ISME

www.nature.com/ismej

ARTICLE

Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates

Zongjun Li[1,9], Xiangnan Wang[1,9], Yu Zhang[1,9], Zhongtang Yu[2,9], Tingting Zhang[1,9], Xuelei Dai[1], Xiangyu Pan[1], Ruoxi Jing[1,3], Yueyang Yan[1], Yangfan Liu[1], Shan Gao[1], Fei Li[4], Youqin Huang[1,4], Jian Tian[5], Junhu Yao[1], XvPeng Xing[6], Tao Shi[1], Jifeng Ning[7], Bin Yao[5], Huoqing Huang[5]✉ and Yu Jiang[1,8]✉

Understanding the biodiversity and genetics of gut microbiomes has important implications for host physiology and industrial enzymes, whereas most studies have been focused on bacteria and archaea, and to a lesser extent on fungi and viruses. One group, still underexplored and elusive, is ciliated protozoa, despite its importance in shaping microbiota populations. Integrating single-cell sequencing and an assembly-and-identification pipeline, we acquired 52 high-quality ciliate genomes of 22 rumen morphospecies from 11 abundant morphogenera. With these genomes, we resolved the taxonomic and phylogenetic framework that revised the 22 morphospecies into 19 species spanning 13 genera and reassigned the genus *Dasytricha* from *Isotrichidae* to a new family *Dasytrichidae*. Comparative genomic analyses revealed that extensive horizontal gene transfers and gene family expansion provided rumen ciliate species with a broad array of carbohydrate-active enzymes (CAZymes) to degrade all major kinds of plant and microbial carbohydrates. In particular, the genomes of *Diplodiniinae* and *Ophryoscolecinae* species encode as many CAZymes as gut fungi, and ~80% of their degradative CAZymes act on plant cell-wall. The activities of horizontally transferred cellulase and xylanase of ciliates were experimentally verified and were 2–9 folds higher than those of the inferred corresponding bacterial donors. Additionally, the new ciliate dataset greatly facilitated rumen metagenomic analyses by allowing ~12% of the metagenomic sequencing reads to be classified as ciliate sequences.

The ISME Journal (2022) 16:2775–2787; https://doi.org/10.1038/s41396-022-01306-8

INTRODUCTION

The genomes of hundreds of thousands of gut prokaryotes have been assembled, and they have greatly helped advance our understanding of their taxonomic and functional diversity and important roles in host health and nutrition [1–4]. Trichostome ciliates (of the phylum *Ciliophora*, class *Litostomatea*) are ubiquitous in the gut of vertebrates [5, 6], and they play crucial roles in modulating the gut microbiome through predation, competition, and symbiosis [7, 8]. However, genomic research on gut ciliates has lagged [1], and so far only *Entodinium caudatum*, one of the most common rumen ciliate species, has been subjected to genomic study [9]. The lack of genomic information on ciliates has become an obstacle for holistic studies of the gut microbiome, rumen microbiome in particular [10–12], as ciliates can account for up to 50% of the total rumen microbial biomass and play important role in rumen functions [7, 13].

Despite their prominence in the rumen ecosystem and extensive previous research for nearly two centuries [14], much remains to be learned about rumen ciliates [15, 16]. Based on morphological features, rumen ciliates are classified into two clades: the family *Isotrichidae* in the order *Vestibuliferida* (with cilia covering the entire cell surface) and the family *Ophryoscolecidae* in the order *Entodiniomorphida* (with localized zones of cilia including one or two adoral cilia zones) [5, 7]. It has been recognized, however, that taxonomic identification and classification of ciliates based on morphological features are challenging [5, 7, 17], as the same species can exhibit a few different morphotypes [18]. The 18 S rRNA gene of ciliates has been used to aid species identification and protozoal community analyses [19–21], but it has several limitations in proposing a robust taxonomic framework because it is too conserved to allow for distinguishing species or uncovering cryptic species [22] and results in topological discrepancies even at the subfamily level [6, 23].

It has also been challenging to determine or investigate the actual metabolism of rumen ciliates because of the lack of and inability to establish axenic cultures [16, 24]. As a result, the

[1]Center for Ruminant Genetics and Evolution, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China. [2]Department of Animal Sciences, The Ohio State University, Columbus, OH 43210, USA. [3]College of Animal Engineering, Yangling Vocational & Technical College, Yangling 712100, China. [4]State Key Laboratory of Grassland Agro-ecosystems, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou 730020, China. [5]State Key Laboratory of Animal Nutrition, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China. [6]Key Laboratory of Animal Biotechnology of the Ministry of Agriculture, College of Veterinary Medicine, Northwest A&F University, Yangling 712100, China. [7]College of Information Engineering, Northwest A&F University, Yangling 712100, China. [8]Center for Functional Genomics, Institute of Future Agriculture, Northwest A&F University, Yangling 712100, China. [9]These authors contributed equally: Zongjun Li, Xiangnan Wang, Yu Zhang, Zhongtang Yu, Tingting Zhang. ✉email: huanghuoqing@caas.cn; yu.jiang@nwafu.edu.cn

Received: 22 February 2022 Revised: 4 August 2022 Accepted: 9 August 2022
Published online: 19 August 2022

previous knowledge of rumen ciliate metabolism can only be inferred from studies using monocultures (that contain bacteria and other microbes) or defaunation (removal of rumen protozoa chemically or through physical isolation), both of which introduce confounding effects of other microbes and residual effects [16, 25]. Recent metatranscriptomic, single-cell transcriptomic (*Entodinium furca, Diplodinium dentatum*, and *Isotricha intestinalis*), and genomic (*Ent. caudatum*) studies provided new insights into the metabolism of rumen ciliates [9–11, 26]. However, the functional diversification and evolutionary mechanisms underlying most rumen ciliate genera and species remain to be determined. Additionally, the unique genome structures of ciliates (e.g., nuclear dimorphism, chromosomal fragmentation) make them valuable models to understand many important fundamental molecular and cellular processes and features, such as catalytic RNA, telomerase, and histone acetylation [27, 28]. However, most knowledge comes from studies on a few free-living aerobic ciliate species within the classes *Oligohymenophorea* and *Spirotrichea* [29, 30], while rumen ciliates, which are anaerobic and symbiotic, are exclusively found in the class *Litostomatea* [5].

In this study we sequenced, de novo assembled, and analyzed 52 high-quality single-cell amplified genomes (SAGs) from 22 rumen ciliate morphospecies spanning 11 abundant morphogenera. An assembly-and-identification pipeline was developed to help improve the quality and accuracy of the assembled ciliate genomes. This unparalleled genome catalog of rumen ciliates greatly facilitated the understanding of their taxonomy, phylogeny, metabolism, and ecology.

MATERIALS AND METHODS

Single-ciliate-cell isolation, identification, and sequencing

All the experimental procedures on animals were conducted following the guidelines of the Regulations for the Administration of Affairs Concerning Experimental Animals (Ministry of Science and Technology, China, 2013) and were approved by the Northwest A&F University Animal Care and Use Committee.

Fresh rumen fluid samples were collected from 14 ruminally fistulated ruminants (6 Holstein cattle, 3 Qinchuan beef steers, and 5 Guanzhong dairy goats) all kept in Yangling, Shannxi Province, China. All the animals were healthy without any medical treatment at the time of rumen fluid collection. Individual ciliate cells of different morphospecies (Fig. S1) were purposely isolated from rumen fluid samples using glass micropipettes under an inverted microscope (Nikon, TI-FL, Japan), and individual ciliate cells of the same morphospecies with slightly distinct morphologies were also purposely isolated to obtain potential cryptic species. The collected ciliate cells were repeatedly transferred and washed using MB9 buffer [31] until no other ciliate cells could be found microscopically, and then were transferred to individual microtubes each containing 2.5 µl of cell lysis buffer (buffer RLT Plus, Qiagen). A total of 69 single ciliate cells representing 22 morphospecies (Fig. 1A and Table S1) were isolated. Of the 69 cells, 24 were subjected to parallel whole-genome amplification and whole-transcriptome amplification for sequencing of both the genomes and transcriptomes using G&T-seq [32, 33], while the remaining 45 were subjected to only whole-genome amplification using multiple displacement amplification [34] for genomic sequencing (Table S1). Illumina TruSeq sequencing libraries were prepared for each of the single-cell amplified genomes and transcriptomes and then sequenced (paired-end, 2 × 150 bp) on a NovaSeq 6000 platform (Illumina, Inc., USA). The single-cell amplification and sequencing were conducted by Annoroad Gene Technology Co., Ltd. (Beijing, China).

The rumen ciliates were identified morphologically with light microscopy, scanning electron microscopy, and confocal laser scanning microscopy [35, 36]. The morphological description and micrographs of the isolated ciliate species are summarized in Figs. S1–3. The taxonomic identification of the collected single ciliates was confirmed by sequencing analysis of their 18 S rRNA genes after the genome assembly. The ciliates in the rumen samples (Table S2) were counted using a Sedgewick-Rafter chamber as described previously [35].

Genome assembly and ciliate sequence identification

Because rumen ciliate cells carry other microbes as endosymbionts or engulfed microbes (primarily bacteria) and because the DNA of the latter can also be amplified and sequenced, we developed an assembly-and-identification pipeline to recover and identify ciliate macronuclear sequences, which includes five steps (Fig. 1B). Firstly, we used three assemblers to improve the assembly quality (Fig. S4). Briefly, the cleaned genome sequencing reads of each single-ciliate cell were assembled using Megahit (v1.2.1) [37] and SPAdes (v3.13.1) [38] separately with the default parameters. The contigs assembled by the two assemblers were then merged to extend their lengths using quickmerge (v1.2.1) [39].

