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

Running title: The rumen metagenome across four ruminant species

Authors: Laura Glendinning [1]*, Buğra Genç[2#], R John Wallace[2], Mick Watson[1]

[1]Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, Midlothian, United Kingdom

[2]Rowett Institute, University of Aberdeen, Ashgrove Rd W, Aberdeen, Aberdeenshire, United Kingdom

#Current address: Ondokuz Mayis University Veterinary Medicine Faculty Department of Laboratory Animals, Atakum, Samsun, Turkey

*Corresponding author: Laura Glendinning (laura.glendinning@roslin.ed.ac.uk). The Roslin Institute, Easter Bush Campus, Midlothian, United Kingdom, EH25 9RG. +44 0131 6519220.

Abstract:

Background: The rumen microbiota comprises a community of microorganisms which specialise in the degradation of complex carbohydrates from plant-based feed. These microbes play a highly important role in ruminant nutrition and could also act as sources of industrially useful enzymes. In this study, we performed a metagenomic analysis of samples taken from the ruminal contents of cattle (*Bos Taurus*), sheep (*Ovis aries*), reindeer (*Rangifer tarandus*) and red deer (*Cervus elaphus*).

Results: We constructed 391 metagenome-assembled genomes originating from 16 microbial phyla. We compared our genomes to other publically available microbial genomes and found that they contained 279 novel species. We also found significant differences between the microbiota of different ruminant species in terms of the abundance of microbial taxonomies, carbohydrate-active enzyme genes and KEGG orthologs. However, we found that the vast majority of carbohydrate-active enzyme were present in all of our sample types, which may indicate that there is a core set of these enzymes which are present across ruminants and are independent of diet and environmental conditions.

Conclusions: We present a dataset of rumen-derived genomes which in combination with other publicly-available rumen genomes can be used as a reference dataset in future metagenomic studies.

31 **Introduction:**

32 The microbial communities which inhabit the rumen contain a mixture of bacteria, fungi, protozoa,
33 viruses and archaea and through fermentation are able to convert complex plant carbohydrates into
34 short-chain volatile fatty acids. The microbial pathways present have a large impact on feed
35 efficiency (1-3), alongside other important production traits such as milk and fat yield (4, 5).
36 Understanding the processes by which food is digested in the rumen may allow us to improve feed
37 efficiency in ruminants (1), either by the production of enzymes isolated from microbes or by
38 manipulating the microbiota through the use of pre- or probiotics (6). There are also other potential
39 industrial uses for the proteins produced by ruminal microbes, for example in processing biofuels,
40 bioremediation, processing pulp/paper and textile manufacturing (7-10). Ruminants are also the
41 largest source of anthropogenic methane emissions and gaining a greater understanding of which
42 microbes are important in methane production could lead to improved methane mitigation
43 strategies (6, 11-15).

44 While inroads have been made towards culturing members of the ruminal microbiota (16, 17) there
45 are still many members which have not been characterised. Metagenomics is a powerful tool which
46 allows us to examine the entire genetic repertoire of the rumen microbiota without the need for
47 culturing. Our group has previously published thousands of metagenome assembled genomes
48 (MAGs) from cattle rumen samples (18, 19) and hundreds of genomes from chicken caecal samples
49 (20), many of which were identified as novel species.

50 Several studies have examined the rumen microbiota using metagenomic techniques in cattle and
51 sheep; however, less effort has been made to characterise the microbiota of other ruminant species
52 which may be less commercially-important but which could harbour microbes which could be
53 industrially useful. For example, wild ruminants are likely to consume a far more diverse diet than
54 farm-raised individuals, and are therefore likely to contain microbes which are able to digest
55 different substrates. In this paper we analyse rumen metagenomic data from four ruminant species:

56 cattle (*Bos Taurus*), sheep (*Ovis aries*), red deer (*Cervus elaphus*) and reindeer (*Rangifer tarandus*).

57 We compare the microbiota of these species taxonomically and functionally and construct 391

58 named rumen-uncultured genomes (RUGs), representing 372 putative novel strains and 279 putative

59 novel species.

60 **Materials and methods:**

61 **Ethics approval**

62 Cattle projects were carried out under Home Office PPL 30/2579. Sheep experimentation was
63 carried out under the conditions set out by UK Home Office licence no. 604028, procedure reference
64 number 8.

65

66 **Experimental design**

67 Reindeer (*Rangifer tarandus*: Grazing mixed vegetation) and red deer (*Cervus elaphus*: Grazing mixed
68 vegetation) were shot in the wild, and ruminal digesta samples were collected immediately. Samples
69 were taken from Holstein cattle (*Bos Taurus*: Fed total mixed ration (once a day)) and Finn-Dorset
70 cross sheep (*Ovis aries*: Grazing mixed pasture) via a rumen cannula. Sheep sampling was performed
71 as described in McKain et al. (21). Samples were collected from the bovine rumen in the following
72 locations: top near cannula, middle at the front of the rumen, middle towards the back of the
73 rumen and bottom (approximately 45cm down from the entrance to the rumen). Digesta samples
74 were mixed with buffer containing glycerol as a cryoprotectant (21). The mixtures were kept on ice
75 for 1-2 hours then frozen at -20°C. DNA extraction was performed using repeated bead beating plus
76 column filtration, as described in (22). Shotgun sequencing was performed on an Ilumina Hiseq
77 2000, producing an average of 1626 million paired reads per sample, of 100bp or 151bp in length.

78

79 **Bioinformatics**

80 Illumina adaptors were removed using trimmomatic (23). IDBA-UD (24) with the options --
81 num_threads 16 --pre_correction --min_contig 300 was used to perform single sample assemblies.
82 After indexing using BWA index, BWA-MEM was used to map reads to assemblies (25). BAM files
83 were created by SAM tools (26) and coverage was calculated using the command

84 jgi_summarize_bam_contig_depths from the MetaBAT software package (27). A coassembly was
85 carried out on all samples using MEGAHIT (28) with the options --continue --kmin-1pass -m 100e+10
86 --k-list 27,37,47,57,67,77,87 --min-contig-len 1000 -t 16. After filtering out reads which were <2kb,
87 indexing and mapping were performed as for single assemblies.

88 Metagenomic binning was carried out using MetaBAT2 with the options --minContigLength 2000, --
89 minContigDepth 2. From the single-assemblies, 1691 bins were created and from the co-assembly
90 2508 bins were created. Completeness and contamination of bins were calculated using CheckM
91 (options: lineage_wf, -t 16, -x fa), and the bins were dereplicated using dRep (29) (options:
92 dereplicate_wf -p 16 -comp 80 -con 10 -str 100 –strW 0). Thus, bins were discarded if their
93 completeness was <80% or if they had contamination >10%. The dereplicated 'winning' bins are
94 referred to below as RUGs. MAGpy was used to compare the RUGs to public datasets (30). Trees
95 produced by MAGpy were rerooted at the branch between archaea and bacteria using Figtree (31)
96 and visualised using GraPhlAn (32). Taxonomies were assigned using GTDB-Tk (33). For submission to
97 public repositories, our RUGs were named as the lowest taxonomic level at which NCBI and GTDB-Tk
98 matched. The taxonomies assigned to RUGs were manually checked against the taxonomic tree and
99 improved accordingly.

100 Carbohydrate-active enzymes (CAZymes) were identified using dbCAN2 (version 7, 24th August 2018)
101 by comparing RUG proteins to the CAZy database (34). RUG proteins were compared to the KEGG
102 database (downloaded on Sept 15th 2018) (35) using DIAMOND (36). KEGG hits for which the
103 alignment length was ≥ 90% of the query length were retained. The likely KEGG ortholog (KO) group
104 for each RUG protein was inferred from the DIAMOND search results and the KEGG database.
105 CAZyme and KEGG ortholog abundances were calculated as the sum of the reads mapping to RUG
106 proteins within each group after using DIAMOND to align reads to the RUG proteins. PULpy was used
107 to identify polysaccharide utilisation loci (37).

