A catalog of microbial genes from the bovine rumen reveals the determinants of herbivory

Junhua Li[1, 2, 3], Huanzi Zhong[1, 2], Yuliaxis Ramayo-Caldas[4, 5], Nicolas Terrapon[6, 7], Vincent Lombard[6, 7], Gabrielle Potocki-Veronese[8], Jordi Estellé[4], Milka Popova[9], Ziyi Yang[1, 2], Hui Zhang[1, 2], Fang Li[1, 2], Shanmei Tang[1, 2], Weineng Chen[1], Bing Chen[1, 2], Jiyang Li[1], Jing Guo[1, 2], Cécile Martin[9], Emmanuelle Maguin[10], Xun Xu[1, 2] , Huanming Yang[1, 11], Jian Wang[1, 11], Lise Madsen[1,12, 13], Karsten Kristiansen[1, 13*], Bernard Henrissat[6, 7, 13], Stanislav D. Ehrlich[14, 15*], Diego P. Morgavi[9*]

[1]BGI-Shenzhen, Shenzhen 518083, China. [2]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China. [3]School of Bioscience & Bioengineering, South China University of Technology, Guangzhou, China. [4]INRA, Institut National de la Recherche Agronomique, Génétique Animale et Biologie Intégrative, AgroParisTech, Université Paris-Saclay, 78350, Jouy-en-Josas, France. [5]Animal Breeding and Genetics Program, Institute for Research and Technology in Food and Agriculture (IRTA), Torre Marimon, Caldes de Montbui, 08140, Spain. [6]CNRS UMR 7257, Aix-Marseille University, 13288 Marseille, France. [7]INRA, USC 1408 AFMB, 13288 Marseille, France. [8]LISBP, Universite de Toulouse, CNRS, INRA, INSA, 31077, Toulouse, France. [9]INRA, UMR Herbivores, Université Clermont Auvergne, VetAgro Sup, F-63122 Saint-Genès Champanelle, France. [10]INRA, Micalis Institute, AgroParisTech, Université Paris-Saclay, 78350, Jouy-en-Josas, France. [11]James D. Watson

22 Institute of Genome Sciences, Hangzhou 310058, China. [12]Laboratory of Genomics and Molecular Biomedicine, Institute of Marine Research (IMR), Postboks 1870 Nordnes, 5817
24 Bergen, Norway. [13]Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen Ø, Denmark. [13]Department of
26 Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia. [14]MGP MetaGenoPolis, INRA, Université Paris-Saclay, 78350 Jouy en Josas, France, [15]Centre for
28 Host Microbiome Interactions, Dental Institute, King's College London, UK.

30 *Corresponding authors:
Diego Morgavi (diego.morgavi@inra.fr); Stanislav Ehrlich (stanislav.ehrlich@inra.fr);
32 Karsten Kristiansen (kk@bio.ku.dk)

34 **Running title**: Gene catalog of the rumen microbiome

Keywords: rumen; metagenome; herbivory; carbohydrate-active enzymes

36

ABSTRACT

38 The rumen microbiota is a unique feature of ruminants and thorough knowledge of the genetic potential of rumen symbiotic microbes provides opportunities for improving the sustainability
40 of ruminant production systems. Using deep metagenome sequencing we identified 13,825,880 non-redundant prokaryote genes from the bovine rumen and constructed 324 high
42 quality metagenomic species. These metagenomic species were prevalent in the rumen of 77 cattle fed different feeds whereas previously identified rumen microbial genomes were less
44 abundant. Compared to human, pig and mouse gut metagenome catalogs, the rumen is richer in functions and microbial species associated to the degradation of plant cell wall material and
46 production of methane. Genes coding for enzymes that deconstruct plant polysaccharides showed a particularly high richness that is otherwise impossible to infer from available
48 genomes or shallow metagenomics sequencing. These data bring new insights on functions, enzymes and microbes of the rumen, critical to understand phenotypes and biological
50 processes.

INTRODUCTION

52 Ruminant production contributes to livelihood and to food and nutritional security in many regions of the world. Milk and meat from ruminants are important sources of protein and
54 micronutrients in the human diet but often criticized as unsustainable because of the low conversion efficiency of plant feeds into animal foods (Aiking 2014) and also due to their
56 high environmental footprint. However, when the feed conversion efficiency of protein and energy contained in milk and meat is calculated based on the ingestion of human-inedible
58 protein and energy the output is higher that the input, particularly in forage-based production systems (Gill et al. 2010; Ertl et al. 2016). The transformation of feeds, not suitable for human
60 consumption, into highly nutritious protein and energy products is carried out by gastrointestinal symbiotic microbes, particularly those residing in ruminants' forestomach –

62 the rumen. Rumen microbes are essential for ruminants allowing them to thrive in agricultural land not suitable for crops and to consume agricultural byproducts unfit for other livestock
64 species. The enhanced functions provided by the microbiome are key for the characteristic adaptability and robustness of ruminants to cope with nutritional and climatic stresses
66 (Morgavi et al. 2013). Improving our understanding of the rumen microbiome provide opportunities for knowledge-based strategies aiming at enhancing efficacy in ruminant
68 production while minimizing its negative effect on the environment. In this study, we established a catalog of rumen prokaryotes genes that enabled us to decipher functions of the
70 microbiome, in particular the capacity to deconstruct structural carbohydrates from forages, and we explored the effect of feed on the microbiome composition and functions.

72 # RESULTS

Constitution of the bovine rumen prokaryotic gene catalog

74 To build a bovine rumen prokaryotic gene catalog, we collected total rumen content samples from five Holstein cows and five Charolais bulls. To reduce the ecosystem complexity and to
76 improve metagenome assemblies, rumen ciliated protozoa were depleted from the samples before microbial DNA extraction. A total of 1,206 Gb of raw metagenomic sequencing data
78 were generated with an average 111 Gb clean data for each animal. This sequencing depth, much greater than that used for other gene catalogs (5-10 Gb per sample (Li et al. 2014; Xiao
80 et al. 2015; Xiao et al. 2016)) was necessary to enable the assembly of the more complex rumen microbiome. After *de novo* assembly, open reading frames (ORF) prediction and
82 removal of redundancy, 13,825,880 non-redundant genes were obtained with an average length of 716 base pairs (bp) and 39% of these genes were complete ORFs (Supplementary
84 Table 1).

