Chromosomal level assembly and population sequencing of the Chinese tree shrew genome

Yu Fan1,2, Mao-Sen Ye1,3, Jin-Yan Zhang1,3, Ling Xu1,2, Dan-Dan Yu1,2, Tian-Le Gu1,3, Yu-Lin Yao1,3, Jia-Qi Chen4, Long-Bao Lv4, Ping Zheng3,4,5,6,7, Dong-Dong Wu2,8, Guo-Jie Zhang2,8, Yong-Gang Yao1,3,4,5,∗

1 Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming Yunnan 650223, China
2 Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming Yunnan 650223, China
3 Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming Yunnan 650204, China
4 Kunming Primate Research Center of the Chinese Academy of Sciences, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming Yunnan 650223, China
5 KIZ–CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming Yunnan 650223, China
6 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming Yunnan 650223, China
7 Yunnan Key Laboratory of Animal Reproduction, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming Yunnan 650223, China

ABSTRACT

Chinese tree shrews (Tupaia belangeri chinensis) have become an increasingly important experimental animal in biomedical research due to their close relationship to primates. An accurately sequenced and assembled genome is essential for understanding the genetic features and biology of this animal. In this study, we used long-read single-molecule sequencing and high-throughput chromosome conformation capture (Hi-C) technology to obtain a high-quality chromosome-scale scaffolding of the Chinese tree shrew genome. The new reference genome (KIZ version 2: TS_2.0) resolved problems in presently available tree shrew genomes and enabled accurate identification of large and complex repeat regions, gene structures, and species-specific genomic structural variants. In addition, by sequencing the genomes of six Chinese tree shrew individuals, we produced a comprehensive map of 12.8 M single nucleotide polymorphisms and confirmed that the major histocompatibility complex (MHC) loci and immunoglobulin gene family exhibited high nucleotide diversity in the tree shrew genome. We updated the tree shrew genome database (TreeshrewDB v2.0: http://www.treeshrewdb.org) to include the genome annotation information and genetic variations. The new high-quality reference genome of the Chinese tree shrew and the updated TreeshrewDB will facilitate the use of this animal in many different fields of research.

Keywords: Tupaia belangeri; Chromosomal level assembly genome; Population sequencing; Database
INTRODUCTION

Tree shrews (Tupaia belangeri) are widely distributed throughout South Asia, Southeast Asia (Fuchs & Corbach-Söhle, 2010), and South and Southwest China (Peng et al., 1991). They possess many unique characteristics that are useful in biomedical research models, such as small adult body size (100–150 g), easy and low cost maintenance, short reproductive cycle (~6 weeks), moderate life span (6–8 years), high brain-to-body mass ratio, and very close relationship to primates (Fan et al., 2013; Xiao et al., 2017; Xu et al., 2012; Yao, 2017; Zheng et al., 2014). Hilberto, tree shrews have been used in a wide variety of studies, including research on viral infection (Amako et al., 2010; Guo et al., 2018; Kock et al., 2001; Li et al., 2014; Li et al., 2016; Yang et al., 2013), visual cortex function (Bosking et al., 2002; Lee et al., 2016; MacEvoy et al., 2009; Mooser et al., 2004; Veit et al., 2014), brain development and aging (Fan et al., 2018; Wei et al., 2017), and neuropsychiatric disorders induced by social stress (Fuchs, 2005; Meyer et al., 2001). Previously, we successfully sequenced the genome of the Chinese tree shrew (Tupaia belangeri chinensis) using Illumina short-read sequencing (KIZ version 1: TS_1.0) and showed their close relationship to non-human primates, thereby settling a long-running debate regarding the phylogenetic position of tree shrews within eutherian mammals (Fan et al., 2013). Furthermore, to advance the use of the tree shrew genome, we developed a user-friendly tree shrew database (TreeShrewDB: www.treeShrewdb.org) (Fan et al., 2014). The successful genome sequencing (Fan et al., 2013) and genetic manipulation of tree shrews (Li et al., 2017) have opened up new avenues for the wide usage of this species in biomedical research (Yao, 2017).

Accurate genome sequencing and assembly are essential for understanding phylogenetic relationships and genome and phenome evolution (Kronenberg et al., 2018). Despite the fact that short-read sequencing technologies remain the most popular methods used to generate high-throughput data at relatively low cost (Schatz et al., 2010), whole-genome assembly of mammalian genomes based on these older sequencing technologies contains many problems, including assembly gaps and incomplete gene models (Sohn & Nam, 2018). For instance, approximately 50% of the human genome comprises non-random repeat elements (Cordaux & Batzer, 2009) and a complex sequence structure, which is a major challenge in reference genome assembly (Phillippy et al., 2008). Although our earlier Chinese tree shrew genome (KIZ version 1: TS_1.0) produced in 2013 (Fan et al., 2013) had high sequencing coverage (79x), the assembled genome still contained 223 607 gaps (including 65 222 gaps in the genomic region), and thus did not fully meet research needs. Single-molecule sequencing technology can generate reads tens of kilobases in size and can span most repeat sequences, which allows for complete reference genome assembly (Bickhart et al., 2017; Chaisson et al., 2015). High-throughput chromosome conformation capture (Hi-C) technology can be used to study the three-dimensional architecture of genomes and can order, orient, and anchor contigs into chromosome-scale scaffolds (Burton et al., 2013).

Here, we applied both long-read single-molecule sequencing and Hi-C technology to obtain a new reference genome for the Chinese tree shrew. We also generated a single nucleotide polymorphism (SNP) map of the tree shrew by whole-genome sequencing of six individuals. We updated the TreeShrewDB v2.0 (http://www.treeShrewdb.org) to incorporate the new reference genome and population genetic variations.

MATERIALS AND METHODS

Tissue samples and genome sequencing

A male Chinese tree shrew from the Experimental Animal Center of the Kunming Institute of Zoology, Chinese Academy of Sciences, was used for single-molecule, real-time (SMRT) long-read sequencing (PacBio) and Hi-C sequencing. Ear tissues of six Chinese tree shrews were used for whole-genome sequencing using Illumina HiSeq X Ten (USA). This study was approved by the Institutional Review Board of the Kunming Institute of Zoology, Chinese Academy of Sciences (KIZ-SYDW-20101015-001 and KIZ-SMKX-20160315-001).