108 Statistical analyses were carried out within R (version 3.5.1) (38). The gplots package was used to
109 construct heatmaps (39). The complete linkage method for hierarchical clustering (R hclust function)
110 was used to cluster variables within heatmaps. The ggplot2 (40) package was used to construct
111 scatter plots and NMDS graphs. The vegan package (41) was used to create NMDS axes using the
112 Bray–Curtis dissimilarity. The Adonis function from the vegan package was used to perform
113 PERMANOVA analyses and DeSeq2 (42) was used to calculate differences in coverage for individual
114 CAZymes, KEGG orthologs and RUGs. UpSet graphs were constructed using the UpSetR package (43).
115 Taxonomies were assigned to paired sequence reads with Kraken (44) using a custom kraken
116 database consisting of RefSeq complete genomes with our RUGs and the rumen superset (19)
117 added. Prior to statistical analyses (excluding DeSeq2) and graph construction, data was subsampled.
118 For RUGs, subsampling to the lowest sample coverage was performed. CAZymes and KEGG orthologs
119 were subsampled to the lowest sample abundance.

120

121 **Data availability**

122 The paired-read fastq files generated for this study are available in the European Nucleotide Archive
123 under project PRJEB34458. RUG fasta files are available in Edinburgh DataShare
124 (https://doi.org/10.7488/ds/2640).

125 **Results:**

126 **Construction of RUGs from rumen sequencing data**

127 We produced 979G of Illumina sequencing data from 12 samples then performed a metagenomic
128 assembly of single samples and a co-assembly of all samples. This created a set of 391 dereplicated
129 genomes (99% ANI) with estimated completeness ≥80% and estimated contamination ≤10%
130 (**Supplementary Figure 1**). 284 of these genomes were produced from the single-sample assemblies
131 and 107 were produced from the co-assemblies. 172 genomes were >90% complete with
132 contamination <5%, and would therefore be defined as high-quality draft genomes by Bower et al.
133 (45). The distribution of these RUGs between our samples can be found in **Dataset 1** (based on
134 coverage). **Dataset 2** contains the predicted taxonomic assignment for each RUG while **Figure 1**
135 shows a phylogenetic tree of the genomes. The tree is dominated by the *Bacteroidota* (136 RUGs: All
136 order *Bacteroidales*) and the *Firmicutes_A* (121 RUGs), followed by lesser numbers of the
137 *Firmicutes_C* (40 RUGs), *Synergistota* (20 RUGs: All family *Aminobacteriaceae*), *Firmicutes* (19 RUGs),
138 *Proteobacteria* (15 RUGs), *Cyanobacteriota* (9 RUGs: All family *Gastranaerophilaceae*),
139 *Actinobacteriota* (7 RUGs), *Euryarchaeota* (7 RUGs: All family *Methanobacteriaceae*), *Spirochaetota*
140 (5 RUGs), *Elusimicrobiota* (3 RUGs: All family *Endomicrobiaceae*), *UBP6* (3 RUGs: All genus *UBA1177*),
141 *Fibrobacterota* (2 RUGs: All genus *Fibrobacter*), *Riflebacteria* (2 RUGs: All family *UBA8953*),
142 *Chloroflexota* (1 RUGs: family *Anaerolineaceae*) and *Desulfobacterota* (1 RUGs: genus *Desulfovibrio*).
143 All members of the phylum *Firmicutes_A* belonged to the *Clostridia* class: orders *4C28d-15* (n=9),
144 *CAG-41* (n=3), *Christensenellales* (n=4), *Lachnospirales* (n=56), *Oscillospirales* (n=45),
145 *Peptostreptococcales* (n=2) and *Saccharofermentanales* (n=2). *Firmicutes_C* contains the orders
146 *Acidaminococcales* (n=8) and *Selenomonadales* (n=32). The phylum *Firmicutes* contained the orders
147 *Acholeplasmatales* (n=3), *Erysipelotrichales* (n=1), *Izimaplasmatales* (n=1), *ML615J-28* (n=1),
148 *Mycoplasmatales* (n=1). *RFN20* (n=7) and *RF39* (n=5), The *Actinobacteria* contained the orders
149 *Actinomycetales* (n=1) and *Coriobacteriales* (n=6). The *Proteobacteria* phylum contains the orders

150 *Enterobacterales* (n=4), *Paracaedibacterales* (n=1), *RF32* (n=8) and *UBA3830* (n=2). The
151 *Spirochaetota* contains the orders *Sphaerochaetales* (n=1) and *Treponematales* (n=4).

152 After sub-sampling, we found that samples from different ruminant species clustered significantly
153 separately by abundance of RUGs (PERMANOVA: P = 3e-05). This may be due to the fact that the
154 vast majority of RUGs were only found in a single host species (**Figure 2**), including 111 RUGs in red
155 deer, 78 RUGs in reindeer, 40 RUGs in cow and 31 RUGs in sheep. Only 3 RUGs were found in ≥1X
156 average coverage in all species: uncultured *Bacteroidaceae sp. RUG30019*, uncultured *Prevotella sp.*
157 *RUG30028* and uncultured *Prevotella sp. RUG30114*, suggesting a lack of a cross-species core gut
158 microbiome at the strain level.

159 We compared our RUGs to microbial genomes which had previously been sequenced from the
160 rumen to determine if we had discovered any novel strains or species. We dereplicated our RUGs at
161 99% and 95% ANI to a "superset" of genomes containing rumen RUGs previously produced by our
162 group (19), Hess et al. (10), Parks et al. (46), Solden et al. (47) and Svartström et al. (48) and the
163 genomes from the Hungate collection (16). After dereplication at 99% and the removal of any RUGS
164 with ≥99% ANI to an existing genome (as assigned by GTDB-Tk) or which clustered with members of
165 the superset, 372 of our RUGs remained, representing putative novel strains. After dereplication at
166 95% and the removal of any RUGS with ≥95% ANI to an existing genome (assigned by GTDB-tk) or
167 which clustered with members of the superset, 279 of our RUGs remained, representing putative
168 novel species. The majority of these species originated from single-sample assemblies: 110 from red
169 deer samples, 68 from reindeer samples, 23 from sheep samples and 1 from cattle samples,
170 suggesting that many novel microbial species remain to be discovered from non-cattle ruminant
171 hosts. These novel species are taxonomically diverse, with members belonging to the phyla
172 *Bacteroidota* (n = 97), *Firmicutes_A* (n = 85), *Firmicutes_C* (n = 27), *Firmicutes* (n = 16), *Synergistota*
173 (n = 14), *Proteobacteria* (n = 11), *Cyanobacteriota* (n = 9), *Actinobacteriota* (n = 5), *Spirochaetota* (n =

174 4), *Euryarchaeota* (n = 3), *Elusimicrobiota* (n = 3), *Riflebacteria* (n = 2), *Chloroflexota* (n = 1),
175 *Desulfobacterota* (n = 1) and *UBP6* (n = 1).

176 31 of our total RUGs were able to be taxonomically identified to species level and these contain
177 bacteria which are commonly isolated from the rumen including novel strains of *Bacteroidales*
178 *bacterium UBA1184* (46), *Bacteroidales bacterium UBA3292* (46), *Butyrivibrio fibrisolvens*,
179 *Escherichia coli*, *Fibrobacter sp. UWB2* (49), *Lachnospiraceae bacterium AC3007* (16),
180 *Lachnospiraceae bacterium UBA2932* (46), *Methanobrevibacter sp. UBA188* (46),
181 *Methanobrevibacter sp. UBA212* (46), *Prevotella sp. UBA2859* (46), *Ruminococcaceae bacterium*
182 *UBA3812* (46), *Ruminococcus sp. UBA2836* (46), *Sarcina sp. DSM 11001* (16), *Selenomonas sp.*
183 *AE3005* (16), *Succiniclasticum ruminis* and *Succinivibrio dextrinosolvens*.

184

185 **Comparing microbial taxonomies, CAZymes and KEGG orthologs between ruminant species**

186 We assigned taxonomies to paired sequence reads using our custom kraken database containing
187 RefSeq complete genomes, our RUGs, and a superset of rumen isolated microbial genomes. After
188 subsampling we compared the abundance of members of the microbiota in different ruminant
189 species at multiple taxonomic levels. Averaging reads across rumens species, the vast majority of
190 reads mapped to bacteria (Sheep: 97%, Cow: 97%, Reindeer: 92%, Red deer: 98%) with smaller
191 amounts of archaea (Sheep: 2.3%, Cow: 2.1%, Reindeer: 6.3%, Red deer: 1.9%) and Eukaryota
192 (Sheep: 0.22%, Cow: 1.3%, Reindeer: 1.8%, Red deer: 0.56%). In all ruminants *Bacteroidota* was the
193 most abundant phylum (Sheep: 64%, Cow: 65% Reindeer: 54% Red deer: 52%), with *Firmicutes* being
194 the second most abundant (Sheep: 29%, Cow: 26% Reindeer: 26% Red deer: 38%). Using
195 PERMANOVA, significant differences in the abundance of taxonomies between ruminant species
196 were found at both high (Kingdom: P = 0.01058, Phylum: P = 0.00017) and low (Family: P = 1e-05,
197 Genus: P = 3e-05) taxonomic levels (**Supplementary figure 2**).

