bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Insights into rumen microbial biosynthetic gene cluster diversity through genome-resolved metagenomics

Christopher L. Anderson[1] and Samodha C. Fernando*[1]

[1]Department of Animal Science, University of Nebraska, Lincoln, NE 68583

*Corresponding author: Samodha C. Fernando (samodha@unl.edu)

Abstract

Ruminants are critical to global food security as they transform lignocellulosic biomass into high-quality protein products. The rumen microbes ferment feed to provide necessary energy and nutrients for the ruminant host. However, we still lack insight into the metabolic processes encoded by most rumen microbial populations. In this study, we implemented metagenomic binning approaches to recover 2,809 microbial genomes from cattle, sheep, moose, deer, and bison. By clustering genomes based on average nucleotide identity, we demonstrate approximately one-third of the metagenome-assembled genomes (MAGs) to represent species not present in current reference databases and rumen microbial genome collections. Combining these MAGs with other rumen genomic datasets permitted a phylogenomic characterization of the biosynthetic gene clusters (BGCs) from 8,160 rumen microbial genomes, including the identification of 5,346 diverse gene clusters for nonribosomal peptide biosynthesis. A subset of *Prevotella* and *Selenomonas* BGCs had higher expression in steers with lower feed efficiency. Moreover, the microdiversity of BGCs was fairly constant across types of BGCs and cattle breeds. The reconstructed genomes expand the genomic representation of rumen microbial lineages, improve the annotation of multi-omics data, and link microbial populations to the production of secondary metabolites that may constitute a source of natural products for manipulating rumen fermentation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Main

With the expected population growth and changes in food consumption patterns, ruminant agriculture is critical to meeting global demands for animal products[1]. The rumen microbial community is central to the conversion of indigestible plant biomass into food products via conversion of complex carbohydrates to volatile fatty acids (VFAs) that provides the ruminant animal with approximately 70% of its caloric requirements[2]. Consequently, rumen microbes are paramount to ruminant health and productivity. Advancing the understanding of the structure-function relationship of the rumen microbiome is critical to improving ruminant agriculture.

Recent investigations of the rumen microbiome have expanded rumen microbial genomic databases[3–5]; however, the genomic characterization of rumen microbes is far from complete. The Hungate1000 project provided high-quality reference genomes for hundreds of cultured rumen microbial strains[3]. Nevertheless, the majority of rumen microbial populations remain elusive to current culturing strategies. In a recent cultivation experiment with defined and undefined media, 23% of rumen microbial operational taxonomic units were recovered[6]. Metagenomic binning approaches have been employed to bypass the cultivation bottleneck and generate rumen microbial population genomes[4,5,7,8]. Stewart *et al.* reconstructed 4,941 metagenome-assembled genomes (MAGs) from cattle and highlighted the carbohydrate-active enzyme diversity residing in uncultivated taxa[4,5]. However, a notable fraction of metagenomic reads from previous studies did not map to genomes from Stewart *et al.* and the Hungate1000 collection[4,5], suggesting many rumen microbial species are yet to be characterized. Increasing the number of reference genomes for rumen microbes by identifying MAGs across different ruminant species would enhance our understanding of structure-function relationships within the rumen microbiome and improve metagenomic inference.

Secondary metabolites are involved in a broad range of functions, such as antimicrobial agents and mediating microbial interactions[9]. Given the evidence linking the transmission of antibiotic resistance from livestock to humans[10,11], there is a need to reduce the use of antimicrobial feed additives by identifying alternatives[12,13]. Due to the intense competition for nutrient resources, the rumen microbiome may provide novel opportunities to develop alternatives using endogenous

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

antimicrobial peptides and probiotic microbial species[14]. A previous analysis found 45.4% of 229 rumen genomes encoded at least one bacteriocin gene cluster[15]. Secondary metabolites also have ecological roles in intercellular communication[9,16]. Thus, gaining fundamental knowledge on secondary metabolism in the rumen is critical to reduce the use of antimicrobial feed additives and expand our understanding of host-microbe and microbe-microbe interactions.

Here, we used publicly available metagenomes from ruminants (cattle, deer, moose, bison, and sheep) in combination with new cattle rumen metagenomic datasets to reconstruct 2,809 MAGs. The MAGs expand the genomic representation of rumen microbial lineages and provide unique genomic insights into rumen microbial physiology. Moreover, we present a phylogenetic characterization of the secondary metabolite biosynthetic gene clusters (BGCs) of rumen microbial genomes and demonstrate the vast potential present within the rumen microbiome for the discovery of novel metabolites and probiotics to improve animal health and productivity.

Results

2,809 draft MAGs from the rumen ecosystem

We amassed 3.3 terabase pairs (Tbp) of data from 369 publicly available and 66 new rumen metagenome datasets (Supplementary Table 1). The metagenomes were from cattle (335 samples, 2.2 Tbp), sheep (75 samples, 888.4 gigabase pairs (Gbp)), moose (9 samples, 108.8 Gbp), deer (8 samples, 62.9 Gbp), and bison (8 samples, 52.3 Gbp). Metagenomes were assembled independently to reduce the influence of strain variation and improve the recovery of closely related genomes[17,18]. Following refinement, dereplication, and filtering of resulting population genomes, we identified 2,809 non-redundant MAGs satisfying the following criteria: dRep[19] genome quality score ≥60, ≥75% complete, ≤10% contamination, N50 ≥5 kbp, and ≤500 contigs.

The median estimated completeness and contamination of the MAGs were 89.7% and 0.9%, respectively (Fig. 1a and Supplementary Table 2). Further, recovered MAGs had a median genome size of 2.2 Mbp, a median of 131 contigs, and a median N50 of 28.3 kbp (Fig. 1b). The proposed minimum information about a MAG (MIMAG) specifies high-quality draft genomes to have an

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

estimated ≥90% completeness, ≤5% contamination, and contain 23S, 16S, and 5S rRNA genes[20]. It remains challenging to reconstruct rRNA genes from short metagenomic reads due to the high sequence similarity of rRNA genes in closely related species. As a result, despite high estimated completeness and low contamination rates, only 20 MAGs meet the MIMAG standards for a high-quality draft genome. We identified a 16S rRNA gene in 197 of the MAGs. The remaining MAGs are characterized as medium-quality MAGs under the MIMAG standards.

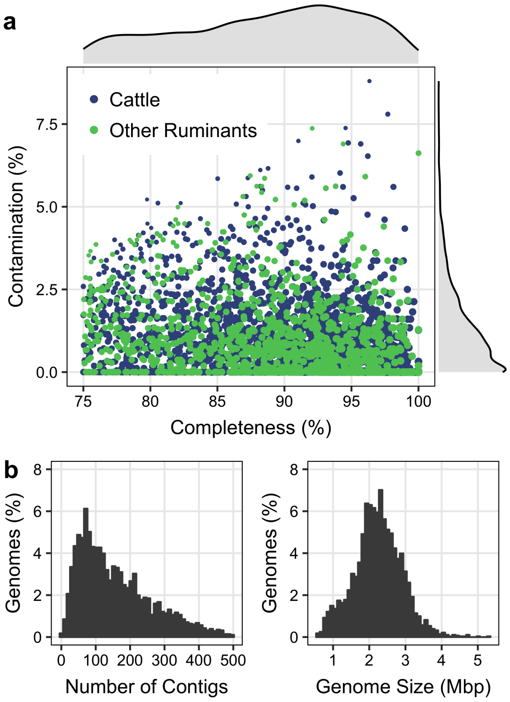

Fig. 1: Genomic properties of 2,809 rumen MAGs. a) CheckM completeness and contamination estimates for the 2,809 population genomes recovered from rumen metagenomes. The size of the point on the scatter plot corresponds to the dRep genome quality score, where Quality = Completeness − (5 · Contamination) + (Contamination · (Strain Heterogeneity / 100)) + 0.5 · (log(N50). The reported MAGs meet the following minimum criteria: genome quality score ≥60, ≥75% complete, ≤10% contamination, N50 ≥5 kbp, and ≥500 contigs. **b)** The frequency distribution of the number of contigs and genome sizes of reconstructed MAGs.

