



OPEN

Metagenomic analysis of the cow, sheep, reindeer and red deer rumen

Laura Glendinning^{1✉}, Buğra Genç^{2,3}, R. John Wallace² & Mick Watson¹

The rumen microbiota comprises a community of microorganisms which specialise in the degradation of complex carbohydrates from plant-based feed. These microbes play a highly important role in ruminant nutrition and could also act as sources of industrially useful enzymes. In this study, we performed a metagenomic analysis of samples taken from the ruminal contents of cow (*Bos Taurus*), sheep (*Ovis aries*), reindeer (*Rangifer tarandus*) and red deer (*Cervus elaphus*). We constructed 391 metagenome-assembled genomes originating from 16 microbial phyla. We compared our genomes to other publically available microbial genomes and found that they contained 279 novel species. We also found significant differences between the microbiota of different ruminant species in terms of the abundance of microbial taxonomies, carbohydrate-active enzyme genes and KEGG orthologs. We present a dataset of rumen-derived genomes which in combination with other publicly-available rumen genomes can be used as a reference dataset in future metagenomic studies.

The microbial communities which inhabit the rumen contain a mixture of bacteria, fungi, protozoa, viruses and archaea, and through fermentation are able to convert complex plant carbohydrates into short-chain volatile fatty acids. The metabolic pathways present have a large impact on feed efficiency^{1–3}, alongside other important production traits such as milk and fat yield^{4,5}. Understanding the processes by which food is digested in the rumen may allow us to improve feed efficiency in ruminants¹, either by the production of enzymes isolated from microbes⁶ or by manipulating the microbiota through the use of pre- or probiotics⁷. There are also other potential industrial uses for the enzymes produced by ruminal microbes, for example in processing biofuels, bioremediation, processing pulp/paper and textile manufacturing^{8–11}. Ruminants are also a large source of animal agriculture related methane emissions and gaining a greater understanding of which microbes are important in methane production could lead to improved methane mitigation strategies^{7,12–16}.

While inroads have been made towards culturing members of the ruminal microbiota^{17,18} there are still many members which have not been characterised. Metagenomics is a powerful tool which allows us to examine the entire genetic repertoire of the rumen microbiota without the need for culturing. Using this approach, we have obtained thousands of metagenome assembled genomes from cow rumen samples^{19,20} and hundreds of genomes from chicken caecal samples²¹, many of which were identified as novel species.

Several studies have examined the rumen microbiota using metagenomic techniques in cows and sheep. However, less effort has been made to characterise the microbiota of other ruminant species that may be less agriculturally-important but which could harbour microbes that produce a diverse biochemical reservoir of enzymes of industrial interest²². For example, wild ruminants are likely to consume a far more diverse diet than farm-raised individuals, and are therefore likely to contain microbes which are able to digest different substrates. In this paper we analyse rumen metagenomic data from four ruminant species: cow (*Bos Taurus*), sheep (*Ovis aries*), red deer (*Cervus elaphus*) and reindeer (*Rangifer tarandus*). We compare the microbiota of these species taxonomically and functionally and construct 391 named rumen-uncultured genomes (RUGs), representing 372 putative novel strains and 279 putative novel species.

Results

Construction of RUGs from rumen sequencing data. We produced 979G of Illumina sequencing data from 4 cows, 2 sheep, 4 red deer and 2 reindeer samples, then performed a metagenomic assembly of single samples and a co-assembly of all samples. This created a set of 391 dereplicated genomes (99% ANI (average nucleotide identity)) with estimated completeness $\geq 80\%$ and estimated contamination $\leq 10\%$ (Fig. 1). 284 of these genomes were produced from the single-sample assemblies and 107 were produced from the co-assemblies. 172 genomes were $>90\%$ complete with contamination $<5\%$, and would therefore be defined as

¹Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, Midlothian, UK. ²Rowett Institute, University of Aberdeen, Ashgrove Rd W, Aberdeen, Aberdeenshire, UK. ³Present address: Department of Laboratory Animals, Veterinary Medicine Faculty, Ondokuz Mayıs University, Atakum, Samsun, Turkey. ✉email: laura.glendinning@roslin.ed.ac.uk

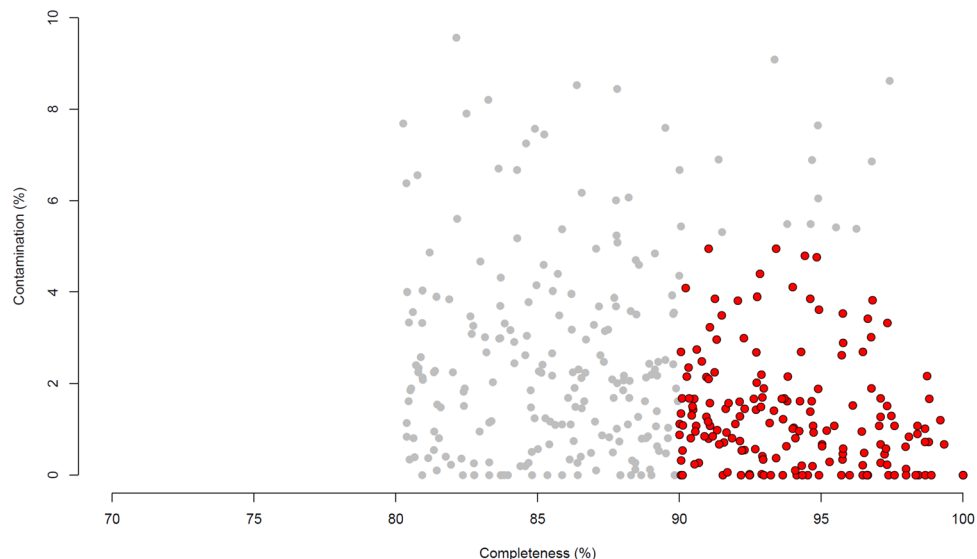


Figure 1. Contamination and completeness (defined by CheckM software) of 391 dereplicated metagenome-assembled-genomes from rumen samples. Grey: genomes which are 80–90% complete with 5–10% contamination. Red: genomes which are >90% complete with <5% contamination.