198 We also compared the abundance of genes encoding for specific carbohydrate-active enzymes
199 (CAZymes) between species. These enzymes are responsible for the synthesis, binding and
200 metabolism of carbohydrates. The carbohydrate esterases (CEs), glycoside hydrolases (GHs),
201 glycosyltransferases (GTs) and polysaccharide lyases (PLs) act to degrade cellulose, hemicellulose
202 and other carbohydrates which could otherwise not be digested by the host. Non-catalytic
203 carbohydrate-binding modules (CBMs) bind to specific carbohydrates, increasing the efficiency of
204 enzymatic degradation (50). The auxiliary activities (AAs) redox enzymes are reclassified CBMs which
205 are lytic polysaccharide monooxygenases (51). In our samples we found the following numbers of
206 these CAZyme families: 6 AAs redox enzymes, 39 CBMs, 14 CEs, 191 GHs, 61 GTs and 27 PLs. The ten
207 most abundant GHs in the different ruminant species were: for cows GH2, GH3, GH31, GH97, GH28,
208 GH51, GH43_10, GH105, GH10 and GH95; for sheep GH2, GH3, GH28, GH31, GH97, GH32, GH51,
209 GH77, GH78 and GH95; for red deer GH2, GH3, GH31, GH97, GH77, GH32, GH51, GH109, GH28 and
210 GH78; and for reindeer GH2, GH3, GH92, GH109, GH97, GH13, GH31, GH78, GH28 and GH77. The
211 abundance of CAZymes in all samples can be seen in **Figure 3**. Different ruminant species were found
212 to have significantly differently abundant CAZyme genes (PERMANOVA: P = 1e-05, **Figure 4**).
213 However, it should be noted that the vast majority of CAZyme families were found in all sample
214 types (**Figure 5**), indicating that there exists a set of CAZymes which are present across ruminant
215 species consuming different diets and living in vastly different conditions.

216 DeSeq2 was used to identify specific CAZymes which were significantly more abundant in one
217 ruminant species vs another (**Dataset 3**). Those CAZymes which were consistently more abundant in
218 specific species when compared to other species are listed in **Supplementary tables 1-4**.

219 CAZymes are often found organised into Polysaccharide Utilization Loci (PUL) which comprise a set
220 of genes that enable the binding and degradation of specific carbohydrates or multiple
221 carbohydrates. We used the software PULpy to predict PULs which were present in our
222 *Bacteroidales* RUGs. Of the 136 RUGs which belong to the taxonomy *Bacteroidales*, 112 contain

223 putative PULs. Within these RUGs we identified 970 PULs, with numbers of PULs per RUG ranging
224 from 1 to 35. The largest quantity of PULs originating from one RUG was 35 from uncultured
225 *Bacteroidales sp. RUG30227*; these encoded a wide range of CAZymes. This RUG was more abundant
226 in reindeer samples than samples from other ruminants. Of the 970 PULs, 332 of these were a single
227 susC/D pair. A summary of identified PULs can be found in **Dataset 4** (**Figure 6**).

228 We also examined the abundance of genes which belonged to specific KEGG orthologs. KEGG
229 orthologs represent a wide range of molecular functions and are defined by a network-based
230 classification. We found that, as for CAZymes, ruminant species clustered significantly by the
231 abundance of genes with specific KEGG orthologs (PERMANOVA: P = 1e-05, **Figure 7**) and that the
232 vast majority of orthologs were found in all ruminant species (**Figure 8**). However, the large amount
233 of orthologs (n=729) which were only found in the two domesticated species (cattle and sheep) is
234 also worthy of note. DeSeq2 was used to identify many KEGG orthologs which were significantly
235 more abundant in one ruminant species vs another (**Dataset 5**). Those orthologs which were
236 consistently more abundant in specific ruminant species (Adjusted p-value <0.05) are listed in
237 **Dataset 6**.

238 **Discussion:**

239 The rumen microbiota plays a crucial role in the ability of ruminants to efficiently digest feed while
240 the rumen microbiota and their products also have a potential use in diverse industrial applications.
241 The ruminal microbiota of red deer and reindeer have previously been studied using 16S rRNA gene
242 sequencing (52-55). However, metagenomic studies of these species are limited, with only one study
243 in reindeer (56) and no studies in red deer.

244 In this study we constructed 391 rumen microbial genomes from metagenomic data from cattle,
245 sheep, red deer and reindeer. We assigned taxonomies to our RUGs using GTDB-Tk rather than NCBI
246 based taxonomies as this improves the classification of uncultured bacteria due to the use of a
247 genome-based taxonomy (33). We have also previously found less need to manually correct
248 taxonomic assignments when using GTDB-Tk (20). Our microbes predominantly belonged to the
249 *Bacteroidota* and *Firmicutes_A*, with lesser numbers of 14 other phyla. We dereplicated our
250 genomes alongside a superset of rumen bacterial genomes (19) and used the results output by
251 GTDB-Tk to identify RUGs which represent novel microbial strains and species. Amongst our
252 genomes we identified 372 novel strains and 279 novel species. These microbes were taxonomically
253 diverse, belonging to 15 phyla. Only 31 RUGs were assigned an identity at species level.

254 The vast majority of our total RUGs were only present on average at ≥1 coverage in one ruminant
255 species. However, when assigning taxonomies to reads using kraken we found that many
256 taxonomies were shared between sample types. When comparing the abundance of sequences
257 which belong to different taxonomies within our data we found that species separated by the
258 microbial taxa present at high and low taxonomic levels. We are aware that the sample sizes for our
259 study are small and therefore any conclusions about differences between the microbiota of
260 ruminant species should be drawn cautiously. However, our data are supportive of the hypothesis
261 that there are host species-specific rumen microbiota at the strain and species level but that these
262 differences do not necessarily translate into large differences in the types of CAZymes expressed.

263 While we found that there were significant differences between the abundances of CAZymes
264 between ruminant species, most CAZymes were present in all ruminant species. These results also
265 reflected those which we found when analysing the abundance of KEGG orthologs. We also
266 identified 970 PULs in our *Bacteriodales* RUGs, with numbers of PULs per genome ranging from 1 to
267 35. The RUG containing 35 PULs was found most abundantly in reindeer samples, emphasising the
268 potential for the discovery of novel carbohydrate-active enzymes in lesser studied ruminant species,
269 as also highlighted by a previous study which identified multiple PULs in metagenomic samples from
270 reindeer (56). Unfortunately due to the nature of our samples, with red deer and reindeer samples
271 originating from animals eating a non-regimented diet, we are not able to provide metadata as to
272 the exact nutritional composition of our animals' diets, therefore a more in depth analysis of dietary
273 carbohydrates vs CAZyme/PUL abundance is not possible.

274 In conclusion, metagenomic binning is a highly powerful technique; however, while we believe that
275 sequencing the microbiota and the construction of metagenome-assembled genomes is an
276 important first step in characterising the rumen microbiota we feel that in the future it highly
277 important that culture based studies are performed in order for us to gain a functional
278 understanding of the microbiota and to design interventions to improve livestock productivity or to
279 produce industrially useful microbial products. In conclusion, we present a dataset of RUGs from
280 four ruminant species which can be used as a reference dataset in future metagenomic studies and
281 to aid in the design of culture based studies.

Acknowledgements:

The Rowett Institute is funded by the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish Government. BG thanks the Higher Education Council of the Turkish Republic for the award of a travelling fellowship. We thank Bob Mayes and Dave Hamilton of the James Hutton Institute for their permission and help in sampling the sheep digesta, Kevin Shingfield for the provision of reindeer digesta samples, and Euan Munro for red deer digesta samples. We also thank Nest McKain of RI for technical assistance. The Roslin Institute forms part of the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council, including institute strategic programme and national capability awards to The Roslin Institute (BBSRC: BB/P013759/1, BB/P013732/1, BB/J004235/1, BB/J004243/1), and the Technology Strategy Board (TS/J000108/1, TS/J000116/1).

Conflict of Interest:

The authors declare no competing financial interests.

296 **References**:

297 1. Lima J, Auffret MD, Stewart RD, Dewhurst RJ, Duthie CA, Snelling TJ, et al. Identification of
298 rumen microbial genes involved in pathways linked to appetite, growth, and feed conversion
299 efficiency in cattle. Front Genet. 2019;**10**: 701.

