nature biotechnology

RESOURCE

https://doi.org/10.1038/s41587-019-0202-3

OPEN

Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery

Robert D. Stewart[1], Marc D. Auffret[2], Amanda Warr[1], Alan W. Walker[3], Rainer Roehe[2] and Mick Watson[1*]

Ruminants provide essential nutrition for billions of people worldwide. The rumen is a specialized stomach that is adapted to the breakdown of plant-derived complex polysaccharides. The genomes of the rumen microbiota encode thousands of enzymes adapted to digestion of the plant matter that dominates the ruminant diet. We assembled 4,941 rumen microbial metagenome-assembled genomes (MAGs) using approximately 6.5 terabases of short- and long-read sequence data from 283 ruminant cattle. We present a genome-resolved metagenomics workflow that enabled assembly of bacterial and archaeal genomes that were at least 80% complete. Of note, we obtained three single-contig, whole-chromosome assemblies of rumen bacteria, two of which represent previously unknown rumen species, assembled from long-read data. Using our rumen genome collection we predicted and annotated a large set of rumen proteins. Our set of rumen MAGs increases the rate of mapping of rumen metagenomic sequencing reads from 15% to 50–70%. These genomic and protein resources will enable a better understanding of the structure and functions of the rumen microbiota.

Ruminants convert human-inedible, low-value plant biomass into products of high nutritional value, such as meat and dairy products. The rumen, which is the first of four chambers of the stomach, contains a mixture of bacteria, archaea, fungi and protozoa that ferment complex carbohydrates, including lignocellulose and cellulose, to produce short-chain fatty acids (SCFAs) that the ruminant uses for homeostasis and growth. Rumen microbes are a rich source of enzymes for plant biomass degradation for use in biofuel production[1–3], and manipulation of the rumen microbiome offers opportunities to reduce the cost of food production[4].

Ruminants are important for both food security and climate change. For example, methane is a byproduct of ruminant fermentation, released by methanogenic archaea, and an estimated 14% of methane produced by humans has been attributed to ruminant livestock[5]. Methane production has been directly linked to the abundance of methanogenic archaea in the rumen[6], offering possibilities for mitigating this issue through selection[7] or manipulation of the microbiome. Two studies have reported large collections of rumen microbial genomes. Stewart et al. assembled 913 draft MAGs (named rumen-uncultured genomes (RUGs)) from the rumens of 43 cattle raised in Scotland[8], and Seshadri et al. reported 410 reference archaeal and bacterial genomes from the Hungate collection[9]. As isolate genomes, the Hungate genomes are generally higher quality and, crucially, the corresponding organisms exist in culture and so can be grown and studied in the lab. However, we found that addition of the Hungate genomes increased read classification by only 10%, as compared to an increase of 50–70% when the RUGs were used, indicating large numbers of undiscovered microbes in the rumen.

We present a comprehensive analysis of more than 6.5 terabases of sequence data from the rumens of 283 cattle. Our catalog of rumen genomes (named RUG2) includes 4,056 genomes that were not present in Stewart et al.[8], and brings the number of rumen genomes assembled to date to 5,845. We also present a metagenomic assembly of nanopore (MinION) sequencing data (from one rumen sample) that contains at least three whole bacterial chromosomes as single contigs. These genomic and protein resources will underpin future studies on the structure and function of the rumen microbiome.

Results

Metagenome-assembled genomes from the cattle rumen. We sequenced DNA extracted from the rumen contents of 283 beef cattle (characteristics of the animals sequenced are in Supplementary Data 1), producing over 6.5 terabytes of Illumina sequence data. We operated a continuous assembly-and-dereplication pipeline, which means that newer genomes of the same strain (>99% average nucleotide identity (ANI)) replaced older genomes if their completeness and contamination statistics were better. All 4,941 RUGs we present here have completeness ≥80% and contamination ≤10% (Supplementary Fig. 1).

All the RUGs were analyzed using MAGpy[10] and their assembly characteristics, putative names and taxonomic classifications are given in Supplementary Data 2. Sourmash[11], DIAMOND[12] and PhyloPhlAn[13] outputs, which reveal genomic and proteomic similarity to existing public data, are given in Supplementary Data 3. A phylogenetic tree of the 4,941 RUGs, alongside 460 public genomes from the Hungate collection, is presented in Fig. 1 and Supplementary Data 4. The tree is dominated by large numbers of genomes from the Firmicutes and Bacteroidetes phyla (dominated by Clostridiales and Bacteroidales, respectively), but also contains many new genomes from the Actinobacteria, Fibrobacteres and Proteobacteria phyla. Clostridiales (2,079) and Bacteroidales (1,081) are the dominant orders, with Ruminoccocacae (1,111) and

[1]The Roslin Institute and the Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, UK. [2]Scotland's Rural College, Edinburgh, UK. [3]The Rowett Institute, University of Aberdeen, Aberdeen, UK. *e-mail: mick.watson@roslin.ed.ac.uk

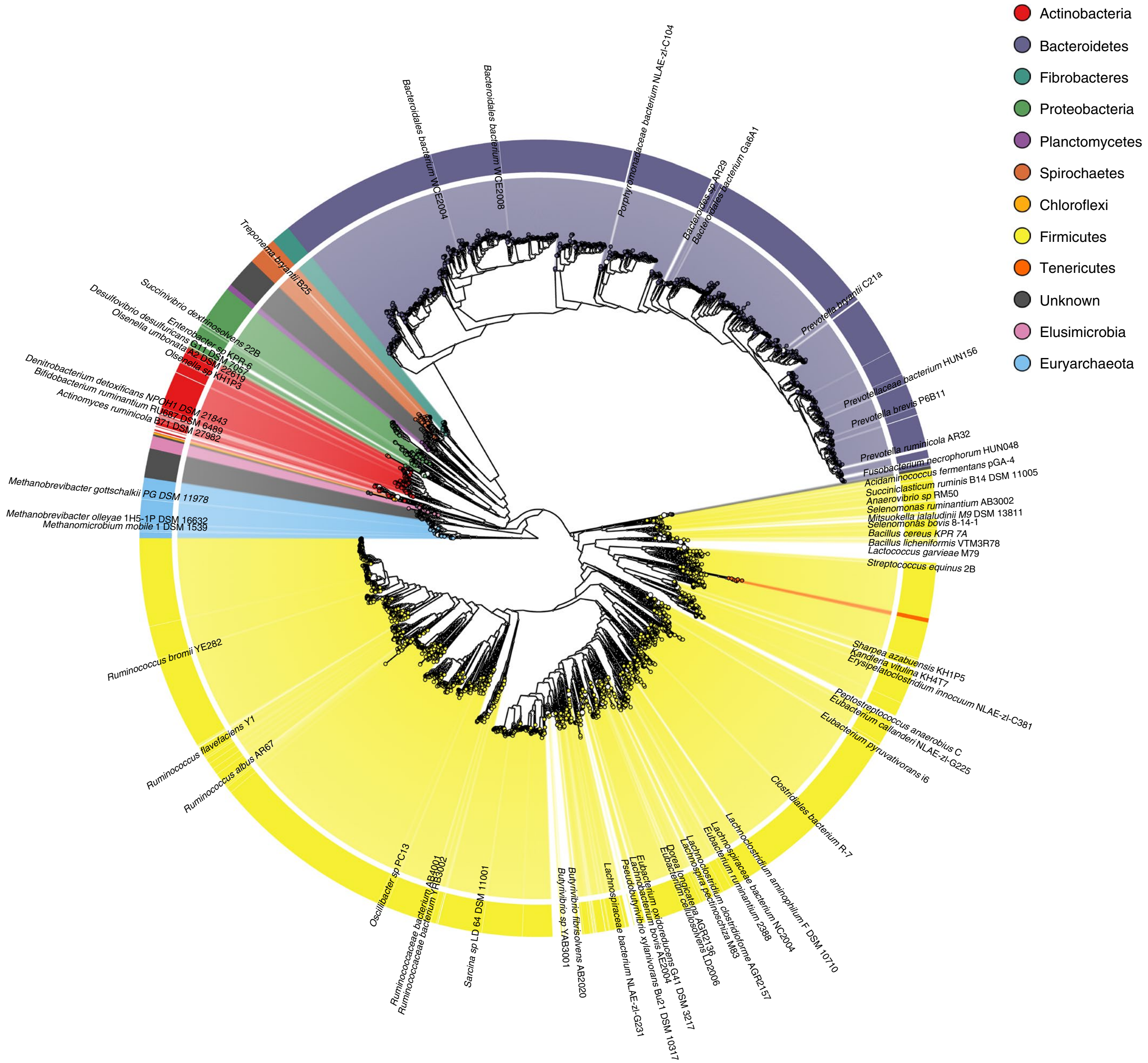
Actinobacteria
Bacteroidetes
Fibrobacteres
Proteobacteria
Planctomycetes
Spirochaetes
Chloroflexi
Firmicutes
Tenericutes
Unknown
Elusimicrobia
Euryarchaeota
Bacteroidales bacterium WCE2004
Bacteroidales bacterium WCE2008
Porphyromonadaceae bacterium NLAE-zl-C104
Bacteroides sp AR29
Bacteroidales bacterium Ga6A1
Prevotella bryantii C21a
Prevotellaceae bacterium HUN156
Prevotella brevis P6B11
Prevotella ruminicola AR32
Fusobacterium necrophorum HUN048
Acidaminococcus fermentans pGA-4
Succiniclasticum ruminis B14 DSM 11005
Anaerovibrio sp RM50
Selenomonas ruminantium AB3002
Mitsuokella jalaludinii M9 DSM 13811
Selenomonas bovis 8-14-1
Bacillus cereus KPR 7A
Bacillus licheniformis VTM3R78
Lactococcus garvieae M79
Streptococcus equinus 2B
Sharpea azabuensis KH1P5
Kandleria vitulina KH4T7
Erysipelatoclostridium innocuum NLAE-zl-C381
Peptostreptococcus anaerobius C
Eubacterium callanderi NLAE-zl-G225
Eubacterium pyruvativorans i6
Clostridiales bacterium R-7
Lachnoclostridium aminophilum F DSM 10710
Lachnospiraceae bacterium NC2004
Eubacterium ruminantium 2388
Lachnoclostridium clostridioforme AGR2157
Lachnospira pectinoschiza M83
Eubacterium cellulosolvens LD2006
Dorea longicatena AGR2136
Eubacterium oxidoreducens G41 DSM 3217
Lachnobacterium bovis AE2004
Pseudobutyrivibrio xylanivorans Bu21 DSM 10317
Lachnospiraceae bacterium NLAE-zl-G231
Butyrivibrio fibrisolvens AB2020
Butyrivibrio sp YAB3001
Sarcina sp LD 64 DSM 11001
Ruminococcaceae bacterium AB4001
Ruminococcaceae bacterium YRB3002
Oscillibacter sp PC13
Ruminococcus albus AR67
Ruminococcus flavefaciens Y1
Ruminococcus bromii YE282
Methanomicrobium mobile 1 DSM 1539
Methanobrevibacter olleyae 1H5-1P DSM 16632
Methanobrevibacter gottschalkii PG DSM 11978
Actinomyces ruminicola B71 DSM 27982
Bifidobacterium ruminantium RU687 DSM 6489
Denitrobacterium detoxificans NPOH1 DSM 21843
Olsenella sp KH1P3
Olsenella umbonata A2 DSM 22619
Desulfovibrio desulfuricans G11 DSM 7057
Enterobacter sp KPR-6
Succinivibrio dextrinosolvens 22B
Treponema bryantii B25

