ARTICLE IN PRESS

Science Bulletin xxx (xxxx) xxx

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Ruminant-specific genes identified using high-quality genome data and their roles in rumen evolution

Chunyan Chen[a,1], Yuan Yin[a,1], Haorong Li[a,1], Botong Zhou[a], Jiong Zhou[a], Xiaofang Zhou[a], Zhipeng Li[b], Guichun Liu[a,c], Xiangyu Pan[d,e], Ru Zhang[a], Zeshan Lin[a], Lei Chen[a,c,*], Qiang Qiu[a,*], Yong E. Zhang[f,g,h,*], Wen Wang[a,c,h,*]

[a] *School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710072, China*
[b] *College of Animal Science and Technology, Jilin Agricultural University, Changchun 130118, China*
[c] *State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China*
[d] *Department of Gastroenterology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China*
[e] *Guangdong Cardiovascular Institute, Guangzhou 510080, China*
[f] *Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*
[g] *University of Chinese Academy of Sciences, Beijing 100049, China*
[h] *Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China*

ARTICLE INFO

Article history:
Received 24 August 2021
Received in revised form 26 October 2021
Accepted 13 December 2021
Available online xxxx

Keywords:
Ruminant
High-quality ruminant genome
New genes
Rumen
New organ evolution

ABSTRACT

Ruminants comprise a highly successful group of mammals with striking morphological innovations, including the presence of a rumen. Many studies have shown that species-specific or lineage-specific genes (referred to as new genes) play important roles in phenotypic evolution. In this study, we identified 1064 ruminant-specific genes based on the newly assembled high-quality genomes of representative members of two ruminant families and other publically available high-quality genomes. Ruminant-specific genes shared similar evolutionary and expression patterns with new genes found in other mammals, such as primates and rodents. Most new genes were derived from gene duplication and tended to be expressed in the testes or immune-related tissues, but were depleted in the adult brain. We also found that most genes expressed in the rumen were genes predating sheep–sperm whale split (referred to as old genes), but some new genes were also involved in the evolution of the rumen, and contributed more during rumen development than in the adult rumen. Notably, expression levels of members of the ruminant-specific *PRD-SPRRII* gene family, which are subject to positive selection, varied throughout rumen development and may thus play important roles in the development of the keratin-rich surface of the rumen. Overall, this study generated two novel ruminant genomes and also provided novel insights into the evolution of new mammalian organs.

1. Introduction

Newly evolved genes (or new genes) refer to genomic loci that have originated relatively recently in the process of evolution [1–3]. Because new genes only exist in certain species or evolutionary lineages, they are also referred to as species- or lineage-specific genes [1–3], and may play important roles in the evolution of species [1–6]. New genes in mammals have mainly been studied in primates and rodents, such as humans and mice. The major characteristics of new genes in primates and rodents include: most new genes were derived from gene duplication [7,8]; because of relaxation of functional constraints or positive selection, most of them were not subject to significant natural selection [3,7]; new genes had lower expression abundance and a narrower expression distribution compared to ancient genes, possibly due to their suboptimal regulatory context [9–11], and they tend to be testis- or male-biased under the action of multiple forces (e.g., faster-X hypothesis), especially in the case of new X-linked genes [7,12]. The human genome encodes hundreds of primate-specific genes, some of which are functionally related to the evolution of important human traits and organs, including the human brain [10,13]. However, despite the importance of new genes, studies in other mammals are currently sparse.

* Corresponding authors.
E-mail addresses: chen_lei@nwpu.edu.cn (L. Chen), qiuqiang@nwpu.edu.cn (Q. Qiu), zhangyong@ioz.ac.cn (Y.E. Zhang), wenwang@nwpu.edu.cn (W. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2022.01.023
2095-9273/© 2022 Published by Elsevier B.V. on behalf of Science China Press.

Ruminants are terrestrial herbivores, including important livestock species such as cows, buffalo, yak, sheep, goats, and reindeer. They are one of the most successful mammalian groups in the world, and are both numerous and widely distributed throughout diverse habitats. Ruminants have unique biological characteristics, including a multi-chambered stomach that can return semi-digested food to the mouth for further chewing and digestion. The multi-chambered stomach includes an abomasum, rumen, omasum, and reticulum, of which the rumen is the most important evolutionary innovation in ruminants. Many studies have suggested that the rumen, omasum, and reticulum may have originated from the esophagus [14,15].

We recently explored the genetic mechanism underlying specific traits in ruminants using comparative genomics, and identified several new genes in ruminants [14,15]. However, these previous studies relied largely on genomes assembled using second-generation sequencing data, which may include many gaps. High-quality genomic data with fewer gaps are therefore needed to allow the reliable and comprehensive identification of new genes [9]. Ruminants include six families: Tragulidae, Antilocapridae, Giraffidae, Moschidae, Cervidae, and Bovidae, of which high-quality third-generation sequencing genomes have so far been released for Bovidae, Cervidae, and Giraffidae. We previously showed that Antilocapridae and Giraffidae belong to a sister group, representing the oldest branch of extant pecoran families [14]. The Moschidae is the closest sister group to Bovidae, while the Tragulidae is the most basal ruminants [14]. High-quality genomes for Tragulidae and Moschidae are thus important for identifying new genes in ruminants. In addition, it is unknown if new genes in ruminants follow a similar pattern to new genes in primates and rodents, and whether these new genes contribute to the evolution of new mammalian organs, such as the rumen. In this study, we therefore analyzed the newly sequenced genomes for lesser mouse-deer and forest musk deer, representing the Tragulidae and Moschidae lineages, respectively, to identify ruminant-specific new genes. We performed transcriptional and evolutionary analyses of these genes to determine if they had similar patterns to those identified in primates and rodents [3,7–12]. We also compared the expression of the new genes in the developing compared with the adult rumen. Finally, we investigated the functional role of the *PRD-SPRRII* gene family in the function of the rumen.

2. Materials and methods

2.1. Genome sequencing, assembly, and annotation

Fibroblast cells from a lesser mouse-deer (*Tragulus kanchil*) and blood from a forest musk deer (*Moschus berezovskii*) were collected from the Kunming Cell Bank and Sichuan Institute of Musk Deer Breeding, respectively, in accordance with the ethical guidelines for Animal Care and Use of Northwestern Polytechnical University. High-quality DNA was extracted from 5×10^7 lesser mouse-deer cells and 20 mL of forest musk deer blood using a DNeasy Blood & Tissue Kit (Qiagen, Valencia, USA) according to the manufacturer's instructions. Some of this DNA was then used to construct Nanopore libraries, which were sequenced using GridION X5 sequencers (Oxford Nanopore Technologies, Oxford, UK), and the remainder of the DNA was used to construct second-generation sequencing (NGS) libraries with an insert size of 400 bp, which were sequenced using the Illumina NovaSeq platform (Illumina, San Diego, USA). The same cell samples were used to construct high-throughput chromosome conformation capture (Hi-C) libraries, which were subsequently sequenced using the Illumina NovaSeq platform (Illumina).

The Nanopore reads for lesser mouse-deer and forest musk deer were first assembled into contigs using NextDenovo software (v2.2-beta0, with seed cut-off parameter = 24,000) (https://github.com/Nextomics/NextDenovo, last accessed June 2019). The Nextpolish tool (v1.1.0, with task = best) in the NextDenovo suite was then used to polish the contigs using all the Nanopore and NGS Illumina reads. The Hi-C data were mapped to the polished genomes using Juicer (v1.6) [16] and chromosomes were assembled using 3D-DNA software (v180922 with default parameters) [17]. The position and orientation of the contigs on the chromosomes were reviewed manually, refined, and corrected using Juicebox Assembly Tools [18]. Finally, we annotated repetitive sequences and gene models for the lesser mouse-deer and forest musk deer genomes using the pipeline of Chen et al. [14].