300 2. Ben Shabat SK, Sasson G, Doron-Faigenboim A, Durman T, Yaacoby S, Miller MEB, et al.
301 Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants.
302 Isme J. 2016;**10**: 2958-72.

303 3. Patil RD, Ellison MJ, Wolff SM, Shearer C, Wright AM, Cockrum RR, et al. Poor feed efficiency
304 in sheep is associated with several structural abnormalities in the community metabolic network of
305 their ruminal microbes. J Anim Sci. 2018;**96**: 2113-24.

306 4. Jami E, White BA, Mizrahi I. Potential role of the bovine rumen microbiome in modulating
307 milk composition and feed efficiency. PLoS One. 2014;**9**: e85423.

308 5. Scharen M, Frahm J, Kersten S, Meyer U, Hummel J, Breves G, et al. Interrelations between
309 the rumen microbiota and production, behavioral, rumen fermentation, metabolic, and
310 immunological attributes of dairy cows. J Dairy Sci. 2018;**101**: 4615-37.

311 6. Huws SA, Creevey CJ, Oyama LB, Mizrahi I, Denman SE, Popova M, et al. Addressing global
312 ruminant agricultural challenges through understanding the rumen microbiome: Past, present, and
313 future. Front Microbiol. 2018;**9**: 33.

314 7. Solbak AI, Richardson TH, McCann RT, Kline KA, Bartnek F, Tomlinson G, et al. Discovery of
315 pectin-degrading enzymes and directed evolution of a novel pectate lyase for processing cotton
316 fabric. J Biol Chem. 2005;**280**: 9431-8.

317 8. Singh B, Gautam SK, Verma V, Kumar M, Singh B. Metagenomics in animal gastrointestinal
318 ecosystem: Potential biotechnological prospects. Anaerobe. 2008;**14**: 138-44.

319 9. Ufarte L, Laville E, Duquesne S, Morgavi D, Robe P, Klopp C, et al. Discovery of carbamate
320 degrading enzymes by functional metagenomics. PLoS One. 2017;**12**: e0189201.

321 10. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, et al. Metagenomic discovery
322 of biomass-degrading genes and genomes from cow rumen. Science. 2011;**331**: 463-7.
323 11. Wallace RJ, Rooke JA, McKain N, Duthie CA, Hyslop JJ, Ross DW, et al. The rumen microbial
324 metagenome associated with high methane production in cattle. BMC Genomics. 2015;**16**: 839.
325 12. Difford GF, Plichta DR, Lovendahl P, Lassen J, Noel SJ, Hojberg O, et al. Host genetics and the
326 rumen microbiome jointly associate with methane emissions in dairy cows. PLoS Genet. 2018;**14**:
327 e1007580.
328 13. Shi WB, Moon CD, Leahy SC, Kang DW, Froula J, Kittelmann S, et al. Methane yield
329 phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res.
330 2014;**24**: 1517-25.
331 14. Auffret MD, Stewart R, Dewhurst RJ, Duthie CA, Rooke JA, Wallace RJ, et al. Identification,
332 comparison, and validation of robust rumen microbial biomarkers for methane emissions using
333 diverse *Bos taurus* breeds and basal diets. Front Microbiol. 2018;**8**: 2642.
334 15. Roehe R, Dewhurst RJ, Duthie CA, Rooke JA, McKain N, Ross DW, et al. Bovine host genetic
335 variation influences rumen microbial methane production with best selection criterion for low
336 methane emitting and efficiently feed converting hosts based on metagenomic gene abundance.
337 PLoS Genet. 2016;**12**: 20.
338 16. Seshadri R, Leahy SC, Attwood GT, Teh KH, Lambie SC, Cookson AL, et al. Cultivation and
339 sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol.
340 2018;**36**: 359-67.
341 17. Creevey CJ, Kelly WJ, Henderson G, Leahy SC. Determining the culturability of the rumen
342 bacterial microbiome. Microb Biotechnol. 2014;**7**: 467-79.
343 18. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913
344 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun. 2018;**9**: 870.

345 19. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941
346 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat
347 Biotechnol. 2019;**37**: 953-61.
348 20. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel
349 bacterial genomes from the chicken caecum. preprint at bioRxiv. 2019: 699843.
350 21. McKain N, Genc B, Snelling TJ, Wallace RJ. Differential recovery of bacterial and archaeal 16S
351 rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. J Microbiol Methods.
352 2013;**95**: 381-3.
353 22. Yu ZT, Morrison M. Improved extraction of PCR-quality community DNA from digesta and
354 fecal samples. Biotechniques. 2004;**36**: 808-12.
355 23. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data.
356 Bioinformatics. 2014;**30**: 2114-20.
357 24. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and
358 metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;**28**: 1420-8.
359 25. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv.
360 2013: preprint arXiv:1303.3997.
361 26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
362 Alignment/Map format and SAMtools. Bioinformatics. 2009;**25**: 2078-9.
363 27. Kang DWD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
364 reconstructing single genomes from complex microbial communities. PeerJ. 2015;**3**: 15.
365 28. Li DH, Liu CM, Luo RB, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for
366 large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;**31**:
367 1674-6.
368 29. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic
369 comparisons that enables improved genome recovery from metagenomes through de-replication.
370 Isme J. 2017;**11**: 2864-8.

371 30. Stewart RD, Watson M, Auffret MD, Roehe R, Snelling TJ. MAGpy: A reproducible pipeline for
372 the downstream analysis of metagenome-assembled genomes (MAGs). Bioinformatics. 2018;**35**:
373 2150–2.
374 31. Rambaut A. FigTree v1. 4. http://treebioedacuk/software/figtree20122012.
375 32. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical
376 representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015;**3**: e1029.
377 33. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A
378 standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.
379 Nat Biotechnol. 2018;**36**: 996-1004.
380 34. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-
381 Active EnZymes database (CAZy): An expert resource for glycogenomics. Nucleic Acids Res. 2009;**37**:
382 D233-D8.
383 35. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res.
384 2000;**28**: 27-30.
385 36. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat
386 Methods. 2015;**12**: 59-60.
387 37. Stewart RD, Auffret MD, Roehe R, Watson M. Open prediction of polysaccharide utilisation
388 loci (PUL) in 5414 public *Bacteroidetes* genomes using PULpy. preprint in bioRxiv. 2018: 421024.
389 38. Team RC. R: A language and environment for statistical computing. 2013.
390 39. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R
391 programming tools for plotting data. https://rdrrio/cran/gplots/2019.
392 40. Wickham H. ggplot2: elegant graphics for data analysis. http://ggplot2org20162016.
393 41. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Minchin DMPR, et al. vegan:
394 Community Ecology Package. https://CRANR-projectorg/package=vegan20132018.
395 42. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-
396 seq data with DESeq2. Genome Biol. 2014;**15**: 550.

397 43. Conway JR, Lex A, Gehlenborg N. UpSetR: An R package for the visualization of intersecting
398 sets and their properties. Bioinformatics. 2017;**33**: 2938-40.
399 44. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact
400 alignments. Genome Biol. 2014;**15**: R46.
401 45. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al.
402 Minimum information about a single amplified genome (MISAG) and a metagenome-assembled
403 genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;**35**: 725-31.
404 46. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of
405 nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol.
406 2017;**2**: 1533-42.
407 47. Solden LM, Naas AE, Roux S, Daly RA, Collins WB, Nicora CD, et al. Interspecies cross-feeding
408 orchestrates carbon degradation in the rumen ecosystem. Nat Microbiol. 2018;**3**: 1274-84.
409 48. Svartstrom O, Alneberg J, Terrapon N, Lombard V, de Bruijn I, Malmsten J, et al. Ninety-nine
410 de novo assembled genomes from the moose (*Alces alces*) rumen microbiome provide new insights
411 into microbial plant biomass degradation. Isme J. 2017;**11**: 2538-51.
412 49. Neumann AP, Suen G. The phylogenomic diversity of herbivore-associated *Fibrobacter* spp.
413 is correlated to lignocellulose-degrading potential. mSphere. 2018;**3**: e00593-18.
414 50. Shoseyov O, Shani Z, Levy I. Carbohydrate binding modules: Biochemical properties and
415 novel applications. 2006;**70**: 283-95.
416 51. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the enzymatic
417 repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol Biofuels. 2013;**6**:
418 41.
419 52. Ostbye K, Wilson R, Rudi K. Rumen microbiota for wild boreal cervids living in the same
420 habitat. FEMS Microbiol Lett. 2016;**363**: fnw233.