The majority of bacterial MAGs belonged to phyla Firmicutes or Bacteroidota (2,326; Fig. 2a and Supplementary Table 2). Additionally, we assembled 12 bacterial genomes from the superphylum Patescibacteria. At lower taxonomic ranks, Lachnospiraceae (415) and *Prevotella* (398) were the dominant family and genus identified among the assembled bacterial genomes. The most prevalent archaeal family and genus were Methanobacteriaceae (45) and *Methanobrevibacter* (35), respectively (Fig. 2b). The recovered MAGs represent several new taxonomic lineages, as four genomes could not be classified at the rank of order, 16 at the rank of family, and 243 at the genus rank.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

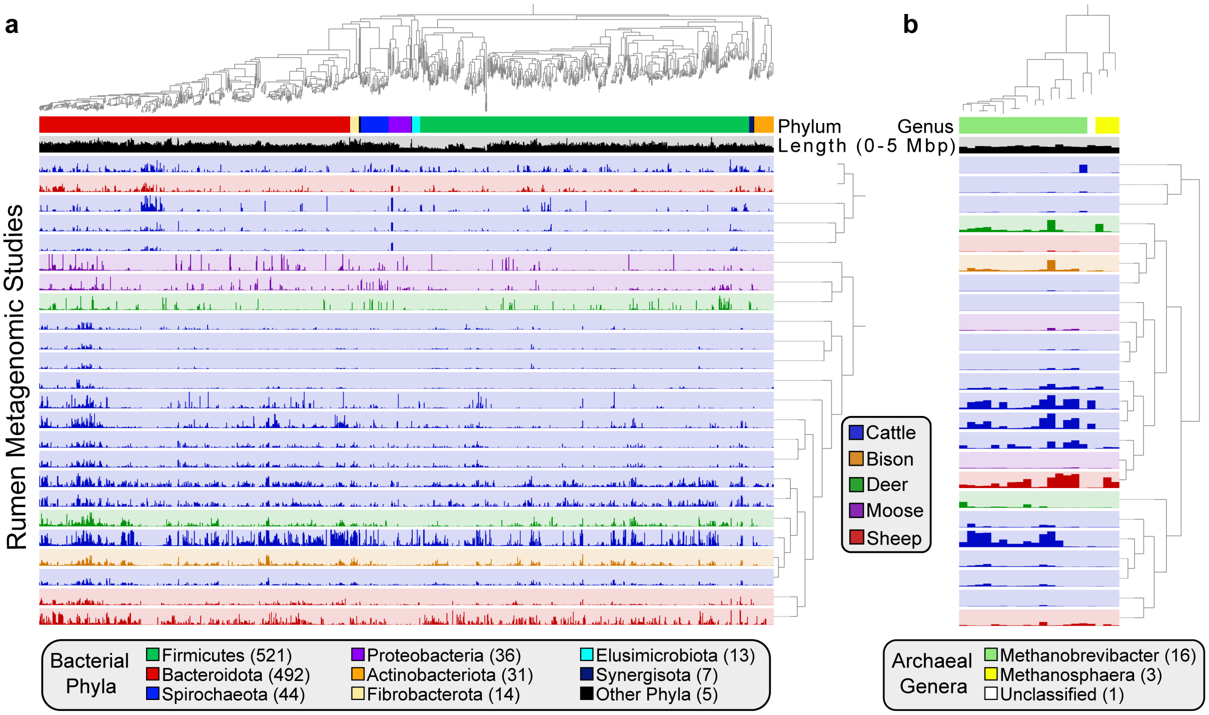

Fig. 2: Phylogenetic relationships and coverage patterns of near-complete MAGs. a) Phylogenomic analysis of 1,163 near-complete (≥90% complete, ≤5% contamination, and N50 ≥15 kbp) bacterial MAGs and **b)** 20 near-complete archaeal MAGs inferred from the concatenation of phylogenetically informative proteins. Layers below the genomic trees designate bacterial phylum or archaeal genus based on GTDB taxonomic assignments, genomic size (0-5 Mbp), and the mean number of bases with ≥1X coverage in a rumen metagenomic dataset (layer color indicates the ruminant the data was collected from). The mean number of bases with ≥1X coverage was used as input for hierarchical clustering of rumen metagenomic datasets based on Euclidean distance and Ward linkage. The bacterial and archaeal phylogenetic trees are provided as Supplementary File 1 and Supplementary File 2, respectively.

Species-level overlap between reference genomes, the Hungate1000 Collection, and rumen MAGs

To further characterize the assembled genomes, we compared the MAGs to other rumen specific genome collections, specifically genomes generated from the Hungate1000 project[3] and MAGs identified from the Stewart *et al.* studies[4,5]. We clustered genomes based on approximate species-level thresholds (≥95% ANI) and calculated the intersection between MAGs in the current study and the Hungate1000 Collection (410 genomes)[3], MAGs from Stewart *et al.* (4,941 genomes)[4,5], and a dereplicated genome collection from the GTDB (22,441, see methods)[21], which includes environmental MAGs[22]. It should be noted that we used the raw data from the first of the Stewart

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

et al. studies[4] (Supplementary Table 1), but with different assembly and binning approaches. Approximately one-third of the MAGs (1,007) did not exhibit ≥95% ANI with a genome in any of the three genomic collections (Fig. 3a). When considering the pairwise intersections between the datasets, 98 (3.5%), 933 (33.2%), and 1,438 (51.2%) of the MAGs in the current study had ≥95% ANI with a genome in the Hungate1000 Collection[3], GTDB[21], and Stewart *et al.*[4,5], respectively. One hundred twenty-one (29.5%), 552 (2.5%), and 3,125 (63.2%) of the genomes from the Hungate1000 Collection[3], GTDB[21], and Stewart *et al.*[4,5] displayed ≥95% ANI with a MAG from the current study. Together, these results indicate that we recovered a majority of previous rumen genomic diversity with additional new lineages not previously identified from the rumen ecosystem.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