2.2. Identification of new genes and inferred mechanisms of origin

We identified new genes in different branches of ruminants using the newly assembled lesser mouse-deer and forest musk deer genomes (Tragulidae and Moschidae, respectively), as well as representative third-generation genome assemblies for three other ruminant families: Giraffidae (giraffe, *Giraffa camelopardalis*) [19], Cervidae (black muntjac, *Muntiacus crinifrons*) [20], and Bovidae (sheep, Oar_rambouillet_v1.0, accession number GCA_002742125.1, Baylor College of Medicine Human Genome Sequencing Center, Sheep Genome Project). We also used high-quality third-generation genome assemblies of six mammal species as an outgroup: sperm whale (ASM283717v2) [21], pig (Sscrofa11.1) [22], horse (EquCab3.0) [23], cat (Felis_catus_9.0) [24], human (GRCh38.p13), and Virginia opossum (*Didelphis virginiana*) (www.dnazoo.org) [17,25]. We then identified ruminant-specific genes according to the synteny-based pipeline (SBP) [7,9]. We used the sheep genome as the focal reference to identify new genes and adopted the stringent criterion that each new gene assigned to a gene-age group did not have a corresponding syntenic locus in any outgroup species. The SBP initially involved whole-genome alignment of collected genomes with the reference genome using Lastz (v1.04.03) [26], with different Lastz parameters set according to the divergence time of the query species from sheep (Table S1 online). For the Lastz alignment output file, we generated the reciprocal alignment results as described by Zhang et al. [7,9] and http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto. The reciprocal alignment output axt format files were used for subsequent gene-age annotation. The ages of all coding genes were inferred from the distribution of homologous sequences on the phylogenetic tree (Fig. 1a). The age of each gene was marked using branch numbers ranging from 0 (for genes shared by all mammals) to 10 (for genes present only in sheep). If multiple exons or transcripts of a gene had different ages, we used the oldest exon or transcript to represent the age of that gene. In addition, if 70% of the gene length aligned with repetitive sequences, orthologous sequences were unevenly distributed in the phylogenetic tree, or if genes were on an unanchored contig, these genes were filtered out and were not included in subsequent analyses. In addition, we used forest musk deer as the reference species to identify new genes.

We then inferred the mechanism of origin for all the identified new genes, as described previously [8,27]. There are three possible types of origin mechanisms of ruminant-specific genes: DNA-mediated duplication, RNA-mediated duplication (retrogenes), and *de novo* birth. We further screened the results to obtain more reliable *de novo* genes by extending the previously published pipeline [6]; i.e., we further ensured the absence of homologous proteins in non-ruminant species in the NCBI NR database and of annotated protein domains in the Pfam database [28].

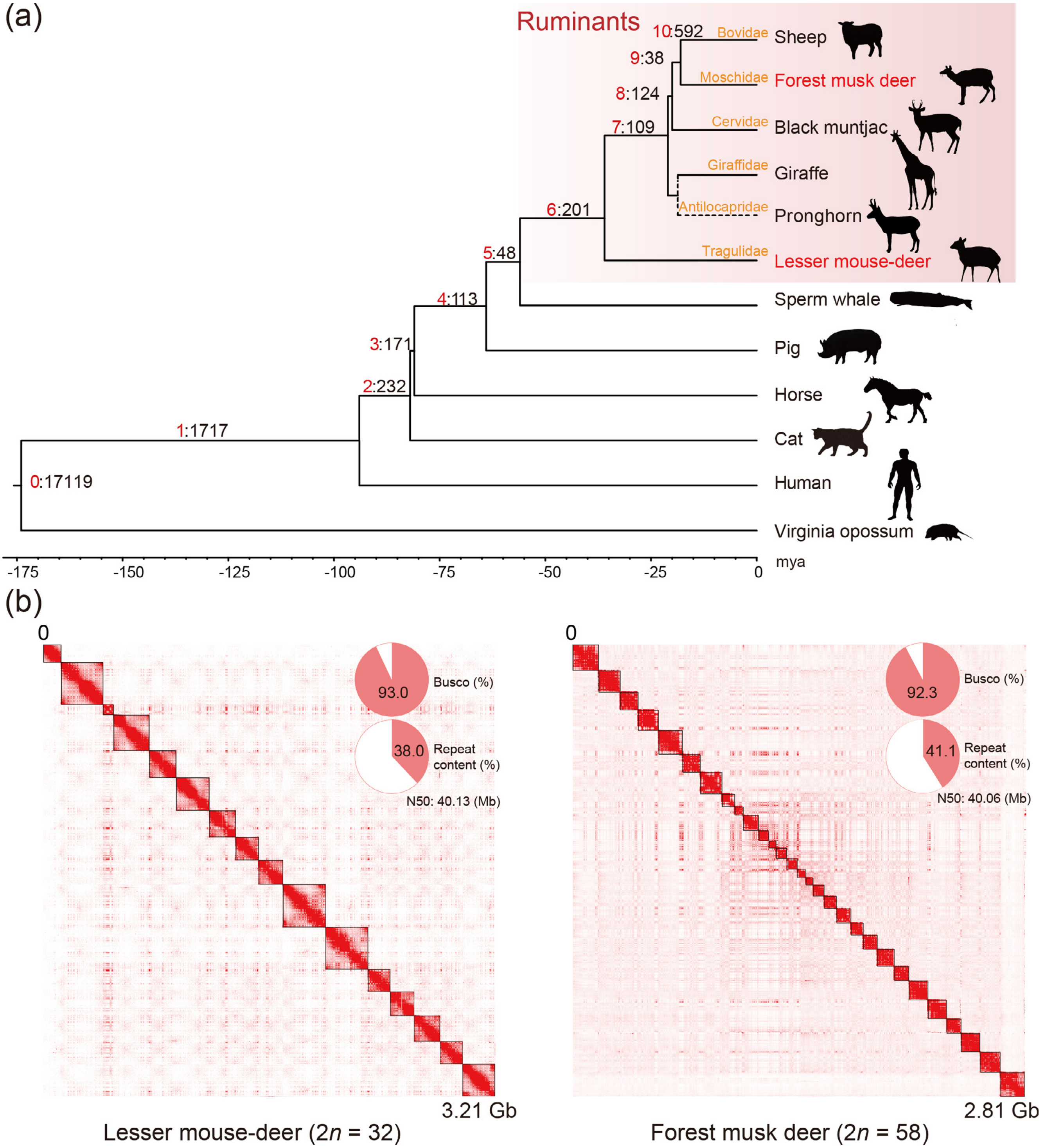


Fig. 1. New genes identified in ruminants. (a) Distribution of new genes in the phylogenetic tree. Ruminants are indicated by red gradient blocks, with sheep as the reference genome. Eleven distinct gene-age groups (branches 0–10) were specified in the SBP, based on the phylogenetic context. Genes postdating the sheep and sperm whale split (branches 6–10) were regarded as ruminant-specific genes. Forest musk deer and lesser mouse-deer are marked in red in the phylogenetic tree. The six ruminant families are marked on the branches of the corresponding representative species. Antilocapridae and Giraffidae belong to a sister group, and given that we did not have high-quality sequences for Antilocapridae, indicated by a dotted line, we selected giraffes as a representative of Giraffidae for the identification of new genes. (b) Hi-C heatmaps for lesser mouse-deer and forest musk deer generated by Juicebox Assembly Tools [18]. Each lesser mouse-deer or forest musk deer chromosome is framed in black box. Pie charts represent the proportion of conserved BUSCO gene sets and repeat content. Contig N50 and assembled genome size are shown.

2.3. Analysis of selection pressures on new genes

Most new genes originated via duplication and are relatively young. We therefore examined the selection pressures on the identified new genes by comparing their sequences with their closest paralogs and calculating the ratio of non-synonymous substitution rate (*Ka*) to synonymous substitution rate (*Ks*) between the paralogs [29]. We first aligned the protein sequences of the new genes and their paralogous copies using BLAST (bl2seq 2.2.26) [30], and then transformed the results into codon-level alignment using the PAL2NAL tool [31]. We then calculated the *Ka*/*Ks* values and performed the likelihood ratio test (LRT) with the assumption of *Ka*/*Ks* = 0.5, using codeml in the PAML4.9 package [32]. We also performed a branch model test using the codeml program with the default parameters model = 0, NSsites = 0, and codon frequency model F3X4. We then filtered out alignments with *Ks* values outside the 1.5-fold interquartile range, given that candidate paralogs with outlier *Ks* values may not represent *bona fide* paralogous genes.

