MICROBIAL GENOMICS
Bases to Biology

Research Paper

Sequence-based analysis of the genus *Ruminococcus* resolves its phylogeny and reveals strong host association

Alex J. La Reau,[1] Jan P. Meier-Kolthoff[2] and Garret Suen[1]

[1]Department of Bacteriology, University of Wisconsin – Madison, Madison, WI 53706, USA

[2]Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, 38124 Braunschweig, Germany

Correspondence: Garret Suen (gsuen@wisc.edu)

DOI: 10.1099/mgen.0.000099

It has become increasingly clear that the composition of mammalian gut microbial communities is substantially diet driven. These microbiota form intricate mutualisms with their hosts, which have profound implications on overall health. For example, many gut microbes are involved in the conversion of host-ingested dietary polysaccharides into host-usable nutrients. One group of important gut microbial symbionts are bacteria in the genus *Ruminococcus*. Originally isolated from the bovine rumen, ruminococci have been found in numerous mammalian hosts, including other ruminants, and non-ruminants such as horses, pigs and humans. All ruminococci require fermentable carbohydrates for growth, and their substrate preferences appear to be based on the diet of their particular host. Most ruminococci that have been studied are those capable of degrading cellulose, much less is known about non-cellulolytic non-ruminant-associated species, and even less is known about the environmental distribution of ruminococci as a whole. Here, we capitalized on the wealth of publicly available 16S rRNA gene sequences, genomes and large-scale microbiota studies to both resolve the phylogenetic placement of described species in the genus *Ruminococcus*, and further demonstrate that this genus has largely unexplored diversity and a staggering host distribution. We present evidence that ruminococci are predominantly associated with herbivores and omnivores, and our data supports the hypothesis that very few ruminococci are found consistently in non-host-associated environments. This study not only helps to resolve the phylogeny of this important genus, but also provides a framework for understanding its distribution in natural systems.

Keywords: *Ruminococcus*; host range; phylogeny; phylogenomics; 16S rRNA; host diet.

Abbreviations: BME, balanced minimum evolution; dDDH, digital DNA–DNA hybridization; GBDP, genome BLAST distance phylogeny; IMG/M, Integrated Microbial Genomes and Microbiomes; OTU, operational taxonomic unit.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

Data Summary

The 16S rRNA and *recA* gene sequences used for the single and dual-locus phylogenetic analyses were obtained from the National Center for Biotechnology Information and are detailed in Table S1(available in the online Supplementary Material). The single-locus phylogenetic tree of the *recA* gene can be found in Figure S1 (available in the online Supplementary Material). The 56 genomes of *Ruminococcus* and related genera used for the phylogenomic analysis were obtained from the US Department of Energy Joint Genome Institute's Integrated Microbial Genomes and Microbiomes database, and are detailed in Table S2, and the core set of genes from these genomes is detailed in Table S3 (available in the online Supplementary Material). An additional genome was added to this set (*Eubacterium contortum* ATCC 25540) for the genome BLAST distance phylogeny analysis and was obtained from the PATRIC database (genome ID 39482.3; see Data Bibliography). All pairwise digital DNA-DNA hybridization values between genomes used in the phylogenomic analysis are detailed in Table S4 (available in the online Supplementary Material). A genome

Received 8 August 2016; Accepted 22 November 2016

© 2016 The Authors. Published by Microbiology Society

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).

blast distance phylogeny of the nucleotide sequences of these genomes can be found in Figure S2 (available in the online Supplementary Material). The 16S rRNA gene sequences used in the phylogenetic analysis of undescribed ruminococci are detailed in Table S5 (available in the online Supplementary Material). The 47 microbiota studies used for the *Ruminococcus* distributional analyses are detailed in Table S6 (available in the online Supplementary Material; also see Data Bibliography), while the raw proportional data by host is detailed in Table S7. All pairwise comparisons of percent relatedness for Ruminococcus spp. and select related species are detailed in Table S8 (available in the online Supplementary Material).

Introduction

Recent work has demonstrated that the composition of the gut microbiota of mammals is substantially diet driven, with herbivores, omnivores and carnivores harbouring distinct microbial communities (Ley *et al.*, 2008a). These communities are often dominated by bacteria in the phyla *Firmicutes* and *Bacteroidetes*, which are known to form intricate mutualisms with their hosts. Importantly, these bacteria have profound implications for host health (Ley *et al.*, 2008b), such as in humans where they modulate metabolism (Li *et al.*, 2008) and immune system function (Round & Mazmanian, 2009). In other systems like herbivorous ruminants, these microbial communities degrade and ferment dietary cellulosic-based biomass into nutritive short-chain fatty acids (Flint *et al.*, 2008). One important member of both ruminant and human microbial communities is the bacterial genus *Ruminococcus*. For example, the abundance of *Ruminococcus bromii* in humans has been shown to be stimulated by a diet high in resistant starch (Walker *et al.*, 2011). Moreover, some members are now considered as 'keystone' species (Ze *et al.*, 2012; Moraïs *et al.*, 2016), and several occur as prominent members of the 'core gut microbiome' found in a majority of humans (Qin *et al.*, 2010). Aside from their presence in humans, other members are abundant and active in the degradation and fermentation of dietary polysaccharides in ruminant mammals (Leschine, 1995).