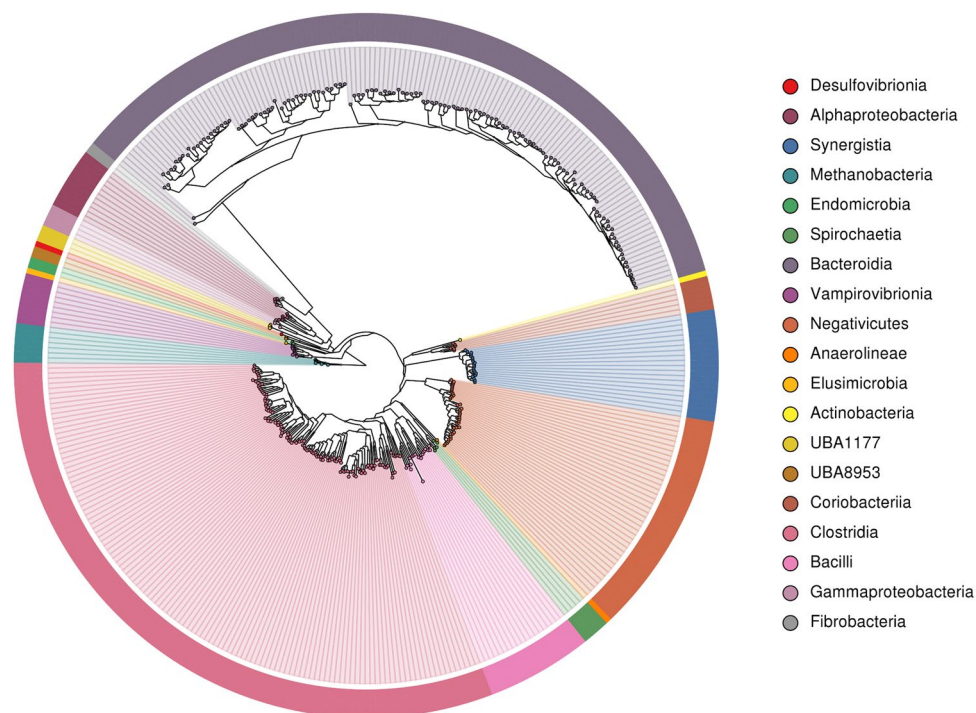


Figure 2. Phylogenetic tree of the 391 draft microbial genomes from rumen samples, labelled by taxonomic class. Taxonomies were defined by GTDB-Tk. The tree was produced by MAGpy, using GraPhlAn²⁴ (v.0.9.7), and rerooted at the branch between archaea and bacteria.

high-quality draft genomes by Bower et al.²³. The distribution of these RUGs between our samples can be found in Supplementary data S1 (based on coverage). Supplementary data S2 contains the predicted taxonomic assignment for each RUG while Fig. 2 shows a phylogenetic tree of the genomes.

The tree is dominated by the *Bacteroidota* (136 RUGs: All order *Bacteroidales*) and the *Firmicutes_A* (121 RUGs), followed by lesser numbers of the *Firmicutes_C* (40 RUGs), *Synergistota* (20 RUGs: All family *Aminobacteriaceae*), *Firmicutes* (19 RUGs), *Proteobacteria* (15 RUGs), *Cyanobacteriota* (9 RUGs: All family *Gastranaerophilaceae*), *Actinobacteriota* (7 RUGs), *Euryarchaeota* (7 RUGs: All family *Methanobacteriaceae*), *Spirochaetota*

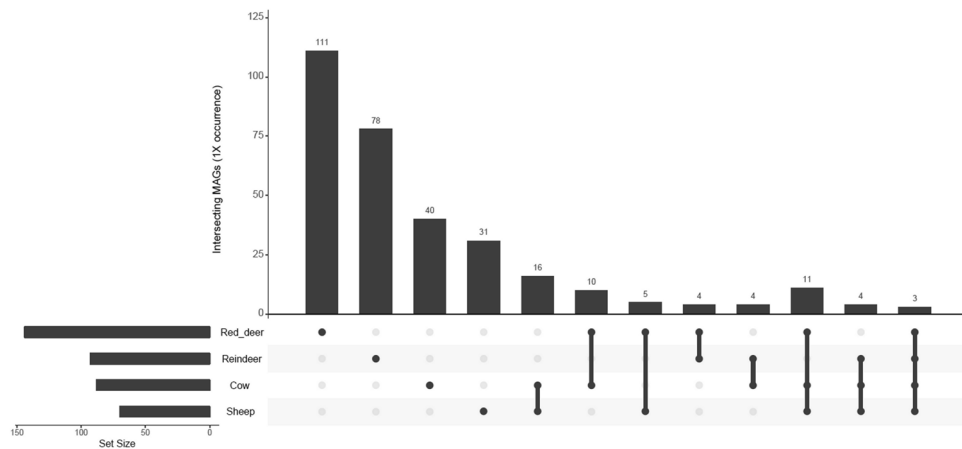


Figure 3. UpSetR graph showing the number of shared microbial genomes at average $\times 1$ coverage (after sub-sampling to equal depth) within four ruminant species.

(5 RUGs), *Elusimicrobiota* (3 RUGs: All family *Endomicrobiaceae*), *UBP6* (3 RUGs: All genus *UBA1177*), *Fibrobacterota* (2 RUGs: All genus *Fibrobacter*), *Riflebacteria* (2 RUGs: All family *UBA8953*), *Chloroflexota* (1 RUGs: family *Anaerolineaceae*) and *Desulfobacterota* (1 RUGs: genus *Desulfovibrio*). All members of the phylum *Firmicutes_A* belonged to the *Clostridia* class: orders *4C28d-15* (n = 9), *CAG-41* (n = 3), *Christensenellales* (n = 4), *Lachnospirales* (n = 56), *Oscillospirales* (n = 45), *Peptostreptococcales* (n = 2) and *Saccharofermentanales* (n = 2). *Firmicutes_C* contains the orders *Acidaminococcales* (n = 8) and *Selenomonadales* (n = 32). The phylum *Firmicutes* contained the orders *Acholeplasmatales* (n = 3), *Erysipelotrichales* (n = 1), *Izimaplasmatales* (n = 1), *ML615J-28* (n = 1), *Mycoplasmatales* (n = 1). *RFN20* (n = 7) and *RF39* (n = 5), The *Actinobacteria* contained the orders *Actinomycetales* (n = 1) and *Coriobacteriales* (n = 6). The *Proteobacteria* phylum contains the orders *Enterobacteriales* (n = 4), *Paracaedibacteriales* (n = 1), *RF32* (n = 8) and *UBA3830* (n = 2). The *Spirochaetota* contains the orders *Sphaerochaetales* (n = 1) and *Treponematales* (n = 4).