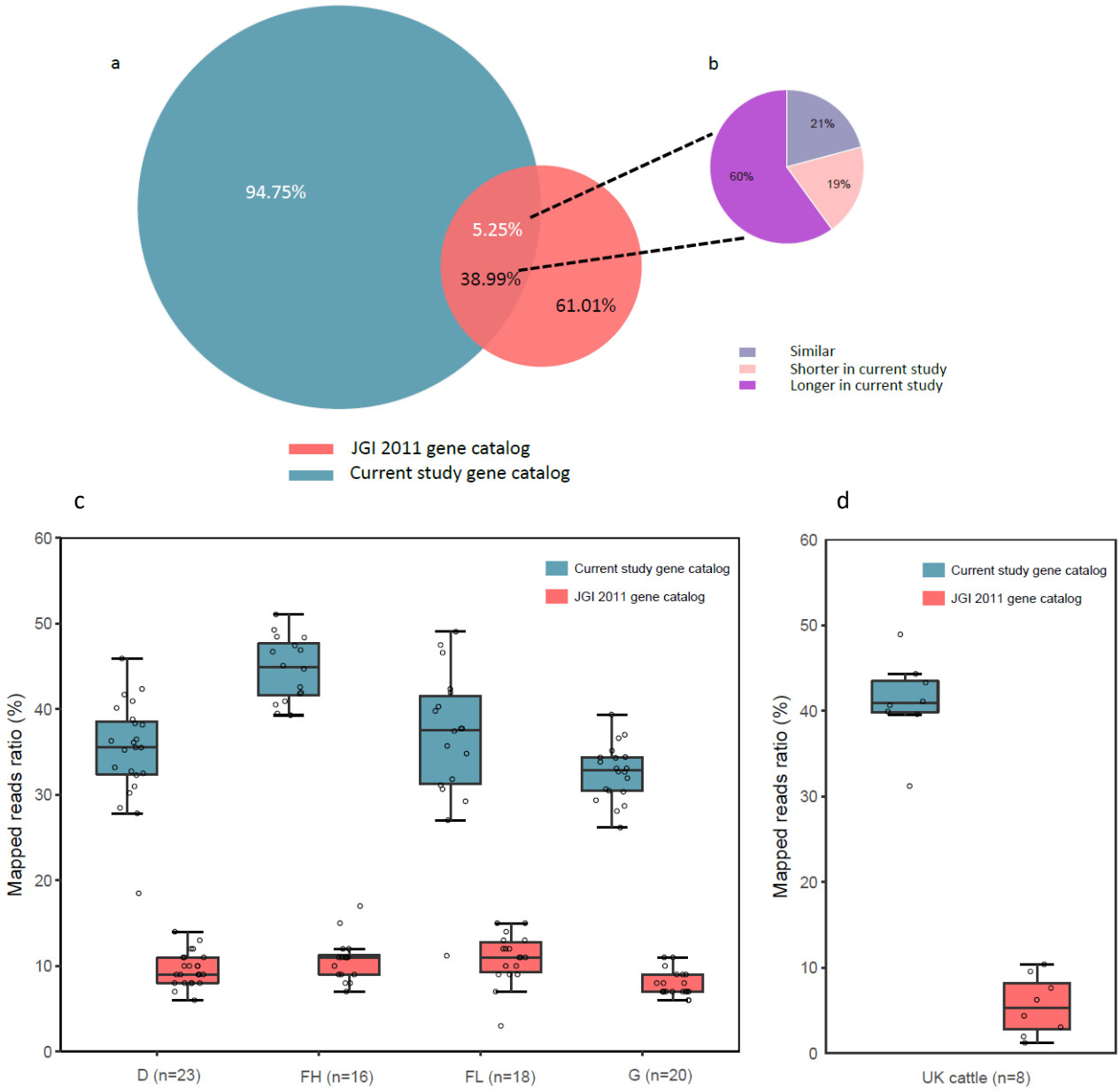

Figure 1. (a) Identity of current study genes compared to Hess et al. (2011). (b) Differences in gene length between studies. Green, the length of genes is longer in the current study; Blue, the length of genes is similar; Red, the length of genes is shorter in the current study. (c) Percentage of total reads in the current study (n=77 samples) in four different diet groups that could be mapped to current study gene catalog and JGI 2011 gene catalog. (d) Percentage of total reads in unrelated studies of 8 UK cattle samples that could be mapped to the current study gene catalog and JGI 2011 gene catalog.

Compared to the largest rumen gene catalog published to date by Hess et al. (2011), the number of non-redundant genes discovered in this study is more than 5 fold larger; shared genes were in most cases also longer (Figure 1a & b and Supplementary Table 2). Thus, the mapping reads from 77 additional rumen samples obtained in this study and eight published rumen samples from UK cattle (Wallace et al. 2015) increased from ~10% using the previous catalog(Hess et al. 2011) to ~40% (11-51%) (Figure 1 c & d). This confirms that the

representativeness of the rumen catalog has been greatly improved, even if the mapping efficiency was relatively low, as compared to 80% for the human gut microbiome (Li et al. 2014).

Comparison of gastrointestinal microbiomes: bovine rumen versus Human, pig and mouse

Genes were taxonomically classified using CARMA3 (Gerlach and Stoye 2011) and compared to genes from human, mouse and pig gut catalogs (Li et al. 2014; Xiao et al. 2015; Xiao et al. 2016). Up to 42.7% of rumen genes could be annotated to known phyla. This value is similar to pig gut (41.3%) but lower than the human (55.9%) and mouse gut metagenomes (59.6%) (Supplementary Figure 1 and Supplementary Table 3). Firmicutes and Bacteroidetes were predominant in all catalogs representing 84-94% of assigned genes and in accord with the expected gastrointestinal-associated microbial communities in mammals. For the rumen, however, the proportion of Firmicutes and that of Bacteroidetes was lower and higher, respectively, than for the other three catalogs. Other enriched phyla (>2%) in the rumen catalog were the Spirochaetes, Proteobacteria, Euryarchaeota, Actinobacteria and Fibrobacteres that, with the exception of Proteobacteria, were more abundant in the rumen than in the other catalogs (Supplementary Figure 2). At the genus level, only 8.7% of rumen genes could be annotated; a value similar to that of the other two animal catalogs but lower than that of human (16.8%), reflecting a more extensive characterization of human-associated microbes. However, the top 10 enriched genera in the rumen showed distinct abundance patterns compared with the same genera in other catalogs (Supplementary Figure 3 and Supplementary Table 3 & 4). These differences in symbiotic microbial genera likely reflect dissimilarities in dietary lifestyles, anatomical localization of the gut fermentation compartment and are indicative of predominant functions, i.e. methane production and plant fiber degradation for ruminants. *Prevotella* was the most abundant rumen genus with 39% of

116 genus-annotated genes assigned. Other abundant genera were *Treponema*, *Butirivibrio*, *Methanobrevibacter* and *Ruminococcus* that were absent or at lower proportions in other
118 catalogs, particularly in human.

Based on abundance profiles and clustering methods we identified 324 metagenomic
120 species (MGS), with an average size of 1.8 Mbp (minimum threshold of 1 Mbp; see Methods for more information on these MGSs). More than half (173) were medium-quality and 23
122 were high-quality drafts (>90% completion, <5% contamination) (Bowers et al. 2017). These rumen MGSs were also compared to pig and mouse MGSs. For the rumen, 314 were
124 annotated to the Bacteria domain and one to Archaea. For the bacterial MGSs, 44.4% could be annotated to the order level but only 3.7% (12 MGSs) to the genus level; 10 of these
126 belonged to *Prevotella*. The 324 MGSs were present in all four diets groups from our validation cohort; about 10% of reads from the 77 cattle dataset mapped to these MGSs. For
128 genomes from the Hungate 1000 project (Kelly), which are representative of the diversity of cultured rumen bacteria and archaea, mapping rate was 5.4%. In contrast, only 0.1% of reads
130 mapped to the metagenomic species described by Hess et al. (2011) (Supplementary Figure 4). Only 23 MGS were similar to genomes from the Hungate 1000 project, highlighting the
132 novelty represented by our MGS as more than 90% have no close cultured representative (Supplementary Table 5).

134 **Carbohydrate active enzymes in the bovine rumen metagenome**

The efficient deconstruction of structural plant polysaccharides by symbiotic gastrointestinal
136 microbes is what sets ruminants apart from other livestock species. We have therefore analyzed carbohydrate active enzymes (CAZymes) in the rumen ecosystem to obtain insights
138 into this important function for nutrition and health of cattle.

Glycoside hydrolases (GHs) and polysaccharide lyases (PLs) are the most relevant classes
140 of CAZymes as they orchestrate the breakdown of plant material and of diverse polysaccharides which are encountered in the rumen ecosystem, i.e. host, fungal, and bacterial
142 glycans. GHs and PLs are classified into sequence-based families (145 GH and 26 PL familes; www.cazy.org accessed in June 2017) that display a pronounced specificity for a
144 glycan category, thereby offering a functional readout of the degradative power of an ecosystem. The rumen catalog reported here encodes ~290,000 GHs (97.4%) and PLs (2.6%)
146 modules, which belong to 114 distinct GHs families and 18 PLs families. In the rumen catalog, the substrate specificity of the most abundant GH families reflects the prominent
148 glycan sources of herbivores: starch (GH13, GH77 and GH97, by decreasing abundance), pectins and hemicelluloses (GH43, GH28, GH10, GH51, GH9 and GH78, by decreasing
150 abundance). In contrast, only one of the 15 most abundant families, namely GH25 lysozymes, targets a non-plant substrate (peptidoglycan). Additionally, three of the five most
152 abundant families (GH3, GH2 and GH5) represent enzymes active on a wide range of substrates, not necessarily from plant origin. Two of these families (GH2 and GH3) contain
154 exo-glycosidases that act on the oligosaccharides produced by depolymerases, a broad function that may explain their abundance.

156 Dockerins domains (DOCs) are key building blocks of cellulosomes and amylosomes complexes. The DOC sequences are found in modular proteins and help the protein to which
158 they are appended to bind cohesin domains (COHs) found as repeats in large proteins named scaffoldins. This system allows the spatial grouping of numerous binding and enzymatic
160 modules into large assemblies for a synergistic action of their components in the immediate vicinity of the bacterial cell. In the rumen catalog, more than 12,000 dockerin modules were
162 identified. Intriguingly, some proteins harbored many dockerin modules, up to 13 modules in a single sequence, without any other recognizable functional module. The function of such

164 polydockerin proteins is unknown, and polydockerin proteins were not observed in reconstructed MGS (max. of two DOCs in a protein). In the literature, dockerin modules
166 initially detected in cellulosomes, have been investigated in relation to their co-occurrence with CAZymes in these cellulosome complexes (Brulc et al. 2009; Turnbaugh et al. 2010).
168 Surprisingly, our analysis of the rumen catalog reveals that only ~24% of the DOC-containing proteins carry a CAZyme domain. The remaining DOC-containing proteins were subjected to
170 a Pfam domain annotation, which identified proteases (4%) and some lipases (<0.3%), while a third of DOC-containing proteins are attached to non-catalytic modules, likely involved in the
172 binding of these non-carbohydrate substrates. More importantly, the last third did not have any match to any Pfam domain (Supplementary Figure 5).