Ruminococcus species are defined as strictly anaerobic, Gram-positive, non-motile cocci that do not produce endospores and require fermentable carbohydrates for growth (Rainey, 2009b). They were initially described from the isolation of *Ruminococcus flavefaciens* from the bovine rumen (Sijpesteijn, 1948). *Ruminococcus* is currently considered a polyphyletic genus, with species members belonging to two separate families: the *Ruminococcaceae* and the *Lachnospiraceae* (Rainey & Janssen, 1995). The type species of the genus, *R. flavefaciens*, belongs to the *Ruminococcaceae*. Moreover, several former *Ruminococcus* species have been reclassified to the genus *Blautia* (family *Lachnospiraceae*), based on 16S rRNA gene sequence data (Lawson & Finegold, 2015; Liu *et al.*, 2008), and it has been suggested that only species of the *Ruminococcaceae* be considered as 'true *ruminococci*' (Rainey & Janssen, 1995).

Impact Statement

In this study, we used the wealth of publicly available 16S rRNA, genomic and large-scale microbiota sequence data to address fundamental questions about the diversity, phylogenetic relationships and environmental distribution of the genus *Ruminococcus*. Ruminococci have been studied for decades since their discovery, but very little work has been done to resolve the phylogeny of these bacteria. Furthermore, all described isolates have been obtained from host sources, leading to the hypothesis that the genus is strictly host associated. To this end, we leveraged public sequence databases to generate multiple phylogenies of currently described *Ruminococcus* spp. and resolved the evolutionary relatedness of species members, in addition to identifying clades containing potentially novel isolates. We also used numerous microbiota studies to explore the environmental distribution of *Ruminococcus* sequences to show that this genus is highly diverse, and that novel species likely exist in disparate host environments. Finally, we provide strong evidence that *Ruminococcus* is a strictly host-associated genus, due to its virtual absence in all environmental datasets considered.

Under this definition, there are only six described species of *Ruminococcus* to date. Some species are cellulolytic, including the rumen isolates *R. flavefaciens* and *Ruminococcus albus* (Hungate, 1957), and the recently described human isolate *Ruminococcus champanellensis* (Chassard *et al.*, 2012), which is the only known bacterial species isolated from the human colon capable of degrading crystalline cellulose (Moraïs *et al.*, 2016). Others are non-cellulolytic and utilize polysaccharides like resistant starches in the case of *R. bromii* (Ze *et al.*, 2012), or selectively use various plant hemicelluloses in the case of *Ruminococcus callidus* (Lay *et al.*, 2005) and '*Ruminococcus bicirculans*' (Wegmann *et al.*, 2014). Substantial work has been done on cellulolytic *Ruminococcus* isolates from the bovine rumen due to their potential application in biofuels and their importance in animal health (Dassa *et al.*, 2014; Christopherson *et al.*, 2014; Pavlostathis *et al.*, 1988). Strains from other ruminant (Orpin *et al.*, 1985; Krause *et al.*, 1999) and non-ruminant (Julliand *et al.*, 1999) sources have also been described, but no isolate has been reported from a non-host (i.e. environmental) source to date. Presently, much less is known about the non-cellulolytic, non-ruminant, host-associated ruminococci isolates.

Here, we used publicly available 16S rRNA sequences, genomes and microbiota data to demonstrate that the genus *Ruminococcus* has unexplored diversity and a broad host distribution. Our phylogenomic analysis of the genomes of *Ruminococcus* spp. and several related taxa confirms the polyphyletic nature of the genus, with ruminococci falling into distinct, distantly related clades (i.e. the *Ruminococcaceae* and *Lachnospiraceae*). We also present evidence that *Ruminococcus* species are predominantly associated with herbivores and omnivores, relative to carnivores, and that significantly abundant *Ruminococcus* populations are absent in non-host-associated environments.

Methods

Dual-locus phylogenetic analysis of described *Ruminococcus* spp. Full-length or near full-length sequences of two highly conserved genes – 16S rRNA and *recA* – were obtained for each formally described *Ruminococcus* type species (both *Ruminococcaceae* and *Lachnospiraceae*), other related genera (see Table S1 for details) and the outgroup species *Eubacterium acidaminophilum* from GenBank. In the case that both a 16S rRNA and *recA* gene sequence were not available for a type strain, another strain of the same species was used. Gene sequences for *recA* were obtained from genome data for each strain when they were not available as stand-alone sequences in GenBank.

Full-length and near full-length sequences were grouped into three libraries: 16S rRNA, *recA*, and a combined 16S rRNA and *recA* library. The single-locus libraries were imported into MEGA6 (Tamura *et al.*, 2013), aligned using CLUSTALW with default parameters and trimmed to 1282 and 959 bp for the 16S rRNA and *recA* genes, respectively. A dual-locus library was created by concatenating the two alignments into a 16S rRNA–*recA* alignment. All alignments were then exported into MrBayes (v3.2.3) (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) and Bayesian phylogenetic analyses were performed (ngen=10 000 000) on each library. The resulting trees were visualized using FigTree v1.4.2 (A. Rambaut; http://tree.bio.ed.ac.uk/software/figtree).

Phylogenomic analyses of ruminococci and related genera. The available genome sequences for bacteria characterized as ruminococci, as well as for several related non-ruminococci (Table S2), were analysed using the US Department of Energy Joint Genome Institute's Integrated Microbial Genomes and Microbiomes (IMG/M) database (Markowitz *et al.*, 2012). All predicted genes in the genome sequence for *R. albus* 7 (Suen *et al.*, 2011) (IMG/M Genome ID: 649633094) were used as a reference to conduct an analysis using the Phylogenetic Profiler for Single Genes tool against all other genomes, using default settings, to generate a list of orthologues shared between these genomes (Table S3). All nucleotide sequences for orthologues shared between each genome were obtained from the IMG/M database, concatenated and aligned using MAFFT version 7 (Katoh *et al.*, 2002; Katoh & Standley, 2013). A Bayesian phylogeny was then generated from this alignment using MrBayes (ngen=100 000).