We carried out three analyses to test if the *de novo* genes were constrained by natural selection. We first calculated the *Ka*/*Ks* and *P* values between the sheep *de novo* genes and their homologous genes in the goat genome, and if these values were <1 and $P < 0.05$, the genes were considered to be under selective pressure. Second, we aligned the protein sequences of homologs of sheep *de novo* genes in other ruminant species with the sheep *de novo* genes. After removing low-quality alignment results, we calculated the selective pressures on different evolutionary branches and also conducted LRT analysis.

2.4. Estimation of rate of gene origin

Most new genes are generated by duplication. We therefore estimated the rate of gene gain by accounting for the gene-loss rate, following conventional practice [33]. We initially performed an all-to-all self-alignment of all proteins in sheep using BLAST [30], set the e-value to $<10^{-10}$, and retained the best gene pairs. We then used the codon-level alignments of each gene pair to estimate the numbers of nucleotide substitutions per silent site (*dS*) using the maximum-likelihood procedure in the PAML package. We excluded gene pairs with a $dS > 0.25$, because these duplications could be too old. We then divided *dS* into 0.01 intervals and counted the number of gene pairs in each interval. The natural logarithms of gene pair number and *dS* were fitted linearly using the "lm" method in the geom_smooth function of ggplot2 package. According to the fitting result, we identified 116.7 duplicate genes with a $dS < 0.01$ in sheep. We then estimated the rate of duplicate gene origin assuming the molecular clock of sheep (2.1×10^{-9} per site per year) [14].

2.5. Expression analysis of new genes

We collected a total of 64 transcriptome datasets for sheep [34], derived from 13 adult tissues including the rumen, omasum, testes, etc. We also retrieved rumen-development RNA-seq data, including three samples from the rumens of 60-d-old embryos from Pan et al. [15], and eight samples from newborn lambs prior to feeding, and at 1 and 8 weeks after birth, from Clark et al. [34], covering different stages of rumen development. Adaptors and low-quality bases were removed using Trimmomatic (v0.30) [35]. The splice-aware mapper STAR (v2.4.0 k) [36] was used to align the reads with the Oar_rambouillet_v1.0 sheep genome, guided by the NCBI annotation. We then specified "quantMode" to provide coordinate information on the transcripts in the alignment, which served as the input for RSEM (v1.2.19) quantification software [37]. We employed the transcript per million (TPM) values generated by RSEM in the subsequent analyses.

We checked the sex bias of the genes by comparing their expression levels in the testis and ovary. We downloaded RNA-seq data for sheep testis and ovary tissues from NCBI (PRJEB19199) to align with the Oar_rambouillet_v1.0 sheep genome, guided by the NCBI annotation, using STAR (v2.4.0 k) [36]. After calculating the read counts using HTSeq [38], genes that were significantly differentially expressed ($P < 0.05$) between testis and ovary tissues were identified using the DESeq2 in R packages [39]. We classified genes with significantly higher expression in testis and ovary as male- and female-biased genes, respectively.

2.6. Functional analysis of new genes

We calculated the graded tissue specificity index τ [40] and selected genes with $\tau > 0.85$ and rank ≥5 as tissue bias-expressed genes. We analyzed the expression characteristics of the ruminant-specific genes by conducting weighted gene co-expression network analysis (WGCNA) [41] using RNA-seq data for 13 adult tissues. We removed the batch effect using the removeBatchEffect function in the R package limma [42], and then normalized the expression using the normalizeQuantiles function in limma. We also filtered out genes with a coefficient of variance (CV) (variance/mean) <0.3, because these genes were regarded as uninformative and are excluded in standard WGCNA practices [43].

We ran WGCNA using previously described parameters [44], specifying a soft threshold of 12, as the smallest value resulting in a scale-free R^2 fit of 0.85. Modules were numbered in order of decreasing number of gene members. Modules with a tissue Pearson's correlation coefficient >0.6 and *P*-value <0.01 were considered as tissue-biased modules, which mainly contained tissue-biased genes. For modules with an excess number of ruminant-specific genes, we converted sheep genes to human genes by gene name and performed Gene Ontology (GO) term enrichment analysis with the Metascape server [45], and presented the top five GO Biological Processes for these modules.

Regarding the analysis of RNA-seq data for rumen development, we normalized the expression level of genes in these samples using the R package limma and obtained the median expression levels of each gene at each stage. We then filtered out genes with an average expression level <0.5 and genes with little variation (variance/mean ratio < 0.3) among the four developmental stages [9,41]. The remaining genes were grouped into four clusters using Mfuzz [46], based on their expression levels during the developmental stages. For genes in each cluster, we converted sheep genes to human genes based on gene name, and then performed GO term enrichment analysis with Metascape.

2.7. Analysis of *PRD-SPRRII* gene family

The *PRD-SPRRII* gene family plays important roles in rumen development. To gain insights into the evolution of important new genes, we therefore subjected members of this gene family to further analysis. We determined the copy numbers of the gene family in other species using the sheep *PRD-SPRRII* protein sequence as the query, and obtained the corresponding protein sequences from the genome sequences of other species, including forest musk deer, black muntjac, lesser mouse-deer, giraffe, horse, sperm whale, and pig, using exonerate [47]. For each search result, we parsed the longest open reading frame that was not disturbed by a stop codon or frameshift and with >90% coverage. We then aligned the acquired protein sequences in Phylip format using MAFFT (v7.305b) [48] in the G-INS-i model, and selected the best maximum likelihood model using modeltest-ng [49]. We constructed a gene tree with raxml-ng [50] using the best model, TIM3 + G4. We then translated the protein sequence alignment to a codon sequence alignment using PAL2NAL (v14) [31]. Finally, we ran codeml in the PAML 4.8 package [32] to calculate *Ka*/*Ks* for the whole *PRD-SPRRII* and its parental *SPRRII* gene family [51] branch, and performed LRT with the assumption of *Ka*/*Ks* = 1, respectively. We also analyzed the mapping results of available ATAC-seq data for rumen epithelial cells from sheep embryos at 60 days [15,52] to identify possible regulatory elements adjacent to *PRD-SPRRII* and *SPRRII*.

3. Results and discussion

3.1. Novel high-quality genome assemblies allowed identification of ruminant-specific genes

We identified ruminant-specific genes using publically available data and our newly generated genomic data. Representative high-quality and chromosome-level genomes derived from long-read sequencing were previously only available for three of the six Ruminantia families: Bovidae, Cervidae, and Giraffidae, and high-quality genomic data for members of the Tragulidae and Moschidae were still lacking. Moschidae (represented here by the forest musk deer, *M. berezovskii*) is the closest sister group to sheep and other bovids, while Tragulidae (represented here by the lesser mouse-deer, *T. kanchil*) is the most basal ruminant taxons (Fig. 1a) [14]. These phylogenetically representative genomes are thus important for identifying ruminant-specific genes and helping to resolve the timing of their origin. In this study, we acquired and sequenced 185 Gb and 210 Gb Nanopore data for the lesser mouse-deer and forest musk deer, respectively, to assemble contigs. The contig N50 lengths of these draft genomes reached ~40 Mb (Fig. 1b), which were much longer than the contigs N50 of the preciously published short read-based ruminant genomes, with an average value <1 Mb [14], indicating higher continuity. Using the abundant Hi-C data (~100 folds), we improved the lesser mouse-deer and forest musk deer draft genomes to chromosome level (Fig. 1b and Fig. S1 online). Gene models were annotated for both genomes and covered over 92% of the conserved Benchmarking Universal Single-Copy Orthologs (BUSCO) gene sets (Fig. 1b), suggesting high integrity of the genome assemblies. In contrast, the average BUSCO score for previous ruminant genomes was around 87% [14]. We also annotated 38% and 41% of the respective genomic sequences in the two genomes as repeat sequences, similar to the reported proportions in other ruminants [19,22,23] (Fig. 1b). The chromosomal assemblies of lesser mouse-deer (*Tragulus kanchil*) and forest musk deer (*Moschus berezovskii*) were deposited in the US National Center for Biotechnology Information (NCBI) database under the project accession number PRJNA752261.