After sub-sampling, we found that samples from different ruminant species clustered significantly separately by abundance of RUGs (PERMANOVA: $P = 3e - 05$). This may be due to the fact that the vast majority of RUGs were only found in a single host species (Fig. 3), including 111 RUGs in red deer, 78 RUGs in reindeer, 40 RUGs in cow and 31 RUGs in sheep. Only 3 RUGs were found in $\geq 1X$ average coverage in all species: uncultured *Bacteroidaceae* sp. *RUG30019*, uncultured *Prevotella* sp. *RUG30028* and uncultured *Prevotella* sp. *RUG30114*.

We compared our RUGs to microbial genomes which had previously been sequenced from the rumen to determine if we had discovered any novel strains or species. We dereplicated our RUGs at 99% and 95% ANI to a “superset” of genomes containing rumen RUGs previously produced by our group²⁰, Hess et al.¹¹, Parks et al.²⁵, Solden et al.²⁶ and Svartström et al.²⁷ and the genomes from the Hungate collection¹⁷. After dereplication at 99% and the removal of any RUGs with $\geq 99\%$ ANI to an existing genome (as assigned by GTDB-Tk) or which clustered with members of the superset, 372 of our RUGs remained, representing putative novel strains. After dereplication at 95% and the removal of any RUGs with $\geq 95\%$ ANI to an existing genome (assigned by GTDB-Tk) or which clustered with members of the superset, 279 of our RUGs remained, representing putative novel species. The majority of these species originated from single-sample assemblies: 110 from red deer samples, 68 from reindeer samples, 23 from sheep samples and 1 from cow samples, suggesting that many novel microbial species remain to be discovered from non-cow ruminant hosts. These novel species are taxonomically diverse, with members belonging to the phyla *Bacteroidota* (n = 97), *Firmicutes_A* (n = 85), *Firmicutes_C* (n = 27), *Firmicutes* (n = 16), *Synergistota* (n = 14), *Proteobacteria* (n = 11), *Cyanobacteriota* (n = 9), *Actinobacteriota* (n = 5), *Spirochaetota* (n = 4), *Euryarchaeota* (n = 3), *Elusimicrobiota* (n = 3), *Riflebacteria* (n = 2), *Chloroflexota* (n = 1), *Desulfobacterota* (n = 1) and *UBP6* (n = 1).

31 of our total RUGs were able to be taxonomically identified to species level and these contain bacteria which are commonly isolated from the rumen including novel strains of *Bacteroidales bacterium* *UBA1184*²⁵, *Bacteroidales bacterium* *UBA3292*²⁵, *Butyrivibrio fibrisolvens*, *Escherichia coli*, *Fibrobacter* sp. *UWB2*²⁸, *Lachnospiraceae bacterium* *AC3007*¹⁷, *Lachnospiraceae bacterium* *UBA2932*²⁵, *Methanobrevibacter* sp. *UBA188*²⁵, *Methanobrevibacter* sp. *UBA212*²⁵, *Prevotella* sp. *UBA2859*²⁵, *Ruminococcaceae bacterium* *UBA3812*²⁵, *Ruminococcus* sp. *UBA2836*²⁵, *Sarcina* sp. *DSM 11001*¹⁷, *Selenomonas* sp. *AE3005*¹⁷, *Succiniclacticum ruminis* and *Succinivibrio dextrinosolvens*.

Comparing microbial taxonomies, CAZymes and KEGG orthologs between ruminant species. We assigned taxonomies to paired sequence reads using our custom kraken database containing RefSeq complete genomes, our RUGs, and the superset of rumen isolated microbial genomes. After subsampling we compared the abundance of members of the microbiota in different ruminant species at multiple taxonomic levels. Averaging reads across rumens species, the vast majority of reads mapped to bacteria (Sheep: 97%, Cow: 97%, Reindeer: 92%, Red deer: 98%) with smaller amounts of archaea (Sheep: 2.3%, Cow: 2.1%, Reindeer: 6.3%, Red deer: 1.9%) and Eukaryota (Sheep: 0.23%, Cow: 1.3%, Reindeer: 1.8%, Red deer: 0.56%). Eukaryota reads