A whole-genome-sequence-based phylogenomic analysis was conducted for those genomes obtained from the IMG/M database at the nucleotide level using the genome BLAST distance phylogeny (GBDP) method (Henz *et al.*, 2005; Meier-Kolthoff *et al.*, 2013a), including the inference of branch support (Meier-Kolthoff *et al.*, 2014). BLAST+ (Camacho *et al.*, 2009) was used as local alignment tool with default settings, and subsequent calculations of intergenomic distances were carried out with an *e*-value filter of 10^{-8}, the trimming algorithm and formula d_5 (Meier-Kolthoff *et al.*, 2014). The same settings were used for a complementary GBDP analysis of the entire sets of genes at the amino acid level, as conducted in an earlier study (Lagkouvardos *et al.*, 2016). ORF calling was carried out using the gene finding program Prodigal (Hyatt *et al.*, 2010). All balanced minimum evolution (BME) trees were reconstructed via FASTME 2.1.4 with SPR postprocessing (Lefort *et al.*, 2015). To further infer potential affiliation to the same species, all pairwise digital DNA–DNA hybridization (dDDH) values and their confidence intervals were calculated with the Genome-to-Genome Distance Calculator (GGDC 2.1; freely available at http://ggdc.dsmz.de) (Meier-Kolthoff *et al.*, 2013a) under the National Center for Biotechnology Information (NCBI)-BLAST setting (Table S4).

Phylogenetic analysis of undescribed ruminococci. A search was performed through the NCBI nucleotide sequence database using the search terms, '*Ruminococcus* [ORGANISM] AND 16S [TEXT WORD] AND 250:1600 [SEQUENCE LENGTH]' (on 2/3/2015) to obtain all 16S rRNA gene sequences classified to the genus. Sequences representing contigs from whole-genome shotgun projects were removed to reduce redundancy, leaving a final *Ruminococcus* sequence library of 345 sequences. The closelyrelated *Eubacterium acidaminophilum* was used as an outgroup. The full sequence dataset is presented in Table S5.

This library was then processed in mothur (v.1.35.1) (Schloss *et al.*, 2009) using the following commands (indicated in italics) with default parameters except where indicated. Sequences were aligned (*align.seqs, flip=t*) to the Silva 16S/18S rRNA non-redundant sequence database (SSU ref NR; release 119; 534 968 total sequences). Sequences ≥900 bp in length (*screen.seqs*) were retained, followed by removal of duplicate sequences (*unique.seqs*), leaving 160 total sequences. Aligned sequences were then filtered (*filter.seqs*) and trimmed to 807 bp in length. A distance matrix was created (*dist.seqs*), and used to estimate the number of operational taxonomic units (OTUs) at various percentage-similarity cut-offs (*cluster, method=furthest*). A representative sequence for each OTU was chosen (*get.oturep*) and used to construct a Bayesian tree of OTUs at 97 % similarity in MrBayes (ngen=10 000 000).

Mining of published datasets for *Ruminococcus* 16S rRNA gene sequences. A broad survey of 16S rRNA microbiota sequencing studies encompassing host-associated (various animal and plant hosts) and non-host-associated (marine, freshwater and soil) environments was performed. These studies varied in sequencing methodology (i.e. 16S rRNA clone libraries, 454 pyrosequencing and

Illumina platforms), and are detailed in Table S6. Raw sequences were obtained for each study and imported into separate sequence libraries. Each library was then processed in mothur by first aligning the sequences as described above, followed by a taxonomic classification (*classify.seqs*) using default parameters and the Silva reference taxonomy provided through mothur. A classification to the genus *Ruminococcus* (of family *Ruminococcaceae*) for a sequence was considered positive only for bootstrap values ≥80 (Wang *et al.*, 2007). Since the studies varied in sequencing methodologies, we could not compare relative abundances across sequence sources, but rather we determined the overall trends of distribution using proportional data for the presence of *Ruminococcus* sequences for each dataset (total *Ruminococcus* sequences/total sequences in dataset). These data are detailed in Table S7.

Results

Resolved phylogeny of the ruminococci

In order to obtain a highly resolved view of the relationship of all currently described '*Ruminococcus*' species, we created four separate phylogenetic trees at varying levels of genomic resolution (Fig. 1) including: (a) a Bayesian phylogenetic tree of 16S rRNA gene sequences, (b) a dual-locus Bayesian phylogenetic tree using the 16S rRNA and *recA* gene sequences, (c) a phylogenomic Bayesian tree using 275 orthologues shared amongst 41 *Ruminococcus* genomes and 15 genomes of taxa from closely related genera, and (d) a BME GBDP analysis of the entire set of genes at the amino acid level for all genomes. All four trees showed confirmation of the reported split between *Ruminococcaceae* (Rainey, 2009b) and *Lachnospiraceae* (Rainey, 2009a; Cornick & Stanton, 2009) species in the genus *Ruminococcus* (Rainey & Janssen, 1995) with high confidence (i.e. high posterior probability values at each node for Bayesian trees and high branch support values for the BME tree with a mean of 93.7 %). These findings were also separately confirmed for the *recA* single-locus tree (Fig. S1). Moreover, these trees detailed the phylogenetic relationship between all described *Ruminococcus* species within both families. For example, the true ruminococci all form a monophyletic group within the *Ruminococcaceae* in all of our trees with the exception of *R. bromii* strains, which are deeply rooted and form a clade with more closely related *Clostridium* spp. Furthermore, the ruminococci that were previously reclassified to the genus *Blautia* all form a monophyletic clade with each other, *Ruminococcus gauvreauii* and several other undescribed *Ruminococcus* spp. (Fig. 1b–d), although this is not as resolved at the 16S rRNA level alone (Fig. 1a). All trees showed nearly identical topologies for all taxa examined. However, minor topological changes were seen for the GBDP nucleotide analysis (Fig. S2) in which some *Ruminococcaceae* taxa were found to group in the *Lachnospiraceae* clade, though branch support values for these discrepancies were low.