We use the sheep genome as the focal reference to identify new genes during ruminant evolution for the following reasons: (1) sheep represent a model species of ruminants, with high-quality gene annotation (BUSCO score, 99.4%) among ruminant genomes (Table S2 online); (2) relatively more functional genomic data (e.g., transcriptome data) are available for sheep; and (3) bovid species have high economic values. We first dated the emergence of genes based on our previous pipelines [7,9], and performed whole-genome synteny inference against ten outgroups, including the aforementioned two newly assembled genomes (Materials and methods). Among 20,464 protein-coding genes in sheep genomes, our synteny-based analysis identified 1064 ruminant-specific genes, which could be further divided into five age groups or phylostrata (branches 6–10 in Fig. 1a, Table S3 online). In contrast, 2281 genes (branches 1–5) were shared by more placental mammal groups, and 17,119 genes (branch 0) predated the split of placental mammals and opossums. Following conventional practice [33], we estimated that the rate of duplicate gene origin in sheep was 0.0012 per gene per million years (24.7 genes per million years), which is within the range estimated for vertebrates (~0.001–0.01 per gene per million years) [53]. In addition, we identified 620 forest musk deer-specific genes (Table S4 online). The number of Moschidae-specific genes was more or less the same as that of sheep-specific genes.

The reference-based SBP would identify more new genes in the reference genomes, especially species-specific new genes. In this study, we chose sheep as the reference for the above reasons, and dense representative species were not available for each family. Although this would not have a major effect on the identification of new genes in ancestral nodes of ruminants, this approach may miss many other species-specific new genes in non-sheep, especially non-bovid species. Future studies using genome data from more species and sub-lineages (e.g., subfamilies and genera) and new reference-free methods will allow us to reveal a more comprehensive picture of the origin of new genes in Ruminantia.

3.2. Ruminant-specific genes showed similar evolutionary and transcriptional patterns to primate- and rodent-specific genes

It is unknown if the evolutionary and transcriptional patterns of new genes in primates and rodents apply to all mammals [3,7–12]. Notably, however, the current results indicated consistent patterns in ruminants.

First, the ruminant-specific gene set included 912 (85.7%) DNA-mediated duplicates, 129 (12.1%) RNA-mediated duplicates (retrogenes), and 23 (2.2%) *de novo* genes (Fig. 2a). Alignment of these *de novo* genes with homologous sequences showed disablers such as frame-disrupting indels or premature termination codons (in Fig. 2b and Fig. S2 online). Most of the ruminant new genes thus originated from DNA-mediated duplications (85.7%), consistent with the results for primates and rodents [27].

Second, only some (525/1041) of the duplicated new genes were available for *Ka/Ks* analysis and for inferring the selection force (Materials and methods). Among them, 51 (51/525 = 9.7%) had been under significant purifying selection (Table S5 online), which was within the same magnitude as primate-specific genes (4%) [9]. Although both relaxation of functional constraints and positive selection may explain this pattern, 81% (862/1064) of ruminant-specific genes, including 81% (844/1041) of duplicate new genes, were expressed in at least one tissue (TPM > 1, Table S6 online), indicating that most duplicate genes may be functional.

Third, using the RNA-seq data profiling for 13 adult sheep tissues [34], we found that genes showed age-dependent expression profiles, as for primate- and rodent-specific genes [9,10]. The youngest genes, i.e., ruminant-specific genes (branches 6–10) were thus transcribed with a median TPM value <1 in four tissues, while the medium-age and oldest gene groups were transcribed at a median TPM value > 1 in four tissues, and >100 in all 13 tissues, respectively (Fig. 2c, d).

Fourth, comparison of the gene expression levels in testis and ovary showed that a higher proportion of new genes had male-biased expression on the X chromosome than genes on autosomes (Fig. 2e, Table S7 online). These results supported the idea that new genes on the X chromosome were more likely to be male-biased [7], suggesting that young recessive male-benefit genes on the X chromosome may have been fixed at an early stage of their emergence in ruminants.

Overall, these results suggest that the evolution and expression patterns of ruminant-specific genes are similar to those in other mammals, such as primates and rodents. Although each of these mammalian taxa has very specific phenotypes, the evolution of new genes thus follows general rules, indicating that other mammals may also have similar patterns of new gene evolution.

3.3. Roles of new genes in rumen evolution

The rumen is the most innovative organ in ruminants, and first appeared in a ruminant ancestor about 35 million years ago [14,54]. The relatively recent origin of the rumen and the availability of high-quality genomes and expression data provide a unique opportunity to understand the genetic basis of the evolution of