421 53. Salgado-Flores A, Hagen LH, Ishaq SL, Zamanzadeh M, Wright ADG, Pope PB, et al. Rumen
422 and cecum microbiomes in reindeer (*Rangifer tarandus tarandus*) Are changed in response to a
423 lichen diet and may affect enteric methane emissions. PLoS One. 2016;**11**: e0155213.
424 54. Sundset M, Edwards J, Cheng Y, Senosiain R, Fraile M, Northwood K, et al. Molecular
425 diversity of the rumen microbiome of Norwegian reindeer on natural summer pasture. Microb Ecol.
426 2009;**57**: 335-48.
427 55. Qian WX, Ao WP, Jia CH, Li ZP. Bacterial colonisation of reeds and cottonseed hulls in the
428 rumen of Tarim red deer (*Cervus elaphus yarkandensis*). Antonie Van Leeuwenhoek. 2019;**112**: 1283-
429 96.
430 56. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, et al. Metagenomics
431 of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci.
432 PLoS One. 2012;**7**: e38571.

433

434 **Figure legends:**

435 **Figure 1:** Phylogenetic tree of the 391 draft microbial genomes from rumen samples, labelled by
436 taxonomic class. Taxonomies were defined by GTDB-Tk.

437 **Figure 2:** UpSetR graph showing the number of shared microbial genomes at average 1X coverage
438 (after sub-sampling to equal depth) within four ruminant species.

439 **Figure 3:** Heatmap showing the abundance of CAZyme families within ruminal contents samples
440 from four different species. Euclidean clustering was used to cluster CAZymes and samples.

441 **Figure 4:** NMDS of ruminal samples clustered by abundance of CAZymes, using Bray–Curtis
442 dissimilarity values (PERMANOVA; P = 1e-05).

443 **Figure 5:** UpSetR graph showing the number of shared CAZyme families at average 1X coverage
444 within four ruminant species.

445 **Figure 6**: Several predicted polysaccharide utilisation loci (PULs). A: The most common PUL
446 consisting of only a susC/D pair. B-E: The 4 largest PULs identified in our dataset.

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

449 **Figure 8**: UpSetR graph showing the number of shared KEGG orthologs families at average 1X
450 coverage within four ruminant species.

451 **Dataset legends:**

452 **Dataset 1**: Average coverage of RUGs in all samples. Coverage was calculated by mapping the RUG
453 scaffolds to adaptor-trimmed Illumina reads.

454 **Dataset 2**: Description of each RUG, including taxonomy, novelty, genome completeness, genome
455 contamination, genome size, number of ambiguous bases, number of scaffolds, number of contigs,
456 N50 (scaffolds), N50 (contigs), mean scaffold length, mean contig length, longest contig and GC
457 content.

458 **Dataset 3**: CAZymes which were more abundant in sheep, cattle, reindeer or red deer (DeSeq2:
459 Adjusted p-value <0.05).

460 **Dataset 4**: Summary of polysaccharide utilisation loci found in *Bacteroidales* RUGs.

461 **Dataset 5**: KEGG orthologs which were more abundant in sheep, cattle, reindeer or red deer
462 (DeSeq2: Adjusted p-value <0.05).

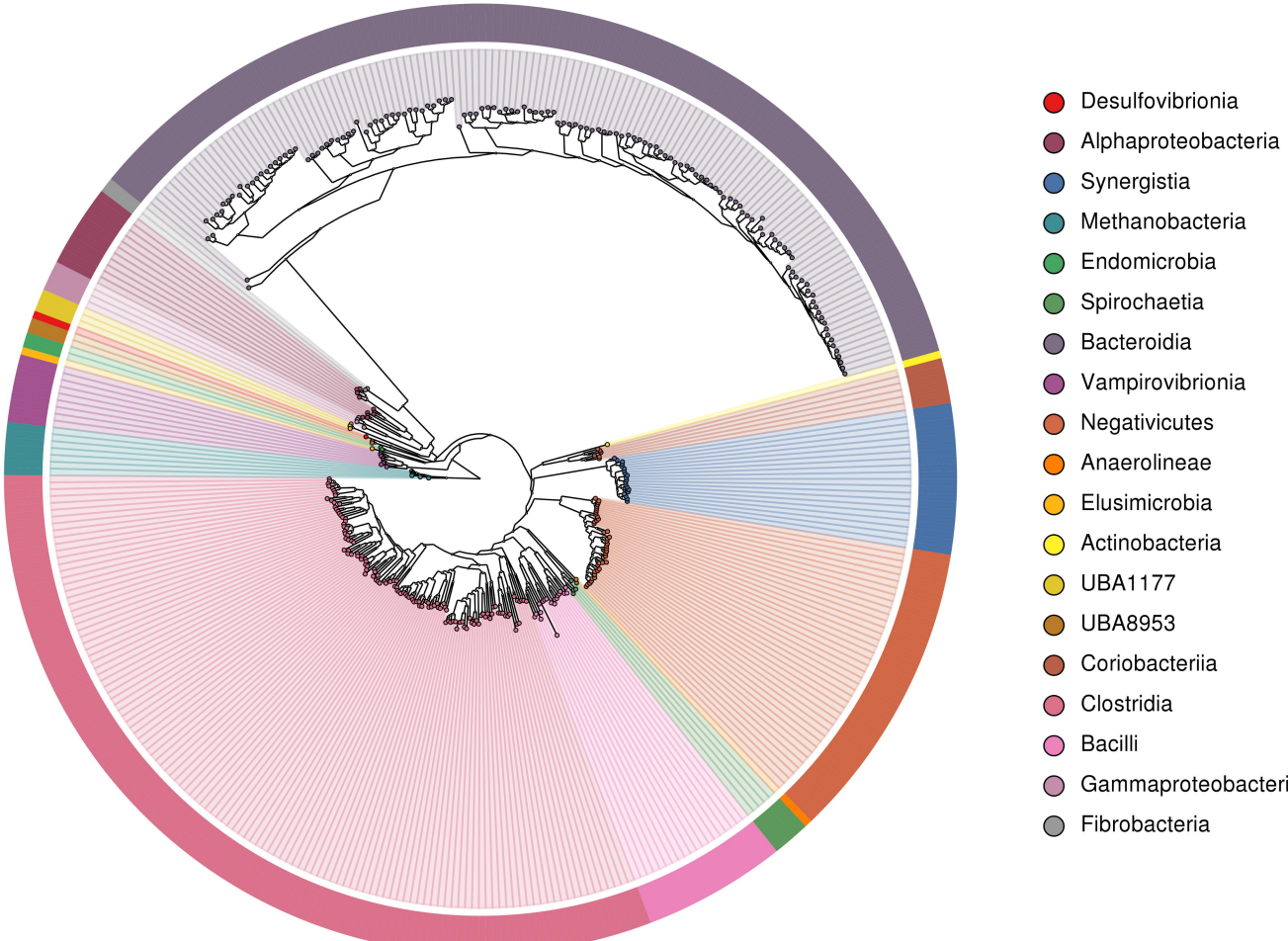
463 **Dataset 6**: KEGG orthologs which were consistently more abundant in sheep, cattle, reindeer or red
464 deer (DeSeq2: Adjusted p-value <0.05).