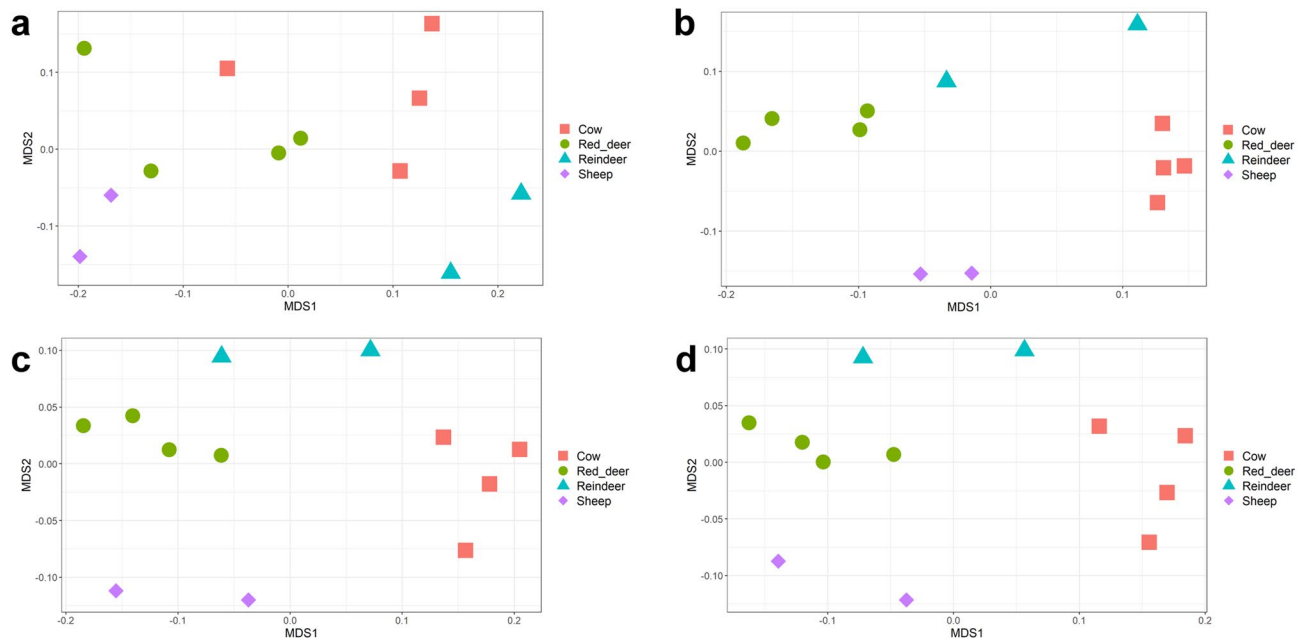


Figure 4. NMDS of ruminal samples clustered by abundance of taxonomies, using Bray–Curtis dissimilarity values. (a) Kingdom (PERMANOVA; $P=0.01058$), (b) Phylum (PERMANOVA; $P=0.00017$), (c) Family (PERMANOVA; $P=1e-05$), (d) Genus (PERMANOVA; $P=3e-05$).

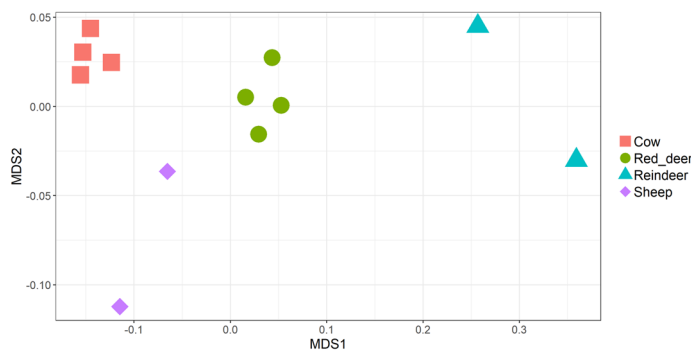


Figure 5. NMDS of ruminal samples clustered by abundance of CAZymes, using Bray–Curtis dissimilarity values (PERMANOVA; $P=1e-05$).

originated primarily from fungi and protists. In all ruminants, *Bacteroidetes* was the most abundant phylum (Sheep: 64%, Cow: 65% Reindeer: 54% Red deer: 52%), with *Firmicutes* being the second most abundant (Sheep: 29%, Cow: 26% Reindeer: 26% Red deer: 38%). Using PERMANOVA, significant differences in the abundance of taxonomies between ruminant species were found at both high (Kingdom: $P=0.01058$, Phylum: $P=0.00017$) and low (Family: $P=1e-05$, Genus: $P=3e-05$) taxonomic levels (Fig. 4).

We also compared the abundance of genes encoding for specific CAZymes between species. These enzymes are responsible for the synthesis, binding and metabolism of carbohydrates. The carbohydrate esterases (CEs), glycoside hydrolases (GHs), glycosyltransferases (GTs) and polysaccharide lyases (PLs) act to degrade cellulose, hemicellulose and other carbohydrates which could otherwise not be digested by the host. Non-catalytic carbohydrate-binding modules (CBMs) bind to specific carbohydrates, increasing the efficiency of enzymatic degradation²⁹. The auxiliary activities (AAs) redox enzymes are reclassified CBMs which are lytic polysaccharide monoxygenases³⁰. In our samples we found the following numbers of these CAZyme families: 6 AAs redox enzymes, 39 CBMs, 14 CEs, 191 GHs, 61 GTs and 27 PLs. The ten most abundant GHs in the different ruminant species were: for cows GH2, GH3, GH31, GH97, GH28, GH51, GH43_10, GH105, GH10 and GH95; for sheep GH2, GH3, GH28, GH31, GH97, GH32, GH51, GH77, GH78 and GH95; for red deer GH2, GH3, GH31, GH97, GH77, GH32, GH51, GH109, GH28 and GH78; and for reindeer GH2, GH3, GH92, GH109, GH97, GH13, GH31, GH78, GH28 and GH77. Different ruminant species were found to have significantly differently abundant CAZyme genes (PERMANOVA: $P=1e-05$, Fig. 5). However, it should be noted that the vast majority of CAZyme families were found in all sample types (Fig. 6), indicating that there exists a set of CAZymes which are present across ruminant species consuming different diets and living in vastly different conditions.

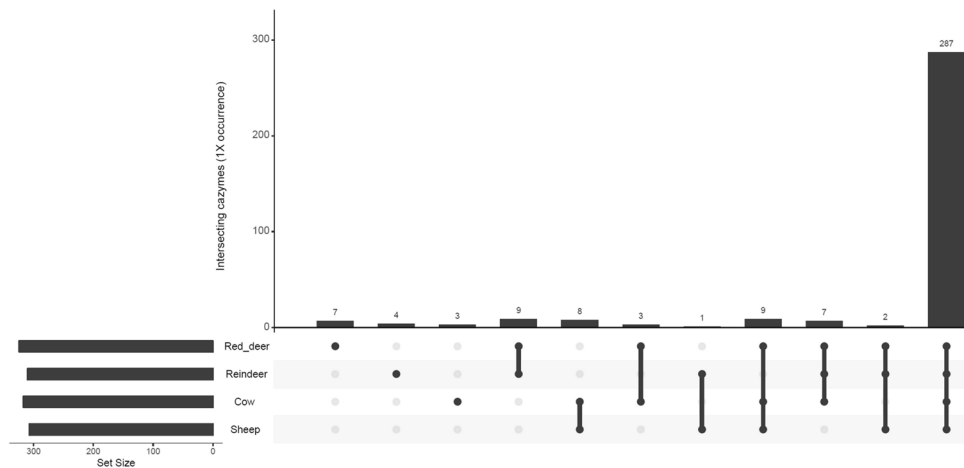


Figure 6. UpSetR graph showing the number of shared CAZyme families at average $\times 1$ coverage within four ruminant species.

