

PacBio full-length 16S rRNA gene sequencing processed with Emu and GTDB provides the highest taxonomic resolution for rumen bacteriome profiling

Hanbeen Kim¹, Mi Zhou¹, Limei Lin¹, Wen Zhu^{1,2}, Tim A. McAllister³, Le Luo Guan^{1,*}

¹ Faculty of Land and Food Systems, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada

² College of Animal Science and Technology, Anhui Agricultural University, Hefei, Anhui 230036, PR China

³ Agriculture and Agri-Food Canada, Lethbridge Research Centre, Lethbridge, AB T1J 4B1, Canada

*Corresponding author: Faculty of Land and Food Systems, The University of British Columbia, 2205 East Mall, Vancouver, BC V6T 1Z4, Canada. E-mail: leluo.guan@ubc.ca

Abstract

Although full-length 16S rRNA gene sequencing has substantially improved taxonomic resolution compared to short-read approaches, a high proportion of unclassified taxa are reported in rumen microbiome studies. This limitation is largely driven by platform-specific analytical workflows and the insufficient representation of rumen-associated lineages in commonly used reference databases. Here, we identified the optimal combination of sequencing platform, analytical workflow, and reference database to improve rumen bacteriome classification. We analyzed short-read and full-length 16S rRNA gene sequences from rumen samples collected from two beef cattle populations. Short-read sequences were generated using Illumina NextSeq2000 and processed with QIIME2. Full-length sequences were generated using PacBio Revio (PacBio-16S) and Nanopore MinION (ONT-16S); PacBio-16S data were analyzed using QIIME2 and Emu, while Nanopore data were analyzed using EPI2ME and Emu. Five reference databases were evaluated across all analytical approaches: SILVA 138.2, SILVA 138.2 with Hungate1000 collection, NCBI, Greengenes2, and GTDB. The comparisons showed that PacBio-16S (Emu) achieved the highest proportion of classified reads among all platform-specific workflows, while GTDB consistently produced the highest number of non-redundant classified taxa. *Prevotella*, a dominant rumen genus, was abundant in Illumina and PacBio-16S datasets but was underrepresented in ONT-16S workflows. Species-level analyses further demonstrated that PacBio-16S (Emu) reliably provided more consistent and high-resolution identification of *Prevotella* species under GTDB across two beef populations. Overall, our results demonstrate that sequencing platform, workflow choice, and database selection strongly influence rumen bacteriome profiles. We recommend PacBio-16S (Emu) under GTDB as the most reliable workflow for achieving high-resolution taxonomic classification of rumen bacteriome.

Keywords rumen, bacteriome, full-length 16S rRNA gene sequencing, taxonomic resolution, reference database limitation, Pacific biosciences, Oxford nanopore technologies, Emu, GTDB

Main body

Third-generation sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) have demonstrated improved taxonomic resolutions and reduced ambiguities compared to short-read methods by capturing the full-length 16S rRNA gene sequences [1]. However, despite these technical advantages, several studies using full-length 16S rRNA gene sequencing to profile the rumen microbiome have reported a substantial proportion of unclassified bacterial reads at the species level [2–4]. This outcome could reflect the limited representation of rumen-associated lineages in conventional reference databases [5]. SILVA remains one of the most widely used reference databases for 16S rRNA gene-based studies, but it does not readily support species-level classification [6]. Although the incorporation of

rumen-specific genomes, such as those in the Hungate1000 collection [7] has improved classification accuracy [5], representation of rumen bacteria is still limited. Alternative databases including NCBI, Greengenes 2 (GG2), or GTDB, incorporate environmentally metagenome-assembled genomes and have progressively expanded taxonomic coverage beyond traditional reference collections, enabling improved classification of full-length 16S rRNA gene amplicon sequencing [8]. However, it is unclear how these reference databases could improve the taxonomic classification of rumen bacteria. In addition to the reference databases, processing bioinformatic pipelines also presents a challenge. PacBio full-length 16S rRNA gene (PacBio-16S) reads are commonly processed using the DADA2 denoise-ccs and Naïve Bayes classifier within QIIME2 [9], while for ONT full-length 16S rRNA gene (ONT-16S) sequences, EPI2ME (wf-16s) platform is often used