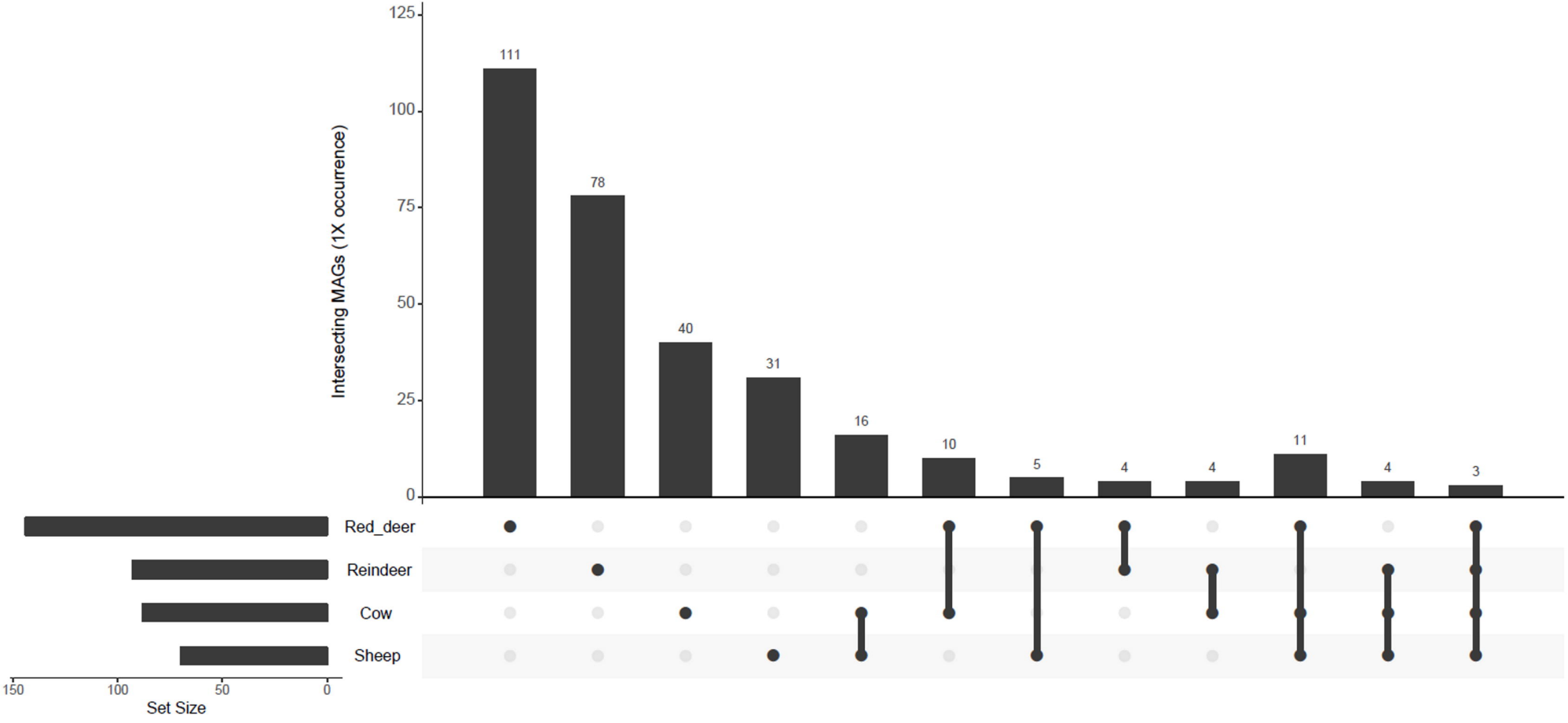
465 **Supplementary figure 1**: Contamination and completeness (defined by CheckM software) of 391
466 dereplicated metagenome-assembled-genomes from rumen samples. Grey: genomes which are 80-
467 90% complete with 5-10% contamination. Red: genomes which are >90% complete with <5%
468 contamination.

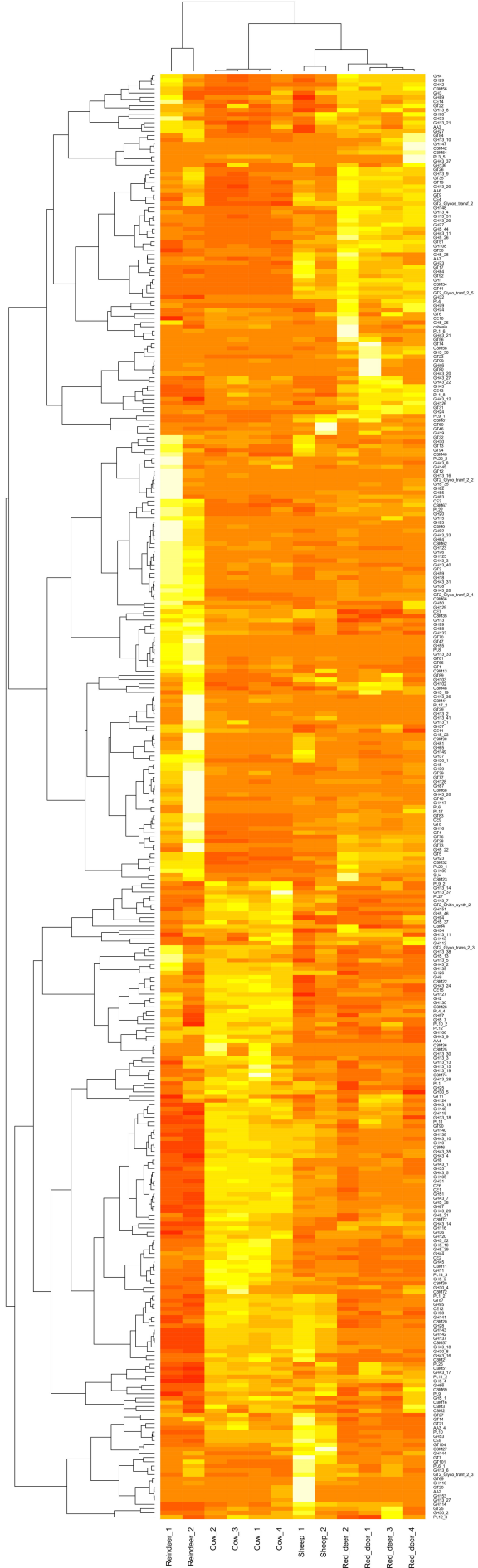
469 **Supplementary figure 2**: NMDS of ruminal samples clustered by abundance of taxonomies, using
470 Bray–Curtis dissimilarity values. A: Kingdom (PERMANOVA; P = 0.01058), B: Phylum (PERMANOVA;
471 P= 0.00017), C: Family (PERMANOVA; P = 1e-05), D: Genus (PERMANOVA; P = 3e-05).

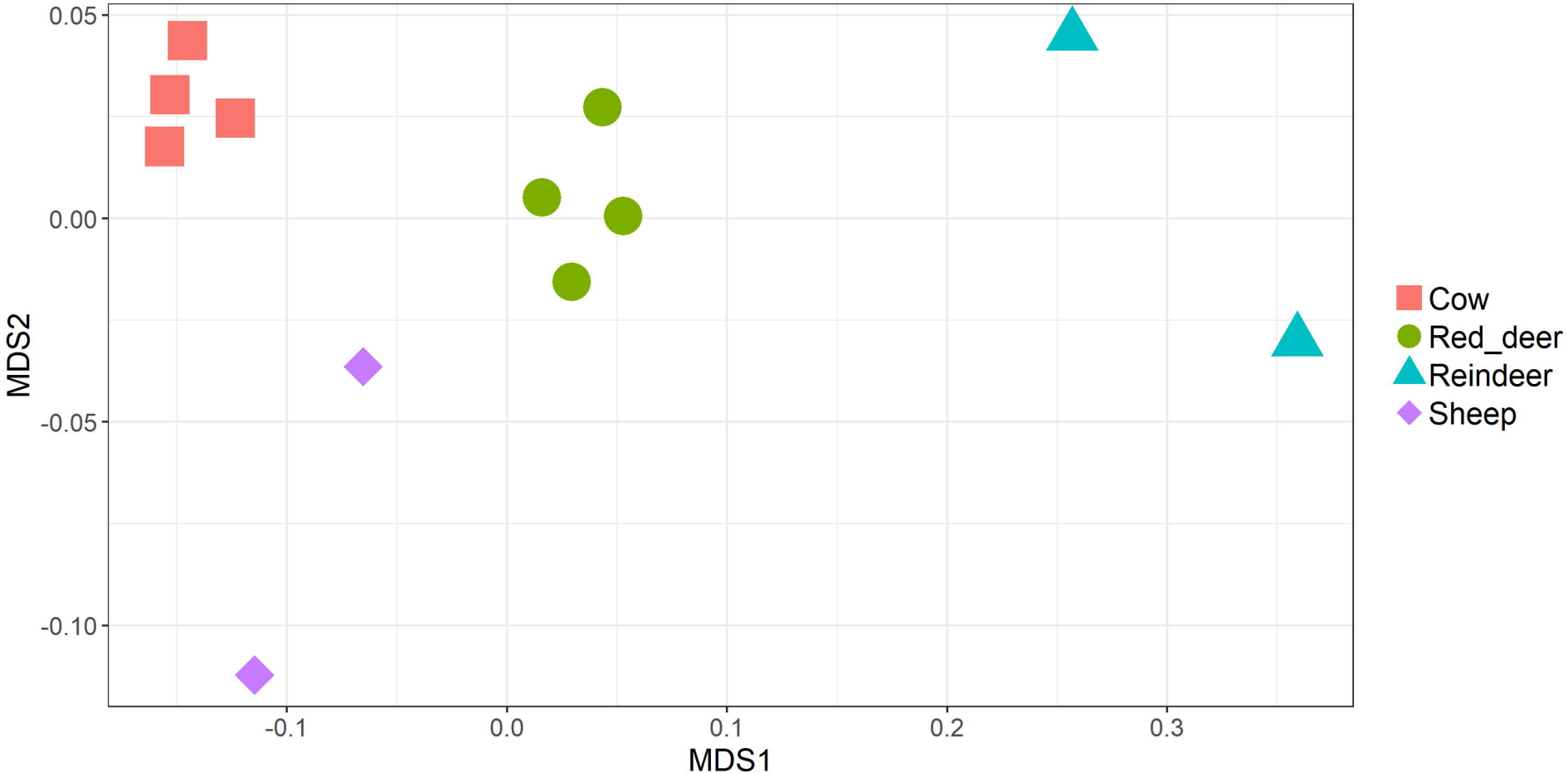
472 **Supplementary table 1**: CAZymes which were consistently more abundant in cattle.
473 **Supplementary table 2**: CAZymes which were consistently more abundant in sheep.
474 **Supplementary table 3**: CAZymes which were consistently more abundant in red deer.