Secondly, the sequence characteristics of rumen ciliates were identified as follows: (1) The ends (30 bp) of single-cell assembled contigs were searched for the telomeric repeats of ciliates using MEME (v4.12.0) [40], and the macronuclear telomeric repeats of the collected morphospecies were all identified as (5′-CCCCAAT)n. (2) The assembled contigs that were capped with at least 1.5 telomeric repeats without any mismatch at both ends (referred to as complete chromosomes, $n = 446{,}195$) and at one end (incomplete chromosomes, $n = 794{,}710$) were extracted using a python script [41]. (3) The complete chromosomes were considered as sequences of rumen ciliates, and they displayed a low GC content (21.6% on average, ranging from 7.2% to 43.1%), with a mean GC content of 6.2% (0% to 20% among 99.8% of the complete chromosomes) in their subtelomeric regions (the 70 bp regions immediately adjacent to the telomere). The incomplete chromosomes and telomereless contigs that had a GC content higher than 44% (based on the highest GC content of the complete chromosomes) were discarded to remove potential non-ciliate contamination sequences. Compared with the complete chromosomes, more of the incomplete chromosomes had a high subtelomeric GC content, suggesting potential contamination sequences among the incomplete chromosomes. One percent of the incomplete chromosomes that had the highest subtelomeric GC content was discarded to further remove potential contamination sequences. The subtelomeric GC content of remaining incomplete chromosomes was ≤20%.

Thirdly, we constructed a sequence dataset of rumen ciliate macronuclei using the contigs assembled from the single-cell sequencing ($n = 69$, 1.7 Tb) and published rumen metagenomes ($n = 901$, ~17 Tb, Table S3) to help identify the telomereless contigs. Briefly, the publicly available rumen metagenomes were downloaded from the NCBI SRA and then assembled using Megahit. The telomere-capped contigs (both complete and incomplete chromosomes) were extracted and subjected to decontamination by removing those with a GC content higher than 44% or a subtelomeric GC content higher than 20%. The decontaminated contigs from the rumen metagenomes and those from the single-cell sequencing were combined, and the redundant contigs were removed by cd-hit-est (v4.6.6) with the parameter "-c 0.99 -n 10 -G 0 -aS 0.80" applied, resulting in an 18 Gb dataset of rumen ciliate macronuclear sequences.

Fourthly, the above ciliate macronuclear sequence dataset was combined with the genomes of rumen prokaryotes and gut fungi (Table S4) to serve as a reference for identifying the telomereless ciliate contigs resulting from single-cell sequencing using blastn (v2.6.0+). The telomereless contigs that best matched to ciliate reference sequences with a > 85% sequence identity over >500 bp and a > 50% coverage of the reference or query were considered as ciliate contigs.

Finally, the telomereless contigs picked above and the telomere-capped contigs with >3X read depth were retained in SAGs. Genome completeness of each SAG was assessed using BUSCO v5 [42] with OrthoDB v10 [43] using the predicted protein-coding genes as input, with those having ≥80% [2] of the 171 Alveolata conserved marker genes considered "high-quality" SAGs. Only the high-quality SAGs ($n = 52$, Table S5) were further analyzed.

Gene prediction

To aid gene prediction, we extracted the polyadenylated eukaryotic mRNAs of rumen fluid samples ($n = 14$) for metatranscriptomic sequencing after ciliate enrichment by filtration (150 µm) and centrifugation (500 × *g* for 5 min). The paired reads of the single-cell transcriptomes and metatranscriptomes that could map [44] to the SAGs with >95% identity and >80% coverage were assigned to each ciliate species for de novo gene prediction of SAGs (Fig. 1B). The total transcriptomic reads of each species were de novo assembled using Trinity (v2.8.4) [45]. The complete transcripts (having the untranslated region at both the 5′ and 3′ ends) that could align (>98% identity and >90% coverage) to a complete chromosome of SAGs were selected to train the de novo gene prediction

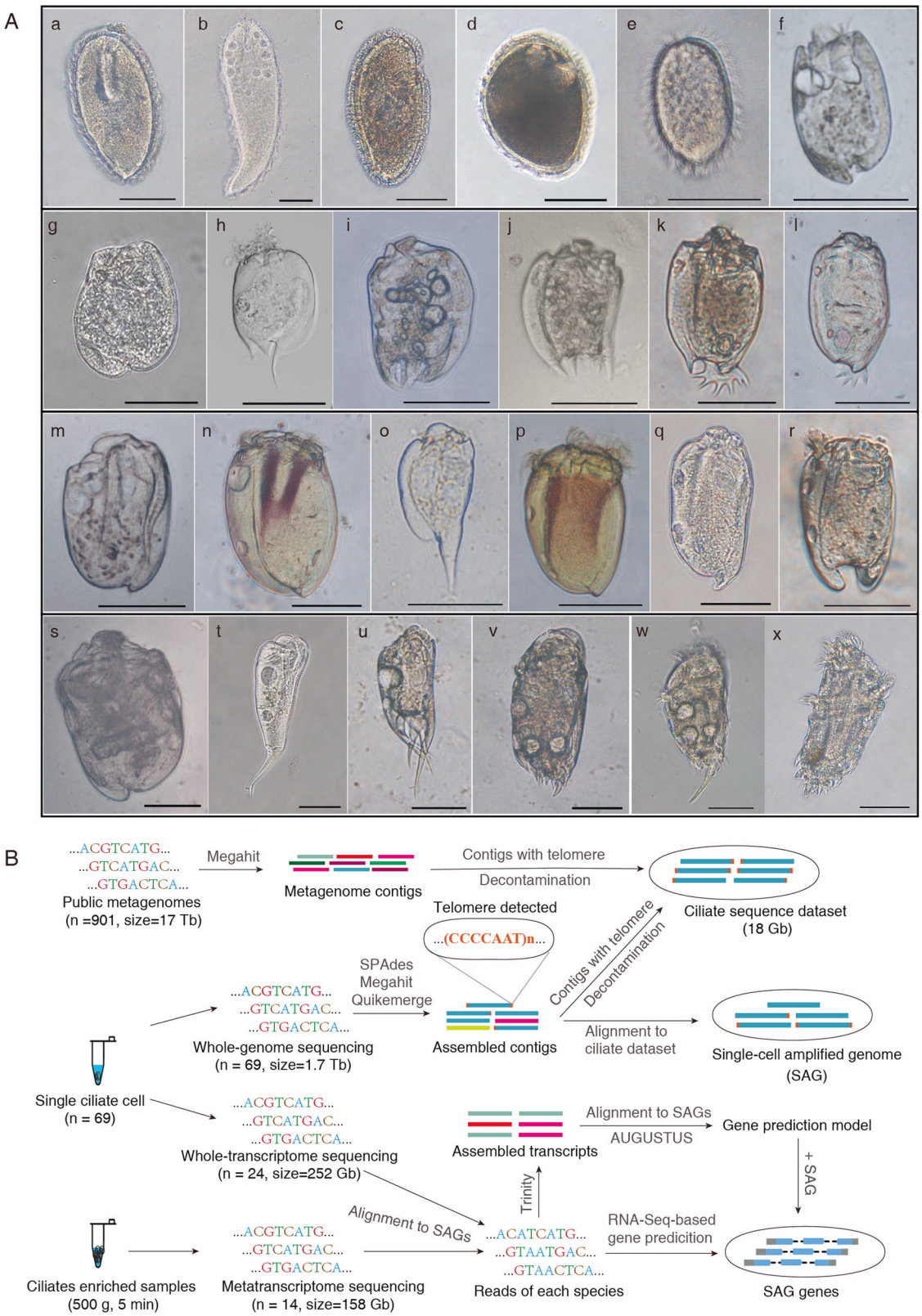


Fig. 1 Generation of genome catalog of rumen ciliates. Light micrographs of the 22 rumen ciliate morphospecies and two cryptic species examined in this study (**A**) and an assembly-and-identification pipeline for recovering single-cell genomes and gene prediction (**B**). **a** *Iso. prostoma*. **c** *Iso. intestinalis*. **e** *Das. ruminantium*. **f** *Ent. longinucleatum*. **g** *Ent. bursa*. **h** *Ent. caudatum*. **i** *Dip. anisacanthum*. **j** *Dip. dentatum*. **k** *Dip. flabellum monospinatum*. **l** *Dip. flabellum aspinatum*. **m** *Eno. triloricatum*. **n** *Met. minomm*. **o** *Ere. rostratum*. **p** *Ost. gracile*. **q** *Ost. venustum*. **r** *Ost. mammosum*. **s** *Pol. multivesiculatum*. **t** *Epi. caudatum*. **u** *Epi. cattanei*. **v** *Oph. bicinctus*. **w** *Oph. caudatus*. **x** *Oph. purkynjei*. *Iso.* sp. YL-2021a (**b**) and *Iso.* sp. YL-2021b (**d**) were two cryptic species of *Isotrichidae*. Micrographs **n** and **p** show stained skeletal plates. Scale bars, 50 μm (a-x). More details of the rumen ciliates are shown in Fig. S1.

model using AUGUSTUS (v3.2.3) [46]. The best model was that of *Polyplastron multivesiculatum* with 93.5% sensitivity and 85.7% specificity at the exon level. RNA-Seq-based gene predictions of SAGs were performed using AUGUSTUS and the best model (http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat). The protein sequences were translated from the predicted genes using the universal genetic code as previously suggested [47]. The tRNAs-coding and rRNA genes were identified using tRNAscan-SE (v2.0.5) [48] and Barrnap (v0.9, https://github.com/tseemann/barrnap), respectively.

Circumscribing species

The whole-genome average nucleotide identity (ANI) and alignment coverage between SAGs were calculated using pyani (v0.2.10) [49] with the default parameters. The circumscription between intraspecies and interspecies of rumen ciliates was set according to the gap among the ANI values [50, 51].

Phylogenomic tree construction and taxonomic rank normalization

A total of 113 single-copy homologous genes were identified among the representative genomes of the rumen ciliate species and *Tetrahymena thermophila* genomes (as an outgroup) by OrthoFinder (v2.5.2) [52]. The corresponding amino acid sequences of these single-copy genes were aligned and then concatenated into supergenes to construct a maximum likelihood (ML) phylogenetic tree using RAxML (v.8.2.9) [53] with the options of "-f a -N 1000 -m PROTGAMMAJTTF" applied. One rooted species tree was also inferred from all gene families across the 52 SAGs using OrthoFinder with the STAG and STRIDE algorithms [52].

The concatenated single-copy protein-based phylogeny served as the basis for high taxonomic rank (at and above genus) normalization using the relative evolutionary divergence (RED) values calculated with PhyloRank (v0.1.0) [54]. The RED intervals for normalizing taxonomic ranks were defined as the median RED value at each rank ± 0.1. The taxa that fell outside of their RED distribution were corrected by comparing the RED values with the RED intervals of taxonomic ranks.