Received: 12 May 2026. Revised: 26 May 2026. Accepted: 27 May 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of the International Society for Microbial Ecology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

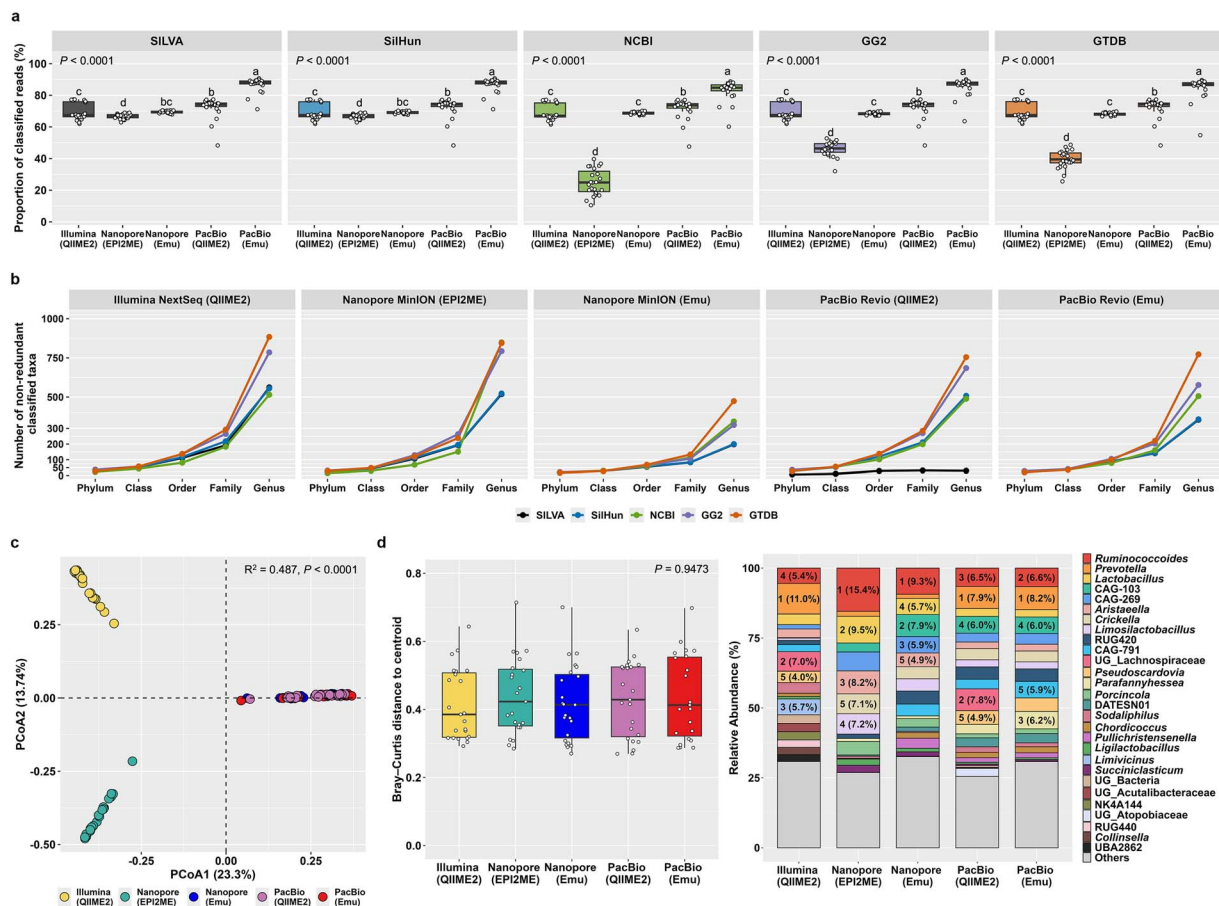


Figure 1 Impact of sequencing platform and reference database on taxonomic resolution of the rumen bacteriome from 23 Kinsella composite hybrid steers. (a) Proportion of classified reads (%) across sequencing platforms using different reference databases. Different letters indicate significant differences among platforms within each database. (b) Number of non-redundant classified taxa across taxonomic ranks for each sequencing platform and reference database. Non-redundant classified taxa were defined as the number of unique taxonomic labels at each rank after excluding unassigned taxa. (c) Principal coordinates analysis based on bray-Curtis dissimilarity at the genus level under GTDB classification. (d) Beta diversity dispersion measured as the distance to centroid based on bray-Curtis dissimilarity. (e) Taxonomic distribution of major genera across sequencing platforms under GTDB classification. Major genera were defined as those with a prevalence $\geq 50\%$ of samples within each sequencing platform and a relative abundance $\geq 2.0\%$ in at least one platform. Only these major genera are shown. “Others” represents remaining taxa not included among the major genera. Numbers within bars indicate rank and relative abundance of the top five genera within each platform.

for taxonomic assignment. Additionally, Emu applies an expectation-maximization algorithm to reduce false positives and negatives arising from both ONT and PacBio reads [10]. In this study, we conducted a comparative analysis of sequencing platforms, bioinformatic pipelines, and reference databases to identify best practices for taxonomic profiling of the rumen bacteriome.

