. Indeed, previously reported tree shrew genes that show altered expression upon viral infection (Gu et al., 2019a, 2021; Xu et al., 2016, 2020a, 2020b) could be confirmed. We identified several important targets showing a universal regulator effect, such as *PURA* and *IFI35*. The *PURA* gene encodes Pur-alpha, which has a repeated nucleic acid binding domain (Daniel & Johnson, 2018), and is reported to be regulated by transcription start sites Ⅰ and Ⅱ (Wortman et al., 2010). PURA is known to activate the John Cunningham virus in the glial cells of many acquired immunodeficiency syndrome patients (Krachmarov et al., 1996). In addition, *IFI35* is an interferon-stimulated gene that negatively regulates RIG-I antiviral signals to support vesicular stomatitis viral replication (Das et al., 2014) and enhances H5N1 influenza disease symptoms (Gounder et al., 2018). We speculate that *in vivo* overexpression of both *PURA* and *IFI35* may create tree shrew models more permissive to different viruses, including HCV and HBV, which have no feasible animal models at present.

In summary, we generated an improved tree shrew genome annotation using comprehensive RNA-seq and ISO-seq datasets. The updated version of the tree shrew genome annotation (TS_3.0) fixed some of the issues with previous versions, such as TS_1.0 (Fan et al., 2013) and TS_2.0 (Fan et al., 2019). Detailed annotation of the genes, gene families, and alternative splicing events in the tree shrew genome, as well as cross-comparison of expression patterns among different tissues and species, further illuminated the unique and common genetic features of tree shrews and provided further evidence of the considerable potential of tree shrews in biomedical research.

DATA AVAILABILITY

The TS_3.0 genome annotation data and newly generated RNA-seq and ISO-seq data are available from the Tree shrew Database (http://www.treeshrewdb.org/download/). Related data were also deposited in GSA (accession No. PRJCA006366).

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Y.G.Y. and M.S.Y. conceived and designed the experiments. B.L.L. provided living tree shrews and tissues. D.D.Y. and L.X. isolated tree shrew primary cells and performed viral infection and RNA extraction. Q.Y.Z. provided the transcriptome data of lung tissues from IAV-infected tree shrews. M.S.Y., J.Y.Z., M.X., and Y.F. collected transcriptome data and performed genome annotation and transcriptome analyses. M.S.Y. and Y.G.Y. wrote the manuscript. All authors read and approved the final version of the manuscript.

REFERENCES

Amako Y, Tsukiyama-Kohara K, Katsume A, Hirata Y, Sekiguchi S, Tobita Y, et al. 2010. Pathogenesis of hepatitis C virus infection in *Tupaia belangeri*. *Journal of Virology*, **84**(1): 303–311.

Baralle FE, Giudice J. 2017. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, **18**(7): 437−451.

Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, et al. 2019. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. *BMC Genomics*, **20**(1): 344.

Bennett AJ, Panicker S. 2016. Broader impacts: international implications and integrative ethical consideration of policy decisions about US chimpanzee research. *American Journal of Primatology*, **78**(12): 1282−1303.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15): 2114−2120.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. 2007. UniProtKB/Swiss-Prot. *Methods in Molecular Biology*, **406**: 89−112.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5): 525−527.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, **36**(5): 411−420.

Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen CY, Shao Y, et al. 2019. Gene expression across mammalian organ development. *Nature*, **571**(7766): 505−509.

Chen G, Shi TL, Shi LM. 2017. Characterizing and annotating the genome using RNA-seq data. *Science China Life Sciences*, **60**(2): 116−125.

Dahariya S, Paddibhatla I, Kumar S, Raghuwanshi S, Pallepati A, Gutti RK. 2019. Long non-coding RNA: classification, biogenesis and functions in blood cells. *Molecular Immunology*, **112**: 82−92.

Daniel DC, Johnson EM. 2018. *PURA*, the gene encoding Pur-alpha, member of an ancient nucleic acid-binding protein family with mammalian neurological functions. *Gene*, **643**: 133−143.

Das A, Dinh PX, Panda D, Pattnaik AK. 2014. Interferon-inducible protein IFI35 negatively regulates RIG-I antiviral signaling and supports vesicular stomatitis virus replication. *Journal of Virology*, **88**(6): 3103−3113.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**(10): 1269−1271.

Dimanico MM, Klaassen AL, Wang J, Kaeser M, Harvey M, Rasch B, et al. 2021. Aspects of tree shrew consolidated sleep structure resemble human sleep. *Communications Biology*, **4**(1): 722.

Djureinovic D, Fagerberg L, Hallström B, Danielsson A, Lindskog C, Uhlén M, et al. 2014. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Molecular Human Reproduction*, **20**(6): 476−488.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1): 15−21.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research*, **47**(D1): D427−D432.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**(1): 238.

Fan Y, Huang ZY, Cao CC, Chen CS, Chen YX, Fan DD, et al. 2013. Genome of the Chinese tree shrew. *Nature Communications*, **4**: 1426.

Fan Y, Luo RC, Su LY, Xiang Q, Yu DD, Xu L, et al. 2018. Does the genetic feature of the Chinese tree shrew (*Tupaia belangeri chinensis*) support its potential as a viable model for Alzheimer's disease research?. *Journal of Alzheimer's Disease*, **61**(3): 1015−1028.

Fan Y, Ye MS, Zhang JY, Xu L, Yu DD, Gu TL, et al. 2019. Chromosomal level assembly and population sequencing of the Chinese tree shrew genome. *Zoological Research*, **40**(6): 506−521.

Fan Y, Yu DD, Yao YG. 2014. Tree shrew database (TreeshrewDB): a genomic knowledge base for the Chinese tree shrew. *Scientific Reports*, **4**: 7145.

Fitzpatrick D. 1996. The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex. *Cerebral Cortex*, **6**(3): 329−341.

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, et al. 2019. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biology*, **17**(1): 108.

Fu BQ, Xu XL, Wei HM. 2020. Why tocilizumab could be an effective treatment for severe COVID-19?. *Journal of Translational Medicine*, **18**(1): 164.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, **8**(6): 469−477.

Ge GZ, Xia HJ, He BL, Zhang HL, Liu WJ, Shao M, et al. 2016. Generation and characterization of a breast carcinoma model by PyMT overexpression in mammary epithelial cells of tree shrew, an animal close to primates in evolution. *International Journal of Cancer*, **138**(3): 642−651.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science*, **352**(6281): aae0344.

Gounder AP, Yokoyama CC, Jarjour NN, Bricker TL, Edelson BT, Boon ACM. 2018. Interferon induced protein 35 exacerbates H5N1 influenza disease through the expression of IL-12p40 homodimer. *PLoS Pathogens*, **14**(4): e1007001.

Gu TL, Yu DD, Fan Y, Wu Y, Yao YL, Xu L, et al. 2019a. Molecular identification and antiviral function of the guanylate-binding protein (GBP) genes in the Chinese tree shrew (*Tupaia belangeri chinesis*). *Developmental & Comparative Immunology*, **96**: 27−36.

Gu TL, Yu DD, Li Y, Xu L, Yao YL, Yao YG. 2019b. Establishment and characterization of an immortalized renal cell line of the Chinese tree shrew (*Tupaia belangeri chinesis*). *Applied Microbiology and Biotechnology*, **103**(5): 2171−2180.

Gu TL, Yu DD, Xu L, Yao YL, Zheng X, Yao YG. 2021. *Tupaia* guanylate-binding protein 1 interacts with vesicular stomatitis virus phosphoprotein and represses primary transcription of the viral genome. *Cytokine*, **138**: 155388.