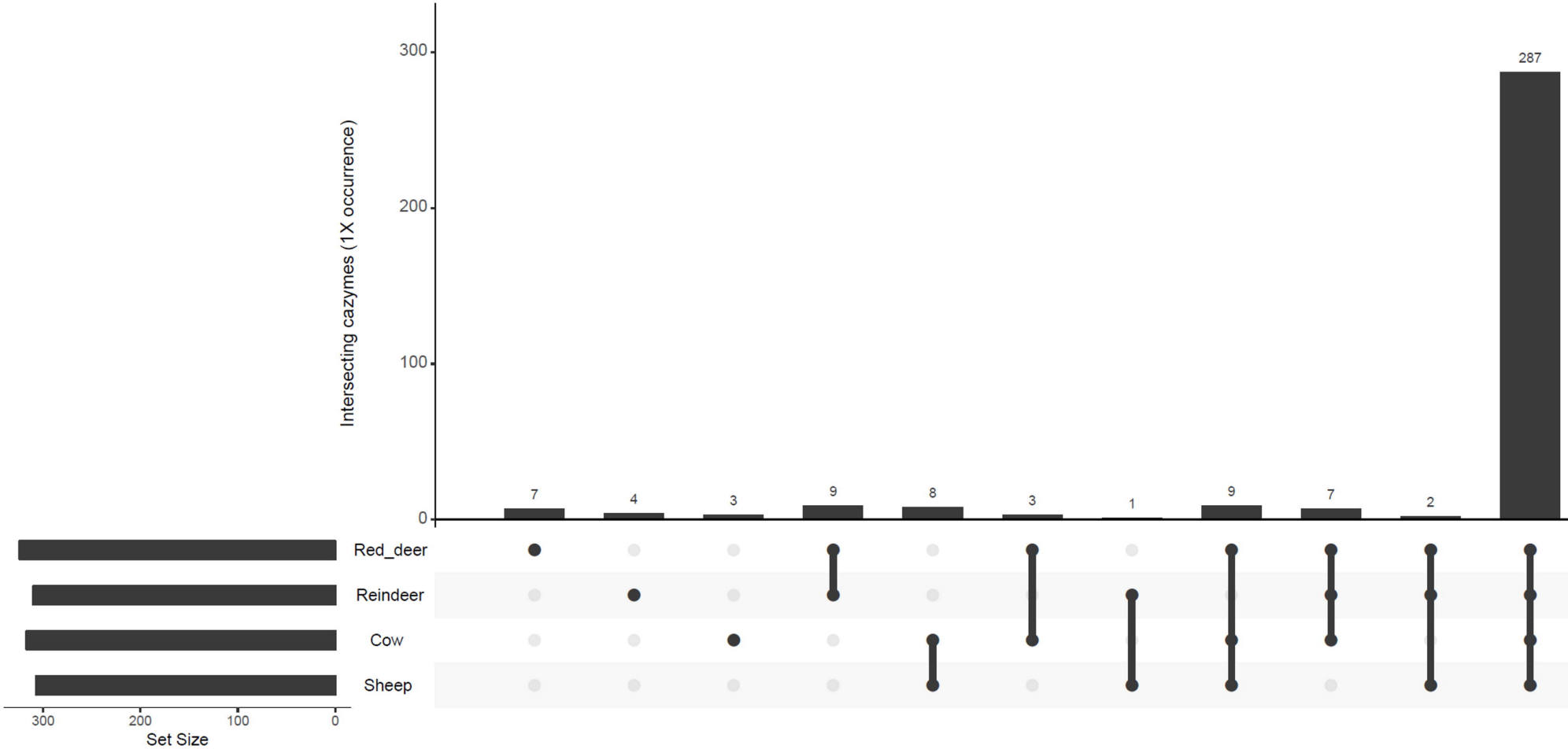
475 **Supplementary table 4**: CAZymes which were consistently more abundant in reindeer.

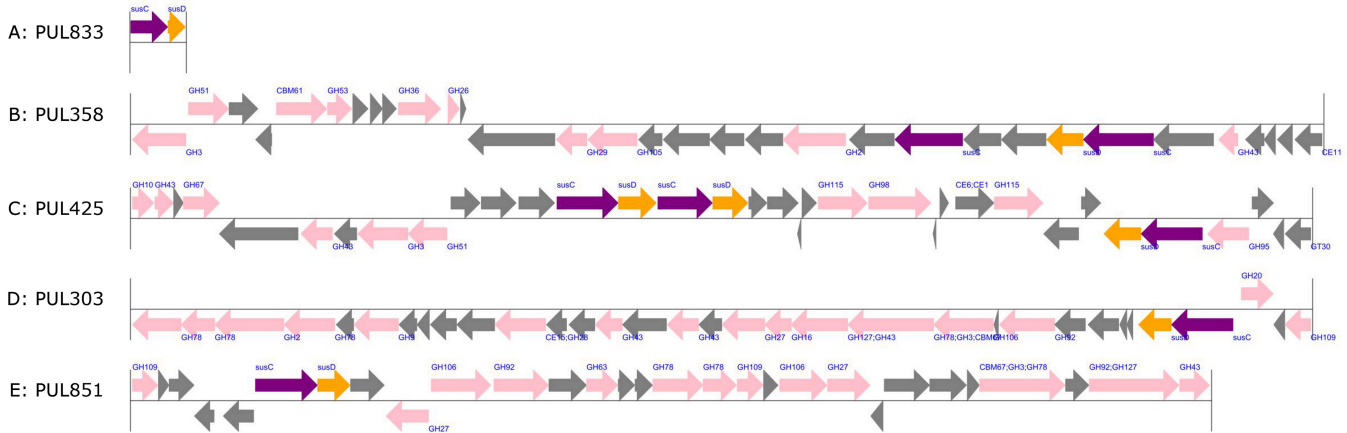

Desulfovibrionia
Alphaproteobacteria
Synergistia
Methanobacteria
Endomicrobia
Spirochaetia
Bacteroidia
Vampirovibrionia
Negativicutes
Anaerolineae
Elusimicrobia
Actinobacteria
UBA1177
UBA8953
Coriobacteriia
Clostridia
Bacilli
Gammaproteobacteria
Fibrobacteria

