BMC Genomic Data

DATA NOTE

Open Access

A curated database of rumen ciliate protozoal 18S rRNA gene sequences for metataxonomic applications

Katie Lawther[1,2*], Ilma Tapio[3], Arturo Vera-Ponce de León[4], Velma T. E. Aho[4], Sharon A. Huws[1] and Nicholas J. Dimonaco[1*]

Abstract

Objectives Protozoa are key members of the rumen microbiome playing significant roles in nutrient cycling and methane production, yet are understudied. As rumen metataxonomic studies increasingly incorporate protozoal primers, the lack of curated dedicated reference databases limits accurate classification. This dataset was developed to address that gap and support future protozoa-focused rumen microbial analyses.

Data description The curated dataset comprises 228 rumen ciliate protozoal 18S rRNA gene sequences sourced from publicly available datasets. Sequences were processed to remove redundancy and standardise naming. The final database spans 23 families, 53 genera, and 100 species, and is suitable for use in metataxonomic pipelines, including QIIME2. It provides a valuable resource for researchers aiming to improve taxonomic resolution of protozoal communities in rumen environments.

Keywords Protozoa, Ciliates, 18S, rRNA gene, Rumen, Metataxonomy

Objective

The rumen microbiome is a complex and dynamic ecosystem essential to ruminant digestion and health. While bacterial and archaeal constituents have been extensively studied, the eukaryotic members, particularly fungi and protozoa, remain comparatively under-characterised. This disparity stems from challenges with axenic culturing of rumen eukaryotes and the complexity of their genomes [1]. Consequently, metagenomic studies have been heavily biased toward prokaryotic members and metataxonomic approaches have traditionally focused on 16S ribosomal ribonucleic acid (rRNA) gene sequencing.

Despite making up nearly half of the rumen microbial biomass, the biological and metabolic roles of protozoa remain poorly defined in vivo [1, 2]. Defaunation studies, the deliberate removal of protozoa from the rumen, report an increase in microbial protein flow by up to 30% and a reduction of methane emissions by 11%, underscoring their influence on nutrient cycling and greenhouse gas production [2]. Protozoa are major hydrogen producers via their hydrogenosomes, which are specialised organelles that generate H_2 during fermentation. This hydrogen production facilitates interspecies transfer

*Correspondence:
Katie Lawther
katiejlawther@gmail.com
Nicholas J. Dimonaco
nicholas@dimonaco.co.uk
[1]Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, Belfast, UK
[2]Laboratory of Microbiology, Wageningen University and Research, Wagneningen, Netherlands
[3]Genomics and Breeding, Natural Resources Institute Finland (Luke), Jokioinen, Finland
[4]Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

to methanogens, with many methanogenic archaea found in close physical association with protozoal cells [3]. While protozoa thus play a central role in methane formation, their impact on methanogen community structure is complex and variable [3].

Considering the importance of protozoa in the rumen microbiome, it is increasingly common to include members of eukaryotic taxa such as ciliate protozoa or anaerobic fungi in rumen metataxonomic studies. However, there is still a lack of appropriate databases specific to protozoal species equivalent to those available for bacterial and archaeal species [4–7]. Therefore, this database was built to support future rumen metataxonomic studies with tools such as QIIME2 [8].

Data description

These data were sourced from three publications.

1. One hundred and sixty-eight 18S rRNA gene sequences were sourced directly from the supplementary material provided in Kittelmann et al. [9]
2. Sixty-five single amplified genome (SAG) assemblies were sourced from Li et al. [10], PRJNA777442
3. One macronuclear genome of *Entodinium caudatum* was sourced from Park et al. [11], PRJNA380643

Barrnap v0.9 [12] was used with default parameters and –kingdom euk to extract ribosomal RNA genes and those identified as 18S were selected. The other rRNA genes were excluded (28S, 5S_8S and 5S). The 18S rRNA gene sequences from all three sources were then combined (Data file 1 [13]). All sequences were converted to DNA (U-T) and then to remove redundancy in the database, CD-HIT (–est) [14] was used to cluster the sequences at 100% identity and 100% length.