The sequence data of 23 rumen samples collected from Kinsella Composite hybrid (KC) steers [11] were generated and processed using short-read (Illumina NextSeq2000 with QIIME2) and full-length 16S rRNA gene sequencing (PacBio Revio with QIIME2 and Emu; and ONT MinION with EPI2ME and Emu). Species-level resolution was further evaluated using full-length 16S rRNA datasets from both KC steers ($n = 23$) [11] and Angus bulls ($n = 24$) [12], enabling assessment of the consistency of classification patterns across independent rumen populations. Taxonomic profiling was then performed using five reference databases—SILVA 138.2 (SILVA), SILVA 138.2 supplemented with the Hungate1000 collection (SilHun), NCBI 16S RefSeq (NCBI, downloaded 2025.06.17), GG2 (Release 2024.9), and GTDB (Release 10-RS226, v226). The performance of each sequencing platform and database was

evaluated based on the average proportion of classified reads, diversity metrics, cross-method concordance, and differential abundance analyses. Detailed animal information, experimental methods, and bioinformatic workflows are provided in the Supplementary Information and Fig. S1. Although the ONT-16S exhibited a relatively lower proportion of bases with Q-scores ≥ 30 compared to the other platforms, all three sequencing platforms exhibited high read-level quality metrics (Tables S1–S3). When comparing classified read proportions, PacBio-16S (Emu) data yielded a significantly higher proportion of classified reads than other approaches across all reference databases (Fig. 1a; $P < .0001$). Due to limitations in species-level identification for short-read amplicon data [1], comparisons among sequencing platforms were conducted at the genus level. Rarefaction curves based on the observed genera indicated adequate sequencing depth across all platforms (Fig. S2). Across databases, GTDB consistently yielded the highest number of non-redundant classified taxa from phylum to genus levels (Fig. 1b), as well as the highest number of observed genera, followed by GG2 (Fig. S3; $P < .0001$). However, a greater number of detected genera does not necessarily indicate improved taxonomic

