nature COMMUNICATIONS

ARTICLE

Received 6 Sep 2012 | Accepted 20 Dec 2012 | Published 5 Feb 2013

DOI: 10.1038/ncomms2416

OPEN

Genome of the Chinese tree shrew

Yu Fan[1,2,*], Zhi-Yong Huang[3,*], Chang-Chang Cao[3], Ce-Shi Chen[1], Yuan-Xin Chen[3], Ding-Ding Fan[3], Jing He[3], Hao-Long Hou[3], Li Hu[3], Xin-Tian Hu[1], Xuan-Ting Jiang[3], Ren Lai[1], Yong-Shan Lang[3], Bin Liang[1], Sheng-Guang Liao[3], Dan Mu[1,2], Yuan-Ye Ma[1], Yu-Yu Niu[1], Xiao-Qing Sun[3], Jin-Quan Xia[3], Jin Xiao[3], Zhi-Qiang Xiong[3], Lin Xu[1], Lan Yang[3], Yun Zhang[1], Wei Zhao[3], Xu-Dong Zhao[1], Yong-Tang Zheng[1], Ju-Min Zhou[1], Ya-Bing Zhu[3], Guo-Jie Zhang[1,3,5], Jun Wang[3,4,5,6] & Yong-Gang Yao[1]

Chinese tree shrews (*Tupaia belangeri chinensis*) possess many features valuable in animals used as experimental models in biomedical research. Currently, there are numerous attempts to employ tree shrews as models for a variety of human disorders: depression, myopia, hepatitis B and C virus infections, and hepatocellular carcinoma, to name a few. Here we present a publicly available annotated genome sequence for the Chinese tree shrew. Phylogenomic analysis of the tree shrew and other mammalians highly support its close affinity to primates. By characterizing key factors and signalling pathways in nervous and immune systems, we demonstrate that tree shrews possess both shared common and unique features, and provide a genetic basis for the use of this animal as a potential model for biomedical research.

[1] Key Laboratory of Animal Models and Human Disease Mechanisms of Chinese Academy of Sciences and Yunnan Province, Kunming Institute of Zoology, Kunming, Yunnan 650223, China. [2] University of Chinese Academy of Sciences, Beijing 100039, China. [3] BGI-Shenzhen, Shenzhen 518083, China. [4] Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, DK-2200, Copenhagen, Denmark. [5] Department of Biology, University of Copenhagen, DK-2200, Copenhagen, Denmark. [6] King Abdulaziz University, 21589 Jeddah, Saudi Arabia. * These authors contributed equally to this work and should be treated as co-first authors. Correspondence and requests for materials should be addressed to G.-J.Z. (email: zhanggj@genomics.org.cn) or to J.W. (email: wangj@genomics.org.cn) or to Y.-G.Y. (email: ygyaozh@gmail.com).

The tree shrew (*Tupaia belangeri*), currently placed in the order Scandentia, has a wide distribution in South Asia, Southeast Asia and Southwest China[1]. For several decades, owing to a variety of unique characteristics ideal in an experimental animal (for example, small adult body size, high brain-to-body mass ratio, short reproductive cycle and life span, low cost of maintenance, and most importantly, a claimed close affinity to primates) the tree shrew has been proposed as a viable animal model alternative to primates in biomedical research and drug safety testing[2].

Currently, there are many attempts to employ tree shrew to create animal models for studying hepatitis C virus (HCV)[3] and hepatitis B virus (HBV) infections[4], myopia[5], as well as social stress and depression[6,7]. Recent studies of aged tree shrew brain suggested that tree shrew is also a valid model for aging research[8] and learning behaviours[9]. Despite marked progress in using tree shrews as an animal model, tree shrews are studied only in a handful of laboratories worldwide, partially because there is no pure breed of this animal, limited access to this animal resource and lack of specific reagents. Moreover, a great number of obstacles to furthering these studies remain, especially the lack of a high-quality genome and an overall view of gene expression profiling that leave several key questions unanswered: (a) How closely related are tree shrews to primates; (b) do tree shrews share similarity of key signalling pathways to primates and be fully used as an adjunct to primates; and (c) what are the unique biological features of the tree shrew? The answers to these questions provide the information foundation needed to expedite current efforts in making the Chinese tree shrew a viable model animal, and to design and develop new animal models for human diseases, drug screening and safety testing.

In this study, we presented a high-quality genome sequence and the annotation of Chinese tree shrew. Comparison of tree shrew and other genomes, including human, revealed a closer relationship between tree shrew and primates. We identified several genetic features shared between tree shrew and primates, as well as the unique genetic changes that corresponds to their unique biological features. The data provided here are a useful resource for researches using tree shrew as an animal model.

Table 1 | Global statistics of the Chinese tree shrew genome.

	Insert size (bp)	Total data (Gb)	Sequence coverage (X)
(a) Sequencing			
Paired-end library	170–800 bp	187.09	58.47
	2–40 × 10^3	66.36	20.74
	Total	253.45	79.20
	N50 (Kb)	**Longest (Kb)**	**Size (Gb)**
(b) Assembly			
Contig	22	188	2.72
Scaffold	3,656	19,270	2.86
	Number	**Total length (Mb)**	**Percentage of genome**
(c) Annotation			
Repeats	4,843,686	1,001.9	35.01
Genes	22,063	743.4	25.98
CDS	166,392	31.0	1.08

Results

Genome sequencing of the Chinese tree shrew. To address the phylogenetic relationship and genetic divergence of tree shrew and human, and also facilitate the application of the Chinese tree shrew as an animal model for biomedical research, we generated a reference genome assembly from a male Chinese tree shrew (*Tupaia belangeri chinensis*) from Kunming, Yunnan, China. The assembly was generated with 79× high-quality Illumina reads from 19 paired-end libraries with various insert sizes from 170 bp to 40 kb (Supplementary Table S1), and has a contig N50 size of 22 kb and a scaffold N50 size of 3.7 Mb (Table 1). The total assembled size of the genome is about 2.86 Gb, close to the 3.2 Gb genome size estimated from the K-mer calculation (Supplementary Fig. S1). Repetitive elements comprise 35% of the tree shrew genome (Supplementary Table S2). Unlike primate genomes, which are characterized by a large number of Alu/SINE elements, the tree shrew genome has a marginal proportion of this element but contains over a million copies of a tree shrew-specific transfer RNA-derived SINE (Tu-III) family, representing the dominated transposon that makes up 14% of the entire genome (Supplementary Table S3).

To aid the gene annotation of the tree shrew genome, we generated a high-depth transcriptome data from seven tissues including the brain, liver, heart, kidney, pancreas, ovary and testis collected from two Chinese tree shrews (Supplementary Methods 1). The genome was then annotated with a method integrating the homologous prediction, *ab initio* prediction and transcription-based prediction methods (Supplementary Methods 3.2). A non-redundant reference gene set included 22,063 protein-coding genes of which 17,511 genes show one-to-one orthology with other mammals, while the remaining genes display complicated orthologous relationships.

We compared the major parameters of our genome assembly with the recently released tree shrew genome by Broad Institute (http://www.ensembl.org/Tupaia_belangeri/Info/Index; abbreviated as Broad version in the below text), and found that our assembly has great advantages than the Broad version (Supplementary Table S4). First, the Broad version only provided very low coverage (2X) for the tree shrew genome, whereas we offered very high depth (~79X) coverage to guarantee a high accuracy for the genome at the single-base level. Second, our assembly is more complete than the Broad version. The contiguous non-gap sequences covered over 85% of our tree shrew genome, while the Broad version covered <67% of the genome. A more complete assembly allows us to perform a comprehensive analysis for the genomic features of this animal and to systemically compare with other species (see below). Third, our assembly provided over 20 times longer than the Broad version in the scaffold size. The assembly with longer scaffolds and contig scan allows us to produce a more complete individual gene model and a long gene synteny, which is very useful for cross-species comparisons. Finally, with the availability of our high-quality assembly, we generated a significantly improved annotation for the tree shrew genome, which contains 22,063 genes and is closer to the human gene number. In contrast, the gene annotation of the Broad version was based on the homological prediction and only includes 15,414 genes (most of them are partial genes). In addition, our gene models are supported by the high-depth transcriptome data. Over 95% of our gene models have complete open reading frames, while only <40% of gene models in the Broad version are complete. Overall, we provided a high-quality genome together with the well-annotated genes, which would be a very useful resource for the scientific community.

Evolutionary status of the tree shrew. The entire tree shrew genome sequence offers essential information needed to settle ongoing debates on the exact phylogenetic position of this species

in Euarchontoglires[10,11]. Analyses of the mitochondrial genome showed that the tree shrew had a closer relationship to Lagomorpha than to Dermoptera or primates,[11] and molecular cytogenetic data supported a Scandentia–Dermoptera sister clade[10]. However, available evidence from multiple nuclear genes suggests a closer affinity of tree shrews and primates (including human)[12,13]. In a recent study by Hallström *et al.*[14] based on 3,000 genes for phylogenetic analysis, tree shrew was grouped with Glires (including Rodentia and Lagomorpha), suggesting a closer affinity of tree shrew with mouse or rabbits. However, this placement was insufficiently supported thus even unresolved. Genome sequencing of the Chinese tree shrew and comparison with 14 other species, including 6 primate species, on the basis of 2,117 single-copy genes showed that the tree shrew was first clustered with primate species with a high bootstrap support by all phylogenetic signals, including coding sequences with all codon positions and peptide sequences (Fig. 1 and Supplementary Fig. S2). This result helped to clarify potential controversy regarding the phylogenetic position of tree shrew within eutherian mammals reconstructed on the basis of mitochondrial DNA genome[11], genome-wide comparative chromosome map[10] and multilocus nuclear sequences[12,13]. It should be mentioned that we observed an unexpected deep split between our tree shrew and the one sequenced by the Broad Institute (Supplementary Fig. S2). If this was not caused by the potential sequence quality owing to the low coverage of the Broad version, one would expect that the divergence of tree shrew from different geographic regions may be more complex albeit they were grouped as one species (*Tupaia balangeri*).

We estimated the divergence time among these 15 mammalian genomes (Fig. 1). The tree shrew seems to have diverged from the clade encompassing the six primate species around 90.9 million years ago, whereas the rodent clade diverged from the primate clade relatively earlier, around 96.4 million years ago. The close affinity of tree shrews to non-human primates, as demonstrated by the clustering pattern in the phylogenetic tree and relatively smaller divergence time, directly settles controversies regarding the phylogenetic position of tree shrews within Euarchontoglires as well as supports rationale for using tree shrews as an adjunct and alternative to primates as animal models.

Genetic relationship of tree shrews and humans. The genetic basis of primate uniqueness and phenotypic distinctions is under intense scrutiny. The clustering of tree shrew and primates within the Euarchonta clade is consistent with the observation that the tree shrew genes have an overall higher similarity in proteins with humans than rodents (Supplementary Fig. S3). The closer relationship between tree shrew and primates raises an interesting question: what primate genes emerged from the Euarchonta clade and are shared in the tree shrew genome? These genes may encode functional proteins that shape similar phenotypic characters between tree shrews and primates. From multiways gene synteny of humans, tree shrews and mice (Supplementary Methods 4.5), we identified 28 genes previously considered primate specific present in the tree shrew genome that are likely to have originated in the Euarchonta clade (Supplementary Table S5). One such example is the psoriasin protein, a potent chemotactic inflammatory protein that has an important role in the innate defence against bacteria on the surface of the body[15], which has duplicated twice within the Euarchonta and formed three tandem duplicated gene clusters in both tree shrews and other primates, including humans. Another example is the NKG2D–ligand interaction, a powerful mechanism to activate natural killer cells and T cells that regulates immune recognition and responses during infection, cancer and autoimmunity[16]. The NKG2D ligands are induced in response to a variety of stress stimuli but these ligands belong to diverse families in humans and mice[17]. Tree shrews possess the same ligand families as humans, consisting of a major histocompatibility complex (MHC) class I-related chain (*MIC*) gene and the ULBP (UL16-binding protein) family (Supplementary Fig. S4), and they have six members in the ULBP family, similar to humans[18]. This observation suggests that the tree shrews' immune system may employ the same indicators as in humans to cue the elimination of infected, stressed and damaged cells.

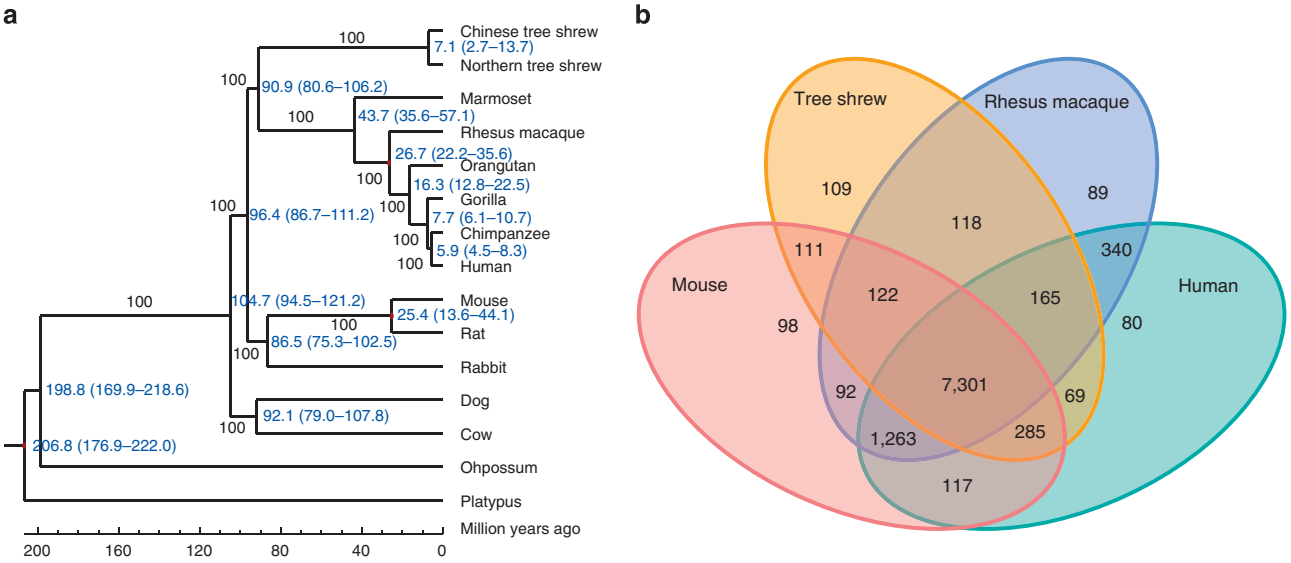


Figure 1 | Relationship of the Chinese tree shrew and related mammals. (**a**) Consensus phylogenetic tree of 15 (sub)species based on 2,117 single-copy genes. The topology was supported by all phylogenetic resources including full-coding sequences, first, second, third codon positions, and amino acids from the orthologous genes. Bootstrap values were calculated from 1,000 replicates and marked in each note. The divergence times for all notes were estimated using three notes with fossil records as calibration times and marked in each note with error range. (**b**) Venn diagram of Chinese tree shrew gene families with human, rhesus macaque and mouse.

Unique genetic features of the tree shrew. By comparing primate and rodent genomes, we identified several lineage-specific genetic changes that potentially contributed to the tree shrews' adaptations. A total of 162 gene families underwent specific expansion in tree shrews (Supplementary Methods) with the immunoglobulin lambda variable gene family showing the most striking expansion, 67 copies in tree shrews but only 36 copies in the human genome (Fig. 2a). The immunoglobulins can block and promote elimination of the pathogen antigens, and accordingly, this expansion could provide an immediate selective advantage to tree shrews. To further investigate specific gene loss or pseudogenization in tree shrews, we compared the gene synteny of the tree shrew, human and mouse genomes. We identified a total of 11 (potential) gene loss and 144 pseudogenes in the tree shrew genome (Supplementary Table S6 and Supplementary Data 1). Of particular interest, the prostate-specific transglutaminase 4 (*TGM4*), which expresses as a seminal fluid protein, was lost in tree shrews. This protein participates in the formation or dissolution of seminal coagulum, a process that has an important role in sperm competition[19]. The absence of *TGM4* may be consistent with the observed tree shrew mating system, for example, *Tupaia tana* species and a few other tupaiids are generally considered behavioural monogamy[1,20], so competitive postmating is lacking in males of this species. Premature stop codon mutations or frame-shift mutations may also lead to functional loss of some important genes in the tree shrew, for example, the *NADPH* oxidase (*NOX1*) gene, which has an important role in cellular defence against acidic stress[21], was disrupted by a premature stop codon in tree shrews. The pseudogenization of this gene suggests that tree shrews may have reduced levels of reactive oxygen species in the arterial wall in conditions like hypertension, hypercholesterolaemia, diabetes and aging, as well as infection.