Divergence time estimation and gene family expansion

Divergence times of the rumen ciliate species were estimated based on the ML tree via the Bayesian relaxed molecular clock approach using the MCMCtree program in the PAML package (v.4.9) [55]. The calibrated points of the most recent common ancestor (MRCA) of rumen ciliate species (<135 million years ago) and *Ophryoscolecidae* species (<55 million years ago) were obtained as described by Vďačný et al. [6].

The orthologous gene families among the 52 SAGs were identified using OrthoFinder [52]. To reduce the limitation of single-cell genomes, the maximum gene number in each gene family of each species was used in the CAFE (v5.0) [56] to infer the expansion and contraction of the gene families.

Functional annotation

Functional annotation (KEGG, GO, and COG) of the predicted proteins was performed using eggNOG mapper (v2.0.1) with the Diamond mapping mode, based on the eggNOG 5.0 orthology data [57]. Protein domains were annotated using pfam_scan.pl by comparing them to the Pfam_A v33.1 database [58]. Carbohydrate-active enzymes (CAZymes) were annotated using dbCAN2 [59] against the CAZyme database v9. The identified CAZymes were then categorized with their corresponding EC numbers by aligning their sequences against the Uniprot_sprot (release-2020_06) and CAZyme databases using Diamond (v0.9.21.122) with an *E*-value cutoff of e-3. To compare the numbers of CAZymes and CAZyme profiles of the rumen ciliates with those of rumen bacteria ($n = 2405$), gut fungi ($n = 8$), and non-gut *Alveolata* ciliates ($n = 13$) (Table S6), the CAZymes of the latter three groups microbes were also annotated.

Identification of CAZymes with horizontal gene transfer signature

The rumen ciliate CAZymes that were probably acquired via horizontal gene transfer (HGT) were identified by combining similarity searches with phylogenetic analysis against the NCBI nr database (2020-10-26) and one in-house microbial (including rumen prokaryotes, gut fungi, and non-gut *Alveolata*, Table S4) protein dataset following the pipeline and cut-offs of previous studies [26, 60]. To infer the likely number of times and the likely phylogenetic locations of HGT events, we extracted each type of degradative CAZymes (with the same EC number and within the same family) from rumen ciliates and the plausible donors and built a phylogenetic tree using IQ-TREE (v2.1.4-beta) [61] with the default parameters. When a horizontally transferred CAZyme was present in two or more sister branches, that CAZyme was considered to have been acquired via HGT in their MRCA.

Enzymatic activity assay of horizontally transferred cellulase and xylanase

In vitro enzymatic assays were used to verify the enzymatic activity of ciliate CAZymes that were inferred to be acquired via HGT and compare that with the corresponding CAZymes of the likely HGT donors. The horizontally transferred ciliate CAZymes that were only present in most of the species of *Diplodiniinae* and *Ophryoscolecinae* (the HGT events likely occurred in their MRCA and thus the HGT donors were likely rumen bacteria according to their divergence time [6]) and their best matched CAZymes of the likely HGT donors were selected as candidates. The candidate ciliate CAZymes were further screened according to their transcriptional expression and localization on complete chromosomes. Ultimately, one cellulase (EC 3.2.1.4 in GH5) from *Ophryoscolex caudatus* and its likely HGT donor (*Treponema bryantii* NK4A124), and one xylanase (EC 3.2.1.8 in GH10) from *Ostracodinium gracile* and its likely HGT donor (*Ruminococcus flavefaciens* YRD2003) were chosen and tested. These two pairs of CAZymes (four genes) were codon-optimized according to the codon preference of *Pichia pastoris* [62], which was used as the overexpression platform. Construction of recombinant plasmids, heterologous expression, and protein purification followed the procedures previously described [63]. The activities of the cellulase and xylanase were measured as previously described [64, 65]. Each experiment was performed with three reaction replicates to determine the mean ± standard error of the hydrolytic activities (U/mg protein) of the enzymes.

Enzymatic activity assay of lysozyme

One GH19 (it includes lysozyme and chitinase) protein, which could not be annotated to any EC numbers, of *Oph. caudatus* was chosen and tested for its substrate and activity also using the overexpression platform of *P. pastoris* as described above. The enzymatic substrate assay showed that it was a lysozyme. Its enzymatic activity was measured as previously described [66]. The antibacterial ability of this lysozyme was tested by inhibition of *Micrococcus luteus* (ATCC 4698) and *Escherichia coli* (CVCC 3367), as a representative of Gram-negative and -positive bacteria, respectively, using zone assay on agar plates [67].

Taxonomic classification of metagenomic sequences

The sequencing reads of 901 publicly available rumen metagenomes (Table S3) were taxonomically classified using Kraken (v2.0.7-beta) [68] against three in-house databases: (1) a public dataset containing microbial genomes ($n = 166{,}583$) from RefSeq (release 201) plus the genomes of rumen prokaryotes ($n = 7034$) and gut fungi ($n = 8$) (Table S4); (2) the public dataset plus the 52 SAGs; (3) the public dataset plus the rumen ciliate dataset (including the 52 SAGs and the telomere-capped contigs from the 901 metagenomes).

Statistical analysis

The data of two groups were statistically compared by the two-tailed Wilcoxon test unless otherwise noted. Bonferroni-corrected p-value < 0.05 was set as the significance threshold. The standard errors were provided when means were compared. Principal coordinate analysis (PCoA) based on Bray–Curtis dissimilarity was done using the vegan package (https://github.com/vegandevs/vegan/) of R (v3.6.2, http://www.R-project.org). Permutational multivariate analysis of variance was performed using the Adonis function in the vegan package of R to compare the statistical divergence across families/subfamilies of ciliates.

RESULTS AND DISCUSSION

A single-cell amplified genome catalog of rumen ciliates

By integrating single-cell sequencing and an assembly-and-identification pipeline, we constructed the first genome catalog of rumen ciliates. The 69 collected single ciliate cells represented 22 morphospecies (Figs. 1 and S1–3) across 11 prevalent and

abundant morphogenera, which represent ~97% of the genus-level abundance across our and previous surveys of rumen ciliates using microscopic enumeration (Table S2). Of the 69 SAGs we assembled (Table S5), 52 were considered "high-quality" with 80–91% genome completeness, 10 were considered "medium-quality" with 52–78% genome completeness, while the remaining 7 were considered "low-quality" with 8–47% genome completeness. Among the 52 high-quality SAGs, the telomere-capped contigs accounted for 14% to 92% (mean 61%) of the total contigs (Table S5). The highest number of complete chromosomes among the SAGs was found in the SAGs of *Oph. caudatus* (referred to as *Ophryoscolex caudatus* SAGT3, $n = 26{,}497$), which also had 33,578 incomplete chromosomes and 14,581 telomereless contigs (Table S5). Therefore, the number of chromosomes of *Oph. caudatus* was at least 43,286 (assuming two incomplete chromosomes would represent one complete chromosome), which was much higher than the known number of chromosomes in free-living ciliates, such as *Oxytricha trifallax* (16,000) and *Halteria grandinella* (23,000) [29, 69]. Compared with previous studies on fungi and marine stramenopiles [70, 71], the present study was more successful in recovering the ciliate genomes from single-cell sequencing probably because ciliates typically have mini-/nano-chromosomes and a high ploidy [29, 69]. Metagenomic binning can help recover genomes of uncultured microbes [72], but our results showed that it would be not applicable to recover rumen ciliate genomes because the copy numbers of their chromosomes are not uniform (varying by multiple orders of magnitude) within single genomes (Fig. S5) and binning depends on different sequencing depths among inter-genomes and only the contigs with similar sequencing depth could be binned into one bin [72].

A robust genome-based taxonomy and phylogeny

A robust taxonomy is essential to communicate scientific results and describe biodiversity [51, 54]. Using the 52 high-quality SAGs, we established a robust genome-based taxonomic framework of rumen ciliates that overcomes the uncertainty inherent in the morphology-based taxonomy.

Whole-genome ANI has been widely used to quantitatively circumscribe species of prokaryotes [50, 51]. We found a clear ANI discontinuity among the 52 SAGs: >97% or <92% (Fig. 2a, b), suggesting that the ANI was also suitable for circumscribing species of rumen ciliates. ANI > 97% and <92% were considered as intraspecies and interspecies boundaries for rumen ciliates, respectively. These boundaries of ciliates are much clearer and greater than those of prokaryotes (with ANI > 95% and <95% as intraspecies and interspecies boundaries, respectively) [50, 51], suggesting that the intraspecific genetic conservation and interspecific genetic divergence of ciliates are greater than those of prokaryotes. Based on the above ANI cutoff values, the 22 collected morphospecies were reclassified to 19 species, with four groups of synonymic species in *Ophryoscolecidae* and two cryptic species of *Isotrichidae* being identified (Figs. 2a and S6). The synonymic *Ophryoscolecidae* species (ANI > 97%, Fig. 2a) have different morphologies (Figs. 1A and S1). For example, morphospecies *Oph. caudatus* has a long main caudal spine and three circles of secondary caudal spines (Fig. S1), whereas its synonymic morphospecies *Ophryoscolex purkynjei* has a short main caudal spine and *Ophryoscolex bicinctus* has two circles of secondary caudal spines (Fig. S3). Similarly, *Diplodinum anisacanthum* (main caudal spines being incurved) and *Dip. dentaum* (main caudal spines being direct) were synonymic species, and so were *Diplodinum flabellum aspinatum* (no dorsal spine) and *Diplodinum flabellum monospinatum*, and *Ost. gracile* (no small ventral lobe) and *Ostracodinium venustum* (Figs. 2a and S1). These results suggest that the shape of the main caudal spine and the number of secondary caudal spines are not reliable morphological features for the classification of species of *Ophryoscolecidae*. On the other hand, despite being morphologically similar to other morphospecies, two new cryptic species of *Isotrichidae* were identified (ANI < 92%, Fig. 2a): tentatively named as *Isotricha* sp. YL-2021a and *Isotricha* sp. YL-2021b. The former was a cryptic species (Fig. 1A, b) of *Isotricha prostoma* and had a long body (up to 278 μm), while the latter was a cryptic species (Fig. 1A–d) of *Iso. intestinalis* and had the largest mean length of chromosomes (22.5 kb, with a range from 0.6 to 426 kb) among the 19 ciliate species. Additionally, a discontinuity of the 18 S rRNA gene sequence identity was noted among the SAGs, and 98% identify was considered as species boundary for rumen ciliates (Fig. S6). The 18 S rRNA gene sequence identity among the synonymic morphospecies was all >99%, while that among the four *Isotrichidae* species was all <98% (Fig. S6), which supports the genome-based taxonomic revision at the species level.