Our phylogenomic analysis revealed that the 275 orthologues shared between all ruminococci genomes were mostly housekeeping genes (e.g. ribosomal proteins, tRNA synthetases) and genes involved in other highly conserved cell functions (e.g. chaperones, amino acid biosynthesis, ABC transporters). Among these orthologues, two were known carbohydrate-active enzymes (CAZymes) annotated as being involved in starch phosphorylation (Table S3): a glycogen/starch/alpha-glucan phosphorylase (*Rumal_0466*) and a maltodextrin phosphorylase (*Rumal_2782*).

Finally, we used pairwise dDDH values to estimate the number of discrete species amongst the genomes we analysed (Table S4). Of the 41 *Ruminococcus* genomes examined, we found 35 separate clades (27 in *Ruminococcaceae* and 8 in *Lachnospiraceae*) that likely represent distinct species. Among the 27 in the *Ruminococcaceae*, 4 had no current species representatives. Conversely, the *R. flavefaciens* strains separated into 13 potential species and *R. albus* separated into 5. Both *R. bromii* strains appear to delineate into two different species using this analysis.

Expanded phylogenetic analysis describes the diversity of the ruminococci

With an established reference phylogeny, we sought to further probe the unexplored diversity of the genus beyond the sequenced isolates. To accomplish this, we performed a search of all 16S rRNA gene sequences available in GenBank designated as belonging to the genus *Ruminococcus*, and generated a library of 345 total sequences (Table S5), 160 of which passed our filtering criteria. Due to the nomenclature discrepancies between families in this group, this dataset contains *Ruminococcus* sequences from both the *Ruminococcaceae* and the *Lachnospiraceae*. We evaluated this sequence library using an OTU analysis, which defines phylogenetic relationships at different taxonomic levels based on the percentages of sequence identity independent of sequence counts. Our OTU analysis, as performed in mothur (Schloss *et al.*, 2009), revealed that the described species (see Fig. 1a) fell into distinct clades at 97 % 16S rRNA identity (Fig. 2). At this level, we identified 44 OTUs, 20 of which belonged to the family *Ruminococcaceae* (Table 1). Of these, eight represented potentially novel OTUs from different host sources, as well as one apparent non-host sequence from an Antarctic intertidal sediment (Yu *et al.*, 2010). As seen in our protein-based GBDP analysis (Fig. 1d), *R. albus* and *R. flavefaciens* had multiple strains that did not group into a shared OTU at a 97 % sequence similarity level.

Our analysis also yielded insights into *Ruminococcus* spp. host specificity. For example, the OTU containing *R. flavefaciens* C94^T contained only sequences from multiple ruminants (Table 1). An identical trend was observed for the other *R. flavefaciens* strains, as well as for *R. albus* 7. Conversely, those OTUs containing *R. champanellensis*, *R. callidus* and '*R. bicirculans*' contained only sequences from human sources. Finally, the OTU containing *R. bromii* ATCC 27255 showed a broader host range, including both humans, pigs and bovines.