Han YY, Wang WG, Jia J, Sun XM, Kuang DX, Tong PF, et al. 2020. WGCNA analysis of the subcutaneous fat transcriptome in a novel tree shrew model. *Experimental Biology and Medicine*, **245**(11): 945−955.

He L, Frost MR, Siegwart JT Jr, Norton TT. 2014. Gene expression signatures in tree shrew choroid during lens-induced myopia and recovery. *Experimental Eye Research*, **123**: 56−71.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, **34**(8): 2115−2122.

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, **47**(3): 199−208.

Ji XJ, Li P, Fuscoe JC, Chen G, Xiao WZ, Shi LM, et al. 2020. A comprehensive rat transcriptome built from large scale RNA-seq-based annotation. *Nucleic Acids Research*, **48**(15): 8320−8331.

Jiang S, Cheng SJ, Ren LC, Wang Q, Kang YJ, Ding Y, et al. 2019. An expanded landscape of human long noncoding RNA. *Nucleic Acids Research*, **47**(15): 7842−7856.

Jones SA, Hunter CA. 2021. Is IL-6 a key cytokine target for therapy in COVID-19?. *Nature Reviews Immunology*, **21**(6): 337−339.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, **45**(D1): D353−D361.

Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei LP, et al. 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, **45**(W1): W12−W16.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei LP, et al. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, **35**(W1): W345−W349.

Krachmarov CP, Chepenik LG, Barr-Vagell S, Khalili K, Johnson EM. 1996. Activation of the JC virus Tat-responsive transcriptional control element by association of the Tat protein of human immunodeficiency virus 1 with cellular protein Purα. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(24): 14112−14117.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, **35**(6): 1547−1549.

Lee KS, Huang XY, Fitzpatrick D. 2016. Topology of ON and OFF inputs in visual cortex enables an invariant columnar architecture. *Nature*, **533**(7601): 90−94.

Levy AM, Fazio MA, Grytz R. 2018. Experimental myopia increases and scleral crosslinking using genipin inhibits cyclic softening in the tree shrew sclera. *Ophthalmic and Physiological Optics*, **38**(3): 246−256.

Li CH, Yan LZ, Ban WZ, Tu Q, Wu Y, Wang L, et al. 2017. Long-term propagation of tree shrew spermatogonial stem cells in culture and successful generation of transgenic offspring. *Cell Research*, **27**(2): 241−252.

Li RF, Yuan B, Xia XS, Zhang S, Du QL, Yang CG, et al. 2018. Tree shrew as a new animal model to study the pathogenesis of avian influenza (H9N2) virus infection. *Emerging Microbes & Infections*, **7**(1): 166.

Lin DL, Cherepanova NA, Bozzacco L, MacDonald MR, Gilmore R, Tai AW. 2017. Dengue virus hijacks a noncanonical oxidoreductase function of a cellular oligosaccharyltransferase complex. *mBio*, **8**(4): e00939−e00917.

Lin JN, Chen GF, Gu L, Shen YF, Zheng MZ, Zheng WS, et al. 2014. Phylogenetic affinity of tree shrews to Glires is attributed to fast evolution rate. *Molecular Phylogenetics and Evolution*, **71**: 193−200.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**(12): 550.

Lu H, Cherepanova NA, Gilmore R, Contessa JN, Lehrman MA. 2019. Targeting STT3A-oligosaccharyltransferase with NGI-1 causes herpes simplex virus 1 dysfunction. *The FASEB Journal*, **33**(6): 6801−6812.

Lu T, Peng HM, Zhong LP, Wu P, He J, Deng ZM, et al. 2021. The tree shrew as a model for cancer research. *Frontiers in Oncology*, **11**: 653236.

Luo MT, Fan Y, Mu D, Yao YG, Zheng YT. 2018. Molecular cloning and

characterization of *APOBEC3* family in tree shrew. *Gene*, **646**: 143−152.
McGonigle P, Ruggeri B. 2014. Animal models of human disease: challenges in enabling translation. *Biochemical Pharmacology*, **87**(1): 162−171.
McInnes L, Healy J, Melville J. 2020. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv: 1802.03426.
Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ, et al. 2020. COVID-19: consider cytokine storm syndromes and immunosuppression. *The Lancet*, **395**(10229): 1033−1034.
Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. 2017. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Research*, **27**(1): 27−37.
Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2020. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*: btaa1022
Ni RJ, Huang ZH, Luo PH, Ma XH, Li T, Zhou JN. 2018. The tree shrew cerebellum atlas: systematic nomenclature, neurochemical characterization, and afferent projections. *Journal of Comparative Neurology*, **526**(17): 2744−2775.
Nudelman G, Frasca A, Kent B, Sadler KC, Sealfon SC, Walsh MJ, et al. 2018. High resolution annotation of zebrafish transcriptome using long-read sequencing. *Genome Research*, **28**(9): 1415−1425.
Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**(1): 46−58.
Page RD. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**(9): 819−820.
Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Research*, **9**: 304.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3): 290−295.
Petry HM, Bickford ME. 2019. The second visual system of the tree shrew. *Journal of Comparative Neurology*, **527**(3): 679−693.
Phillips JR, Khalaj M, McBrien NA. 2000. Induced myopia associated with increased scleral creep in chick and tree shrew eyes. *Investigative Ophthalmology & Visual Science*, **41**(8): 2028−2034.
Ponjavic J, Oliver PL, Lunter G, Ponting CP. 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genetics*, **5**(8): e1000617.
Purugganan MD, Jackson SA. 2021. Advancing crop genomics from lab to field. *Nature Genetics*, **53**(5): 595−601.
Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, **17**(1): 47−62.
Robinson NB, Krieger K, Khan FM, Huffman W, Chang M, Naik A, et al. 2019. The current state of animal models in research: a review. *International Journal of Surgery*, **72**: 9−13.
Rodriguez JM, Pozo F, di Domenico T, Vazquez J, Tress ML. 2020. An analysis of tissue-specific alternative splicing at the protein level. *PLoS Computational Biology*, **16**(10): e1008287.
Rose-John S. 2018. Interleukin-6 family cytokines. *Cold Spring Harbor Perspectives in Biology*, **10**(2): a028415.
Salmela L, Rivals E. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, **30**(24): 3506−3514.
Sanada T, Tsukiyama-Kohara K, Shin IT, Yamamoto N, Kayesh MEH, Yamane D, et al. 2019. Construction of complete *Tupaia belangeri* transcriptome database by whole-genome and comprehensive RNA sequencing. *Scientific Reports*, **9**(1): 12372.
Savier E, Sedigh-Sarvestani M, Wimmer R, Fitzpatrick D. 2021. A bright future for the tree shrew in neuroscience research: summary from the inaugural Tree Shrew Users Meeting. *Zoological Research*, **42**(4): 478−481.
Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods in Molecular Biology*, **1962**: 227−245.
Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, **31**(11): 1009−1014.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19): 3210−3212.
Stark R, Grzelak M, Hadfield J. 2019. RNA sequencing: the teenage years. *Nature Reviews Genetics*, **20**(11): 631−656.
Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, **19**(1): 40.
Tu Q, Yang D, Zhang XN, Jia XT, An SQ, Yan LZ, et al. 2019. A novel pancreatic cancer model originated from transformation of acinar cells in adult tree shrew, a primate-like animal. *Disease Models & Mechanisms*, **12**(4): dmm038703.
Ule J, Blencowe BJ. 2019. Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Molecular Cell*, **76**(2): 329−345.
Vabret N, Britton GJ, Gruber C, Hegde S, Kim J, Kuksin M, et al. 2020. Immunology of COVID-19: current state of the science. *Immunity*, **52**(6): 910−941.
Wang GL, Park HJ, Dasari S, Wang SQ, Kocher JP, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, **41**(6): e74.
Wang K, Wang DH, Zheng XM, Qin A, Zhou J, Guo BY, et al. 2019. Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nature Communications*, **10**(1): 4714.
Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, **35**(3): 543−548.
Wei S, Hua HR, Chen QQ, Zhang Y, Chen F, Li SQ, et al. 2017. Dynamic changes in DNA demethylation in the tree shrew (*Tupaia belangeri chinensis*) brain during postnatal development and aging. *Zoological Research*, **38**(2): 96−102.
Wortman MJ, Hanson LK, Martínez-Sobrido L, Campbell AE, Nance JA, García-Sastre A, et al. 2010. Regulation of *PURA* gene transcription by three promoters generating distinctly spliced 5-prime leaders: a novel means of fine control over tissue specificity and viral signals. *BMC Molecular Biology*, **11**: 81.
Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016a. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology*, **1418**: 283−334.
Wu XY, Xu HB, Zhang ZG, Chang Q, Liao SS, Zhang LQ, et al. 2016b. Transcriptome profiles using next-generation sequencing reveal liver changes in the early stage of diabetes in tree shrew (*Tupaia belangeri chinensis*). *Journal of Diabetes Research*, **2016**: 6238526.
Xu L, Yu DD, Fan Y, Peng L, Wu Y, Yao YG. 2016. Loss of RIG-I leads to a functional replacement with MDA5 in the Chinese tree shrew. *Proceedings*