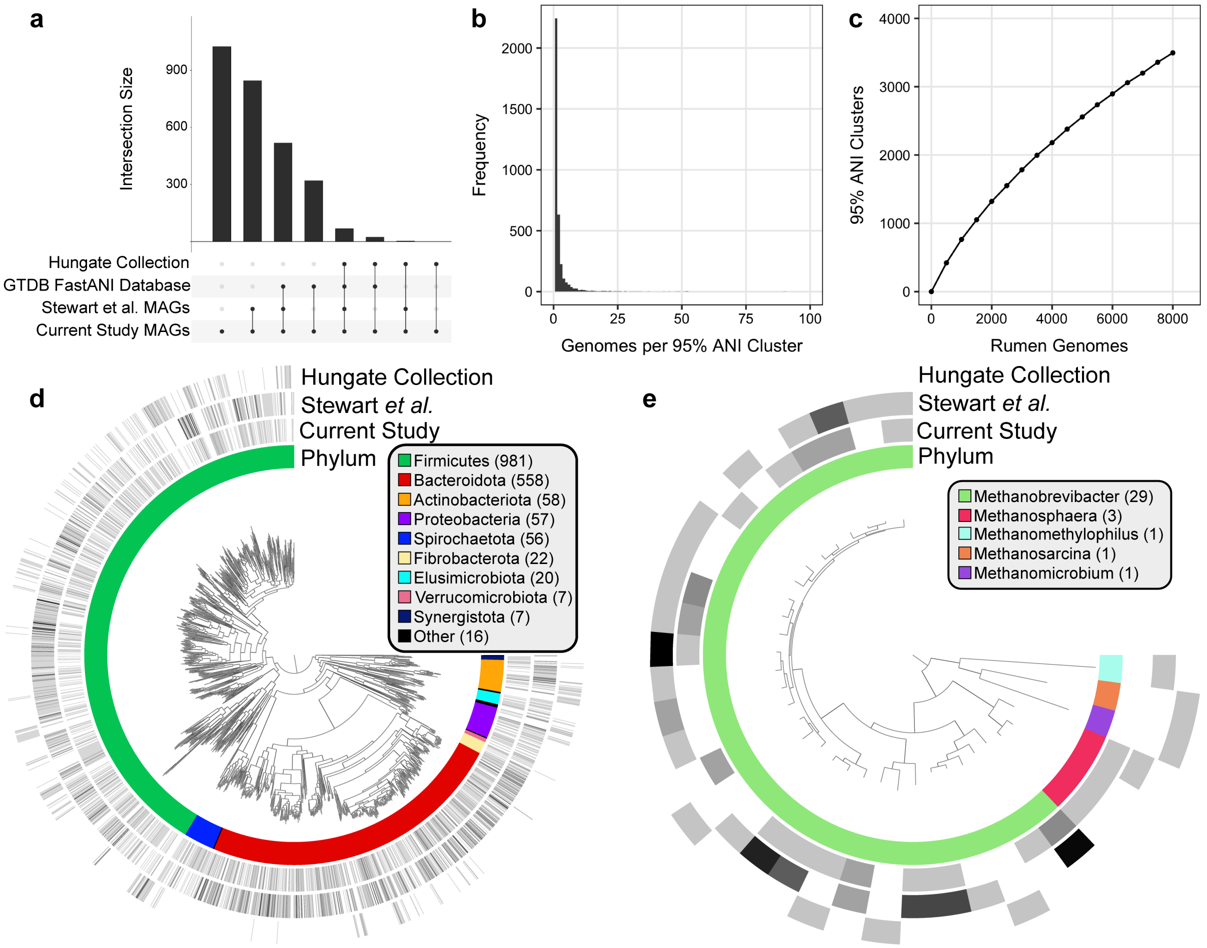

Fig. 3: Genomes sharing ≥95% ANI between databases and the characterization of rumen-specific 95% ANI clusters. a) The approximate number of species overlapping amongst rumen-specific and reference genomic datasets. Genomes demonstrating ≥95% ANI were considered to be shared between two datasets. Presented are a subset of intersections in which a MAG from the current study was the query genome. **b)** The number of genomes comprising each of the 3,541 95% ANI clusters generated from 8,160 rumen microbial genomes in the current study, the Hungate1000 Collection[3], and Stewart *et al.* studies[4,5]. **c)** Rarefaction analysis based on subsampling 95% ANI clusters at steps of 500 genomes indicates the 8,160 genomes from recently published rumen genomic collections still only represent a fraction of expected microbial species diversity in the rumen ecosystem. Phylogenomic relationships of the 1,781 near-complete bacterial **(d)** and 35 near-complete archaeal **(e)** representative genomes with the highest dRep genome quality score from the 3,541 95% ANI clusters generated from 8,160 rumen-specific genomes. Near-complete genomes were defined as being ≥90% complete, having ≤5% contamination, and contig N50 ≥15 kbp. Layers surrounding the genomic trees indicate the bacterial phyla or archaeal genera and the log normalized number of genomes from each rumen genomic collection belonging to the same 95% ANI cluster. The bacterial and archaeal phylogenetic trees are provided as Supplementary File 3 and Supplementary File 4, respectively.

We applied an additional clustering approach to identify the approximate number of species represented by the rumen-specific genomes assembled in this study, in the Hungate1000 Collection[3], and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Stewart *et al.*[4,5]. A 95% ANI threshold yielded 3,541 clusters from the combination of the datasets (Supplementary Table 3). Of the 3,541 clusters, 2,024 contained a MAG from the current study, and 1,135 were composed exclusively of MAGs from the current study. In comparison, 2,175 and 286 clusters were comprised of genomes from Stewart *et al.*[4,5] and the Hungate1000 Collection[3], respectively. The majority of 95% ANI clusters (2,166) are only comprised of a single genome (Fig. 3b). Furthermore, a rarefaction curve suggests the 8,160 genomes from the genomic collections analyzed here only represent a fraction of the estimated microbial species diversity in the rumen (Fig. 3c). The genome with the best dRep score from each cluster was used to generate a phylogenetic tree highlighting the species diversity within each rumen genomic collection and represents the vast diversity of rumen bacterial (Fig. 3d) and archaeal (Fig. 3e) genomes published to date.

As stated previously, the median genome size of reconstructed MAGs was 2.2 Mbp, smaller than the median size of genomes from the Hungate1000 project (3.1 Mbp)[3]. For the purpose of providing an assessment at a finer resolution, genome sizes of MAGs and Hungate1000 genomes[3] belonging to the same 95% ANI cluster were compared (Supplementary Figure 1). Adjusted sizes of MAGs and Hungate1000 genomes that are ≥95% complete displayed a regression coefficient of 0.96 with a slope of 0.86, indicating the binning process likely did not lead to extensive losses and systematic biases in the reconstructed genomes. Instead, it further highlights that current culturing approaches have not brought large portions of rumen microbial diversity into culture and putatively supports previous findings from the human gut that revealed genome-reduction in uncultured bacteria[23].

Rumen metagenome classification rates using reference and rumen-specific genomes

Utilizing an approach similar to Stewart *et al.*[4,5], we investigated the influence of MAGs on rates of metagenomic read classification. The baseline for read classification was the standard Kraken database containing bacterial, archaeal, fungal, and protozoal RefSeq genomes[24]. Each rumen-specific dataset was incrementally added to the Kraken RefSeq genomic database in the following order to build new databases: the Hungate1000 Collection[3], MAGs from Stewart *et al.*[4,5], and MAGs from the current study. Each individual and collective database was used for classification

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of sample reads that underpinned metagenomic binning and from a rumen metagenomic dataset not used in the reconstruction of MAGs[25]. MAGs from the current work classified more reads from deer, moose, and sheep metagenomes, while the more numerous MAGs from Stewart *et al.*[4,5] classified more reads from bison and cattle metagenomes (Fig. 4a). The addition of MAGs improves classification relative to databases primarily based on cultured isolates, like the Hungate1000 Collection[3] (Fig. 4b). Using the combination of all reference and rumen-specific genomes, the median classification rate on an independent set of cattle metagenomes was 62.6%.

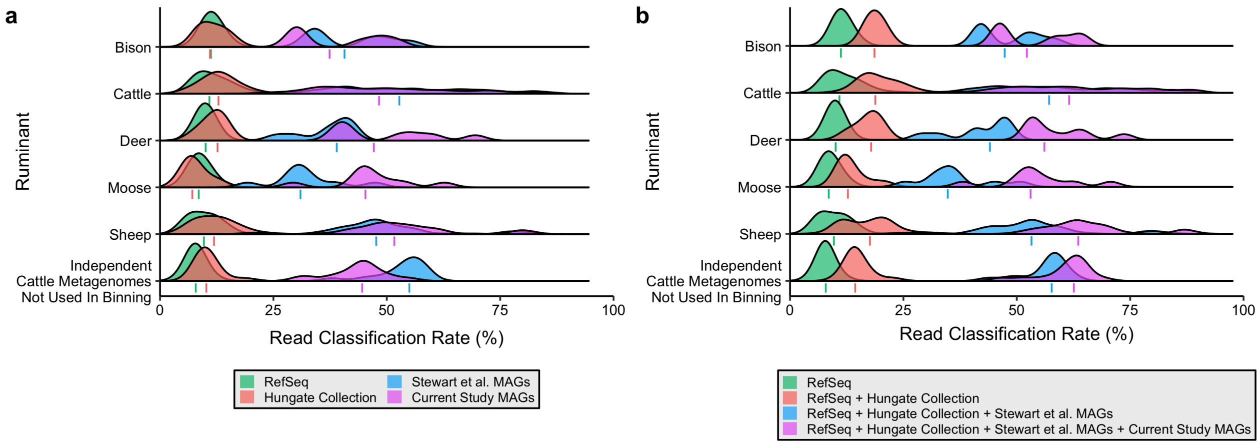

Fig. 4: Rumen MAGs improve metagenomic classification. Classification rates of reads from the 435 metagenomes used to bin MAGs and 16 cattle metagenomes not applied in genome binning using a combination of genomes from RefSeq, the Hungate1000 Collection[3], Stewart *et al.* studies[4,5], and the current study as databases. The four genomic databases were utilized to classify reads independently **(a)** or used to incrementally build larger databases for classification **(b)**. A database including rumen MAGs from the Stewart *et al.*[4,5] studies and the current study improved classifications rates for bison, cattle, deer, moose, sheep, and independent cattle metagenomes a median 33.3%, 42.1%, 40.9%, 40.1%, 45.0%, and 46.8% compared to a database of mainly isolate genomes from RefSeq and the Hungate1000 collection. The lines denote the median proportions of sample reads classified by the dataset or combinations of datasets.

Phylogenetic characterization of biosynthetic gene clusters

Microbial genome mining is a powerful tool for natural product discovery. We sought to explore the extent of secondary metabolite diversity coded by the MAGs in the current study, the Hungate1000 Collection[3], and Stewart *et al.* MAGs[4,5]. We identified 14,814 BGCs encoded by the 8,160 rumen-specific genomes using antiSMASH[26] (Fig. 5a and Supplementary Table 4). The majority of BGCs were nonribosomal peptide synthetases (NRPS, 5,346), followed by aryl

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

polyenes (2,800), sactipeptides (2,126), and bacteriocins (1,943). Only a few polyketide synthetases (PKS) were identified (75). Firmicutes harbored the vast majority of clusters for NRPS, sactipeptide, lantipeptide, lassopeptide, and bacteriocin synthesis (Fig. 5b). At lower taxonomic ranks, DTU089 (979), Bacteroidaceae (934), and Lachnospiraceae (923) coded for the bulk of NRPS gene clusters. Moreover, Acidaminococcaceae genomes contained 21.2% of identified bacteriocins and *Ruminococcus spp.* possessed the bulk of sactipeptides and lantipeptides. Archaea were predicted to code 737 biosynthetic gene clusters, including an average of 3.8 NRPS gene clusters per genome (Fig. 5a).