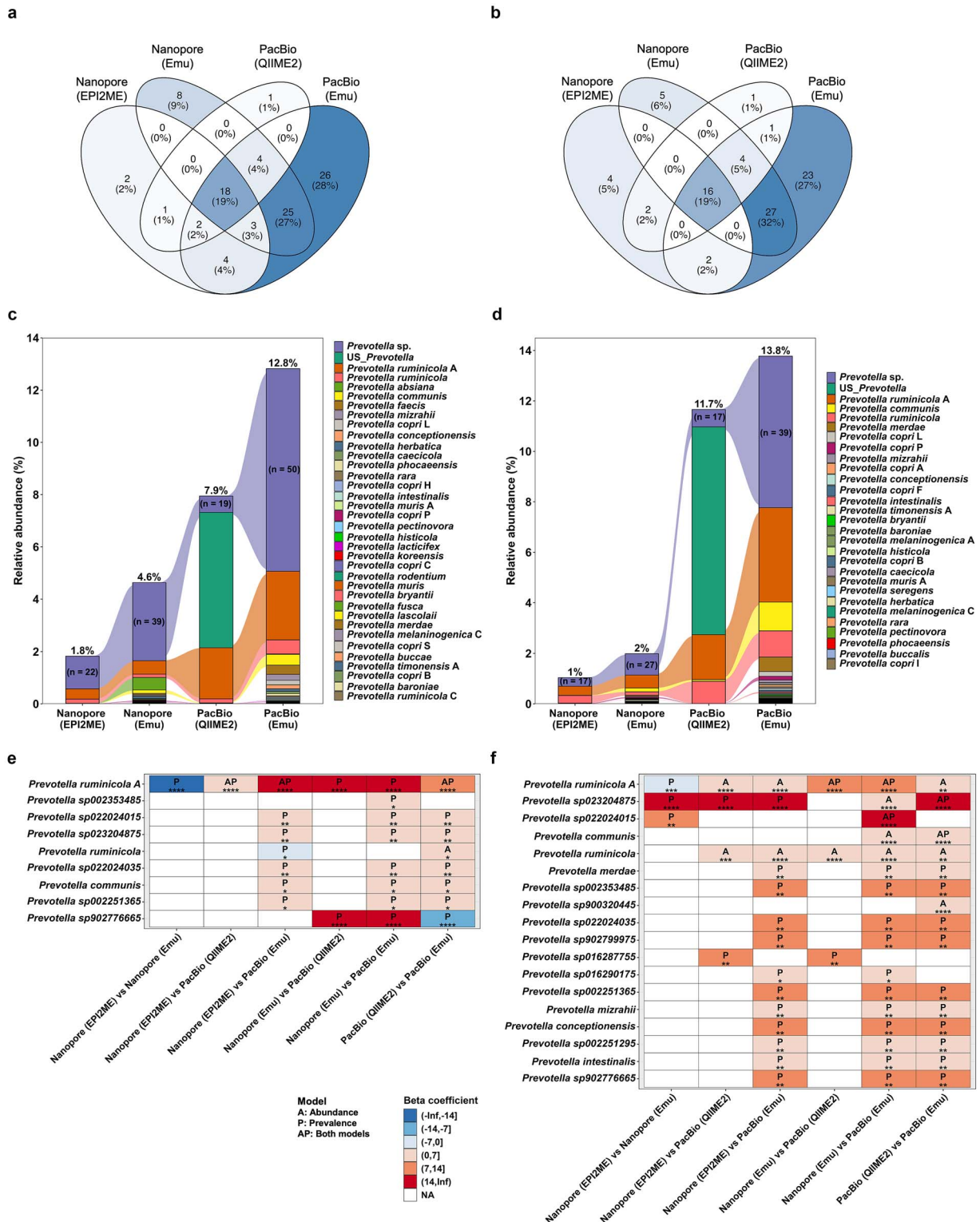


Figure 2 Comparison of *Prevotella* species across sequencing platforms based on the GTDB (release 10-RS226) database. (a–b) Venn diagrams showing the number and proportion (%) of shared and unique *Prevotella* species identified across sequencing platforms in (a) 23 Kinsella composite hybrid steers and (b) 24 Angus bulls. (c–d) relative abundance profiles of *Prevotella* species across sequencing platforms in (c) 23 Kinsella composite hybrid steers and (d) 24 Angus bulls. Bars represent cumulative relative abundance, with individual colors indicating species-level contributions. (e–f) pairwise differential abundance and prevalence analysis of *Prevotella* species across sequencing platforms based on MaAsLin3. Only *Prevotella* species with a prevalence $\geq 50\%$ of samples within each sequencing platform and a relative abundance $\geq 0.05\%$ in at least one platform were included in the analysis. Significant associations ($q < 0.05$) are visualized. Tiles are colored according to beta coefficients, and labels indicate the dominant model type: A, abundance; P, prevalence; AP, both models with concordant direction. When abundance and prevalence models showed opposite directions, the model with the larger absolute beta coefficient was selected. US_*Prevotella*, unclassified *Prevotella* species.