Nervous system of the tree shrew. Tree shrews have a high brain-to-body mass ratio and a well-developed brain structure resembling primates[1]. Available evidence indicates that tree shrews could be used in depression research[6]. A dominant and subordinate relationship could be created between two male tree shrews in visual and olfactory contact, with the subordinate animal showing a remarkable alteration of physiological, brain

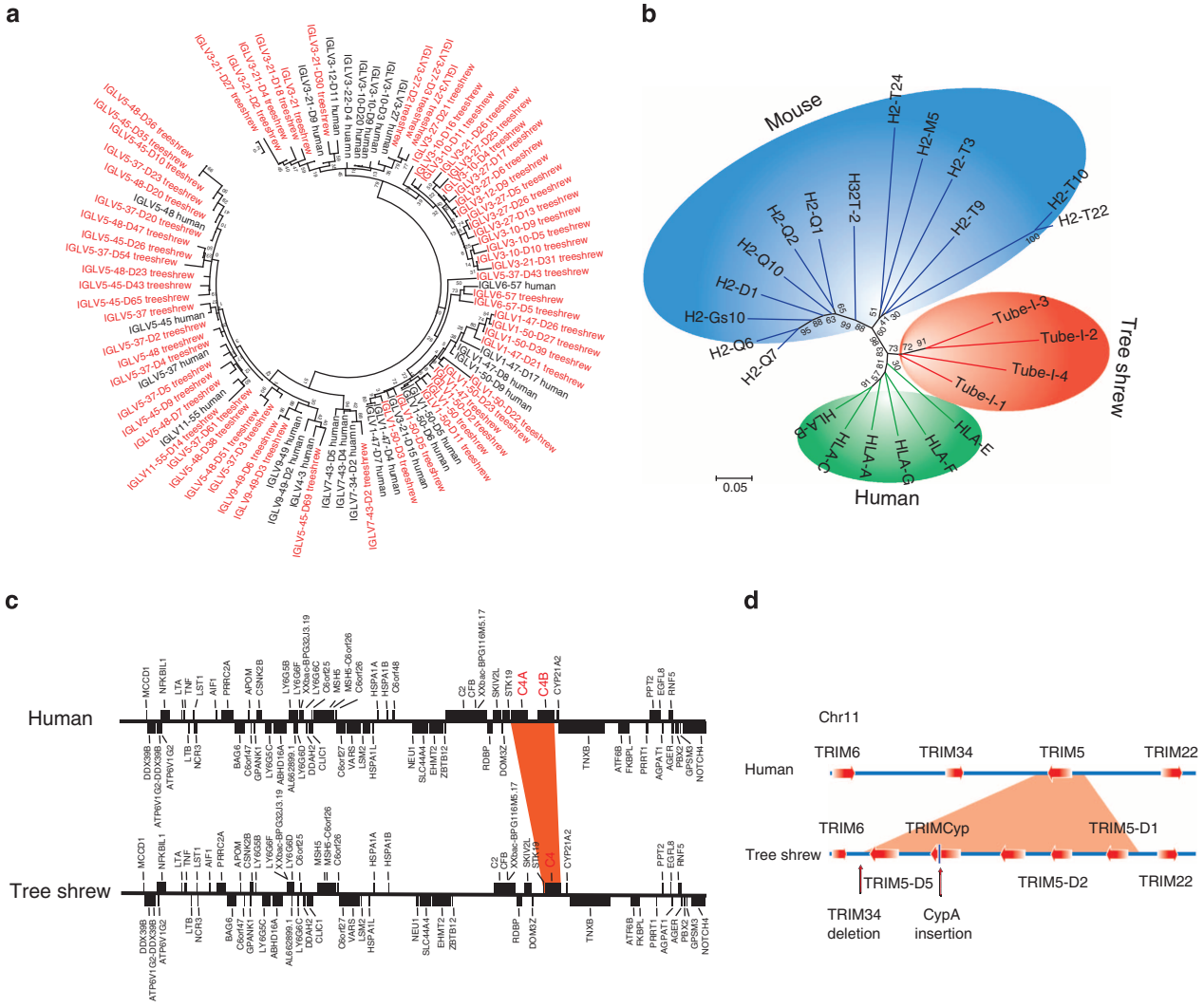


Figure 2 | Immune system in Chinese tree shrew and compared with human and mouse. (**a**) Specific expansion of the immunoglobulin lambda variable (*IGLV*) gene family in the tree shrew. Gene IDs in red are tree shrew genes. (**b**) Phylogenetic relationship of MHC-class I genes in human, tree shrew and mouse. (**c**) Highly conserved gene synteny of MHC-class III region between human and tree shrew. Black bar represents the gene in each species. (**d**) Trim gene cluster in tree shrew and human. Tree shrews have lost *TRIM34* while had multiple specific duplication of *TRIM5*, one of which was inserted by CypA transposon, leading to a fused transcript Trim-Cyp.

functional and behavioural activities that are similar to those observed in depressed patients[6]. In humans, the polymorphism of the serotonin transporter promoter is reputedly associated with the stress disorder and depression susceptibility[22]. However, tree shrews lack this polymorphism region[23], a finding confirmed by our genome sequencing, implying a potentially different regulation of this gene in stress reactions between tree shrews and humans. Excepting this difference, we detected all 23 known neurotransmitter transporters (Supplementary Table S7) in the tree shrew genome that have known roles responsible for the corresponding features of depression[24]. Studies have demonstrated that antidepressants function in patients by suppressing the activity of neurotransmitter transporters[25]. In tree shrews, these transporters are highly conserved in amino-acid sequence with their human counterparts, with the exception of glycine transporter type 1 protein, which shows a relatively fast rate of evolution in the tree shrew lineage (Supplementary Fig. S5). The existence of complete and conserved neuro-transmitter transporters in tree shrews provides a genetic basis for making tree shrews an attractive model for experimental studies of psychosocial stress[6] and evaluation of pharmacological effect of antidepressant drugs.

Similar to primates, tree shrews have an especially well-developed visual system, colour vision and eye structure[1]. A recent study reported that there is a close homology between cholinergic mechanisms in tree shrew and primate visual cortices[26]. Experiments on tree shrews suggest that the subordinate relationship caused by social stress is mediated by visual rather than olfactory cues[27], coinciding with our finding that several olfactory genes have been pseudogenized and the relatively small number of observed olfactory receptors ($n = 690$) in tree shrews as compared with in rodents ($n = \sim 1{,}000$) (Supplementary Methods). The well-developed eye structure of tree shrews has also created substantial interest in using tree shrews as a model in ophthalmological studies, especially for myopia[5].

To provide a genetic basis for the tree shrews' visual system, we systemically scanned the genes involved in visual system. The tree shrew genome encompasses the orthologues of almost all the 209 known visually related human genes, but lacks two cone photoreceptors, the middle wave-length sensitive proteins, which are specifically duplication genes in catarrhines and lead to the trichromacy in higher primates[28]. The absence of the middle wave-length sensitive proteins is consistent with the fact that tree shrews, similar to some lower primates, lack the green pigment and possess dichromats[29]. As most tree shrew species are diurnal and spend the entire night for sleep in their nests, they do not rely on dim-light visuals[29]. The evolutionary rate testing suggested that the rod photoreceptor rhodopsin, which is responsible for the night vision, had a faster evolutionary rate in the tree shrew lineage (Supplementary Fig. S6), suggesting a looser evolutionary constraint of dim-light vision because of their adaptation to the diurnal life. Mutation p.F45L of rhodopsin can cause retinitis pigmentosa, an incurable night blindness disease in humans[30]. Interestingly, we detected a unique p.F45C substitution in tree shrew species (Supplementary Fig. S7), which implies a potentially functional degeneration of this gene in tree shrews. This finding corroborates earlier observations of heavily cone-dominated retina structures with only a small proportion of rod photoreceptors in tree shrews[31]. In addition, we checked the presence of genes regulating the circadian photoreceptor, including both rod–cone photoreceptive systems and non-visual photoreceptive systems, in tree shrew and compared their sequence identity between tree shrew and human. We identified an overall high amino-acid sequence identity (except for enzyme acetylserotonin *O*-methyltransferase) for genes that are involved in photopigment, phototransduction or synthesis of melatonin, which acts as a circadian rhythm regulator[32] (Supplementary Table S8). This pattern may explain why most tree shrews are day-active.

Immune system of the tree shrew. Hepatitis B is an inflammatory liver disease caused by HBV, which has infected about 2 billion people globally and with an annual death toll estimated at 600,000 (ref. 33). Hepatitis C is caused by the HCV, another worldwide infectious disease[34]. Except for chimpanzees, there are many reports that tree shrew and its hepatocytes could be infected with human HBV[4] and HCV[3]. Hence, the property of genes involved in immunity response of viral infection demonstrated by tree shrews further contributes to their preferred choice as an attractive model for studying viral hepatitis and hepatocellular carcinoma[35]. Here, the available tree shrew genome data offer a distinct advantage to scan these immune genes involved in viral hepatitis.

The MHC has a central role in immune responsiveness and susceptibility to various autoimmune and infection diseases. However, so far there is limited information for tree shrew MHC sequences[36,37]. Even though the fragment nature of MHC region and sequencing of the MHC in tree shrew are still incomplete, several points can be distilled from the genome data. First, the entire MHC region of tree shrews is conserved with that of humans, both in the organization of MHC and the gene syntenic order. Second, tree shrews bear at least four genes that belong to MHC class I genes, which are homologous to HLA class I gene and one MIC (Supplementary Fig. S8). Phylogenetic tree analysis clusters tree shrew genes into a separated group diverging from human class I gene group, implying tree shrews have a unique MHC class I locus formed by paralogous amplification (Fig. 2b). Intriguingly, one class I gene in tree shrews is located in the HLA-A region and has well synteny with human locus. However, its functional orthologue with HLA class I gene requires further experimental inspection (Supplementary Fig. S8). The MHC class II region of the tree shrew encompasses homologous of all human class II genes, including the classical class II gene *HLA-DP*, *HLA-DQ* and *HLA-DR*, as well as non-classical class II genes *HLA-DM* and *HLA-DO* (Supplementary Fig. S9 and Supplementary Table S9). The class III region in tree shrews is the most conserved region with humans in gene syntentic alignment. However, in contrast with humans and mice that both obtained two copies of C4 by lineage-specific duplication[38], tree shrews only have one C4 gene in this region (Fig. 2c and Supplementary Fig. S10).

We next investigated the property of gene interaction pathways involved in viral infection. Current studies suggest that a total of 163 human genes were reported to respond in HBV and HCV infection[39,40]. The counterparts of most of those genes are present in the tree shrew genome and shared a relatively high sequence identity with human (Fig. 3 and Supplementary Data 2), with the exception of *DDX58*. Tree shrews have lost *DDX58*, which functions to trigger the transduction cascade involving in the signalling pathway mediated by the *MAVS*, resulting in the activation of NF-κB and is essential for the production of interferon in response to virus, including HCV[41]. The functional loss of *DDX58* in tree shrews suggests that the interruption of immune response may serve as one potential reason causing the capable HCV infection in this animal. Interestingly, other subpathways involved in HCV infection show relatively lower cross-species genetic diversity than that of the MAVS–NF-κB signalling pathway (Fig. 3), in which recurrent viral antagonism has led to a convergent

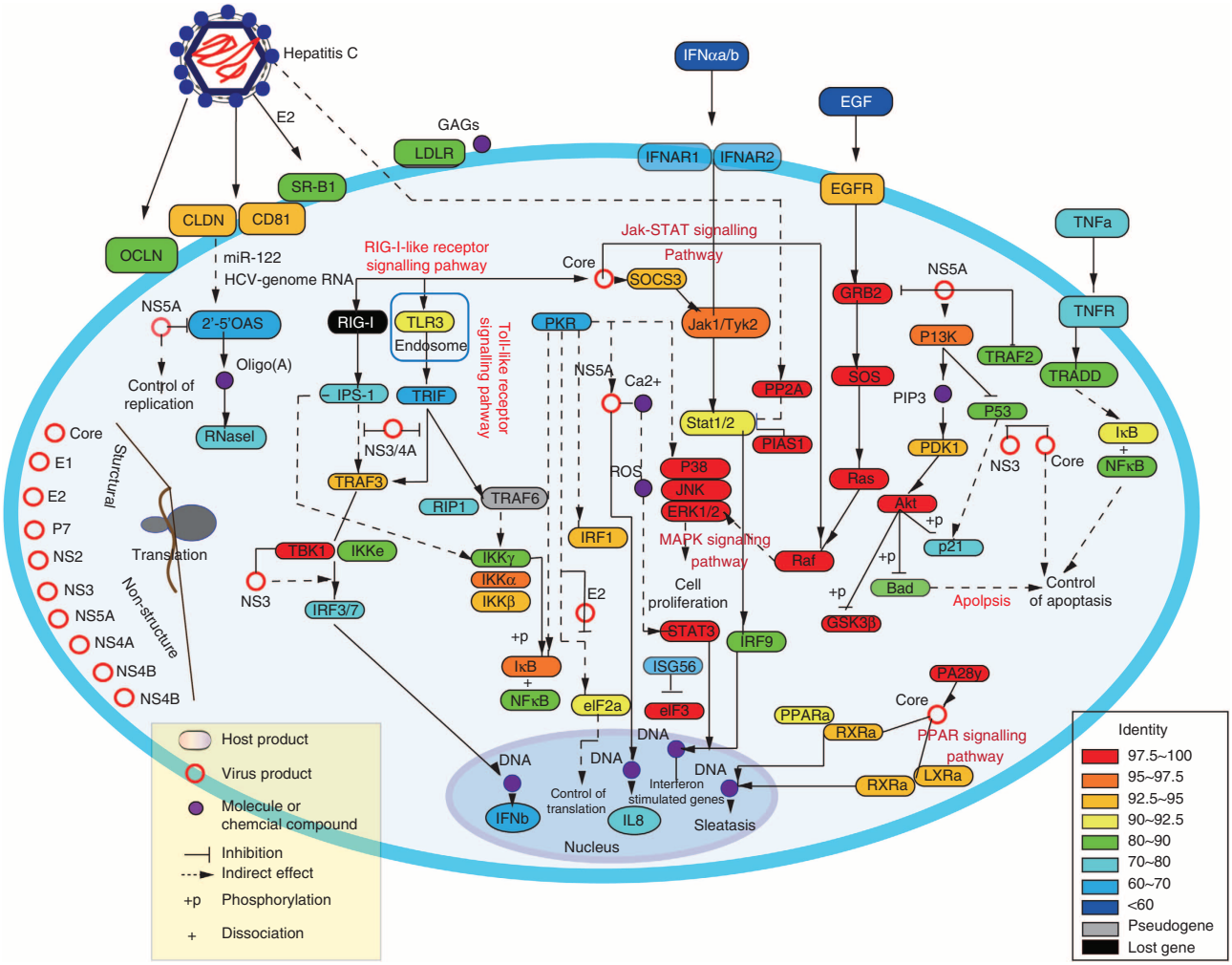


Figure 3 | Genetic divergence of genes involved in HCV infection pathway between human and Chinese tree shrew. Colours represent the degree of sequence identity at the amino-acid level.

evolution of escape from hepaciviral antagonism in primates[42]. Note that a recent study by Tong *et al.*[40] provided functional data that tree shrew CD81, SR-BI, claudin-1 and occludin support HCV infection.

Although HBV is classified as a double-stranded DNA virus, it behaves similarly to a retrovirus and replicates by reverse transcription of an RNA intermediate[43]. *TRIM5*, one of the host restriction factors blocking retroviral replication[44], is located in a gene cluster in human with three other closely related *TRIM* genes, including *TRIM6*, *TRIM34* and *TRIM22*. Genes in this cluster have also been suggested to inhibit the activity of HBV[45]. In the tree shrew, this gene cluster displays a dynamic evolutionary episode (Fig. 2d and Supplementary Fig. S11) as it has achieved five Trim5 copies with four encoding validated open reading frames by several lineage-specific tandem duplication events. Astonishingly, similar to some primates[46,47], one of the *TRIM5* copy has a CypA retrotranposition and form a TrimCyp chimera transcript, which was validated by reverse transcriptase PCR (Supplementary Fig. S12). The appearance of TrimCyp independently in several primate species and tree shrews implies the potential importance of this fused transcript. The *TRIM34* gene in the cluster, which also has function in retrovirus restriction[48], however, has apparently been lost in tree shrews, though tree shrews may potentially have remedied the loss of *TRIM34* with the expansion of *TRIM5*.

The current analysis for all related and essential genes involved in HBV and HCV infection (Fig. 3 and Supplementary Data 2) provided helpful information for us to explain why this animal could be used to create animal model for viral infection. Although we did not provide independent infection experiments (either the animal or primary hepatocytes) to prove its susceptibility, the plenty of previous reports on HBV[4] and HCV[3] infection would certify tree shrews' susceptibility to these viruses. Nonetheless, the findings for the absence of *DDX58* gene and other unique gene features in tree shrew would account for the distinct immune response involved in viral hepatitis.

Drug-targeted domain in tree shrews. The cytochrome P450 (CYP) superfamily encodes the major enzymes involved in drug metabolism, activation and interaction[49]. In general, tree shrews have a more similar number of genes in CYP subfamilies with humans than mice do (Supplementary Table S10). For example, mice have substantially expanded in CYP2 family with 46 members, while humans and tree shrews have fewer copy numbers.

Finally, we sought to assess the genetic divergence of hepatitis drug-targeted genes between tree shrews and humans. A total of 42 genes are known targets for hepatitis drugs, such as halothane, theophylline and meperidine[50,51]. Only one gene, neuropeptide S

receptor 1 (*NPSR1*), has lost its targeted domain (7tm-1) of halothane in tree shrews owing to the frame-shift mutation (Supplementary Fig. S13). All other druggable genetic components can be found in tree shrews and show high conservation in sequence with human orthologues (Supplementary Table S11). The average diversity of the hepatitis drug-targeted domains between humans and tree shrews is about 5%. The conservation of the drug targets, together with the conserved signalling pathways in tree shrews and humans, would encourage the use of tree shrews as a substitution for human patients in pharmacokinetics evaluation of drug disposition, targets and side effects.

Discussion

Despite the fact that tree shrew has been proposed as a valid experimental animal to replace primates in biomedical research and drug safety testing[2], there are limited usages of this animal in the field owing to many reasons. The publicly available annotated genome sequence of the Chinese tree shrew we generated offers an opportunity to decipher the genetic basis of the tree shrews' suitability as an animal model for studying depression, myopia and viral infection[3–7]. Although we did not provide further experimental evidence to solidify the speculations deducted from the comparative genomics, the unique genetic features that we discerned from the genome of Chinese tree shrew has provided insightful information for us to understanding the biology of this animal. By comparing the overall genomic profile of tree shrews and other related mammalians, particularly those of the commonly used experimental animals like rats and mice, we showed that tree shrews had a relatively closer affinity to non-human primates, settling a long-running dispute regarding the phylogenetic position of tree shrew within the Euarchontoglires clade. We likewise characterized the key classes of molecules relevant to the tree shrew nervous and immune systems, demonstrating the genetic basis of this animal as a rising model for biomedical research. The availability of this new genomic data provides a valuable resource and tool for functional genomic and pharmacogenomic studies on tree shrews while also facilitating increasing use of the tree shrew as an animal model in broader fields.

Methods

Source of samples. A male Chinese tree shrew (*Tupaia belangeri chinensis*) from Yunnan, China, was used for DNA isolation and sequencing. RNA samples from the brain, liver, heart, kidney, pancreas and testis of another male Chinese tree shrew and from the ovary of one female individual were collected for transcriptome sequencing. All experiments on animals involved in this study have been approved by the Kunming Institute of Zoology Institutional Review Board.