NRPS exhibit high molecular and structural diversity resulting in a wide array of biological activities. The diversity of NRPS, combined with their proteolytic stability and selective bioactivity, has resulted in the development of many NRPS as antimicrobials and other therapeutic agents[27]. Given the prevalence of NRPS among the recovered MAGs (Fig. 5a), the peptides appear to be important bioactive metabolites in the rumen. To gain fundamental insight into the phylogenetic diversity of rumen NRPS, we built a network based on BGC similarity using BiG-SCAPE[28]. BiG-SCAPE uses protein domain content, order, copy number, and sequence identity to calculate a distance metric. We assessed the similarity of NRPS gene clusters identified in Firmicutes, Bacteroidota, and Euryarchaeota, as these three phyla coded for 96.4% of assembled NRPS gene clusters from rumen genomes. With a BiG-SCAPE similarity threshold of 0.3, the resulting network consisted of 3,436 nodes (NRPS BGCs on contigs $\geq$10 kbp) and 79,112 edges (Fig. 5c and Supplementary Table 5). As expected, the network analysis depicted high inter- and intra-phylum genetic diversity among the NRPS gene clusters. The median intra-phylum, -family, and -genus similarity was 0.40, 0.44, and 0.46, respectively, while the median inter-phylum, -family, and -genus similarity was 0.32, 0.34, and 0.34, respectively. Further, only 2.6% of edges were inter-phylum and 69.0% were intra-family. Of the 6,594 Euryarchaeota edges, 8.1% were Euryarchaeota-Firmicutes (median similarity of 0.32) and 2.0% of edges were Euryarchaeota-Bacteroidota (median similarity of 0.31). To further examine the phylogenetic relationships of rumen Euryarchaeota NRPS, we clustered 265 NRPS gene clusters ($\geq$10 kbp) from 85 near-complete Euryarchaeota genomes at a higher similarity threshold of 0.75, yielding 57 NRPS clusters (Fig. 5d). The distribution of NRPS clusters amongst

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the genomes suggests there exists a strong relationship between methanogen phylogeny and NRPS similarity. Only *Methanobrevibacter* genomes contain NRPS gene clusters, and genomes of the same species often possessed many of the same NRPS clusters (see genomes highlighted in blue in Fig. 5d). However, there are instances in which closely related methanogens code for a contrasting pattern of NRPS clusters or no NRPS clusters at all (see genomes highlighted in red in Fig. 5d).

We aligned previously published rumen metatranscriptome data from steers characterized as having high and low feed efficiency to the BGCs to demonstrate if the identified BGCs are active and to explore potential ecological roles of secondary metabolites. Despite data from the metatranscriptome study not being applied to reconstruct genomes in the current study, we identified the expression of 554 gene clusters from rumen-specific genomes in the 20 metatranscriptomes ($\geq$100 aligned reads). Metatranscriptome read count data were normalized independently for each genome to better account for the variation in taxonomic composition across samples[29]. Genome-specific normalization resulted in the identification of 17 differentially expressed gene clusters between steers with high and low feed efficiency (DESeq2[30] false discovery rate adjusted P <0.05; Supplementary Table 6). Of the 17 differentially expressed BGCs, 16 exhibited higher expression levels in the rumen samples from less efficient steers with higher residual feed intake. Further, *Prevotella* and *Selenomonas* coded for 12 of the differentially expressed BGCs (70.6%). All of the differentially expressed *Selenomonas* BGCs were sactipeptides (n = 7), while the *Prevotella* BGCs were more diverse and included NRPS and aryl polyenes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

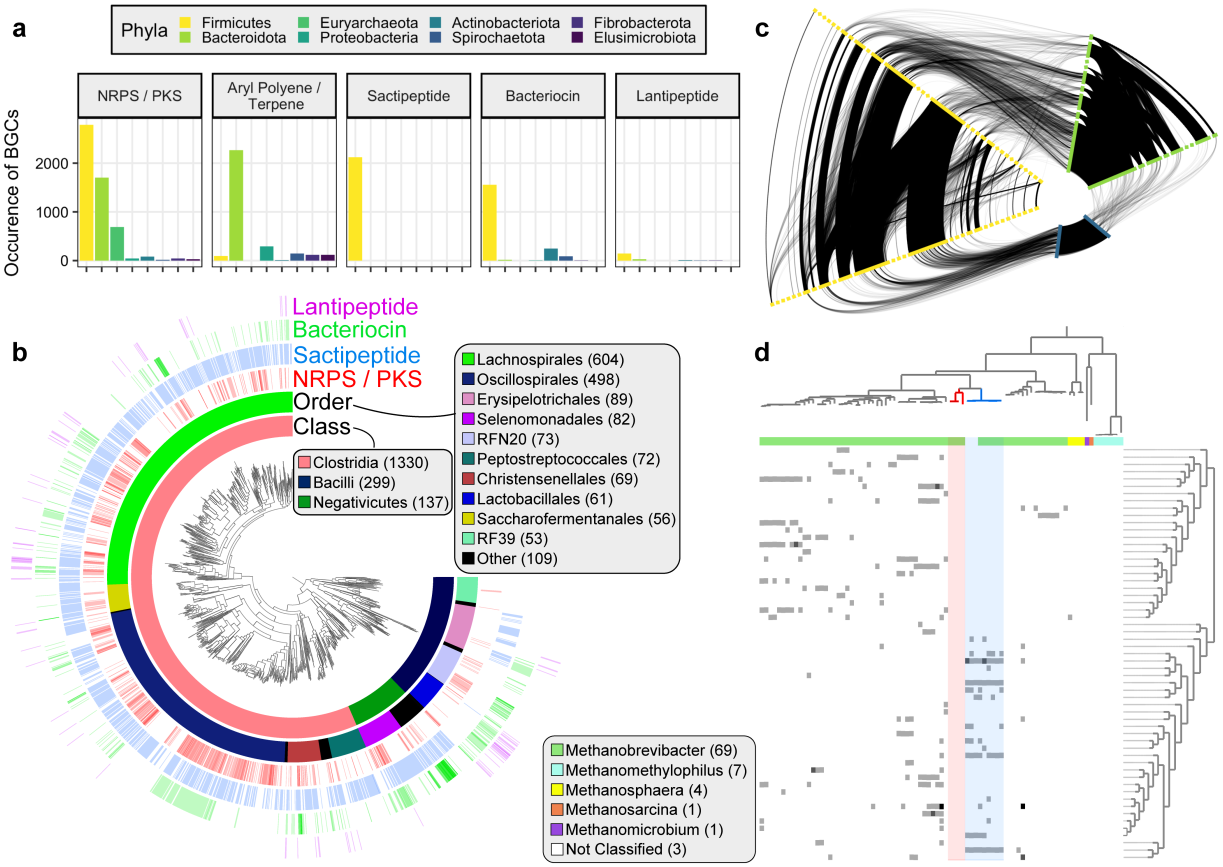

Fig. 5: Characterization of BGCs from 8,160 rumen genomes and MAGs. **a)** Number and types of BGCs identified from select phyla in genomes from the Hungate1000 Collection[3], Stewart *et al.* studies[4,5], and the current study. **b)** Phylogenomic analysis of 1,766 near-complete Firmicutes genomes inferred from the concatenation of phylogenetically informative proteins. The inner layer surrounding the genomic tree designates taxonomic annotations, while the remaining layers depict the log normalized number of BGCs in the genome with the ascribed function. Bacterial class and order labels are displayed for those lineages in which more than 50 genomes were identified. Near-complete genomes were defined as being ≥90% complete, having ≤5% contamination, and contig N50 ≥15 kbp. The phylogenetic tree is provided as Supplementary File 5. **c)** A relational network of NRPS gene cluster similarity in Firmicutes, Bacteroidota, and Euryarchaeota. Edge weight represents the similarity of two BGCs, as determined by BiG-SCAPE. Edges are only shown for BGCs with ≥0.3 BiG-SCAPE similarity. Nodes from each phylum are duplicated to illustrate intra-phylum relationships and nodes along a given axis are ordered alphabetically by taxonomic family. **d)** The association between genome phylogeny and the similarity of NRPS gene clusters coded by near-complete Euryarchaeota genomes. BGCs designated as NRPS were clustered with BiG-SCAPE. The relationship between NRPS clusters was portrayed through the hierarchical clustering of pairwise inter-cluster similarities. The number of NRPS clusters coded by each genome (range of 0-3) is presented alongside the assigned genus. A group of *Methanobrevibacter* genomes, likely of the same species (≥95% ANI), possessed very similar NRPS clusters (highlighted in blue). Yet, phylogenetically closely related genomes, belonging to two different 95% ANI clusters, did not code for any identified NRPS gene clusters (highlighted in red). The phylogenetic tree is based on the concatenation of 122 phylogenetically informative archaeal proteins and is available as Supplementary File 6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Microdiversity of BGCs and MAGs