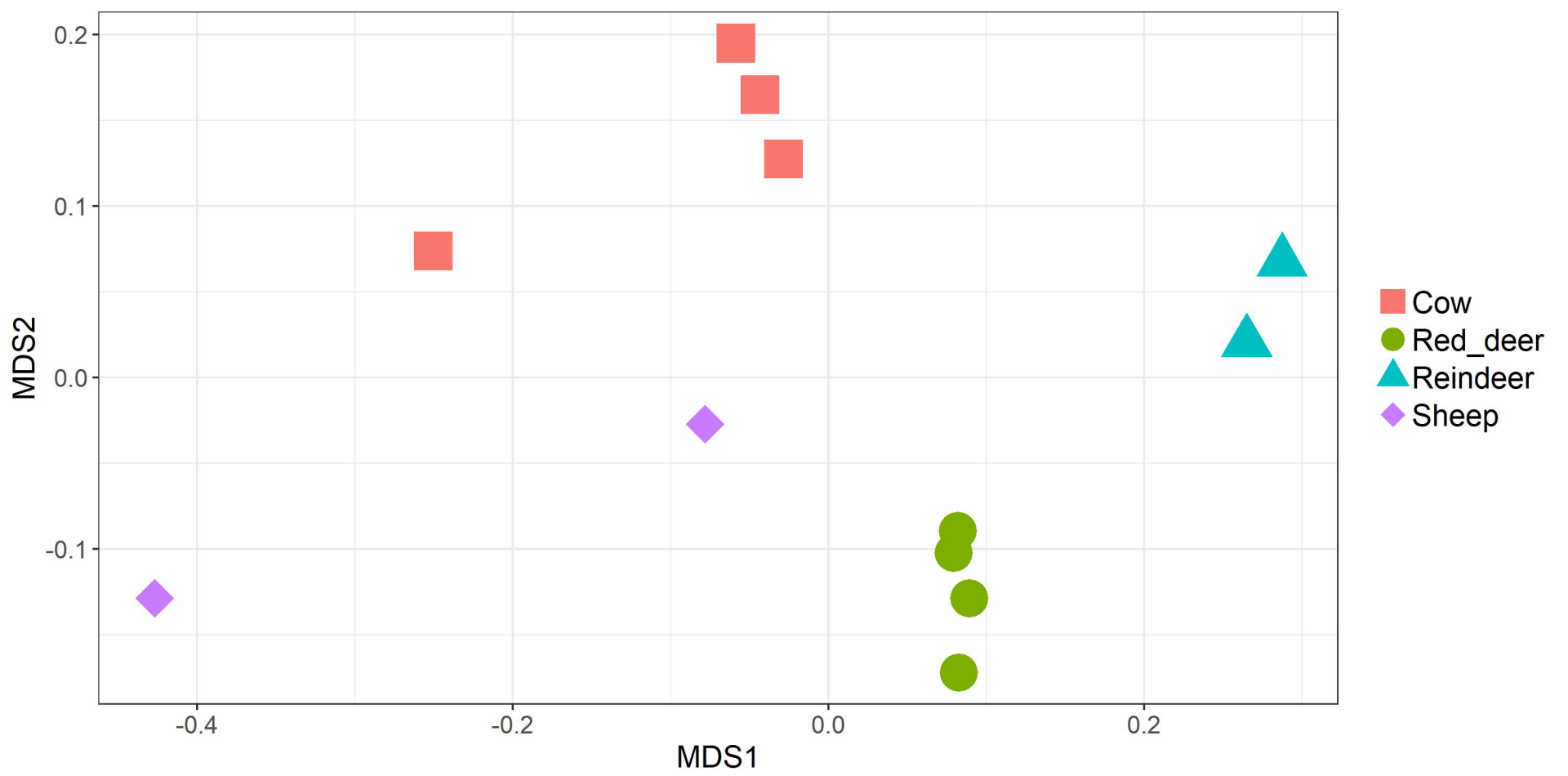

Intersecting MAGs (1X occurrence)
125
100
75
50
25
0
111
78
40
31
16
10
5
4
4
11
4
3
Red_deer
Reindeer
Cow
Sheep
150
100
50
0
Set Size

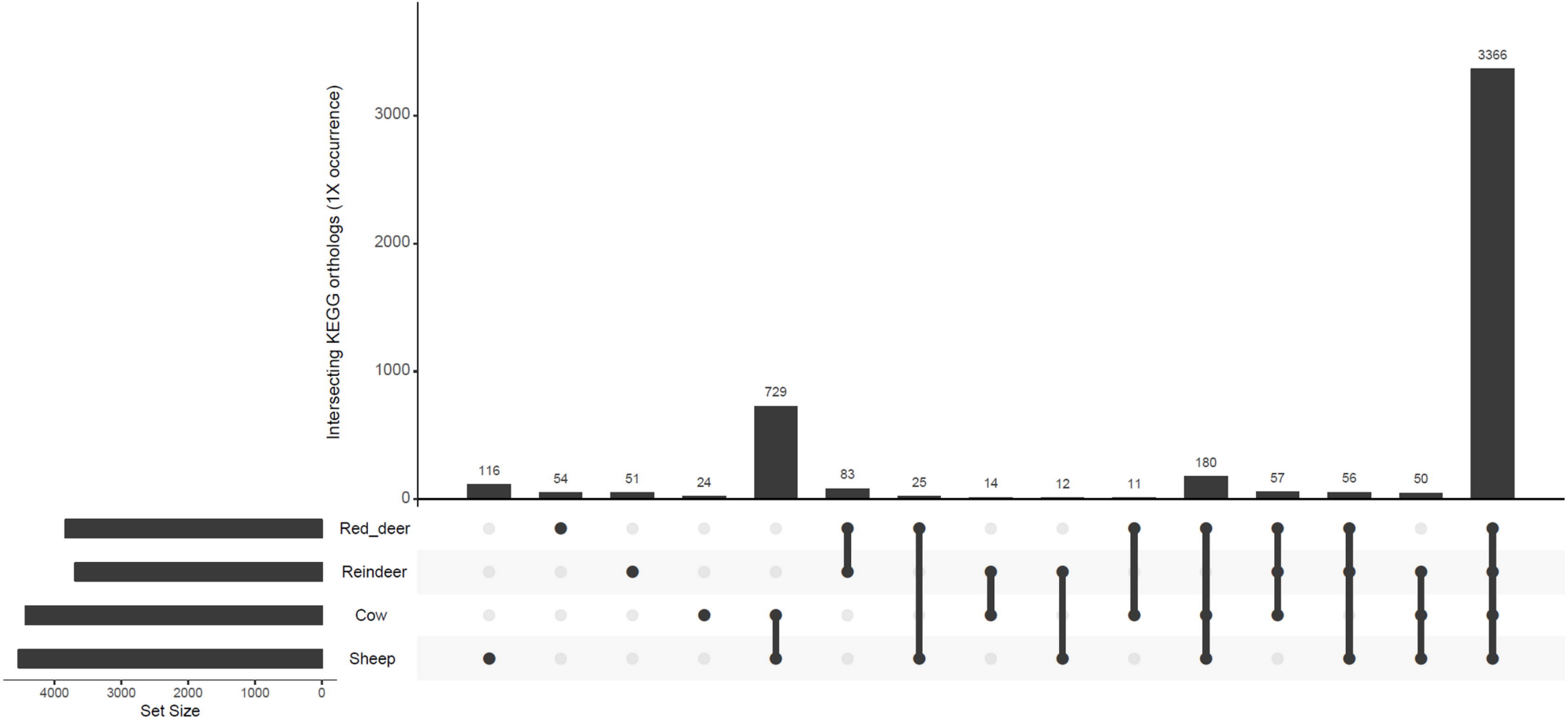

Reindeer_1
Reindeer_2
Cow_3
Cow_3
Cow_1
Cow_4
Sheep_1
Sheep_2
Red_deer_2
Red_deer_1
Red_deer_3
Red_deer_4


0.05
0.00
-0.05
-0.10
MDS2
-0.1
0.0
0.1
0.2
0.3
MDS1
Cow
Red_deer
Reindeer
Sheep


Intersecting cazymes (1X occurrence)
300
200
100
0
287
7
4
3
9
8
3
1
9
7
2
Red_deer
Reindeer
Cow
Sheep
300
200
100
0
Set Size


A: PUL833
susC
susD
B: PUL358
GH51
CBM61
GH53
GH36
GH26
GH3
GH29
GH105
GH2
susE
susD
susC
GH43
CE11
C: PUL425
GH10
GH43
GH67
susC
susD
susC
susD
GH115
GH98
CE6;CE1
GH115
GH43
GH3
GH51
susD
susC
GH95
GT30
D: PUL303
GH20
GH78
GH78
GH2
GH16
GH3
CE16;GH26
GH43
GH43
GH27
GH16
GH127;GH43
GH78;GH3;CBM6
GH106
GH92
susD
susC
GH109
E: PUL851
GH109
susC
susD
GH106
GH92
GH63
GH78
GH78
GH109
GH106
GH27
CBM67;GH3;GH78
GH92;GH127
GH43
GH27


0.2
0.1
0.0
-0.1
MDS2
-0.4
-0.2
0.0
0.2
MDS1
Cow
Red_deer
Reindeer
Sheep


Intersecting KEGG orthologs (1X occurrence)
3000
2000
1000
0
116
54
51
24
729
83
25
14
12
11
180
57
56
50
3366
Red_deer
Reindeer
Cow
Sheep
4000
3000
2000
1000
0
Set Size