The current taxonomy for rumen ciliates at or above the genus level also lacks molecular evidence, as evidenced by the 18 S rRNA gene-based phylogeny placing the subfamilies *Entodiniinae* and *Ophryoscolecinae* as two separate lineages within the subfamily *Diplodiniinae* [6, 23]. Using the SAGs, we firstly resolved the phylogenetic disarray by constructing an ML tree using 113 concatenated single-copy proteins (Fig. 2d) and a rooted tree inferred with all gene families (Fig. S7). The two trees had similar topologies and 100% bootstrap support for each node. Both phylogenetic inferences show *Entodiniinae* representing the oldest branch in *Ophryoscolecidae* and *Ophryoscolecinae* as a sister to *Diplodiniinae*, which is consistent with previous morphological inference [73]. The concatenated single-copy protein-based phylogeny served as the basis for the taxonomic rank normalization using RED. Comparing the RED values with the RED intervals of taxonomic ranks (median RED ± 0.1), a new family and two new genera were proposed (Fig. 2c): The genus *Dasytricha* was reassigned from *Isotrichidae* to a new family *Dasytrichidae*, which was treated as a synonymy of *Isotrichidae* in the past [5]. Additionally, *Entodinium bursa* and *Ent. caudatum* were reassigned from *Entodinium* to a new genus *Entodinium_A*, and *Iso. prostoma* and *Iso.* sp. YL-2021a were reassigned from *Isotricha* to a new genus *Isotricha_A* (Fig. 2d). On the other hand, none of the genera in the *Ophryoscolecinae* and *Diplodiniinae* was reassigned, suggesting that the size and number of skeletal plates are reliable morphological features for the classification of genera. Ultimately, the 19 identified ciliate species were assigned into two orders, three families, and 13 genera (Fig. 2d).

Based on both ANI and RED, we substantially revised the taxonomy of rumen ciliates by reclassifying 10 of the 22 collected morphospecies to a new species, genus, or family. The genome-based taxonomy could corroborate or reject the taxonomic value of some morphological features of rumen ciliates, which will help to revise the morphology-based taxonomy and to identify new species. Additionally, we provided a robust phylogenetic backbone that could help understand the evolutionary processes of specific traits in rumen ciliates. For example, according to the parsimony principle of evolution [74] and the phylogenetic relationships (Fig. 2d), the skeletal plate, which is a newly evolved organelle storing amylopectin in the clades of *Ophryoscolecinae* and *Diplodiniinae* [7], probably arose from a single origin and was lost in *Diplodinium*.

Genome architecture and gene family diversity

Invasion into the rumen was one of the most important events in the evolutionary history of *Trichostome* ciliates [6]. As shown above, not only rumen *Ophryoscolecidae*, as previously suggested [6], but also *Isotrichidae* have radiated in speciation. Molecular clock analyses showed that explosive radiation of both *Ophryoscolecidae* and *Isotrichidae* occurred about 5–35 million years ago (Fig. S8), which is consistent with the duration of the rapid radiation of ruminants [75], suggesting concurrent co-evolution and co-speciation between ruminants and rumen ciliates. Explosive speciation can result from a series of genomic changes

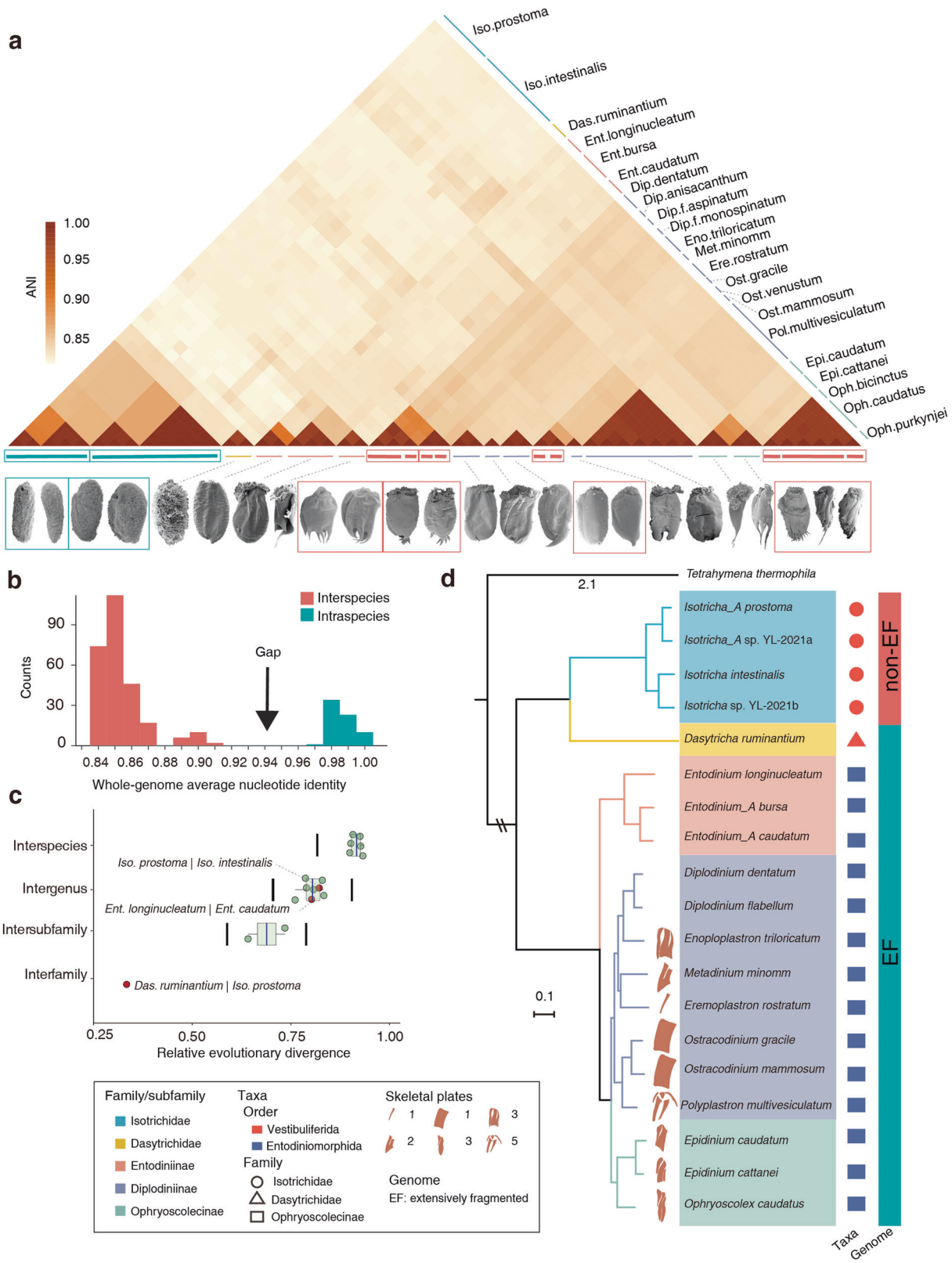


Fig. 2 Genome-based taxonomy and phylogeny of rumen ciliates. a A heatmap showing the average nucleotide identity (ANI) among the 52 SAGs. The scanning electron microscopy images of 22 morphospecies and two cryptic species are shown. Synonymic morphospecies are wrapped in red boxes. Morphospecies and its cryptic species are wrapped in blue boxes. **b** ANI distributions among the 52 SAGs. **c** Rank normalization through relative evolutionary divergence (RED). The RED interval for each rank is shown by two vertical black lines, median ± 0.1. The reassigned taxa are indicated in red dot. **d** The maximum likelihood phylogenetic tree of 19 rumen ciliate species and *T. thermophila* (as the outgroup) based on 113 concatenated single-copy proteins. All nodes have a 100% bootstrap support. The size and number of skeletal plates and the extensively fragmentation (EF) or non-EF of genome are labeled for ciliate species.