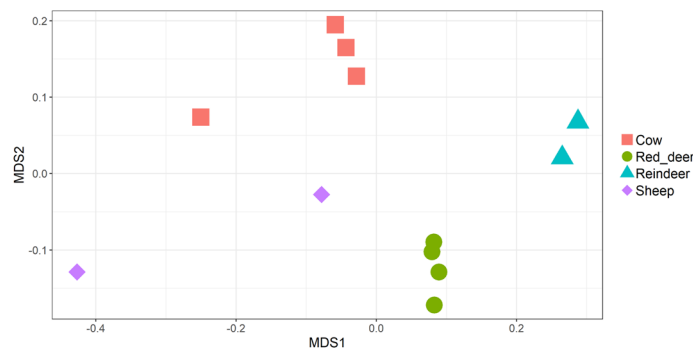


Figure 7. NMDS of ruminal samples clustered by abundance of KEGG orthologs, using Bray–Curtis dissimilarity values (PERMANOVA; $P = 1e-05$).

DeSeq2 was used to identify specific CAZymes which were significantly more abundant in one ruminant species versus another (Supplementary data S3). Those CAZymes which were consistently more abundant in specific species when compared to other species are listed in Supplementary tables S1–S4.

CAZymes are often found organised into Polysaccharide Utilization Loci (PUL) which comprise a set of genes that enable the binding and degradation of specific carbohydrates or multiple carbohydrates. We used the software PULpy to predict PULs which were present in our *Bacteroidales* RUGs. Of the 136 RUGs which belong to the taxonomy *Bacteroidales*, 112 contain putative PULs. Within these RUGs we identified 970 PULs, with numbers of PULs per RUG ranging from 1 to 35. The largest quantity of PULs originating from one RUG was 35 from uncultured *Bacteroidales* sp. *RUG30227*; these encoded a wide range of CAZymes. This RUG was more abundant in reindeer samples than samples from other ruminants. Of the 970 PULs, 332 of these were a single susC/D pair. A summary of identified PULs can be found in Supplementary data S4 and Supplementary fig S1.

We also examined the abundance of genes which belonged to specific KEGG orthologs. KEGG orthologs represent a wide range of molecular functions and are defined by a network-based classification. We found that, as for CAZymes, ruminant species clustered significantly by the abundance of genes with specific KEGG orthologs (PERMANOVA: $P = 1e-05$, Fig. 7) and that the vast majority of orthologs were found in all ruminant species (Fig. 8). However, the large amount of orthologs ($n = 729$) which were only found in the two domesticated species (cows and sheep) is also worthy of note. It should also be noted that the two sheep samples did not cluster visually to the same extent as the samples originating from the other ruminant species (Fig. 7). DeSeq2 was used to identify many KEGG orthologs which were significantly more abundant in one ruminant species vs another (Supplementary data S5). Those orthologs which were consistently more abundant in specific ruminant species (Adjusted p value < 0.05) are listed in Supplementary data S6.

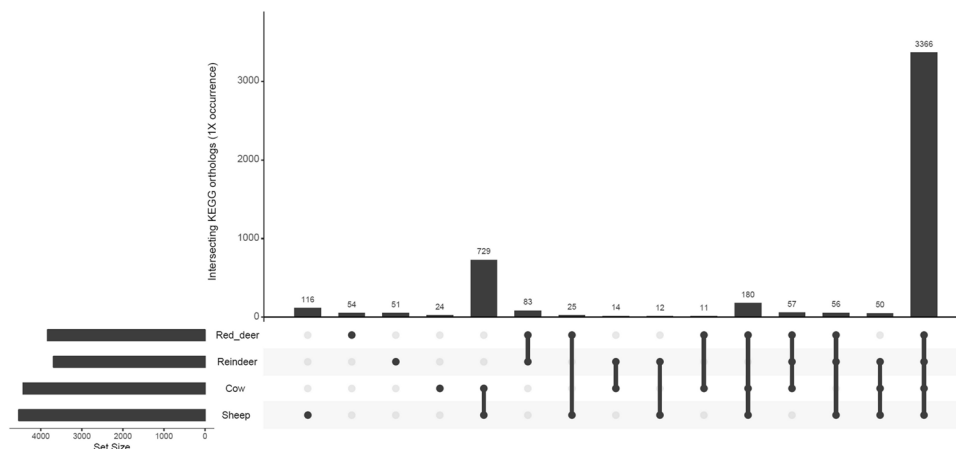


Figure 8. UpSetR graph showing the number of shared KEGG orthologs families at average $\times 1$ coverage within four ruminant species.

Discussion

The rumen microbiota plays a crucial role in the ability of ruminants to efficiently digest feed while the rumen microbiota and their products also have a potential use in diverse industrial applications. The ruminal microbiota of red deer and reindeer have previously been studied using 16S rRNA gene sequencing^{31–34}. However, metagenomic studies of these species are limited, with only one study in reindeer³⁵ and no studies in red deer.

In this study we constructed 391 rumen microbial genomes from metagenomic data from cows, sheep, red deer and reindeer. We assigned taxonomies to our RUGs using GTDB-Tk rather than NCBI based taxonomies as this improves the classification of uncultured bacteria due to the use of a genome-based taxonomy³⁶. We have also previously found less need to manually correct taxonomic assignments when using GTDB-Tk²¹. Our microbes predominantly belonged to the *Bacteroidota* and *Firmicutes_A*, with lesser numbers of 14 other phyla. We dereplicated our genomes alongside a superset of rumen bacterial genomes²⁰ and used the results output by GTDB-Tk to identify RUGs which represent novel microbial strains and species. Amongst our genomes we identified 372 novel strains and 279 novel species. These microbes were taxonomically diverse, belonging to 15 phyla. Only 31 RUGs were assigned an identity at species level.

The vast majority of our total RUGs were only present in one ruminant species. However, we found that at higher taxonomic levels taxonomies were shared between sample types. When comparing the abundance of taxonomies between samples we found that ruminant species clustered separately by both higher (kingdom and phylum) and lower (family and genus) taxonomic levels. We are aware that the sample sizes for our study are small and unequal numbers of samples were included per group, therefore any conclusions about differences between the microbiota of ruminant species should be drawn cautiously. However, our data are supportive of the hypothesis that there are host species-specific rumen microbiota at the strain and species level, potentially due to the co-evolution of the microbiome and host³⁷, but that these differences do not necessarily translate into large differences in the types of CAZymes expressed. While we found that there were significant differences between the abundances of CAZymes and KEGG orthologs between ruminant species, most CAZymes and KEGG orthologs were present in all ruminant species. These findings may indicate that while the microbial strains and species present in the rumen differ between ruminant species, these microbes perform similar metabolic roles. That function is more highly conserved than taxa across samples has also been documented in humans³⁸.