accuracy or resolution. Although Illumina-QIIME2 with GTDB detected a greater number of genera, this may partly reflect inflated diversity estimates associated with short-read amplicon sequencing. Previous study has reported that short-read regions may also overestimate bacterial diversity compared with full-length 16S rRNA gene sequencing [13]. In contrast, PacBio full-length sequencing combined with Emu provided more consistent taxonomic profiles across databases and may therefore better reflect biologically robust genus-level classification. The superior performance of GTDB is attributable to its expanded genomic representation [14]. However, caution is warranted when interpreting these results, as the inclusion of MAGs may introduce limitations, including genome incompleteness and potential contamination, which may affect the reliability of taxonomic assignments. In the beta diversity, Illumina-derived profiles were largely separated from long-read sequencing platforms across most databases, except SILVA where separation was less distinct (Fig. S4, $P < .0001$). Long-read workflows showed relatively tighter clustering, particularly under SilHun, NCBI, and GTDB classifications (Fig. 1c and Fig. S4). *Prevotella* is widely recognized as a dominant and functionally important genus in the rumen [15, 16]. Under GTDB classification, *Prevotella* was the most predominant genus in Illumina (QIIME2) and PacBio-16S (QIIME2 and Emu), whereas *Ruminococcoides* was dominant in both EPI2ME and Emu workflows for ONT-16S (Fig. 1d). *Prevotella* was also detected among the major genera under GG2 and SilHun classifications, although with lower relative abundance. In contrast, it was not identified as a dominant genus under SILVA and NCBI classifications, indicating substantial database-dependent variation in taxonomic assignments (Fig. S5). Based on its superior performance in taxonomic coverage and cross-platform consistency, GTDB was selected for subsequent species-level analyses. PacBio-16S (Emu) enabled the capture of a higher effective number of species and a greater number of major bacterial species, with consistent patterns observed across both datasets (Fig. S6). Spearman correlation of effective Shannon diversity was used to assess concordance in alpha diversity estimates across sequencing platforms and analytical workflows. The highest correlation was observed between PacBio-16S workflows (QIIME2 and Emu), while Nanopore workflows (EPI2ME and Emu) also showed strong agreement (Fig. S7). These results suggest that the choice of sequencing platform has a stronger impact on overall community structure than the choice of analysis pipeline. Interestingly, Emu-based analyses exhibited relatively strong correlations between PacBio and Nanopore datasets, indicating the analysis pipeline can partially mitigate platform-specific differences. Although the distribution of major genera varied between KC steers and Angus bulls, *Prevotella* consistently remained the most abundant genus in both PacBio-16S workflows under GTDB, whereas *Ruminococcoides* and *Aristaeella* were predominant in ONT-16S workflows in KC steers and Angus bulls, respectively (Fig. S8A). In addition, differential abundance analysis revealed that *Prevotella* was the only genus that consistently exhibited concordant differential patterns across both datasets, remaining significantly enriched in PacBio-16S relative to ONT-16S regardless of analytical workflow (Fig. S8B). We further investigated species-level variation within the *Prevotella* to assess how sequencing platform and analytical workflow influence taxonomic resolution. Notably, PacBio-16S (Emu) consistently identified a greater number of *Prevotella* species (KC steers, $n=82$ and Angus bulls $n=73$) and captured a broader distribution of species-level abundances (KC steers = 12.8% and Angus bulls = 13.8%) compared to other workflows (Fig. 2a–d and Tables S4–S5). In addition, *Prevotella* species showed