174 The CAZyme profile in the rumen catalog was compared to the mouse, pig and human reference gut catalogs (Li et al. 2014; Xiao et al. 2015; Xiao et al. 2016). Despite important
176 differences in the size of these catalogs, similar trends could be observed on, for example, the ratio of DOCs or GHs plus PLs over the catalog size, or the most abundant GH families
178 (Supplementary Table 6). The number of distinct GHs/PLs is also very similar, and a detailed analysis highlighted 101 GHs families common to all four catalogs, while only five GH
180 families were specific to a single catalog (Supplementary Figure 6). These specific families were closely related to hosts' diets. In accord with herbivory, 305 GH45 cellulase modules
182 were found in the rumen catalog against none in human, mouse catalogs, and only 12 for the pig. In contrast, we identified families GH70 and GH68, transglycosidases acting on sucrose,
184 and GH47, processing N-glycan, that are absent in the rumen but present in other catalogs. For instance, 94, 24 and 6 GH70 modules were found in the human, pig and mouse catalogs,
186 respectively, whereas the rumen had zero occurrence.

The specific adaptation of the rumen catalog to herbivory was confirmed by comparing its
188 GH+PL family counts against the human catalog after normalization (Figure 2). The most

enriched GH families in the rumen are involved in the degradation of plant polysaccharides while the more depleted families of GHs are those degrading animal (host) glycans. These observations are not only in accord with cattle normal diet but they are also in agreement with the absence of a glycoprotein-rich mucus lining the rumen as opposed to the lower gastrointestinal tract. Finally, we also observed that multiple DOC module duplications seem to be more frequent and intense in the rumen as up to 13 DOC repeats in a single protein were found for the rumen catalog, compared to only six in the human, four in the pig and two in the mouse catalogs.

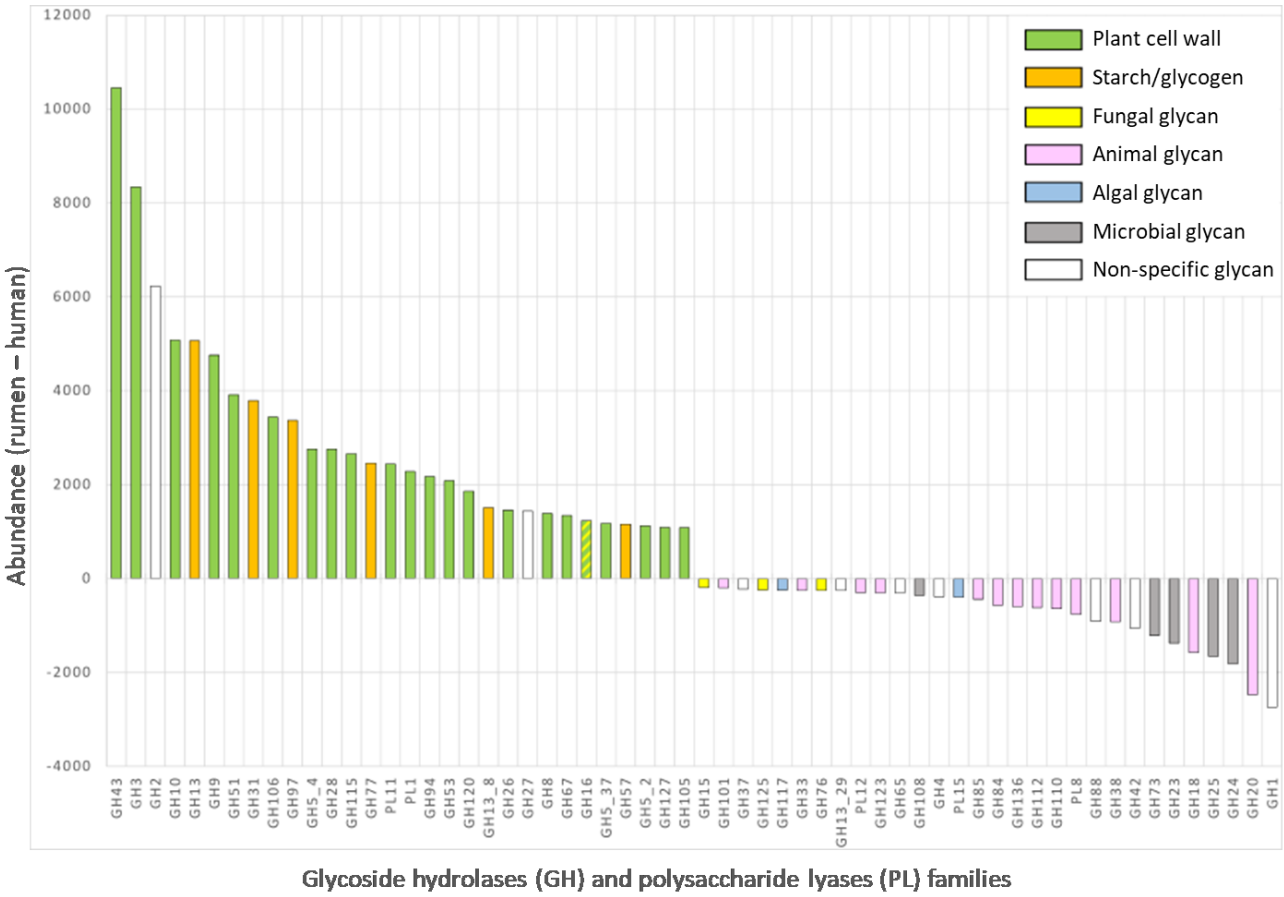

Figure 2. Enrichment or depletion of glycoside hydrolases and polysaccharide lyases in the bovine rumen as compared to human gut. Human counts were normalized to rumen catalog size before comparison.

CAZyme-encoding genes were also annotated in the 324 MGSs. Remarkably the most abundant families in the MGSs are for plant cell wall breakdown and correspond closely to the most abundant families in the non-redundant catalog. The CAZyme profiles of each

generated MGS were thus determined and subjected to a hierarchical clustering analysis
202 (Supplementary Figure 7) that revealed that MGSs group together roughly to their predicted taxonomy, even despite large differences in repertoire size within each phylum. Hereafter, we
204 detailed several strategies for carbohydrate foraging that evolved in the different bacterial phyla. Among the predicted Firmicutes, MGSs encoding cellulosomes and amylosomes
206 displayed a readily recognizable profile characterized by the presence of several DOC and COH domains along with several GHs families containing cellulases (GH5, GH44, GH48,
208 and GH124) and amylases (GH13 with associated CBM26), respectively. We also identified Bacteroidetes MGSs that contained a few DOC domains but, interestingly, none of these
210 MGSs appeared to contain a recognizable COH domain. The presence of dockerin domains not associated to cohesins in Bacteroidetes MGSs was recently reported in the moose rumen
212 microbiome (Svartstrom et al. 2017). The role of the dockerins in Bacteroidetes is unclear but the conspicuous absence of cohesins suggests that they may not be for the assembly of a *bona*
214 *fide* cellulosome or that the Bacteroidetes cohesins are so distantly related from their clostridial counterparts that they cannot be recognized.

216 Confirming previous reports in the literature (Kaoutari et al. 2013), the largest CAZyme repertoires dedicated to plant degradation were found among the predicted Bacteroidetes,
218 which represent the majority of the 324 reconstructed genomes. In Bacteroidetes, CAZymes are often grouped in distinct *Polysaccharide Utilization Loci* (PULs) around *susC* and *susD*
220 marker genes to build up specific depolymerization machineries capable of deconstructing in a synergistic manner even the most complex polysaccharides (White et al. 2014; Ndeh et al.
222 2017). In this context, it is interesting to note the clustering of families GH137 to GH143 recently shown to catalyze the breakdown of type II rhamnogalacturonan (Ndeh et al. 2017) in
224 the CAZyme profile heatmap (Supplementary Figure 7). Inspection of the predicted PULs in

the Bacteroidetes MGSs revealed the presence of degradation machineries dedicated to pectin
226 (type II rhamnogalacturonan), starch, or barley β-glucan (Supplementary Figure 8).

Other MGSs with distinctive CAZymes were those assigned to Proteobacteria and
228 Fibrobacteres that despite their small number (eight and six respectively) form tight groups. Predicted Proteobacteria were characterized by the presence of families GH84 and GH103
230 along with an important diversity of GH13 subfamilies. In contrast, the Fibrobacteres show the presence of several families known to degrade cellulose and β-glucans (*e.g.* GH5, GH45,
232 and GH55). Focusing on CAZymes from *Fibrobacter* spp. present in the catalogue revealed an astonishingly strain-level diversity for this genus. We compared the CAZymes present in
234 *Fibrobacter succinogenes* type species (Suen et al. 2011) against all Fibrobacter CAZymes in the catalog. There were 1262 hits with ≥ 90% identity to 135 of the 175 *Fibrobacter*
236 *succinogenes* CAZymes whereas only 19 of them had a 100% identity with the type strain. Up to 465 and 375 of these genes were differentially abundant in the Holstein and Charolais
238 groups, respectively (Supplementary Table 7). Zooming on a particularly important endoglucanase enzyme, i.e. GH45, reveals its presence in all 77 animals receiving different
240 diets. Animals harbored between 4 to 13 GH45 variants and each gene was present in 25 to up 99% of all animals; however, the type strain, at 79%, was not the variant most commonly
242 present.