Phylogenetic analyses of BGC often revealed high inter-species diversity (i.e, methanogen NRPS in Fig. 5d). We next investigated patterns of sub-species microdiversity in rumen BGCs. In order to reduce the influence of study-to-study effects, we focused on the microdiversity of MAGs across 282 metagenomes in the Stewart *et al.* studies[4,5]. MAGs with ≥50% of its genome covered by at least 5 reads were considered as detected in a sample and used for microdiversity analyses. The within-sample microdiversity of genes and genomes were assessed using InStrain[31]. Our phylogenetic analysis identified that different classes of BGCs are enriched in certain lineages (Fig. 5a and Fig. 5b). As a result, the nucleotide diversity values for genes were normalized using the mean genome-wide nucleotide diversity for each MAG to account for lineage-specific evolutionary processes and more accurately compare patterns of microdiversity in BGCs across lineages. There were significant differences in the nucleotide diversity of genes from the four major classes of BGCs identified in rumen-specific genomes (Kruskal-Wallis H = 1795.5, ε^2 = 0.001, P <2.2 × 10^{-16}; Fig. 6a), but the effect size (ε^2) between BGC types was negligible. Outliers with high microdiversity were bacteriocin genes from *RC9* and *UBA3207 sp.* as well as NRPS genes from *CAG-710* and *UBA9715 sp.* Additionally, we explored the association of genome-wide and secondary metabolism gene microdiversity with cattle breed. The mean nucleotide diversity of MAGs (Kruskal-Wallis H = 1027.5, ε^2 = 0.0265, P <2.2 × 10^{-16}; Fig. 6b) and the normalized nucleotide diversity of genes from BGCs (Kruskal-Wallis H = 403.84, ε^2 = 0.0003, P <2.2 × 10^{-16}; Fig. 6c) were both significantly different between the four breeds. The effect size (ε^2) of microdiversity difference between breeds was much larger for the genome-wide comparison than for genes from BGCs. This finding raised the question if genes from BGCs have different nucleotide diversity relative to other genes. We found that genes across all BGCs had lower normalized nucleotide diversity compared to all other genes from investigated MAGs (Wilcoxon rank-sum W = 6.11 × 10^{13}, Vargha and Delaney's A = 0.507, P <2.2 × 10^{-16}; Fig. 6d). The raw nucleotide diversity values were higher for genes in BGCs than other genes (Wilcoxon rank-sum W = 5.801 × 10^{13}, Vargha and Delaney's A = 0.481, P <2.2 × 10^{-16}). Regardless, again we find the effect size of the difference to be very small though. Together, microdiversity analyses suggest rumen microbial BGC diversity is comparable across the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

prevalent BGC classes, breeds, and similar to other genes.

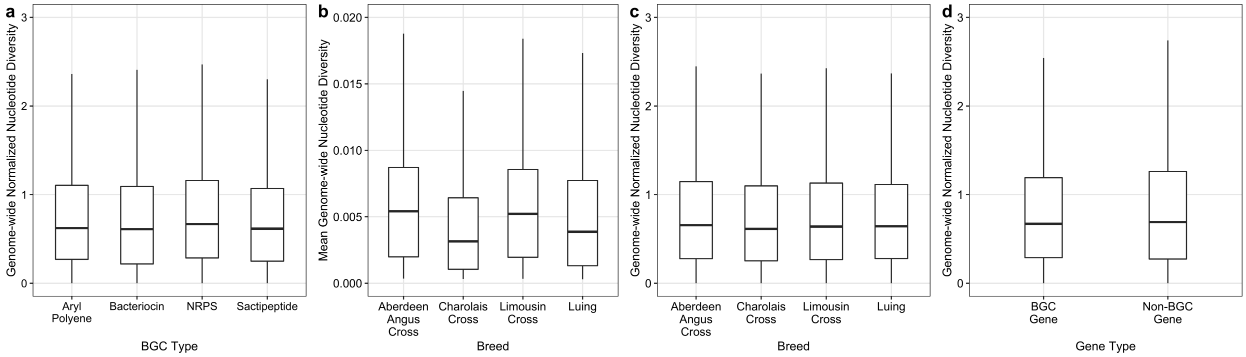

Fig. 6: Comparison of the microdiversity of MAGs and BGCs from cattle metagenomes in the Stewart *et al.* studies[4,5]. The within-sample nucleotide diversity of BGCs was statistically different between BGC types, but the effect size of the difference was small ($\varepsilon^2 = 0.001$) **(a)**. The difference in nucleotide diversity across breeds was greater for MAGs ($\varepsilon^2 = 0.0265$) **(b)** than for genes in BGCs ($\varepsilon^2 = 0.0003$) **(c)**. Additionally, the effect size of the difference between the normalized nucleotide diversity of genes from BGCs and other genes was small (Vargha and Delaney's $A = 0.507$) **(d)**. Genome-wide normalized nucleotide diversity is the nucleotide diversity of a gene relative to the mean nucleotide diversity of the MAG. The genome-wide normalized nucleotide diversity metric was used to reduce the influence of lineage-specific evolutionary processes, allowing for a more accurate comparison of gene nucleotide diversity across microbial populations. The same conclusions were identified using the raw nucleotide diversity in the place of genome-wide normalized nucleotide diversity. The outlier points have been removed from the boxplots for clarity.

Discussion

Ruminant agriculture is critical to the global food system. However, with land constraints and associated environmental impacts, ruminant production systems will need to become more efficient and sustainable to feed a growing population. Due to the importance of microbial processes in ruminant health and production, rumen microbes are central to nearly all aspects of ruminant agriculture[32]. Actionable insights into the roles of rumen microbes have lagged though, partly due to a lack of genomic references that underpin analyses and contextualize community data.

We reconstructed 2,809 metagenome-assembled population genomes from several ruminant species to advance our understanding of the structure-function relationship of the rumen microbial ecosystem. Nearly half of the MAGs are estimated to be ≥90% complete with minimal contamination. Based on pairwise ANI comparisons, the MAGs in this study constitute approximately 2,024 species

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(95% ANI clusters), greatly expanding the genomic representation of rumen microbial lineages. Moreover, clustering the genomic data reported in this study with genomes from the Hungate1000 Collection[3] and Stewart *et al.* studies[4,5] suggest there are at least 3,541 rumen microbial species with a draft reference genome now. It is worth emphasizing that some of the MAGs reported in this study may have been reported as part of metagenomic binning efforts in previous studies (Supplementary Table 1). However, the aggregating of data from multiple studies and contrasting assembly and binning approaches implemented in the current study may have yielded improved bins.

Approximately one-third of the resolved MAGs did not have a species-level representative in the compared genomic databases. Among the fraction of genomes that did exhibit high similarity, only 3.7% of MAGs formed a cluster with 29.8% of genomes in the Hungate1000 Collection[3]. Further, 64.6% of the Hungate1000 Collection genomes did not cluster with a MAG from the current study or the Stewart *et al.* studies[4,5], implying metagenomic binning did not recover some of the cultured rumen isolates. The poor reconstruction of isolate genomes may be because the Hungate1000 strains are in low abundance in vivo or have high intra-species diversity. An examination of culturing with defined and undefined media found that a large number of cultured OTUs were not detected in the 16S rRNA gene profile from the same rumen sample or were unique to a culture plate, suggesting cultured OTUs often constituted rare rumen microbial populations[6]. The addition of rumen MAGs to classification indices may improve statistical power and allow for a more accurate interpretation of shallow rumen metagenomic datasets[33]. Therefore, the MAGs presented here are valuable for interpreting future and previously sequenced rumen metagenomic datasets and serving as a scaffold for other multi-omics data.

Moreover, we genomically linked microbial populations to the coding and expression of BGCs to demonstrate the utility of genome-resolved metagenomics in the rumen ecosystem. This analysis identified 14,814 gene clusters from 8,160 rumen-specific genomes, indicating the rumen is a rich resource for secondary metabolites. Previous investigations of rumen secondary metabolites have primarily focused on bacteriocin production. A similar genome mining approach revealed 46 bacteriocin gene clusters from 33 rumen bacterial strains[15]. Roughly half the clusters were related

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to lantipeptide biosynthesis. In this study, we have considerably expanded the phylogenetic diversity of known rumen bacteriocins and related peptides, identifying 4,326 putative bacteriocins, sactipeptide, lantipeptide, and lassopeptide clusters. The recovered MAGs increase the number of bacteriocins native to the rumen ecosystem and available for targeted isolation and functional screening to develop novel probiotics and alternatives for antibiotics in ruminant production.