Fig. 1 | Phylogenetic tree of 4,941 RUGs from the cattle rumen, additionally incorporating rumen genomes from the Hungate collection. The tree was produced from concatenated protein sequences using PhyloPhlAn[13], and subsequently drawn using GraPhlAn[45]. Labels show Hungate genome names, and were chosen to be informative but not overlap.

Lachnospiraceae (640) constituting the dominant families within Clostridiales and Prevotellaceae (521) consituting the dominant family within Bacteroidales.

The Genome Taxonomy Database (GTDB) proposed a new bacterial taxonomy based on conserved concatenated protein sequences[14], and we include the GTDB-predicted taxa for all RUGs (Supplementary Data 3). A total of 4,763 RUGs had <99% ANI with existing genomes, and 3,535 had <95% ANI with existing genomes and therefore represent potential new species.

Of the 4,941 genomes, 144, were classified to the species level, 1,092 were classified to the genus level, 3,188 were classified to the family level, 4,084 were classified to the order level, 4,514 were classified to the class level, 4,801 were classified to the phylum level and 4,941 were classified to the kingdom level. Of the genomes classified at the species level, 43 represented genomes derived from uncultured strains of *Ruminococcus flavefaciens*, 42 represented genomes from uncultured strains of *Fibrobacter succinogenes*, 18 represented genomes from uncultured strains of *Sharpea azabuensis* and 10 represented genomes from uncultured strains of *Selenomonas ruminantium*. These species belong to genera known to play an important role in rumen homeostasis[15].

We assembled 126 archaeal genomes, 111 of which were species of *Methanobrevibacter*. There are two other members of the Methanobacteriaceae family, which were both predicted to be members of the *Methanosphaera* genus by GTDB. Nine of the archaeal RUGs had sourmash hits to *Candidatus* Methanomethylophilus sp. 1R26; a further three had weak sourmash hits to Methanogenic archaeon ISO4-H5; and the remaining archaeal genome had no sourmash hits, and weaker DIAMOND hits to the same genome (Methanogenic archaeon ISO4-H5). All 13 were predicted to be members of the genus *Candidatus* Methanomethylophilus by GTDB, but this is based on similarity to only two genomes, both of which have uncertain phylogenetic lineages. If *Candidatus* Methanomethylophilus is a true genus, then our dataset increases the number of sequenced genomes from 2 to 15.

Genome quality statistics were measured by analyzing single-copy core genes (Supplementary Fig. 1). There are different standards for the definition of MAG quality. Bowers et al.[16] describe high-quality drafts as having ≥90% completeness and ≤5% contamination; 2,417 of the RUGs met these criteria. Alternatively, Parks et al.[17] define a quality score as completeness − (5 × contamination) and exclude any MAG with a score less than 50; 4,761 of the RUGs met this criterion, although, whilst the MAGs from Parks et al. could have completeness as low as 50%, the genomes presented here were all ≥80% complete. The RUGs ranged in size from 456 kilobases (kb) to 6.6 megabases (Mb), with N50 values (50% of assembled bases in contigs larger than the N50 value) ranging from 4.5 kb to 1.37 Mb. The average number of tRNA genes per RUG was 16.9, and 446 of the RUGs had all 20. As assemblies of Illumina metagenomes struggle to assemble repetitive regions, most of the RUGs did not contain a 16S rRNA gene—464 RUGs encoded a fragment of the 16S rRNA gene, and 154 encoded at least one full-length 16S rRNA gene.

The coverage of each RUG in each sample is provided in Supplementary Data 5. Using a cut-off of 1× coverage, most RUGs (4,863) were present in more than one animal, 3,937 RUGs were present in more than ten animals and 225 RUGs were present in more than 200 animals. One RUG was present in all animals, RUG11026, which was a member of the Prevotellaceae family.

A near-complete single-contig Proteobacteria genome. Metagenomic assembly of Illumina data often results in highly fragmented assemblies, but RUG14498, an uncultured Proteobacteria species (genome completeness 87.91% and contamination 0%), had 136 of 147 single-copy genes present with no duplications in a single contig of just over 1 Mb in size. Proteobacteria with small genomes (<1.5 Mb in size) were relatively common ($n=67$) in our dataset and have also been found in other large metagenome assembly projects[17]. The Proteobacteria genomes we present encode proteins with only 45–60% amino acid identity with proteins in UniProt TREMBL[18]. We compared our single-contig Proteobacteria assembly with nine Proteobacteria with similarly sized genomes assembled by Parks et al.[17] (Supplementary Fig. 2). ANI, which is often used to delineate new strains and species, between the nine UBA genomes and RUG14498 was revealing. UBA2136, UBA1908, UBA3307, UBA3773 and UBA3768 had no detectable level of identity with any other genome in the set; UBA4623, UBA6376, UBA6864 and UBA6830 all had greater than 99.4% ANI with one another, indicating that they are highly similar strains of the same species. UBA4623, UBA6376, UBA6864 and UBA6830 also had around 77.8% ANI with RUG14498, suggesting that the single-contig RUG14498 is a high-quality, near-complete whole genome of a new Proteobacteria species. The single-contig RUG14498 was assembled by IDBA_ud from sample 10678_020. IDBA_ud exploits uneven depth in metagenomic samples to improve assemblies. RUG14498 was the tenth most abundant genome in 10678_020, and other genomes of similar depth in that sample were taxonomically unrelated, enabling IDBA_ud to assemble almost the entire genome in a single contig.

RUG14498 had a single full-length 16S rRNA gene (1,507 base pairs). The top hit in GenBank (97% identity across 99% of the length) was accession AB824499.1, a sequence from an uncultured bacterium from the rumen of Thai native cattle and swamp buffaloes. The top hit in SILVA[19] was to the same sequence, only this time annotated as an uncultured *Rhodospirillales*. Together, these results support the conclusion that RUG14498 represents a new Proteobacteria species. Low amino acid identity to known proteins limited our ability to predict function and metabolic activity; nevertheless, RUG14498 encodes 73 predicted CAZymes, including 42 glycosyl transferases and 19 glycosyl hydrolases, suggesting a role in carbohydrate synthesis and metabolism.

New microbial genomes from the rumen microbiome. We compared the 4,941 RUGs to the Hungate collection and to our previous dataset[8] (Fig. 2). Of the 4,941 RUGs, 149 had >95% protein identity with Hungate members and 271 had >90%; this left 4,670 RUGs with <90% protein identity with Hungate members. Of the 4,941 RUGs, 2,387 had <90% protein identity with genomes in Stewart et al., and more than 1,100 RUGs had <70% protein identity with genomes in Stewart et al. Many of the RUGs with the lowest protein identity to publicly available genomes could not be classified beyond the phylum level, and some are classified as simply uncultured bacterium.