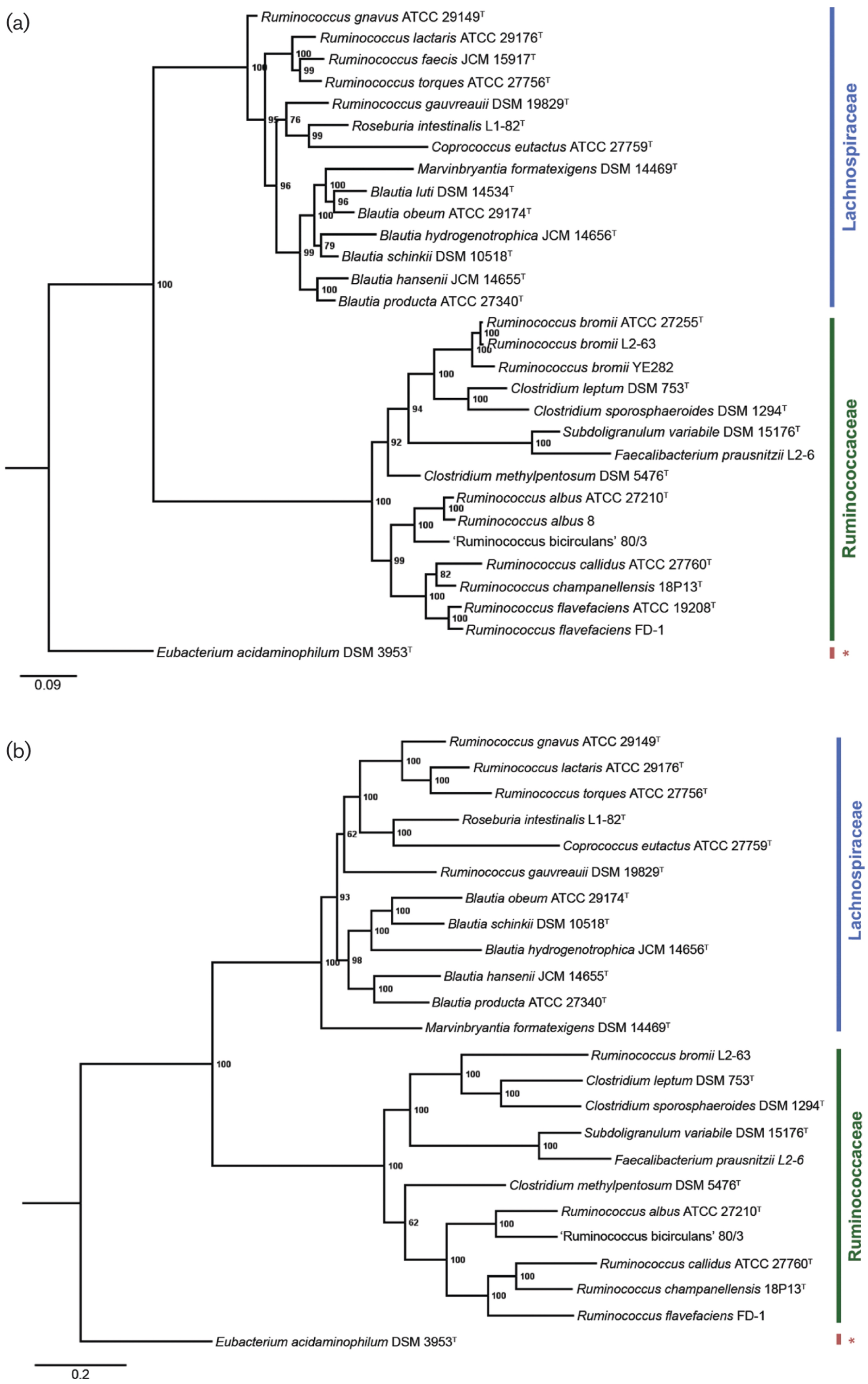
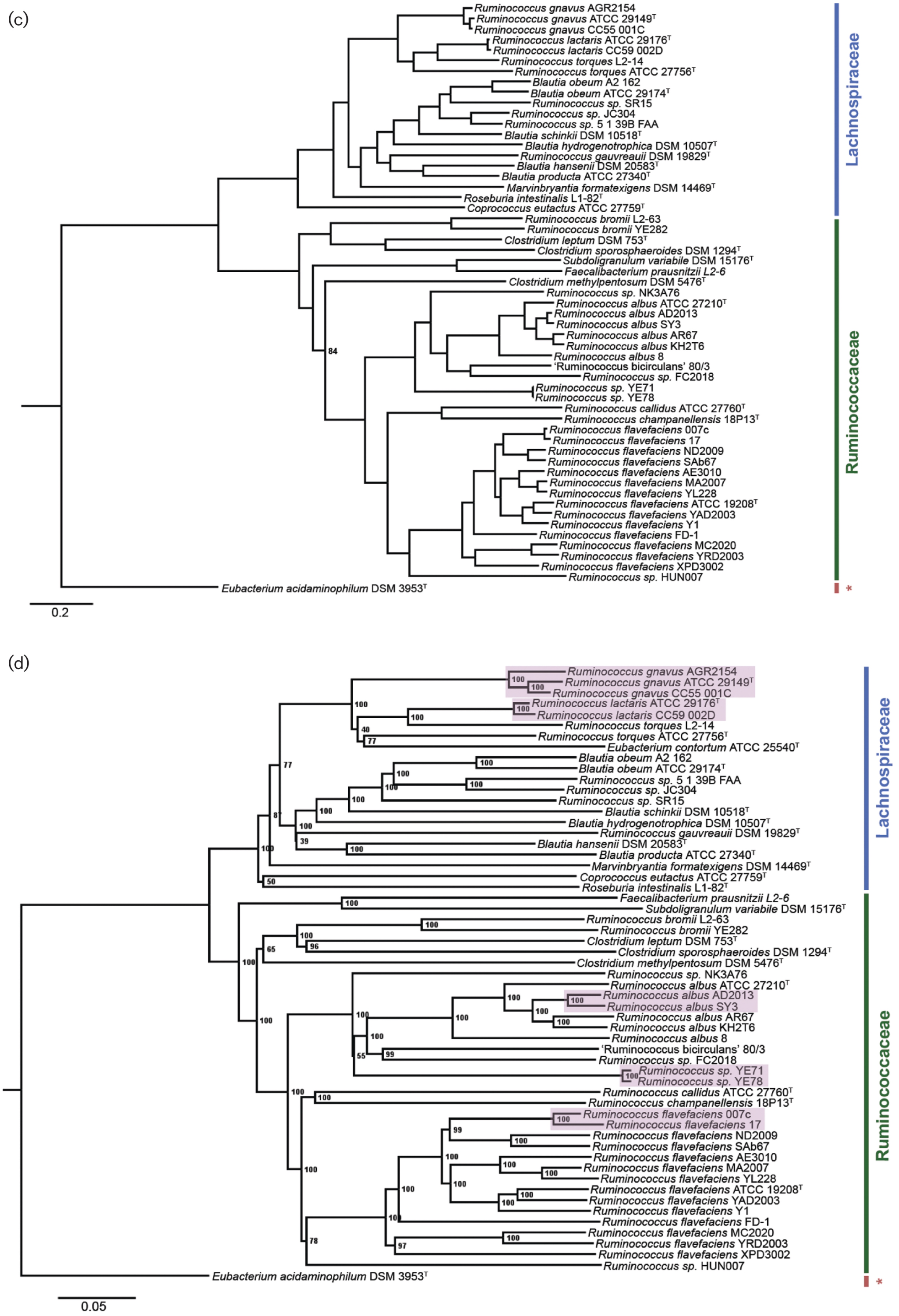