of the National Academy of Sciences of the United States of America, **113**(39): 10950−10955.

Xu L, Yu DD, Ma YH, Yao YL, Luo RH, Feng XL, et al. 2020c. COVID-19-like symptoms observed in Chinese tree shrews infected with SARS-CoV-2. *Zoological Research*, **41**(5): 517−526.

Xu L, Yu DD, Peng L, Wu Y, Fan Y, Gu TL, et al. 2020a. An alternative splicing of *tupaia* STING modulated anti-RNA virus responses by targeting MDA5-LGP2 and IRF3. *The Journal of Immunology*, **204**(12): 3191−3204.

Xu L, Yu DD, Yao YL, Gu TL, Zheng X, Wu Y, et al. 2020b. *Tupaia* MAVS is a dual target during hepatitis c virus infection for innate immune evasion and viral replication via NF-κB. *The Journal of Immunology*, **205**(8): 2091−2099.

Xu XP, Chen HB, Cao XM, Ben KL. 2007. Efficient infection of tree shrew (*Tupaia belangeri*) with hepatitis C virus grown in cell culture or from patient plasma. *The Journal of General Virology*, **88**(Pt9): 2504−2512.

Yan H, Zhong GC, Xu GW, He WH, Jing ZY, Gao ZC, et al. 2012. Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. *eLife*, **1**: e00049.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, **13**(5): 329−342.

Yao YG. 2017. Creating animal models, why not use the Chinese tree shrew (*Tupaia belangeri chinensis*)?. *Zoological Research*, **38**(3): 118−126.

Yao YG, Chen YB, Liang B. 2015. The 3rd symposium on animal models of primates - the application of non-human primates to basic research and translational medicine. *Journal of Genetics and Genomics*, **42**(6): 339−341.

Yao YL, Yu DD, Xu L, Fan Y, Wu Y, Gu TL, et al. 2019. Molecular characterization of the 2', 5'-oligoadenylate synthetase family in the Chinese tree shrew (*Tupaia belangeri chinensis*). *Cytokine*, **114**: 106−114.

Yim HS, Cho YS, Guang XM, Kang SG, Jeong JY, Cha SS, et al. 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*, **46**(1): 88−92.

Yu DD, Wu Y, Xu L, Fan Y, Peng L, Xu M, et al. 2016. Identification and characterization of toll-like receptors (TLRs) in the Chinese tree shrew (*Tupaia belangeri chinensis*). *Developmental & Comparative Immunology*, **60**: 127−138.

Yu DD, Xu L, Liu XH, Fan Y, Lü LB, Yao YG. 2014. Diverse interleukin-7 mRNA transcripts in Chinese tree shrew (*Tupaia belangeri chinensis*). *PLoS One*, **9**(6): e99859.

Yu GC, Wang LG, Han YY, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS:A Journal of Integrative Biology*, **16**(5): 284−287.

Zhang P, Chen JS, Li QY, Sheng LX, Gao YX, Lu BZ, et al. 2020a. Neuroprotectants attenuate hypobaric hypoxia-induced brain injuries in cynomolgus monkeys. *Zoological Research*, **41**(1): 3−19.

Zhang XM, Yu DD, Wu Y, Gu TL, Ma N, Dong SZ, et al. 2020b. Establishment and transcriptomic features of an immortalized hepatic cell line of the Chinese tree shrew. *Applied Microbiology and Biotechnology*, **104**(20): 8813−8823.

Zhao Y, Wang JB, Kuang DX, Xu JW, Yang ML, Ma CX, et al. 2020. Susceptibility of tree shrew to SARS-CoV-2 infection. *Scientific Reports*, **10**(1): 16007.

Zheng YT, Yao YG, Xu L. 2014. Basic Biology and Disease Models of Tree Shrews. Kunming: Yunnan Science and Technology Press, 1–475. (in Chinese)

Zhou YG, Fu BQ, Zheng XH, Wang DS, Zhao CC, Qi YJ, et al. 2020. Pathogenic T-cells and inflammatory monocytes incite inflammatory storms in severe COVID-19 patients. *National Science Review*, **7**(6): 998−1002.

Figure S1. Protein sequence alignment of the tree shrew IL7. The red sequence IDs indicate sequences obtained by molecular cloning and deposited in GenBank in our previous study (Yu et al., 2014), and the blue sequence IDs indicate sequences obtained by using the TS_3.0 genome annotation.