We generated long-insert (20–40 kb) genomic libraries based on standard SMRT sequencing protocols developed by Pacific Biosciences (PacBio). The libraries were sequenced using the PacBio RS II instrument with the P6-C4 sequencing reagent. Brain tissue from the same individual was used to construct the Hi-C libraries. Briefly, minced brain tissue was fixed in 2% formaldehyde for 10 min and then lysed in 2.5 mol/L glycine. Cross-linked genomic DNA was digested with Mbol (#B7024, New England Biolabs, UK). Sticky ends were filled with nucleotides, one of which was biotinylated. Ligation was performed under extremely dilute conditions favoring intramolecular ligation events: the Mbol site was lost and a NheI (#R3131, New England Biolabs, UK) site was created. DNA was purified and sheared, and biotinylated junctions were isolated using streptavidin beads. Interacting fragments were sequenced by Illumina HiSeq X Ten (USA).

For whole-genome sequencing of the six tree shrew individuals, short-insert read (300 bp) genomic libraries were constructed using the Illumina TruSeq Nano DNA Library Prep Kits (USA) and sequenced using the Illumina HiSeq X Ten (USA).

Genome assembly and quality evaluation

We applied Canu (Koren et al., 2017) to correct the SMRT reads, then used smartdenovo (https://github.com/ruanjue/smartdenovo) to perform de novo assembly. The assembly was error-corrected using Quiver (Chin et al., 2013) and Pilon (Walker et al., 2014) based on alignment of 30-fold Illumina paired-end reads. The Hi-C sequencing reads were aligned to the assembled contigs using the bowtie2 end-to-end algorithm (Langmead & Salzberg, 2012). We used Lachesis (Burton et al., 2013) to cluster, order, and direct the assembled contigs onto 31 pseudo-chromosomes (TS_2.0 assembly), which was arbitrarily defined based on the number of haploid
chromosomes of the tree shrew (Liu et al., 1989). A total of 4,104 benchmarking universal single-copy orthologs in the mammalian dataset of the Benchmarking with Universal Single-Copy Orthologs (BUSCO) (Simao et al., 2015) were mapped to the assembled contigs using Tblastn (Altschul et al., 1997) to assess overall assembly quality. We also used the whole-genome sequencing data of the male Chinese tree shrew to assess the quality of the TS_2.0 assembly. In brief, we aligned the reads to the TS_2.0 assembly and previous TS_1.0 assembly (Fan et al., 2013) using BWA (Li & Durbin, 2009). We called genetic variants (SNPs and indels (insertions and deletions)) using FreeBayes (Garrison & Marth, 2012) and the structural variants (SVs) using Lumpy-SV (Layer et al., 2014), respectively. The feature response curve (FRC) (Vezzi et al., 2012) was estimated based on the aligned reads. The quality value (QV) was calculated as described previously (Bickhart et al., 2017):

\[
QV = -\log_{10}\left(\frac{S}{B}\right)
\]

where, S indicates the cumulative length of all SNPs and indels identified using FreeBayes (Garrison & Marth, 2012) that had a probability of being heterozygous greater than 0.5, and B indicates the number of base pairs in the assembly that had at least 3x sequencing coverage.

Annotation of repeats in genome

We employed Tandem Repeats Finder v4.09 (Benson, 1999) to annotate the tandem repeats in the TS_2.0 assembly. The transposable elements (TEs) were identified based on a combination of de novo and homology-based predictions, as described in our previous study (Fan et al., 2013). Briefly, the RepeatModeler (Chen, 2004) was used to construct a de novo repeat library. We used RepeatMasker and RepeatProteinMask (Chen, 2004) to identify different types of TEs by aligning the TS_2.0 assembly with the known RepBase library (Chen, 2004) and the constructed de novo repeat library.

Gene prediction and annotation

A total of 13 RNA-seq datasets from our previous studies (Fan et al., 2013, 2018) were cleaned using Trimmomatic (Bolger et al., 2014), then aligned to the TS_2.0 assembly using TopHat2 (Kim et al., 2013). The cleaned reads were also de novo assembled using Trinity (Grabherr et al., 2011). The above RNA-seq assemblies were further combined using PASA (Haas et al., 2008).

For homology-based gene prediction, we downloaded protein sequences of humans (Homo sapiens), chimpanzees (Pan troglodytes), macaques (Macaca mulatta) and mice (Mus musculus) from Ensembl (release 71;https://asia.ensembl.org/index.html), which display more accurate annotation of gene models. These protein sequences were mapped to the TS_2.0 assembly using TblastN (Altschul et al., 1997). GeneWise (Birney et al., 2004) was used to define gene models. For ab initio gene prediction, Augustus (Stanke & Waack, 2003), Genescan (Salamov & Solovyev, 2000), SNAP (Korf, 2004), and GeneMark (Besemer & Borodovsky, 2005) were used to predict coding genes.

We employed EVidenceModeler (Haas et al., 2008) to combine the RNA-seq, cDNA, and protein alignments with different weights (RNA-seq>cDNA/protein>ab initio gene predictions) to achieve a comprehensive and non-redundant reference gene set. This gene set was further updated using PASA (Haas et al., 2008), followed by annotation based on the best matches derived from the protein sequence alignments described in the SwissProt and TrEMBL databases (O’Donovan et al., 2002) using Blastp (with default parameters) (Altschul et al., 1997). We annotated motifs and domains of proteins using InterPro (Mulder & Apweiler, 2007) to search publicly available databases, including Pfam (http://pfam.sanger.ac.uk/), PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php), PROSITE (http://prosite.expasy.org/), ProDom (http://prodom.prabi.fr/prodom/current/html/home.php), and SMART (http://smart.embl-heidelberg.de/). Descriptions of gene products, such as Gene Ontology (Ashburner et al., 2000) information, were retrieved from InterPro (Mulder & Apweiler, 2007). Pathway information was obtained by blasting the above reference gene set with the KEGG database (Kanehisa & Goto, 2000), with the best hit for each gene used for the annotation.

Gene synteny map among different species

We used the human (hg38;https://www.ncbi.nlm.nih.gov/grc/human), macaque (rheMac3 (Yan et al., 2011)), and mouse (GRCm38;https://www.ncbi.nlm.nih.gov/grc/mouse) genomes and TS_2.0 to build a gene synteny map, as described previously (Fan et al., 2013). Briefly, the gene synteny map was constructed on the basis of orthologous genes. We did not use the whole genome alignment due to great sequence diversity among the species. The longest human, macaque, tree shrew, and mouse transcripts were chosen to represent each gene with alternative splicing variants. All protein sequences from the four species were aligned against the same protein set using BlastP with a similarity cutoff threshold of e-value=1×10^-6. With the human protein set as a reference, we found the best hit for each protein in the other species, with a criterion that more than 30% of the aligned sequence showed identity above 30%. Reciprocal best-match pairs were defined as orthologs. Orthologs not in the gene synteny blocks were removed from further analysis. For example, for three continuous genes (A, B, and C) in the human genome, if all three orthologs could be identified between humans and tree shrews based on the cutoff threshold described above, and the B gene in the tree shrew genome was not between genes A and C, or located in other scaffolds or other places within the same scaffold, then the B gene was removed. Using this method, we identified four-way gene synteny relationships for humans, macaques, tree shrews, and mice. The gene order information of the human genome was used to identify the macaque, tree shrew, and mouse genomic SVs.