(a)
Ruminococcus gnavus ATCC 29149T
Ruminococcus lactaris ATCC 29176T
Ruminococcus faecis JCM 15917T
Ruminococcus torques ATCC 27756T
Ruminococcus gauvreauii DSM 19829T
Roseburia intestinalis L1-82T
Coprococcus eutactus ATCC 27759T
Marvinbryantia formatexigens DSM 14469T
Blautia luti DSM 14534T
Blautia obeum ATCC 29174T
Blautia hydrogenotrophica JCM 14656T
Blautia schinkii DSM 10518T
Blautia hansenii JCM 14655T
Blautia producta ATCC 27340T
Ruminococcus bromii ATCC 27255T
Ruminococcus bromii L2-63
Ruminococcus bromii YE282
Clostridium leptum DSM 753T
Clostridium sporosphaeroides DSM 1294T
Subdoligranulum variabile DSM 15176T
Faecalibacterium prausnitzii L2-6
Clostridium methylpentosum DSM 5476T
Ruminococcus albus ATCC 27210T
Ruminococcus albus 8
'Ruminococcus bicirculans' 80/3
Ruminococcus callidus ATCC 27760T
Ruminococcus champanellensis 18P13T
Ruminococcus flavefaciens ATCC 19208T
Ruminococcus flavefaciens FD-1
Eubacterium acidaminophilum DSM 3953T
100
99
100
95
76
99
96
100
96
100
79
99
100
100
100
100
100
100
94
92
100
100
100
99
82
100
100
Lachnospiraceae
Ruminococcaceae
*
0.09
(b)
Ruminococcus gnavus ATCC 29149T
Ruminococcus lactaris ATCC 29176T
Ruminococcus torques ATCC 27756T
Roseburia intestinalis L1-82T
Coprococcus eutactus ATCC 27759T
Ruminococcus gauvreauii DSM 19829T
Blautia obeum ATCC 29174T
Blautia schinkii DSM 10518T
Blautia hydrogenotrophica JCM 14656T
Blautia hansenii JCM 14655T
Blautia producta ATCC 27340T
Marvinbryantia formatexigens DSM 14469T
Ruminococcus bromii L2-63
Clostridium leptum DSM 753T
Clostridium sporosphaeroides DSM 1294T
Subdoligranulum variabile DSM 15176T
Faecalibacterium prausnitzii L2-6
Clostridium methylpentosum DSM 5476T
Ruminococcus albus ATCC 27210T
'Ruminococcus bicirculans' 80/3
Ruminococcus callidus ATCC 27760T
Ruminococcus champanellensis 18P13T
Ruminococcus flavefaciens FD-1
Eubacterium acidaminophilum DSM 3953T
100
100
100
62
100
93
100
100
100
98
100
100
100
100
100
100
100
62
100
100
100
100
100
Lachnospiraceae
Ruminococcaceae
*
0.2

Fig. 1. (a) A rooted Bayesian phylogeny (ngen=10 000 000) of 16S rRNA gene sequences of *Ruminococcus* spp. and related genera. (b) A rooted Bayesian phylogeny (ngen=10 000 000) of concatenated 16S rRNA and *recA* gene sequences for *Ruminococcus* spp. and related genera. (c) A rooted multi-locus Bayesian phylogeny (ngen=100 000) of 275 concatenated orthologous genes (see Table S7) shared among 56 genomes of *Ruminococcus* and closely related species available in the Integrated Microbial Genomes database. (d) A BME tree inferred by phylogenomic GBDP analysis at the amino acid level for coding genes in genomes of *Ruminococcus* spp. and related genera. All posterior probability values <100 are shown on Bayesian phylogenies a–c, and all branch support values are shown for the GBDP analysis (mean support of 93.7 % for the full tree). All type strains are indicated with a superscript T. The red asterisks indicates family *Eubacteriaceae* for *E. acidaminophilum*, the outgroup in all trees presented. The purple highlighting indicates taxa inferred as the same species according to the GBDP analysis.

Fig. 1. (cont.)