Common functions and Influence of diet on the bovine rumen microbiome

244 To investigate how different feeds affected the rumen microbiota in beef and dairy cattle we examined beef Charolais bulls fed either a fattening diet high (n=16) or low (n=18) in starch
246 and lipids and Holstein cows fed a corn silage and concentrate diet (n=23) or grazed a natural prairie (n=20). By using this 77-sample dataset, differences in α-diversity were observed
248 between diets at the gene level. Animals fed fresh grass had the highest α-diversity and richness compared to other diets containing conserved feeds. Particularly, animals on

250 fattening diets had a lesser α-diversity. In contrast, the fattening diet rich in starch and PUFA had the highest β-diversity and/or had the highest disparity in interquartile range (box in the
252 boxplot) for all indices (Figure 3). The rumen microbiome of animals fed this diet also exhibited the highest dispersion on ordination analyses at the gene level (Supplementary
254 Figure 9). Such changes, akin to the described Anna Karenina principle (Zaneveld et al. 2017) for microbiomes, probably reflected divergences in individual microbiomes (and hosts)
256 reaction to PUFAs and may underlie a stress response to the diet.

Genes were annotated to known functions (KEGG and CAZy) and taxonomical
258 information. For functions, there were 43.3% of genes could be classified into KEGG orthology and 2.1% assigned to feed carbohydrate degradation. A total of 5,893 unique
260 KEGG orthologs (KOs) and 45,683 unique CAZy enzymes and binding modules were identified. Comparing the annotated genes for KEGG and CAZy functions showed a large
262 overlap among groups with 91% and 94% of shared genes, respectively (Supplementary Figure 10). In contrast with overall gene abundance, the highest α-diversity was observed for
264 the corn silage diet group (Figure 3). To assess the functions encoded by the minimal rumen metagenome we identified genes and KOs that were shared by all individuals in the group of
266 77 cattle. We found common sets of non-redundant genes, functions, genera and MGSs that were shared by all 77 rumen samples (Figure 4 and Supplementary Figure 11). The core gene
268 set shared by all animals represented only <0.1% (6051 to 12075 depending on the calculation method) of the nearly 14 M non-redundant genes in the catalog, whereas about 63% of the
270 KO functions (~3700) were shared indicating the high redundancy of genes for similar functions. Compared to all annotated KO, this minimal KO set was significantly enriched in
272 pathways related to metabolism (amino acids, carbohydrate, nucleotides and metabolism of cofactors and vitamins), cellular processes (motility), and genetic information processing
274 (translation) (Supplementary Figure 11b). Concerning the diversity of genera found in the

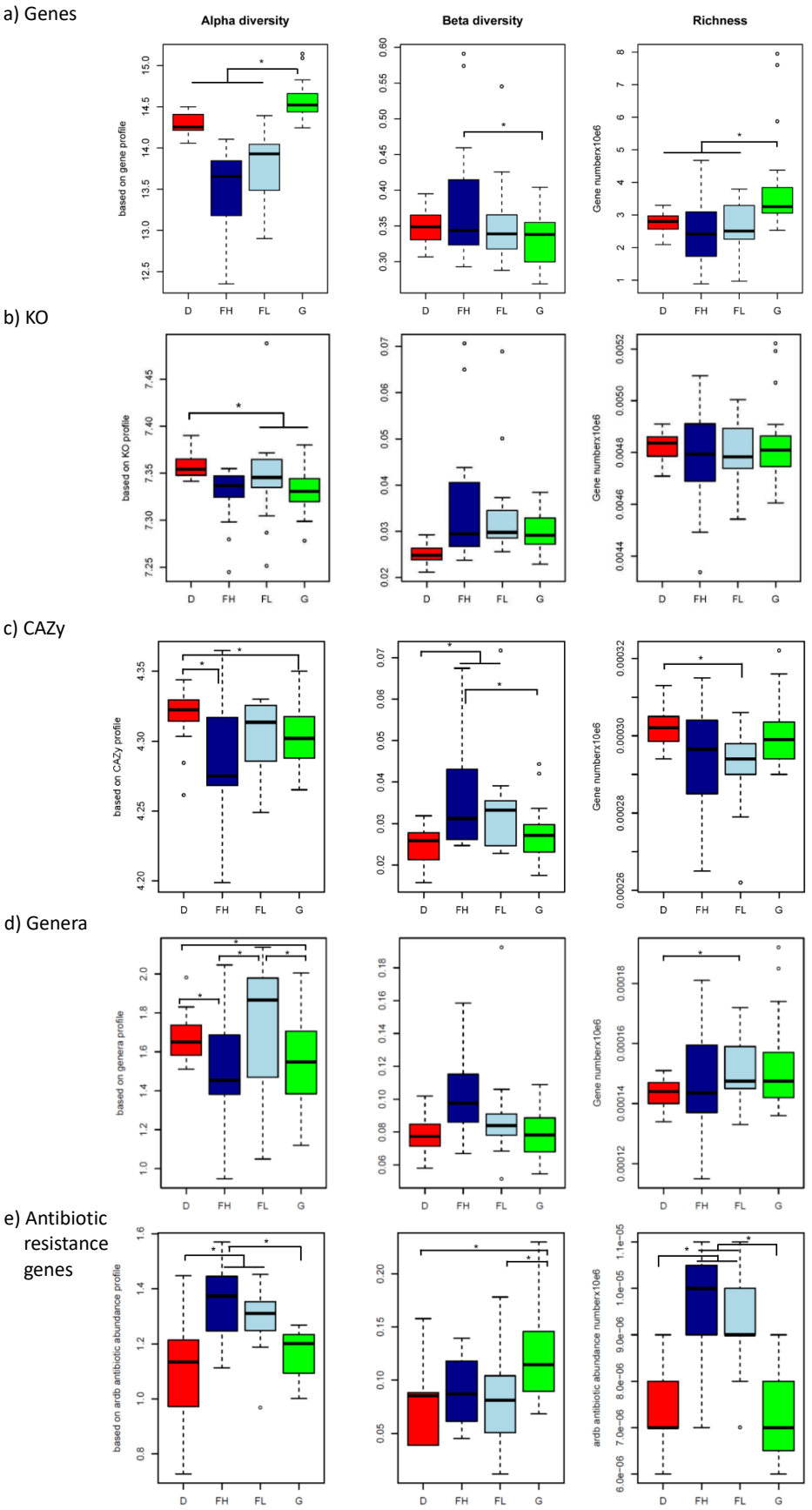

Figure 3. Effect of diet on diversity indexes of the bovine rumen microbiome. Comparison of Alpha diversity, Beta diversity and Richness at Gene (A), KO (B), CAZy (C), Genera (D) and antibiotic resistance gene (E) levels among cattle fed: dairy (red, n=23), fattening high-start (dark blue, n=16) fattening low-starch (light blue, n=18) and grazing (green, n=20) diets. * indicates P<0.05.

different groups, there was also a relatively large overlap. Out of 242 genera identified by the
278 taxonomic analysis described above, 182 (75%) were present in all four groups but only 67
(27%) were shared by all animals (Figure 4 and Supplementary Figure 10). This overlap was
280 maximal for MGSs identified in this study, which were present in virtually all individuals
(Figure 4). The presence of common functions may explain the plasticity of the microbiota
282 and adaptability of ruminants to digest various types of feeds even after sudden dietary
changes. To get a better understanding of the functional changes induced by diet in these
284 microbial communities we analyzed the abundance of genes in the 77-sample dataset for
functions, genera and MGS. To avoid possible confounding effect of breed and sex, the
286 differential abundance analysis was performed within each breed. For Holstein, greater
changes in the relative abundance of genes were observed; ~43% difference in KEGG and
288 CAZy functions (Supplementary Table 8 and Supplementary Figure 12 a & b). For CAZy,
146 catabolic families exhibited indeed differences in abundance between the corn silage and
290 grazing groups (Supplementary Figures 12a and 13, Supplementary Table 9). Most of the
differences related to functions were due to increases in the relative abundance of genes in
292 cows fed the corn silage diet rather than the presence of different genes. Notwithstanding, the
greatest contrast was observed for families targeting fructans and sucrose that were more
294 abundant in the grazing group. Particularly for family GH32 (P = 7.6E-12) whose higher
abundance could be related to the high contents of sucrose and fructans in grasses
296 (Ghasempour et al. 1998; Vijn and Smeekens 1999) included in the grazing diet. The other
CAZy families differing in abundance were all more abundant in the corn silage-diet group.
298 Interestingly, these results highlight the ability of ruminal bacteria to be equally capable of
using glycans from plants as well as from microbial origin such as bacterial peptidoglycans,
300 bacterial exopolysaccharides and fungal cell walls. Corn silage, the main constituent of the
diet, is a fermented feed with an abundant epiphytic microbiota composed of