Whole genome sequencing and SNP calling

Low-quality raw short reads were removed using Trimmomatic v0.32 (Bolger et al., 2014) with the parameters "LEADING:3
Population genetic analyses

Nucleotide diversity (π) and Tajima’s D value (Tajima, 1989) were estimated using VCFTools based on the six wild Chinese tree shrews, with a sliding window of 100 kb in each genome. For each coding gene, we estimated the population genetic parameters, including π, Watterson theta estimate (θw) (Watterson, 1975), Tajima’s D (Tajima, 1989), Fu and Li’s D (Fu & Li, 1993, Fu and Li’s F (Fu & Li, 1993), and Fay and Wu’s H (Fay & Wu, 2000), using in-house perl scripts. Manhattan plot analysis was performed using the R package qqman (https://cran.r-project.org/web/packages/qqman/index.html).

mRNA expression analysis

The raw RNA-seq data (Supplementary Table S1) were trimmed to remove sequencing adapters and reads containing one or more Ns>5% or of low quality (more than 20% of the base’s qualities were less than 10). The filtered reads were aligned to the reference genome TS_2.0 assembly using HISAT2 (Kim et al., 2015). The HTSeq-count (Anders et al., 2015) was used to count aligned reads mapped to the above reference genome set. We calculated the FPKM (fragments per kilobase per million mapped reads) value using in-house perl script to quantify mRNA expression as follows:

\[
\text{FPKM} = \frac{10^6 C}{NL/10^3}
\]

where, FPKM refers to the mRNA expression level of gene A, C is the number of fragments uniquely aligned to gene A, N is the number of reads uniquely aligned to all genes, and L is the base number in the coding region of gene A.

For co-expression analysis, we used reported RNA-seq data from seven tree shrew brain tissues (Fan et al., 2013, 2018) to calculate Pearson’s correlation coefficients for each gene pair. A co-expression gene pair was defined by a Pearson’s correlation coefficient cut-off of 0.8.

Tree shrew database

Our developed tree shrew database (TreeshrewDB v2.0) runs on a dual-processor server with an Ubuntu operating system and is implemented under the LAMP (Linux-Apache-MySQL-Perl) software stack. The Chinese tree shrew genome TS_2.0 assembly, gene set, gene annotation, and other information are stored in the MySQL, and are administrated with the help of phpMyAdmin. The web interfaces were developed using various computer languages such as HTML, CSS, JavaScript, and Perl.

RESULTS

Assembly of reference genome and quality evaluation

We generated ~55x (148.58 Gb) whole-genome sequence coverage for the sampled adult male Chinese tree shrew using SMRT long-read sequencing technology (PacBio). After filtering poor-quality reads, we programmed a combination method of de novo assembly to generate a high-quality tree shrew genome (KIZ version 2: TS_2.0, with a size of 2.67 Gb). The new assembly produced a total of 3 344 sequence contigs, with a 112-fold reduction in the number of contigs compared to that of our previous assembly (KIZ version 1: TS_1.0) based on short reads (Fan et al., 2013) (Table 1). The contig N50 of TS_2.0 was 3.2 Mb and exhibited remarkable improvement (146-fold) compared with that of the previous assembly (TS_1.0) (Fan et al., 2013). Nearly 60% of contigs (1 963/3 344) were longer than 100 kb and accounted for 97.4% of the assembled genome. The longest contig was 16.2 Mb (Table 1; Figure 1). We used BUSCO analysis, which is a powerful tool for assessing genome assembly and annotation completeness with single-copy orthologs (Simao et al., 2015), to evaluate the quality of the TS_2.0 contigs. About 91.4% of the 4 104 core genes in the mammalian dataset were complete BUSCO genes (Table 2). These tests all showed that the newly assembled tree shrew genome contigs had superior quality to those in recently reported ape genomes (Kronenberg et al., 2018).

We also generated ~264 x (705 Gb) Hi-C data (Table 3) to cluster the contigs into chromosome-scale scaffolds. A total of 1 728 contigs (comprising 96.2% of the assembled genome sequence) were anchored into 31 pseudo-molecules, whereas 1 616 contigs (102 Mb, 3.8% of assembled genome sequence) were unanchored (Table 4). The final chromosome-scale scaffolding of the de novo genome assembly of the Chinese tree shrew (TS_2.0) had a scaffold N50 length of 104 Mb, which is much more complete than the previous TS_1.0 assembly (Fan et al., 2013) (Table 1).

To compare the long-read tree shrew genome assembly (TS_2.0) in this study with the previous short-read assembly (TS_1.0) (Fan et al., 2013), we generated ~30x coverage Illumina paired-end read sequences from another tree shrew and aligned it to both assemblies. The identified SNPs and indels were used to estimate assembly accuracy. The TS_2.0 assembly (quality value=28.56) had a higher quality value, as estimated using the number of non-matching base calls from FreeBayes (Bickhart et al., 2017; Garrison & Marth, 2012), than that of the TS_1.0 assembly (quality value=26.75) (Table 5). In addition, the TS_2.0 assembly had 3-fold fewer SVs than that of TS_1.0 (Fan et al., 2013), thus suggesting fewer
assembly using tBlastn

0

2

(2015 Simao et al.

benchmarking universal single-copy orthologs of the

4

104 BUSCO

A total:

187 505  

19 269 909

–

–

–

305

1 418

2

3 655 608

3 042 664

50 199

36 335

234

3 020 651

67 915

427

1 648 848

91 810

573

Table 1  Comparison of Chinese tree shrew assembly quality between assemblies TS_1.0 and TS_2.0

| Version | Item   | Contig length (bp) | Scaffold length (bp) | Contig No. | Scaffold No. |
|---------|--------|---------------------|----------------------|------------|--------------|
|         | Total  | 2 719 442 484      | 2 861 790 358        | 374 120    | 150 513      |
|         | Max_length | 187 505          | 19 269 909           | –          | –            |
|         | >2 000 bp | –                  | –                    | 180 802    | 4 525        |
|         | >100 kb  | –                  | –                    | 305        | 1 418        |
| Short-read assembly (KIZ version 1: TS_1.0) | N50    | 22 000              | 3 655 608            | 36 335     | 234          |
|         | N60    | 17 500              | 3 042 664            | 50 199     | 319          |
|         | N70    | 13 431              | 2 302 651            | 67 915     | 427          |
|         | N80    | 9 571               | 1 648 848            | 91 810     | 573          |
|         | Total  | 2 667 337 536       | 2 667 507 236        | 3 344      | 1 647        |
|         | Max_length | 16 177 999         | 224 450 918          | –          | –            |
|         | >2 000 bp | –                  | –                    | 3 344      | 1 647        |
|         | >100 kb  | –                  | –                    | 1963       | 281          |
| Long-read assembly (KIZ version 2: TS_2.0) | N50    | 3 217 288           | 104 643 080          | 229        | 10           |
|         | N60    | 2 462 062           | 94 037 081           | 323        | 13           |
|         | N70    | 1 641 093           | 71 760 103           | 457        | 16           |
|         | N80    | 995 871             | 57 328 337           | 664        | 20           |

Table 2  Assessment of assembly completeness in Chinese tree shrew using BUSCO

| Parameter                  | No. | Percentage (%) |
|---------------------------|-----|----------------|
| Complete genes            | 3 749 | 91.40         |
| Complete and single-copy genes | 3 696 | 90.10         |
| Complete and duplicated genes | 53   | 1.30          |
| Fragmented genes          | 184  | 4.50          |
| Missing genes             | 171  | 4.10          |
| Total genes               | 4 104 | –             |

Table 3  Statistics of Hi-C data for mapping

| Parameter                  | Hi-C data |
|---------------------------|-----------|
| Clean data                | 705 Gb    |
| Clean paired-end reads    | 2 351 150 069 |
| Unmapped paired-end reads | 47 514 976 |
| Unmapped paired-end reads rate (%) | 2.02 |
| Paired-end reads with singleton | 303 352 692 |
| Paired-end reads with singleton rate (%) | 12.9 |
| Multi mapped paired-end reads | 443 734 174 |
| Multi mapped ratio (%)    | 18.87     |
| Unique mapped paired-end reads | 1 556 548 227 |
| Unique mapped ratio (%)   | 66.2      |

Contig: Contiguous length of genomic sequence in which the order of bases has a high confidence level. Gaps occur where reads from two sequenced ends of at least one fragment overlap with other reads in two different contigs. Scaffolds are composed of contigs and gaps. N50: N50 statistic defines assembly quality in terms of contiguity. Given a set of contigs ranked by contig size, N50 is defined as the size of the shortest contig, which adds contigs with larger size to reach 50% of total genome length. –: Not available.

assembly errors (Table 5). Quality evaluation using the FRC method (Vezzi et al., 2012) also showed TS_2.0 to be a better assembly (Table 5).

About 73% of the gaps (163 220 gaps, 93.10 Mb) in the TS_1.0 assembly (Fan et al., 2013) were filled by the 49.15 Mb long-read sequences in the TS_2.0 assembly. Among these fully closed gaps, 65 222 were located in the genomic regions. Only 39 gaps in TS_2.0 were fully closed by TS_1.0 (Table 6). We note that 4 112 gaps in TS_1.0 had flanking sequences that were mapped to separate pseudo-chromosomes in TS_2.0, indicating assembly errors in TS_1.0 (Fan et al., 2013).

The updated genome assembly is available at TreeshrewDB v2.0 (http://www.treeshrewdb.org) and has been deposited in GSA (accession No. PRJCA001472; http://gsa.big.ac.cn/) (Wang et al., 2017).

Repeat content in tree shrew genome

Repeat content in a genome poses a daunting difficulty for sequence assembly (Kronenberg et al., 2018). The function of repeat content has also begun to be recognized (Chuong et al., 2018). Repeat content in tree shrew genome has also been recognized (Kronenberg et al., 2018). The function of repeat content has also begun to be recognized (Chuong et al., 2018).
Figure 1 Assembly, annotation, and nucleotide diversity of Chinese tree shrew genome
A: Contig length distribution of long-read assembly (TS_2.0) in comparison with short-read assembly (TS_1.0) (Fan et al., 2013). B: Circos plot showing genome-wide distribution profiles of genes, SNPs, and indels across Chinese tree shrew genome, and values of population genetic parameters (π and Tajima’s D). C: Manhattan plot of nucleotide diversity (π) at gene level based on SNPs located in coding regions of six wild tree shrews. Top 30 genes are shown in plot, with a cut-off π value of 0.025.
We used Lachesis (Burton et al., 2013) to cluster, order, and direct the assembled contigs onto 31 pseudo-chromosomes, which were defined according to number of haploid chromosomes of the tree shrew (Liu et al., 1989). Contig No.: Number of contigs assembled onto each chromosome by Hi-C. Total anchored: total number of contigs that could be anchored into 31 pseudo-chromosomes. Unanchored: total number of contigs that could not be anchored into 31 pseudo-chromosomes.

We note that some long tandem repeats within the genic regions were located in gaps in the assembled contigs. For instance, Fan et al. (2010) and later studies also identified long tandem repeats (up to 150 kb) with each repeat being less than 2 kb. We also defined 4,111 satellites each greater than 20 kb with the exception of the QV score, lower counts are indicative of better assembly.

highest proportion in TS_2.0 (18.54% of genome size; Table 8), similar to that of L1 in the human genome (Beck et al., 2010). The tree shrew specific tRNA-derived Tu-III family (Beck et al., 2010), accounted for 15.17% of genome size in TS_2.0 (Table 8). The reason for the unusually high prevalence of the tRNA-derived Tu-III family in the tree shrew genome remains to be determined. Because of the improvement in genome quality, we were able to identify 127 long transposable elements (each >20 kb). We also defined 4,111 satellites (total length of 131 Mb) in TS_2.0. Among them, 4,293,990 were short tandem repeats (each <150 bp) and 1,512 were long tandem repeats (each >5 kb). The longest tandem repeat was mapped to a non-coding region in the end of pseudo-chromosome 26 and had a length of 168.3 kb (period size= 359 bp, copy number=471). We note that some long tandem repeats within the genic regions were located in gaps in the genome assembly (Fan et al., 2013). For instance, a long
tandem repeat (period size=1 917 bp, copy number=26) overlapped with the coding sequence of the OS9 gene (osteosarcoma amplified 9, endoplasmic reticulum lectin), which plays a key role in the endoplasmic reticulum stress response associated with hypoxia (Satoh et al., 2010). Therefore, these long tandem repeats in the tree shrew genome may be functional. However, focused studies are required for their characterization.