Genome sequencing and assembly. DNA and RNA sequencing libraries were constructed using standard Illumina libraries prep protocols. Tree shrew genomes were assembled *de novo* by the *de Bruijn* graph-based assembler SOAPdenovo 1.05 (ref. 52). First, low-quality reads or those with potential sequencing errors were removed or corrected by K-mer frequency-based methods. SOAPdenovo1.05 constructed the *de Bruijn* graph by splitting the reads from short insert size libraries (170–800 bp) into 41-mers and then merging the 41-mers, after which the contigs (which exhibited unambiguous connections in *de Bruijn* graphs) were collected. All reads were aligned onto the contigs for scaffold building using the paired-end information. This paired-end information was subsequently used to link contigs into scaffolds, step-by-step, from short insert sizes to long insert sizes. Some intra-scaffold gaps were filled by local assembly using the reads in a read pair, where one end uniquely aligned to a contig while the other end was located within the gap.

Genome annotation. We employed RepeatMasker 3.3.0 (ref. 53) to identify and classify transposable elements by aligning the tree shrew genome sequences against a library of known repeats, Repbase (http://www.girinst.org/repbase/), with default parameters. We used the same pipeline and parameters to re-run the repeat annotation in human, mouse, rat and dog genomes, which were downloaded from Ensembl (release 60). To predict genes in the tree shrew genome, we used both homology-based and *de novo* methods. For the homology-based prediction, human and mouse proteins were downloaded from Ensembl (release 60) and mapped onto the genome using TblastN 2.2.18. Then, homologous genome sequences were aligned against the matching proteins using GeneWise 2-2-0 (ref. 54) to define gene models. For *de novo* prediction, Augustus 2.5.5 (ref. 55) and Genscan 1.0 (ref. 56) were employed to predict coding genes, using appropriate parameters. RNA-seq data were mapped to genome using Tophat 1.4.1 (ref. 57), and transcriptome-based gene structures were obtained by cufflinks 1.3.0 (http://cufflinks.cbcb.umd.edu/). Finally, homology-based, *de novo*-derived gene sets and transcript gene sets were merged to form a comprehensive and non-redundant reference gene set using GLEAN 2.0 (http://sourceforge.net/projects/glean-gene/), removing all genes with sequences <50 amino acid as well as those that only had weak *de novo* support.

Phylogenetic analysis. We used TreeFam 7.0 (http://www.treefam.org/) to define gene families among 15 mammalian genomes: human, chimpanzee, gorilla, orangutan, rhesus macaque, marmoset, Chinese tree shrew, northern tree shrew, rabbit, mouse, rat, dog, cow, opossum and platypus. We carried out the same pipeline and parameters used in our previously published study[58]. We obtained 18,823 gene families and 2,117 single-copy orthologues. The 2,117 single-copy gene families were used to reconstruct the phylogenetic tree. CDS sequences from each single-copy family were aligned by MUSCLE 3.7 (http://www.ebi.ac.uk/Tools/msa/muscle/) with the guidance of aligned protein sequences and concatenated to one super gene for each species. Codons 1, 2, 3 and $1+2$ sequences were extracted from CDS alignments and used as input for building trees, along with protein, CDS sequences. Then, RAxML 7.2.8 (http://sco.h-its.org/exelixis/software.html) was applied for these sequence sets to build phylogenetic trees under the GTR + gamma model for nucleotide sequences and BLOSUM62 + gamma model for protein sequences. We used 1,000 rapid bootstrap replicates to assess the branch reliability in RAxML 7.2.8. Using MCMCtree in PAML 4.4 (ref. 59), we determined split times with approximate likelihood calculation. The alpha parameter for gamma rates at sites was set as that computed by baseml in the initial step. The MCMC process of PAML 4.4 MCMCtree was run to sample 1 million times with sample frequency set to 50, after a burn-in of 5 millions iterations. The 'fine-tune' parameters were set as '0.00356 0.02243 0.00633 0.12 0.43455' to make acceptance proportions fall in interval (20 and 40%). For other parameters we used the defaults. We applied Tracer 1.4 (http://beast.bio.ed.ac.uk/) to check convergence. Two independent runs were performed to confirm the convergence. Gene family expansion analysis was performed using CAFE 2.1 (http://sites.bio.indiana.edu/~hahnlab/Software.html). In CAFE, a random birth and death model was proposed to study gene gain and loss in gene families across a user-specified phylogenetic tree. A global parameter λ, which described both the gene birth (λ) and death ($\mu = -\lambda$) rates across all branches in tree for all gene families was estimated using maximum likelihood. Then, the conditional *P*-value was calculated for each gene family, and families with conditional *P*-values less than threshold (0.05) were considered as having accelerated rate of expansion and contraction.

Shared gene and loss gene identification. To identify genes tree shrews and primates shared, we first elucidated the orthologous relationship among tree shrew, mouse and human proteins. The longest transcript from the Ensembl database (release 60) was chosen to represent each gene with alternative splicing variants. We then subjected all the proteins to BlastP analysis with the similarity cutoff threshold of *e*-value $= 1e^{-5}$. With the human protein set as a reference, we found the best hit for each tree shrew protein in other species, with the criteria that $>30\%$ of the aligned sequence showed an identity above 30%. Reciprocal best-match pairs were defined as orthologues. Then gene order information was used to filter the false-positive orthologues caused by draft genome assembly and annotation. The orthologues not in gene synteny blocks were removed from further analysis. Previously identified primate-specific genes[60] were mapped on to the synteny map. Primate genes with tree shrew orthologues in the synteny map but which were absent in mice were considered candidate-shared genes between primates and tree shrews. We also performed the manual check for all candidate genes. From the synteny map, we also observed genes specifically missing in tree shrews that should have been lost in this species. To further confirm this finding, we manually checked and annotated the genes in the tree shrew genome, and filtered those located in gap regions.

Pseudogene identification. To detect homozygous pseudogenes in the tree shrew genome *in silico*, we first aligned all the human cDNA (Ensembl release-56) onto the tree shrew genomes using BLAT with parameters (–extendThroughN -fine). The best hit regions of each gene with 1-kb flanking sequence were cut down and re-aligned with their corresponding human orthologous protein sequence using GeneWise 2-2-0 (ref. 54) with parameters (-genesf -tfor -quiet), which could help define the detail exon–intron structure of each gene. All genes containing frame shifts or premature stop codons reported by GeneWise were considered candidate pseudogenes. We then carried out a series of filtering processes: (1) the reported frame shifts or premature stop codons were due to the flaw in algorithm of GeneWise that were filtered; (2) the candidate pseudogenes with obvious splicing errors near their frame shifts or premature stop codons were filtered; and (3) the candidate pseudogenes due to assembly error or heterozygosis were filtered.

Finally, we compared all candidate pseudogenes with the alternative splicing forms in human.

References

1. Peng, Y. Z. *et al. Biology of Chinese Tree Shrews (Tupaia belangeri chinensis)* (Yunnan Science and Technology Press, 1991).
2. Cao, J., Yang, E. B., Su, J. J., Li, Y. & Chow, P. The tree shrews: adjuncts and alternatives to primates as models for biomedical research. *J. Med. Primatol.* **32,** 123–130 (2003).
3. Zhao, X. *et al.* Primary hepatocytes of *Tupaia belangeri* as a potential model for hepatitis C virus infection. *J. Clin. Invest.* **109,** 221–232 (2002).
4. Yan, R. Q. *et al.* Human hepatitis B virus and hepatocellular carcinoma. I. Experimental infection of tree shrews with hepatitis B virus. *J. Cancer Res. Clin. Oncol.* **122,** 283–288 (1996).
5. Norton, T. T., Amedo, A. O. & Siegwart, Jr J. T. Darkness causes myopia in visually experienced tree shrews. *Invest. Ophthalmol. Vis. Sci.* **47,** 4700–4707 (2006).
6. Fuchs, E. Social stress in tree shrews as an animal model of depression: an example of a behavioral model of a CNS disorder. *CNS Spectr.* **10,** 182–190 (2005).
7. van Kampen, M., Kramer, M., Hiemke, C., Flugge, G. & Fuchs, E. The chronic psychosocial stress paradigm in male tree shrews: evaluation of a novel animal model for depressive disorders. *Stress* **5,** 37–46 (2002).
8. Yamashita, A., Fuchs, E., Taira, M., Yamamoto, T. & Hayashi, M. Somatostatin-immunoreactive senile plaque-like structures in the frontal cortex and nucleus accumbens of aged tree shrews and Japanese macaques. *J. Med. Primatol.* **41,** 147–157 (2012).
9. Bartolomucci, A., de Biurrun, G., Czeh, B., van Kampen, M. & Fuchs, E. Selective enhancement of spatial learning under chronic psychosocial stress. *Eur. J. Neurosci.* **15,** 1863–1866 (2002).
10. Nie, W. *et al.* Flying lemurs—the 'flying tree shrews'? Molecular cytogenetic evidence for a Scandentia-Dermoptera sister clade. *BMC Biol.* **6,** 18 (2008).
11. Xu, L., Chen, S. Y., Nie, W. H., Jiang, X. L. & Yao, Y. G. 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. *J. Genet. Genomics* **39,** 131–137 (2012).
12. Janecka, J. E. *et al.* Molecular and genomic data identify the closest living relative of primates. *Science* **318,** 792–794 (2007).
13. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
14. Hallstrom, B. M. & Janke, A. Mammalian evolution may not be strictly bifurcating. *Mol. Biol. Evol.* **27,** 2804–2816 (2010).
15. Glaser, R. *et al.* Antimicrobial psoriasin (S100A7) protects human skin from *Escherichia coli* infection. *Nat. Immunol.* **6,** 57–64 (2005).
16. Eagle, R. A. & Trowsdale, J. Promiscuity and the single receptor: NKG2D. *Nat. Rev. Immunol.* **7,** 737–744 (2007).
17. Gleimer, M. & Parham, P. Stress management: MHC class I and class I-like molecules as reporters of cellular stress. *Immunity* **19,** 469–477 (2003).
18. Kondo, M. *et al.* Comparative genomic analysis of mammalian NKG2D ligand family genes provides insights into their origin and evolution. *Immunogenetics* **62,** 441–450 (2010).
19. Brillard-Bourdet, M. *et al.* Amidolytic activity of prostatic acid phosphatase on human semenogelins and semenogelin-derived synthetic substrates. *Eur. J. Biochem.* **269,** 390–395 (2002).
20. Munshi-South, J., Bernard, H. & Emmons, L. Behavioral monogamy and fruit availability in the large treeshrew (*Tupaia tana*) in Sabah, Malaysia. *J. Mammal.* **88,** 1427–1438 (2007).
21. Rueckschloss, U., Duerrschmidt, N. & Morawietz, H. NADPH oxidase in endothelial cells: impact on atherosclerosis. *Antioxid. Redox Signal* **5,** 171–180 (2003).
22. Caspi, A. *et al.* Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* **301,** 386–389 (2003).
23. Lesch, K. P. *et al.* The 5-HT transporter gene-linked polymorphic region (5-HTTLPR) in evolutionary perspective: alternative biallelic variation in rhesus monkeys. *J. Neural. Transm.* **104,** 1259–1266 (1997).
24. Iversen, L. Neurotransmitter transporters and their impact on the development of psychopharmacology. *Br. J. Pharmacol.* **147**(Suppl 1): S82–S88 (2006).
25. Richelson, E. Interactions of antidepressants with neurotransmitter transporters and receptors and their clinical relevance. *J. Clin. Psychiatry* **64**(Suppl 13): 5–12 (2003).
26. Bhattacharyya, A., Biessmann, F., Veit, J., Kretz, R. & Rainer, G. Functional and laminar dissociations between muscarinic and nicotinic cholinergic neuromodulation in the tree shrew primary visual cortex. *Eur. J. Neurosci.* **35,** 1270–1280 (2012).
27. Gould, E., McEwen, B. S., Tanapat, P., Galea, L. A. & Fuchs, E. Neurogenesis in the dentate gyrus of the adult tree shrew is regulated by psychosocial stress and NMDA receptor activation. *J. Neurosci.* **17,** 2492–2498 (1997).
28. Dulai, K. S., von Dornum, M., Mollon, J. D. & Hunt, D. M. The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. *Genome Res.* **9,** 629–638 (1999).
29. Hunt, D. M. *et al.* Molecular evolution of trichromacy in primates. *Vision Res.* **38,** 3299–3306 (1998).
30. Sung, C. H. *et al.* Rhodopsin mutations in autosomal dominant retinitis pigmentosa. *Proc. Natl Acad. Sci. USA* **88,** 6481–6485 (1991).
31. Immel, J. H. *The Tree Shrew Retina: Photoreceptors and Retinal Pigment Epithelium* (University of California, 1981).
32. Hattar, S. *et al.* Melanopsin and rod-cone photoreceptive systems account for all major accessory visual functions in mice. *Nature* **424,** 76–81 (2003).
33. World Health Organization (who.int). Hepatitis B Key Facts.Fact sheet No. 204 Revised August 2008.(updated July 2012) Available from http://www.who.int/mediacentre/factsheets/fs204/en/.
34. Shepard, C. W., Finelli, L. & Alter, M. J. Global epidemiology of hepatitis C virus infection. *Lancet Infect. Dis.* **5,** 558–567 (2005).
35. Yan, R. Q. *et al.* Human hepatitis B virus and hepatocellular carcinoma. II. Experimental induction of hepatocellular carcinoma in tree shrews exposed to hepatitis B virus and aflatoxin B1. *J. Cancer Res. Clin. Oncol.* **122,** 289–295 (1996).
36. Oppelt, C., Wutzler, R. & von Holst, D. Characterisation of MHC class II DRB genes in the northern tree shrew (*Tupaia belangeri*). *Immunogenetics* **62,** 613–622 (2010).
37. Flugge, P., Fuchs, E., Gunther, E. & Walter, L. MHC class I genes of the tree shrew *Tupaia belangeri*. *Immunogenetics* **53,** 984–988 (2002).
38. Blanchong, C. A. *et al.* Genetic, structural and functional diversities of human complement components C4A and C4B and their mouse homologues, Slp and C4. *Int. Immunopharmacol.* **1,** 365–392 (2001).
39. Wang, F. S. Current status and prospects of studies on human genetic alleles associated with hepatitis B virus infection. *World J. Gastroenterol.* **9,** 641–644 (2003).
40. Tong, Y. *et al. Tupaia* CD81, SR-BI, claudin-1, and occludin support hepatitis C virus infection. *J. Virol.* **85,** 2793–2802 (2011).
41. Sumpter, Jr R. *et al.* Regulating intracellular antiviral defense and permissiveness to hepatitis C virus RNA replication through a cellular RNA helicase, RIG-I. *J. Virol.* **79,** 2689–2699 (2005).
42. Patel, M. R., Loo, Y. M., Horner, S. M., Gale, Jr. M. & Malik, H. S. Convergent evolution of escape from hepaciviral antagonism in primates. *PLoS Biol.* **10,** e1001282 (2012).
43. Miller, R. H. & Robinson, W. S. Common evolutionary origin of hepatitis B virus and retroviruses. *Proc. Natl Acad. Sci. USA* **83,** 2531–2535 (1986).
44. Sebastian, S. & Luban, J. TRIM5alpha selectively binds a restriction-sensitive retroviral capsid. *Retrovirology* **2,** 40 (2005).
45. Gao, B., Duan, Z., Xu, W. & Xiong, S. Tripartite motif-containing 22 inhibits the activity of hepatitis B virus core promoter, which is dependent on nuclear-located RING domain. *Hepatology* **50,** 424–433 (2009).
46. Ribeiro, I. P. *et al.* Evolution of cyclophilin A and TRIMCyp retrotransposition in New World primates. *J. Virol.* **79,** 14998–15003 (2005).
47. Newman, R. M. *et al.* Evolution of a TRIM5-CypA splice isoform in old world monkeys. *PLoS Pathog.* **4,** e1000003 (2008).
48. Li, X. *et al.* Unique features of TRIM5alpha among closely related human TRIM family members. *Virology* **360,** 419–433 (2007).
49. Guengerich, F. P. Cytochrome p450 and chemical toxicology. *Chem. Res. Toxicol.* **21,** 70–83 (2008).
50. Kendrick, S. F., Henderson, E., Palmer, J., Jones, D. E. & Day, C. P. Theophylline improves steroid sensitivity in acute alcoholic hepatitis. *Hepatology* **52,** 126–131 (2010).
51. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids. Res.* **39,** D1035–D1041 (2011).
52. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).
53. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4,** Unit 4, 10 (2004).
54. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
55. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**(Suppl 2): ii215–ii225 (2003).
56. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10,** 516–522 (2000).
57. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
58. Kim, E. B. *et al.* Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* **479,** 223–227 (2011).
59. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
60. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* **9,** e1001179 (2011).

Acknowledgements

We thank Professors Wen Wang, Ya-Ping Zhang and Peng Shi for helpful comments regarding this project, and Dr Wen-Hui Lee for preparing RNA samples. This work was funded in part by grants from Chinese Academy of Sciences (KSCX2-EW-R-11, KSCX2-EW-R-12 and KSCX2-EW-J-23), the National 863 Project of China (No. 2012AA021801) and Yunnan Province (2009CI119). X.-T.H., L.X. and Y.-G.Y. were supported by the Strategic Priority Research Program (B) of the Chinese Academy of Sciences (XDB0202).

Author contributions

G.-J.Z., J.W. and Y.-G.Y. managed the project. Y.F., Z.-Y.H., Z.-Q.X., X.-Q.S., Y.-X.C., W.Z., Y.-B.Z., L.Y., D.-D.F., X.-T.J., J.-Q.X., J.X., S.-G.L., Y.-S.L., H.-L.H., J.H., C.-C.C. and L.H. performed the genome assembly, gene annotation, repeats annotation, evolution analysis, transcriptome analysis, pseudogene, immune gene and druggable domain analyses. G.-J.Z. and Y.-G.Y. wrote the manuscript with significant contribution of Y.F., Z.-Y.H. and other authors in list. C.-S.C., X.-T.H., R.L., B.L., Y.-Y.M., Y.-Y.N., L.X., Y.Z., X.-D.Z., Y.-T.Z., J.-M.Z. and Y.-G.Y. financially supported this work, provided many suggestions, revised the manuscript and contributed equally to this work. D.M. performed PCR-based experiments. The following authors were listed in alphabetical order: Chang-Chang Cao, Ce-Shi Chen, Yuan-Xin Chen, Ding-Ding Fan, Jing He, Hao-Long Hou, Li Hu, Xin-Tian Hu, Xuan-Ting Jiang, Ren Lai, Yong-Shan Lang, Bin Liang, Sheng-Guang Liao, Dan Mu, Yuan-Ye Ma, Yu-Yu Niu, Xiao-Qing Sun, Jin-Quan Xia, Jin Xiao, Zhi-Qiang Xiong, Lin Xu, Lan Yang, Yun Zhang, Wei Zhao, Xu-Dong Zhao, Yong-Tang Zheng, Ju-Min Zhou, Ya-Bing Zhu.