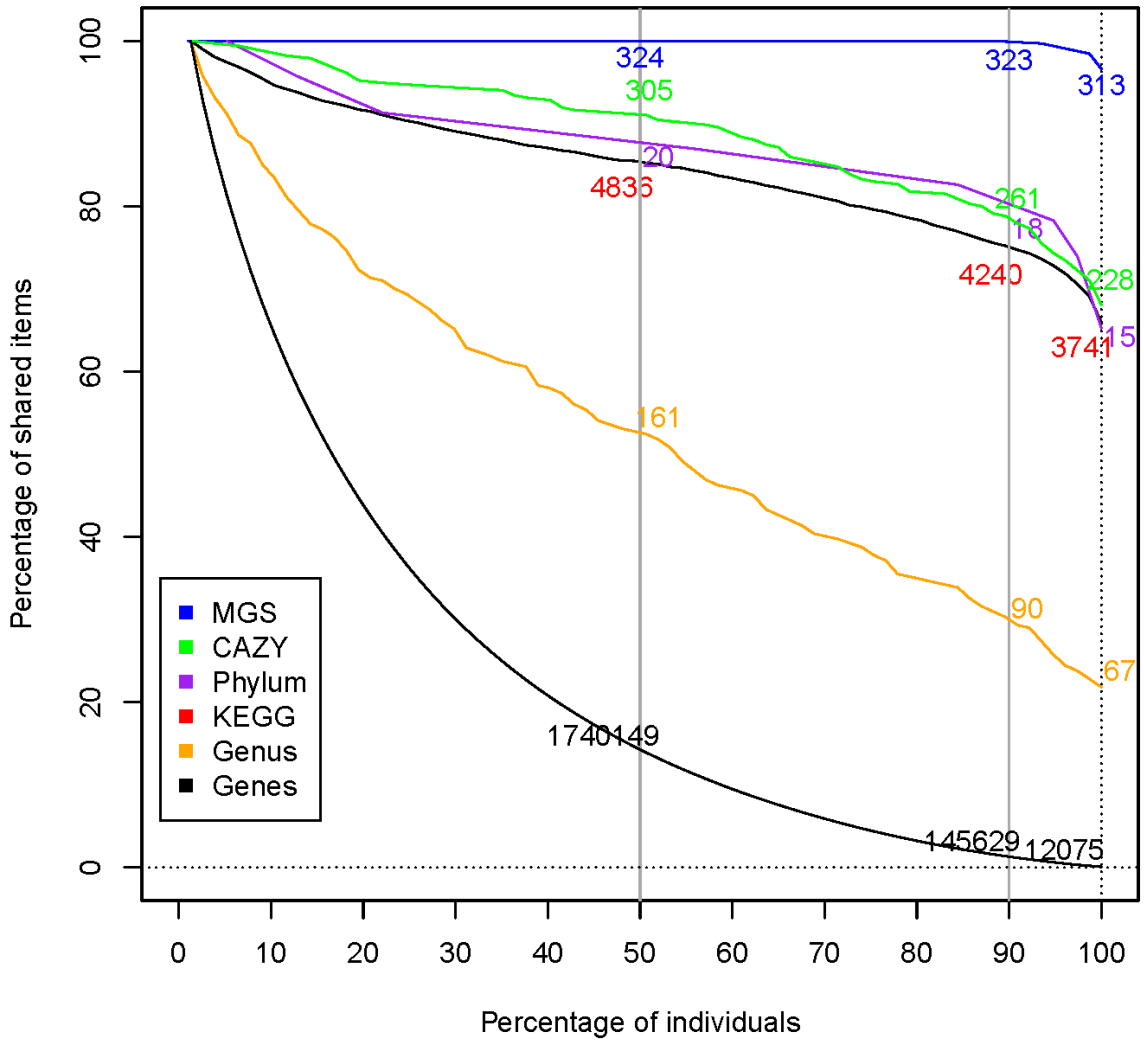

Figure 4. Size of the shared microbiome features among cattle (n= 77) fed four different diets for the number of genes (black), genera (orange), phyla (purple), MGSs (blue), KEGG pathways (red), and CAZy (green). The percentages of shared items and animals are represented on the y and x axes, respectively. The absolute numbers for each item are indicated at the intercept between the percentages of items and animals at the thresholds of 50, 90 and 100%. Only ~1% of genes were shared by 90% of the cattle, whereas close to 80% of KO and CAZy functions were shared by 90% of the cattle, suggesting gene redundancy for similar functions. To note the presence of most MGSs assembled in this work in 90% of the cattle.

302 exopolysaccharide-producing lactic acid bacteria, fungi and yeasts (Storm et al. 2010; Cheli et al. 2013). Accordingly, the CAZome of the corn silage-diet group was oriented towards

304 degradation of starch, a nutrient abundant in corn silage and practically absent in the grazing diet. Forty-two CAZy families targeting plant cell wall polysaccharides were also

306 overabundant in the corn silage-diet group. This could reflect the diversity of fiber structures that ruminal bacteria have to face when cows are fed with such a diversified diet in terms of

308 plant fractions and botanical origins (whole corn plant and soybean meal in the corn silage diet against a natural prairie, composed predominantly of grasses in the grazing diet). Finally,

310 the overabundance of CAZy families targeting animal glycans in the silage-fed cohort was

striking since no glycoprotein-rich mucus is secreted in the bovine rumen as opposed to the
312 lower gastrointestinal tract (Hoorens et al. 2011). It is possible that this difference reflects
that CAZy families targeting animal glycans harbor numerous enzymes that are not fully
314 characterized and may be able to act on plant or even fungal glycans, which contain a panel of
osidic constituents that are very similar to that of animal glycans. Enzyme promiscuity may
316 indeed confer metabolic flexibility and an ecological advantage to certain microbes in the gut
ecosystem.

318 For genera and MGSs, up to 44% (106 genera) and 58% (188 MGSs) of the total detected
were differently abundant in the microbial communities of the two cows' groups
320 (Supplementary Table 10). *Fibrobacter* and *Ruminoccocus* were more abundant in the corn
silage diet group whereas *Prevotella*, *Butyrivibrio* and *Methanobrevibacter* were more
322 abundant in the grazing group.

For Charolais on fattening diets differing in starch content, less than 5% differences were
324 observed in the abundance of genes for functions or genera. Only eight CAZy families
exhibited differences in abundance between the two Charolais groups, the differences in
326 abundances being less significant than for the Holstein groups ($p > 0.004$) (Supplementary
Figures 12a and 13, Supplementary Table 9). The absence of marked variations in the
328 abundance of glycoside-degrading enzymes between the two fattening diets reflects indeed
their similar composition. The differences in starch content was not great enough to
330 drastically impact the carbohydrate harvesting functions of the ruminal microbiota, at least at
the gene level. Similarly, smaller differences in the abundance of genera and MGSs were
332 detected between these two diets (Supplementary Table 8, Supplementary Figure 12d).

Metadata collected on the Holstein and Charolais animals were analyzed using a vector
334 fitting method on the top of the bidimensional NMDS ordination (Supplementary Table 11).

Diet had a significant effect on metagenome gene distribution, particularly in the Holstein
336 group (r^2= 0.68, P = 0.0001), but also variables such as live weight, feed intake and rumen volatile fatty acids were significant. Protozoal numbers were also a significant variable
338 explaining the distribution of genes in the metagenome of animals, underpinning their importance as key members of the rumen ecosystem and modulators of the prokaryotic
340 community.

Antibiotic resistance genes

342 We evaluated the presence of antibiotic resistance genes (ARGs) in the rumen gene catalogue as previously reported (Xiao et al. 2016). Forty-two ARGs encoding resistance to 27
344 antibiotics were detected in the catalog. The most abundant resistances were to tetracycline and bacitracin with Charolais animals harboring globally a higher proportion of these genes
346 (Supplementary Figure 14). It is noted that antibiotics as growth promoters were never used on these animals. In both the bovine rumen and the porcine gut (Xiao et al. 2016), the most
348 abundant ARGs confer resistance to tetracycline and bacitracin. The diversity of ARG is low compared to pig feces where resistance to up to 52 antibiotics was reported, even in farms
350 with no use of growth promoting antibiotics (Xiao et al. 2016).

DISCUSSION

352 Ruminants are extraordinary bioreactors, engineered by nature to use recalcitrant plant biomass—a renewable resource— as feedstock for growth and for production of useful
354 products. This ability is a microbial attribute that was important in domestication and that today has a renewed interest due to human population increases, resource scarcity and climate
356 change issues. The reference gene catalog from the rumen microbiota reported here is a useful resource for future metagenomics studies to decipher the functions and interactions of
358 this complex ecosystem with feeds and the host animal. Comparison with human, mouse and

pig gut catalogs shows the distinct character and potential of the rumen ecosystem. As
360 opposed to the microbiome of single-stomached animals including humans, the rumen
microbiome harbors a plethora of genes coding for glycoside hydrolases (CAZymes) that
362 degrade structural polysaccharides. Information on these enzymes that deconstruct biomass
plant material and are essential for transforming recalcitrant feeds into meat and milk is also
364 useful for the design of improved processes for the biofuel industry (Weimer et al. 2009; Liao
et al. 2016).