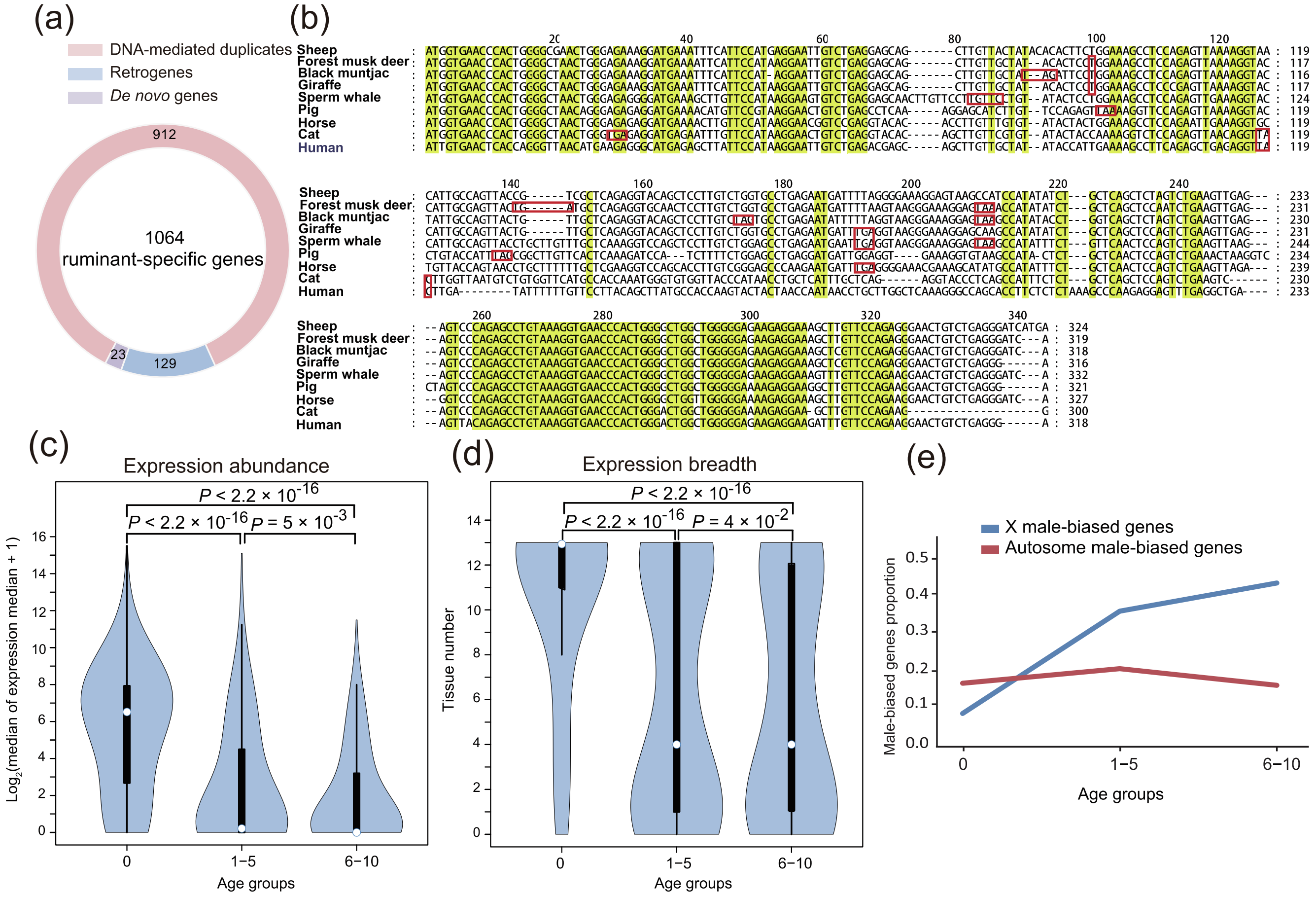


Fig. 2. Evolution and expression patterns of ruminant-specific genes. (a) Numbers and proportions of 1064 ruminant-specific genes originating from DNA-mediated duplication, RNA-mediated duplication, and *de novo* birth, respectively. (b) Sequence alignment of candidate *de novo* gene *LOC114110053* across mammals. Open reading frame information was based on the sheep gene. Sites with identical sequences in all species are indicated in yellow. Premature termination codons and frame-disrupting indels are labeled in red. Violin plots of expression profiles of ruminant-specific genes across gene-age groups, showing (c) $\log_2$-based median expression in 13 adult sheep tissues and (d) numbers of tissues in which genes were expressed. Black bars indicate interquartile range, violin curve indicates the probability density of the data, and white dot shows the median value. The genes were divided into three gene-age groups as defined in Fig. 1a. *P*-values between age groups according to Wilcoxon's test are also presented. (e) Proportion of genes with male-biased expression in each gene-age group. Blue and red lines indicate X chromosome and autosomes, respectively. The age groups in (c), (d), and (e) were depicted based on the phylogenic relationship of extant species (Fig. 1a). Group 0 indicates the oldest genes shared by vertebrates. Groups 1–5 refer to genes that appeared predating sheep–sperm whale split. Groups 6–10 refer to ruminant-specific genes.

new organs, and especially the role of new genes. The function of the adult rumen has been shown to depend mainly on genes that predate sheep–sperm whale split (referred to as old genes) recruited from other organs [14,15]; however, information on the roles of new genes in rumen evolution has been limited to studies of just a few new genes [14]. We identified a comprehensive new gene set using the most comprehensive high-quality ruminant genome and transcriptome data (Fig. 1a). Although most (14,585/14,968 = 97.4%) genes expressed in the adult rumen were old genes (branches 0–5), new genes (383/14,968 = 2.6%) also played roles in rumen evolution. The proportion of old genes (306/317 = 96.5%) with rumen-biased expression was also much higher than that of new genes (11/317 = 3.5%). We further revealed the roles of new genes in the evolution of the rumen and ruminants by co-expression network analysis [48] based on the above RNA-seq dataset (Materials and methods), and inferred the potential functions of these new genes according to their co-expression partners. We identified a total of 24 co-expression modules (Fig. S3 online), of which five (ME1, 2, 7, 14, and 15) consisted mainly of tissue-biased genes (Fig. 3a and Fig. S3 online; Materials and methods). These modules included cerebellum-biased ME1, testis-biased ME2, and immune cell-biased ME15. GO term enrichment analysis identified functions consistent with the expression biases of these genes (Fig. 3a, Tables S8–S10 online). The distribution of ruminant-specific genes echoed the situation in primates and rodents: (1) new genes were enriched in the testis (Fig. 3a), male reproduction [3,9], and immune response [9]; (2) new genes were depleted in the adult brain [9]; and (3) most modules (14/23, 61%) contained broadly expressed genes with fewer ruminant-specific genes than the genome average, and this difference was significant for five modules (ME9, 10, 16, 21, and 23) (Fig. 3a, Fig. S3 online), which was also consistent with the known tendency of new genes to be tissue-specific (Fig. 2b, c) [9,10]. We further investigated the functional significance of new genes by checking the proportion of transcription factors in each module in the WGCNA results, and found that transcription factors were depleted in the new gene enriched modules, such as ME2, 3, and 15. This result suggested that new genes tended to have downstream functions (Fig. S4 online).

In addition to this general pattern, we considered how ruminant-specific genes contributed to the function and development of the rumen. The rumen and other tissues, such as the omasum and esophagus, had similar expression patterns in housekeeping gene modules, including ME5, 10, 13, and 16. These tissues may share the same genetic elements, and therefore have fewer new genes. This result was consistent with the previous