Figure S2. Biological process enrichment of testis-specific genes

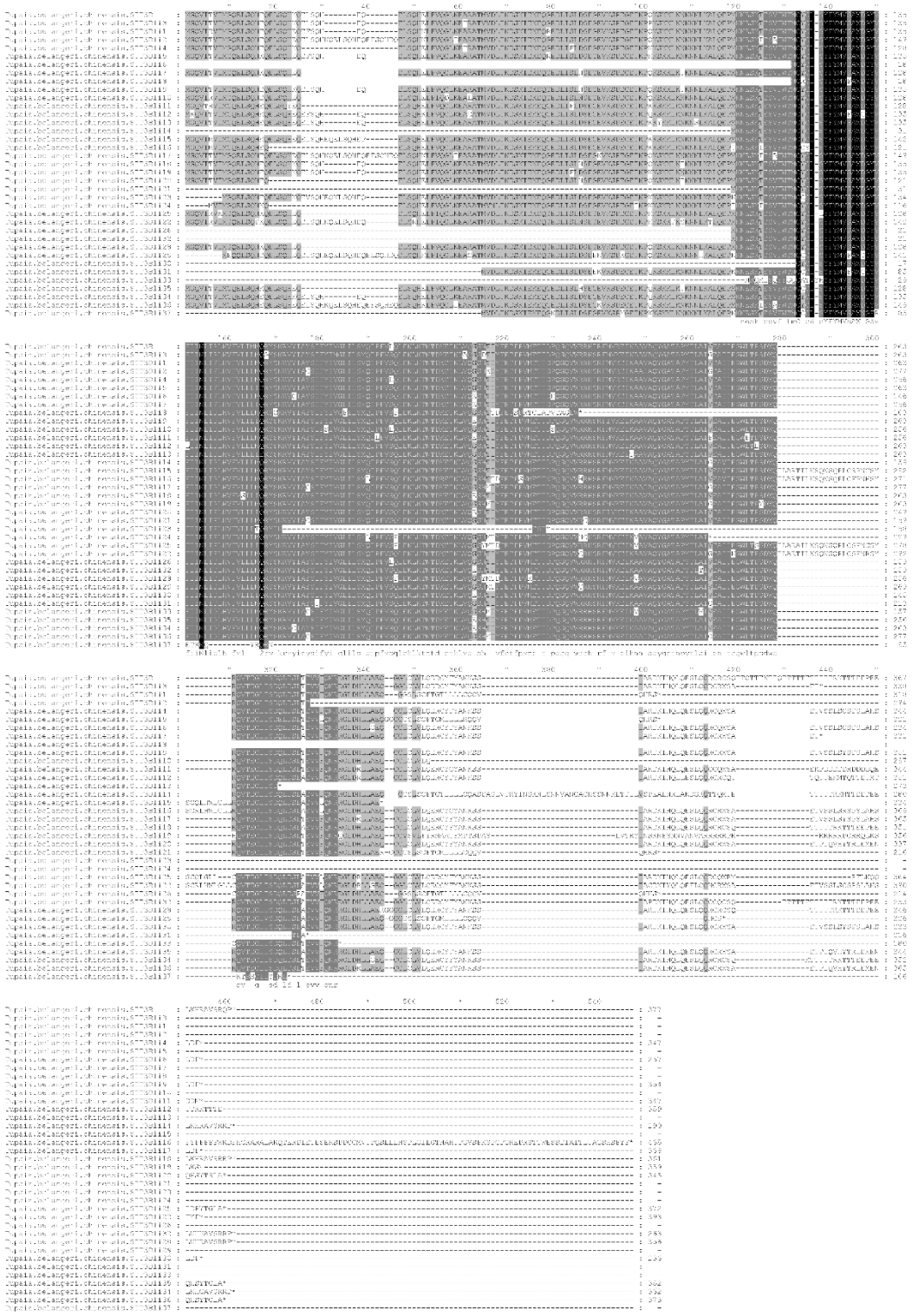

Figure S3. Protein sequence alignment of tree shrew STT3B gene family. The longest protein sequence for each gene family member was chosen for the alignment.

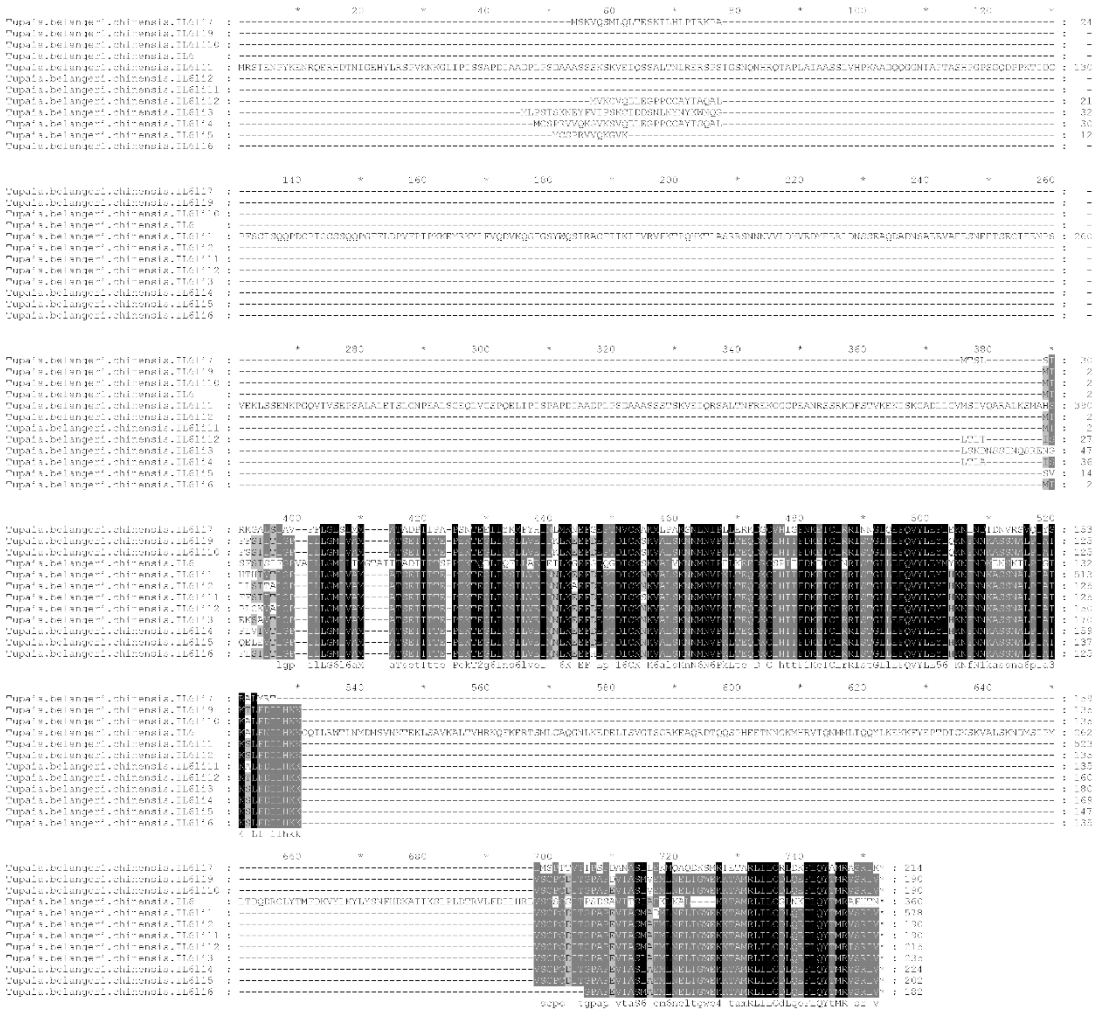

Figure S4. Protein sequence alignment of tree shrew IL6 gene family. The longest protein sequence for each gene family member was chosen for the alignment.