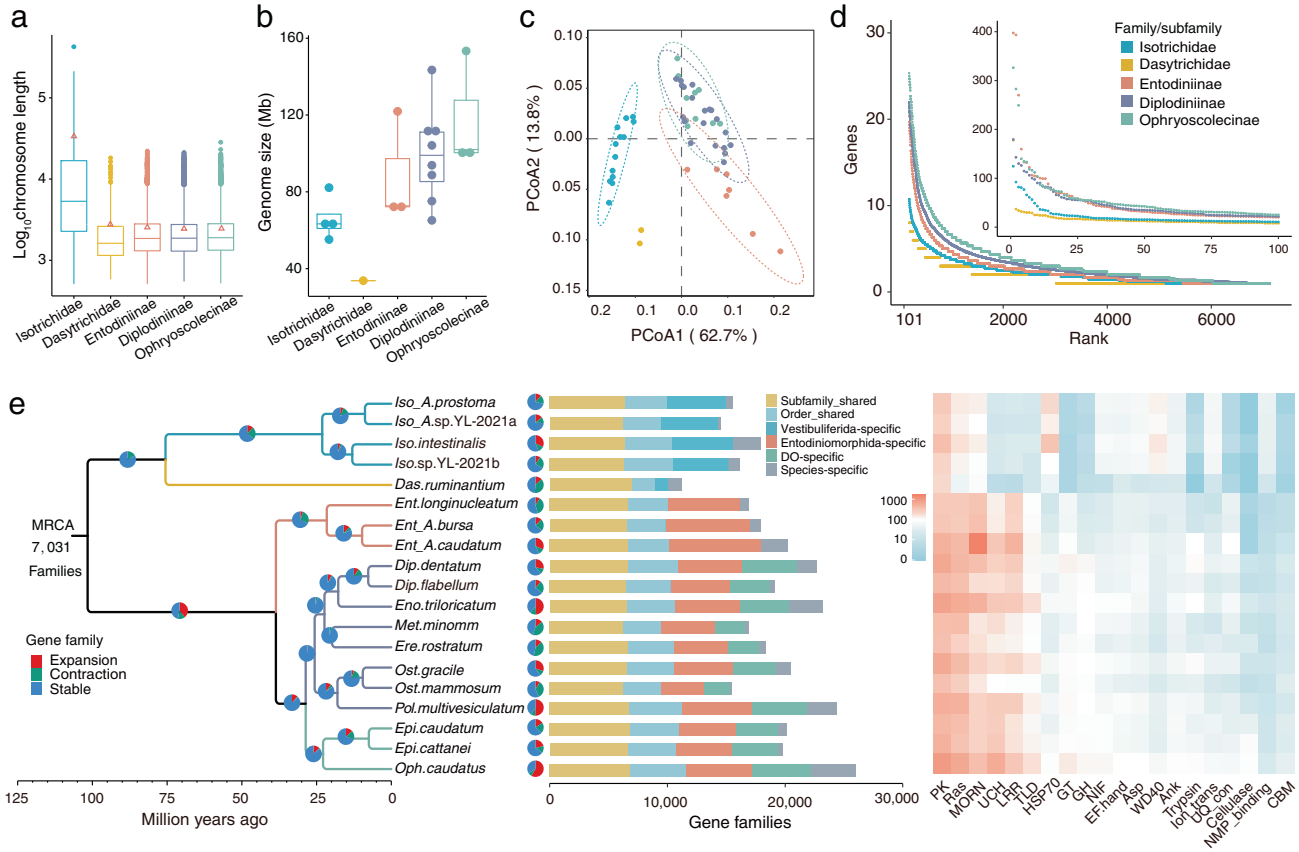


Fig. 3 Genome characteristics of rumen ciliates. a Chromosome length (the red triangles indicate N50) and **b** Genome size distributions across families/subfamilies. **c** Functional divergence based on the clans of Pfam (*p*-value of permutational multivariate analysis of variance <0.001). **d** Rank curves showing gene number divergence in the families/subfamilies-shared gene families. **e** Gene family divergence: From left to right: a time phylogenetic tree of rumen ciliates with gene family expansions and contractions; numbers of gene families in each of the 19 species shared within or specific to a taxon; a heat map showing the top 10 categories of domains found in order-specific gene families of each species. MRCA most recent common ancestor, DO *Diplodiniinae* and *Ophryoscolecinae*.

and innovations in architecture and function [76, 77], and the genome catalog and phylogenetic relationships of the rumen ciliates provided new insights into their early diversification.

Two notably different macronuclear chromosome architectures exist among classes of free-living ciliates in the environment: "long" versus "nano" sized chromosomes [41, 78, 79]. To our knowledge, we discovered for the first time that these two chromosome architectures of macronuclear genomes could exist in a single ciliate subclass (Figs. 2d and 3a): Both *Ophryoscolecidae* and *Dasytrichidae* had an extensively fragmented (EF) macronuclear genome with most of its chromosomes being gene-sized nanochromosomes with a mean length of 2.3 kb, while *Isotrichidae* had a non-EF macronuclear genome with most of its chromosomes each encoding several genes with a mean length of 14.3 kb (Fig. S9). Compared with the previous inter-class investigations [78, 79], the genomes of rumen ciliates will provide a more allied model for investigating the unsolved molecular mechanisms underlying chromosome fragmentation. Like *Dasytrichidae* and *Ophryoscolecidae*, *Spirotrichea* and *Haptoria*, which are a sister class and a sister subclass of rumen ciliates, respectively, both had EF genomes [41]. Thus, based on the parsimony principle of evolution [74], the genome of the MRCA of rumen ciliates was inferred as EF genome, and the non-EF genome of *Isotrichidae* was inferred as an independent origin. Similarly, some of the genomic architecture features of *Isotrichidae* were obviously divergent from those of *Dasytrichidae* and *Ophryoscolecidae* (Table S5), despite *Dasytrichidae* and *Isotrichidae* being in the same order. For example, *Isotrichidae* had a lower GC content (18.7 ± 0.3% vs. 21.9 ± 0.3%), longer gene-coding regions (1783 ± 55 bp vs. 1642 ± 26 bp), and longer gene introns (118 ± 2 bp vs. 71 ± 2 bp) than *Dasytrichidae* and *Ophryoscolecidae* (Figs. S9, 10). These results suggest that *Isotrichidae* has likely gone through a series of independent evolutionary processes in forming their genomic architecture early and during the diversification of *Isotrichidae* into a modern family.

We annotated 64,189 gene families from the 52 SAGs (Figs. 3e and S10), of which, 40,898 were shared by at least two ciliate species, and the remaining 23,291 were species-specific (ranging from 110 in *Ost. mammosum* to 3818 in *Oph. caudatus*). Of the species-shared gene families, 14,871 (36.4%) were found in both *Vestibuliferida* and *Entodiniomorphida*, while the remaining gene families were specific to *Vestibuliferida* (6091) or *Entodiniomorphida* (19,936). Of the *Entodiniomorphida*-specific gene families, 10,043 arose at the clades of *Diplodiniinae* and *Ophryoscolecinae*. This suggests that extensive lineage-specific gene families arose early and during the diversification of rumen ciliates into modern orders. Additionally, among the 7031 gene families that were shared by five ciliate families/subfamilies (Fig. S10), more gene families expanded in *Entodiniomorphida* (2784) than in *Vestibuliferida* (53) (Fig. 3e). As a result, species in *Entodiniomorphida*, in particular those of *Diplodiniinae* and *Ophryoscolecinae*, had larger and more diverse gene families, and larger genome sizes than the species in *Vestibuliferida* (Fig. 3b, d), which might have further induced the functional divergence across orders and families (Fig. 3c). Comparing the differential investment in KEGG pathways and Gene Ontology terms across those clades (Figs. S11, 12), we found that the members of

the *Isotrichidae* possessed a higher proportion of genes related to organismal systems and environmental adaptation (e.g., sensory, immune, detoxification, and antioxidant activity), while the genes of *Ophryoscolecidae* were primarily related to degradation of substrates (e.g., carbohydrates, lipids, and xenobiotics). Indeed, a comparison of the domains of order-specific genes between *Isotrichidae* and *Ophryoscolecidae* showed that *Isotrichidae* had a higher predominance of HSP70 and WD40, which are related to stress resistance [80] and biogenesis of motile cilia [81], respectively, while *Ophryoscolecidae* had a higher predominance of CAZymes-related domains (GT, GH, CBM and cellulase) (Fig. 3e). Additionally, the species-shared and the lineage-specific proteins identified by the SAGs may be explored to control rumen ciliates as a potential approach to improve dietary nitrogen utilization and lower methane emissions by ruminants as demonstrated by others [13, 24, 82].

A broad array of CAZymes encoded by the rumen ciliate SAGs

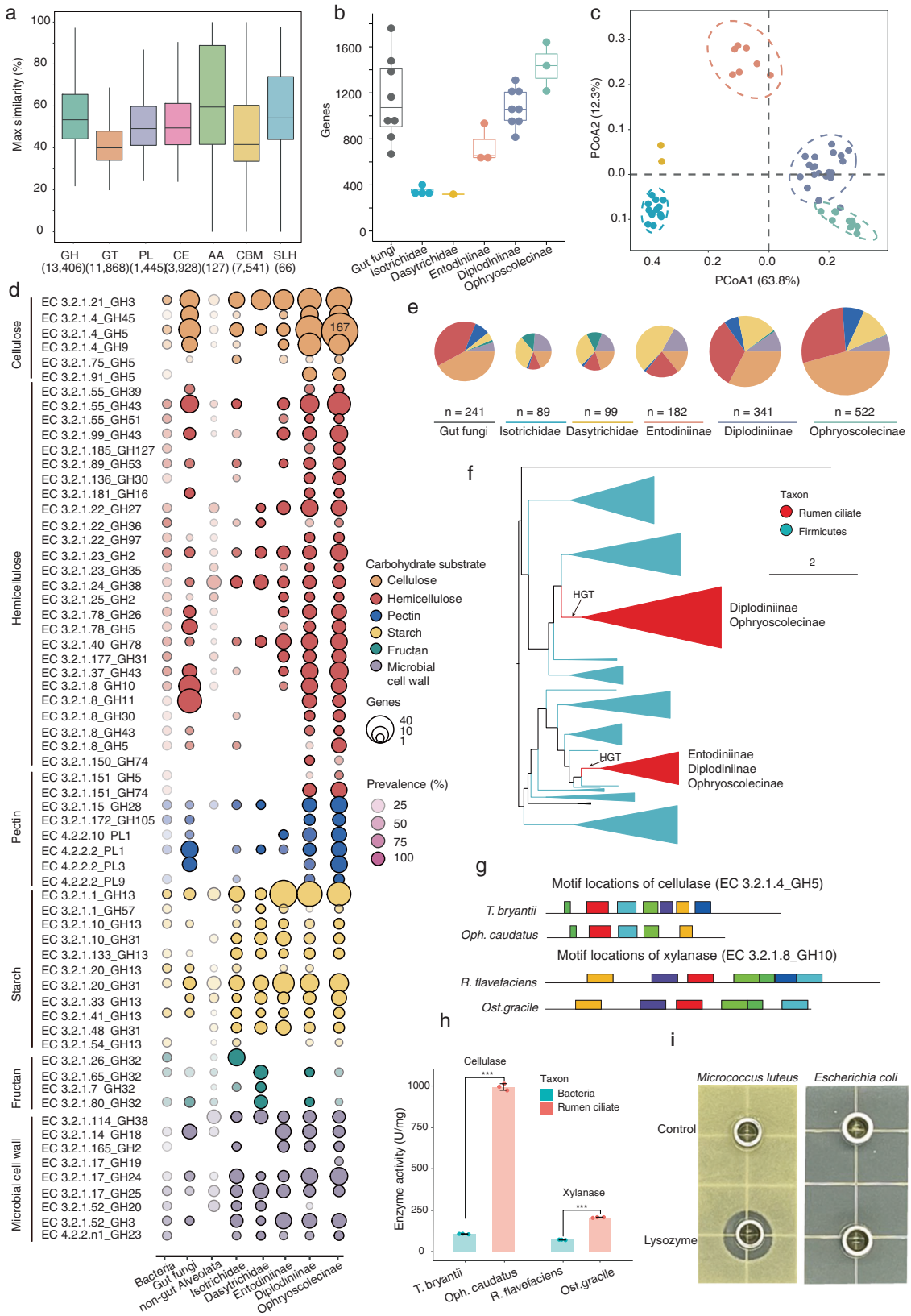
The gut microbes are of huge industrial interest due to their ability to release fermentable sugars from lignocellulose [72, 83, 84]. Here, we found that the rumen ciliates were also a rich source of biomass-degrading enzymes. A total of 43,511 CAZyme proteins were predicted in the 52 SAGs, and they were clustered into 33,693, 25,185, and 21,468 clusters by cd-hit at 99%, 95%, and 90% identity, respectively. Of the 33,693 clustered CAZymes, only 357 were highly similar (≥95% amino acid sequence identity) with any of the proteins in public databases (i.e., nr, env_nr, m5nr, and UniProt TrEMBL) and the rumen bacterial and gut fungal protein datasets (Fig. 4a), indicating that 99% of the ciliate CAZymes can be considered new. In particular, the species of *Diplodiniinae* and *Ophryoscolecinae* encoded as many CAZymes as gut fungi, the latter of which possesses the largest number of CAZyme genes yet known in nature [83]. And the species of *Diplodiniinae* and *Ophryoscolecinae* encoded more CAZymes than *Entodiniinae* (by two-fold) and *Isotrichidae* and *Dasytrichidae* (by five-fold) (Fig. 4b) with clear profile-divergences (Fig. 4c), largely attributed to the extensive gene family arising and expansion as shown above.