We compiled a database comprising all RUG genomes, the Hungate collection genomes[9] and rumen MAGs from Hess et al.[1], Parks et al.[17], Solden et al.[20] and Svartström et al.[21] that we name the rumen superset. The rumen superset was dereplicated at both 99% (strain level) and 95% (species level) ANI. At 95% ANI, the rumen superset was reduced to 2,690 clusters, representing species-level bins. Of these clusters, 2,078 contained only RUG genomes, and therefore represent putative new rumen microbial species identified in this study. Fifty-eight clusters contained both Hungate and RUG genomes, and 268 clusters contained only Hungate genomes (Supplementary Data 6). At 99% ANI, the rumen superset was reduced to 5,574 clusters, representing strain-level bins. Of these clusters, 4,845 contained only RUG genomes, and may represent putative new rumen microbial strains (Supplementary Data 7). Supplementary Figure 3 shows how the various rumen MAG sets overlap at 95% ANI after dereplication.

We calculated an estimate of the completeness of the RUG2 dataset using the Chao 1 estimator[22] (we were only able to do this for our own dataset, as the estimate was based on the number of times species were observed at different frequencies, and we did not have these values for other datasets). Dereplicating all RUG genomes at 95% ANI gave us 2,180 species-level bins. Of these, 948 were singletons (that is, were observed exactly once), and 410 were doublets (that is, were observed exactly twice). Using the Chao 1 formula, we predicted 3,276 species, we therefore estimate that we have discovered 66.54% of the species present in our samples.

We assessed the impact of using rumen genomic data on the read classification rates of several public datasets using three databases—the first, our custom rumen kraken database, consisted of RefSeq complete genomes and the Hungate collection[23,24]; the second was the same database plus only the RUGs; and the third was the same database plus the rumen superset (which includes the RUGs). We classified the following five datasets—our own (Stewart et al.[8]), a dataset we previously published (Wallace et al.[6]), data from 14 cattle from a study on niche specialization (Rubino et al.[25]), data from a methane emissions study of sheep (Shi et al.[26]) and data from a recent metagenomic study of moose (Svartström et al.[21]) (Supplementary Fig. 4).

The classification rate was increased by using either the RUG or rumen superset database, although using the rumen superset resulted in only a marginal increase in most cases. We improved read classification rates from 15% to 70%, with more than a quarter of our samples achieving a classification rate of 80% or higher. These rates were comparable with read classification rates for the human microbiome as reported by Pasolli et al.[27].

Strain-level analysis of methane emissions in sheep. Previously Shi et al.[26] found no significant changes in microbiota community structure between low-methane-emitting (LME) and high-methane-

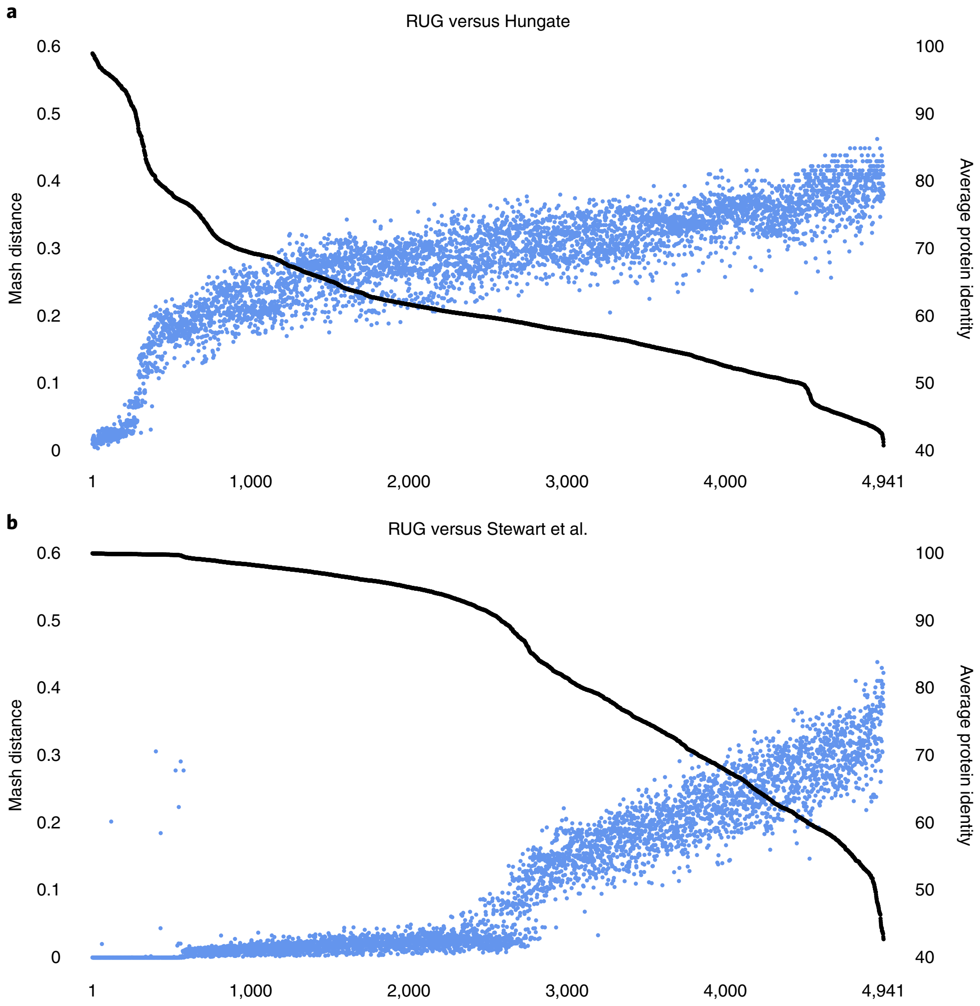

Fig. 2 | A comparison of the RUG dataset with the Hungate collection and previously published data. a,**b**, A comparison of the 4,941 RUGs with the Hungate collection (**a**) and our previously published data from Stewart et al.[8] (**b**). The black line indicates the average percentage protein identity with the closest match (right-hand *y* axis), and blue dots indicate the mash distance ($k = 100,000$) between each RUG and the closest match in the comparison dataset (a measure of dissimilarity between two DNA sequences). As expected, a high protein identity relates to a low mash distance, and vice versa. The RUGs are sorted independently by average protein identity for **a** and **b**. There is a clear inflection point in Fig. 2b, roughly half way along the *x* axis, where the protein identity dips below 90% and the mash distance rises, neatly demonstrating the novelty represented by our new larger dataset.

emitting (HME) sheep, although there were differences in gene expression between the two groups. We reanalyzed the dataset from Shi et al. using our rumen metagenomic data; specifically, we used our custom kraken database consisting of RefSeq genomes and the rumen superset to classify reads at the level of kingdom, phylum, family, genus and species, and tested differences between LME and HME sheep. While we found no significant differences at the level of kingdom, we found significant and profound differences at every other taxonomic level tested (Supplementary Tables 1–5 and Supplementary Figs. 5–9). At the genus level, *Sharpea*, *Kandleria*, *Fibrobacter* and *Selenomonas* were associated with LME sheep and *Elusimicrobium* was associated with HME sheep (Supplementary Table 4). At the species level, we found that 340 species differed significantly between LME and HME sheep (Supplementary Table 5), including 11 species of *Bifidobacterium* and 6 species of *Olsenella* that were significantly more proportionally abundant in LME sheep and 9 species of *Desulfovibrio* that were significantly more proportionally abundant in HME sheep. *Fibrobacter succinogenes*, an important rumen microbe known to be heavily involved in the degradation of plant fibers, was also significantly different between the two groups, and was associated with LME sheep. Some of these microbes were previously identified as differentially proportionally abundant between LME and HME sheep[15,28] using marker-gene sequencing, but our results provide greater resolution and reveal the genome sequences involved.