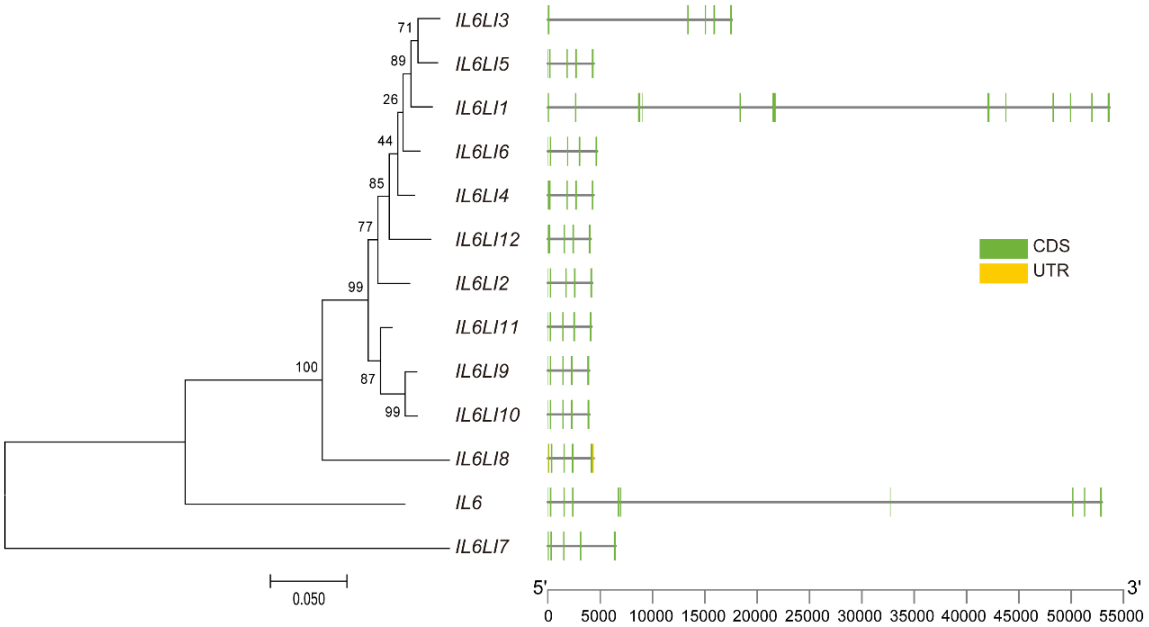

Figure S5. Gene structure and phylogenetic relationship among tree shrew *IL6* gene family members. (*left*) The maximum likelihood tree was constructed using MEGA7, with the ML methods (K2+G model) and 1,000 bootstraps. (*right*) Gene structure based on the TS_3.0 genome annotation. CDS, coding sequence; UTR, untranslated region.

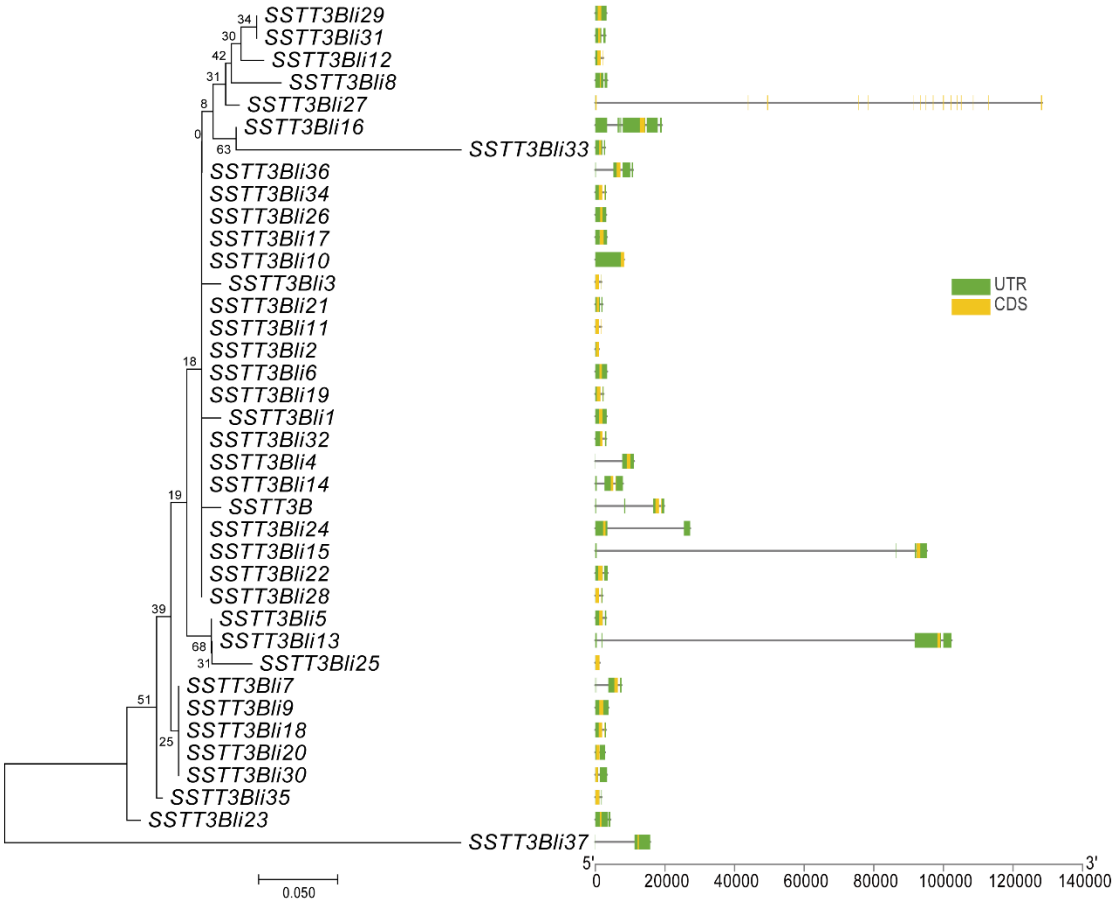

Figure S6. Gene structure and phylogenetic relationship among tree shrew *STT3B* family members. (*left*) The maximum likelihood tree was constructed using MEGA7, with the ML methods (K2+G model) and 1,000 bootstraps. (*right*) Gene structure based on the TS_3.0 genome annotation. CDS, coding sequence; UTR, untranslated region.

Table S1. Information of publicly available tree shrew RNA-seq datasets

Library type	Description	No. of datasets	Source	Accession number	No. of datasets after QC	Reference
Poly(A)$^+$	Pancreatic cancer tissues	3	GEO	PRJNA434450	3	(Tu et al., 2019)
Poly(A)$^+$	Tree shrew adipose tissues	12	GEO	PRJNA310673	12	(Han et al., 2020)
Poly(A)$^+$	Tree shrew liver tissues	2	GEO	PRJNA282350	2	(Wu et al., 2016)
Poly(A)$^+$	Tree shrew PBMC	6	GEO	PRJNA473782	6	-
Poly(A)$^+$	Tree shrew brain RNA-seq	6	GEO	PRJNA416241	6	(Fan et al., 2018)
Poly(A)$^+$	Tree shrew multi organ RNA-seq	7	GEO	PRJNA170104	7	(Fan et al., 2013)
Poly(A)$^+$	Tree shrew immortalized hepatocytes cell lines	21	GSA	PRJCA002538	21	(Zhang et al., 2020)
Poly(A)$^+$	Tree shrew hepatocytes RNA-seq	7	GEO	PRJNA87013	7	(Yan et al., 2012)
Poly(A)$^+$	differentiating germ cell and cell line	10	TSDB	TSDB2017R01- TSDB2017R15	6	(Li et al., 2017)
Poly(A)$^+$	Mixed tissues	1	GEO	PRJNA200450	1	(Lin et al., 2014)
Poly(A)$^+$ & Ribo$^-$	HBV infected tree shrew liver	21	DDBJ	DRR155071-DRR155099	21	(Sanada et al., 2019)
Poly(A)$^+$	Tree shrew glioblastoma	5	GSA	PRJCA000125	5	(Tong et al., 2017)

Poly(A)$^+$, poly A enriched library construction; Ribo$^-$, ribosome depletion library construction. PBMC, peripheral blood mononuclear cell; HBV, hepatitis B virus; GEO, gene expression omnibus (https://www.ncbi.nlm.nih.gov/geo/); GSA, genome sequence archive (https://ngdc.cncb.ac.cn/gsa/); TSDB, tree shrew database (http://www.treeshrewdb.org/); DDBJ, DNA Data Bank of Japan (https://www.ddbj.nig.ac.jp/index-e.html).