366 The type of diet modulated as expected the abundance of genes and the metagenome profile
of individuals. However, more than 90% of genes coding for functions (KO and CAZy) were
368 shared among animals receiving different diets. This large functional diversity might be the
key that allows ruminants to feed on a variety of dietary sources and to adapt to seasonal or
370 production-imposed dietary changes. The 13.8M genes catalog produced in this work, despite
being significantly larger than gut bacterial catalogs from other species (Li et al. 2014; Xiao et
372 al. 2015; Xiao et al. 2016) does only partially cover the diversity present in the rumen
microbiome indicating the higher complexity of this ecosystem. The catalog needs to be
374 expanded with additional data, particularly the inclusion of ciliated protozoa and fungi to
reflect the overall diversity. Nevertheless, this catalog and the 324 uncultured assembled
376 genomes are an important instrument to characterize and understand the biological functions
of the rumen microbiome. This information is essential to enhance the sustainability of
378 ruminant production.

METHODS

380 This study was conducted using the animal facilities at the French National Institute for
Agricultural Research (INRA) in Theix and Bourges, France. Procedures on animals used in

382 this study complied with the guidelines for animal research of the French Ministry of Agriculture and all other applicable National and European guidelines and regulations.

384 **Rumen Sampling**

Total rumen content samples from 10 animals used for deep sequencing metagenome were 386 taken at the experimental slaughterhouse of the INRA Centre Auvergne-Rhône-Alpes. Total rumen content samples from 77 animals were also collected. These 77 animals, from two 388 different genetic stocks, were fed diets characteristics of beef and milk production systems. Beef cattle, represented by Charolais breed were fed fattening diets high (n=16) and low 390 (n=18) in starch and lipids; whereas Holstein dairy cows were fed a corn silage and concentrate diet (n=23) or grazed a natural prairie (n=20) (Supplementary Table 12). Rumen 392 samples from these animals were also collected at the experimental slaughterhouse except for the grazing group. Cows from this latter group were fitted with rumen cannula and samples 394 were taken from live animals.

Sample handling and DNA extraction

396 The 10 rumen samples used for deep sequencing were depleted from eukaryotes using washing and centrifugation steps. Rumen contents were filtered through a 400 µm nylon 398 monofilament mesh. The filtrate was centrifuged at 300 *g*, 5 min to decant protozoa and the supernatant (fraction A) was stored at 4 °C. Fifty grams from the filtered rumen content 400 retentate were mixed with 100 ml of anaerobic phosphate saline buffer (PBS), mixed manually for 5 min by gentle inversion, centrifuged at 300 *g*, 5 min to decant protozoa and the 402 supernatant, passed through a 100-µm filter (fraction B), was stored at 4 °C. The pellet was mixed with 75 ml anaerobic, ice-chilled 0.15% Tween 80 in PBS and incubated on ice for 2.5 404 h to detach microbes attached to feed particles. At the end of the incubation, contents were vortexed for 15 s and centrifuged at 500 *g*, 15 min. The supernatant (fraction C) was mixed 406 with fraction B and 50 ml of fraction A and centrifuged at 20,000 *g*, 20 min, 4 °C. The

supernatant was decanted and the microbial pellet was exposed to an osmotic shock to lyse
408 any remaining eukaryote (mainly protozoal) cells followed by an endonuclease treatment.
Briefly, the pellet was suspended in water (Millipore Waters Milli Q purification unit) and
410 incubated for 1 h at room temperature followed by DNase treatment (Benzonase, Novagen) as
described (Hunter et al. 2011). The suspension was filtered through a 10 µm monofilament
412 textile, collected by centrifugation as before, suspended in an appropriate volume of PBS and
stored at -80 °C until DNA extraction. DNA was extracted following the method described
414 by Yu and Morrison (2004). Samples from 77 animals were extracted directly from whole
rumen contents using the same extraction method.

416 **DNA library construction and sequencing**

Paired-end (PE) metagenomic libraries were constructed and sequenced following Illumina
418 HiSeq2000's instruction. Quality control and bovine DNA removal (by aligning reads to *Bos*
taurus genome Btau_4.0 (Elsik et al. 2016)) for each sample were independently processed
420 using MOCAT pipeline as previously described (Kultima et al. 2012). On average, 111.3 Gb
of high-quality bases was generated for each of the 10 deep sequencing samples and 3.43 Gb
422 (median ~2.5 Gb) for each of the 77 samples (Supplementary Table 12). The averaged
proportion of high-quality bases among all raw bases from each sample was 92.29%.

424 **Public data use**

Three public rumen microbial datasets used in this study include: (i) a cow rumen microbiome
426 sequenced at DOE's Joint Genome Institute (JGI) in 2011 (JGI 2011), which consists of
268 Gb of metagenomics sequences, 2,547,270 predicted genes and 15 uncultured microbial
428 genomes assembled from the cow rumen (Hess et al. 2011) (NCBI accession number
SRA023560), (ii) 8 rumen metagenomics samples from beef steers (Wallace et al. 2015)
430 (European Bioinformatics Institute (EBI), PRJEB10338), (iii) 409 reference rumen microbial
genome sequences and one rumen metagenomic dataset from the Hungate1000 Project

432 (Kelly) (Supplementary Table 13) (Integrated Microbial Genomes (IMG), JGI Proposal Id: 612 / 300816).

434 Three public gut microbial gene datasets from human[1] (GigaDB, doi:10.5524/100064), mouse (Xiao et al. 2015) (GigaDB, doi:10.5524/100114) and pig (Xiao et al. 2016) (EBI, 436 PRJEB11755) were also collected.

Construction of the rumen microbial gene catalog

438 High quality reads from 10 deep sequenced samples were processed in MOCAT toolkit (Kultima et al. 2012) including de novo individual assembly (SOAPdenovo v1.06 (Li et al. 440 2010), -K 47). The assembled contigs with length equal to or greater than 100 bp were followed by gene prediction (MetaGeneMark (Zhu et al. 2010), –M 100 –A) and redundant 442 genes were removed (CD-HIT (Li and Godzik 2006), ≥ 95% identity and ≥ 90% overlap), resulting in a non-redundant rumen microbial gene catalog containing 13,825,880 genes 444 (Supplementary Table 1).

Evaluation of current rumen microbial gene catalog

446 To assess the representative of our rumen gene catalog, we used the largest rumen gene catalog published to date by Hess, et al.(Hess et al. 2011). First, the genes with gaps were 448 filtered as follows: genes were broken when meet '*N*' base, a subset for each interrupted gene was obtained, retaining only the longest sub-gene as representative of the original gene. A 450 total of 2.46 M genes without gaps were obtained, termed 'JGI-2011-gene-catalog' and used for following analysis (Supplementary Table 2).

452 Further, 13.83 M genes from current study and 2.46 M genes from JGI were pooled together to identify shared genes using CD-HIT (Li and Godzik 2006). The comparison of 454 gene length between the two catalogs was conducted as previously described (Li et al. 2014).

Gene catalog annotation

456 Taxonomic assignments of genes from rumen, mouse, pig and human guts were performed using CARMA3(Gerlach and Stoye 2011) on the basis of BLASTP (Altschul et al. 1990) (
458 V2.2.24) against the NCBI-NR database (v20130906 for rumen, mouse, pig guts; v20160219 for human gut) (Supplementary Table 3). Microbiotas from these four species were compared
460 at different taxonomic levels. Functional annotation based on Kyoto Encyclopedia of Genes and Genomes (KEGG) database was performed using an in-house pipeline (Li et al. 2014).
462 Annotation of the carbohydrate-active enzymes (CAZymes) of each catalog was performed by comparing the predicted protein sequences to those in the CAZy database and to Hidden
464 Markov models (HMMs) built from each CAZy family (Lombard et al. 2014), following a procedure previously described for other metagenomics analyses (Svartstrom et al. 2017). In
466 order to allow a direct comparison of the results, annotation of antibiotic resistance genes (ARGs) was done as previously reported in the pig metagenome catalogue (Xiao et al. 2016)
468 by using the ARDB database (Liu and Pop 2009).

Construction relative abundance profiles of genes, KOs, ARG and CAZY enzymes

470 The gene profiles of 77 rumen samples were generated by aligning high-quality clean reads to the current 13.83M gene catalogue (SOAP2, ≥95% identity) (Qin et al. 2012). Gene relative
472 abundance was estimated as described previously (Qin et al. 2010). The relative abundance of each KEGG orthologous group (KO), ARGs and CAZy enzyme was calculated from the
474 abundance of its genes (Li et al. 2014).

Characterization of total and minimal metagenome

476 We computed the total and shared number of genes, KO and CAZy functions present in random combinations of *n* individuals (with n=2 to 77, 100 replicates per bin) (Qin et al.
478 2010). Furthermore, we used a permutation test to identify the second-level KEGG functions that were significantly enriched or depleted in the minimal KO set compared with the total

480 KO set. We first calculated the contribution of second-level functions using the following formula:

482 $$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}}, P_j = \frac{\sum_i p_{ij}}{N}$$

Where f_{ij} is the number of second level function j from the KO i ; p_{ij} is the relative

484 contribution of second level function j in the KO i ; N is the number of KO in the KO set; P_j is the relative contribution of function j in the KO set.

486 Randomly sampling 999 times in all annotated KO set, simulated the distribution of each function. Calculating the position of this function contribution ratio of minimal KO set under

488 the distribution of all annotated KO set. A p value of less than 0.01 was regarded as significant (Supplementary Figure 11b).