Kraken classifies data at different levels of the NCBI taxonomy; unfortunately, this does not give data on the RUGs that do not yet have specific NCBI taxonomy IDs. Therefore, to estimate the abundance of individual strains, we aligned reads directly to the rumen superset, and used the number of reads designated as primary alignments as a proxy for the relative abundance of each genome. At a false discovery rate ≤ 0.05, 1,709 genomes showed differentially proportional abundance between LME and HME sheep (Supplementary Data 8 and Supplementary Fig. 10). In Supplementary Fig. 10, LME and HME sheep are clearly separated along principal component 1, which explained 58% of the variance in the data. Supplementary Data 8 lists the differentially abundant genomes. Of note were the large numbers of previously uncharacterized Lachnospiraceae species associated with LME sheep and 22 strains of *S. azabuensis* that all had higher proportional abundance in LME sheep (all 18 *S. azabuensis* RUGs and 4 *S. azabuensis* strains from the Hungate collection). These results agree with previous studies based on

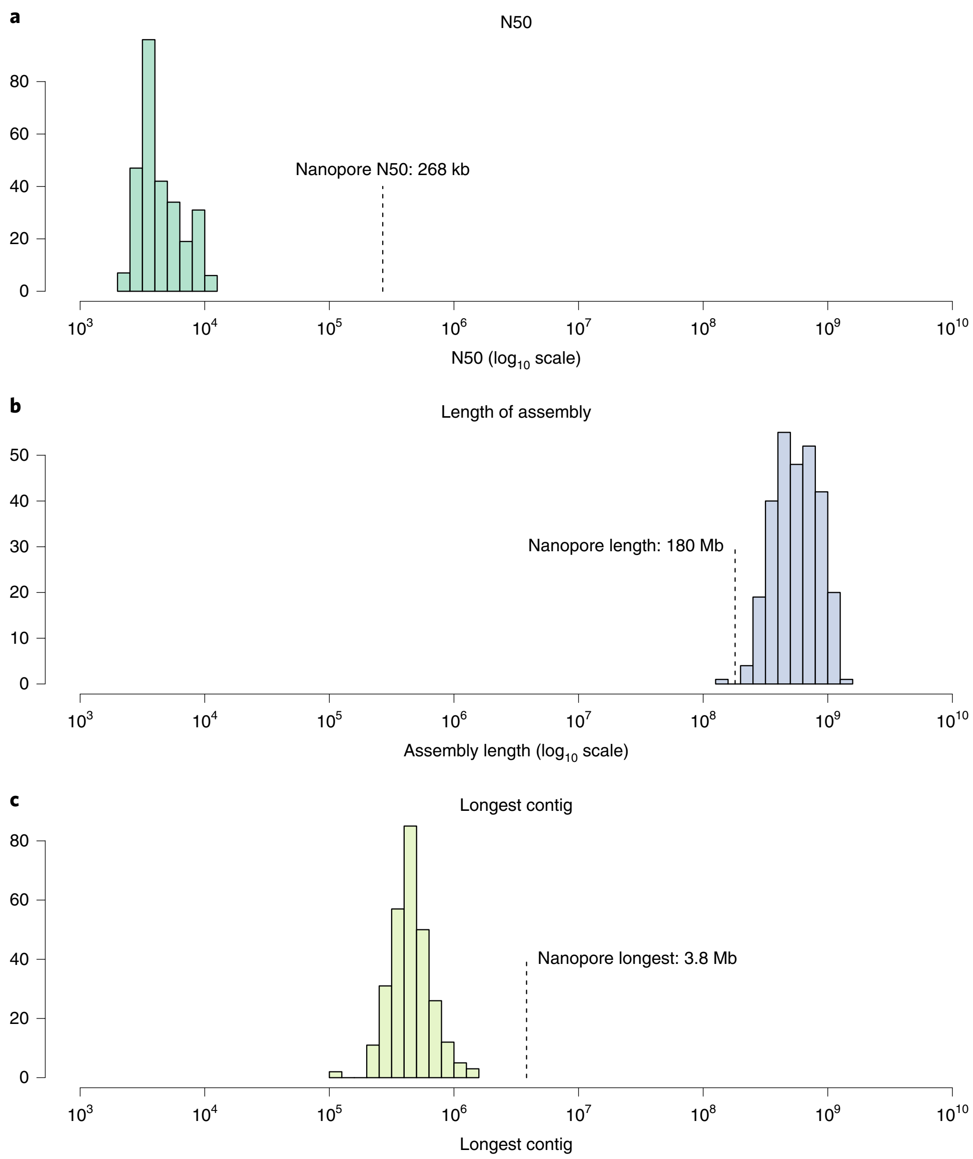

Fig. 3 | A comparison of Illumina and nanopore metagenomic assembly statistics. The colored histograms show the distribution of statistics for 282 Illumina assemblies, and the single nanopore assembly is highlighted. **a**, N50 values. **b**, Total length of the assembly. **c**, Length of the longest contig. The nanopore assembly N50 of 268 kb was over 56 times longer than that for the average Illumina assembly (4.7 kb), the Illumina assemblies were often longer (average of 600 Mb), the nanopore assembly (at 178 Mb in length) was not the shortest of the assemblies we produced and the nanopore assembly produced the longest contig at 3.8 Mb, seven times longer than the average for the Illumina assemblies (479 kb) and 2.74 times longer than the longest single Illumina contig (1.38 Mb; one of 13 contigs from the 99.19% complete uncultured *Bacteroidia* bacterium RUG14538). In terms of a direct comparison, the Illumina-only assembly of the same sample had an N50 of 12.2 kb, a total length of 247 Mb and a longest contig of 358 kb.

marker-genes[15], and our dataset increases the number of publicly available genomes for *S. azabuensis* from 4 to 22. Large numbers of uncharacterized Ruminococcaceae and Bacteroidia were also associated with HME sheep. Multiple strains of uncharacterized Proteobacteria, including RUG14498 described above, were more proportionally abundant in HME sheep, and *Fibrobacter* strains were almost all associated with LME sheep.

The relationship between proportional abundance of archaea and methane emissions is not simple. Most archaeal strains were present at similar abundance in LME and HME sheep (Supplementary Data 8). RUGs representing new strains of *Methanobrevibacter* were often more abundant in HME sheep. The RUG with the most striking proportional abundance was RUG12825, which is likely a member of the *Methanosphaera* genus, and was more abundant in LME sheep. The complex relationship between relative abundance of methanogens and methane emissions may underlie our inability to find significant differences in overall archaeal proportional abundance.

That notwithstanding, these data represent a strain-level view of methane emissions in sheep, and support the hypothesis that there are major fundamental changes in rumen metagenomic relative abundance associated with the extremes of low and high methane emissions.

Global rumen census updated. The global rumen census attempted to determine the core rumen microbiome by using 16S rRNA sequencing of rumen samples from 742 individual animals from around the world, comprising eight ruminant species[29]. *Prevotella*, *Butyrivibrio* and *Ruminococcus*, as well as unclassified Lachnospiraceae, Ruminococcaceae, Bacteroidales and Clostridiales, were the dominant rumen bacteria and may

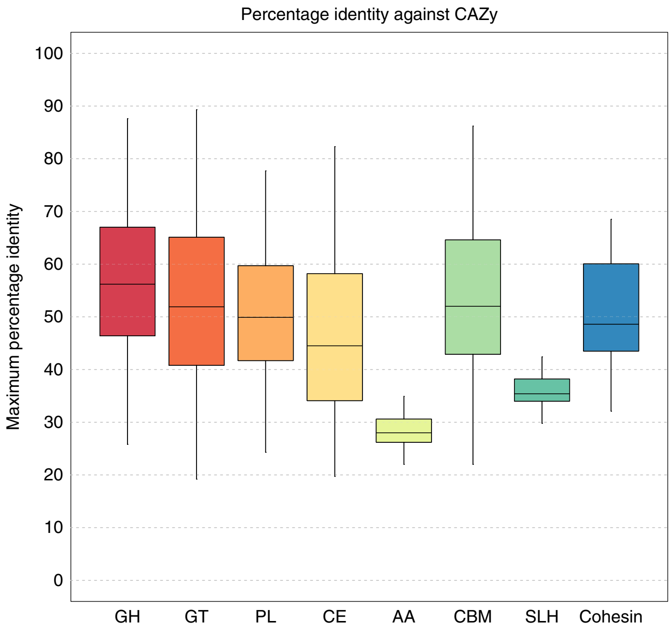

Fig. 4 | Maximum percentage identity between CAZyme-predicted proteins from the RUGs and the CAZy database. GH, glycoside hydrolase ($n = 235,001$); GT, glycosyl transferase ($n = 120,494$); PL, polysaccharide lyase ($n = 6,834$); CE, carbohydrate esterase ($n = 55,523$); AA, auxiliary activities; CBM, carbohydrate-binding module ($n = 23,928$); SLH, S-layer homology domain ($n = 150$); cohesin, cohesin domain ($n = 80$). Center lines indicate the median value; boxes show the interquartile range; and whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box.

represent a core bacterial rumen microbiome. The same species were abundant in our data (Supplementary Data 5). We also found that many Proteobacteria were highly abundant, including *Succinivibrio* (Supplementary Data 5). This is noteworthy because Proteobacteria were found to be highly abundant in many of the samples from the rumen census, but were not highlighted as being part of the core rumen microbiome.

To further characterize the proportional abundance of Proteobacteria, we used the rumen superset database to classify data from this study, Wallace et al.[6], Rubino et al.[25], Shi et al.[15] and Svartström et al.[21] (Supplementary Fig. 11). Proteobacteria were present in all datasets; they were abundant in cattle datasets, but less so in moose and sheep. Given the high proportional abundance of Proteobacteria in many samples, and their consistent presence in all of the samples we tested, we suggest adding Proteobacteria to the core bacterial rumen microbiome that was proposed by Henderson et al.[29].