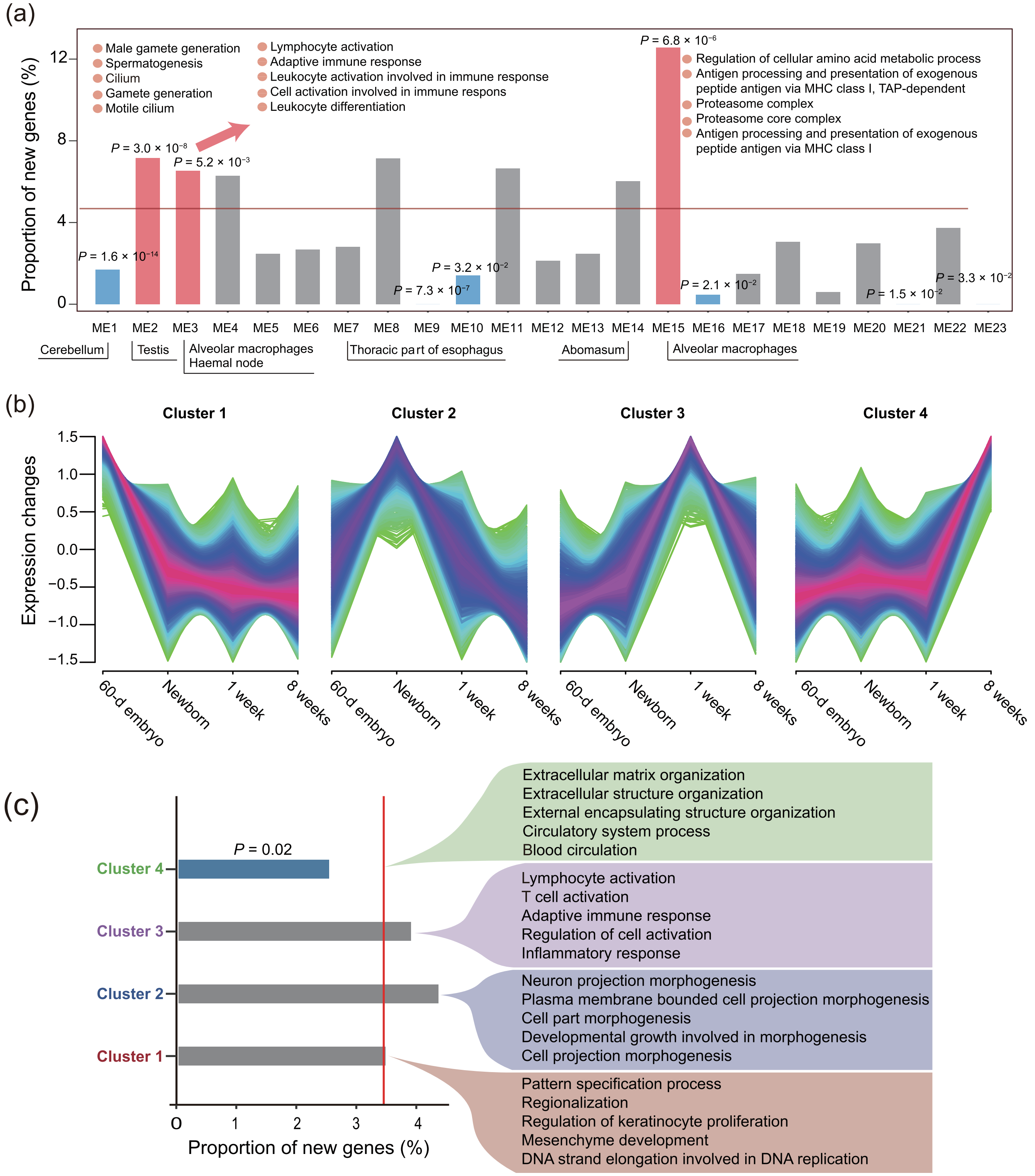


Fig. 3. Expression landscape of ruminant-specific genes in adult sheep tissues and rumen at different developmental stages. (a) Proportion of ruminant-specific genes within each co-expression module defined using transcriptomic data for 13 adult sheep tissues. The genome-wide percentage of ruminant-specific genes (4.7%) is indicated by the horizontal red line, and modules with significant increased or decreased expression (Bonferroni-corrected $P < 0.05$) of ruminant-specific genes are indicated by red and blue blocks, respectively, and non-significant modules by the gray bar. Corresponding multi-testing adjusted P-values, which come from the comparison of the percentage of ruminant-specific genes in each module with genome-wide percentage of ruminant-specific genes are shown above the boxes. The five tissue-biased modules are labeled with corresponding tissue names. The top five significantly enriched GO terms for both testis and alveolar macrophage-biased modules are shown. (b) Developmental expression profiles of genes found to be highly expressed in rumens from 60-d embryos, newborn lambs, and lambs 1 week and 8 weeks after birth (designated Clusters 1–4, respectively; obtained from soft clustering analysis using Mfuzz). Light blue and green lines indicate expression profiles of genes with low membership values, and red and dark blue lines indicate profiles of genes with high membership values. (c) Bar plots of ruminant-specific genes enriched in derived modules and associated functional enrichment analysis in the four clusters. Vertical red line indicates the average percentage of ruminant-specific genes (3.5%), and the top five significantly enriched GO terms in each cluster are shown. P-values come from the comparison of the percentage of ruminant-specific genes in each cluster with the average percentage of ruminant-specific genes are shown above the boxes.

report [15] indicating that the rumen was originated from the esophagus. We further analyzed the transcriptomic data for four rumen-developmental stages: 60-d-old embryos [15], neonatal lambs at parturition euthanized immediately prior to first feed, 1-week-old lambs fed with milk, and 8-week-old lambs at the stage when rumen function is established [34]. We clustered transcriptomic data into four clusters using Mfuzz, and showed that the new-gene expression patterns in each cluster differed among the developmental stages (Fig. 3b). More new genes were expressed in the developing rumen than in the mature rumen (8 weeks after birth), although the difference was not significant (Bonferroni-corrected $P > 0.05$) (Fig. 3c). This was consistent with the depletion of new gene expression in the adult rumen (Fig. 3a). This trend echoed the importance of new genes in the development of human brains, compared with the adult brain [9,55].

The current results present a more complete picture of the pattern of new gene evolution in ruminants, compared with the previous sporadic discoveries of new ruminant genes [14], as well as providing more evidence for the roles of new genes in the evolution of new mammalian organs. Our results also showed that genes co-opted from other tissues may play more pervasive roles than new genes in the evolution of new organs. Further regulatory studies are needed to reveal the comprehensive genetic basis underlying phenotypic novelties, especially the roles of new regulatory elements in addition to young genes. This comprehensive picture will allow us to establish the general pattern and rules underlying the origin and evolution of new mammalian organs. In addition, more data and analyses are also needed to understand the evolution of other ruminant traits, such as the origin and evolution of cervid antlers, as the only mammalian organ that can be fully regenerated, and which is thus of great medical significance [56].

3.4. The ruminant-specific gene family PRD-SPRRII was preferentially expressed in the rumen and was subject to adaptive radiation

We further analyzed the *PRD-SPRRII* gene family, because of the emergence of many new duplicates during ruminant evolution and their possible roles in the development of the keratin-rich surface of the rumen [54]. Ruminant-specific *PRD-SPRRII* and *PRD-SPRRII*-like genes are homologous to the *SPRRII* gene family [51], with the major difference that the amino termini of *PRD-SPRRII* and *PRD-SPRRII*-like genes are enriched in proline and histidine [57]. Among 13 adult sheep tissues, *PRD-SPRRII* genes were only expressed in the rumen, reticulum, and omasum, with higher expression in the rumen (Fig. 4a). However, information on the evolution of this new gene family and its contribution to rumen development remains limited.

We therefore conducted an in-depth analysis of this gene family by analyzing the mechanism of gene origin and using the sheep *PRD-SPRRII* protein sequence as a template to predict *PRD-SPRRII*-like sequences in other ruminant genomes. We found four main results. First, the *PRD-SPRRII* family originated by DNA-level duplication of an *SPRRII* gene followed by cumulative mutations in one duplicate, leading to enrichment of proline and histidine at the N-terminus of the encoded protein, with subsequent amplification in ruminants (Fig. 4b), and the highest copy number in the sheep genome (Fig. 4c). The structure and function of the rumen continued to evolve in different ruminant lineages. For example, bovid species can digest hard forage and have the highest copy number of lysozyme C genes [14], while cervids tend to eat soft leaves. Second, members of the *PRD-SPRRII* and *SPRRII* gene families formed respective clusters, indicating that the *PRD-SPRRII* gene family has diverged from its parental genes at the sequence level (Fig. 4d). In addition, the *PRD-SPRRII* and *SPRRII* gene families have been subjected to different selection pressures, as revealed by branch tests, with the former being subject to marginally significant positive natural selection ($P = 0.087$, Fig. 4d) and the latter to negative selection ($P = 1.4 \times 10^{-4}$, Fig. 4d), suggesting that the *PRD-SPRRII* gene family may be amplified by adaptive radiation [58]. We analyzed the available ATAC-seq data for rumen epithelial cells from 60-d sheep embryos [15,52] to identify possible regulatory elements adjacent to *PRD-SPRRII* and *SPRRII.* We detected several ATAC-seq peaks around the upstream 2 kb regions of these new genes, which were absent from the equivalent regions of old *SPRRII* genes (Table S11 online). More regulatory data are needed to clarify the evolution of regulatory elements associated with new genes. Third, members of this gene family showed divergent expression during rumen development (Fig. 4e). Most *PRD-SPRRII* copies retained the expression pattern of their parental genes at different developmental stages of the rumen; for example, *PRD-SPRRII* and *PRD-SPRRII*-like-1/3/4/6 were not expressed in the rumen at the 60-d embryo stage, but their expression increased at 1 and 8 weeks after birth, while *PRD-SPRRII*-like-5 showed distinctly different expression compared with other copies, being expressed most strongly in the newborn stage.

The *PRD-SPRRII* family thus originated and was amplified during ruminant evolution, possibly driven by functional divergence at different developmental stages of the rumen. A new gene may undergo two evolutionary scenarios, i.e., continuous positive selection after its origin, in the form of pervasive positive selection or one-time, episodic positive selection. The *PRD-SPRRII* family seems to fit the first scenario, with their radiation shaped by diversifying positive selection. Most new genes are expected to initially evolve neutrally, and then experience positive selection after the emergence of a beneficial mutation [59], with many duplicate genes becoming pseudogenes during this process.