Table S2. Sample information for different tree shrew transcriptome datasets

ID	Gender	Age (months)	Tissues	Sequencing strategy	No. of datasets
27d2	female	21	Brain, Cortex, Hippocampus, Heart, Liver, Spleen, Kidney, Small intestine, Ovary	RNA-seq & ISO-seq (pooled samples)	9
U23	male	29	Testis	RNA-seq	1
U22	female	17	Brain, Cortex, Hippocampus, Heart, Liver, Spleen, Lung, Kidney, Small intestine, Muscle, Ovary	RNA-seq & ISO-seq (pooled samples)	11
16c7	male	32	Brain, Cortex, Hippocampus, Heart, Liver, Spleen, Lung, Kidney, Small intestine, Muscle, Testis	RNA-seq	11
16c9	male	28	Brain, Cortex, Hippocampus, Heart, Liver, Spleen, Lung, Kidney, Small intestine, Muscle, Testis	RNA-seq	11
21	male	67	Liver	Ribo$^-$ RNA-seq	1
28a4	male	63	Liver	Ribo$^-$ RNA-seq	1
U31	male	6	Liver	Ribo$^-$ RNA-seq	1
U32	male	7	Liver	Ribo$^-$ RNA-seq	1

ISO-seq, long-read isoform sequencing; Ribo$^-$, ribosome depletion RNA-SEQ library construction.

Table S3. Information of RNA-seq datasets of virus-infected cells and tissues

Library type	Description	No. of datasets	Experiment design
Poly(A)$^+$	HCV infected TPH	18	HCV infected and uninfected (NC) samples were collected at 24, 48, and 72 hours post infection
Poly(A)$^+$	NDV infected TSPRC	18	NDV infected and uninfected (NC) samples were collected at 3, 9, and 12 hours post infection
Poly(A)$^+$	EMCV infected TSPRC	9	EMCV infected and uninfected (NC) samples were collected at 3, 9, and 12 hours post infection
Poly(A)$^+$	HSV-1 infected TSPRC	8	HSV-1 infected and uninfected (NC) samples was collected at 6 and 12 hours post infection
Poly(A)$^+$	SeV infected TSPRC	9	SeV infected and uninfected (NC) samples were collected at 3, 9, and 12 hours post infection
Ribo$^-$	IAV infected lung	24	Sample were collected at 0, 3, and 5 days post IAV infection

Poly(A)$^+$, poly A enriched library construction; Ribo$^-$, ribosome depletion library construction.
TSPRC, tree shrew primary renal cells. TPH, Tree shrew primary hepatocytes.
HCV, Hepatitis C virus; NDV, Newcastle disease virus; EMCV, Encephalomyocarditis virus; HSV-1, herpes simplex virus type 1; SeV, Sendai virus; IAV, influenza A virus.

Table S4. Sequences obtained by molecular cloning

Gene	GenBank accession number	Reference
GBP1	AG214685.1	(Gu et al., 2019)
GBP2	AG214686.1	
GBP4	AG214687.1	
GBP5	AG214688.1	
GBP7	AG214689.1	
IL7 (variant 1)	AFA42954.1	(Yu et al., 2014)
IL7 (variant 2)	AFA42955.1	
IL7 (variant 3)	AFA42956.1	
IL7 (variant 4)	AFA42957.1	
IL7 (variant 5)	AFA42958.1	
IL7 (variant 6)	AFA42959.1	
TLR1	GT354316.1	(Yu et al., 2016)
TLR2	GT354317.1	
TLR3	GT354318.1	
TLR4	GT354319.1	
TLR5	GT354320.1	
TLR6	GT354321.1	
TLR7	GT354322.1	
TLR8	GT354323.1	
TLR9	GT354324.1	
TLR11	GT354325.1	
TLR12	GT354326.1	
OAS1	MH512001	(Yao et al., 2019)
OAS2	MH512002	
OASL1	MH512003	
OASL2	MH512004	
APOBEC3_A3A	KU053484.1	(Luo et al., 2018)
APOBEC3_A3C	KU053485.2	
APOBEC3_A3F	KU053486.3	
APOBEC3_A3G	KU053487.4	

GBP, guanylate binding protein; *IL7*, interleukin 7; *TLR*, toll like receptor; *OAS*, 2'-5'-oligoadenylate synthetase; *OASL*, 2'-5'-Oligoadenylate synthetase like; *APOBEC3*, apolipoprotein b mRNA editing enzyme catalytic subunit 3.

Table S5. Genome information of 11 vertebrate species used in this study

Species	Genome version	Weblink
Homo sapiens	GRCh38.p13	https://asia.ensembl.org/Homo_sapiens/Info/Index
Pan troglodytes	Pan_tro_3.0	https://asia.ensembl.org/Pan_troglodytes/Info/Index
Gorilla gorilla gorilla	gorGor4	https://asia.ensembl.org/Gorilla_gorilla/Info/Index
Macaca mulatta	Mmul_10	https://asia.ensembl.org/Macaca_mulatta/Info/Index
Rattus norvegicus	Rnor_6.0	https://asia.ensembl.org/Rattus_norvegicus/Info/Index
Mus musculus	GRCm39	https://asia.ensembl.org/Mus_musculus/Info/Index
Sus scrofa	Sscrofa11.1	https://asia.ensembl.org/Sus_scrofa/Info/Index?db=core
Bos taurus	ARS-UCD1.2	https://asia.ensembl.org/Bos_taurus/Info/Index
Canis lupus familiaris	CamFam3.1	https://asia.ensembl.org/Canis_lupus_familiaris/Info/Index
Oryctolagus cuniculus	OryCun2.0	https://asia.ensembl.org/Oryctolagus_cuniculus/Info/Index

Table S6. Assessment of annotation completeness in Chinese tree shrew using *Eukaryota* BUSCOs and *mammalian* BUSCOs

Version	Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs
***Eukaryota* BUSCOs**					
TS_1.0	221	210	11	20	14
TS_2.0	235	228	7	8	12
TS_3.0	250	236	14	4	1
Tupchi_1.0	216	207	9	16	23
TupaiaBase	195	187	8	30	30
Human	253	246	7	1	1
Mouse	250	244	6	2	3
***mammalian* BUSCOs**					
TS_1.0	7568	7484	84	666	992
TS_2.0	7519	7457	62	332	1375
TS_3.0	8559	8250	177	304	363
Tupchi_1.0	7884	7814	70	453	889
TupaiaBase	6080	6023	57	640	2506
Human	8931	8810	121	31	264
Mouse	8916	8839	77	18	292

BUSCO: Benchmarking with Universal Single-Copy Orthologs (Seppey et al., 2019; Simão et al., 2015; Waterhouse et al., 2018). A total of 255 benchmarking universal single-copy orthologs of the *Eukaryota* dataset were retrieved from BUSCO. The three versions of tree shrew genome were established by our group: TS_1.0 (Fan et al., 2013), TS_2.0 (Fan et al., 2019), and TS_3.0 (current version of tree shrew annotation). Tupchi_1.0, NCBI version of tree shrew annotation (https://www.ncbi.nlm.nih.gov/assembly/GCF_000334495.1/). The TupaiaBase was established by (Sanada et al., 2019). Human (https://asia.ensembl.org/Homo_sapiens/Info/Index) and Mouse (https://asia.ensembl.org/Mus_musculus/Info/Index/) datasets were taken from the public sources.