Long-read assembly of complete bacterial chromosomes. We analyzed a single sample (10572_0012) using a MinION sequencer and a comparison of Illumina and MinION assembly statistics is presented in Fig. 3. Three flow cells produced 11.4 gigabases of data with an N50 value of 11,585 base pairs. The mean read length was 6,144 base pairs, which is short in comparison to other reports[30,31]. We attribute this to short DNA fragments and nicks caused by the bead beating step during DNA extraction. We assembled long reads using Canu[32], to form an assembly of 178 Mb in length with an N50 of 268 kb. Regardless of length, Canu predicted 31 of the contigs to be circular. These circular contigs might represent putative plasmids or other circular chromosomes.

One problem with single-molecule sequencing technologies is the presence of post-assembly insertions and deletions (indels)[33]. Canu can correct reads but not enough to remove all indels. Detecting sequencing errors without a ground truth dataset is difficult, so we hypothesized that most indels would create premature stop codons and that gene prediction tools (for example, Prodigal[34]) would produce truncated proteins. We examined the ratio between the lengths of predicted proteins and their top hits in UniProt to estimate indels (Supplementary Fig. 12). Although these data indicate multiple errors as compared to the Illumina short-read data, we corrected errors by polishing with one round of Nanopolish and two rounds of Racon. We set up a software pipeline to calculate statistics and produce similar plots for any input genome or metagenome called IDEEL.

Statistics for all contigs ≥500 kb and all contigs predicted to be circular are provided in Supplementary Data 9. The nanopore assembly contained several single contigs that we predict are complete, or near-complete, circular whole chromosomes.

Prevotella copri nRUG14950 (tig00000032) was a single contig of 3.8 Mb, which most closely resembled *Prevotella copri* DSM 18205, and which showed high similarity to RUG14032. *Prevotella copri* nRUG14950 was predicted to be 98.48% complete by CheckM[35], with a contamination score of 2.03%, whereas RUG14032 was estimated to be 96.62% complete with a contamination score of 1.35%. Comparative alignments between *Prevotella copri* nRUG14950, RUG14032 and *Prevotella copri* DSM 18205 are shown in Supplementary Fig. 13. There was a clear and striking relationship between *Prevotella copri* nRUG14950 and RUG14032. These two genomes, both estimated to be nearly complete, were assembled from different samples using different techniques, and sequenced with different sequencing technologies. Our assembly of *Prevotella copri* nRUG14950 consisted of only one contig and was estimated to be 98.48% complete, representing the most continuous chromosomal assembly of *Prevotella copri*, despite having been assembled from a metagenome.

Selenomonas spp. nRUG14951 was a single contig of 3.1 Mb in length that was predicted to be circular, with completeness and contamination statistics of 98.13% and 0.16%, respectively. The most similar RUG was RUG10160, with a mean of 94% protein identity. RUG10160 was estimated to be 97.66% complete and 0% contaminated. However, the closest public reference genome was *Selenomonas ruminantium* GACV-9, part of the Hungate collection, which had only ~64% protein identity with *Selenomonas* spp. nRUG14951. There was a good whole-genome alignment between *Selenomonas* spp. nRUG14951 and RUG10160 (Supplementary Fig. 14), albeit with some evidence of rearrangements and some small sections of the genome that were only captured by the nanopore assembly.

We also identified Lachnospiraceae bacterium nRUG14952, which had a 2.5-Mb circular, near-complete genome (95.46%), a second RUG13141 (which had 96% protein identity to nRUG14952) and a more distantly related public reference genome (Lachnospiraceae bacterium KHCPX20, which has 63% protein identity to nRUG14952). The nanopore-assembled genome Lachnospiraceae bacterium nRUG14952 contained several genome regions that were absent from RUG13141 (Supplementary Fig. 15).

nRUG14951 and nRUG14952 represent entire bacterial chromosomes assembled as single contigs and are the first genome assemblies for these species. The remainder of the nanopore assembly contained highly continuous contigs that represent large portions of previously unsequenced bacterial chromosomes. These results taken together demonstrate the power of long reads for assembling complete chromosomes from complex metagenomes.

To assess the advantage of having complete chromosomal assemblies, we annotated the three nanopore whole genomes and the three genomes of their closely related RUGs (Supplementary Data 10). The three complete nanopore genomes contained five, seven and three full-length 16S gene sequences, whereas all three RUGs

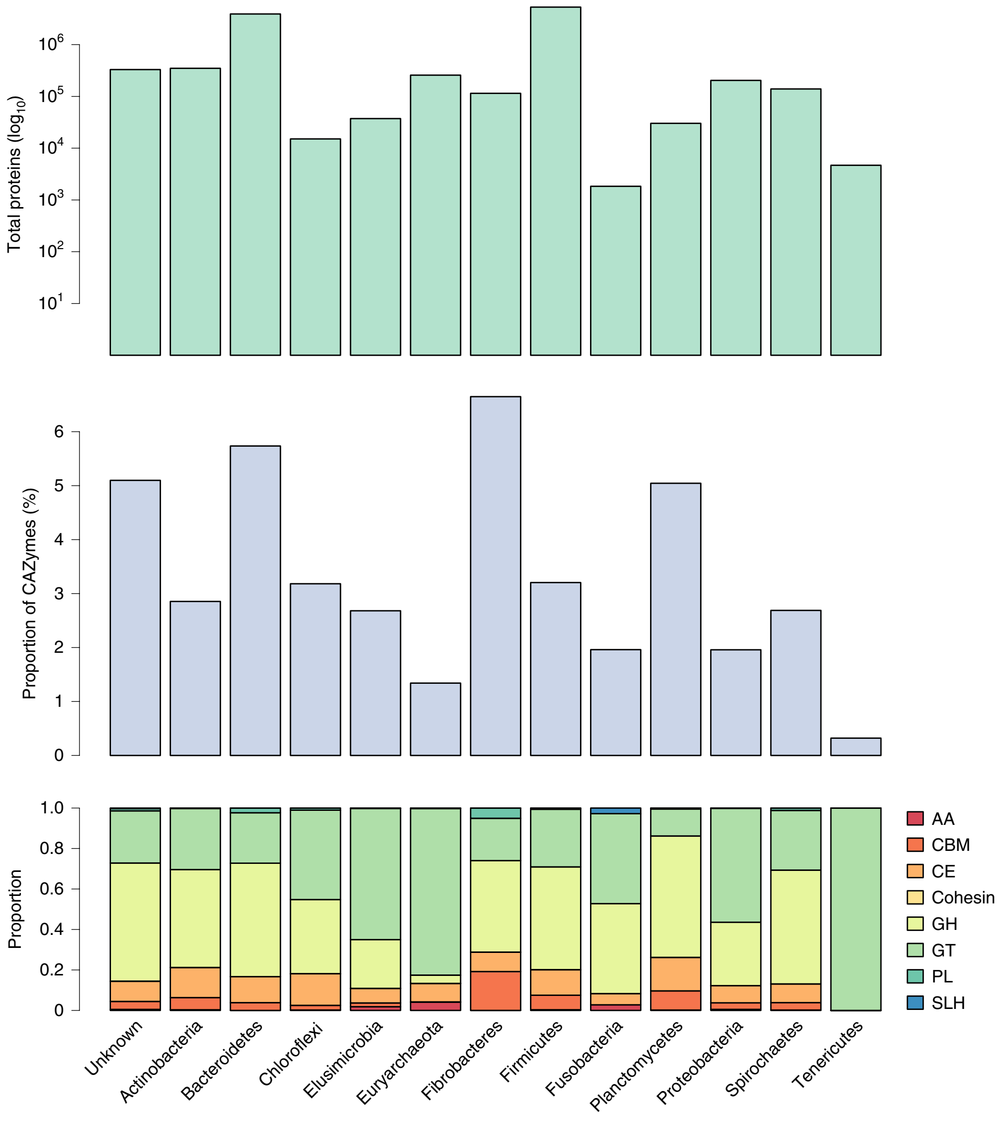

Fig. 5 | Taxonomic and functional distribution of proteins. Top, total number of proteins for 12 phyla and the group of unknown bacteria. Middle, percentage of the proteome predicted to be CAZymes. Bottom, distribution of eight CAZyme classes as a proportion of the total number of predicted CAZymes.

contained none. In addition, the three nanopore genomes were massively enriched for IS family transposase proteins as compared to their RUG counterparts. Transposases are associated with insertion sequences in bacterial genomes, and catalyze the transposition of mobile elements[36]. Finally, in all cases, the nanopore assemblies had more annotated clusters of orthologous genes, suggesting that they have more complete functional annotation than their short-read counterparts.

A protein database for rumen microbial proteomics. We put together a non-redundant dataset of rumen proteins from the 4,941 RUGs and 460 publicly available genomes from the Hungate collection (10.69 million proteins), following the model of UniRef[37] and clustering the protein set at 100% (9.45 million clusters), 90% (5.69 million clusters) and 50% (2.45 million clusters) identity to form RumiRef100, RumiRef90 and RumiRef50, respectively.

To assess the protein-level difference between our dataset and other rumen MAG datasets, we took RumiRef100 and added over 900,000 predicted proteins from the rumen superset. We clustered these at 90% identity, which resulted in 6.24 million protein clusters. Of these, 5 million clusters contained at least one RUG protein, 4.74 million contained only RUG proteins and 3.67 million were singletons that contained only RUG proteins.