Additional information

Accession codes: The Chinese tree shrew whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number ALAR00000000. The version described in this paper is the first version, ALAR01000000. All short read data have been deposited into the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA055299. Raw sequencing data of the transcriptome have been deposited in the Gene Expression Omnibus with the accession number GSE39150.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Fan, Y. *et al.* Genome of the Chinese tree shrew. *Nat. Commun.* 4:1426 doi: 10.1038/ncomms2416 (2013).

Supplementary Information: Genome of the Chinese tree shrew

Yu Fan [a,c †], Zhi-Yong Huang [b,†], Chang-Chang Cao [b,*], Ce-Shi Chen [a,*], Yuan-Xin Chen [b,*], Ding-Ding Fan [b,*], Jing He [b,*], Hao-Long Hou [b,*], Li Hu [b,*], Xin-Tian Hu [a,*], Xuan-Ting Jiang [b,*], Ren Lai [a,*], Yong-Shan Lang [b,*], Bin Liang [a,*], Sheng-Guang Liao [b,*], Dan Mu [a,c,*], Yuan-Ye Ma [a,*], Yu-Yu Niu [a,*], Xiao-Qing Sun [b,*], Jin-Quan Xia [b,*], Jin Xiao [b,*], Zhi-Qiang Xiong [b,*], Lin Xu [a,*], Lan Yang [b,*], Yun Zhang [a,*], Wei Zhao [b,*], Xu-Dong Zhao [a,*], Yong-Tang Zheng [a,*], Ju-Min Zhou [a,*], Ya-Bing Zhu [b,*], Guo-Jie Zhang [b,‡], Jun Wang [b, d, e, ‡], Yong-Gang Yao [a,‡]

[a]Key Laboratory of Animal Models and Human Disease Mechanisms of Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Kunming, Yunnan 650223, China

[b] BGI-Shenzhen, Shenzhen 518083, China

[c] University of Chinese Academy of Sciences, Beijing 100039, China

[d] Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark

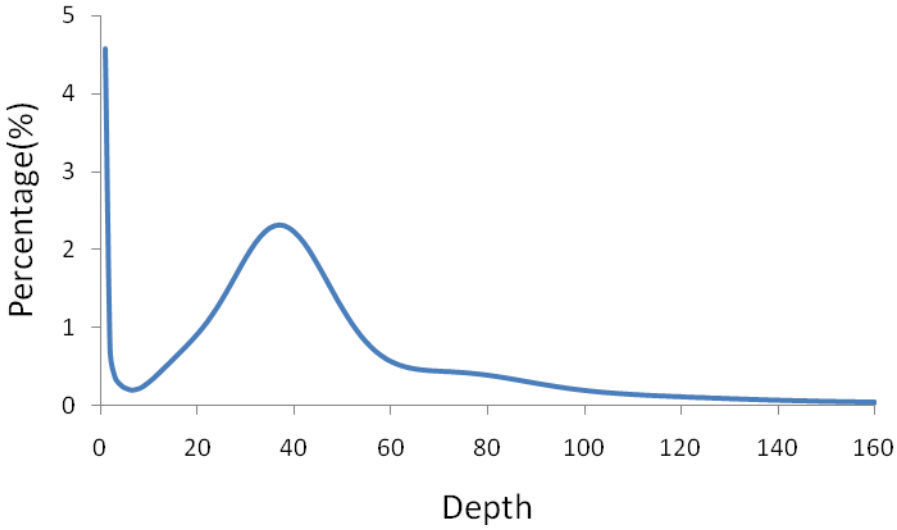
[e] Department of Biology, University of Copenhagen, Copenhagen, Denmark

[†] These authors contributed equally to this work and should be treated as co-first author

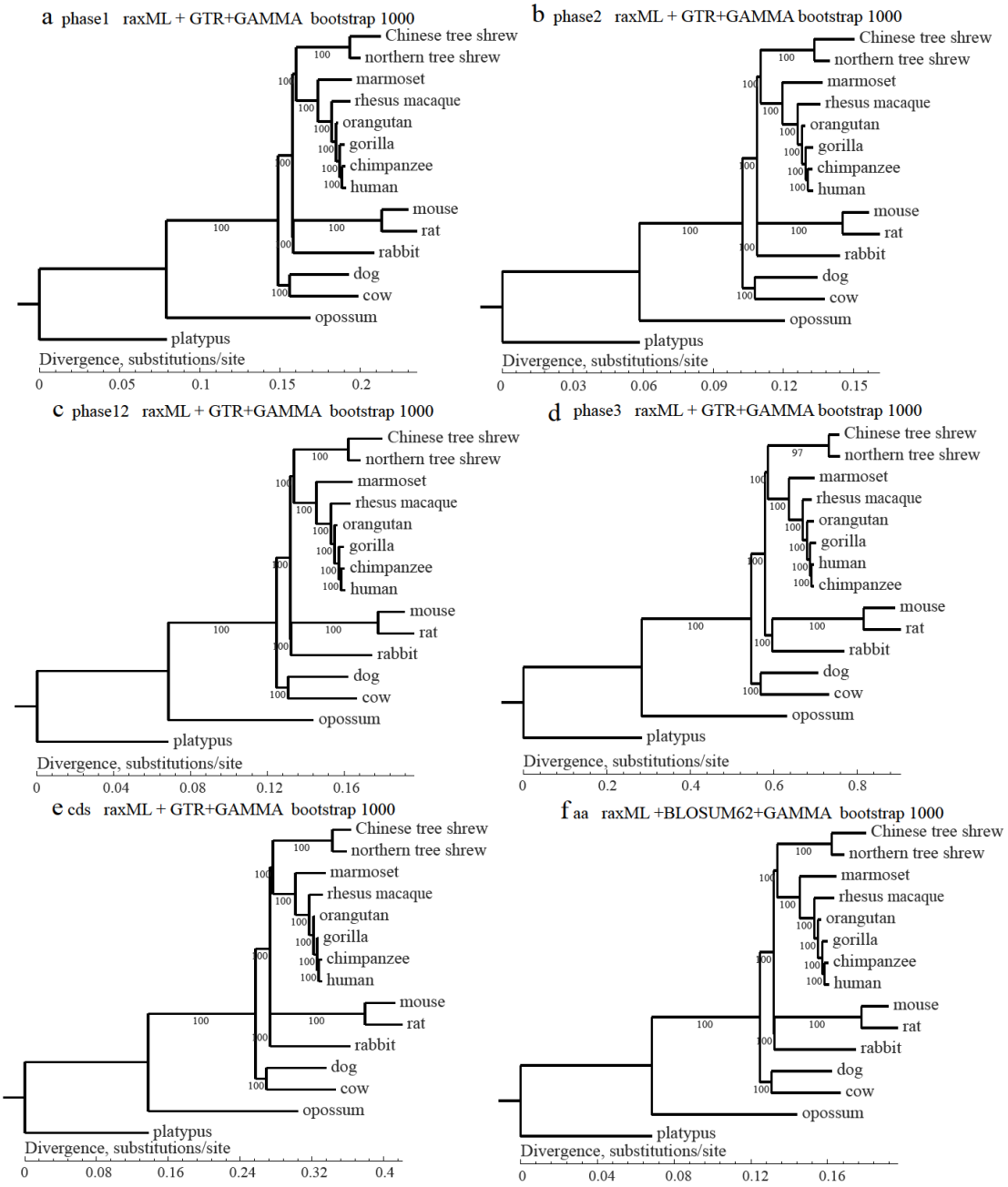
[*] These authors were listed in alphabetical order

[‡] Corresponding authors: Dr. Guo-Jie Zhang: zhanggj@genomics.org.cn; Dr. Jun Wang: wangj@genomics.org.cn; Dr. Yong-Gang Yao: ygyaozh@gmail.com

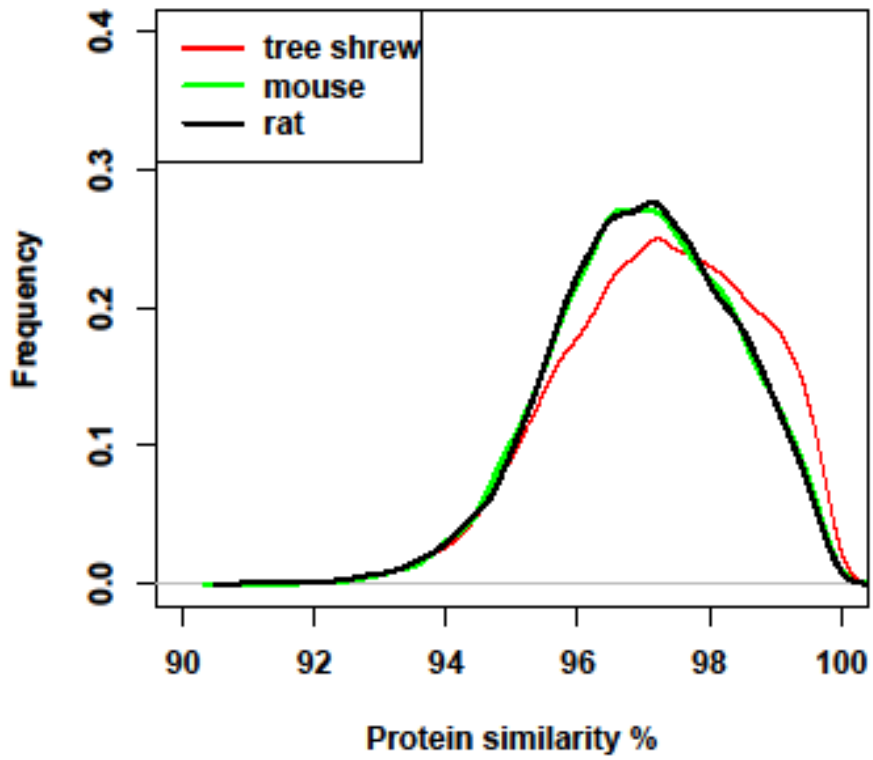
Supplementary Figures



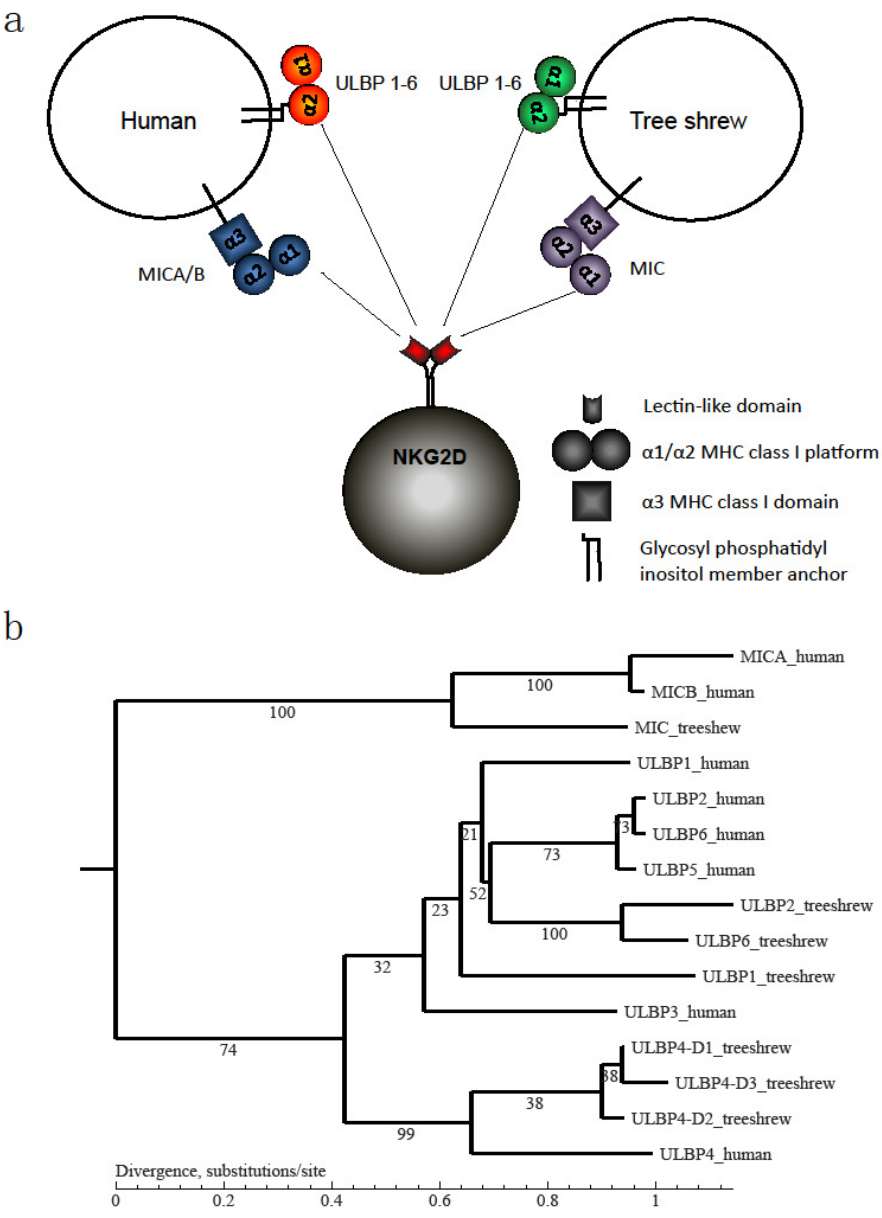
Supplementary Figure S1. 17-k-mer estimation of genome size. Genome size of Chinese tree shrew was estimated to be 3.2Gb based on reads from short insert size libraries.



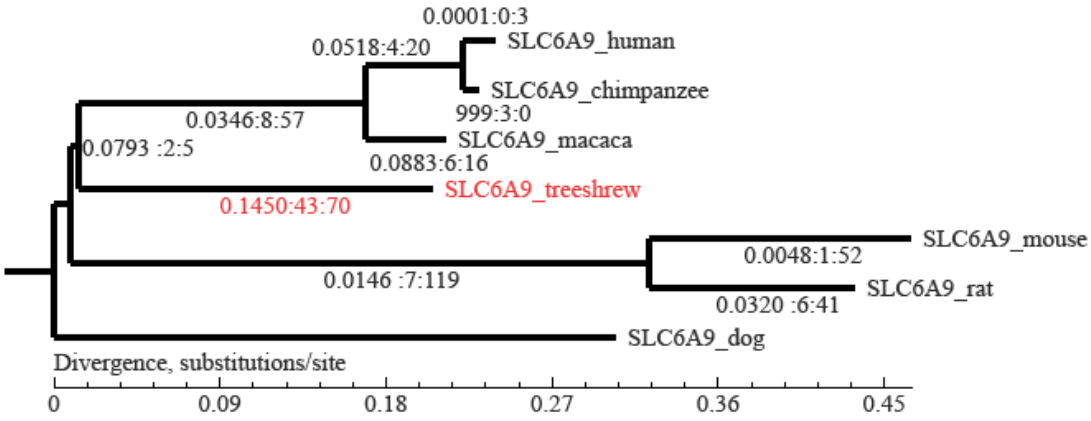
Supplementary Figure S2. Phylogenetic tree of 15 species. Trees were constructed with RAxML under GTR/JTT+GAMMA model based on CDS, peptide, codon1, 2, 3, and 1+2 sequences, bootstrap values are labeled on the species lines. (a) Tree constructed with RAxML under JTT+ GAMMA model based on phase 1 codon. (b) Tree constructed with RAxML under GTR+ GAMMA model based on phase 2 codon. (c) Tree constructed with RAxML under GTR+ GAMMA model based on phase 1 and 2 codon. (d) Tree constructed with RAxML under GTR+ GAMMA model based on phase 3 codon. (e) Tree constructed with RAxML under GTR+ GAMMA model based on CDS. (f) Tree constructed with RAxML under BLOSUM62+ GAMMA model based on protein. The bootstrap values were calculated based on 1,000 replications.



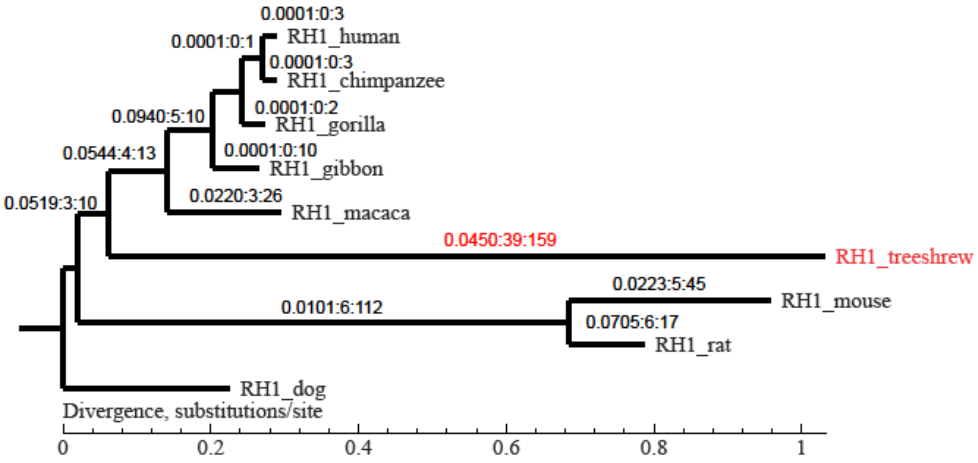
Supplementary Figure S3. Overall protein similarity distribution of 1:1 orthologs between human and tree shrew, mouse and rat. To calculate the percent protein similarity between two sequences, the number of identical residues was divided by the total number of alignment positions.