significant and concordant enrichment patterns in PacBio-16S relative to ONT-16S, with PacBio (Emu) providing the most consistent and robust signals across comparisons (Fig. 2e–f). Consistent with our findings, several studies have reported the potential underestimation of *Prevotella* in ONT-16S data [17–19]. Several technical factors may contribute to the observed underrepresentation of *Prevotella* species in ONT-16S. In particular, differences in PCR conditions (e.g. types of Taq polymerase, PCR cycles) and primer sequence (e.g. degenerate primers) could influence amplification biases across taxa [20]. As the primer information for the ONT 16S Barcoding Kit is not publicly disclosed, their taxonomic coverage cannot be directly evaluated, representing a limitation of this study. Future studies are needed to evaluate how PCR conditions and primer sequences in ONT-16S workflows contribute to biases in species-level resolution. Another limitation of this study is the absence of mock community validation. Although mock communities are useful for benchmarking accuracy, standardized rumen-specific mock communities remain limited, and available standards do not fully represent rumen microbial complexity. In addition, this study was based on two beef cattle cohorts, which may limit the generalizability of the findings to other ruminant systems. In conclusion, our findings demonstrate that PacBio (Emu) classified using GTDB provides improved species-level resolution and more consistent taxonomic profiling, particularly for ecologically important rumen bacterial taxa such as *Prevotella*.

Acknowledgements

The authors acknowledge the staff at Roy Berg Kinsella Ranch, University of Alberta (Alberta, Canada) and our laboratory members from both the Department of Agriculture, Food and Nutritional Science at the University of Alberta and the Faculty of Land and Food Systems at The University of British Columbia for their assistance with animal handling and sample collection.

Author contributions

H.K., M.Z., and L.L.G. designed the experiment and bioinformatic procedures. H.K. generated the raw sequencing data. L.L.G. provided the resources and reagents and supervised the research. H.K. implemented the experimental procedures and bioinformatic pipeline, and performed the statistical analyses, prepared the figures, and drafted the initial manuscript. H.K., M.Z., L.L., W.Z., T.A.M., and L.L.G. contributed to the revision and finalization of the manuscript. All authors reviewed and approved the final version of the manuscript.

Supplementary material

Supplementary material is available at *ISME Communications* online.

Conflicts of interest

The authors declare no conflicts of interest related to this work.

Funding

This work was funded by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery, NSERC New Frontier Research Global Program, NSERC-SSHRC Sustainable Agriculture Research Initiative (ALLRP588541–23), and Beef Cattle Research Council Cluster

(FDE.18.21C). This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliance.can.ca; RRG #5356). LLG is supported by NSERC Canadian Research Chair (Tier 1) program.

Data availability

The 16S rRNA gene sequencing data are accessible under the NCBI BioProject (PRJNA1378457).

Consent for publication

All authors have consented to publication.