4. Conclusions

We used new high-quality genome assembly data to obtain a comprehensive and accurate ruminant-specific gene set using the SBP [7,9]. We conducted evolutionary and expression analyses and showed that ruminant-specific genes generally shared similar patterns to those in other mammals, such as humans and mice. The rumen is the hallmark innovative trait in ruminants. Our results showed that the rumen mostly recruited old genes from other organs/tissues, but new genes also contributed to its evolution, especially during its development. Detailed analysis showed that the rumen-specific *PRD-SPRRII* gene family has undergone significant amplification in ruminants, and its members have been subject to functional differentiation during rumen development. To the best of our knowledge, this is the first research on the roles of new genes in the origin of a new mammalian organ. Overall, the results provide a valuable data resource for studies of ruminant biology, and also confirm the occurrence of similar new gene evolution patterns among different mammalian taxa. The current findings regarding the involvement of new genes in rumen development and function also shed light on the origin of new mammalian organs.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32030016), the Natural Science Foundation of Shaanxi Province (2020JQ-149), the State Key Laboratory of Genetics Resources and Evolution (GREKF19-11), the Innovation Founda-

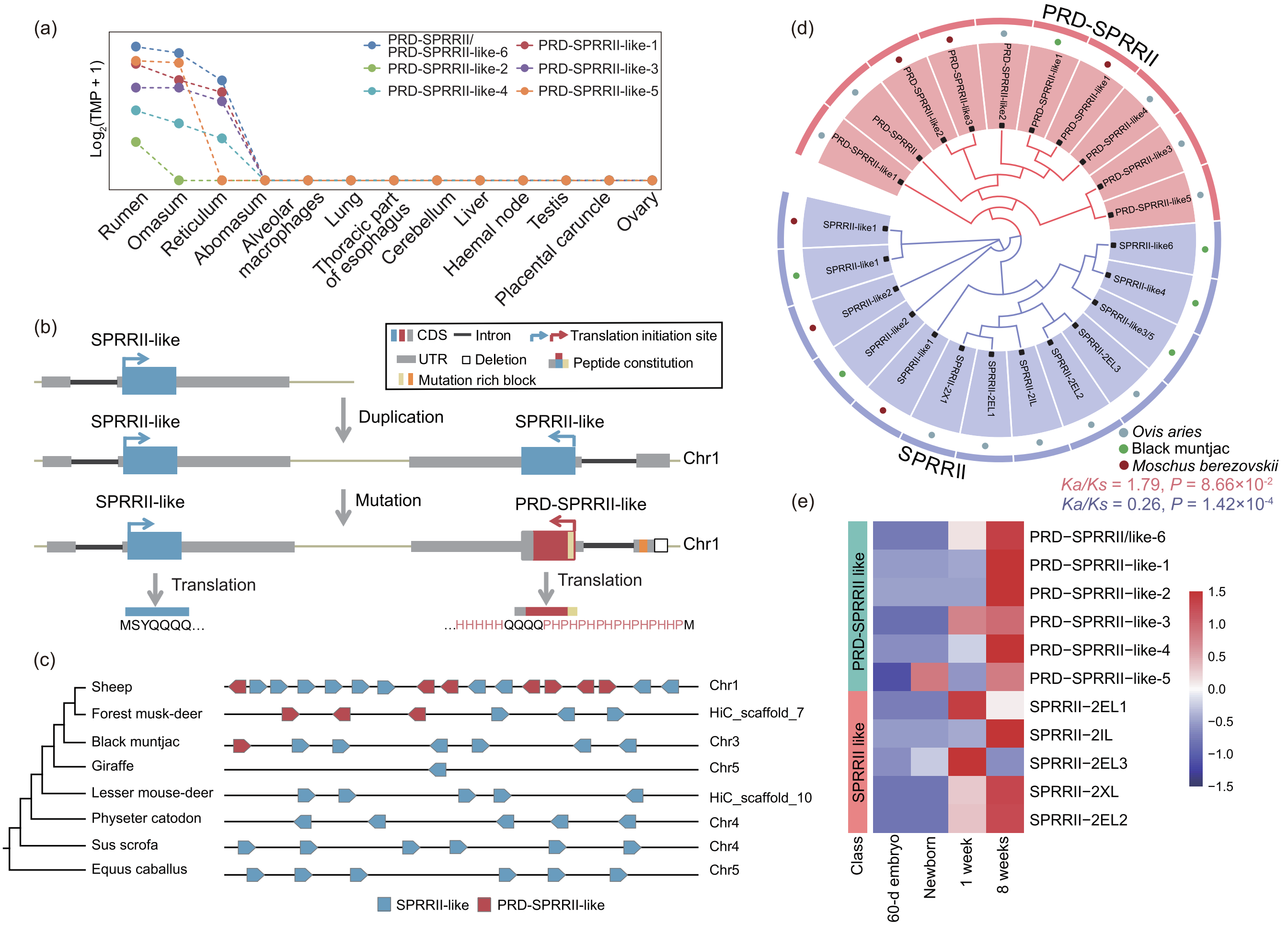


Fig. 4. Mechanisms of origin and expression patterns of *PRD-SPRRII* and *SPRRII*-like genes. (a) Expression profile of ruminant-specific *PRD-SPRRII* gene family in 13 adult sheep tissues. Expression values for different members of the family are shown in different colors. None of the gene members were expressed in the abomasum or subsequent tissues. (b) Mechanism of origin of *PRD-SPRRII* and *PRD-SPRRII*-like genes. *PRD-SPRRII* and *PRD-SPRRII*-like genes originated from DNA duplication, followed by mutations that cumulatively enriched the 5′-terminal of one duplicate copy in terms of proline and histidine. Other regions, such as the first exon, also accumulated numerous mutations. (c) *PRD-SPRRII* copy numbers in different species on the phylogenetic tree. Red and blue rectangles represent the coding sequence region, and arrows indicate transcription direction. (d) Gene tree of *PRD-SPRRII* and *SPRRII* gene families in ruminants marked with different colored circles. The entire *PRD-SPRRII* gene family and selection pressure are marked in red, and the corresponding *SPRRII* gene family is in lilac. (e) Expression profiles of *PRD-SPRRII* and *SPRRII* gene family members during rumen development. Heatmap shows the expression of *PRD-SPRRII* and *SPRRII* gene family members during rumen development. The abscissa represents different rumen-development stages. Color code bar indicates relative expression level. *PRD-SPRRII* and *PRD-SPRRII*-like-6 sequences were too similar for accurate quantification, and their expression values were therefore merged.

tion for Doctor Dissertation of Northwestern Polytechnical University to Zeshan Lin, and the Shaanxi Program for Support of Top-notch Young Professionals to Lei Chen. We thank Huijing Ma, Dan Zhang, Daqi Yu, Hao Yuan, Tianhan Su, and Yufei Zhang from the Institute of Zoology, Chinese Academy of Sciences, and Yi Shao from the Chinese Institute for Brain Research (CIBR) for their valuable advice on computational analysis. We thank Yongxin Li, Kun Wang, and Wenjie Zhang from the School of Ecology and Environment, Northwestern Polytechnical University, for their valuable suggestions on preparing the cover letter.

Author contributions

Chunyan Chen, Yuan Yin, and Haorong Li performed the computational analyses. Lei Chen, Botong Zhou, Jiong Zhou, Xiaofang Zhou, Zhipeng Li, Zeshan Lin, and Guichun Liu took part in sample collection and genome assembly. Ru Zhang and Xiangyu Pan provided some important analysis suggestions. Chunyan Chen, Yuan Yin, Lei Chen, Qiang Qiu, Yong E. Zhang, and Wen Wang wrote and revised the manuscript. Wen Wang and Yong E. Zhang conceived and designed the project.

Appendix A. Supplementary materials

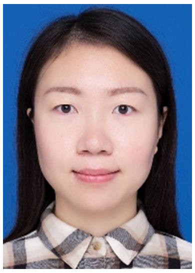
Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2022.01.023.

References

[1] Long M, Betrán E, Thornton K, et al. The origin of new genes: glimpses from the young and old. Nat Rev Genet 2003;4:865–75.
[2] Ding Y, Zhou Q, Wang W. Origins of new genes and evolution of their novel functions. Annu Rev Ecol Evol Syst 2012;43:345–63.
[3] Long M, VanKuren NW, Chen S, et al. New gene evolution: little did we know. Annu Rev Genet 2013;47:307–33.
[4] Zhang YE, Long M. New genes contribute to genetic and phenotypic novelties in human evolution. Curr Opin Genet Dev 2014;29:90–6.
[5] VanKuren NW, Long M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. Nat Ecol Evol 2018;2:705–12.
[6] Zhang L, Ren Y, Yang T, et al. Rapid evolution of protein diversity by *de novo* origination in Oryza. Nat Ecol Evol 2019;3:679–90.