The output from cd-hit-est (.clstr file - Data file 2 [15]), was then manually examined, and for clusters containing multiple identical sequences (length and sequence were identical), only the first sequence in the cluster was kept as a 'representative' sequence. The remaining identical sequences were removed and recorded (38 sequences in total, Data file 3 [16]). The sequences that were retained then underwent one of two filtering steps based on their lineages or names.

1. Where IDs (NCBI accession) differed but lineages (provided by the publications) were the same, the sequences were assigned a group number (groups 0–22). The retained *representative* sequence was then renamed to include the allocated group number (all group member information can be found in *Data file 4* [17]).
2. Where the species name differed, the species name was removed and generalised to `s__`

For example:

```
>Cluster 14
>AB535662|k__Eukaryota;...;g__Ostracodinium;s__Ostracodinium_gracile *
>AB536718|k__Eukaryota;...;g__Ostracodinium;s__Ostracodinium_trivesiculatum... at +/100%

Was generalised to:
>AB_group14|k__Eukaryota;...;g__Ostracodinium;s__
```

More information and examples regarding the renaming process can be found at the project's GitHub repository https://github.com/lawkj/Protozoal_18S_rRNA_Database.

This curated database (v1.0.0) contains 228 rumen ciliate protozoal 18S rRNA gene sequences, representing 23 families, 53 genera, and 100 species (RNA sequences Data file 5 [18] and DNA sequences Data file 6 [19], taxonomy Data file 7 [20]).

In addition to the curated 18S rRNA gene sequences, we provide several supporting files formatted for direct use with QIIME2 [8], which were produced using QIIME2 v2024.10.1 with sklearn v1.4.2. Taxonomy information is available both as a plain text file (Data file 7 [20]), and as a QIIME2 'artifact' file (Data file 8 [21]), while the full set of 18S rRNA gene sequences, provided in DNA format, is also included as a QIIME2 'artifact' file (Data file 9 [22]). We also provide a pre-trained Naive Bayes classifier [23] compatible with QIIME2 (Data file 10 [24]) and a QIIME2-formatted BLAST database (Data file 11 [25]). The FASTA sequence files (Data files 5 and 6 [18, 19]) and tab-delimited taxonomy file (Data file 7 [20]) follow standard formats compatible with mothur [26] and other sequence search and classification tools.

Together, this range of provided files enables straightforward integration of the RumenProtozoaDB into QIIME2, mothur and other 18S rRNA gene sequence analysis workflows. Please see Table 1 for full details and links to the data.

Table 1 Overview of data files

Label	Data file	File type	Data repository
Data file 1	`RumenProtozoaDBv1.0.0.18S.rawsequences`	`fasta`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [13]
Data file 2	`RumenProtozoaDBv1.0.0.18S.clsr`	`xlsx`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [15]
Data file 3	`RumenProtozoaDBv1.0.0.18S.filtered prerename`	`fasta`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [16]
Data file 4	`RumenProtozoaDBv1.0.0.18S.group info`	`csv`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [17]
Data file 5	`RumenProtozoaDBv1.0.0.18S.RNAsequences`	`fasta`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [18]
Data file 6	`RumenProtozoaDBv1.0.0.18S.DNAsequences`	`fasta`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [19]
Data file 7	`RumenProtozoaDBv1.0.0.18S.taxonomy`	`txt`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [20]
Data file 8	`2024.10.RumenProtozoaDBv1.0.0.18S.taxonomy`	`qza`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [21]
Data file 9	`2024.10.RumenProtozoaDBv1.0.0.18S.sequences`	`qza`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [22]
Data file 10	`2024.10.RumenProtozoaDBv1.0.0.18S.nb.sklearn-1.4.2.qza`	`qza`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [24]
Data file 11	`2024.10.RumenProtozoaDBv1.0.0.18S.blastdb.qza`	`qza`	Zenodo (https://doi.org/10.5281/zenodo.18324737) [25]

Limitations

These data are limited in size, drawing from only three independent studies; however, this reflects the totality of publicly available data for rumen-associated protozoal 18S rRNA gene sequences and genomes. A limitation of this database is that it focuses exclusively on the 18S rRNA region and does not cover alternative regions, such as ITS or 28S, which may also be used for protozoal taxonomic classification. Should additional data become publicly available, the authors intend to expand the database and update the GitHub repository accordingly.