490 **Construction of metagenomic species (MGS) and taxonomic assignment**

To recover the draft bacterial and archaeal genomes from the 10 deep sequenced samples, we

492 developed an in-house pipeline that comprises three steps as indicated in **Supplementary Figure 18** and described below.

494 *1. Construction of Scaftig-Linkage Groups (SLGs)*

We generated a scaftig abundance profile by aligning high-quality clean reads from 77

496 rumen samples to assembled scaftigs from samples (Qin et al. 2012). Scaftig relative abundance was determined using the same method for gene abundance (Qin et al. 2012). The

498 highly co-abundance correlated scaftigs from each deep sequencing sample were binned into scaftig-linkage groups (SLGs) using the previous pipeline (Qin et al. 2012) with modified

500 parameters as follows, an edge was assigned between two scaftigs sharing Pearson correlation coefficient > 0.7 and the minimum edge density between a join was set as 0.99. A total of

502 745 preliminary SLGs with length > 1Mbp were generated for further analysis.

2. Filtering of Preliminary SLGs based on GC content and assembly outputs

504 For all preliminary SLGs, we then examined their specificity by plotting the GC content versus reads aligned depth of each scaftig. In this step, 520 SLGs containing sole GC cluster
506 were treated as 'qualified' and retained for the step 3. For the remaining 225 SLGs, 184 presented a scattered GC distribution and were discarded whereas the 41 SLGs containing
508 two or more GC clusters were further processed. First, those SLGs with scaftig N50 <2000bp were considered as too fragmented and discarded. Then, multiple GC clusters in remaining
510 SLGs were separated by DBSCAN(M. Ester et al. 1996) (Eps<=0.10, MinPts>=49). After splitting and filtering, we retained 55 'qualified' SLGs that had a coverage depth greater than
512 20×.

3. Reconstruction of metagenomic species

514 In order to improve the completeness and remove the redundancy of multiple metagenome assemblies from 10 deep sequencing samples, we performed hierarchical clustering for these
516 575 qualified SLGs based on their scaftigs nucleotide identity calculated by MUMi (Deloger et al. 2009). The MUMi distance between two SLGs (a and b) was defined as:

518 $$\mathrm{MUMi} = \frac{1 - \frac{L_{unmap\ length\ of\ a} + L_{unmap\ length\ of\ b}}{L_{total\ length\ of\ a} + L_{total\ length\ of\ b}}}{M}$$

Where $M = 2 \times \min\left(L_{total\ length\ of\ a}, L_{total\ length\ of\ b}\right) / \left(L_{total\ length\ of\ a} +
520 L_{total\ legnth\ of\ b}\right)$, $L_{total\ length\ of\ a}$ is the length of SLG a, and $L_{unmap\ length\ of\ a}$ is the length of unmapped sequence compared with SLG b. The threshold for generating a species level
522 metagenomic species (MGS) at 0.54 for MUMi as previously suggested (Backhed et al. 2015). Two hundred and eighteen qualified SLGs could not be clustered with other SLGs and
524 were defined as singleton-MGSs. The remaining 357 qualified SLGs were clustered into 105

candidate-MGSs. We performed overlap-based assembling on the scaftigs for each of these
526 105 candidate-MGSs respectively, using Phrap with default parameters. To get reliable contiguous sequences for each candidate-MGS, the overlaps between two scaftigs less than
528 500bp were considered as unreliable and re-broken.

The 105 reconstructed candidate-MGSs were further examined using GC patterns using the
530 same method mentioned in step 2 above. Eighty out of the 105 candidate-MGSs containing sole GC cluster were retained as combined-MGS. The remaining 25 candidate-MGSs
532 containing two or more clusters were split into sub-MGSs using the same method mentioned in step 2 above. In order to preserve the most comprehensive genomic information for these
534 sub-MGSs, sequences from each sub-MGS was aligned back to its original SLGs. If the sub-MGS covered 90% or more sequences of its original SLG, it would be retained as a revised-
536 MGS. Otherwise, its original SLG will replace the corresponding sub-MCs and be considered as a revised-MGS. This splitting step finally obtained 31 revised-MGSs.

538 After filtering the total sequence size of 218 singleton-MGSs, 80 combined-MGSs and 31 revised-MGSs with the criterion of > 1Mbp, we finally obtained 324 MGSs for rumen
540 microbiota including 224 singleton-MGSs and 100 combined-MGSs (Supplementary Table 14). We used the same pipeline described above for the gene catalog for the ORF prediction
542 and taxonomic annotation of MGS genes. We used CheckM (Parks et al. 2015) to estimate of the completeness, contamination and heterogeneity of metagenomic species (Supplementary
544 Table 14). MGSs were assigned a taxonomic level annotation if more than 50% of its genes were assigned at a given taxonomic level (including genes with no match) (Supplementary
546 Table 15). The MGS relative abundance of 77 rumen samples was calculated from the relative abundance of its aligned genes.

548 **Quality assessment and taxonomic annotation of MGSs**

CheckM software (Parks et al. 2015) was used to calculate the completeness and
550 contamination of these MGSs. The median percentage of completeness was high at 62.5% with a low, 2.6% contamination. The combined-MGSs showed higher completeness but also
552 slightly higher levels of contamination and strain heterogeneity than singleton-MGSs (Supplementary Figures 15, Supplementary Figures 16). Taxonomic annotation for rumen
554 MGSs was performed using CARMA3 on the basis of BLASTP against the NCBI-NR database (v20130906) and compared with MGSs from pig and mice (Supplementary Table
556 14, Supplementary Figure 17).

Cluster distribution by diet at species level

558 The relative MGS abundance profile (matrix of 324 × 77) obtained above was analyzed to highlight differences induced by diet. As we found when coverage of a MGS is less than 0.1
560 the depth of this MGS is close to 0 (Supplementary Figure 19). This result is caused by the noise and non-conducive to the MGS clustering. Therefore, when the coverage value was less
562 than this threshold value, then we set the value of depth equal to 0.

Ordination and differentially abundance analyses

564 Breed and diet distribution were visualized in ordination analyses based on two-dimensional non-metric multidimensional scaling (Shepard 1962). Dissimilarity between pairs of samples
566 was calculated using Bray–Curtis dissimilarity index (Bray and Curtis 1957). Vegan R package (Oksanen et al. 2016) was also employed to estimate the diversity indexes
568 corresponding to richness, alpha (Shannon index) and beta diversity (Whittaker). The '*envfit'* function of Vegan was used to determine whether phenotype information corresponding to the
570 77 samples contribute to the overall pattern of the rumen microbiome structure. The significance of the environmental factors was assessed after 9999 permutations.

572 The relative abundance of the 13,825,880 non-redundant genes was collapsed into taxonomic (Phylum and Genus) and functional levels (KEEG and CAZy). Procrustes rotation
574 analysis was performed to compare the ordinations obtained at different levels. Identified KOs were mapped to KEGG and visualized using the Interactive Pathway Explorer (iPath2.0)
576 web-based tool (Yamada et al. 2011). To estimate a core, the overlapping number of Genus, CAZy and KOs between Holstein and Charolais breeds was compared.

578 To avoid confounding factors such as: sex, breed and age, the differentially abundance analysis was performed within breeds. Therefore, for each breed, diet comparison was done
580 based on a Zero-Inflated Gaussian mixture model as implemented in the fitZig function of the metagenomeSeq R package (Paulson et al. 2013). Correction for multiple testing was done,
582 and the cut-off of the differential abundance was set at FDR ≤ 0.05.

DATA ACCESS

584 Metagenomic sequencing data generated in this study have been deposited in EBI database under the accession code PRJEB23561. The data of assembled scaftigs, the rumen gene catalog,
586 the rumen MGS catalog, and the abundance profile tables generated in this study have been deposited in *GigaScience Database* (doi: to be completed).

588 Note: accession numbers are in process.

ACKNOWLEDGMENTS

590 This work had the financial support of Animal Physiology and Livestock Systems Division of INRA, the INRA metaprogramme Meta-omics of Microbial Ecosystems (MEM), and BGI-
592 Shenzhen. This study was supported by the National Natural Science Foundation of China (No.31601073), the Shenzhen Municipal Government of China (No.
594 JSGG20160229172752028), the Shenzhen Key Laboratory of Human commensal

microorganisms and Health Research (No.CXB201108250098A). The work on CAZy was
596 supported by a European Union's Seventh Framework Program (FP/2007/2013)/European Research Council (ERC) Grant Agreement 322820 to BH. Work by Metagenopolis was
598 supported from grant ANR-11-DPBS-0001. YRC's salary was funded by the European Union, in the framework of the Marie-Curie FP7 COFUND People Programme, through the
600 award of an AgreenSkills' fellowship (grant number 267196) linked to the MEM METALIT project.

602 The authors acknowledge technical support for animal care, sampling and analytical measures on live animals provided by the personnel at INRA's experimental units of
604 Herbipôle (P. Faure and D. Roux) and Bourges, and the Herbivores research unit (Y. Rochette, F. Anglard, and B. Sepchat). Particular thanks to D. Graviou for DNA extraction
606 and biochemical analysis.