(c)
Ruminococcus gnavus AGR2154
Ruminococcus gnavus ATCC 29149T
Ruminococcus gnavus CC55 001C
Ruminococcus lactaris ATCC 29176T
Ruminococcus lactaris CC59 002D
Ruminococcus torques L2-14
Ruminococcus torques ATCC 27756T
Blautia obeum A2 162
Blautia obeum ATCC 29174T
Ruminococcus sp. SR15
Ruminococcus sp. JC304
Ruminococcus sp. 5 1 39B FAA
Blautia schinkii DSM 10518T
Blautia hydrogenotrophica DSM 10507T
Ruminococcus gauvreauii DSM 19829T
Blautia hansenii DSM 20583T
Blautia producta ATCC 27340T
Marvinbryantia formatexigens DSM 14469T
Roseburia intestinalis L1-82T
Coprococcus eutactus ATCC 27759T
Ruminococcus bromii L2-63
Ruminococcus bromii YE282
Clostridium leptum DSM 753T
Clostridium sporosphaeroides DSM 1294T
Subdoligranulum variabile DSM 15176T
Faecalibacterium prausnitzii L2-6
Clostridium methylpentosum DSM 5476T
Ruminococcus sp. NK3A76
Ruminococcus albus ATCC 27210T
Ruminococcus albus AD2013
Ruminococcus albus SY3
Ruminococcus albus AR67
Ruminococcus albus KH2T6
Ruminococcus albus 8
84
'Ruminococcus bicirculans' 80/3
Ruminococcus sp. FC2018
Ruminococcus sp. YE71
Ruminococcus sp. YE78
Ruminococcus callidus ATCC 27760T
Ruminococcus champanellensis 18P13T
Ruminococcus flavefaciens 007c
Ruminococcus flavefaciens 17
Ruminococcus flavefaciens ND2009
Ruminococcus flavefaciens SAb67
Ruminococcus flavefaciens AE3010
Ruminococcus flavefaciens MA2007
Ruminococcus flavefaciens YL228
Ruminococcus flavefaciens ATCC 19208T
Ruminococcus flavefaciens YAD2003
Ruminococcus flavefaciens Y1
Ruminococcus flavefaciens FD-1
Ruminococcus flavefaciens MC2020
Ruminococcus flavefaciens YRD2003
Ruminococcus flavefaciens XPD3002
Ruminococcus sp. HUN007
Eubacterium acidaminophilum DSM 3953T
0.2
Lachnospiraceae
Ruminococcaceae
*
(d)
Ruminococcus gnavus AGR2154
Ruminococcus gnavus ATCC 29149T
Ruminococcus gnavus CC55 001C
Ruminococcus lactaris ATCC 29176T
Ruminococcus lactaris CC59 002D
Ruminococcus torques L2-14
Ruminococcus torques ATCC 27756T
Eubacterium contortum ATCC 25540T
Blautia obeum A2 162
Blautia obeum ATCC 29174T
Ruminococcus sp. 5 1 39B FAA
Ruminococcus sp. JC304
Ruminococcus sp. SR15
Blautia schinkii DSM 10518T
Blautia hydrogenotrophica DSM 10507T
Ruminococcus gauvreauii DSM 19829T
Blautia hansenii DSM 20583T
Blautia producta ATCC 27340T
Marvinbryantia formatexigens DSM 14469T
Coprococcus eutactus ATCC 27759T
Roseburia intestinalis L1-82T
Faecalibacterium prausnitzii L2-6
Subdoligranulum variabile DSM 15176T
Ruminococcus bromii L2-63
Ruminococcus bromii YE282
Clostridium leptum DSM 753T
Clostridium sporosphaeroides DSM 1294T
Clostridium methylpentosum DSM 5476T
Ruminococcus sp. NK3A76
Ruminococcus albus ATCC 27210T
Ruminococcus albus AD2013
Ruminococcus albus SY3
Ruminococcus albus AR67
Ruminococcus albus KH2T6
Ruminococcus albus 8
'Ruminococcus bicirculans' 80/3
Ruminococcus sp. FC2018
Ruminococcus sp. YE71
Ruminococcus sp. YE78
Ruminococcus callidus ATCC 27760T
Ruminococcus champanellensis 18P13T
Ruminococcus flavefaciens 007c
Ruminococcus flavefaciens 17
Ruminococcus flavefaciens ND2009
Ruminococcus flavefaciens SAb67
Ruminococcus flavefaciens AE3010
Ruminococcus flavefaciens MA2007
Ruminococcus flavefaciens YL228
Ruminococcus flavefaciens ATCC 19208T
Ruminococcus flavefaciens YAD2003
Ruminococcus flavefaciens Y1
Ruminococcus flavefaciens FD-1
Ruminococcus flavefaciens MC2020
Ruminococcus flavefaciens YRD2003
Ruminococcus flavefaciens XPD3002
Ruminococcus sp. HUN007
Eubacterium acidaminophilum DSM 3953T
100
100
100
100
100
40
77
77
100
100
100
100
100
100
87
100
39
100
100
50
100
100
100
65
96
100
100
100
100
100
100
100
100
55
99
100
100
100
100
100
99
100
100
100
100
100
100
100
100
100
78
97
100
0.05
Lachnospiraceae
Ruminococcaceae
*

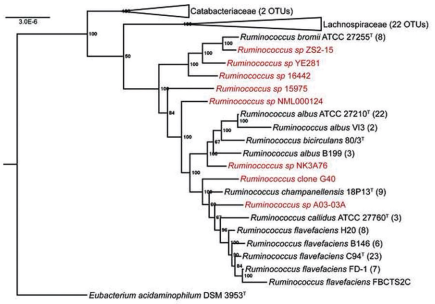

3.0E-6
Catabacteriaceae (2 OTUs)
Lachnospiraceae (22 OTUs)
Ruminococcus bromii ATCC 27255^T (8)
Ruminococcus sp ZS2-15
Ruminococcus sp YE281
Ruminococcus sp 16442
Ruminococcus sp 15975
Ruminococcus sp NML000124
Ruminococcus albus ATCC 27210^T (22)
Ruminococcus albus VI3 (2)
Ruminococcus bicirculans 80/3^T
Ruminococcus albus B199 (3)
Ruminococcus sp NK3A76
Ruminococcus clone G40
Ruminococcus champanellensis 18P13^T (9)
Ruminococcus sp A03-03A
Ruminococcus callidus ATCC 27760^T (3)
Ruminococcus flavefaciens H20 (8)
Ruminococcus flavefaciens B146 (6)
Ruminococcus flavefaciens C94^T (23)
Ruminococcus flavefaciens FD-1 (7)
Ruminococcus flavefaciens FBCTS2C
Eubacterium acidaminophilum DSM 3953^T
100
50
100
100
100
100
84
100
100
67
100
100
96
100
84
100

Fig. 2. A Bayesian phylogeny of representative 16S rRNA OTUs at 97 % sequence similarity with *E. acidaminophilum* used as an outgroup. All posterior probability values are shown (ngen=10 000 000). Numbers in parentheses indicate the number of sequences that clustered into the OTU shown. Clades for *Ruminococcus* sequences that did not fall in the family *Ruminococcaceae* have been collapsed. Sequence names are coloured in black (strains of described species) and red (uncultured or undescribed strains).