We predicted the substrates of the glycoside hydrolases (GH) and polysaccharide lyases (PL) encoded by the 52 SAGs based on their corresponding EC number (EC 3.2.1.- or EC 4.2.2.-) (Table S6). We found that each species of the rumen ciliates possessed a broad array of GH and PL enzymes, which could allow for the degradation of plant cell wall (structural carbohydrates, including cellulose, hemicellulose, and pectin), plant storage carbohydrates (fructan and starch) and microbial cell wall (chitin and peptidoglycan) (Fig. 4d and Table S6). Moreover, the SAGs, in particular the SAGs of *Diplodiniinae* and *Ophryoscolecinae*, encoded multiple enzymes to synergistically degrade a single kind of polysaccharides and the resulting oligosaccharides. For example, the species in *Ophryoscolecidae* possessed endo-β-1,4-glucanase (EC 3.2.1.4), β-glucosidase (EC 3.2.1.21), endo-1,6-β-glucosidase (EC 3.2.1.75), and β-1,4-cellobiosidase (EC 3.2.1.91), allowing them to synergistically degrade cellulose. These results corroborate the previous reports that rumen ciliates contribute to fiber degradation with their own diverse CAZymes [12, 13, 85]. These results also advance our understanding of the trophic niche breadth of rumen ciliates at the species level beyond what has been gained from previous metatranscriptomic and biochemical studies [12, 13, 85]. For example, *Dasytricha ruminantium* was thought to be non-cellulolytic [86], but we identified two types of cellulase-encoding genes (EC 3.2.1.4 in GH5 and EC 3.2.1.21 in GH3) in its genome. Future research is needed to verify if these cellulase genes encode functional enzymes.

Variable investment tradeoffs in degradative CAZymes were found across rumen ciliate families/subfamilies (Fig. 4d, e). Indeed, the majority of the degradative CAZymes of *Diplodiniinae* (72%) and *Ophryoscolecinae* (82%) was invested in degrading plant cell wall, followed by CAZymes degrading plant storage carbohydrates (19% and 12%, respectively) and microbial cell wall (9% and 6%, respectively), while that of *Isotrichidae* (45%), *Dasytrichidae* (43%), and *Entodiniinae* (46%) were mainly invested in degrading plant storage carbohydrates (Fig. 4e), followed by CAZymes degrading plant cell wall (31%, 37%, and 37%, respectively) and microbial cell wall (24%, 19%, and 17%, respectively). The investment portfolios of degradative CAZymes corroborate that rumen ciliates acquire their energy mainly from the use of plant materials. The investment portfolios of degradative CAZymes of *Diplodiniinae* and *Ophryoscolecinae* were most similar to that of gut fungi, with 89% of their degradative CAZymes acting on plant cell wall (Fig. 4e). The investment tradeoffs also varied across families/subfamilies in the CAZymes degrading each kind of polysaccharides. For example, *Isotrichidae* and *Dasytrichidae* invested more heavily (13 ± 1.8%) in CAZymes degrading fructan than *Ophryoscolecidae* (0.4 ± 0.1%, Fig. 4e). These observations were consistent with the substrate preferences determined experimentally among rumen ciliates [7] but provided much more detailed and quantified information. The variable investment tradeoffs in degradative CAZymes could not only help avoid an overlapping niche across families/subfamilies but also enable robust degradation of plant materials [87] and facilitate their coexistence in the rumen [88]. Taken together, the broad array of CAZymes, variable substrate preferences, and herbivory in rumen ciliates revealed by the SAGs may contribute to their ecological success in the rumen ecosystem.

Compared with rumen ciliates, non-gut *Alveolata* species ($n = 13$) possess a much smaller number and lower diversity of genes encoding degradative CAZymes in their genomes, and they are incapable of digesting xylan-related hemicellulose (xylan, xyloglucan, and arabinoxylan) or pectin (Fig. 4d and Table S6). These results suggest that most of the degradative CAZyme-coding genes of rumen ciliates were not likely obtained by vertical inheritance. Previous studies [9, 13, 26] suggest that rumen ciliates might have acquired many of their CAZyme-coding genes through inter-kingdom HGT. This is in line with our findings that about 63% of the rumen ciliate CAZyme-coding genes (Fig. S13) was likely acquired via HGT (based on high amino acid sequence similarity and phylogenetic nesting), with 55% of them being from bacteria (mostly *Firmicutes* and *Bacteroidota*) and 8% from fungi. Compared with those of *Entodiniomorphida*, more of the CAZyme-coding genes of *Vestibuliferida* (7 ± 0.4% vs.12 ± 1.9%) were likely acquired via HGT from fungi (Fig. S13). The extensive inter-kingdom transferred CAZyme-coding genes in the rumen ciliate genomes were likely attributed to the typically frequent HGT in phagotrophic, unicellular eukaryotes [89], and subsequent the gene expansion or inter-ciliate transfer. Most types of the ciliate degradative CAZymes (with the same EC number and within the same family) might have more than one HGT event, which could occur in the early or during the diversification of rumen ciliates (Figs. 4f and S14). For example, the rumen ciliates might have recruited five GH families (GH5, GH10, GH11, GH30, and GH43) of xylanases (EC 3.2.1.8) from bacteria and fungi via at least 26 HGT events (15 from *Firmicutes*, 6 from *Bacteroidota*, 4 from other bacteria, and 1 from fungi), with 15 HGT events being species-specific, 1 each *Ophryoscolecinae*- and *Ophryoscolecidae*-specific, 7 shared by *Diplodiniinae* and *Ophryoscolecinae*, and 2 shared by *Entodiniomorphida and Vestibuliferida*.

Most of the CAZyme-coding genes that were likely acquired via HGT were structurally different from those of their probable donors. For instance, compared with the respective CAZymes of their likely donors, the rumen ciliate EC 3.2.1.8 (GH43) lacked the CBM22 domain, and the rumen ciliate EC 3.2.1.4 (GH5) and EC 3.2.1.8 (GH10) each lacked one motif, while the rumen ciliate EC 3.2.1.91 (GH5) gained one motif (Figs. 4g and S15). Moreover, the in vitro enzymatic assays (after cloning and overexpression in *P. pastoris*, see methods) showed that the horizontally transferred ciliate cellulase (EC 3.2.1.4 in GH5) and xylanase (EC 3.2.1.8 in GH10) were functional and had nine- and two-fold higher activities, respectively, than those of the



predicted bacterial donors (Fig. 4h). Additionally, about 25% of the degradative CAZymes could not be annotated to any EC numbers. To verify if they are potentially functional CAZymes, we evaluated the substrate and activity of one GH19 protein of *Oph. caudatus* also after cloning and overexpression in *P. pastoris*. This GH protein was a lysozyme (with a specific activity of 36,007 U/mg) and it inhibited the growth of *M. luteus* (Gram-positive) but not *E. coli* (Gram-negative) (Fig. 4i). Taken together, all the three ciliate CAZymes were

Fig. 4 CAZyme profiles of rumen ciliates. a Maximum amino acid sequence similarities of the CAZymes identified in the rumen ciliate SAGs compared to the public databases for seven classes of CAZyme proteins: GH glycoside hydrolase, GT glycosyl transferase, PL polysaccharide lyases, CE carbohydrate esterases, AA auxiliary activities, CBM carbohydrate-binding modules, SLH S-layer homology modules. The numeric values in parentheses refer to the numbers of genes of each class. **b** Number of CAZyme-coding genes in the representative genomes of rumen ciliate families/subfamilies and gut fungi. **c** A PCoA plot showing the CAZyme profiles across the five ciliate families/subfamilies (the color schemes are the same as in **b**), and the *p*-value of permutational multivariate analysis of variance is <0.001. **d**, **e** The mean number and prevalence of degradative CAZymes (with the same EC number and within the same family) encoded in the genomes of rumen bacteria ($n = 2405$), gut fungi ($n = 8$), non-gut *Alveolata* ($n = 13$), and the five families/subfamilies of rumen ciliates ($n = 19$). The mean number of each type of degradative CAZymes of rumen bacteria was set as 1 in **d**. More details of degradative CAZymes encoded in each genome are presented in Table S6. **f** An example for the xylanase (EC 3.2.1.8 in GH10) of rumen ciliates likely acquired via horizontal gene transfer (HGT) from rumen bacteria. **g**, **h** Structural and activity divergences of cellulase and xylanase between rumen ciliates and its likely bacterial donors. *** represents a statistical significance of *p*-value <0.001. **i** Inhibition zone assays of one ciliate lysozyme in GH19.