We thank the direction of INRA divisions Microbiology and Food Chain, Animal
608 Genetics, and Science for Food and Bioproduct Engineering for their support. We also thank E. Forano, and M. Naves (INRA) for helpful discussions during the setup of the project.

610 **Author contributions**

JL, HZ, SDE, and DPM designed the work and managed the project. JL and HZ designed the
612 analyses, and analyzed and interpreted the sequencing data. JL, HZ, SDE, GPV, BH, NT, VL, YRC, JE, and DPM wrote the manuscript. GPV, BH, NT, and VL conducted data
614 analysis on CAZy and were involved in the interpretation of data. YRC and JE conducted integrative data analysis, implemented ARG analysis and were involved in the interpretation
616 of data. MP and CM were involved in the implementation of animal studies, samples and metadata collection. ZY, HZ, ST, and FL performed data analyses, constructed and annotated
618 the MGS catalog. WC, BC, and JLI performed data analyses, constructed and annotated the

gene catalog. JG contributed to the experimental development and discussion for filtering
620 rumen samples. MP, EM, XX, HY, LM, and JW contributed to text revision and discussion.
KK interpreted the data, revised the paper. DPM coordinated the project. All authors
622 approved the submitted versions and agree to be accountable for all aspects of the work.

DISCLOSURE DECLARATION

624 The authors declare no competing financial interests.

REFERENCES

626 Aiking H. 2014. Protein production: planet, profit, plus people? *Am J Clin Nutr* **100**: 483S-489S.

628 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

630 Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H et al. 2015. Dynamics and stabilization of the human gut microbiome during
632 the first year of life. *Cell Host Microbe* **17**: 690-703.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz
634 F, Jarett J, Rivers AR, Eloe-Fadrosh EA et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG)
636 of bacteria and archaea. *Nat Biotech* **35**: 725-731.

Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern
638 Wisconsin. *Ecol Monogr* **27**: 325-349.

Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards
640 RE, Frank ED, Emerson JB, Wacklin P et al. 2009. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases.
642 *Proc Natl Acad Sci USA* **106**: 1948-1953.

Cheli F, Campagnoli A, Dell'Orto V. 2013. Fungal populations and mycotoxins in silages:
644 From occurrence to analysis. *Anim Feed Sci Tech* **183**: 1-16.

Deloger M, El Karoui M, Petit MA. 2009. A genomic distance based on MUM indicates
646 discontinuity between most bacterial species and genera. *J Bacteriol* **191**: 91-99.

Elsik CG, Unni DR, Diesh CM, Tayal A, Emery ML, Nguyen HN, Hagen DE. 2016. Bovine
648 Genome Database: new tools for gleaning function from the Bos taurus genome. *Nucl Acids Res* **44**: D834-D839.

650 Ertl P, Knaus W, Zollitsch W. 2016. An approach to including protein quality when assessing the net contribution of livestock to human food supply. *Animal* **10**: 1883-1889.

652 Gerlach W, Stoye J. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucl Acids Res* **39**: e91-e91.

654 Ghasempour HR, Gaff DF, Williams RPW, Gianello RD. 1998. Contents of sugars in leaves of drying desiccation tolerant flowering plants, particularly grasses. *Plant Growth*
656 *Regul* **24**: 185-191.

Gill M, Smith P, Wilkinson JM. 2010. Mitigating climate change: the role of domestic
658 livestock. *Animal* **4**: 323-333.

Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F,
660 Zhang T et al. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**: 463-467.

662 Hoorens PR, Rinaldi M, Li RW, Goddeeris B, Claerebout E, Vercruysse J, Geldhof P. 2011. Genome wide analysis of the bovine mucin genes and their gastrointestinal
664 transcription profile. *BMC Genomics* **12**: 1-12.

Hunter SJ, Easton S, Booth V, Henderson B, Wade WG, Ward JM. 2011. Selective removal
666 of human DNA from metagenomic DNA samples extracted from dental plaque. *J Basic Microbiol* **51**: 442-446.

668 Kaoutari AE, Armougom F, Gordon JI, Raoult D, Henrissat B. 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev*
670 *Microbiol* **11**: 497-504.

Kelly B. The HUNGATE1000. A catalogue of reference genomes from the rumen
672 microbiome. Vol 2016.

Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B,
674 Qin J et al. 2012. MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7**: e47656.

676 Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T et al. 2014. An integrated catalog of reference genes in the human gut
678 microbiome. *Nat Biotech* **32**: 834-841.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010.
680 De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**: 265-272.

682 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

684 Liao JC, Mi L, Pontrelli S, Luo S. 2016. Fuelling the future: microbial engineering for the production of sustainable biofuels. *Nat Rev Microbiol* **14**: 288-304.

686 Liu B, Pop M. 2009. ARDB--Antibiotic Resistance Genes Database. *Nucl Acids Res* **37**: D443-447.

688 Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucl Acids Res* **42**: D490-
690 D495.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu. 1996. A density-based algorithm for discovering
692 clusters in large spatial databases with noise. In *Proceedings of Second International Conference on Knowledge Discovery and Data Mining,* , pp. 226-231, Portland, OR.

694 Morgavi DP, Kelly WJ, Janssen PH, Attwood GT. 2013. Rumen microbial (meta)genomics and its application to ruminant production. *Animal* **7 s1**: 184-201.

696 Ndeh D, Rogowski A, Cartmell A, Luis AS, Baslé A, Gray J, Venditto I, Briggs J, Zhang X, Labourel A et al. 2017. Complex pectin metabolism by gut bacteria reveals novel
698 catalytic functions. *Nature* **544**: 65-70.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, R. B.
700 O'Hara, Simpson GL, Solymos P et al. 2016. vegan: Community Ecology Package. R package version 2.4-1. https://CRAN.R-project.org/package=vegan.

702 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and
704 metagenomes. *Genome Res* **25**: 1043-1055.

Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial
706 marker-gene surveys. *Nat Meth* **10**: 1200-1202.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez
708 F, Yamada T et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59-65.

710 Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**:
712 55-60.

Shepard RN. 1962. The analysis of proximities: Multidimensional scaling with an unknown
714 distance function. I. *Psychometrika* **27**: 125-140.

Storm IM, Kristensen NB, Raun BM, Smedsgaard J, Thrane U. 2010. Dynamics in the
716 microbiology of maize silage during whole-season storage. *Journal of Applied Microbiology* **109**: 1017-1026.

718 Suen G, Weimer PJ, Stevenson DM, Aylward FO, Boyum J, Deneke J, Drinkwater C, Ivanova NN, Mikhailova N, Chertkov O et al. 2011. The complete genome sequence
720 of *Fibrobacter succinogenes* S85 reveals a cellulolytic and metabolic specialist. *PLoS One* **6**.

722 Svartstrom O, Alneberg J, Terrapon N, Lombard V, de Bruijn I, Malmsten J, Dalin A-M, El Muller E, Shah P, Wilmes P et al. 2017. Ninety-nine de novo assembled genomes
724 from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation. *The ISME journal* **11**: 2538-2551.

726 Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, Niazi F, Affourtit J, Egholm M, Henrissat B, Knight R et al. 2010. Organismal, genetic, and transcriptional
728 variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA* **107**: 7503-7508.

730 Vijn I, Smeekens S. 1999. Fructan: more than a reserve carbohydrate? *Plant Physiol* **120**: 351-360.

732 Wallace R, Rooke J, McKain N, Duthie C-A, Hyslop J, Ross D, Waterhouse A, Watson M, Roehe R. 2015. The rumen microbial metagenome associated with high methane
734 production in cattle. *BMC Genomics* **16**: 839.

Weimer PJ, Russell JB, Muck RE. 2009. Lessons from the cow: What the ruminant animal
736 can teach us about consolidated bioprocessing of cellulosic biomass. *Bioresour Technol* **100**: 5323-5331.

738 White BA, Lamed R, Bayer EA, Flint HJ. 2014. Biomass utilization by gut microbiomes. *Annu Rev Microbiol* **68**: 279-296.

740 Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen AØ, Kjeldsen NJ, Liu C et al. 2016. A reference gene catalogue of the pig gut microbiome.
742 *Nat Microbiol* **1**: 16161.

Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, Li X, Long H, Zhang J, Zhang D et al.
744 2015. A catalog of the mouse gut metagenome. *Nat Biotech* **33**: 1103-1108.

Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P. 2011. iPath2. 0: interactive pathway
746 explorer. *Nucl Acids Res* **39**: W412-W415.

Yu Z, Morrison M. 2004. Improved extraction of PCR-quality community DNA from digesta
748 and fecal samples. *Biotechniques* **36**: 808-812.

Zaneveld JR, McMinds R, Vega Thurber R. 2017. Stress and stability: applying the Anna
750 Karenina principle to animal microbiomes. *Nat Microbiol* **2**: 17121.

Zhu W, Lomsadze A, Borodovsky M. 2010. Ab initio gene identification in metagenomic
752 sequences. *Nucl Acids Res* **38**: e132.