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.

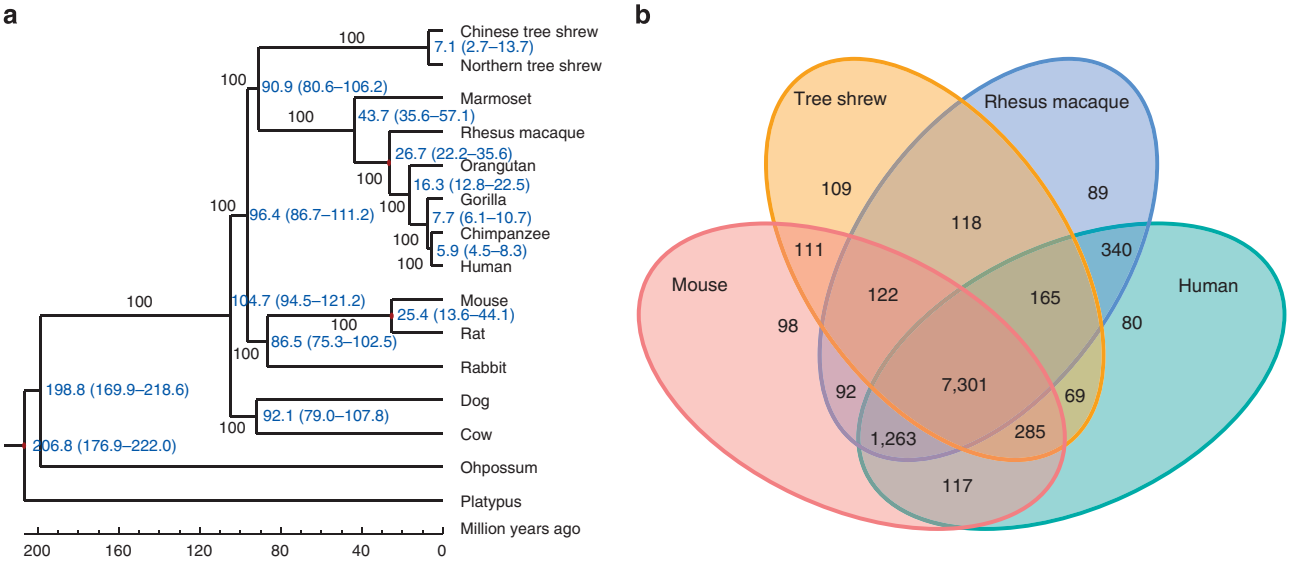

Figure 1 | Relationship of the Chinese tree shrew and related mammals. (**a**) Consensus phylogenetic tree of 15 (sub)species based on 2,117 single-copy genes. The topology was supported by all phylogenetic resources including full-coding sequences, first, second, third codon positions, and amino acids 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