Given the abundance of NRPS gene clusters harbored by recovered genomes, we explored the diversity of this family of natural products through a relational network based on the BiG-SCAPE implemented distance metric. The network analysis confirmed that NRPS BGCs have immensely diverse gene content and highlighted that approximately 70% of the network edges were between BGCs of the same taxonomic family. We further identified 687 NRPS gene clusters encoded by 125 archaeal genomes. Archaeal NRPS have been described previously, notably in *Methanobrevibacter ruminantium*[34]. A 2014 genomic survey found only three instances of archaeal NRPS in classes Methanobacteria and Methanomicrobia[35], and a recent analysis identified 73 BGCs from 203 archaeal genomes[36]. Phylogenetic analyses suggest that archaeal NRPS were acquired through horizontal transfer from bacteria[34,35]. Our network analysis appears to support this hypothesis as we established that there are NRPS in Euryarchaeota with high similarity to NRPS in different Firmicutes families. Given the proposed roles of NRPS in signaling and intercellular communication in ecosystems, it has been suggested that methanogen NRPS may be involved in perpetrating syntrophic interactions that are important for interspecies hydrogen transfer[34]. We noted methanogen genomes of the same species often contain very similar NRPS gene clusters, while other closely related genomes could lack NRPS gene clusters altogether. As such, we hypothesize that methanogens without NRPS may typically exist as symbionts of protozoa or other microbes and have lost the need to produce the compound. It is difficult to confidently assess the expression of populations in low abundance, but future work should aim to establish the expression patterns of methanogen NRPS gene clusters under various conditions. In addition to predicting thousands of BGCs from MAGs, we also demonstrated a subset of BGCs that were expressed in rumen samples from high and low efficient steers. The differentially expressed BGCs were mainly sactipeptides encoded by *Selenomonas* and aryl polyenes and NRPS encoded by *Prevotella*. Host-associated microbes may

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mediate important interactions through the production of secondary metabolites[37]. *Prevotella* and *Selenomonas* populations are often linked to feed efficiency. Our approach using genome-resolved metagenomics and organism-specific normalization suggests secondary metabolites may play a role in this association. Further, the findings fit the emerging hypothesis that inefficient cattle have higher microbial diversity and produce a broader range of less usable metabolites for the animal's energy needs[38,39].

Inter-species diversity of BGCs appeared to be high in the rumen, while sub-species microdiversity analyses suggest strain-level BGC diversity may be more constant across samples. The majority of genes within BGCs had similar nucleotide diversity as other genes, with a few outliers that displayed very high diversity. We know little regarding the relationship between genetic and functional diversity of BGCs in the rumen. As such, future work may focus on obtaining a better understanding of the evolutionary processes shaping the microdiversity patterns of BGCs. The mean genome-wide nucleotide diversity of sub-species MAGs was more different across breeds than it was for genes of BGCs, suggesting host genetics may influence microdiversity.

In this study, we have provided a phylogenomic characterization of rumen-specific genomes that may serve as a foundation for future in silico and laboratory experiments to better explore the rumen as a source for alternative peptides and metabolites to modulate rumen fermentation. The genomes reported here and in other recent genetic explorations of the rumen microbiome appear to only provide a glimpse into rumen microbial diversity. Moving forward, we anticipate using the combination of cultured and uncultured genomes to populate a bottom-up systems biology framework that advances towards mechanistic understandings and modeling dynamics of the rumen microbial ecosystem.

Methods

Rumen metagenomic datasets

We used 435 metagenomes for assembly and metagenomic binning (Supplementary Table 1). Rumen metagenomic studies with sufficient depth and quality were identified from the Sequence Read

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Archive, European Nucleotide Archive, and MG-RAST in early 2018. All publicly available metagenomes were sequenced on Illumina next-generation sequencing platforms. The remaining metagenomic datasets were previously unpublished.

The first two unpublished metagenomic datasets were from an 84-day growing study utilizing 120 steers and subsequent 125-day finishing study with 60 steers at the University of Nebraska Agriculture Research and Development Center, as described previously[40]. The University of Nebraska-Lincoln Institutional Animal Care and Use Committee approved animal care and management procedures. From the original 120 animals in the growing study, 23 animals across different treatment groups were randomly selected for metagenomic sequencing. Sixty of the steers were utilized in a finishing study to evaluate the influence of dietary nitrate and sulfate on methane emissions and animal performance. From this study, 27 animals across different treatment groups were selected randomly for metagenomic sequencing. In both studies, sampling was conducted via esophageal tubing and snap-frozen with liquid nitrogen. Total DNA was extracted from rumen samples with the PowerMax Soil DNA Isolation Kit (MO BIO Laboratories, Inc.) according to the manufacturer's protocols. Metagenomes were prepared with the Nextera XT DNA Library Prep Kit and sequenced on the Illumina HiSeq platform using 150 bp paired-end sequencing. Raw data from these two datasets is associated with NCBI BioProject PRJNA627299 (Supplementary Table 1).

Paz et al. characterized the rumen microbiomes of 125 heifers and 122 steers to identify bacterial operational taxonomic units linked to feed efficiency[41]. From this cohort, 16 steers displaying divergent feed efficiency phenotypes were selected for metagenomic sequencing. In brief, rumen samples were collected through esophageal tubing and snap-frozen in liquid nitrogen. Total DNA was extracted from rumen samples with the PowerMax Soil DNA Isolation Kit (MO BIO Laboratories, Inc.) according to the manufacturer's protocols. Metagenomes were prepared using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) and sequenced on the Illumina MiSeq platform (600 cycles, MiSeq Reagent Kit v3). Raw data from this study is associated with NCBI BioProject PRJNA627251 (Supplementary Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Quality control of metagenomes

Initial quality control of sequencing reads and adapter trimming were performed using BBDuk of the BBTools software suite (version 38.16; parameters: ktrim=r, k=23, mink=11, hdist=1)[42]. VSEARCH (version 2.0.3) was used to remove sequences based on the presence of ambiguous bases (–fastq_maxns 0), minimum read length (range from –fastq_minlen 36 to –fastq_minlen 100 based on sample median read length and sequencing technology), and the maximum expected error rate (–fastq_maxee_rate 0.02 or –fastq_maxee_rate 0.025 depending on quality of the sequencing data and technology)[43]

Assembly and metagenomic binning

Paired-end and single-end sequences from each sample were assembled independently with MEGAHIT (version 1.1.1; parameters: –min-contig-len 1000, –k-min 27, –k-step 10)[44]. No co-assemblies were performed. We applied a maximum k-mer size of 87 (–k-max 87) for samples in which the longest read length was ≤100 bp. For samples with longer read lengths, we employed a maximum k-mer size of 127 (–k-max 127). The single-sample assemblies were input for both single-sample and multi-sample binning strategies with MetaBAT followed by re-assembly and dereplication[45]. Reads from each sample were mapped to assembled contigs (minimap2, parameters: –ax sr)[46]. The resulting alignments were used to bin contigs with a minimum length of 2,000 bp for single-sample binning and 2,500 for multi-sample binning strategies. Due to the total size of the collected datasets, the multi-sample binning was conducted independently for cattle (335 metagenomes) and other ruminant metagenomic datasets (100 metagenomes). Estimates of the completeness and contamination of the resulting bins were assessed using the lineage-specific workflow (lineage_wf) of CheckM (version 1.0.11)[47]. Bins ≥50% complete were re-assembled with SPAdes (version 3.13.0; –careful parameter)[48]. MAGs stemming from the single-sample binning pipeline were re-assembled only with reads from that same sample. MAGs reconstructed through multi-sample binning were re-assembled from the sample with the most reads aligning to the bin and from all reads aligning to the bin. The quality of re-assembled bins was assessed with CheckM. The best assembly (original or re-assembly) was retained based on the dRep quality

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

score, where Genome Quality = Completeness − (5 · Contamination) + (Contamination · (Strain Heterogeneity / 100)) + 0.5 · (log(N50)[19]. Contigs with divergent genomic properties (GC content and tetranucleotide frequency) were identified and removed with RefineM to reduce genome bin contamination[22]. Refined genomes from single-sample and multi-sample binning strategies were pooled and dereplicated with dRep at a threshold of 99% ANI[19]. Genomes meeting the following thresholds were retained: dRep quality score ≥60; N50 ≥5 kbp; ≤500 contigs; genome size ≥500 kbp; CheckM contamination estimate ≤10%; and CheckM completeness estimate ≥75%. Near-complete genomes were defined as MAGs with CheckM completeness estimate ≥90%, CheckM contamination estimate ≤5%, and N50 ≥15 kbp.

Taxonomic and functional annotations of MAGs

Taxonomy was assigned to MAGs using the classify workflow (classify_wf) of the Genome Taxonomy Database Toolkit (GTDB-Tk 0.2.2) with the associated Genome Taxonomy Database (release 86 v3)(Parks 2018). In short, the GTDB-Tk classifies each genome based on ANI to a curated collection of reference genomes, placement in the bacterial or archaeal reference genome tree, and relative evolutionary distance. For consistency, genomes from the Hungate1000 project[3] and Stewart *et al.*[4,5] were also assigned taxonomy with the GTDB-Tk. Depending on the taxonomic annotation, MAGs were functionally annotated with Prokka by evoking either the –kingdom Bacteria or –kingdom Archaea parameter (version 1.13.7)[49]. Prokka annotations were used to sum the number and types of tRNAs and rRNAs in each MAG (Supplementary Table 2).