Table 6 Comparison of transposable elements in Chinese tree shrews between short-read assembly (KIZ version 1: TS_1.0) and long-read assembly (KIZ version 2: TS_2.0)

| Type      | Long-read assembly (TS_2.0) | Short-read assembly (TS_1.0) |
|-----------|-----------------------------|-----------------------------|
| DNA       | 96.8                        | 3.6                         |
| LINE      | 553.3                       | 20.8                        |
| SINE      | 663.1                       | 24.9                        |
| LTR       | 138.0                       | 5.2                         |
| Other     | 0.0005                      | 0.0                         |
| Unknown   | 68.0                        | 2.6                         |
| Total     | 1 310.5                     | 49.1                        |

DNA: Deoxyribonucleic acid transposon. LINE: Long interspersed nuclear element. SINE: Short interspersed nuclear element. LTR: Long terminal repeat.

Gene annotation updates

We combined the homology-based, de novo, and transcriptome-based methods (Haas et al., 2008) to predict protein-coding genes in the TS_2.0 assembly and identified a total of 23 568 non-redundant protein-coding genes (Fan et al., 2013) (Table 9; Figure 1B). Among these genes, the majority (22 907 genes) were supported by the reported RNA-seq data in our recent studies (Fan et al., 2013, 2018). The newly updated gene set had longer coding sequences, which were, on average, composed of more exons (Table 9) compared with the TS_1.0 gene set (Fan et al., 2013). The gaps in 2 091 exons in TS_1.0 (Fan et al., 2013) were all filled in TS_2.0, thus providing better annotation information for the genes. For instance, LILRB3 (leukocyte immunoglobulin like receptor B3), which binds to the major histocompatibility complex (MHC) class I molecules on antigen-presenting cells to inhibit stimulation of immune response (Huang et al., 2010), was complete in TS_2.0, but less than 50% of this gene sequence was retrieved in TS_1.0 (Fan et al., 2013). BMP8A (bone morphogenetic protein 8a), which plays a role in the development of the reproductive system (Wu et al., 2017), exhibited low protein sequence identity (57.22%) with human homolog in TS_1.0 (Fan et al., 2013) due to assembly error and gaps, but the sequence identity reached 88.94% in TS_2.0. In addition, ALOX15 (arachidonate 15-lipoxygenase), which uses polyunsaturated fatty acid substrates to generate various bioactive lipid mediators, such as eicosanoids, heparins, and lipoxins (Kuhn et al., 2018; Singh & Rao, 2019), had only one copy in TS_1.0 (Fan et al., 2013) but four copies in TS_2.0. The updated versions of these important genes have provided a good basis for further specific functional characterization.

Of the annotated genes, 20 811 (88.3%) were functionally classified according to InterPro (Mulder & Apweiler, 2007), GO (Ashburner et al., 2000), KEGG (Kanehisa & Goto, 2000), Swissprot, and TrEMBL (O’Donovan et al., 2002). In addition, 586 genes were newly annotated in TS_2.0 (Table 10). All these genes can be retrieved from TreeshrewDB v2.0.

It should be mentioned that the MHC region (starting from MOG to COL11A2 (Beck et al., 1999) in pseudo-chromosome 3) was completely assembled (Figure 2A) in the TS_2.0 assembly. It was previously difficult to assemble this region using short-read sequencing technologies as it is highly polymorphic and repetitive. There were 412 gaps in the MHC region in TS_1.0 (Fan et al., 2013), which were all filled in TS_2.0. Thus, the tree shrew has more MHC class I genes (n=8 according to TS_2.0) than those identified in humans (n=6), although fewer than those identified in mice (n=12) (Elmer & McAllister, 2012) (Figure 2B).
Table 8 Comparison of transposable element subtypes in Chinese tree shrews between short-read assembly (KIZ version 1: TS_1.0) and long-read assembly (KIZ version 2: TS_2.0)

| TE subtype | Long-read assembly (TS_2.0) | Short-read assembly (TS_1.0) |
|------------|-----------------------------|-----------------------------|
|            | Length (Mb) | % in genome | Length (Mb) | % in genome |
| DNA/En-Spm | 7.92        | 0.30        | 4.87        | 0.17        |
| DNA/HAT    | 34.29       | 1.28        | 33.77       | 1.18        |
| DNA/TcMar  | 52.11       | 1.96        | 26.90       | 0.94        |
| LINE/CR1   | 5.57        | 0.21        | 2.00        | 0.07        |
| LINE/L1    | 494.51      | 18.54       | 267.29      | 9.34        |
| LINE/L2    | 49.48       | 1.86        | 22.04       | 0.77        |
| LINE/Penelope | 2.02   | 0.08        | 2.29        | 0.08        |
| LTR/ERV1   | 37.66       | 1.41        | 31.77       | 1.11        |
| LTR/ERVK   | 18.55       | 0.70        | 8.87        | 0.31        |
| LTR/ERVL   | 78.13       | 2.93        | 68.40       | 2.39        |
| LTR/Gypsy  | 2.59        | 0.10        | 2.86        | 0.1         |
| SINE/Au    | 10.60       | 0.40        | 3.15        | 0.11        |
| SINE/B4    | 3.02        | 0.11        | 1.72        | 0.06        |
| SINE/MIR   | 47.40       | 1.78        | 23.75       | 0.83        |
| SINE/RNA-Lys | 15.04    | 0.56        | 1.14        | 0.04        |
| SINE/Tu-III | 404.49   | 15.17       | 410.09      | 14.33       |

Table 9 Comparison of Chinese tree shrew gene annotation between short-read assembly (KIZ version 1: TS_1.0) and long-read assembly (KIZ version 2: TS_2.0)

| Parameter          | Long-read assembly (TS_2.0) | Short-read assembly (TS_1.0) |
|--------------------|-----------------------------|-----------------------------|
| Total number of genes | 23 568                     | 22 121                     |
| Complete ORFs      | 21 117                      | 21 085                      |
| Annotated genes    | 20 811                      | 20 225                      |
| Average mRNA length | 40 114                     | 33 712                      |
| Average CDS length | 1 527                      | 1 404                      |
| Average exon number | 8.86                      | 7.54                       |
| Average exon length | 172                       | 186                        |
| Average intron length | 4 907                     | 4 937                      |

ORF: open reading frame. CDS: coding-region sequences.