We also identified 970 PULs in our *Bacteroidales* RUGs, with numbers of PULs per genome ranging from 1 to 35. The RUG containing 35 PULs was found most abundantly in reindeer samples, emphasising the potential for the discovery of novel carbohydrate-active enzymes in lesser studied ruminant species, as also highlighted by a previous study which identified multiple PULs in metagenomic samples from reindeer³⁵. Unfortunately due to the nature of our samples, with red deer and reindeer samples originating from animals eating a non-regimented diet, we are not able to provide metadata as to the exact nutritional composition of our animals' diets, therefore a more in depth analysis of dietary carbohydrates vs CAZyme/PUL abundance is not possible.

While several thousand RUGs have previously been published that originate from the rumen microbiota, the vast majority of these originate from cows. By investing more effort in exploring the metagenome of less well studied ruminants we will be able to identify a greater diversity of microorganisms and enzymes of industrial interest. In conclusion, we present a dataset of RUGs from four ruminant species which can be used as a reference dataset in future metagenomic studies and to aid in selection of microbes in culture based studies.

Methods

Ethical approval. Cow projects were carried out under Home Office PPL 30/2579. Sheep experimentation was carried out under the conditions set out by UK Home Office licence no. 604028, procedure reference number 8. Animal experiments were assessed and approved by animal ethics committees of the University of Reading (cows) and James Hutton Institute (sheep), respectively. All methods were carried out in accordance

with the Animals (Scientific Procedures) Act of 1986. The study was carried out in compliance with the ARRIVE guidelines.

Experimental design. Reindeer (*Rangifer tarandus*: Grazing mixed vegetation, $n=2$) and red deer (*Cervus elaphus*: Grazing mixed vegetation, $n=4$) were shot in the wild, and ruminal digesta samples were collected immediately. Samples were taken from Holstein cows (*Bos Taurus*: Fed total mixed ration (once a day), $n=4$) and Finn-Dorset cross sheep (*Ovis aries*: Grazing mixed pasture, $n=2$) via a rumen cannula. Samples were taken from sheep after morning grazing. Sheep sampling was performed as described in McKain et al.³⁹. Cow samples were taken 3 h post feeding. Samples were collected from the bovine rumen in the following locations: top near cannula, middle at the front of the rumen, middle towards the back of the rumen and bottom (approximately 45 cm down from the entrance to the rumen). Digesta samples were mixed with buffer containing glycerol as a cryoprotectant³⁹. The mixtures were kept on ice for 1–2 h then frozen at -20°C . DNA extraction was performed using repeated bead beating plus column filtration, as described by Yu et al.⁴⁰. Illumina TruSeq genomic DNA libraries were constructed and shotgun sequencing was performed on an Illumina HiSeq 2000, producing an average of 1626 million paired reads per sample, of 100 bp or 150 bp in length.

Bioinformatics. Illumina adaptors were removed using trimmomatic⁴¹ (v.0.36). IDBA-UD⁴² (v.1.1.3) with the options `--num_threads 16 --pre_correction --min_contig 300` was used to perform single sample assemblies. After indexing using BWA index (v.0.7.15), BWA-MEM was used to map reads to assemblies⁴³. BAM files were created by SAM tools⁴⁴ (v.1.3.1) and coverage was calculated using the command `jgi_summarize_bam_contig_depths` from the MetaBAT2 (v.2.11.1) software package⁴⁵. A coassembly was carried out on all samples using MEGAHIT⁴⁶ (v.1.1.1) with the options `--continue --kmin-1pass -m 100e+10 --k-list 27, 37, 47, 57, 67, 77, 87 --min-contig-len 1000 -t 16`. After filtering out reads which were < 2 kb, indexing and mapping were performed as for single assemblies.

Metagenomic binning was carried out using MetaBAT2 with the options `--minContigLength 2000, --min-ContigDepth 2`. From the single-assemblies, 1691 bins were created and from the co-assembly 2508 bins were created. Completeness and contamination of bins were calculated using CheckM (options: `lineage_wf, -t 16, -x fa`) (v.1.0.5), and the bins were dereplicated using dRep⁴⁷ (options: `dereplicate_wf -p 16 -comp 80 -con 10 -str 100 -strW 0`) (v.1.1.2). Thus, bins were discarded if their completeness was $< 80\%$ or if they had contamination $> 10\%$. The dereplicated 'winning' bins are referred to below as RUGs. MAGpy⁴⁸ was used to compare the RUGs to public datasets. This Snakemake⁴⁹ pipeline uses CheckM (v.1.0.5); prodigal (v2.6.3)⁵⁰; Pfam_Scan (v.1.6)⁵¹; DIAMOND (v.0.9.22.123)⁵² searching against UniProt TrEMBL⁵³; PhyloPhlAn (v.0.99)⁵⁴ and sourmash (v.2.0.0)⁵⁵ searching against all public bacterial genomes. Taxonomies were assigned to MAGs using GTDB-Tk³⁶. Trees produced by MAGpy were rerooted at the branch between archaea and bacteria using Figtree⁵⁶ (v.1.4.4) and visualised using GraPhlAn²⁴ (v.0.9.7). For submission to public repositories, our RUGs were named as the lowest taxonomic level at which NCBI and GTDB-Tk matched. The taxonomies assigned to RUGs were manually checked against the taxonomic tree and improved accordingly.

Carbohydrate-active enzymes (CAZymes) were identified using dbCAN2 (version 7, 24th August 2018) by comparing RUG proteins to the CAZy database⁵⁷. RUG proteins were compared to the KEGG database (downloaded on Sept 15th 2018)^{58–60} using DIAMOND (v.0.9.21). KEGG hits for which the alignment length was $\geq 90\%$ of the query length were retained. The likely KEGG ortholog group for each RUG protein was inferred from the DIAMOND search results and the KEGG database. CAZyme and KEGG ortholog abundances were calculated as the sum of the reads mapping to RUG proteins within each group after using DIAMOND to align reads to the RUG proteins. PULpy was used to identify polysaccharide utilisation loci⁶¹.