Abbreviations

rRNA Ribosomal ribonucleic acid
SAG Single amplified genome

Acknowledgments

Not applicable.

Author contributions

KL, NJD; study conceptualisation, data analysis and data note draft. VTEA, AVPL, and IT; study conceptualisation and data note review. SAH; data note review.

Funding

Not applicable.

Data availability

All data have been made available at https://github.com/lawkj/Protozoal_18S_rRNA_Database, Zenodo https://doi.org/10.5281/zenodo.18324737.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 11 November 2025 / Accepted: 27 March 2026
Published online: 03 April 2026

References

1. Andersen TO, Altshuler I, León A, Walter JM, McGovern E, Keogh K, et al. Metabolic influence of core ciliates within the rumen microbiome. ISME J. 2023;17(7):1128–40.
2. 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.
3. Morgavi DP, Martin C, Jouany J-P, Ranilla MJ. Rumen protozoa and methanogenesis: not a simple cause–effect relationship. Br J Nutr. 2012;107(3):388–97.
4. McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, et al. Greengenes2 unifies microbial data in a single reference tree. Nat Biotechnol. 2024;42(5):715–18.
5. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal rna gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41(D1):590–96.
6. Seedorf H, Kittelmann S, Henderson G, Janssen PH. RIM-DB: a taxonomic framework for community structure analysis of methanogenic archaea from the rumen and other intestinal environments. PeerJ. 2014;2:494.
7. Henderson G, Yilmaz P, Kumar S, Forster RJ, Kelly WJ, Leahy SC, et al. Improved taxonomic assignment of rumen bacterial 16s rrna sequences using a revised silva taxonomic framework. PeerJ. 2019;7:6496.
8. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. Nat Biotechnol. 2019;37(8):852–57.
9. Kittelmann S, Devente SR, Kirk MR, Seedorf H, Dehority BA, Janssen PH. Phylogeny of intestinal ciliates, including charonina ventriculi, and comparison of microscopy and 18s rrna gene pyrosequencing for rumen ciliate community structure analysis. Appl Environ Microb. 2015;81(7):2433–44.
10. Li Z, Wang X, Zhang Y, Yu Z, Zhang T, Dai X, et al. Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates. ISME J. 2022;16(12):2775–87.

11. 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(3):1416–27.
12. Seemann T. Barrnap 0.9: rapid ribosomal RNA prediction. Github. 2013. https://github.com/tseemann/barrnap.
13. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: raw 18S rRNA gene sequences. Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
14. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–52.
15. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: sequence clustering results. Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
16. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: filtered 18S sequences. Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
17. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: grouping information for sequences. Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
18. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB V1.0.0: RNA 18S sequences.
19. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: DNA 18S sequences. Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
20. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: taxonomy annotations (txt). Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
21. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: taxonomy artifact for QIIME 2 (.qza). Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
22. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: sequence artifact for QIIME 2 (.qza). Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
23. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90.
24. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: Naive Bayes classifier artifact for QIIME 2 (.qza). Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
25. Lawther K, Tapio I, León A, Aho VTE, Huws SA, Dimonaco NJ. RumenProtoZoaDB v1.0.0: BLAST database artifact for QIIME 2 (.qza). Zenodo. 2026. https://doi.org/10.5281/zenodo.18324737.
26. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, communitysupported software for describing and comparing microbial communities. Appl Environ Microb. 2009;75(23):7537–41.

Publisher's Note