All 10.69 million predicted proteins from the RUGs were compared to KEGG[38], 460 public genomes from the Hungate collection, UniRef100, UniRef90 and UniRef50. The mean protein identities of the top hit for these databases were 55.88%, 63.58%, 67.52%, 67.25% and 59.97%, respectively. These data provide a comprehensive and richly annotated protein dataset from the rumen.

The RUG proteins were compared to the CAZy[39] database (31 July 2018) using dbCAN2 (ref. [40]). A total of 442,917 were predicted to be involved in carbohydrate metabolism, including 235,001 glycoside hydrolases, 120,494 glycosyl transferases, 55,523 carbohydrate esterases, 23,928 proteins with carbohydrate-binding modules, 6,834 polysaccharide lyases, 907 proteins with predicted auxiliary activities, 80 proteins with a predicted cohesin domain and 150 proteins with an S-layer homology module (SLH).

The similarity of the predicted CAZymes to the current CAZy database can be seen in Fig. 4. None of the eight classes of carbohydrate-active enzymes displayed an average protein identity greater than 60% indicating that CAZy poorly represents the diversity of CAZymes encoded in the genomes of ruminant microbes.

Of particular note is the class AA 'auxiliary activities', with an average protein identity of less than 30% between CAZy and the RUG CAZymes. AA was created by CAZy to classify ligninolytic enzymes and lytic polysaccharide monooxygenases (LPMOs).

The distribution of CAZymes across 12 different phyla and the group of unknown bacteria can be seen in Fig. 5. The Bacteroidetes (3.9 million) and Firmicutes (5.3 million) together contributed the largest number of proteins to our dataset; however, whereas 5.7% of the proteome of Bacteroidetes was devoted to CAZyme activity, in Firmicutes the figure was 3.2%. Fibrobacteres devoted the highest percentage of their proteome to carbohydrate metabolism (over 6.6%), as was expected owing to their fiber-attached, high-cellulolytic activity. Only a few studies exist on the role of Planctomycetes in the rumen[24,41,42]; however, while they contributed a relatively low number of proteins in our dataset (30,172), just over 5% of those proteins were predicted to be CAZymes, suggesting a role in, and adaptation to, carbohydrate metabolism. Of the 80 cohesin-containing proteins, 79 were encoded by the Firmicutes (the remaining one was encoded by an unknown bacterium), as were 101 of 149 SLH-domain-containing proteins. Both are components of cellulosomes, multienzyme complexes that are involved in fiber degradation, which are encoded by some members of the Clostridiales family.

There were 1,707 Bacteroidetes genomes in the RUGs, and additionally we had a whole genome of *Prevotella copri* from the nanopore assembly. These 1,708 genomes were subjected to prediction of polysaccharide utilization loci (PULz) using our pipeline PULpy[43]. Of the 1,708 genomes, 1,469 were predicted to have at least one PUL, and in total there were 15,629 separate loci involving 88,260 proteins. The highest numbers of PULs per genome were 52 for RUG13980 and 50 for RUG10279; both these were labeled as uncultured Prevotellaceae and both of these genomes are closely related to *Prevotella multisaccharivorax*, which is known to be able to utilize multiple carbohydrate substrates[44].

Discussion

The rumen microbiome has a crucial role in food security and climate change. Recent studies have released more than 1,300 draft and complete rumen genomes. We add 4,941 near-complete, dereplicated metagenome-assembled genomes to these 1,300 existing rumen genomes[9,20,21]. By combining our dataset with publicly available genomes, we assembled a rumen superset of 5,845 publicly available bacterial and archaeal genomes. This set contains 2,690 unique species-level bins (95% ANI), and 2,078 of these 2,690 putative species are RUG2 genomes discovered in this study. The RUG2 dataset and the rumen superset bring read classification rates up to 70% for our own data and 45–55% for other rumen metagenome datasets (some from non-cattle ruminants). The remaining reads are likely to derive from low-abundance bacterial and archaeal species, difficult-to-assemble genomes, and the fungal, protozoan and viral genomes that are not part of this study.

We estimate that we have discovered 65% of rumen species in our samples, representing four important beef cattle breeds, which suggests that there are over 1,000 species yet to be sequenced and assembled. Given that average read classification rates dip from 70% in our own data to 50% in the cattle data of Rubino et al. (Limousin × Friesian cross)[25] and the sheep data of Shi et al. sheep data[26] and to 45% in the moose data[21], there are many species yet to be discovered, and there are likely to be species- and breed-specific rumen microbiomes. We note the high abundance of unclassified Proteobacteria in our data, as well as in the rumen census data, and suggest that these may form part of a core rumen microbiome. Our dataset contains 74 proteobacterial genomes, and we present one near-complete genome in a single contig.

We apply our dataset to reanalyze data on methane emissions in sheep that were published in 2014[26]. Using a combined database of rumen microbial genomes, we reveal fundamental and large-scale differences in rumen metagenomic abundance between LME and HME sheep. These differences occur at almost every taxonomic level tested, and the rumen superset database enabled us to analyze these data at high resolution. While species- and strain-level metagenomic data must always be interpreted with care, there remains a possibility that strains that are not present in the database are driving the observed differences. Nonetheless, we observed consistent patterns suggesting large changes in abundance for numerous species. Our analysis supports subsequent studies of methane emissions in sheep[15,28] by identifying specific strains of bacteria and archaea involved, and revealing their genome sequence. Our analysis confirmed that there was a complex relationship between archaeal abundance and methane emissions, with archaeal species and strains both positively and negatively associated with methane emissions. These insights into metagenomic species- and strain-level aspects of methane emissions will form the basis of future studies.

The main rumen functions rely on the activity of proteins encoded in rumen microbe genomes, and as researchers produce more proteomic data, it is vital that protein reference datasets be available. We present a dataset of large redundant and non-redundant rumen microbial protein predictions, and provide rich annotation using public protein, pathway and enzyme databases. This resource will enable researchers to predict the function of each protein, and better assess the functional consequences of changes in the rumen proteome.

Going forward, it is vital that more rumen bacteria and archaea be brought into culture, to better enable studying the functions of the rumen microbiome. In particular, if we are to design rational interventions to manipulate rumen feed conversion or methane emissions, we will need to understand microbiome structure, the substrates that are utilized by microbiota and how the microbiota interacts with one another and the ruminant host. Sequencing and assembling rumen microbial genomes is an important step toward improved culture collections and future manipulation of the rumen microbiome for human benefit.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0202-3.

Received: 19 November 2018; Accepted: 27 June 2019;
Published online: 2 August 2019

References

1. Hess, M. et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467 (2011).
2. Cowan, D. A. et al. Metagenomics, gene discovery and the ideal biocatalyst. *Biochem. Soc. Trans.* **32**, 298–302 (2004).
3. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front. Genet.* **8**, 23 (2017).
4. Huws, S. A. et al. Addressing global ruminant agricultural challenges through understanding the rumen microbiome: past, present, and future. *Front. Microbiol.* **9**, 2161 (2018).
5. Gerber, P. J et al. *Tackling Climate Change Through Livestock: a Global Assessment of Emissions and Mitigation Opportunities*. (Food and Agriculture Organization of the United Nations (FAO), 2013).
6. Wallace, R. J. et al. The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics* **16**, 839 (2015).
7. 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**, e1005846 (2016).
8. Stewart, R. D. et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).
9. Seshadri, R. et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).

10. Stewart, R. D., Auffret, M., Snelling, T. J., Roehe, R. & Watson, M. MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). *Bioinformatics* **35**, 2150–2152 (2019).
11. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
12. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
13. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
14. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
15. Kamke, J. et al. Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a Sharpea-enriched microbiome characterised by lactic acid formation and utilisation. *Microbiome* **4**, 56 (2016).
16. 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 (2017).
17. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
18. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
19. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
20. Solden, L. M. et al. Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).
21. Svartström, 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 (2017).
22. Chao, A. Nonparametric estimation of the number of classes in a population. *Scand. Stat. Theor. Appl.* **11**, 265–270 (1984).
23. 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 (2018).
24. Auffret, M. D. et al. The rumen microbiome as a reservoir of antimicrobial resistance and pathogenicity genes is directly affected by diet in beef cattle. *Microbiome* **5**, 159 (2017).
25. Rubino, F. et al. Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. *ISME J.* **11**, 932–944 (2017).
26. Shi, W. et al. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res.* **24**, 1517–1525 (2014).
27. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
28. Kittelmann, S. et al. Two different bacterial community types are linked with the low-methane emission trait in sheep. *PLoS One* **9**, e103171 (2014).
29. Henderson, G. et al. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* **5**, 14567 (2015).
30. Risse, J. et al. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* **4**, 60 (2015).
31. Ip, C. L. C. et al. MinION analysis and reference consortium: phase 1 data release and analysis. *F1000Res.* **4**, 1075 (2015).
32. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
33. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126 (2019).
34. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
35. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
36. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
37. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
38. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
39. Cantarel, B. L. et al. The carbohydrate-active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).
40. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).
41. Hoo, S. E. et al. Impact of subacute ruminal acidosis (SARA) adaptation and recovery on the density and diversity of bacteria in the rumen of dairy cows. *FEMS Microbiol. Ecol.* **78**, 275–284 (2011).
42. Kasparovska, J. et al. Effects of isoflavone-enriched feed on the rumen microbiota in dairy cows. *PLoS One* **11**, e0154642 (2016).
43. Stewart, R. D., Auffret, M., Roehe, R. & Watson, M. Open prediction of polysaccharide utilisation loci (PUL) in 5,414 public Bacteroidetes genomes using PULpy. Preprint at https://doi.org/10.1101/421024 (2018).
44. Sakamoto, M., Umeda, M., Ishikawa, I. & Benno, Y. *Prevotella multisaccharivorax* sp. nov., isolated from human subgingival plaque. *Int. J. Syst. Evol. Microbiol.* **55**, 1839–1843 (2005).
45. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).