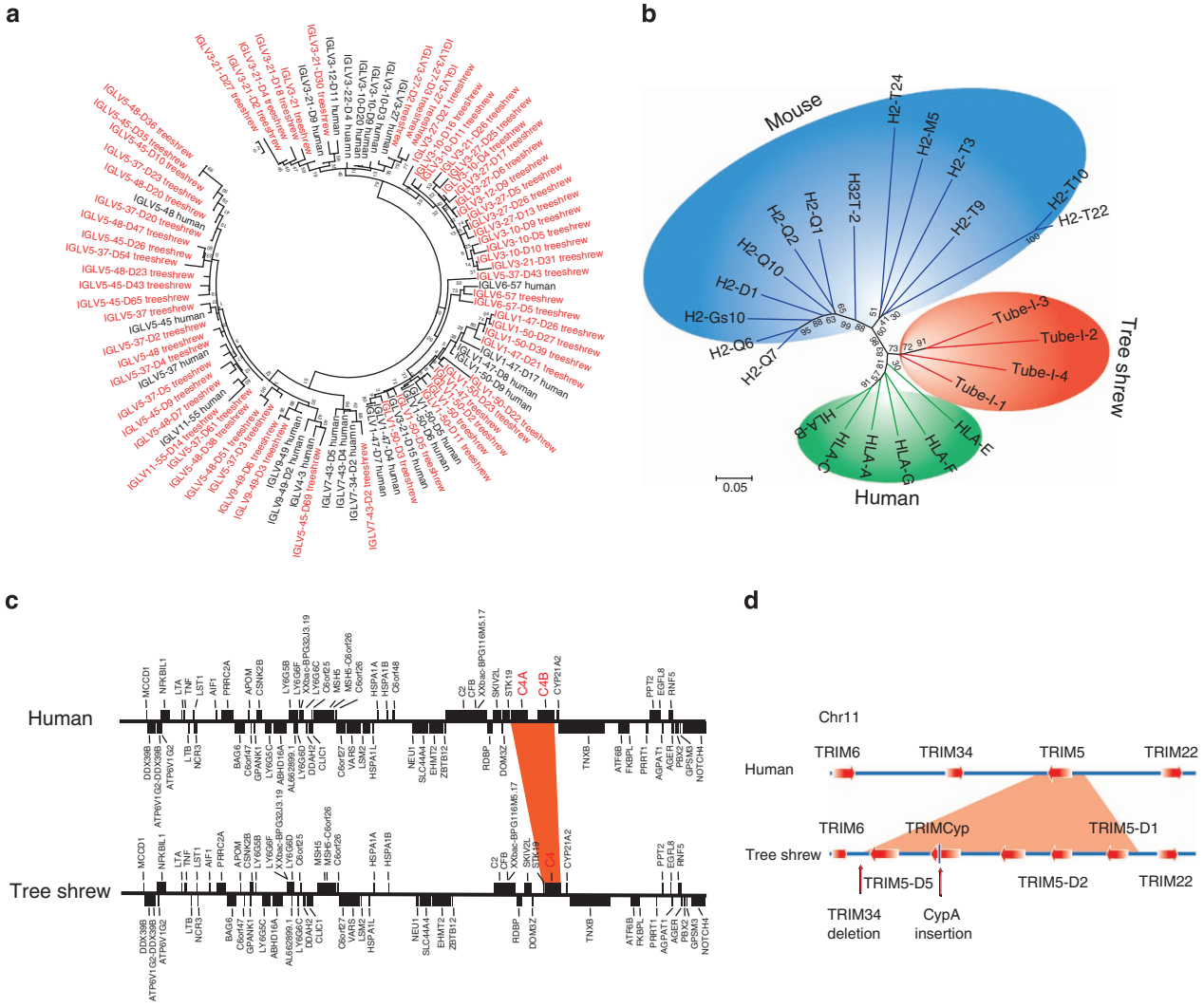

Figure 2 | Immune system in Chinese tree shrew and compared with human and mouse. (**a**) Specific expansion of the immunoglobulin lambda variable (*IGLV*) gene family in the tree shrew. Gene IDs in red are tree shrew genes. (**b**) Phylogenetic relationship of MHC-class I genes in human, tree shrew 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

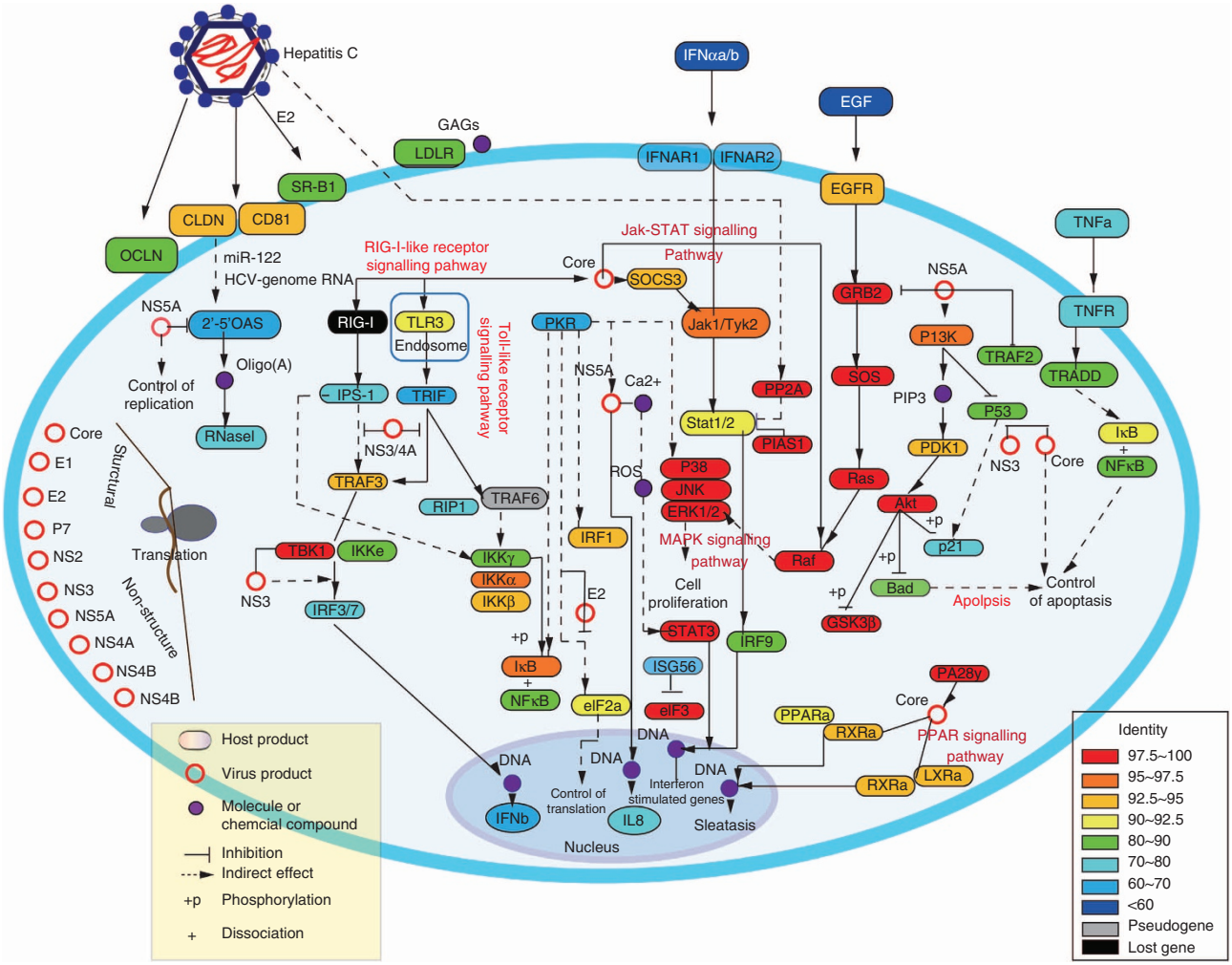

Figure 3 | Genetic divergence of genes involved in HCV infection pathway between human and Chinese tree shrew. Colours represent the degree of sequence identity at the amino-acid level.

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).