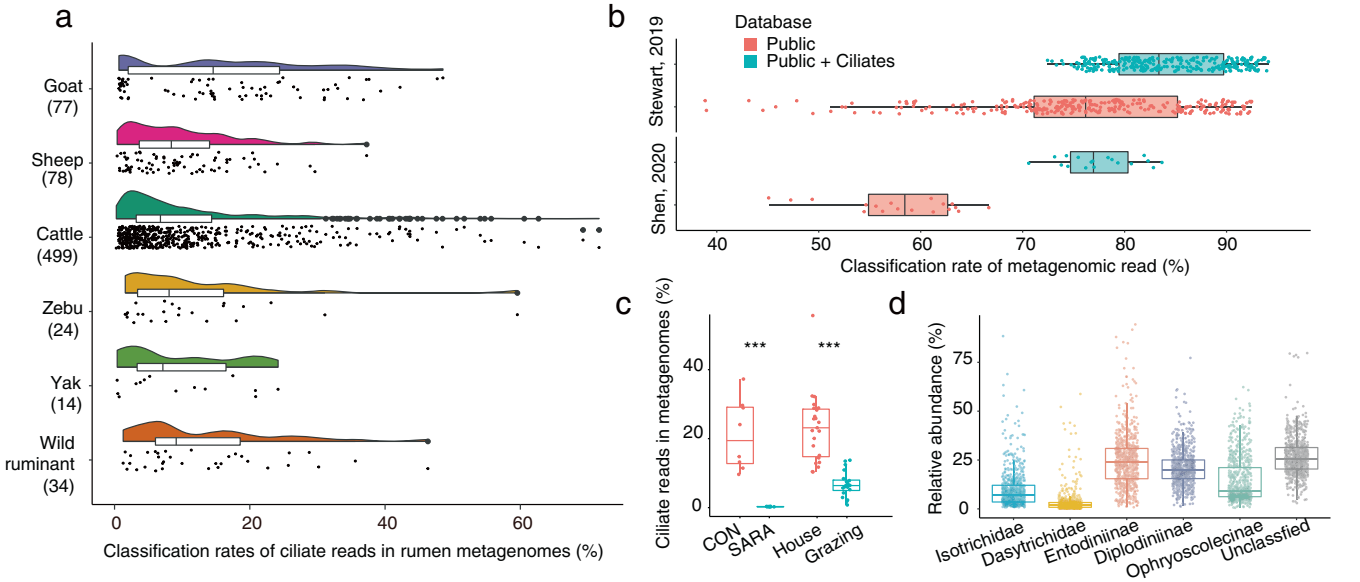


Fig. 5 Classification rates of ciliate reads in rumen metagenomes. a The classification rates of ciliate reads in rumen metagenomes across different hosts fed a diet with <80% concentrate (726 of the 901 metagenomes). **b** Improvement of classification rates of total metagenomic reads in two previous studies [2, 95] enabled by the ciliate dataset. **c** Classification rates of ciliate sequences in the rumen metagenomes of sheep suffering from subacute rumen acidosis (SARA, $n = 8$) or control sheep without SARA (CON, $n = 8$), and dairy cows in confinement (house, $n = 23$) or on pasture (grazing, $n = 20$). The metagenomes were from two previous studies [91, 93]. **d** Relative abundance of rumen ciliates as represented by the percentage of families/subfamilies sequences in the total ciliate sequences.

experimentally verified to be highly active. The rich repertoire of CAZymes of the rumen ciliates may be a rich source of industrial enzymes for biomass conversion and rumen microbiota modulation.

A high-resolution description of rumen metagenomes

The ciliate reads were dark matters in the previous rumen metagenomic analyses because of the lack of ciliate genome sequences [2, 90]. Our new ciliate dataset (including the 52 SAGs and the telomere-capped contigs from public metagenomes) could serve as a reference for the classification of metagenomic ciliate reads and facilitating holistic studies of rumen microbiomes. We re-analyzed 901 publicly available rumen metagenomes (Table S3) against in-house datasets without or with the new ciliate dataset (see methods). Our ciliate dataset on average allowed 12% of the total metagenomic reads to be classified as sequences of rumen ciliates, with a wide range from 0% to 72% among the metagenomes (Fig. 5a). With the use of the ciliate dataset, the classification rates of total metagenomic reads became higher and more convergent across the metagenomes than otherwise. For example, the ciliate dataset increased the total read classification rates of the 285 rumen metagenomes of Stewart et al. [2, 84] from 39–92% to 72–94%, and allowed one-fold more (73% vs. 33%) of the metagenomes reached 80% or higher classification rates (Fig. 5b). We found that the percentage of the ciliate reads in rumen metagenomes differed among hosts, diets, and managements (Fig. 5a, c). Of note, when sheep ($n = 8$) suffered from subacute ruminal acidosis after ingesting a high concentrate diet [91], the percentage of ciliate reads in the rumen metagenomes decreased by 99%, which is consistent with previous experimental findings that ciliates are sensitive to pH [92]. Compared with dairy cows on pasture ($n = 20$), dairy cows in confinement ($n = 23$) [93] had a higher percentage of ciliate reads in their rumen metagenomes (Fig. 5c).

Of the identified ciliate reads in the rumen metagenomes, more than 73% could be classified based on the 52 SAGs, suggesting that the collected ciliates represent the majority of the rumen ciliates community (Fig. 5d). However, as shown by 18 S rRNA gene-based surveys [13], genome-based surveys of rumen ciliates yielded different community structures from microscopic enumeration, which is considered the gold standard. For example, *Entodiniinae* as the most dominant rumen ciliates can account for >90% of the total rumen ciliates in domesticated ruminants (Table S2), but only about 25% of the identified ciliate sequence reads of the rumen metagenomes were assigned to *Entodiniinae* (Fig. 5d). This discrepancy might be attributable to (1) the lack of the genome sequences of some abundant *Entodiniinae* species, such as *Entodinium dilobum* and *Entodinium rostratum* and (2) the genomic ploidies of *Entodiniinae* are likely

lower than those of other rumen ciliates [13]. Since microscopic enumeration of ciliates is laborious and their morphological identification requires a high level of experience [13], more research is needed to identify ciliate images by machine-learning techniques [94] and to establish the correlation between DNA-based analyses and microscopic counting.

DATA AVAILABILITY

All sequencing data and the ciliate genomes assembled in this study have been deposited in the NCBI database with the accession ID PRJNA777442. The GenBank accession ID of the five overexpressed CAZymes are ON513421-ON513423, SEL14292.1, and WP_022932480.1.

REFERENCES

1. Seshadri R, Leahy SC, Attwood GT, Teh KH, Lambie SC, Cookson AL, et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol. 2018;36:359–67.
2. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37:953–61.
3. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2020;39:105–14.
4. Shabat SKB, Sasson G, Doron-Faigenboim A, Durman T, Yaacoby S, Berg Miller ME, et al. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. ISME J. 2016;10:2958–72.
5. Lynn DH. The Ciliated Protozoa: Characterization, Classification, and Guide to the Literature. 3rd ed. Heidelberg: Springer; 2008.
6. Vďačný P. Evolutionary associations of endosymbiotic ciliates shed light on the timing of the marsupial-placental split. Mol Biol Evol. 2018;35:1757–69.
7. Williams AG. Coleman GS. The Rumen Protozoa. 1st ed. New York: Springer-Verlag; 1992.
8. Solomon R, Wein T, Levy B, Eshed S, Dror R, Reiss V, et al. Protozoa populations are ecosystem engineers that shape prokaryotic community structure and function of the rumen microbial ecosystem. ISME J. 2022;16:1187–97.
9. Park T, Wijeratne S, Meulia T, Firkins JL, Yu Z. The macronuclear genome of anaerobic ciliate *Entodinium caudatum* reveals its biological features adapted to the distinct rumen environment. Genomics. 2021;113:1416–27.
10. Söllinger A, Tveit AT, Poulsen M, Noel SJ, Bengtsson M, Bernhardt J, et al. Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. mSystems. 2018;3:e00038–18.
11. Terry SA, Badhan A, Wang Y, Chaves AV, McAllister TA. Fibre digestion by rumen microbiota-a review of recent metagenomic and metatranscriptomic studies. Can J Anim Sci. 2019;99:678–92.
12. Qi M, Wang P, O'Toole N, Barboza PS, Ungerfeld E, Leigh MB, et al. Snapshot of the eukaryotic gene expression in muskoxen rumen-a metatranscriptomic approach. PLoS ONE. 2011;6:e20521.
13. Newbold CJ, de la Fuente G, Belanche A, Ramos-Morales E, McEwan NR. The role of ciliate protozoa in the rumen. Front Microbiol. 2015;6:1313.
14. Gruby D, Delafond H. Recherchessur des animalcules se devéloppant en grand nombredansl' estomacetdan les intestins pendant la digestion des animaux herbivores et carnivores. C R Acad Sci Paris. 1843;17:1304–8.
15. Russell JB. Factors that alter rumen microbial ecology. Science 2001;292:1119–22.
16. Firkins JL. Extending Burk Dehority's perspectives on the role of ciliate protozoa in the rumen. Front Microbiol. 2020;11:17.
17. Hobson PN, Stewart CS. The Rumen Microbial Ecosystem. 2nd ed. London: Blackie Academic & Professional; 1997.
18. Dehority BA. Rumen ciliate protozoa of the blue duiker (*Cephalophus monticola*), with observations on morphological variation lines within the species *Entodinium dubardi*. J Eukaryot Microbiol. 1994;41:103–11.
19. Tymensen L, Barkley C, McAllister TA. Relative diversity and community structure analysis of rumen protozoa according to T-RFLP and microscopic methods. J Microbiol Methods. 2012;88:1–6.
20. Ishaq SL, Wright A-DG. Design and validation of four new primers for next-generation sequencing to target the 18s rRNA genes of gastrointestinal ciliate protozoa. Appl Environ Microbiol. 2014;80:5515–21.
21. Kittelmann S, Seedorf H, Walters WA, Clemente JC, Knight R, Gordon JI, et al. Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. PLoS ONE. 2013;8:e47879.
22. Zhao Y, Yi Z, Gentekaki E, Zhan A, Al-Farraj SA, Song W. Utility of combining morphological characters, nuclear and mitochondrial genes: An attempt to resolve the conflicts of species identification for ciliated protists. Mol Phylogenet Evol. 2016;94:718–29.
23. Moon-van der Staay SY, van der Staay GWM, Michalowski T, Jouany J-P, Pristas P, Javorský P, et al. The symbiotic intestinal ciliates and the evolution of their hosts. Eur J Protistol. 2014;50:166–73.
24. Park T, Meulia T, Firkins JL, Yu Z. Inhibition of the rumen ciliate *Entodinium caudatum* by antibiotics. Front Microbiol. 2017;8:1189.
25. Li Z, Deng Q, Liu Y, Yan T, Li F, Cao Y, et al. Dynamics of methanogenesis, ruminal fermentation and fiber digestibility in ruminants following elimination of protozoa: a meta-analysis. J Anim Sci Biotechnol. 2018;9:89.
26. Feng J, Jiang C, Sun Z, Hua C, Wen J, Miao W, et al. Single-cell transcriptome sequencing of rumen ciliates provides insight into their molecular adaptations to the anaerobic and carbohydrate-rich rumen microenvironment. Mol Phylogenet Evol. 2020;143:106687.
27. Blackburn EH, Gall JG. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. J Mol Biol. 1978;120:33–53.
28. Greider CW, Blackburn EH. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. Nature. 1989;337:331–7.
29. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 2013;11:e1001473.
30. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. eLife. 2016;5:e19090.
31. Or-Rashid MM, Odongo NE, McBride BW. Fatty acid composition of ruminal bacteria and protozoa, with emphasis on conjugated linoleic acid, vaccenic acid, and odd-chain and branched-chain fatty acids1. J Anim Sci. 2007;85:1228–34.
32. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12:519–22.
33. Macaulay IC, Teng MJ, Haerty W, Kumar P, Ponting CP, Voet T. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nat Protoc. 2016;11:2081–103.
34. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA. 2002;99:5261–6.
35. Dehority BA Ciliate protozoa. In: Makkar HPS, McSweeney CS. Methods in Gut Microbial Ecology for Ruminants. Berlin/Heidelberg: Springer-Verlag; 2005. p. 67–78.
36. Dehority BA. Laboratory Manual for Classification and Morphology of Rumen Ciliate Protozoa. 1st ed. Boca Raton: CRC Press; 1993.
37. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 2016;102:3–11.
38. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput. Mol Cell Biol. 2012;19:455–77.
39. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016;44:e147.
40. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.
41. Maurer-Alcala XX, Yan Y, Pilling OA, Knight R, Katz LA. Twisted tales: insights into genome diversity of ciliates using single-cell 'omics. Genome Biol Evol. 2018;10:1927–39.
42. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–54.
43. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47:D807–11.
44. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
45. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
46. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33:W465–7.
47. Tourancheau AB, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A. Genetic code deviations in the ciliates: evidence for multiple and independent events. EMBO J. 1995;14:3262–7.

48. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955–64.
49. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods. 2015;8:12–24.
50. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
51. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079–86.
52. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
53. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.
54. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004.
55. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
56. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 2020;36:5516–8.
57. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14.
58. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2020;49: D412–9.
59. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018; 46:W95–101.
60. Haitjema CH, Gilmore SP, Henske JK, Solomon KV, Groot R, de, Kuo A, et al. A parts list for fungal cellulosomes revealed by comparative genomics. Nat Microbiol. 2017;2:17087.
61. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
62. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 2000; 28:292.
63. You S, Chen C-C, Tu T, Wang X, Ma R, Cai H, et al. Insight into the functional roles of Glu175 in the hyperthermostable xylanase XYL10C-ΔN through structural analysis and site-saturation mutagenesis. Biotechnol Biofuels. 2018;11:159.
64. Bailey MJ, Biely P, Poutanen K. Interlaboratory testing of methods for assay of xylanase activity. J Biotechnol. 1992;23:257–70.
65. Yang H, Zhang Y, Li X, Bai Y, Xia W, Ma R, et al. Impact of disulfide bonds on the folding and refolding capability of a novel thermostable GH45 cellulase. Appl Microbiol Biotechnol. 2018;102:9183–92.
66. He H, Wu S, Mei M, Ning J, Li C, Ma L, et al. A combinational strategy for effective heterologous production of functional human lysozyme in *Pichia pastoris*. Front Bioeng Biotechnol. 2020;8:118.
67. Pan X, Cai Y, Li Z, Chen X, Heller R, Wang N, et al. Modes of genetic adaptations underlying functional innovations in the rumen. Sci China Life Sci. 2021;64:1–21.
68. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.
69. Zheng W, Wang C, Lynch M, Gao S. The compact macronuclear genome of the ciliate *halteria grandinella*: a transcriptome-like genome with 23,000 nanochromosomes. mBio 2021;12:e01964–20.
70. Ahrendt SR, Quandt CA, Ciobanu D, Clum A, Salamov A, Andreopoulos B, et al. Leveraging single-cell genomics to expand the fungal tree of life. Nat Microbiol. 2018;3:1417.
71. Labarre A, López-Escardó D, Latorre F, Leonard G, Bucchini F, Obiol A, et al. Comparative genomics reveals new functional insights in uncultured MAST species. ISME J. 2021;15:1767–81.
72. Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 2011;331:463–7.
73. Puniya AK, Singh R, Kamra DN. Rumen microbiology: from evolution to revolution. 1st ed. Heidelberg: Springer; 2015.
74. Hillis DM, Moritz C, Mable BK. Molecular Systematics. 2nd ed. Sunderland: Sinauer Associates; 1996.
75. Chen L, Qiu Q, Jiang Y, Wang K, Lin Z, Li Z, et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science. 2019;364:eaav6202.
76. Gao F, Roy SW, Katz LA. Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. mBio. 2015;6:e01998–14.
77. De Luca D, Piredda R, Sarno D, Kooistra WHCF. Resolving cryptic species complexes in marine protists: phylogenetic haplotype networks meet global DNA metabarcoding datasets. ISME J. 2021;15:1931–42.
78. Zufall RA, McGrath CL, Muse SV, Katz LA. Genome architecture drives protein evolution in ciliates. Mol Biol Evol. 2006;23:1681–7.
79. Yan Y, Maurer-Alcalá XX, Knight R, Kosakovsky Pond SL, Katz LA. Single-cell transcriptomics reveal a correlation between genome architecture and gene family evolution in ciliates. mBio. 2019;10:e02524–19.
80. La Terza A, Papa G, Miceli C, Luporini P. Divergence between two Antarctic species of the ciliate *Euplotes*, *E.focardii* and *E.nobilii*, in the expression of heat-shock protein 70 genes. Mol Ecol. 2001;10:1061–7.
81. Sharma N, Bryant J, Wloga D, Donaldson R, Davis RC, Jerka-Dziadosz M, et al. Katanin regulates dynamics of microtubules and biogenesis of motile cilia. J Cell Biol. 2007;178:1065–79.
82. Park T, Mao H, Yu Z. Inhibition of rumen protozoa by specific inhibitors of lysozyme and peptidases in vitro. Front Microbiol. 2019;10:2822.
83. Solomon KV, Haitjema CH, Henske JK, Gilmore SP, Borges-Rivera D, Lipzen A, et al. Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. Science. 2016;351:1192–5.
84. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun. 2018;9:870.
85. Findley SD, Mormile MR, Sommer-Hurley A, Zhang X-C, Tipton P, Arnett K, et al. Activity-based metagenomic screening and biochemical characterization of bovine ruminal protozoan glycoside hydrolases. Appl Environ Microbiol. 2011;77: 8106–13.
86. Takenaka A, Tajima K, Mitsumori M, Kajikawa H. Fiber digestion by rumen ciliate protozoa. Microbes Environ. 2004;19:203–10.
87. Rubino F, Carberry C, M Waters S, Kenny D, McCabe MS, Creevey CJ. Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. ISME J. 2017;11:932–44.
88. Brochet S, Quinn A, Mars RA, Neuschwander N, Sauer U, Engel P. Niche partitioning facilitates coexistence of closely related honey bee gut bacteria. eLife. 2021;10:e68583.
89. Andersson JO. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 2005;62: 1182–97.
90. Xie F, Jin W, Si H, Yuan Y, Tao Y, Liu J, et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. Microbiome. 2021;9:137.
91. Ellison MJ, Conant GC, Cockrum RR, Austin KJ, Truong H, Becchi M, et al. Diet alters both the structure and taxonomy of the ovine gut microbial ecosystem. DNA Res. 2014;21:115–25.
92. Hook SE, Steele MA, Northwood KS, Wright A-DG, McBride BW. Impact of high-concentrate feeding and low ruminal pH on methanogens and protozoa in the rumen of dairy cows. Micro Ecol. 2011;62:94–105.
93. Li J, Zhong H, Ramayo-Caldas Y, Terrapon N, Lombard V, Potocki-Veronese G, et al. A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment. GigaScience. 2020;9:giaa057.
94. Shang Z, Wang X, Jiang Y, Li Z, Ning J. Identifying rumen protozoa in microscopic images of ruminant with improved YOLACT instance segmentation. Biosyst Eng. 2022;215:156–69.
95. Shen J, Zheng L, Chen X, Han X, Cao Y, Yao J. Metagenomic analyses of microbial and carbohydrate-active enzymes in the rumen of dairy goats fed different rumen degradable starch. Front Microbiol. 2020;11:1003.

ACKNOWLEDGEMENTS

This study was financially supported by the National Natural Science Foundation of China (31902126, U21A20247, and 31822052) and the China Postdoctoral Science Foundation (2019M663841). We thank the High-Performance Computing Platform of Northwest A&F University and the National Supercomputing Center in Xi'an and Hefei for providing the computing resources for the bioinformatic analyses.

AUTHOR CONTRIBUTIONS

ZL and YJ conceived and supervised the project. ZL, XW, TZ, YL, TS and XX collected the samples. XW, ZL, YL, FL, YH, HH, JN, and JT carried out the experiments. ZL, YZ, XW, TZ, XD, XP, RJ, YY, and SG performed bioinformatic analyses. ZL and ZY wrote the manuscript. ZY, YJ, HH, JY and BY revised the manuscript. All authors have read and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41396-022-01306-8.

Correspondence and requests for materials should be addressed to Huoqing Huang or Yu Jiang.