Inference of genome trees

Phylogenetic trees were inferred with near-complete genomes (CheckM completeness estimate ≥90%, CheckM contamination estimate ≤5%, and N50 ≥15 kbp) using the GTDB-Tk (default parameters for the identify, align, and infer commands). Anvi'o was used to visualize the resulting Newick trees and associated metadata (version 5.5)[50]. We estimated how well a MAG was represented in a sample by calculating the percent of a MAG's bases with at least 1X coverage in the sample. The mean number of bases in a MAG with at least 1X coverage is presented for each

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

metagenomic study and was used to compute the hierarchical clustering of rumen metagenomic datasets (Euclidean distance and Ward linkage).

Similarity of reconstructed MAGs to GTDB reference genomes, the Hungate1000 Collection, and rumen-specific MAGs

Recent analyses support a 95% ANI threshold to delineate microbial species[51,52]. The ANI values of MAGs from the current study and genomes from the GTDB (a curated and dereplicated collection of 22,441 genomes in the GTDB-Tk FastANI database[21]), Hungate1000 project[3], and Stewart *et al.*[4,5] were compared in a pairwise fashion with FastANI (version 1.1)[51]. Genome pairs with ≥95% ANI were denoted as overlapping species between the datasets. We visualized the number of overlapping genomes between each pair of datasets with UpSetR[53,54]. Additionally, genomes from the current study, the Hungate1000 project[3], and Stewart *et al.*[4,5] were clustered at 95% ANI thresholds with dRep[19] to approximate the number of microbial species represented across the rumen genomic collections. The number of genomes belonging to each 95% ANI cluster was used to calculate rarefaction curves in which cluster counts were subsampled without replacement at steps of 500 genomes with 10 replications at each step (QIIME version 1.9)[55].

The average genome size of reconstructed MAGs was smaller than was observed in the Hungate1000 Collection. In order to provide a better comparison of genome sizes across similar species, we evaluated the adjusted genome sizes of MAGs and Hunagte1000 Collection genomes that belonged to the same 95% ANI cluster based on Pearson correlation and linear regression, where Adjusted Genome Size = Genome Size / (Completeness + Contamination).

Classification of Metagenomic Reads

Reads from the 435 rumen metagenomes used to assemble MAGs and reads from 16 samples of an independent cattle metagenomic dataset[25] not used in binning were classified with different databases to assess the value of the reconstructed MAGs to improve metagenomic read classification. Reads were classified with Kraken2 (version 2.0.7; default parameters)[24] using a combination of the Kraken2 standard database containing bacterial, archaeal, fungal, and protozoa RefSeq genomes,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

410 genomes from the Hungate1000 project[3], 4,941 MAGs from Stewart *et al.*[4,5], and the 2,809 MAGs from the current study.

Phylogenetic analysis of biosynthetic gene clusters

BGCs were identified within MAGs, the Hungate1000 collection[3], and the Stewart *et al.* MAGs[4,5] using antiSMASH 4.0)[26]. A network was constructed based on the BiG-SCAPE calculated distances between two BGCs (version "20190604")[28]. In short, BiG-SCAPE combines three approaches to measure the similarity of BGC pairs: 1) the Jaccard Index, which measures the percentage of shared domain types; 2) the Domain Sequence Similarity index that takes into account differences in Pfam domain copy number and sequence identity; 3) the Adjacency Index, a measure of the pairs of adjacent domains that are shared between BGCs. The raw BiG-SCAPE distances were converted to similarities for all analyses. Only NRPS $\geq$10 kbp (71.6% were $\geq$10 kbp) were evaluated and the network analysis was limited to Bacteroidota, Firmicutes, and Euryarchaeota phyla because these three phyla coded for 96.4% of NRPS gene clusters. Two BGCs (nodes in the network) were connected with an edge if the pairwise similarity was $\geq$0.3. We visualized the network as a hive plot with the R tidygraph package to demonstrate the inter- and intra-phylum diversity of NRPS BGCs. Nodes on an axis were ordered by the family of the genome coding the NRPS. Archaeal NRPS were further evaluated by placing BGCs into clusters based on a BiG-SCAPE glocal similarity threshold of 0.75. The distance between clusters was calculated as the mean pairwise similarity between the BGCs of two clusters. The resulting distance matrix was clustered with hierarchical clustering to produce a Newick tree (Euclidean distance and Ward linkage). The number of NRPS from near-complete archaeal genomes (CheckM completeness estimate $\geq$90%, CheckM contamination estimate $\leq$5%, and N50 $\geq$15 kbp) that belong to each BiG-SCAPE cluster were tabulated and visualized alongside a phylogenetic tree inferred with the GTDB-Tk (default parameters for the identify, align, and infer commands)[21,47]. The data were visualized with Anvi'o (version 5.5)[50].

Rumen metatranscriptomic data[56] sequenced from steers with high (10 samples) and low (10 samples) residual feed intake were used to assess the expression of rumen microbial BGCs. ORF abundances for all rumen genomes were quantified with kallisto (version 0.45.0; default parameters)[57].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Kallisto generates pseudo-alignments based on exact k-mer matches. Differences in expression may be attributed to both variations in organism abundance and changes in microbial behavior under different conditions. Taxon-specific scaling of count data should reduce the influence of taxonomic composition changes[29]. Thus, to account for variations in taxonomic composition, count data for each genome were first partitioned and normalized separately with DESeq2 (version 1.24.0)[30]. The genome-specific normalization factors were used to scale raw BGC abundances from the same genome. Normalized BGC counts from each genome were re-combined to identify differentially expressed clusters between steers with high and low feed efficiency with DESeq2[30]. Only genomes with at least one read in all 20 samples (6,630 genomes) and BGCs with a minimum count of 100 reads were included in the analysis (648 BGCs).

Microdiversity analyses were carried out with InStrain (version 1.2.4)[31]. Reads from the 282 Illumina metagenomes described in Stewart *et al.* were mapped to the 4,941 MAGs previously recovered[4,5]. MAGs with an unmasked breadth $\geq$0.5 (i.e., $\geq$50% of the genome has 5X coverage) in a sample were considered to be present in that sample. That is, only genes from detected MAGs were used in subsequent analyses. Of the 4,941 MAGs, 2,926 had an unmasked breadth $\geq$0.5 in at least one sample. Further, genes were only considered present if they had $\geq$5X coverage in a sample in which the MAG was detected. The profile module of InStrain calculates the nucleotide diversity of scaffolds within a given sample. InStrain can use this profile to calculate the nucleotide diversity of genes (profile_genes module) and the mean nucleotide diversity of the genome (genome_wide module). The nucleotide diversity of detected genes was normalized based on the mean genome-wide microdiversity of the MAG (gene nucleotide diversity / genome-wide microdiversity) to reduce lineage-specific effects when comparing the microdiversity of BGCs. The normalized gene nucleotide diversity represents the nucleotide diversity of the gene relative to the nucleotide diversity of the rest of the genome. Statistical differences were assessed with Kruskal-Wallis and Wilcoxon rank-sum tests. All statistical comparisons were also carried out using raw nucleotide diversity values.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data availability

The accessions for all metagenomes analyzed are available in Supplementary Table 1. Metagenomes previously not publicly available were deposited under NCBI BioProject PRJNA627299 and PRJNA627251. The 2,809 reconstructed MAGs are available at: https://doi.org/10.6084/m9.figshare.12164250.

Acknowledgments

This work was completed using the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative. This study was partially supported by USDA National Institute of Food and Agriculture Fellowship 2016-67011-24783 (C.L.A.). This work was also supported by Animal Nutrition, Growth and Lactation grant no. 2018-67015-27496, Effective Mitigation Strategies for Antimicrobial Resistance grant no. 2018-68003-27545, and Multi-state research project accession no. 1000579 from the USDA National Institute of Food and Agriculture awarded to S.C.F.

Author information

Contributions

C.L.A. designed the study, carried out the analyses, and wrote the manuscript. S.C.F. designed the study, interpreted results, and wrote the manuscript.

Corresponding authors

Samodha C. Fernando
Department of Animal Science
University of Nebraska-Lincoln
Lincoln, NE
Phone: 402-472-0518

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

E-mail: samodha@unl.edu

Ethics declarations

The authors declare no competing financial interests.