Table S7. Number of orthologs among 11 species

	Chimpanzee	Cattle	Dog	Gorilla	Human	Rhesus monkey	Mouse	Pig	Rabbit	Rat	Tree shrew
Chimpanzee	-	19 574	19 597	22 487	20 380	20 370	19 930	20 034	19 098	19 656	19 268
Cattle	14 367	-	17 658	17 541	17 835	16 615	18 014	18 246	17 129	17 717	17 201
Dog	14 323	15 187	-	17 385	17 568	16 595	17 689	17 749	16 879	17 413	17 009
Gorilla	17 545	14 445	14 430	-	19 583	19 696	19 417	19 545	18 646	19 142	18 696
Human	16 616	15 313	15 286	16 580	-	17 761	18 291	17 981	17 008	17 791	17 299
Rhesus monkey	15 359	13 991	13 952	15 395	15 366	-	17 961	18 139	17 347	17 744	17 435
Mouse	14 649	15 377	15 199	14 729	15 746	14 175	-	18 401	17 552	20 047	17 736
Pig	14 198	14 916	14 695	14 296	15 085	13 832	14 941	-	19 818	20 105	19 570
Rabbit	13 147	13 985	13 851	13 234	14 061	12 887	14 245	13 502	-	16 820	16 333
Rat	14 149	14 862	14 665	14 171	15 055	13 705	16 579	14 445	13 758	-	17 650
Tree shrew	13 791	14 246	14 165	13 830	14 549	13 390	14 530	13 873	13 040	14 015	-

Below the diagonal, number of one-one orthologue between each species pair; above the diagonal, number of total orthologues between each species pair (one-one, one-many, many-one). The data sources for Chimpanzee (*Pan troglodytes*), Cattle (*Bos taurus*), Dog (*Canis lupus familiaris*), Gorilla (*Gorilla gorilla gorilla*), Human (*Homo sapiens*), Rhesus monkey (*Macaca mulatta*), Mouse (*Mus musculus*), Pig (*Sus scrofa*), Rabbit (*Oryctolagus cuniculus*), and Rat (*Rattus norvegicus*) are listed in Table S5. The Chinese tree shrew (*Tupaia belangeri chinensis*) data are generated in this study.

Table S8. Gene family size changes along the phylogenic tree of 11 mammal species

Nodes	Expansions	Gene gain	Equal	Contractions	Gene loss	Sig Exp	Sig Contra
<11>	582	911	20746	866	1011	86	15
Pan troglodytes <6>	1591	2978	19996	607	657	329	27
Oryctolagus cuniculus <20>	866	1906	14980	6348	6537	53	13
Bos taurus <0>	679	1397	19613	1902	2072	51	16
Macaca mulatta <12>	1113	2179	17956	3125	3192	129	19
Gorilla gorilla gorilla <10>	1408	2518	19612	1174	1235	255	30
<3>	46	52	18533	3615	3632	3	7
<1>	205	323	21609	380	398	32	5
<13>	69	87	21180	945	958	4	8
Tupaia belangeri chinensis <14>	1610	5217	16307	4277	4553	120	22
<17>	468	1140	17649	4077	4216	104	20
Homo sapiens <8>	527	920	19065	2602	3094	76	260
<7>	137	217	21514	543	587	28	29
Sus scrofa <2>	1990	4951	19435	769	871	215	7
<19>	26	34	22146	22	24	6	1
<15>	0	0	21574	620	620	0	0
Canis lupus familiaris <4>	779	1316	19088	2327	2576	20	17
Mus musculus <18>	650	2088	21037	507	628	122	35
<9>	563	703	21285	346	368	65	9
Rat<16>	924	1842	20486	784	908	105	41

Expansions, equal, and contractions: Total number of gene families that experienced expansions, equal and contraction along each branch of the phylogenic tree of the 11 vertebrate species described in the main text, respectively. Gene gain and gene loss, number of genes that gained or lost along each branch of the species tree. Sig Exp: number of gene families that undergone significantly expansion. Sig contra: number of gene families that undergone significantly contraction.

Table S9. Tree shrew gene families that undergone significant size changes

Orthogroups	Count	Rep ID	Orthogroups	Count	Rep ID	Orthogroups	Count	Rep ID
OG0000001	-20	*IGHV3-35*	OG0000176	6	*RPL7A*	OG0001667	15	*NSA2*
OG0000002	-12	*IGKV3-11*	OG0000460	7	*MORF4L1*	OG0000107	16	*RPS6*
OG0000008	-11	*H2BC4*	OG0000782	7	*RPL18*	OG0000801	16	*RPL23A*
OG0000012	-8	*IGKV2-28*	OG0016792	7	*GOLGA2*	OG0000904	16	*NPM1*
OG0000013	-7	*OR4F3*	OG0013864	7	*GAPDH*	OG0000441	16	*PUM3*
OG0000027	-7	*H4*	OG0000095	7	*SETMAR*	OG0001300	17	*MIA3*
OG0000046	-6	*IFNA14*	OG0017249	7	*CBX3*	OG0000102	18	*RPS3A*
OG0000037	-6	*CYP2C19*	OG0000135	7	*HMGN1*	OG0001301	18	*RPL17*
OG0000045	-5	*Novel gene*	OG0000696	7	*HSP90AB3P*	OG0001170	19	*ENV*
OG0000029	-5	*PCDHB1*	OG0000518	7	*UBE2L3*	OG0000071	19	*OR4C6*
OG0000094	-5	*OR10A2*	OG0017248	7	*GAG*	OG0001045	20	*RPL21*
OG0000066	-5	*UGT2B31*	OG0016814	8	*KBTBD7*	OG0000202	21	*KPNA2*
OG0000028	-5	*OR5B12*	OG0000276	8	*BTN1A1*	OG0000054	21	*EIF4A2*
OG0000064	-5	*OR4C12*	OG0000476	8	*HSPD1*	OG0000018	21	*HNRNPA1*
OG0000020	-5	*IGHV1-24*	OG0016812	8	*HMGB2*	OG0000424	23	*LORF2*
OG0000080	-5	*CYP4F11*	OG0016813	8	*RPSA*	OG0000491	25	*ENV*
OG0000099	-4	*RPL39L*	OG0000330	9	*KRT18*	OG0000606	25	*TXNRD1*
OG0000067	-4	*USPL12*	OG0016080	9	*TSORF*	OG0000553	26	*RALGDS*
OG0000082	-4	*PCDHA1*	OG0000306	9	*JAM3*	OG0000030	39	*RPL31*
OG0000019	-4	*CEACAM1*	OG0003475	9	*KRT8*	OG0000425	28	*EEF1A1P5*
OG0000118	-4	*ZNF607*	OG0000009	9	*HLA-E*	OG0000026	29	*HVM45*
OG0000041	-5	*GBP1*	OG0000011	9	*IGLV2-8*	OG0000121	30	*NSA2*
OG0003812	3	*HMGN2*	OG0013871	10	*ANKRD26*	OG0000370	30	*HMGB1*
OG0017921	3	*ST13*	OG0000365	10	*LDHB*	OG0000016	30	*EEF1A1*
OG0017511	3	*OR2I1P*	OG0013872	10	*GAG*	OG0000196	31	*RPL21*
OG0000892	3	*DMD*	OG0005633	10	*RPS3A*	OG0000143	31	*DNAJA1*
OG0001266	3	*LARP7*	OG0013870	10	*TSORF*	OG0000049	32	*HMGN2*