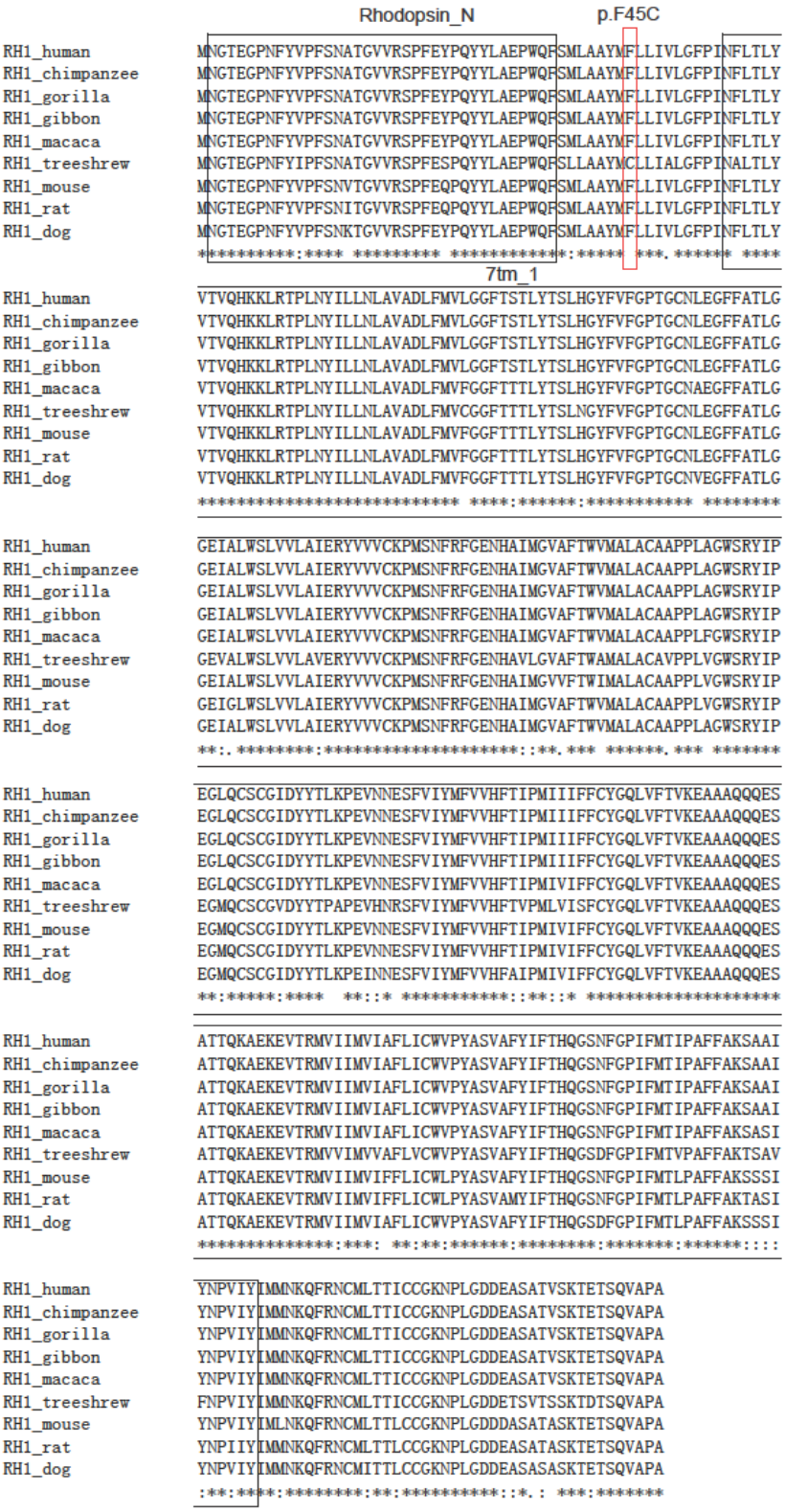
Supplementary Figure S4. NKG2D ligands in human and Chinese tree shrew. (a) ULBP and MIC pathways of human and Chinese tree shrew. (b) Phylogenetic tree of MIC and ULBP gene families in human and Chinese tree shrew. Bootstrap values indicated along the species lines.



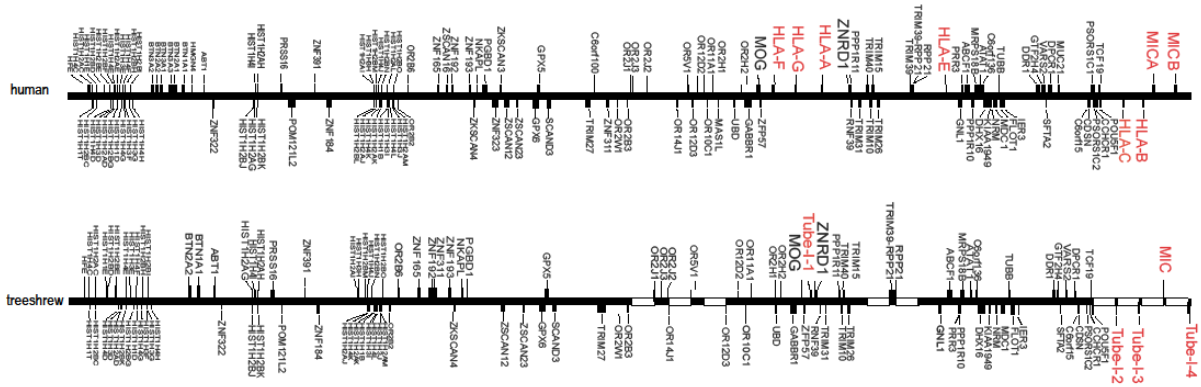
Supplementary Figure S5. Divergence of the *SLC6A9* (*GlyT1*) gene. Branch length is proportional to dS (synonymous substitution rate); dN/dS:N:S of each branch is indicated next to the line. This gene is fast evolved in the Chinese tree shrew compare to other mammals (kaks test [61] p value < 0.05).



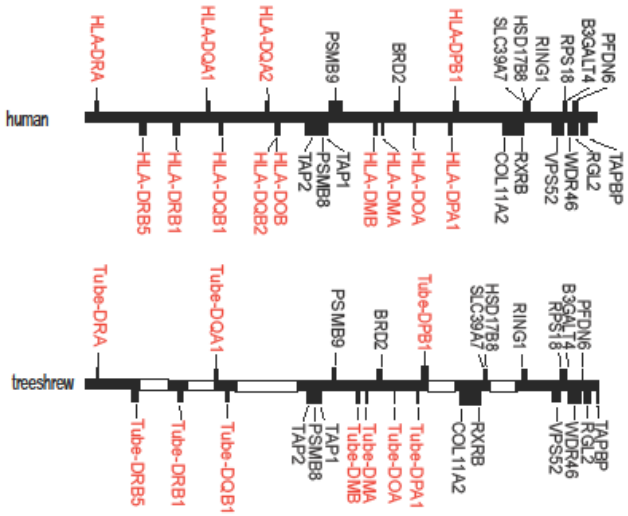
Supplementary Figure S6. Divergence of the *RH1* (*RHO*) gene. Branch length is proportional to dS (synonymous substitution rate); dN/dS:N:S of each branch is indicated next to the line. This gene evolved more quickly in the Chinese tree shrew as compared to other mammals (kaks test [61] p-value < 0.05).



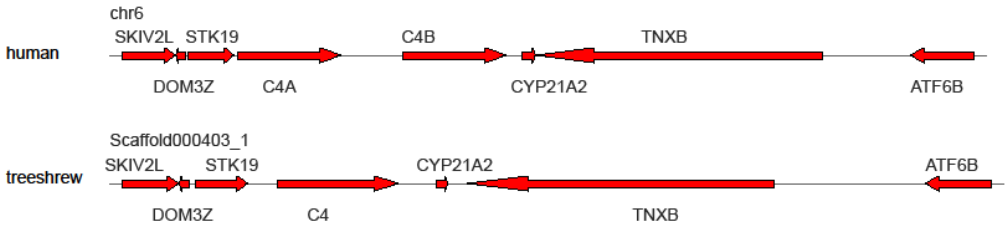
Supplementary Figure S7. Alignment of RH1 in nine mammalian species. Mutant p.F45L in humans may be responsible for autosomal dominant retinitis pigmentosa (ADRP). The Chinese tree shrew has a unique mutation at this site that may be responsible for degeneration of the rod [62].



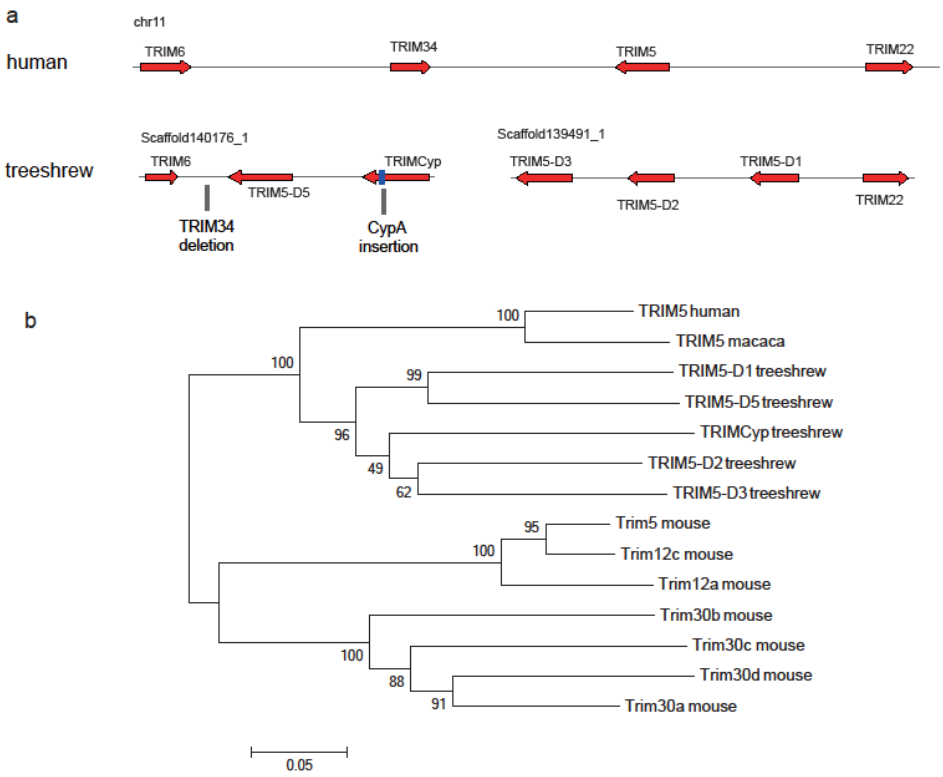
Supplementary Figure S8. Synteny of MHC I between Chinese tree shrew and human. Different scaffolds of the Chinese tree shrew genome are split by white bars; HLA class I gene and MIC are in red.



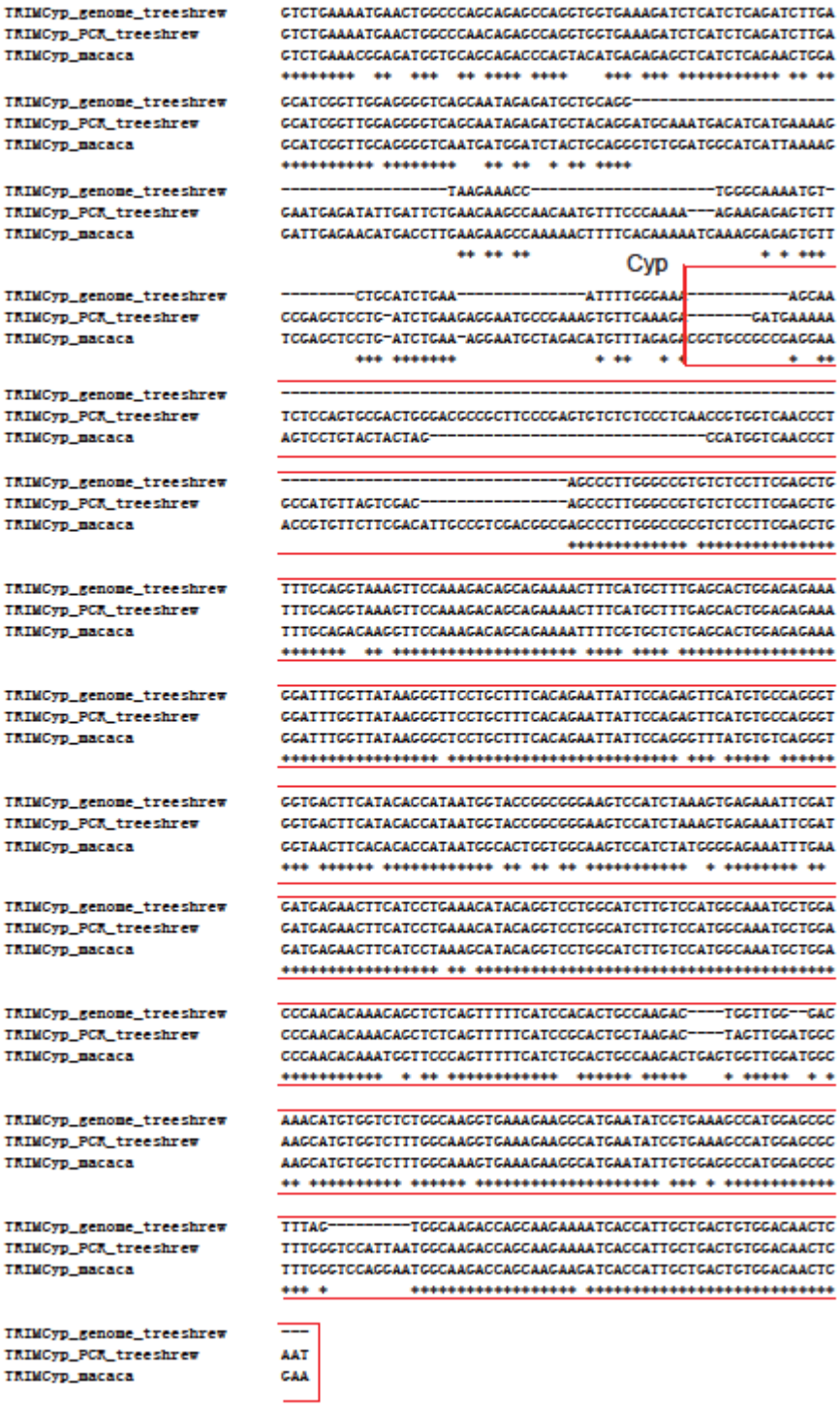
Supplementary Figure S9. Synteny of MHC II between Chinese tree shrew and human. Different scaffolds of the Chinese tree shrew are split by white bars; Class II genes are in red.



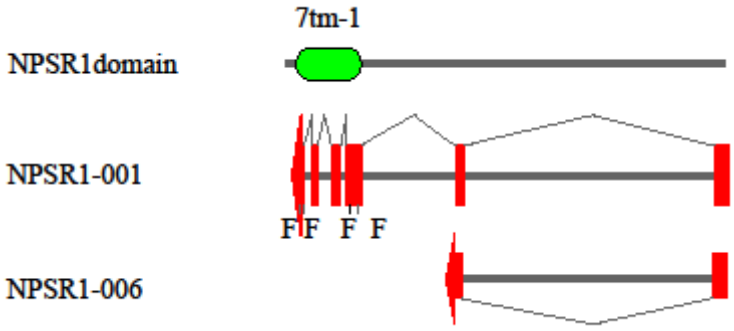
Supplementary Figure S10. Synteny block of the C4 cluster in human and Chinese tree shrew genomes.



Supplementary Figure S11. Chinese tree shrew *TRIM5* gene cluster. (a) *TRIM5* cluster in Chinese tree shrew genome. (b) Phylogenetic tree of the *TRIM5* gene family.



Supplementary Figure S12 Alignment of tree shrew *TRIMCyp* and rhesus macaque (*Macaca mulatta*) *TRIMCyp* genes. TRIMCyp_genome_treeshrew presents Chinese tree shrew *TRIMCyp* cDNA sequence in genome assembly and TRIMCyp_PCR_treeshrew presents the cDNA sequence of Chinese tree shrew *TRIMCyp* by PCR validation.



Supplementary Figure S13. Pseudogenization of *NPSR1* in the Chinese tree shrew genome and loss of 7tm-1 domain. Different transcripts of Chinese tree shrew genes were obtained by the homolog prediction of corresponding human transcripts. F refers to frameshift. The transcript of NPSR1-001 contains a frame shift and a loss of the domain of 7tm-1.

Supplementary Tables

Supplementary Table S1. Data production

Pair-end libraries	Insert Size	Total Data (Gb)	Reads Length (bp)	Sequence coverage (X)	Physical coverage (X)
Solexa Reads	170bp	78.84	100	24.64	20.94
	500bp	57.06	91	17.83	48.99
	800bp	51.19	87	16.00	73.55
	2kb	35.34	49	11.04	225.38
	5kb	15.39	49	4.81	245.38
	10kb	9.22	49	2.88	294.01
	20kb	6	49	1.88	382.65
	40kb	0.41	46	0.13	55.71
Total	-----	253.45	-----	79.20	1,346.60

*assuming the genome size as 3.2 Gb

Supplementary Table S2. Comparison of transposable elements of Chinese tree shrew and other mammalian genomes

Species	Tree shrew		Human		Rhesus macaque		Mouse		Rat		Dog	
TE Class	Length (Mp)	% genome	Length (Mp)	% genome	Length (Mp)	% genome	Length (Mp)	% genome	Length (Mp)	% genome	Length (Mp)	% genome
DNA	76.6	2.7	102.4	3.3	80.1	2.58	63.7	2.3	66.8	2.5	68.3	2.7
LINE	295.2	10.3	543	17.5	395.4	12.77	495.3	18.2	529.5	19.5	418.6	16.54
LTR	113.1	4	257.2	8.3	192.7	6.22	283.9	10.4	220.8	8.1	114.7	4.5
SINE	527.2	18.8	349.4	11.3	279.2	9.02	166.7	6.1	144.6	5.3	226.7	8.9
Other	0.06	0.002	26.4	0.9	5.6	0.18	7.6	0.3	6.8	0.2	0.008	0
Unknown	0.9	0.03	4.8	0.2	2.7	0.09	43.7	1.6	50.7	1.9	1.1	0.04
Total	1,001.90	35	1,257.70	40.5	955.7	30.86	1,017	37.4	978	36	822.8	32.5

Supplementary Table S3. Statistics of transposable elements in six mammalian genomes

Species	Chinese tree shrew		Human		Rhesus macaque		Mouse		Rat		Dog	
TE_Class	Copy Number	% genome	Copy Number	% genome	Copy Number	% genome	Copy Number	% genome	Copy Number	% genome	Copy Number	% genome
SINE/Tu-III	1,854,765	14.33	0	0	0	0	0	0	0	0	0	0
LINE/L1	756,525	9.34	831,144	18.35	683,463	11.42	818,384	17.77	776,312	19.01	845,458	14.88
LTR/ERVL	297,613	2.39	488,118	6	377,608	3.84	472,810	4.64	381,787	3.71	293,343	3.22
DNA/hAT	251,060	1.18	285,124	1.91	250,455	1.31	257,382	0.99	267,001	1	263,952	1.66
DNA/TcMar	179,675	0.94	130,815	1.45	108,408	0.93	65,632	0.29	68,621	0.29	76,337	0.65
SINE/MIR	177,155	0.83	294,277	1.64	246,462	1.14	10,628	0.05	11,165	0.05	254,409	1.42
LTR/ERV1	148,232	1.11	176,672	3.14	148,185	1.92	112,942	1.1	156,084	1.05	97,517	1.04
LINE/L2	130,112	0.77	205,917	1.85	186,712	1.18	56,857	0.28	53,478	0.27	186,192	1.42
DNA/En-Spm	66,221	0.17	67,973	0.22	41,770	0.11	171,360	0.48	185,309	0.52	41,604	0.12
DNA/DNA	43,796	0.14	0	0	20,403	0.06	0	0	0	0	23,850	0.08
LTR/ERVK	37,328	0.31	28,213	0.37	44,133	0.31	278,629	4.4	230,738	3	21,750	0.06
SINE/Alu	33,237	0.11	1,097,378	11.36	882,807	7.79	499,693	2.25	314,869	1.32	88	0
LTR/Gypsy	29,788	0.1	20,855	0.11	35,527	0.12	57,386	0.17	68,499	0.2	37,176	0.15
DNA/Sola	29,191	0.1	25,407	0.1	15,511	0.07	119,518	0.45	140,749	0.56	13,038	0.05
DNA/Maverick	24,748	0.07	16,763	0.04	15,114	0.04	59,360	0.16	66,515	0.18	14,773	0.04
SINE/B4	22,228	0.06	97,447	0.28	79,367	0.18	311,992	1.76	284,608	1.6	136,988	0.45
DNA/MuDR	18,056	0.08	13,428	0.08	9,335	0.04	18,119	0.06	17,820	0.06	9,306	0.03
SINE/tRNA	16,461	0.04	909	0	20,761	0.05	7,788	0.02	8,354	0.02	1,079,722	7.34
LINE/Penelope	15,249	0.08	4,809	0.01	3,604	0.01	28,471	0.09	26,854	0.08	5,166	0.03
LINE/CR1	11,698	0.07	13,527	0.12	19,877	0.11	1,120	0.01	1,274	0.01	22,801	0.15

Supplementary Table S4. Statistics of the Chinese tree shrew genome in our study and Northern tree shrew genome determined by the Broad Institute.