[7] Zhang YE, Vibranovski MD, Landback P, et al. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol 2010;8:e1000494.
[8] Zhou Q, Zhang G, Zhang Y, et al. On the origin of new genes in Drosophila. Genome Res 2008;18:1446–55.
[9] Shao Y, Chen C, Shen H, et al. Gentree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. Genome Res 2019;29:682–96.
[10] Zhang YE, Landback P, Vibranovski M, et al. New genes expressed in human brains: implications for annotating evolving genomes. BioEssays 2012;34:982–91.
[11] Yu D, Shi W, Zhang YE. Underrepresentation of active histone modification marks in evolutionarily young genes. Insect Sci 2017;24:174–86.
[12] Ellegren H. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. Nat Rev Genet 2011;12:157–66.
[13] Heide M, Haffner C, Murayama A, et al. Human-specific *ARHGAP11B* increases size and folding of primate neocortex in the fetal marmoset. Science 2020;369:546–50.
[14] Chen L, Qiu Q, Jiang Y, et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science 2019;364:eaav6202.
[15] Pan X, Cai Y, Li Z, et al. Modes of genetic adaptations underlying functional innovations in the rumen. Sci China Life Sci 2021;64:1–21.
[16] Durand NC, Shamim MS, Machol I, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;3:95–8.
[17] Dudchenko O, Batra SS, Omer AD, et al. *De novo* assembly of the *aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 2017;356:92–5.
[18] Durand NC, Robinson JT, Shamim MS, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 2016;3:99–101.
[19] Liu C, Gao J, Cui X, et al. A towering genome: experimentally validated adaptations to high blood pressure and extreme stature in the giraffe. Sci Adv 2021;7.
[20] Yin Y, Fan Huizhong, Zhou Botong, et al. Molecular mechanisms and topological consequences of drastic chromosomal rearrangements of muntjac deer. Nat Commun 2021;12:6858.
[21] Fan G, Zhang Y, Liu X, et al. The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. Mol Ecol Resour 2019;19:944–56.
[22] Warr A, Affara N, Aken B, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. GigaScience 2020;9:giaa051.
[23] Kalbfleisch TS, Rice ES, DePriest Jr MS, et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol 2018;1:197.
[24] Buckley RM, Davis BW, Brashear WA, et al. A new domestic cat genome assembly based on long sequence reads empowers feline genomic medicine and identifies a novel gene for dwarfism. PLoS Genet 2020;16:e1008926.
[25] Dudchenko O, Shamim MS, Batra SS, et al. The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 2018. https://doi.org/10.1101/254797.
[26] Harris R. Improved pairwise alignment of genomic DNA. Pennsylvania State Univ; 2007. PhD thesis.
[27] Zhang YE, Vibranovski MD, Krinsky BH, et al. Age-dependent chromosomal distribution of male-biased genes in Drosophila. Genome Res 2010;20:1526–33.
[28] El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein family's database in 2019. Nucleic Acids Res 2019;47:D427–32.
[29] Betrán E, Thornton K, Long M. Retroposed new genes out of the X in Drosophila. Genome Res 2002;12:1854–9.
[30] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421.
[31] Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 2006;34:W609–12.
[32] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–91.
[33] Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science 2000;290:1151–5.
[34] Clark EL, Bush SJ, McCulloch MEB, et al. A high-resolution atlas of gene expression in the domestic sheep (*Ovis aries*). PLoS Genet 2017;13:e1006997.
[35] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[36] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.
[37] Li B, Dewey CN. Rsem: Accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinf 2011;12:323.
[38] Anders S, Pyl PT, Huber W. HTSeq–a python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–9.
[39] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
[40] Yanai I, Benjamin H, Shmoish M, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 2005;21:650–9.
[41] Langfelder P, Horvath S. Wgcna: An R package for weighted correlation network analysis. BMC Bioinf 2008;9:559.
[42] Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.
[43] Kang HJ, Kawasawa YI, Cheng F, et al. Spatio-temporal transcriptome of the human brain. Nature 2011;478:483–9.
[44] Parikshak N, Luo R, Zhang A, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell 2013;155:1008–21.
[45] Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10:1523.
[46] Kumar L, Futschik ME. A software package for soft clustering of microarray data. Bioinformation 2007;2:5–7.
[47] Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinf 2005;6:31.
[48] Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–66.
[49] Darriba D, Posada D, Kozlov AM, et al. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol 2020;37:291–4.
[50] Kozlov AM, Darriba D, Flouri T, et al. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 2019;35:4453–5.
[51] Kypriotou M, Huber M, Hohl D. The human epidermal differentiation complex: cornified envelope precursors, s100 proteins and the 'fused genes' family. Exp Dermatol 2012;21:643–9.
[52] Fu W, Wang R, Nanaei HA, et al. 2021. RGD v2.0: a major update of the ruminant functional and evolutionary genomics database. Nucleic Acids Res 2021; gkab887.
[53] Lynch M, Conery JS. The evolutionary demography of duplicate genes. J Struct Funct Genomics 2003;2003:35–44.
[54] Jiang Y, Xie M, Chen WB, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. Science 2014;344:1168–73.
[55] Zhang YE, Landback P, Vibranovski MD, et al. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol 2011;9: e1001179.
[56] Wang Y, Zhang C, Wang N, et al. Genetic basis of ruminant headgear and rapid antler regeneration. Science 2019;364:eaav6335.
[57] Wang L, Baldwin RL, Jesse BW. Identification of two cDNA clones encoding small proline-rich proteins expressed in sheep ruminal epithelium. Biochem J 1996;317:225–33.
[58] Francino MP. An adaptive radiation model for the origin of new gene functions. Nat Genet 2005;37:573–7.
[59] Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 2010;11:97–108.

Chunyan Chen is a Postdoctoral Fellow at Northwestern Polytechnical University. She received her bachelor's degree from Northwest A&F University and her Ph.D. from the Institute of Zoology, Chinese Academy of Sciences. She presently uses synteny information on genome alignment to identify species- or lineage-specific new genes in multiple species, infer the mechanisms of origin of new genes, and analyze the functions of new genes using multi-omics data.



Yuan Yin is a Ph.D. candidate at Northwestern Polytechnical University. She graduated from the School of Life and Technology, Huazhong University of Science and Technology in 2016 with a bachelor's degree in engineering. She presently takes advantage of the technology of comparative genomics, transcriptomics, three-dimensional genomics, and so on to explore the molecular mechanisms and topological consequences of chromosomal fusion in muntjac deer.

Haorong Li is a Ph.D. candidate at the School of Ecology and Environment, Northwestern Polytechnical University. She received her B.S. degree from the College of Life Science and Technology, Huazhong University of Science and Technology. Her research interests focus on evolutionary genetics and genomics.

Lei Chen is a Professor at Northwestern Polytechnical University. He received his bachelor's degree from Wuhan University and doctoral degree from Kunming Institute of Zoology, Chinese Academy of Sciences. He has been working on the phylogenomics, adaptive evolution, and genetic innovations of ungulates by integrating multi-omics data (genome, transcriptome, and regulatome) in a phylogenetic framework, and exploring possible applications in future animal husbandry and medicine.

Qiang Qiu is a Professor at the School of Ecology and Environment, Northwestern Polytechnical University of China. He studies the genetic basis of extreme environmental adaptation, the origin and evolution of new organs and special adaptive traits of ruminants.

Yong Zhang received his Ph.D. degree in bioinformatics from Peking University. He worked as a Postdoctoral Fellow and a bioinformatician at the University of Chicago, and later set up his lab at the Institute of Zoology, Chinese Academy of Sciences. His research focuses on the origins of new genes, including how mutational mechanisms generate new genes, how development recruits new genes, and how mutational mechanisms are transformed as novel molecular tools.

Wen Wang is a Professor at Northwestern Polytechnical University. He graduated from Wuhan University with a bachelor's degree and received his Ph.D. degree from the Chinese Academy of Sciences. He worked as a Postdoctoral Fellow and Research Assistant at the University of Chicago. He has been a Professor at the Chinese Academy of Sciences. His research focuses on evolutionary genomics and genetic innovation mechanisms.