Orthogroups	Count	Rep ID	Orthogroups	Count	Rep ID	Orthogroups	Count	Rep ID
OG0002890	3	*RPS23*	OG0000648	11	*MORF4L2*	OG0000342	33	*HMGB1*
OG0000408	4	*RPL37A*	OG0003756	11	*POL*	OG0000222	35	*GAG-PRO-POL*
OG0002896	4	*LORF2*	OG0001155	11	*VMN2R116*	OG0000083	36	*HMGB1*
OG0016263	4	*TSORF*	OG0000252	11	*BTN1A1*	OG0000253	38	*RPL21*
OG0000637	4	*NLRP1*	OG0005646	11	*RPS3A*	OG0000032	39	*RPL21*
OG0000963	4	*ATP5F1C*	OG0000201	11	*HSP90AB1*	OG0000092	43	*RPL5*
OG0000293	4	*ZNF705A*	OG0000163	11	*RPS4X*	OG0000114	46	*POL*
OG0000709	4	*POL*	OG0003759	12	*IL6*	OG0000005	47	*OR5B3*
OG0000610	4	*EEF1A1*	OG0001945	12	*POL*	OG0000152	38	*STT3B*
OG0000508	4	*RPL7L1*	OG0000601	12	*CACYBP*	OG0000146	51	*LORF2*
OG0000156	5	*RPL17*	OG0003758	12	*RPS3A*	OG0000059	65	*NCL*
OG0000173	5	*NUTM2G*	OG0000050	13	*HSPA1L*	OG0000057	76	*LORF2*
OG0000275	5	*GCSH*	OG0000141	13	*HNRNPC*	OG0000021	84	*ZNF589*
OG0000692	5	*MRPL42*	OG0002899	13	*ENV*	OG0000043	86	*LIN1*
OG0000911	5	*IGLV1-40*	OG0000413	14	*SSB*	OG0000024	129	*LORF2*
OG0016914	5	*SLAMF6*	OG0000077	14	*RPS2*	OG0000022	138	*L1RE1*
OG0000278	5	*IGHV2-5*	OG0000093	14	*VPREB1*	OG0000014	154	*CASR*
OG0000129	6	*RPL36*	OG0001473	15	*ENV*	OG0000004	158	*RPL7*
OG0000294	6	*RPS20*	OG0001948	15	*RNF167*	OG0000007	180	*L1RE1*
OG0017243	6	*RPL9*	OG0001947	15	*GAG*	-	-	-
OG0000920	6	*MICB*	OG0000084	15	*RPL9*	-	-	-

Orthogroups: Group of orthologs among species that were highly identical. The orthogroups were identified by OrthoFinder software (Emms and Kelly, 2019).

Count: number of gene family member that undergone changes; minus means number of genes that contracted in a gene family.

Rep ID: the gene symbols that represented the orthogroups. Note that gene Rep ID may represent multiple orthogroups.

References

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**(1): 238.

Fan Y, Huang ZY, Cao CC, Chen CS, Chen YX, Fan DD, et al. 2013. Genome of the Chinese tree shrew. *Nature Communications*, **4**: 1426.

Fan Y, Luo R, Su LY, Xiang Q, Yu D, Xu L, et al. 2018. Does the genetic feature of the Chinese tree shrew (*Tupaia belangeri chinensis*) support its potential as a viable model for Alzheimer's disease research? *Journal of Alzheimer's Disease*, **61**(3): 1015-1028.

Fan Y, Ye MS, Zhang JY, Xu L, Yu DD, Gu TL, et al. 2019. Chromosomal level assembly and population sequencing of the Chinese tree shrew genome. *Zoological Research*, **40**(6): 506-521.

Gu T, Yu D, Fan Y, Wu Y, Yao YL, Xu L, et al. 2019. Molecular identification and antiviral function of the guanylate-binding protein (*GBP*) genes in the Chinese tree shrew (*Tupaia belangeri chinesis*). *Developmental and Comparative Immunology*, **96**: 27-36.

Han Y, Wang W, Jia J, Sun X, Kuang D, Tong P, et al. 2020. WGCNA analysis of the subcutaneous fat transcriptome in a novel tree shrew model. *Experimental Biology and Medicine*, **245**(11): 945-955.

Li CH, Yan LZ, Ban WZ, Tu Q, Wu Y, Wang L, et al. 2017. Long-term propagation of tree shrew spermatogonial stem cells in culture and successful generation of transgenic offspring. *Cell Research*, **27**(2): 241-252.

Lin J, Chen G, Gu L, Shen Y, Zheng M, Zheng W, et al. 2014. Phylogenetic affinity of tree shrews to Glires is attributed to fast evolution rate. *Molecular Phylogenetics and Evolution*, **71**: 193-200.

Luo MT, Fan Y, Mu D, Yao YG, Zheng YT. 2018. Molecular cloning and characterization of *APOBEC3* family in tree shrew. *Gene*, **646**: 143-152.

Sanada T, Tsukiyama-Kohara K, Shin IT, Yamamoto N, Kayesh MEH, Yamane D, et al. 2019. Construction of complete *Tupaia belangeri* transcriptome database by whole-genome and comprehensive RNA sequencing. *Scientific Reports*, **9**(1): 12372.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods in Molecular Biology*, **1962**: 227-245.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19): 3210-3212.

Tong Y, Hao J, Tu Q, Yu H, Yan L, Li Y, et al. 2017. A tree shrew glioblastoma model recapitulates features of human glioblastoma. *Oncotarget*, **8**(11): 17897-17907.

Tu Q, Yang D, Zhang X, Jia X, An S, Yan L, et al. 2019. A novel pancreatic cancer model originated from transformation of acinar cells in adult tree shrew, a primate-like animal. *Disease Models & Mechanisms*, **12**(4).

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, **35**(3): 543-548.

Wu X, Xu H, Zhang Z, Chang Q, Liao S, Zhang L, et al. 2016. Transcriptome profiles using next-generation sequencing reveal liver changes in the early stage of diabetes in tree shrew

(*Tupaia belangeri chinensis*). *Jounal of Diabetes Research*, **2016**: 6238526.

Yan H, Zhong G, Xu G, He W, Jing Z, Gao Z, et al. 2012. Sodium taurocholate cotransporting polypeptide is a functional receptor for human hepatitis B and D virus. *elife*, **1**: e00049.

Yao YL, Yu D, Xu L, Fan Y, Wu Y, Gu T, et al. 2019. Molecular characterization of the 2',5'-oligoadenylate synthetase family in the Chinese tree shrew (*Tupaia belangeri chinensis*). *Cytokine*, **114**: 106-114.

Yu D, Wu Y, Xu L, Fan Y, Peng L, Xu M, et al. 2016. Identification and characterization of toll-like receptors (TLRs) in the Chinese tree shrew (*Tupaia belangeri chinensis*). *Developmental and Comparative Immunology*, **60**: 127-138.

Yu D, Xu L, Liu XH, Fan Y, Lv LB, Yao YG. 2014. Diverse interleukin-7 mRNA transcripts in Chinese tree shrew (*Tupaia belangeri chinensis*). *PLoS One*, **9**(6): e99859.

Zhang X, Yu D, Wu Y, Gu T, Ma N, Dong S, et al. 2020. Establishment and transcriptomic features of an immortalized hepatic cell line of the Chinese tree shrew. *Applied Microbiology and Biotechnology*, **104**(20): 8813-8823.