Supplementary information

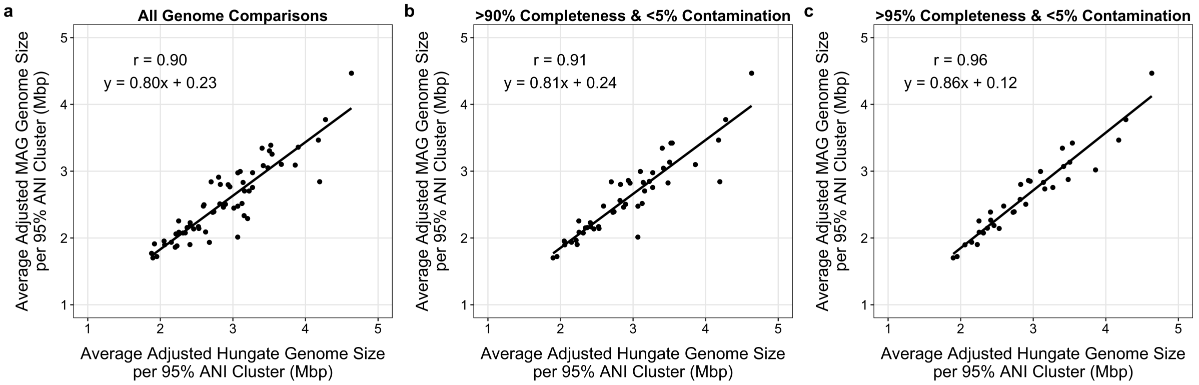

Fig. S1: The genome sizes of MAGs are largely congruent with genome sizes of isolates from the Hungate1000 Collection in the same 95% ANI cluster. Pearson correlation and linear regression analysis were calculated to compare the average adjusted genome size for all MAGs and Hungate1000 genomes of the same 95% ANI cluster **(a)**, those genomes with ≥90% completeness and ≤5% contamination from the same 95% ANI cluster **(b)**, and those genomes with ≥95% completeness and ≤5% contamination from the same 95% ANI cluster **(c)**.

Supplementary Table 1. Characteristics and sources of rumen metagenomic datasets used for the reconstruction of MAGs.

Supplementary Table 2. Taxonomy and genomic properties of recovered rumen MAGs.

Supplementary Table 3. Results from the clustering of MAGs from the current study, the Hungate1000 Collection, and Stewart *et al.* studies based on 95% ANI thresholds with dRep.

Supplementary Table 4. BGC predictions for MAGs presented in the current study, genomes from the Hungate1000 Collection, and MAGs from the Stewart *et al.* studies.

Supplementary Table 5. Nodes and edges of the relational network based on BiG-SCAPE

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

defined similarity between NRPS gene clusters of Bacteroidota, Firmicutes, and Euryarchaeota.

Supplementary Table 6. Genome-specific normalized counts and DESeq2 results for 648 BGCs with $\geq$100 reads and encoded by genomes with at least one read in all 20 metatranscriptomes.

Supplementary File 1. Phylogenetic tree of 1,163 near-complete bacterial MAGs recovered in the current study (Newick format).

Supplementary File 2. Phylogenetic tree of 20 near-complete archaeal MAGs recovered in the current study (Newick format).

Supplementary File 3. Phylogenetic tree of 1,781 near-complete bacterial genomes that were representative genomes of 95% ANI clusters formed from 8,160 rumen-specific microbial genomes (Newick format).

Supplementary File 4. Phylogenetic tree of 35 near-complete archaeal genomes that were representative genomes of 95% ANI clusters formed from 8,160 rumen-specific microbial genomes (Newick format).

Supplementary File 5. Phylogenetic tree of 1,766 near-complete Firmicutes genomes identified from the 8,160 rumen-specific microbial genomes (Newick format).

Supplementary File 6. Phylogenetic tree of 85 near-complete archaeal genomes identified from the 8,160 rumen-specific microbial genomes (Newick format).

References

1. Hunter, M. C., Smith, R. G., Schipanski, M. E., Atwood, L. W. & Mortensen, D. A. Agriculture in 2050: Recalibrating Targets for Sustainable Intensification. *BioScience* **67**, 386–391 (2017).

2. Bergman, E. N. Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiol. Rev.* **70**, 567–590 (1990).

3. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).

5. 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 (2019).

6. Zehavi, T., Probst, M. & Mizrahi, I. Insights Into Culturomics of the Rumen Microbiome. *Front. Microbiol.* **9** (2018).

7. Solden, L. M. *et al.* New roles in hemicellulosic sugar fermentation for the uncultivated Bacteroidetes family BS11. *ISME J.* **11**, 691–703 (2017).

8. 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).

9. Tyc, O., Song, C., Dickschat, J. S., Vos, M. & Garbeva, P. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends in Microbiology* **25**, 280–292 (2017).

10. Marshall, B. M. & Levy, S. B. Food Animals and Antimicrobials: Impacts on Human Health. *Clin Microbiol Rev* **24**, 718–733 (2011).

11. Woolhouse, M., Ward, M., van Bunnik, B. & Farrar, J. Antimicrobial resistance in humans, livestock and the wider environment. *Philos Trans R Soc Lond B Biol Sci* **370** (2015).

12. Cheng, G. *et al.* Antibiotic alternatives: The substitution of antibiotics in animal husbandry? *Front Microbiol* **5** (2014).

13. Boeckel, T. P. V. *et al.* Reducing antimicrobial use in food animals. *Science* **357**, 1350–1352 (2017).

14. Oyama, L. B. *et al.* The rumen microbiome: An underexplored resource for novel antimicrobial discovery. *NPJ Biofilms Microbiomes* **3** (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

15. Azevedo, A. C., Bento, C. B. P., Ruiz, J. C., Queiroz, M. V. & Mantovani, H. C. Distribution and Genetic Diversity of Bacteriocin Gene Clusters in Rumen Microbial Genomes. *Appl Environ Microbiol* **81**, 7290–7304 (2015).

16. Bernier, S. P. & Surette, M. G. Concentration-dependent activity of antibiotics in natural environments. *Front Microbiol* **4** (2013).

17. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

18. 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.e20 (2019).

19. 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 dereplication. *ISME J* **11**, 2864–2868 (2017).

20. 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).

21. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* (2018).

22. 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).

23. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505 (2019).

24. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

25. Sandri, M., Licastro, D., Monego, S. D., Sgorlon, S. & Stefanon, B. Investigation of rumen metagenome in Italian Simmental and Italian Holstein cows using a whole-genome shotgun sequencing technique. *Ital. J. Anim. Sci.* **0**, 1–9 (2018).

26. Blin, K. *et al.* antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36–W41 (2017).

27. Felnagle, E. A. *et al.* Nonribosomal Peptide Synthetases Involved in the Production of Medically Relevant Natural Products. *Mol Pharm* **5**, 191–211 (2008).

28. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* **16**, 60–68 (2020).

29. Klingenberg, H. & Meinicke, P. How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ* **5** (2017).

30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

31. Olm, M. R. *et al.* InStrain enables population genomic analysis from metagenomic data and rigorous detection of identical microbial strains. *bioRxiv* https://doi.org/10.1101/2020.01.22.915579 (2020).

32. Huws, S. A. *et al.* Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future. *Front. Microbiol.* **9** (2018).

33. Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. Correcting index databases improves metagenomic studies. *bioRxiv* https://doi.org/10.1101/712166 (2019).

34. Leahy, S. C. *et al.* The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. *PLoS One* **5**, e8926 (2010).

35. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc*

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Natl Acad Sci U S A **111**, 9259–9264 (2014).

36. Wang, S., Zheng, Z., Zou, H., Li, N. & Wu, M. Characterization of the secondary metabolite biosynthetic gene clusters in archaea. *Computational Biology and Chemistry* **78**, 165–169 (2019).

37. Donia, M. S. *et al.* A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics. *Cell* **158**, 1402–1414 (2014).

38. Shabat, S. K. B. *et al.* Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. *ISME J.* **10**, 2958–2972 (2016).

39. Moraïs, S. & Mizrahi, I. The Road Not Taken: The Rumen Microbiome, Functional Groups, and Community States. *Trends in Microbiology* **27**, 538–549 (2019).

40. Pesta, A. *Dietary Strategies for Mitigation of Methane Production by Growing and Finishing Cattle*. Ph.D. thesis, University of Nebraska (2015).

41. Paz, H. A. *et al.* Rumen bacterial community structure impacts feed efficiency in beef cattle. *J. Anim. Sci.* **96**, 1045–1058 (2018).

42. Brian Bushnell. BBMap. sourceforge.net/projects/bbmap/.

43. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

44. 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).

45. 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).

46. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.105130; this version posted May 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

47. 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).

48. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

49. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

50. Eren, A. M. *et al.* Anvi'o: An advanced analysis and visualization platform for ‘omics data. *PeerJ* **3**, e1319 (2015).

51. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).

52. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5** (2020).

53. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* **20**, 1983–1992 (2014).

54. 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).

55. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* **7**, 335–336 (2010).

56. Li, W. *et al.* Metagenomic analysis reveals the influences of milk containing antibiotics on the rumen microbes of calves. *Arch Microbiol* **199**, 433–443 (2017).

57. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).