Statistics and reproducibility. Statistical analyses were carried out within R (version 3.5.1). The ggplot2⁶² package was used to construct scatter plots and NMDS graphs. The vegan package⁶³ was used to create NMDS axes using the Bray–Curtis dissimilarity. The Adonis function from the vegan package was used to perform PERMANOVA analyses and DeSeq2⁶⁴ was used to calculate differences in coverage for individual CAZymes, KEGG orthologs and RUGs. UpSet graphs were constructed using the UpSetR package⁶⁵. Taxonomies were assigned to paired sequence reads with Kraken⁶⁶ using a custom kraken database consisting of RefSeq complete genomes with our RUGs and the rumen superset²⁰ added. Prior to statistical analyses (excluding DeSeq2) and graph construction, data was subsampled. For RUGs, subsampling to the lowest sample coverage was performed. CAZymes and KEGG orthologs were subsampled to the lowest sample abundance. Significant P values were defined as $P < 0.05$.

Data availability

The paired-read fastq files supporting the conclusions of this article are available in the European Nucleotide Archive repository (<https://www.ebi.ac.uk/ena/browser/view/PRJEB34458>). The RUG fasta files supporting the conclusions of this article are available in the Edinburgh DataShare repository (<https://doi.org/10.7488/ds/2640>).

Received: 21 September 2020; Accepted: 15 December 2020

Published online: 21 January 2021

References

1. Lima, J. et al. Identification of rumen microbial genes involved in pathways linked to appetite, growth, and feed conversion efficiency in cattle. *Front. Genet.* **10**, 701. <https://doi.org/10.3389/fgene.2019.00701> (2019).

2. Ben Shabat, S. K. *et al.* Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. *Isme J.* **10**, 2958–2972. <https://doi.org/10.1038/ismej.2016.62> (2016).
3. Patil, R. D. *et al.* Poor feed efficiency in sheep is associated with several structural abnormalities in the community metabolic network of their ruminal microbes. *J. Anim. Sci.* **96**, 2113–2124. <https://doi.org/10.1093/jas/sky096> (2018).
4. Jami, E., White, B. A. & Mizrahi, I. Potential role of the bovine rumen microbiome in modulating milk composition and feed efficiency. *PLoS ONE* **9**, e85423. <https://doi.org/10.1371/journal.pone.0085423> (2014).
5. Scharen, M. *et al.* Interrelations between the rumen microbiota and production, behavioral, rumen fermentation, metabolic, and immunological attributes of dairy cows. *J. Dairy Sci.* **101**, 4615–4637. <https://doi.org/10.3168/jds.2017-13736> (2018).
6. Ribeiro, G. O., Gruninger, R. J., Badhan, A. & McAllister, T. A. Mining the rumen for fibrolytic feed enzymes. *Anim. Front.* **6**, 20–26. <https://doi.org/10.2527/af.2016-0019> (2016).
7. Huws, S. A. *et al.* Addressing global ruminant agricultural challenges through understanding the rumen microbiome: Past, present, and future. *Front. Microbiol.* **9**, 33. <https://doi.org/10.3389/fmicb.2018.02161> (2018).
8. Solbak, A. I. *et al.* Discovery of pectin-degrading enzymes and directed evolution of a novel pectate lyase for processing cotton fabric. *J. Biol. Chem.* **280**, 9431–9438. <https://doi.org/10.1074/jbc.M411838200> (2005).
9. Singh, B., Gautam, S. K., Verma, V., Kumar, M. & Singh, B. Metagenomics in animal gastrointestinal ecosystem: Potential biotechnological prospects. *Anaerobe* **14**, 138–144. <https://doi.org/10.1016/j.anaerobe.2008.03.002> (2008).
10. Ufarte, L. *et al.* Discovery of carbamate degrading enzymes by functional metagenomics. *PLoS ONE* **12**, e0189201. <https://doi.org/10.1371/journal.pone.0189201> (2017).
11. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467. <https://doi.org/10.1126/science.1200387> (2011).
12. Wallace, R. J. *et al.* The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics* **16**, 839. <https://doi.org/10.1186/s12864-015-2032-0> (2015).
13. Difford, G. F. *et al.* Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLoS Genet.* **14**, e1007580. <https://doi.org/10.1371/journal.pgen.1007580> (2018).
14. Shi, W. B. *et al.* Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res.* **24**, 1517–1525. <https://doi.org/10.1101/gr.168245.113> (2014).
15. Auffret, M. D. *et al.* Identification, comparison, and validation of robust rumen microbial biomarkers for methane emissions using diverse *Bos taurus* breeds and basal diets. *Front. Microbiol.* **8**, 2642. <https://doi.org/10.3389/fmicb.2017.02642> (2018).
16. Roehe, R. *et al.* Bovine host genetic variation influences rumen microbial methane production with best selection criterion for low methane emitting and efficiently feed converting hosts based on metagenomic gene abundance. *PLoS Genet.* **12**, 20. <https://doi.org/10.1371/journal.pgen.1005846> (2016).
17. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367. <https://doi.org/10.1038/nbt.4110> (2018).
18. Creevey, C. J., Kelly, W. J., Henderson, G. & Leahy, S. C. Determining the culturability of the rumen bacterial microbiome. *Microb. Biotechnol.* **7**, 467–479. <https://doi.org/10.1111/1751-7915.12141> (2014).
19. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870. <https://doi.org/10.1038/s41467-018-03317-6> (2018).
20. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961. <https://doi.org/10.1038/s41587-019-0202-3> (2019).
21. Glendinning, L., Stewart, R. D., Pallen, M. J., Watson, K. A. & Watson, M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. *Genome Biol.* **21**, 16. <https://doi.org/10.1186/s13059-020-1947-1> (2020).
22. 22Ishaq, S. & Wright, A.-D. G. in *Encyclopedia of Metagenomics: Environmental Metagenomics* (eds Sarah K. Highlander, Francisco Rodriguez-Valera, & Bryan A. White) 686–693 (Springer, New York, 2015).
23. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731. <https://doi.org/10.1038/nbt.3893> (2017).
24. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029. <https://doi.org/10.7717/peerj.1029> (2015).
25. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542. <https://doi.org/10.1038/s41564-017-0012-7> (2017).
26. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284. <https://doi.org/10.1038/s41564-018-0225-4> (2018).
27. Svartstrom, O. *et al.* Ninety-nine de novo assembled genomes from the moose (*Alces alces*) rumen microbiome provide new insights into microbial plant biomass degradation. *Isme J.* **11**, 2538–2551. <https://doi.org/10.1038/ismej.2017.108> (2017).
28. Neumann, A. P. & Suen, G. The phylogenomic diversity of herbivore-associated *Fibrobacter* spp. is correlated to lignocellulose-degrading potential. *mSphere* **3**, e00593–e00518. <https://doi.org/10.1128/mSphere.00593-18> (2018).
29. Shoseyov, O., Shani, Z. & Levy, I. Carbohydrate binding modules: Biochemical properties and novel applications. *Microbiol. Mol. Biol. Rev.* **70**, 283–295. <https://doi.org/10.1128/mmb.00028-05> (2006).
30. Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels* **6**, 41. <https://doi.org/10.1186/1754-6834-6-41> (2013).
31. Ostbye, K., Wilson, R. & Rudi, K. Rumen microbiota for wild boreal cervids living in the same habitat. *FEMS Microbiol. Lett.* **363**, 233. <https://doi.org/10.1093/femsle/fnw233> (2016).
32. Salgado-Flores, A. *et al.* Rumen and cecum microbiomes in reindeer (*Rangifer tarandus tarandus*) are changed in response to a lichen diet and may affect enteric methane emissions. *PLoS ONE* **11**, e0155213. <https://doi.org/10.1371/journal.pone.0155213> (2016).
33. Sundset, M. *et al.* Molecular diversity of the rumen microbiome of Norwegian reindeer on natural summer pasture. *Microb. Ecol.* **57**, 335–348. <https://doi.org/10.1007/s00248-008-9414-7> (2009).
34. Qian, W. X., Ao, W. P., Jia, C. H. & Li, Z. P. Bacterial colonisation of reeds and cottonseed hulls in the rumen of Tarim red deer (*Cervus elaphus yarkandensis*). *Antonie Van Leeuwenhoek* **112**, 1283–1296. <https://doi.org/10.1007/s10482-019-01260-0> (2019).
35. Pope, P. B. *et al.* Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci. *PLoS ONE* **7**, e38571. <https://doi.org/10.1371/journal.pone.0038571> (2012).
36. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004. <https://doi.org/10.1038/nbt.4229> (2018).
37. Koskella, B. & Bergelson, J. The study of host-microbiome (co)evolution across levels of selection. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 8. <https://doi.org/10.1098/rstb.2019.0604> (2020).
38. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214. <https://doi.org/10.1038/nature11234> (2012).
39. McKain, N., Genc, B., Snelling, T. J. & Wallace, R. J. Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. *J. Microbiol. Methods* **95**, 381–383. <https://doi.org/10.1016/j.mimet.2013.10.009> (2013).
40. Yu, Z. T. & Morrison, M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques* **36**, 808–812. <https://doi.org/10.2144/04365st04> (2004).

41. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170> (2014).
42. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428. <https://doi.org/10.1093/bioinformatics/bts174> (2012).
43. Li, H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM* (2013).
44. Li, H. *et al.* The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
45. Kang, D. W. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, 15. <https://doi.org/10.7717/peerj.1165> (2015).
46. Li, D. H., Liu, C. M., Luo, R. B., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676. <https://doi.org/10.1093/bioinformatics/btv033> (2015).
47. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868. <https://doi.org/10.1038/ismej.2017.126> (2017).
48. Stewart, R. D., Watson, M., Auffret, M. D., Roehe, R. & Snelling, T. J. MAGpy: A reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). *Bioinformatics* **35**, 2150–2152. <https://doi.org/10.1093/bioinformatics/bty905> (2018).
49. Koster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600–3600. <https://doi.org/10.1093/bioinformatics/bty350> (2018).
50. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 11. <https://doi.org/10.1186/1471-2105-11-119> (2010).
51. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **42**, D222–D230. <https://doi.org/10.1093/nar/gkt1223> (2014).
52. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60. <https://doi.org/10.1038/nmeth.3176> (2015).
53. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169. <https://doi.org/10.1093/nar/gkw1099> (2017).
54. Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 11. <https://doi.org/10.1038/ncomms3304> (2013).
55. Brown, C. T. & Irber, L. Sourmash: A library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
56. FigTree v1.4 (2012).
57. Cantarel, B. L. *et al.* The carbohydrate-active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233–D238. <https://doi.org/10.1093/nar/gkn663> (2009).
58. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. <https://doi.org/10.1093/nar/28.1.27> (2000).
59. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951. <https://doi.org/10.1002/pro.3715> (2019).
60. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkaa970> (2020).
61. Stewart, R. D., Auffret, M. D., Roehe, R. & Watson, M. Open prediction of polysaccharide utilisation loci (PUL) in 5414 public *Bacteroidetes* genomes using PULpy. *bioRxiv* <https://doi.org/10.1101/421024> (2018).
62. ggplot2: elegant graphics for data analysis (2016).
63. vegan: Community Ecology Package (2018).
64. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. <https://doi.org/10.1186/s13059-014-0550-8> (2014).
65. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940. <https://doi.org/10.1093/bioinformatics/btx364> (2017).
66. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46. <https://doi.org/10.1186/gb-2014-15-3-r46> (2014).

Acknowledgements

BG thanks the Higher Education Council of the Turkish Republic for the award of a travelling fellowship. We thank Bob Mayes and Dave Hamilton of the James Hutton Institute for their permission and help in sampling the sheep digesta, Kevin Shingfield for the provision of reindeer digesta samples, and Euan Munro for red deer digesta samples. We also thank Nest McKain of RI for technical assistance. The Roslin Institute forms part of the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council, including institute strategic programme and national capability awards to The Roslin Institute (BBSRC: BB/P013759/1, BB/P013732/1, BB/J004235/1, BB/J004243/1), and the Technology Strategy Board (TS/J000108/1, TS/J000116/1). The Rowett Institute is funded by the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish Government. The funding bodies had no role in the study design, the collection, analysis, and interpretation of data or the writing of the manuscript. Sequencing was carried out by Edinburgh Genomics.

Author contributions

L.G. contributed to methodology, formal analysis, data curation, writing (original draft preparation) and visualisation. B.G. and R.J.W. contributed to conceptualisation, methodology, investigation, resources and writing (review and editing). R.J.W. also contributed to supervision and project administration. M.W. contributed to conceptualisation, methodology, formal analysis, writing (review and editing), visualisation, supervision and project administration. All authors 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/s41598-021-81668-9>.

Correspondence and requests for materials should be addressed to L.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021