(a) Sequencing

Species	Sequencing method	Sequence coverage (X)	Contiguous non-gap sequences coverage %
Chinese tree shrew	Next-generation sequencing	79.20	85%
Northern tree shrew	Sanger sequencing	2	67%

(b) Assembly

Species	Contig			Scaffold		
	N50 (Kb)	Longest (Kb)	Size (Gb)	N50 (Kb)	Longest (Kb)	Size (Gb)
Chinese tree shrew	22	188	2.72	3,656	19,270	2.86
Northern tree shrew	2.97	184	2.14	175	2,509	3.67

(c) Annotation

Species	Gene number	Complete ORF number	Complete ORF %
Chinese tree shrew	22,063	21,085	95.57
Northern tree shrew	15,414	6,091	39.52

Supplementary Table S5. Shared Chinese tree shrew and primate genes

Human symbol [a]	Gene name	Tree shrew ortholog
AC103810.1	Uncharacterized proteincDNA FLJ60005, moderately similar to Homo sapiens pleckstrin homology domain containing, family M member 1, mRNA	TREES_T100005357.1
CLC	Charcot-Leyden crystal protein	TREES_T100013793.1
CSH2	chorionicsomatomammotropin hormone 2	TREES_T100003396.1
H2BFM	H2B histone family, member M	TREES_T100005641.1
KRTAP10-6	keratin associated protein 10-6	TREES_T100005268.1
LEUTX	leucine twenty homeobox	TREES_T100013805.1
MRGPRX4	MAS-related GPR, member X4	TREES_T100016013.1
OR1A2	olfactory receptor, family 1, subfamily A, member 2	TREES_T100019311.1
OR2T29	olfactory receptor, family 2, subfamily T, member 29	TREES_T100008236.1
OR4A5	olfactory receptor, family 4, subfamily A, member 5	TREES_T100015967.1
OR4C16	olfactory receptor, family 4, subfamily C, member 16	TREES_T100010814.1
OR4F16	olfactory receptor, family 4, subfamily F, member 16	TREES_T100004225.1
OR4Q3	olfactory receptor, family 4, subfamily Q, member 3	TREES_T100019426.1
OR5M8	olfactory receptor, family 5, subfamily M, member 8	TREES_T100000291.1
OR7D4	olfactory receptor, family 7, subfamily D, member 4	TREES_T100002534.1
OR8B2	olfactory receptor, family 8, subfamily B, member 2	TREES_T100009103.1
RASA4B	RAS p21 protein activator 4B	TREES_T100021318.1
S100A7L2	S100 calcium binding protein A7-like 2	TREES_T100022115.1
SIGLEC9	sialic acid binding Ig-like lectin 9	TREES_T100007284.1
TRBV20OR9-2	T cell receptor beta variable 20/OR9-2 (non-functional)	TREES_T100006136.1
TRBV21OR9-2	T cell receptor beta variable 21/OR9-2 (non-functional)	TREES_T100006137.1
UGT2B28	UDP glucuronosyltransferase 2 family, polypeptide B28	TREES_T100005661.1
ZNF433	zinc finger protein 433	TREES_T100000332.1
DHRS4L2	dehydrogenase/reductase (SDR family) member 4 like 2	TREES_T100008210.1
FOLR3	folate receptor 3 (gamma)	TREES_T100009725.1
APOBEC3G	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G	TREES_T100012038.1
ULBP2	UL16 binding protein 2	TREES_T100018665.1
PYDC1	PYD (pyrin domain) containing 1	TREES_T100001274.1

[a] These "primate-specific" genes are human genes and the ancestors of these genes were originated in the Euarchonta clade.

Supplementary Table S6. Lost genes in the Chinese tree shrew

Human symbol	Gene name	KEGG class
CLEC9A	C-type lectin domain family 9, member A	NA
C1orf110	chromosome 1 open reading frame 110	NA
RXFP1	relaxin/insulin-like family peptide receptor 1	Signaling molecules and interaction
TGM4	transglutaminase 4 (prostate)	NA
CPNE7	copine VII	NA
TOP1MT	topoisomerase (DNA) I, mitochondrial	Replication and repair
SLC17A9	solute carrier family 17, member 9	Membrane transport
PLD4	phospholipase D family, member 4	NA
PRIC285	Peroxisomal proliferator-activated receptor A-interacting complex 285 kDa protein	NA
TRIM34	tripartite motif containing 34	Folding, sorting and degradation
DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	Immune system

Note - The lost gene was defined by comparing Chinese tree shrew to human and mouse. If one gene presents in both human and mouse but is absent in tree shrew, it is considered as a gene specifically lost in Chinese tree shrew.

Supplementary Table S7. Sequence identity of neurotransmitter transporters in Chinese tree shrews and mice relative to humans

Human symbol	Gene name	Tree shrew identity	Mouse identity
SLC17A6(VGLUT2)	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 6	98.61	97.42
SLC17A7(VGLUT1)	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 7	97.03	98.7
SLC17A8(VGLUT3)	solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 8	95.45	93.03
SLC18A1(VMAT1)	solute carrier family 18 (vesicular monoamine), member 1	83.82	82.25
SLC18A2(VMAT2)	solute carrier family 18 (vesicular monoamine), member 2	92.07	92.41
SLC18A3(VAChT)	solute carrier family 18 (vesicular acetylcholine), member 3	94.37	94.13
SLC1A1(EAAT3)	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1	92.08	90.08
SLC1A2(EAAT2)	solute carrier family 1 (glial high affinity glutamate transporter), member 2	90.36	94.9
SLC1A3(EAAT1)	solute carrier family 1 (glial high affinity glutamate transporter), member 3	97.56	96.38
SLC1A6(EAAT4)	solute carrier family 1 (high affinity aspartate/glutamate transporter), member 6	88.25	90.3
SLC1A7(EAAT5)	solute carrier family 1 (glutamate transporter), member 7	92.96	92.32
SLC29A1(ENT1)	solute carrier family 29 (nucleoside transporters), member 1	91.42	79.87
SLC29A2(ENT2)	solute carrier family 29 (nucleoside transporters), member 2	91.78	89.04
SLC29A3(ENT3)	solute carrier family 29 (nucleoside transporters), member 3	78.98	73.89
SLC29A4(ENT4)	solute carrier family 29 (nucleoside transporters), member 4	90.53	90.95
SLC32A1(VGAT)	solute carrier family 32 (GABA vesicular transporter), member 1	98.47	98.67
SLC6A1(GAT1)	solute carrier family 6 (neurotransmitter transporter, GABA), member 1	94.18	98
SLC6A11(GAT3)	solute carrier family 6 (neurotransmitter transporter, GABA), member 11	95.75	94.15
SLC6A12(BGT1)	solute carrier family 6 (neurotransmitter transporter, betaine/GABA), member 12	86.63	87.79
SLC6A2(NET)	solute carrier family 6 (neurotransmitter transporter, noradrenalin), member 2	96.03	94.27
SLC6A3(DAT)	solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	89.66	93.49
SLC6A4(SERT)	solute carrier family 6 (neurotransmitter transporter, serotonin), member 4	95.56	92.09
SLC6A5(GlyT2)	solute carrier family 6 (neurotransmitter transporter, glycine), member 5	91.76	94.1
SLC6A9(GlyT1)	solute carrier family 6 (neurotransmitter transporter, glycine), member 9	90.11	95.49

Supplementary Table S8. Sequence identity of circadian photoreceptor related genes in Chinese tree shrew relative to human

Human symbol	Gene name	Class	Tree shrew identity
OPN4	opsin 4; melanopsin	non-visual photosensitive systems; photopigment	84.03
CRY1	cryptochrome 1 (photolyase-like)	non-visual photosensitive systems; photopigment	98.45
CRY2	cryptochrome 2 (photolyase-like)	non-visual photosensitive systems; photopigment	97.48
GNAT1	guanine nucleotide-binding protein G(t) subunit alpha-1	rod-cone photosensitive systems; mediates phototransduction	90.96
CNGA3	cyclic GMP-gated channel A-subunit 3	rod-cone photosensitive systems; mediates phototransduction	84.74
TPH1	tryptophan hydroxylase 1	regulate circadian rhythms; melatonin synthesis	93.03
TPH2	tryptophan hydroxylase 2	regulate circadian rhythms; melatonin synthesis	97.73
DDC	dopa decarboxylase (aromatic L-amino acid decarboxylase)	regulate circadian rhythms; melatonin synthesis	90.73
AANAT	aralkylamine N-acetyltransferase	regulate circadian rhythms; melatonin synthesis	85.98
ASMT	acetylserotonin O-methyltransferase	regulate circadian rhythms; melatonin synthesis	69.16
MTNR1A	melatonin receptor 1A	regulate circadian rhythms; melatonin receptors	86.59
MTNR1B	melatonin receptor 1B	regulate circadian rhythms; melatonin receptors	84.76

Supplementary Table S9. Sequence identity of the classical and non-classical MHC class II genes in the Chinese tree shrew and human genomes

Human	Tree shrew ortholog	Identity
HLA-DRA	Tube-DRA	81.42
HLA-DRB5	Tube-DRB5	77.43
HLA-DRB1	Tube-DRB1	76.23
HLA-DQA1	Tube-DQA1	82.14
HLA-DQB1	Tube-DQB1	75.67
HLA-DQA2	genome gap	NA
HLA-DQB2	genome gap	NA
HLA-DOB	genome gap	NA
HLA-DMB	Tube-DMB	74.51
HLA-DMA	Tube-DMA	84.44
HLA-DOA	Tube-DOA	75.97
HLA-DPA1	Tube-DPA1	75.79
HLA-DPB1	Tube-DPB1	85.58

Supplementary Table S10. Copy number of CYP gene subfamily in human, Chinese tree shrew and mouse

Gene family	Human	Tree shrew	Mouse
CYP1	3	3	3
CYP2	16	20	46
CYP3	4	1	9
CYP4	12	10	12
CYP5	1	1	1
CYP7	2	2	2
CYP8	2	2	2
CYP11	3	3	3
CYP17	1	1	1
CYP19	1	1	1
CYP20	1	1	1
CYP21	1	1	2
CYP24	1	1	1
CYP26	3	3	3
CYP27	3	3	2
CYP39	1	1	1
CYP46	1	1	1
CYP51	1	1	1

Supplementary Table S11. Sequence identity of hepatitis drug targeted genes in the Chinese tree shrew relative to human

Human symbol	Gene name	Drug	Tree shrew identity
GNG2	guanine nucleotide binding protein (G protein), gamma 2	Halothane	100
KCNJ6	potassium inwardly-rectifying channel, subfamily J, member 6	Halothane	98.57
GRIN2D	glutamate receptor, ionotropic, N-methyl D-aspartate 2D	Meperidine	98.39
GLRA1	glycine receptor, alpha 1	Halothane	98.63
PDE5A	phosphodiesterase 5A, cGMP-specific	Theophylline	97.65
IMPDH1	IMP (inosine 5'-monophosphate) dehydrogenase 1	Ribavirin	98.64
TOP1	topoisomerase (DNA) I	Sodium	98.65
PDE4B	phosphodiesterase 4B, cAMP-specific	Theophylline	98.27
ADORA2A	adenosine A2a receptor	Theophylline	88.51
KCNMA1	potassium large conductance calcium-activated channel, subfamily M, alpha member 1	Halothane	99.19
GRIN3A	glutamate receptor, ionotropic, N-methyl-D-aspartate 3A	Halothane	94.77
KCNJ3	potassium inwardly-rectifying channel, subfamily J, member 3	Halothane	98.29
GRIN2B	glutamate receptor, ionotropic, N-methyl D-aspartate 2B	Meperidine	99.03
OPRK1	opioid receptor, kappa 1	Meperidine	95.6
GRIN1	glutamate receptor, ionotropic, N-methyl D-aspartate 1	Meperidine	95.51
KCNN4	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4	Halothane	91.98
ADORA2B	adenosine A2b receptor	Theophylline	91.01
KCNK9	potassium channel, subfamily K, member 9	Halothane	93.75
ADK	adenosine kinase	Ribavirin	90.79
PDE3A	phosphodiesterase 3A, cGMP-inhibited	Theophylline; Aminophylline	92.49

GRIN2A	glutamate receptor, ionotropic, N-methyl D-aspartate 2A	Meperidine; Halothane	96.52
CPT2	carnitinepalmitoyltransferase 2	Perhexiline	89.45
RHO	rhodopsin	Halothane	88.73
KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2	Amiodarone	96.81
GRIN2C	glutamate receptor, ionotropic, N-methyl D-aspartate 2C	Meperidine	92.3
GABRA1	gamma-aminobutyric acid (GABA) A receptor, alpha 1	Halothane	98.89
CPT1A	carnitinepalmitoyltransferase 1A (liver)	Perhexiline	89.75
PDE4A	phosphodiesterase 4A, cAMP-specific	Theophylline	94.26
ENPP1	ectonucleotidepyrophosphatase/phosphodiesterase 1	Ribavirin	89.16
ATP2C1	ATPase, Ca++ transporting, type 2C, member 1	Halothane	94.18
ATP5D	ATP synthase, H+ transporting, mitochondrial F1 complex, delta subunit	Halothane	91.06
NPSR1	neuropeptide S receptor 1	Halothane	70.93
NT5C2	5'-nucleotidase, cytosolic II	Ribavirin	99.27
TPO	thyroid peroxidase	Methimazole	73.55
CACNA1H	calcium channel, voltage-dependent, T type, alpha 1H subunit	Amiodarone	84.17
ADORA3	adenosine A3 receptor	Aminophylline	79.1
KCNK3	potassium channel, subfamily K, member 3	Halothane	95.08
ADORA1	adenosine A1 receptor	Theophylline; Aminophylline	97.54
ADRB1	adrenergic, beta-1-, receptor	Amiodarone	96.1
CACNA2D2	calcium channel, voltage-dependent, alpha 2/delta subunit 2	Amiodarone	96.87
GRIN3B	glutamate receptor, ionotropic, N-methyl-D-aspartate 3B	Halothane	80.74
MT-ND1	mitochondrially encoded NADH dehydrogenase 1	Halothane	73.82

Supplementary Methods

1 Source of samples

A male Chinese tree shrew (*Tupaia belangeri chinensis*) from Yunnan, China was used for DNA isolation and sequencing. RNA samples from the brain, liver, heart, kidney, pancreas, and testis of another male Chinese tree shrew and from ovary of one female individual (both individuals were taken from our inbreeding population; coefficient of inbreeding, $F = 0.25$), were collected for transcriptome sequencing. All experiments on animals involved in this study have been approved by the Kunming Institute of Zoology institutional review board.

2 Genome sequencing and assembly

2.1 Genome sequencing

To sequence the genome of Chinese tree shrew, we applied a whole genome shotgun strategy and next-generation sequencing technologies using Illumina HiSeq 2000. In order to reduce the risk of non-randomness, we constructed 19 paired-end libraries by sequencing 35 lanes with insert sizes of about 170 base pairs (bp), 350 bp, 500 bp, 800 bp, 2 kbp, 5 kbp, 10 kbp, 20 kbp and 40 kbp. In total, we generated around 365.46Gb of sequence, and 253.45Gb (79x coverage) was retained for assembly after filtering out low-quality and duplicated reads.

2.2 Estimation of genome size using k-mer

A k-mer refers to an artificial sequence division of K nucleotides. A raw sequence read with L bp contains (L-K+1) k-mers if the length of each k-mer is K bp. The frequency of each k-mer can be calculated from the genome sequence reads. Typically, k-mer frequencies plotted against the sequence depth gradient follow a Poisson distribution in any given dataset; whereas sequencing errors may lead to a higher representation of low frequencies. The genome size, G, can be calculate from the formula G=K_num/K_depth, where the K_num is the total number of k-mers, and K_depth denotes the frequency occurring more frequently than the other frequencies. In this work, K was 17, K_num was 118,076,481,727 and K_depth was 37; we therefore estimated the tree shrew genome size to be 3.2 Gb.