Genomic structural variants

The genome TS_2.0 assembly improved sequence continuity and provided an opportunity to explore species-specific genomic SVs in genic regions. We used the human (hg38: https://www.ncbi.nlm.nih.gov/igrc/human), macaque (rheMac3 (Yan et al., 2011)), and mouse genome (GRCm38; https://www.ncbi.nlm.nih.gov/igrc/mouse) genomes and TS_2.0 to construct a synteny map of orthologous genes, using the human genome as a reference. We identified 221 SVs in tree shrews (Supplementary Table S2), 188 SVs in macaques (Supplementary Table S3), and 387 SVs in mice (Supplementary Table S4), suggesting that the tree shrew's genomic structure had an overall higher similarity to that of primates than to that of mice. A detailed comparison of the SVs showed that the tree shrews had a seemingly mosaic pattern with some similarities to rodents and others to
primates. For instance, the tree shrew and primates (human and macaque) had a specific genomic SV in the region starting from MYSM1 to SLC35D1, which was inverted in the mouse genome (Figure 3A). Some SVs existed in the tree shrew and mouse, but primates had different counterparts, such as the region from PRKAB2 to POLR3GL (Figure 3B). Note that in this region, GPR89B (G protein-coupled receptor 89B) and NOTCH2NL (notch 2 N-terminal like A) were only present in the human genome (Figure 3B). The updated TS_2.0 assembly has thus provided more opportunities to understand the evolution of SVs and potentially disrupted genes in the tree shrew genome. The exact reason for the occurrence of species-specific SVs and their potential evolutionary and functional effects await further study.

Genomic sequence variations at population level
To understand genomic sequence variations in the Chinese tree shrew, we analyzed the whole-genome sequencing data of six individuals (each with a sequencing depth of 30x). After mapping to TS_2.0, we identified a total of 12.8 million (M) SNPs in these individuals (Figure 1B), with 293 128 (including 194 751 synonymous and 98 377 non-synonymous SNPs) located in the coding regions.

We estimated population genetic parameters for the Chinese tree shrew using the six captive individuals. We calculated the nucleotide diversity based on SNPs located in coding regions and identified 30 genes with high nucleotide diversity based on a cut-off π value of 0.025 (Figure 1C). Among these genes, five were located in the MHC loci or belonged to the immunoglobulin gene family, suggesting that immune genes may have a relatively high evolutionary rate in tree shrews, although this needs to be validated by analyzing more samples and including non-coding regions (Figure 1C).
Figure 3  Examples of structural variants in mouse, macaque, tree shrew, and human genomes
A: Chinese tree shrews and humans, but not mice, shared a specific genomic structure in the region from MYSM1 to SLC35D1. B: Chinese tree shrews and mice shared a specific genomic structure in the region from PRKAB2 to POLR3GL, which has undergone dramatic changes in humans. GPR89B (G protein-coupled receptor 89B) and NOTCH2NL (notch 2 N-terminal like A) genes, marked in green, were only present in the human genome. These genomes were retrieved from public domains (mouse: GRCm38; https://www.ncbi.nlm.nih.gov/grc/mouse; macaque: rheMac3(Yan et al., 2011); human: hg38; https://www.ncbi.nlm.nih.gov/grc/human) or generated in this study (tree shrew: TS_2.0).
Whether or not this pattern reflects a compensatory effect due to the loss of RIG-I in the tree shrew genome (Fan et al., 2013; Xu et al., 2016; Yao, 2017) remains to be studied. We calculated Tajima’s D (Tajima, 1989) for each gene, and found discrete distribution, with no obvious clustering (Supplementary Figure S1). Results for Fu and Li’s D test (Fu & Li, 1993), Fu and Li’s F test (Fu & Li, 1993), Fay and Wu’s H test (Fay & Wu, 2000) all showed a pattern similar to the Tajima’s D test. Nonetheless, these results should be treated with caution, as they may be biased by the limited sample size.

Tree shrew database updates

Based on the TS_2.0 assembly, we updated TreesheerDB v2.0 (Figure 4) to distribute the new high-quality tree shrew genome and our newly annotated gene and genome information. The main database updates included revision and expansion of genomic data, gene co-expression patterns, population genetic statistics, and improvements to the web interface. Briefly, for the retrieval module, we updated the reference sequence ID, genomic location and map, transcript sequence, and functional annotation based on the new gene set. We added five primate species (gibbon (Nomascus leucogenys), golden snub-nosed monkey (Rhinopithecus roxellana), black snub-nosed monkey (R. bieti), Bolivian squirrel monkey (Saimiri boliviensis boliviensis), and bushbaby (Otolemur garnettii)) in the orthologous gene sets from Ensembl (release 71; https://asia.ensembl.org/index.html) to allow for better comparison for one-to-one homologs. The mRNA expression pattern was upgraded based on RNA-seq data from 26 tissues and/or cells (Supplementary Table S1). For each gene query, it is possible to retrieve basic information on the queried gene, sequence alignment with homologs of other species, mRNA expression levels in tissues/cells, co-expression patterns in brain tissues, sequence variations at the population level, and results of population genetic parameters (including ρ, Watterson theta estimate (θw) (Watterson, 1975), Tajima’s D (Tajima, 1989), Fu and Li’s D (Fu & Li, 1993), Fu and Li’s F (Fu & Li, 1993), and Fay and Wu’s H (Fay & Wu, 2000)) (Figure 4).

We added the new Chinese tree shrew reference genome (TS_2.0) to the TreesheerDB v2.0, which is free to download. The updated gene sequences can be extracted in batches or individually by our homemade ExtractSeq. We incorporated Blast (Altschul et al., 1997) and Genewise (Birney et al., 2004) to show the mapping of genes in the genome. Overall, the updated database now provides a comprehensive annotation of the Chinese tree shrew genome to satisfy the needs of evolutionary analysis and biomedical research.