Acknowledgements

The Rowett Institute and SRUC are core funded by the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish government. 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 (BBSRC; BB/N016742/1, BB/N01720X/1), including institute strategic program and national capability awards to the Roslin Institute (BB/P013759/1, BB/P013732/1, BB/J004235/1, BB/J004243/1, BBS/E/D/30002276); and by the Scottish government as part of the 2016–2021 commission.

Author contributions

M.W., R.R. and A.W.W. conceived of the study and supervised the project. R.D.S. and M.W. carried out all bioinformatics work on the Illumina data and A.W. carried out all bioinformatics work on the nanopore data. M.D.A. carried out all laboratory work, except for DNA clean-up, nanopore library preparation and nanopore sequencing, which were done by A.W. All of the authors contributed ideas, co-wrote the paper, and reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0202-3.

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

Correspondence and requests for materials should be addressed to M.W.

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

Methods

Metagenomic samples. Animal experiments were conducted at the Beef and Sheep Research Centre of Scotland's Rural College (SRUC). The experiment was approved by the Animal Experiment Committee of SRUC and was conducted in accordance with the requirements of the UK Animals (Scientific Procedures) Act 1986.

The data were obtained from three cross breeds (Aberdeen Angus, Limousin and Charolais) and one pure breed (Luing) (Supplementary Data 4). As previously described, the animals were slaughtered in a commercial abattoir where two post-mortem digesta samples (approximately 50 ml) were taken immediately after the rumen was opened to be drained[46,47]. DNA extraction was carried out following the protocol from Yu and Morrison[48] and was based on repeated bead beating with column filtration. Illumina TruSeq libraries were prepared from genomic DNA and sequenced on an Illumina HiSeq 4000 by EdinburghGenomics.

We experienced severe problems when using MinION for rumen microbiome DNA while following standard recommended protocols, and we hope our adapted methods will be of assistance to others. We found that the DNA did not meet the recommended purity for nanopore library preparation following extraction, according to Nanodrop optical density ratios. RNase treatment using Riboshredder and clean-up with methods such as AMPure XP beads were sufficient to obtain optical density ratios within the recommended range, but DNA from these methods typically led to poor or failed sequencing runs. Successful clean-up reaching recommended optical density ratios that led to successful sequencing runs was carried out using RNase treatment with Riboshredder and a phenol–chloroform purification. One-dimensional libraries were prepared starting with 2 μg of DNA per library following Oxford Nanopore's SQK-LSK108 one-dimensional ligation protocol with modifications. The incubation in the end preparation stage of the protocol was extended to 30 min at 20 °C and 30 min at 65 °C, and the incubation in the ligation stage was extended to 15 min at room temperature. The optional repair step for formalin-fixed, paraffin-embedded tissues was also carried out. Three sequencing runs were carried out using FLOMIN-106 flow cells on a MinION MK1b housed in the Watson laboratory at the University of Edinburgh.

Bioinformatics. *Metagenomic assembly and binning.* In total, 282 samples were sequenced for this study generating between 24 and 140 million 150-base-pair paired-end reads per sample. The samples were sequenced in five batches of 48 samples and one batch of 42 samples (this 42-sample batch was the sole basis of Stewart et al.). An additional sample was used for Hi-C sequencing in Stewart et al.[8], and the metagenome-assembled genomes from that sample are included in the dereplicated set.

Unless otherwise stated, all parameters used were the default. Each sample was assembled and binned individually using coverage and content as previously described[8]. In brief, each sample was assembled using idba_ud[49] (v.1.1.3) with the options '--num_threads 16 --pre_correction --min_contig 300'. BWA MEM[50] (v.0.7.15) was used to map reads back to the filtered assembly and Samtools[51] (v.1.3.1) was used to convert to BAM format. Script jgi_summarize_bam_contig_depths from the MetaBAT2[52] (v.2.11.1) package was used to calculate coverage from the resulting BAM files. A co-assembly was also produced for each of the six batches of samples using MEGAHIT[53] (v.1.1.1) with options '--kmin-1pass -m 60e+10 --k-list 27,37,47,57,67,77,87 --min-contig-len 1000 -t 16'.

Metagenomic binning was applied to both single-sample assemblies and the co-assemblies using MetaBAT252 with options '--minContigLength 2000 --minContigDepth 2'. Single-sample binning produced a total of 37,153 bins, and co-assembly binning produced a further 23,335. All 60,743 bins were aggregated and then dereplicated using dRep[54] (v.1.1.2). The dRep dereplication workflow was used with options 'dereplicate_wf -p 16 -comp 80 -con 10 -str 100 -strW 0'. Thus, in prefiltering, only bins assessed by CheckM (v.1.0.5) as having both completeness ≥80% and contamination ≤10% were retained for pairwise dereplication comparison ($n = 10{,}586$). Bin scores were given as completeness − 5 × contamination + 0.5 × log(N50), and only the highest scoring RUG from each secondary cluster was retained in the dereplicated set. For our dataset, 4,941 dereplicated RUGs were obtained.

Note that we operated a continuous dereplication workflow. Therefore all 913 of the RUGs (both MetaBAT2 and Hi-C) we previously published have been merged with the newer RUGs and dereplicated. As a result, while some of the previously published RUGs exist in the newer dataset published here, many have been replaced by newer RUGs of higher quality.

Supplementary Data 5 gives the average depth for each genome in each sample as calculated by script jgi_summarize_bam_contig_depths from the MetaBAT2 (ref. [52]) (v.2.11.1) package.

Metagenomic assignment. The output of metagenomic binning is simply a set of DNA FASTA files containing putative genomes. These were all assessed for completeness and contamination using CheckM[35] (v.1.0.5). The 4,941 best bins were analyzed using MAGpy[10], a Snakemake[55] pipeline that runs a series of analyses on the bins, including CheckM (v.1.0.5); prodigal[34] (v2.6.3) protein prediction; Pfam_Scan[56] (v.1.6); a DIAMOND[12] (v.0.9.22.123) search against UniProt TrEMBL; PhyloPhlAn[13] (v.0.99); and a sourmash (v.2.0.0) search against all public bacterial genomes. The MAGpy results were used to produce a putative taxonomic assignment for each bin as follows:

- If the proportion of proteins assigned to a species is ≥0.9 and the average amino acid identity is ≥0.95, assign to species on the basis of DIAMOND results; else
- If sourmash score is ≥0.8, assign to species on the basis of sourmash results; else
- If PhyloPhlAn probability is high and the level is genus or species, then assign taxonomy on the basis of PhyloPhlAn results; else
- If the proportion of proteins assigned to a genus is ≥0.9 and the average amino acid identity is ≥0.9, assign to genus on the basis of DIAMOND results; else
- If PhyloPhlAn probability is high or medium and the level is genus, then assign to genus on the basis of PhyloPhlAn results; else
- If PhyloPhlAn probability is high or medium and the level is family, then assign to family on the basis of PhyloPhlAn results; else
- If the proportion of proteins assigned to a family is ≥0.8 and the average amino acid identity is ≥0.6, assign to family on the basis of DIAMOND results; else
- If PhyloPhlAn probability is high or medium and the level is order, then assign to order on the basis of PhyloPhlAn results; else
- If the proportion of proteins assigned to a order is ≥0.6 and the average amino acid identity is ≥0.6, assign to order on the basis of DIAMOND results; else
- If PhyloPhlAn probability is high or medium and the level is class, then assign to class on the basis of PhyloPhlAn results; else
- If PhyloPhlAn probability is high or medium and the level is phylum, then assign to phylum on the basis of PhyloPhlAn results; else
- Assign taxonomy on the basis of CheckM lineage

Importantly, at this stage, these are only putative taxonomic assignments. Using these labels, a phylogenetic tree consisting of the RUGs and genomes from the Hungate collection, produced from concatenated protein subsequences using PhyloPhlAn[13] (v.0.99), was visually inspected using FigTree (v.1.4.3), iTol[57] (v.4.3.1) and GraPhlAn[45] (v.0.9.7). Annotations were improved where they could be—for example, where MAGpy had only assigned a taxonomy at the genus level but the genome clustered closely with a Hungate 1,000 genome annotated at the species level, the annotation was updated. The tree was also rerooted manually at the Bacteria–Archaea branch using FigTree.