2.3 Genome assembly

The genome was *de novo* assembled by SOAPdenovo 1.05 [52] (http://soap.genomics.org.cn), a novel short-read assembly method that employs the *de Bruijn* graph algorithm in order to both simplify the task of assembly and reduce computational complexity. First, low-quality reads with or those with potential sequencing errors were removed or corrected using the k-mer frequency. After these quality control and filtering steps, a total of 253.45Gb (or 79X) data were retained for

assembly. SOAPdenovo first constructed the *de Bruijn* graph by splitting the reads from short insert size libraries (170-800 bp) into 41-mers and then merging the 41-mers; contigs which exhibit unambiguous connections in *de Bruijn* graphs were then collected. All reads were aligned on to the contigs for scaffold building using the paired-end information. This paired-end information was then subsequently used to link contigs into scaffolds, step-by-step, from short to long insert sizes. About 187.09Gb (or 57X) data were used to build contigs, while all high-quality data were used to build scaffolds. Some intra-scaffold gaps were filled by local assembly using the reads in a read-pair, where one end uniquely aligned to a contig while the other end was located within the gap. The final total contig size and N50 were 2.72 bp and 22Kbp, respectively. The total scaffold size and N50 were 2.86 Gb and 3.66 Mbp, respectively (Table 1)

3 Genomic feature

3.1 Repeat annotation

We employed RepeatMasker 3.3.0 [53] to identify and classify transposable elements (TEs) by aligning the tree shrew genome sequences against a library of known repeats, Repbase (http://www.girinst.org/repbase/), with default parameters. To better compare the tree shrew genome with other mammals, we used the same pipeline and parameters to re-run the repeat annotations on the human, rhesus macaque, mouse, rat and dog genomes which were downloaded from Ensembl (release 60), as shown in Supplementary Table S2 and Supplementary Table S3.

3.2 Gene annotation

To predict genes in the tree shrew genome, we used both homology-based and *de novo* methods. For the homology-based prediction, human and mouse proteins were downloaded from Ensembl (release 60) and mapped onto the genome using TblastN 2.2.18 [63]. Homologous genome sequences were then aligned against the matching proteins using GeneWise 2-2-0 [54] to define gene models. For *de novo* prediction, Augustus 2.5.5 [55] and Genescan 1.0[56] were employed to predict coding genes, using appropriate parameters. RNA-seq data were mapped to genome using Tophat 1.4.1 [57], and transcriptome-based gene structures were obtained by cufflinks 1.3.0 (http://cufflinks.cbcb.umd.edu/). Finally, homology-based, *de novo* derived gene sets and transcript gene sets were merged to form a comprehensive and non-redundant reference gene set using GLEAN 2.0 (http://sourceforge.net/projects/glean-gene/), removing all genes with sequences less than 50 amino acid as well as those that only had weak *de novo* support. We obtained a reference tree shrew gene set containing 22,063 genes. To determine the orthologous relationship between tree shrew and other mammalian proteins, nucleotide and protein data of five mammals (human, rhesus macaque, mouse, rat and dog) were downloaded from Ensembl (release 60). For genes with alternative splicing variants, the longest transcript was selected to represent each gene. We then subjected human, rhesus macaque, mouse, rat, dog and tree shrew

proteins to BlastP analysis with a similar cutoff threshold of e=1e^{-5}. Using the tree shrew protein set as a reference, we found the best hit for each tree shrew protein in other species, with the criteria that more than 30% of the aligned sequence showed an identity above 30%. Reciprocal best-match pairs were defined as 1:1 orthologs. In total, we detected 17,511 of those 1:1 orthologs between tree shrews and other mammals.

We compared the major parameters of our genome assembly and annotation with the recently released tree shrew genome by Broad Institute. As shown in Supplementary Table S4, our assembly has better quality.

4 Gene evolution

4.1 Gene family cluster

A gene family is a group of similar genes descended from a single gene in the last common ancestor of the targeted species. In this study, we used TreeFam 7.0 (http://www.treefam.org/) [64] to define gene families among 15 mammalian genomes (human, chimpanzee, gorilla, orangutan, rhesus macaque, marmoset, Chinese tree shrew, northern tree shrew, rabbit, mouse, rat, dog, cow, opossum and platypus). We carried out the same pipeline and parameters we used in our previously published studies [58]. We obtained 18,823 gene families and 2,117 single-copy orthologs.

4.2 Phylogenetic analysis

We constructed a phylogenetic tree of the Chinese tree shrew and the 14 other mammalian genomes (human, chimpanzee, gorilla, orangutan, rhesus macaque, marmoset, northern tree shrew, rabbit, mouse, rat, dog, cow, opossum and platypus). Totally 2,117 single-copy gene families obtained before were used to reconstruct the phylogenetic tree. Coding sequences (CDS) from each single-copy family were aligned by MUSCLE 3.7 (http://www.ebi.ac.uk/Tools/msa/muscle/) with the guidance of aligned protein sequences and concatenated to one super gene for each species. Codon 1, 2, 3, and 1+2 sequences were extracted from CDS alignments and used as input for building trees, along with protein, CDS sequences. Then, RAxML 7.2.8 [65] (http://sco.h-its.org/exelixis/software.html) was applied for these sequence sets to build a phylogenetic tree under GTR+gamma for nucleotide sequences and BLOSUM62+gamma model for protein sequences. We employed 1,000 rapid bootstrap replicated to assess the branch reliability in RAxML 7.2.8. All results indicate that tree shrews and primates cluster together (Supplementary Fig. S2 & main text Fig. 1)

4.3 Divergence time estimation

The sequences of 2,117 single-copy gene families of the 15 mammals we had earlier obtained were concatenated and preprocessed for estimating divergence times. We used PAML 4.4 [59] MCMCtree to determine split times with approximate likelihood calculation [66.67]. PAML baseml[59] was firstly taken to compute alpha parameter under

the REV substitution model and substitution rate per time unit. Then the gradient (g) vector and Hessian (H) matrix were estimated which described the shape of the log-likelihood surface around MLE of branch lengths. When performing estimation, the "Correlated molecular clock" and "REV" substitution model were selected. The gamma prior for the overall substitution rate was described by shape and scale parameters which were set as 1 and 11.9, respectively, and were calculated according to substitution rate per time unit. The alpha parameter for gamma rates at sites was set as that computed by baseml in the first step. The MCMC process of PAML MCMCtree was run to sample 1 million times, with sample frequency set to 50, after a burn-in of 5 million iterations. The "finetune" parameters were set as "0.00356 0.02243 0.00633 0.12 0.43455" to make acceptance proportions fall in interval (20%, 40%). Other parameters were set to defaults. Tracer 1.4 in BEAST (http://beast.bio.ed.ac.uk/) was applied to check convergence. Two independent runs were performed to confirm convergence.

4.4 Gene family expansion

Gene family expansion analysis was performed by CAFE 2.1 [68]. In CAFE, a random birth and death model was proposed to study gene gain and loss in gene families across a user-specified phylogenetic tree. A global parameter λ, which described both the gene birth (λ) and death (μ = -λ) rates across all branches of the tree for all gene families, was estimated using maximum likelihood. Then, a conditional p-value was calculated for each gene family, and families with conditional p-values less than the threshold (0.05) were considered to have an accelerated rate of expansion and contraction. Analysis revealed a total of 162 gene families underwent specific expansion in tree shrew.

4.5 Tree shrew and primate shared genes and tree shrew lost genes

Because tree shrews are closely related to primates (Fig. 1 and Supplementary Fig. S2), identifying which genes emerged in their common ancestors is of particular interest in making the tree shrew a valid model. A previous study identified thousands of primate specific genes using the mouse genome as the outgroup [60]. Paired with our study of the tree shrew genome, we can now obtain a better understanding of which genes were been shared by the tree shrew and other primates. To determine the orthologous relationship between tree shrew and human proteins, nucleotide and human and mouse protein data were downloaded from Ensembl (release 60). The longest transcript was chosen to represent each gene with alternative splicing variants. We then subjected all the proteins to BlastP analysis with the similarity cutoff threshold of e-value=$1e^{-5}$. With the human protein set as a reference, we found the best hit for each tree shrew protein in other species, with the criteria that more than 30% of the aligned sequence showed an identity above 30%. Reciprocal best-match pairs were defined as orthologs. Then gene order information was used to filter the false positive orthologs caused by draft genome assembly and annotation. Orthologs not in gene synteny blocks were removed from further analysis. For example, for the continuous 3 genes in human

genome A, B and C, though all of three orthologs can be identified between humans and tree shrews based on the cutoff threshold described above, the B gene in the tree shrew genome was not between A and C, may be from other scaffold or other place within the same scaffold, and should be removed. Using this method we identified three-ways gene synteny relationships for humans, tree shrews and mice. Previously identified primate-specific genes were mapped onto the synteny map. The orthologous genes in primates and tree shrews absent in mice in synteny map were considered as a first appeared gene in the common ancestor of primates and tree shrews. We also performed the manual check for all candidate genes and found 28 such genes (Supplementary Table S5). From the synteny map, we also observed genes specifically missing in tree shrews which should have been lost in this species. To further confirm this finding, we manually checked and annotated the genes in the tree shrew genome, and filtered those located in gap regions. Finally, we identified 11 gene loss events in tree shrews (Supplementary Table S6).

4.6 Tree shrew pseudogene

To detect homozygous pseudogenes in the tree shrew genome *in silico*, we fist aligned all the human genes (cDNA sequences downloaded from Ensembl (release-56)) onto the tree shrew genomes using BLAT with parameters (–extendThroughN -fine). Best hit regions of each gene with 1Kb flanking sequence were cut down and re-aligned with their corresponding human orthologous protein sequence using GeneWise[54] with parameters (-genesf -tfor -quiet), which could help to define the detail exon-intron structure of each gene. All the genes containing frame shifts or premature stop codons as reported by GeneWise were considered candidate pseudogenes. We further carried out a series of filtering processes: (1) To avoid the reported frame shifts or premature stop codons were due to the flaw in algorithm of GeneWise, we also aligned all human proteins to their corresponding loci in the human genome using GeneWise as a control, and genes with frame shifts or premature stop codons in human-to-human alignment reported by GeneWise were filtered; (2) Using the results of human-to-human alignment from GeneWise, candidate pseudogenes with obvious splicing errors near their frame shifts or premature stop codons were filtered; (3) Candidate pseudogenes with a low number of reads covering their frame shift or premature stop codon sites were considered assembly errors. Meanwhile, cases with a considerable number of reads resulting from existing genotype at these sites were treated as heterozygous. Both cases due to either assembly error or heterozygosis were filtered. In total, we identified 144 pseudogenes in the tree shrew genome. By comparison with all alternative splice forms in humans, we found that 89 pseudogenes with the pseudo-mutations can be compensated by alternative splicing.

To detect Olfactory Receptor (OR) functional genes and pseudogenes from the tree shrew genome sequences, we conducted a homology search. Human and mouse OR genes were downloaded from the KEGG database (http://www.genome.jp/kegg, K04257). We then conducted a TblastN search [9] with the E value of 10–20 against the whole human genome sequences by using each of the intact human and mouse OR

genes as a query. We regarded all of the matches detected by the homology search as tree shrew OR functional genes or pseudogenes. GeneWise was then used to detect the right gene structure of tree shrew OR genes, with frame shift or premature stop or lacking complete ORF were detected as pseudogenes. We then scanned the potential OR genes to the InterPro database (http://www.ebi.ac.uk/interpro/) to find if these proteins contain OR domain (IPR000725, blast p-value < 1E-20). We found 690 functional OR genes in tree shrew genome.

To investigate tree shrew visual system, a total of 209 human visual related genes were obtained from the Gene Ontology database (GO0007601: visual perception and other visual related function) for comparative study. We firstly aligned all 209 human cDNA to the tree shrew genome using BLAT with the parameter of (–extend ThroughN -fine), the best hit regions of each gene with 1Kb flanking sequence were cut down and re-aligned with their corresponding human protein sequence using GeneWise[54] (-genesf -tfor -quiet) which could help define the detail exon-intron structure of each gene . Genes located in the right position in synteny blocks were identified as the right copies, otherwise they were filtered as false positives. Finally, nearly all the genes in the tree shrew genome can be detected without any frame shift or premature stop, except for the *OPN1MW* and *OPN1MW2* genes.

5 Drug targeted domain in tree shrew

Hepatitis drug targeted genes were obtained from Drugbank (http://www.drugbank.ca/) using the keyword “hepatitis”. We manually annotated these genes in the tree shrew genome.

Supplementary References

61. Zhao, H., Yang, J.R., Xu, H. & Zhang, J. Pseudogenization of the umami taste receptor gene Tas1r1 in the giant panda coincided with its dietary switch to bamboo. *Mol Biol Evol* **27**, 2669-2673 (2010).
62. Sung, C.H., Schneider, B.G., Agarwal, N., Papermaster, D.S. & Nathans, J. Functional heterogeneity of mutant rhodopsins responsible for autosomal dominant retinitis pigmentosa. *Proc Natl Acad Sci U S A* **88**, 8840-8844 (1991).
63. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
64. Li, H. et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34**, D572-580 (2006).
65. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456-463 (2005).
66. de Magalhães, J.P. et al. Proposal to Sequence an Organism of Unique Interest for Research on Aging: Heterocephalus glaber, the Naked Mole-Rat.
67. Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* **23**, 212-226 (2006).
68. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* **15**, 1153-1160 (2005).
69. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

Supplementary Data 1. Chinese tree shrew pseudogenes.

#human symbol	gene name	KEGG Class	F/S	AS
FASTKD2	FAST kinase domains 2	NA	F	No
BCL2L12	BCL2-like 12 (proline rich)	NA	F	No
EMR3	egf-like module containing, mucin-like, hormone receptor-like 3	Signaling Molecules and Interaction	F	No
KIF12	kinesin family member 12	Cell Motility	F	No
PRR5-ARHGAP8	PRR5-ARHGAP8 readthrough	NA	F	No
PCSK4	proprotein convertase subtilisin/kexin type 4	Folding, Sorting and Degradation;Enzyme Families	F	No
KRT28	keratin 28	Cell Motility	F	No
OR9K2	olfactory receptor, family 9, subfamily K, member 2	Sensory System;Signaling Molecules and Interaction	F	No
OR5B3	olfactory receptor, family 5, subfamily B, member 3	Sensory System;Signaling Molecules and Interaction	F	No
C18orf26	chromosome 18 open reading frame 26	NA	S	No
PVRIG	poliovirus receptor related immunoglobulin domain containing	NA	F	No
SPTBN5	spectrin, beta, non-erythrocytic 5	Cell Motility	F	No
MC5R	melanocortin 5 receptor	Signaling Molecules and Interaction	F	No
OR2D3	olfactory receptor, family 2, subfamily D, member 3	Sensory System;Signaling Molecules and Interaction	S	No
AC022762.1	NA	NA	F	No
AC111177.1	NA	NA	F	No
OR56B4	olfactory receptor, family 56, subfamily B, member 4	Sensory System;Signaling Molecules and Interaction	S	No
OR52N5	olfactory receptor, family 52, subfamily N, member 5	Sensory System;Signaling Molecules and Interaction	F	No
OR51J1	olfactory receptor, family 51, subfamily J, member 1	NA	F&S	No
ARMCX3	armadillo repeat containing, X-linked 3	NA	F&S	No
MPO	myeloperoxidase	Transport and Catabolism	F	No
PXDNL	peroxidasin homolog (Drosophila)-like	NA	S	No
OR11G2	olfactory receptor, family 11, subfamily G, member 2	Sensory System;Signaling Molecules and Interaction	F	No
OR5B17	olfactory receptor, family 5, subfamily B, member 17	Sensory System;Signaling Molecules and Interaction	F	No
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	Immune System;Transport and Catabolism;Immune System Diseases;Signaling Molecules and Interaction;Infectious Diseases;Cardiovascular Diseases;Metabolic Diseases	S	No
CCDC102B	coiled-coil domain containing 102B	NA	F&S	No
KMO	kynurenine 3-monooxygenase (kynurenine 3-hydroxylase)	Amino Acid Metabolism	F	No
ASPM	asp (abnormal spindle) homolog, microcephaly associated	NA	S	No
OR10R2	olfactory receptor, family 10, subfamily R, member 2	Sensory System;Signaling Molecules and Interaction	S	No
DMBT1	deleted in malignant brain tumors 1	Digestive System	F&S	No
ZSWIM1	zinc finger, SWIM-type containing 1	NA	S	No

HNRNPH2	heterogeneous nuclear ribonucleoprotein H2 (H')	Transcription	F&S	No
TRMT2B	TRM2 tRNA methyltransferase 2 homolog B (S. cerevisiae)	NA	F&S	No
NOX1	NADPH oxidase 1	Immune System;Development;Transport and Catabolism	F	No
TNFSF13B	tumor necrosis factor (ligand) superfamily, member 13b	Immune System;Immune System Diseases;Signaling Molecules and	F	No
SLC38A5	solute carrier family 38, member 5	NA	F&S	No
PGPEP1L	pyroglutamyl-peptidase I-like	NA	F	No
IGLV4-69	immunoglobulin lambda variable 4-69	NA	S	No
AC117834.1	NA	NA	F	No
SH2D2A	SH2 domain containing 2A	Signal Transduction	S	No
TTC23	tetratricopeptide repeat domain 23	NA	F	No
GPR112	G protein-coupled receptor 112	Signaling Molecules and Interaction	F	No
PPM1N	protein phosphatase, Mg2+/Mn2+ dependent, 1N (putative)	NA	F	No
MCC	mutated in colorectal cancers	NA	S	No
AC011537.1	NA	NA	F	No
OR6V1	olfactory receptor, family 6, subfamily V, member 1	Sensory System;Signaling Molecules and Interaction	F	No
CTD-3074O7.11	NA	NA	F	No
ARMCX4	armadillo repeat containing, X-linked 4	NA	F	No
PGAM4	phosphoglycerate mutase family member 4	Energy Metabolism;Carbohydrate Metabolism	F	No
RYBP	RING1 and YY1 binding protein	Replication and Repair	F	No
NACA2	nascent polypeptide-associated complex alpha subunit 2	NA	S	No
ZNF260	zinc finger protein 260	NA	F	No
RP11-468E2.1	NA	NA	S	No
RP11-480I12.4	NA	NA	F	No
TRAF6	TNF receptor-associated factor 6	Nervous System;Immune System;Development;Transport and Catabolism;Cancers;Folding, Sorting and Degradation;Signal Transduction;Infectious Diseases	F&S	Yes
PHF16	PHD finger protein 16	NA	F	Yes
ACCN2	amiloride-sensitive cation channel 2, neuronal	Signaling Molecules and Interaction	F	Yes
ATP7B	ATPase, Cu++ transporting, beta polypeptide	NA	F	Yes
PSD4	pleckstrin and Sec7 domain containing 4	Transport and Catabolism	F	Yes
RTN2	reticulon 2	NA	F	Yes
NECAB3	N-terminal EF-hand calcium binding protein 3	NA	F	Yes
HCFC1R1	host cell factor C1 regulator 1 (XPO1 dependent)	NA	S	Yes
ITIH5	inter-alpha-trypsin inhibitor heavy chain family, member 5	NA	F	Yes
AMHR2	anti-Mullerian hormone receptor, type II	Signaling Molecules and Interaction;Enzyme Families;Signal Transduction	F	Yes
PSMC1	proteasome (prosome, macropain) 26S subunit, ATPase, 1	Folding, Sorting and Degradation	F&S	Yes