References

- Johnson JS, Spakowicz DJ, Hong B-Y *et al.* Evaluation of 16s rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019;**10**:5029. <https://doi.org/10.1038/s41467-019-13036-1>
- Saborío-Montero A, López-García A, Gutiérrez-Rivas M *et al.* A dimensional reduction approach to modulate the core ruminal microbiome associated with methane emissions via selective breeding. *J Dairy Sci* 2021;**104**:8135–51. <https://doi.org/10.3168/jds.2020-20005>
- Brede M, Orton T, Pinior B *et al.* Pacbio and illumina miseq amplicon sequencing confirm full recovery of the bacterial community after subacute ruminal acidosis challenge in the rusitec system. *Front Microbiol* 2020;**11**:2020. <https://doi.org/10.3389/fmicb.2020.01813>
- Kang R, Song J, Park JK *et al.* Impact of forage sources on ruminal bacteriome and carcass traits in hanwoo steers during the late fattening stages. *Microorganisms* 2024;**12**:2082. <https://doi.org/10.3390/microorganisms12102082>
- Henderson G, Yilmaz P, Kumar S *et al.* Improved taxonomic assignment of rumen bacterial 16s rRNA sequences using a revised silva taxonomic framework. *PeerJ* 2019;**7**:e6496. <https://doi.org/10.7717/peerj.6496>
- Chuvochina M, Gerken J, Frentrup M *et al.* Silva in 2026: a global core biodata resource for rRNA within the DSMZ digital diversity. *Nucleic Acids Res* 2025;**54**(D1):D334–D341. <https://doi.org/10.1093/nar/gkaf1247>
- Seshadri R, Leahy SC, Attwood GT *et al.* Cultivation and sequencing of rumen microbiome members from the hungate1000 collection. *Nat Biotechnol* 2018;**36**:359–67. <https://doi.org/10.1038/nbt.4110>
- Pascoal F, Duarte P, Assmy P *et al.* Full-length 16s rRNA gene sequencing combined with adequate database selection improves the description of arctic marine prokaryotic communities. *Ann Microbiol* 2024;**74**:29. <https://doi.org/10.1186/s13213-024-01767-6>
- Bolyen E, Rideout JR, Dillon MR *et al.* Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nat Biotechnol* 2019;**37**:852–7. <https://doi.org/10.1038/s41587-019-0209-9>
- Curry KD, Wang Q, Nute MG *et al.* Emu: species-level microbial community profiling of full-length 16s rRNA oxford nanopore sequencing data. *Nat Methods* 2022;**19**:845–53. <https://doi.org/10.1038/s41592-022-01520-4>
- Li F, Hitch TCA, Chen Y *et al.* Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle. *Microbiome* 2019;**7**:6. <https://doi.org/10.1186/s40168-019-0618-5>
- Li F, Li C, Chen Y *et al.* Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. *Microbiome* 2019;**7**:92. <https://doi.org/10.1186/s40168-019-0699-1>
- Wagner J, Coupland P, Browne HP *et al.* Evaluation of pacbio sequencing for full-length bacterial 16s rRNA gene classification. *BMC Microbiol* 2016;**16**:274. <https://doi.org/10.1186/s12866-016-0891-4>
- Parks DH, Chaumeil P-A, Mussig AJ *et al.* GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. *Nucleic Acids Res* 2026;**54**(D1):D743–D754. <https://doi.org/10.1093/nar/gkaf1040>
- Henderson G, Cox F, Ganesh S *et al.* Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci Rep* 2015;**5**:14567. <https://doi.org/10.1038/srep14567>
- Betancur-Murillo CL, Aguilar-Marín SB, Jovel J. Prevotella: a key player in ruminal metabolism. *Microorganisms* 2023;**11**:1. <https://doi.org/10.3390/microorganisms11010001>
- Macip G, Soler-Comas A, Palomeque A *et al.* Comparative analysis of illumina and oxford nanopore sequencing platforms for 16s rRNA profiling of respiratory microbial communities. *Sci Rep* 2025;**15**:33688. <https://doi.org/10.1038/s41598-025-18768-3>
- Yeo K, Connell J, Bouras G *et al.* A comparison between full-length 16s rRNA oxford nanopore sequencing and illumina v3-v4 16s rRNA sequencing in head and neck cancer tissues. *Arch Microbiol* 2024;**206**:248. <https://doi.org/10.1007/s00203-024-03985-7>
- Tort-Miró C, Lorenzo-Rebenaque L, Montoro-Dasi L *et al.* Nanopore versus illumina to study the gut bacterial diversity of sows and piglets between farms with high and low health status. *BMC Vet Res* 2025;**21**:246. <https://doi.org/10.1186/s12917-025-04693-0>
- Silva DP, Epstein HE, Vega Thurber RL. Best practices for generating and analyzing 16s rRNA amplicon data to track coral microbiome dynamics. *Front Microbiol* 2023;**13**:1007877. <https://doi.org/10.3389/fmicb.2022.1007877>