DISCUSSION

The combination of long-read sequencing and long-range chromosome interaction mapping (such as Hi-C) represents the most efficient approach to produce high-quality reference genome assembly (Bickhart et al., 2017; Kronenberg et al., 2018). In this study, we used these techniques to generate an updated reference genome for the Chinese tree shrew (KIZ version 2: TS_2.0) and resolved some of the problems from our earlier tree shrew genome (Fan et al., 2013). The updated TS_2.0 assembly enabled accurate identification of large and complex repeat regions, gene structures, and species-specific genomic SVs in the genic regions. This high-quality tree shrew genome will facilitate the use of this species in both biomedical and basic research, such as annotation and interpretation of RNA-seq data from normal and pathological tissues (Supplementary Table S1), and for a more comprehensive understanding of the evolution of primate-specific SVs and their potential regulatory changes (Kronenberg et al., 2018). For instance, we identified 221 SVs in the genic regions of the Chinese tree shrew genome and found that the overall pattern of SVs in the tree shrew more resembled that of primates than that of rodents (Supplementary Tables S2–4), further confirming the very close relationship between tree shrews and primates (Fan et al., 2013; Yao, 2017). It should be mentioned that the TS_2.0 assembly still misses many large and complex SVs due to the limitations of current sequencing technology and assembly approaches. Moreover, we did not experimentally validate the SVs between TS_1.0 and TS_2.0, which would offer further information regarding the construction of a well-defined reference genome of the Chinese tree shrew. We will continue to refine the tree shrew genome using more data in the future. In general, the new TS_2.0 assembly filled most of the gaps and corrected most assembly errors present in the previous tree shrew genome (Fan et al., 2013), thereby providing better gene annotations. To understand the unique genetic features of the tree shrew genome, such as long tandem repeats, repeat content, and genomic SVs, detailed studies should be carried out in the future.

In our previous study, we built the TreesheerDB (Fan et al., 2014) for easy access to the Chinese tree shrew genome data based on short-read sequencing technology (Fan et al., 2013), which has been visited frequently and used by many researchers. We comprehensively updated TreesheerDB v2 based on the new high-quality reference genome (TS_2.0) generated in this study. We optimized the visualizations of gene annotation and genomic variations of the tree shrew and included results from population genetic parameters for this species. Furthermore, the inclusion of the reported transcriptomic data from 26 tissues and cells (Supplementary Table S1) has enhanced our knowledge of mRNA expression profiling in the Chinese tree shrew. This database will be regularly updated to include recently released genetic data and serve as a platform for data sharing among tree shrew studies and for further elucidation of the genetic features of this animal. We believe that the tree shrew genome assembly TS_2.0 and the updates will meet the increasing needs in the field.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.
Figure 4  Overview of updated tree shrew database (TreeshrewDB version 2.0)

Inclusion of the high-quality reference genome assembly (TS_2.0) in TreeshrewDB version 2.0 provided a comprehensive update of gene annotation information, genomic variations, population genetic features, and mRNA expressions. Population genetic parameters (including \( \pi \), Watterson theta estimate \( \theta_w \) (Watterson, 1975), Tajima’s D (Tajima, 1989), Fu and Li’s D (Fu & Li, 1993), Fu and Li’s F (Fu & Li, 1993), and Fay and Wu’s H (Fay & Wu, 2000)) were estimated based on SNPs located in coding regions in the whole genome sequences of six wild tree shrews. The database is freely accessible at http://www.treeshrewdb.org.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS’ CONTRIBUTIONS

Y.F. and Y.G.Y. conceived and designed the research. J.Q.C. and L.L. participated in the material preparation. Y.F. performed the genome assembly, annotation, and database update. Y.F., M.Y., J.Y.Z., L.X., D.Y., T.G., Y.L.Y., P.Z., D.D.W., and G.J.Z. performed comparative analyses and/or provided critical comments. Y.F. and Y.G.Y. wrote the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank Ian Logan for critical comments and language editing for this manuscript. We are grateful for the three reviewers for helpful comments on the early version of the manuscript.
Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics, Chapter 4: Unit 4.10.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6): 563–569.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics, 18(2): 71–86.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nature Reviews Genetics, 10(10): 691–703.

Elmer BM, McMullin AK. 2012. Major histocompatibility complex class I proteins in brain development and plasticity. Trends in Neurosciences, 35(11): 660–670.

Fan Y, Huang ZY, Cao CC, Chen CS, Chen YX, Fan DD, He J, Hou HL, Hu L, Hu XT, Jiang XT, Lai R, Lang YS, Liang B, Liao SG, Mu D, Ma YY, Niu YY, Sun QX, Xia JQ, Xiao J, Xiong ZQ, Xu L, Yang L, Zhang Y, Zhao W, Zhao XD, Zheng YT, Zhou JM, Zhu YB, Zhang GJ, Wang J, Yao YG. 2013. Genome of the Chinese tree shrew. Nature Communications, 4: 1426.

Fan Y, Luo RC, Su LY, Xiang Q, Yu DD, Xu L, Chen QJ, Bi R, Wu DD, Zheng P, Yao YG. 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, Yu DD, Yao YG. 2014. Tree shrew database (TreeshrewDB): a genomic knowledge base for the Chinese tree shrew. Scientific Reports, 4: 7145.

Fay JC, Li WH. 2000. Hitchhiking under positive Darwinian selection. Genetics, 155(3): 1405–1413.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics, 133(3): 693–709.

Fuchs E. 2005. Social stress in tree shrews as an animal model of depression: an example of a behavioral model of a CNS disorder. CNS Spectrums, 10(3): 182–190.

Fuchs E, Corbach-Söhle S. 2010. Tree Shrews. In: Hubrecht R, Kirkwood J. (eds.). The UFW Handbook on the Care and Management of Laboratory and Other Research Animals (8th Edition). Oxford, UK: Wiley-Blackwell, 262–275.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv: Genomics: arXiv:1207.3907v1.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Haohen N, Ginave H, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Raveg A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7): 644–652.

Guo WN, Zhu B, Ai L, Yang DL, Wang BJ. 2018. Animal models for the study of hepatitis B virus infection. Zoological Research, 39(1): 25–31.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biology, 9(1): R7.

Huang JH, Burke PS, Cung TDH, Pereyra F, Toth I, Walker BD, Borges L, Lichterfeld M, Yu XG. 2010. Leukocyte immunoglobulin-like receptors...
maintain unique antigen-presenting properties of circulating myeloid dendritic cells in HIV-1-infected elite controllers. Journal of Virology, 84(18): 9463–9471.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28(1): 27–30.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12(4): 357–360.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology, 14(4): R36.

Kock J, Nassal M, MacNelly S, Baumert TF, Blum HE, von Weizsacker F. 2001. Efficient infection of primary Tupaia hepatocytes with purified human and woolly monkey hepatitis B virus. Journal of Virology, 75(11): 5084–5089.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5): 722–736.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics, 5: 59.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantisari S, Meyerson OS, Underwood JG, Nelson BJ, Chaissen MP, Dougherty ML, Munson KM, Hastie AR, Diekman H, Hormozdari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Weich AE, Sorensen M, Baker C, Fulton RS, Armstrong J, Graves-Lindsay TA, Denli AM, Hoppe ER, Hsieh P, Hill CM, Pang AWC, Lee J, Lam ET, Dutcher SK, Gage FH, Warf WH, Shendure J, Hausser D, Schneider VA, Cao H, Ventura M, Wilson RK, Patten B, Pollen A, Eichler EE. 2018. High-resolution comparative analysis of great ape genomes. Science, 360(6393): eaar3433.