TTC19	tetratricopeptide repeat domain 19	NA	F	Yes
ACCS	1-aminocyclopropane-1-carboxylate synthase homolog (Arabidopsis)(non-functional)	NA	F	Yes
C1RL	complement component 1, r subcomponent-like	NA	F	Yes
STON2	stonin 2	NA	F	Yes
ALS2CR12	amyotrophic lateral sclerosis 2 (juvenile) chromosome region,	NA	S	Yes
FRRS1	ferric-chelate reductase 1	NA	S	Yes
PEX10	peroxisomal biogenesis factor 10	Transport and Catabolism	F	Yes
KRT4	keratin 4	Cell Motility	F	Yes
RAVER2	ribonucleoprotein, PTB-binding 2	NA	F	Yes
DNALI1	dynein, axonemal, light intermediate chain 1	Cell Motility;Neurodegenerative Diseases	F	Yes
ANKK1	ankyrin repeat and kinase domain containing 1	NA	F	Yes
PTPN7	protein tyrosine phosphatase, non-receptor type 7	Signal Transduction	F	Yes
SLC16A11	solute carrier family 16, member 11 (monocarboxylic acid	Membrane Transport	F	Yes
NIPAL4	NIPA-like domain containing 4	NA	F&S	Yes
ZNF541	zinc finger protein 541	NA	S	Yes
KCNH6	potassium voltage-gated channel, subfamily H (eag-related),	Signaling Molecules and Interaction	F	Yes
DNAJC4	DnaJ (Hsp40) homolog, subfamily C, member 4	Folding, Sorting and Degradation	F	Yes
OR4C3	olfactory receptor, family 4, subfamily C, member 3	Sensory System;Signaling Molecules and Interaction	S	Yes
LCK	lymphocyte-specific protein tyrosine kinase	Immune System;Development;Immune System Diseases;Signaling Molecules and Interaction;Enzyme Families	F	Yes
DNAH11	dynein, axonemal, heavy chain 11	NA	F	Yes
CCDC28A	coiled-coil domain containing 28A	NA	F	Yes
EXOC7	exocyst complex component 7	Endocrine System	F	Yes
ZNF677	zinc finger protein 677	Transcription	F	Yes
C20orf24	chromosome 20 open reading frame 24	NA	S	Yes
ISM2	isthmin 2 homolog (zebrafish)	NA	S	Yes
DCDC5	doublecortin domain containing 5	NA	F	Yes
SPINT1	serine peptidase inhibitor, Kunitz type 1	NA	F	Yes
DCBLD1	discoidin, CUB and LCCL domain containing 1	NA	F	Yes
HELT	helt bHLH transcription factor	NA	F	Yes
KIAA0586	KIAA0586	NA	F&S	Yes
HTR3A	5-hydroxytryptamine (serotonin) receptor 3A	Signaling Molecules and Interaction	F	Yes
TP73	tumor protein p73	Nervous System;Transcription;Infectious Diseases;Cell Growth and Death	S	Yes
KLHDC1	kelch domain containing 1	NA	F	Yes
DSG4	desmoglein 4	NA	F	Yes

ARHGEF18	Rho/Rac guanine nucleotide exchange factor (GEF) 18	NA	F	Yes
MORF4L2	mortality factor 4 like 2	Replication and Repair	F	Yes
FAM63A	family with sequence similarity 63, member A	NA	S	Yes
MYCT1	myc target 1	NA	F	Yes
TRMT1L	TRM1 tRNA methyltransferase 1-like	NA	F	Yes
ROS1	c-ros oncogene 1 , receptor tyrosine kinase	Signaling Molecules and Interaction;Enzyme Families;Signal Transduction	S	Yes
ZNF644	zinc finger protein 644	NA	F	Yes
RBMX	RNA binding motif protein, X-linked	Transcription	S	Yes
USP33	ubiquitin specific peptidase 33	Folding, Sorting and Degradation;Enzyme Families	F	Yes
MRPS18A	mitochondrial ribosomal protein S18A	NA	F	Yes
FOXP3	forkhead box P3	Signaling Molecules and Interaction	F&S	Yes
TAF1C	TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa	Transcription	F&S	Yes
CDK11A	cyclin-dependent kinase 11A	Transcription;Enzyme Families	S	Yes
INTS1	integrator complex subunit 1	Transcription	F	Yes
AKT1S1	AKT1 substrate 1 (proline-rich)	NA	F	Yes
RBMS1	RNA binding motif, single stranded interacting protein 1	NA	F	Yes
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	Cancers;Infectious Diseases;Cell Growth and Death	F	Yes
CABIN1	calcineurin binding protein 1	NA	F	Yes
C2orf73	chromosome 2 open reading frame 73	NA	S	Yes
EPHB2	EPH receptor B2	Development;Signaling Molecules and Interaction;Enzyme Families	F	Yes
TNFSF18	tumor necrosis factor (ligand) superfamily, member 18	Signaling Molecules and Interaction	F	Yes
CKMT1A	creatine kinase, mitochondrial 1A	Amino Acid Metabolism	F	Yes
HIF1AN	hypoxia inducible factor 1, alpha subunit inhibitor	NA	F	Yes
VSIG1	V-set and immunoglobulin domain containing 1	NA	F	Yes
BBC3	BCL2 binding component 3	Infectious Diseases;Neurodegenerative Diseases;Cell Growth and Death	S	Yes
ZBTB7B	zinc finger and BTB domain containing 7B	Transcription;Folding, Sorting and Degradation	F	Yes
C14orf37	chromosome 14 open reading frame 37	NA	F	Yes
OTUB1	OTU domain, ubiquitin aldehyde binding 1	Folding, Sorting and Degradation;Enzyme Families	S	Yes
CASR	calcium-sensing receptor	Signaling Molecules and Interaction	F	Yes
DROSHA	drosha, ribonuclease type III	Translation	F	Yes
SLC4A9	solute carrier family 4, sodium bicarbonate cotransporter, member 9	NA	F	Yes
NCALD	neurocalcin delta	NA	F	Yes
WNK1	WNK lysine deficient protein kinase 1	Enzyme Families;	F	Yes
ZNF623	zinc finger protein 623	NA	F	Yes
C9orf174	chromosome 9 open reading frame 174	NA	F	Yes

DNHD1	dynein heavy chain domain 1	NA	F&S	Yes
CUL7	cullin 7	Folding, Sorting and Degradation	F&S	Yes
ZNF135	zinc finger protein 135	Transcription;	F	Yes
ARHGAP5	Rho GTPase activating protein 5	Immune System;Cell Communication	F	Yes
MYT1	myelin transcription factor 1	NA	F	Yes
TUBA1C	tubulin, alpha 1c	Transport and Catabolism;Cell Motility;Infectious Diseases;Cell	F	Yes
DAZAP2	DAZ associated protein 2	NA	F	Yes
DDIT3	DNA-damage-inducible transcript 3	Transcription;Folding, Sorting and Degradation;Signal Transduction	F	Yes
C12orf28	chromosome 12 open reading frame 28	NA	F	Yes
CLEC5A	C-type lectin domain family 5, member A	Signaling Molecules and Interaction	F	Yes

Supplementary Data 2. HBV and HCC related genes in Chinese tree shrew genome

human symbol	gene name	tree shrew copy number	tree shrew identity	note
TRAF6	TNF receptor-associated factor 6	1	pseudogene	can be compensated by alternative
DDX58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	0	loss	lost in synteny block, and can't find other
IL8	interleukin 8	1	76.34	
OAS1	2'-5'-oligoadenylate synthetase 1, 40/46kDa	1	73.53	
AKT3	v-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)	1	99.77	
ANXA5	annexin A5	1	96.85	
APOH	apolipoprotein H (beta-2-glycoprotein I)	1	84.02	
ARAF	v-raf murine sarcoma 3611 viral oncogene homolog	1	95.17	
ASGR1	asialoglycoprotein receptor 1	1	85.66	
BAD	BCL2-associated agonist of cell death	1	85.31	
CCL3	chemokine (C-C motif) ligand 3	1	74.44	
CCR5	chemokine (C-C motif) receptor 5	1	78.57	
CD81	CD81 molecule	1	96.71	
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	1	78.81	
CHUK	conserved helix-loop-helix ubiquitous kinase	1	96	
CLDN16	claudin 16	1	87.29	
EGF	epidermal growth factor	1	57.98	
EGFR	epidermal growth factor receptor	1	93.13	
EIF2AK1	eukaryotic translation initiation factor 2-alpha kinase 1	1	87.68	
EIF2S1	eukaryotic translation initiation factor 2, subunit 1 alpha, 35kDa	1	90.55	
EIF3E	eukaryotic translation initiation factor 3, subunit E	1	100	
FN1	fibronectin 1	1	89.18	
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	1	95.4	
GRB2	growth factor receptor-bound protein 2	1	99.43	
GSK3B	glycogen synthase kinase 3 beta	1	99.27	
HLA-A	major histocompatibility complex, class I, A	1	80.54	
HLA-B	major histocompatibility complex, class I, B	1	80.61	
HLA-C	major histocompatibility complex, class I, C	1	80	
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	1	75.79	
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	1	70.01	
HLA-DPB1	major histocompatibility complex, class II, DP beta 1	1	64.22	
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	1	75.79	
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	1	75.67	
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog	1	99.46	

IFIT1	interferon-induced protein with tetratricopeptide repeats 1	1	67.65
IFNA1	interferon, alpha 1	1	65.14
IFNAR1	interferon (alpha, beta and omega) receptor 1	1	69.96
IFNAR2	interferon (alpha, beta and omega) receptor 2	1	64.19
IFNB1	interferon, beta 1, fibroblast	1	65.93
IFNG	interferon, gamma	1	65.66
IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells,	1	94.51
IKBKE	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase epsilon	1	89.89
IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	1	88.46
IL10	interleukin 10	1	83.71
IL18	interleukin 18 (interferon-gamma-inducing factor)	1	69.23
IL2	interleukin 2	1	82.89
IL28B	interleukin 28B (interferon, lambda 3)	1	75.63
IL4	interleukin 4	1	37.14
IL6	interleukin 6 (interferon, beta 2)	1	57.81
IRF1	interferon regulatory factor 1	1	93.12
IRF3	interferon regulatory factor 3	1	81.29
IRF9	interferon regulatory factor 9	1	82.95
JAK1	Janus kinase 1	1	96.94
LDLR	low density lipoprotein receptor	1	83.68
MAPK1	mitogen-activated protein kinase 1	1	99.64
MAPK14	mitogen-activated protein kinase 14	1	100
MAPK8	mitogen-activated protein kinase 8	1	99.05
MAVS	mitochondrial antiviral signaling protein	1	70.8
MBP	myelin basic protein	1	80.78
NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells	1	86.61
NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	1	95.9
NR1H3	nuclear receptor subfamily 1, group H, member 3	1	92.15
OCLN	occludin	1	88.25
PDK1	pyruvate dehydrogenase kinase, isozyme 1	1	96.77
PIAS3	protein inhibitor of activated STAT, 3	1	98.69
PIK3R5	phosphoinositide-3-kinase, regulatory subunit 5	1	92.32
PPARA	peroxisome proliferator-activated receptor alpha	1	93.55
PPP2CA	protein phosphatase 2, catalytic subunit, alpha isozyme	1	99.64

PSME3	proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)	1	99.6
RIPK1	receptor (TNFRSF)-interacting serine-threonine kinase 1	1	74.25
RNASEL	ribonuclease L (2',5'-oligoisoadenylate synthetase-dependent)	1	73.85
RXRA	retinoid X receptor, alpha	1	95.08
SCARB1	scavenger receptor class B, member 1	1	88.1
SERPINB3	serpin peptidase inhibitor, clade B (ovalbumin), member 3	1	68.56
SOCS3	suppressor of cytokine signaling 3	1	94.64
SOS1	son of sevenless homolog 1 (Drosophila)	1	99.23
STAT1	signal transducer and activator of transcription 1, 91kDa	1	96.74
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	1	100
TBK1	TANK-binding kinase 1	1	98.04
TICAM1	toll-like receptor adaptor molecule 1	1	67.17
TLR2	toll-like receptor 2	1	71.19
TLR3	toll-like receptor 3	1	82.35
TNF	tumor necrosis factor	1	78.45
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A	1	72.25
TP53	tumor protein p53	1	85.08
TRADD	TNFRSF1A-associated via death domain	1	83.77
TRAF2	TNF receptor-associated factor 2	1	80.92
TRAF3	TNF receptor-associated factor 3	1	97
VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor	1	93.64
AKT1	v-akt murine thymoma viral oncogene homolog 1	1	88.79
AKT2	v-akt murine thymoma viral oncogene homolog 2	1	99.01
BRAF	v-raf murine sarcoma viral oncogene homolog B1	1	98.74
CLDN10	claudin 10	1	94.16
CLDN11	claudin 11	1	97.69
CLDN1	claudin 1	1	93.27
CLDN14	claudin 14	1	93.31
CLDN15	claudin 15	1	82.96
CLDN17	claudin 17	1	91.52
CLDN18	claudin 18	1	91.6
CLDN19	claudin 19	1	97.73
CLDN20	claudin 20	1	79.91
CLDN2	claudin 2	1	96.96
CLDN22	claudin 22	1	77.83
CLDN23	claudin 23	1	73.63

CLDN3	claudin 3	1	90	
CLDN4	claudin 4	1	89	
CLDN5	claudin 5	1	93.1	
CLDN6	claudin 6	1	88.64	
CLDN7	claudin 7	1	94.71	
CLDN8	claudin 8	1	84.65	
CLDN9	claudin 9	1	96.31	
IFNA10	interferon, alpha 10	NA	NA	treeshrew mis-assembled, human tandem
IFNA13	interferon, alpha 13	1	50.27	treeshrew mis-assembled, human tandem
IFNA14	interferon, alpha 14	NA	NA	treeshrew mis-assembled, human tandem
IFNA16	interferon, alpha 16	NA	NA	treeshrew mis-assembled, human tandem
IFNA17	interferon, alpha 17	1	53.23	treeshrew mis-assembled, human tandem
IFNA21	interferon, alpha 21	NA	NA	treeshrew mis-assembled, human tandem
IFNA2	interferon, alpha 2	NA	NA	treeshrew mis-assembled, human tandem
IFNA4	interferon, alpha 4	NA	NA	treeshrew mis-assembled, human tandem
IFNA5	interferon, alpha 5	NA	NA	treeshrew mis-assembled, human tandem
IFNA6	interferon, alpha 6	NA	NA	treeshrew mis-assembled, human tandem
IFNA7	interferon, alpha 7	1	53.23	
IFNA8	interferon, alpha 8	1	50	
IRF7	interferon regulatory factor 7	1	73.65	
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	1	100	
MAPK10	mitogen-activated protein kinase 10	1	100	
MAPK11	mitogen-activated protein kinase 11	1	96.56	
MAPK12	mitogen-activated protein kinase 12	1	95.13	
MAPK13	mitogen-activated protein kinase 13	1	94.64	
MAPK3	mitogen-activated protein kinase 3	1	97.86	
MAPK9	mitogen-activated protein kinase 9	1	97.12	
MIR122	microRNA 122	1	97.65	
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	1	100	
PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	1	99.15	
PIK3CB	phosphoinositide-3-kinase, catalytic, beta polypeptide	1	96.68	
PIK3CD	phosphoinositide-3-kinase, catalytic, delta polypeptide	1	93.66	
PIK3CG	phosphoinositide-3-kinase, catalytic, gamma polypeptide	1	95.72	
PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)	1	97.05	
PIK3R2	phosphoinositide-3-kinase, regulatory subunit 2 (beta)	1	95.5	
PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)	1	91.38	
PPP2CB	protein phosphatase 2, catalytic subunit, beta isozyme	1	98.91	
PPP2R1A	protein phosphatase 2, regulatory subunit A, alpha	1	99.63	

PPP2R1B	protein phosphatase 2, regulatory subunit A, beta	1	96.45
RAF1	v-raf-1 murine leukemia viral oncogene homolog 1	1	98.03
RELA	v-rel reticuloendotheliosis viral oncogene homolog A (avian)	1	91.69
SOS2	son of sevenless homolog 2 (Drosophila)	1	98.2
STAT2	signal transducer and activator of transcription 2, 113kDa	1	87.39
OAS3	2'-5'-oligoadenylate synthetase 3, 100kDa	1	43.16
EIF2AK2	eukaryotic translation initiation factor 2-alpha kinase 2	1	65.78
PIAS1	protein inhibitor of activated STAT, 1	1	98.11
PIAS4	protein inhibitor of activated STAT, 4	1	81.91
EIF2AK4	eukaryotic translation initiation factor 2 alpha kinase 4	1	94.68
EIF2AK3	eukaryotic translation initiation factor 2-alpha kinase 3	1	93.74
OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	1	67.26
IFIT1B	interferon-induced protein with tetratricopeptide repeats 1B	1	72.73
PPP2R2A	protein phosphatase 2, regulatory subunit B, alpha	1	100
PPP2R2B	protein phosphatase 2, regulatory subunit B, beta	1	100
PIAS2	protein inhibitor of activated STAT, 2	1	99.34
PPP2R2D	protein phosphatase 2, regulatory subunit B, delta	1	98.56
PPP2R2C	protein phosphatase 2, regulatory subunit B, gamma	1	99.74
TYK2	tyrosine kinase 2	1	84.69
SLC10A1	solute carrier family 10 (sodium/bile acid cotransporter family), member 1	1	86.21
APOBEC3G	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-	1	44.29