Genome quality and comparative genomics. Genome completeness and contamination was assessed using CheckM (v.1.0.5) (see above). tRNA genes were annotated using tRNAscan-SE (v.2.0.0) and 16S rRNA genes were predicted using barrnap (v.0.9). Whole-genome alignments were calculated with MUMmer[58] (v.3.23) using promer to find matches between genomes. ANI was calculated using FastANI (v.1.1). The RUGs were compared to the Hungate collection and our previous dataset using DIAMOND blastp (v.0.9.22.123) and MASH[59] (v.2.0) with parameters '-k 21 -s100000'.

The rumen superset was dereplicated using dRep as above, with '-sa 0.99' for dereplication at 99% ANI and '-sa 0.95' for dereplication at 95% ANI. Overlaps between sets were plotted with UpSetR[60] (v.1.3.3). Read classification rates were calculated using kraken[61] (v.0.10.5) with parameters '--fastq-input --gzip-compressed --preload --paired'.

Analysis of sheep methane data. Reads from the low and high methane samples from Shi et al. were assigned to different taxonomic levels of the rumen superset database using kraken, as described above. The resulting read counts data were used as input into DESeq2 (v.1.22.2) for differential analysis. Principal-component analysis plots were created using the plotPCA() function within DESeq2, and heat maps were created using the heatmap.2() function within the gplots package (v.3.0.1.1). For strain-level analysis, reads from the low- and high-methane samples from Shi et al. were aligned directly to the rumen superset database using BWA-MEM (v.0.7.15) and the number of primary alignments to each genome was used as input to DESeq2. *P* values for all comparisons were calculated by DESeq2 and adjusted for multiple testing[62].

Rumen census analysis. The average and total depth for each genome in each dataset (Supplementary Data 5) was used as a proxy for abundance in the dataset(s). Kraken (as described above) was used with the rumen superset database to calculate the read abundance of Proteobacteria in all samples.

Assembly and analysis of nanopore sequence data. The nanopore reads were extracted and quality controlled using poRe[63,64] (v.0.24), and assembled using Canu[32] (v.1.8) with default settings and genomeSize = 150 Mb. The resulting assembly was analyzed using MAGpy[10]. The raw assembly was corrected using both Nanopolish[65] (v.0.10.2) and Racon[66] (v.1.3.1) using Illumina data aligned to the nanopore assembly with Minimap2 (v.2.12) using short-read settings (-x sr). Query versus subject length data were extracted and plotted using IDEEL (https://github.com/mw55309/ideel). Whole-genome alignments were calculated using MUMmer79 (v.3.23) using promer to find matches between genomes. The three complete nanopore bacterial genomes and their Illumina counterparts were annotated using Prokka[67] (v.1.13.3). The nanopore assembly was created with a minimum contig length of 1 kb; therefore, the Illumina assemblies were similarly limited before comparison.

Proteome analysis. Proteins were predicted using Prodigal (v.2.6.3) with option '-p meta'. Using DIAMOND, each protein was searched against KEGG (downloaded on 15 September 2018), UniRef100, UniRef90 and UniRef50 (downloaded 3 October 2018), and CAZy (dbCAN2 version, 31 July 2018). The protein predictions were clustered by CD-HIT[68] (v.4.7) at 100%, 90% and 50% identity, mirroring similar methods at UniRef.

All protein predictions were searched against the CAZy database using dbCAN2 (ref. [40]) and HMMER[69] (v.3.1b2), and PULs were predicted for Bacteroidetes RUGs using PULpy[43].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequence reads for all samples are available under European Nucleotide Archive (ENA) project PRJEB31266, except for 10572 that are available under PRJEB21624. All metagenomic assemblies and RUGs have been deposited in ENA under accession PRJEB31266. All protein predictions, clusters and annotation are available at https://doi.org/10.7488/ds/2470.

Code availability

Comparative genomic analysis was carried out using MAGpy[10] (https://github.com/WatsonLab/MAGpy); analysis of PULs was carried out using PULpy[43] (https://github.com/WatsonLab/PULpy); and analysis of indels in nanopore data was carried out using IDEEL (https://github.com/mw55309/ideel).

References

46. Duthie, C.-A. et al. Impact of adding nitrate or increasing the lipid content of two contrasting diets on blood methaemoglobin and performance of two breeds of finishing beef steers. *Animal* **10**, 786–795 (2016).
47. Duthie, C.-A. et al. The impact of divergent breed types and diets on methane emissions, rumen characteristics and performance of finishing beef cattle. *Animal* **11**, 1762–1771 (2017).
48. Yu, Z. & Morrison, M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques* **36**, 808–812 (2004).
49. 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 (2012).
50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
51. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
53. Li, D., Liu, C.-M., Luo, R., 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 (2015).
54. 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 (2017).
55. Koster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
56. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
57. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
58. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
59. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
60. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
61. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.* **57**, 289–300 (1995).
63. Watson, M. et al. poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* **31**, 114–115 (2015).
64. Stewart, R. D. & Watson, M. poRe GUIs for parallel and real-time processing of MinION sequence data. *Bioinformatics* **33**, 2207–2208 (2017).
65. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
66. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
67. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
68. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
69. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

nature research

Corresponding author(s): Mick Watson

Last updated by author(s): Jun 14, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Comparative genomic analysis was carried out using MAGpy (https://github.com/WatsonLab/MAGpy); analysis of PUL was carried out using PULpy (https://github.com/WatsonLab/PULpy); analysis of indels in nanopore data was carried out using IDEEL (https://github.com/mw55309/ideel); other open source software used: idba_ud (v1.1.3), BWA MEM (v0.7.15), Samtools (v1.3.1), MetaBAT2 (v2.11.1), MEGAHIT (v1.1.1), dRep (v1.1.2), CheckM (v1.0.5), prodigal (v2.6.3), Pfam_Scan (v1.6), DIMAOND (v0.9.22.123), PhyloPhlAn (v0.99), Sourmash (v2.0.0), FigTree (v.1.4.3), iTol (v4.3.1) and GraPhlAn (v0.9.7), tRNAscan-SE (v2.0.0), barrnap (v0.9), MUMmer (v3.23), FastANI (v1.1), MASH (v2.0), Kraken (v0.10.5), DESeq2 (v1.22.2), gplots (v3.0.1.1), poRe (v0.24), Canu (v1.8), Nanopolish (v0.10.2), Racon (v1.3.1), Minimap2 (v2.12), using Prokka (v1.13.3), CD-HIT (v4.7), dbCAN (v2), HMMER (v3.1b2)

Data analysis

Comparative genomic analysis was carried out using MAGpy (https://github.com/WatsonLab/MAGpy); analysis of PUL was carried out using PULpy (https://github.com/WatsonLab/PULpy); analysis of indels in nanopore data was carried out using IDEEL (https://github.com/mw55309/ideel); other open source software used: idba_ud (v1.1.3), BWA MEM (v0.7.15), Samtools (v1.3.1), MetaBAT2 (v2.11.1), MEGAHIT (v1.1.1), dRep (v1.1.2), CheckM (v1.0.5), prodigal (v2.6.3), Pfam_Scan (v1.6), DIMAOND (v0.9.22.123), PhyloPhlAn (v0.99), Sourmash (v2.0.0), FigTree (v.1.4.3), iTol (v4.3.1) and GraPhlAn (v0.9.7), tRNAscan-SE (v2.0.0), barrnap (v0.9), MUMmer (v3.23), FastANI (v1.1), MASH (v2.0), Kraken (v0.10.5), DESeq2 (v1.22.2), gplots (v3.0.1.1), poRe (v0.24), Canu (v1.8), Nanopolish (v0.10.2), Racon (v1.3.1), Minimap2 (v2.12), using Prokka (v1.13.3), CD-HIT (v4.7), dbCAN (v2), HMMER (v3.1b2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequence reads for all samples are available under ENA project PRJEB31266, except for 10572 which are available under PRJEB21624. All metagenomic assemblies and RUGs are in the process of being deposited in ENA under accession PRJEB31266. All protein predictions, clusters and annotation are available at DOI: 10.7488/ds/2470.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

As primarily a discovery project, sample size was not important, and we examined as many samples as possible.
Where we carried out statistical analysis, those sample sizes were determined by the authors of those studies (e.g. Shi et al)

Data exclusions

No data were excluded

Replication

As a discovery project, replication is not important

Randomization

As a discovery project, no randomization was required

Blinding

As a discovery project, no blinding was required

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
□	☒	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

The data were obtained from three cross breeds: Aberdeen Angus, Limousin and Charolais and one pure breed: Luing; All animals were male and between 459 and 661 days old

Wild animals

the study did not involve wild animals.

Field-collected samples

the study did not involve field-collected samples

Ethics oversight

Animal experiments were conducted at the Beef and Sheep Research Centre of Scotland's Rural College (SRUC). The experiment was approved by the Animal Experiment Committee of SRUC and was conducted in accordance with the requirements of the UK

Animals (Scientific Procedures) Act 1986.

Note that full information on the approval of the study protocol must also be provided in the manuscript.