Kuhn H, Humeniuk L, Kozlov N, Roigas S, Adel S, Heydeck D. 2018. The evolutionary hypothesis of reaction specificity of mammalian ALOX15 orthologs. Progress in Lipid Research, 72: 55–74.

Langmead B, Salzberg SL. 2012. Fast gapped-read aligner with Bowtie 2. Nature Methods, 9(4): 357–359.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology, 15(6): RB4.

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

Li CH, Yan LZ, Ban WZ, Tu Q, Wu Y, Wang L, Li R, Ji S, Ma YH, Nie WH, Lv LB, Yao YG, Zhao XD, Zheng P. 2017. Long-term propagation of shrew spermatogonial stem cells in culture and successful generation of transgenic offspring. Bioinformatics, 27(2): 241–252.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14): 1754–1760.

Li JP, Liao Y, Zhang Y, Wang JJ, Wang LC, Feng K, Li QH, Li LU. 2014. Experimental infection of tree shrews (Tupaia belangeri) with Coxackie virus A16. Journal of Virology, 35(6): 485–491.

Li PH, Li ZR, Wang EL, Yang R, Xiao Y, Han HB, Lang FC, Li X, Xia YJ, Gao F, Li QH, Fraser NW, Zhou JM. 2016. Herpes simplex virus 1 infection of tree shrews differs from that of mice in the severity of acute infection and viral transcription in the peripheral nervous system. Journal of Virology, 90(2): 790–804.

Liu RQ, Shi LM, Chen YZ. 1989. Comparative studies on chromosomes of 3 subspecies of Tupaia belangeri. Zoological Research, 10(3): 195–200. (In Chinese)

MacEvoy SP, Tucker TR, Fitzpatrick D. 2009. A precise form of divisive suppression supports population coding in the primary visual cortex. Nature Neuroscience, 12(5): 637–645.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garinimela K, Altschuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9): 1297–1303.

Meyer U, van Kampen M, Isovich E, Filigge F, Fuchs E. 2001. Chronic psychosocial stress regulates the expression of both GR and MR mRNA in the hippocampal formation of tree shrews. Hippocampus, 11(3): 329–336.

Mooser F, Bosking WH, Fitzpatrick D. 2004. A morphological basis for orientation tuning in primary visual cortex. Nature Neuroscience, 7(8): 872–879.

Mulder N, Apweiler R. 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods in Molecular Biology, 396: 59–70.

O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Barooshian A, Apweiler R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Briefings in Bioinformatics, 3(3): 275–284.

Peng YZ, Ye ZZ, Zou RJ, Wang YX, Tian BP, Ma YY, Shi LM. 1991. Biology of Chinese Tree Shrews (Tupaia belangeri chinensis). Kunming, China: Yunnan Science and Technology Press.

Phillippy AM, Schatz MC, Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome Biology, 9(3): R55.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Research, 10(4): 516–522.

Satoch T, Chen Y, Hu D, Hanashima S, Yamamoto K, Yamaguchi Y. 2010. Structural basis for oligosaccharide recognition of misfolded glycoproteins by OS-9 in ER-associated degradation. Molecular Cell, 40(6): 905–916.

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. Genome Research, 20(9): 1165–1173.

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.

Singh NK, Rao GN. 2019. Emerging role of 12/15-Lipoxygenase (ALOX15) in human pathologies. Progress in Lipid Research, 73: 28–45.

Sohn JI, Nam JW. 2018. The present and future of de novo whole-genome assembly. Briefings in Bioinformatics, 19(1): 23–40.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, 19(3): 215–225.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3): 585–595.

Veit J, Bhattacharyya A, Kretz R, Rainer G. 2014. On the relation between receptive field structure and stimulus selectivity in the tree shrew primary visual cortex. Cerebral Cortex, 24(10): 2761–2771.

Vezzi F, Narzisi G, Mishra B. 2012. Reevaluating assembly evaluations with feature response curves: GAGE and ambassemblers. PLoS One, 7(12): e52210.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sathikumaran S, Cuomo CA, Zeng QD, Wortman J, Young SK, Earl AM. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One, 9(11): e112963.
Wang K, Li MY, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16): e164.

Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, Tang B, Dong L, Ding N, Zhang Q, Bai Z, Dong X, Chen H, Sun M, Zhai S, Sun Y, Yu L, Lan L, Xiao J, Fang X, Lei H, Zhang Z, Zhao W. 2017. GSA: genome sequence archive. *Genomics, Proteomics & Bioinformatics*, 15(1): 14–18.

Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2): 256–276.

Wei S, Hua HR, Chen QQ, Zhang Y, Chen F, Li SQ, Li F, Li JL. 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.

Wu FJ, Lin TY, Sung LY, Chang WF, Wu PC, Luo CW. 2017. BMP8A sustains spermatogenesis by activating both SMAD1/5 and SMAD2/3 in spermatogonia. *Science Signaling*, 10(477): eaal1910.

Xiao J, Liu R, Chen CS. 2017. Tree shrew (Tupaia belangeri) as a novel laboratory disease animal model. *Zoological Research*, 38(3): 127–137.

Xu L, Chen SY, Nie WH, Jiang XL, Yao YG. 2012. Evaluating the phylogenetic position of Chinese tree shrew (Tupaia belangeri chinensis) based on complete mitochondrial genome: implication for using tree shrew as an alternative experimental animal to primates in biomedical research. *Journal of Genetics and Genomics*, 39(3): 131–137.

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.

Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, Du H, Chen J, Chen R, Zhang P, Huang Z, Thompson JR, Meng Y, Bai Y, Wang J, Zhuo M, Wang T, Huang Y, Wei L, Li J, Wang Z, Hu H, Yang P, Le L, Stenson PD, Li B, Liu X, Ball EV, An N, Huang Q, Zhang Y, Fan W, Zhang X, Li Y, Wang W, Katze MG, Su B, Nielsen R, Yang H, Wang J, Wang X, Wang J. 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature Biotechnology*, 29(11): 1019–1023.

Yang ZF, Zhao J, Zhu YT, Wang YT, Liu R, Zhao SS, Li RF, Yang CG, Li JQ, Zhong NS. 2013. The tree shrew provides a useful alternative model for the study of influenza H1N1 virus. *Virolology Journal*, 10: 111.

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

Zheng YT, Yao YG, Xu L. 2014. Basic Biology and Disease Models of Tree Shrews. Kunming, China: Yunnan Science and Technology Press.