Environmental distribution analysis of the ruminococci

To understand the overall environmental distribution of *Ruminococcus* spp. in greater detail, we leveraged large-scale 16S rRNA-based microbiota studies from various host and non-host environments. Specifically, we aligned sequences for each study to the Silva 16S/18S NR database (SSU ref NR; release 119; 534 968 total sequences) in mothur to obtain a reference taxonomy. We examined a total of 47 studies, encompassing various types of sequencing technologies (e.g. 16S rRNA Sanger clone library, 454 pyrosequencing, Illumina) and different regions of the 16S rRNA gene (Table S6). In total, our dataset included 134 animal hosts (55 herbivores, 47 omnivores and 32 carnivores), 63 plant hosts [57 of which were from Kembel *et al.* (2014)] and 18 environmental sources (including soil, freshwater and marine environments).

From our analysis, we found that herbivores showed the highest representation of *Ruminococcus* sequences (39/55 animals examined; Fig. 3). This included expected sources such as ruminants, but also other herbivorous hosts such as avians (hoatzin), reptiles (gopher tortoise), primates (colobus monkeys, gorillas, orangutan) and other non-ruminant mammals (e.g. rhinoceros, red kangaroo, European rabbit, Linnaeus' two-toed sloth). Omnivores also showed substantial numbers of *Ruminococcus* sequences (18/33 animals examined; Fig. 3) including humans, flying fox, ostrich and a large number of non-human primates (e.g. chimpanzee, bonobo, spider monkey and two lemur species). The carnivores also contained

Table 1. Identified 97 % similarity OTUs of GenBank *Ruminococcus* 16s rRNA gene sequences (*Ruminococcaceae* only)

Representative sequence ordered by sequence source (ruminant herbivores, non-herbivores and mixed sources). Names in black represent OTUs that clustered with described *Ruminococcus* spp. strains. Names in red represent OTUs that did not cluster with described *Ruminococcus* spp. strains.

Representative sequence name (NCBI accession no.)	Isolation source(s)	No. of sequences
Herbivores		
Ruminococcus flavefaciens C94^T (AM915269.1)	Cow, goat, reindeer, moose, sheep	23
Ruminococcus albus ATCC 27210^T (AB538438.1)	Cow, moose, sheep, golden takin	22
Ruminococcus flavefaciens H20 (JF970204.1)	Cow, goat, moose	8
Ruminococcus flavefaciens FD-1 (AM920691.1)	Cow, sheep	7
Ruminococcus flavefaciens B146 (AY445599.1)	Cow, addax	6
Ruminococcus albus B199 (AY445592.1)	Cow, sheep	3
Ruminococcus albus VI3 (HQ404370.1)	Sheep	2
Ruminococcus flavefaciens FBCTS2C (EU445111.1)	Cow	1
Ruminococcus sp. NK3A76 (GU324399.1)	Sheep	1
Ruminococcus sp. YE281 (DQ882650.1)	Cow	1
Ruminococcus clone G40 (JN008429.1)	Goat	1
Non-Herbivores		
Ruminococcus champanellensis 18 P13^T (AB910742.1)	Human	9
Ruminococcus callidus ATCC 27760^T (L76596.1)	Human	3
Ruminococcus sp. ZS2-15 (FJ889653.1)	Antarctic sandy intertidal sediment	1
Ruminococcus bicirculans 80/3^T (HF545617.1)	Human	1
Ruminococcus sp. 16442 (AJ318889.1)	Human	1
Ruminococcus sp. A03-03A (FJ542832.1)	Earthworm	1
Ruminococcus sp. NML000124 (EU815223.1)	Human (blood culture)	1
Ruminococcus sp. 15975 (AJ308104.1)	Human	1
Mixed sources		
Ruminococcus bromii ATCC 27255^T (NR025930.1)	Human, cow, pig	8

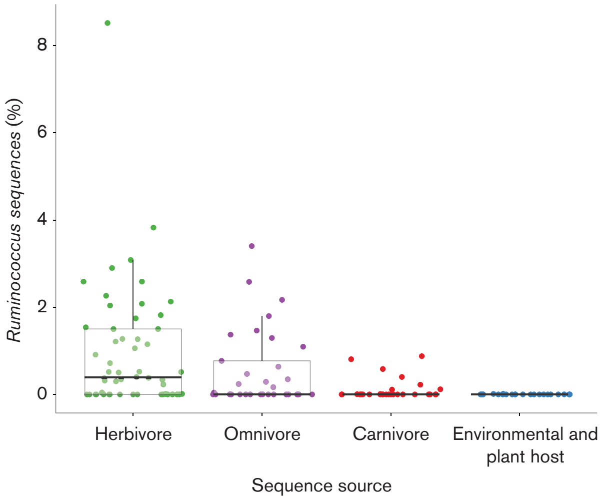


Fig. 3. *Ruminococcus* 16S rRNA gene sequence distribution in various sequencing datasets categorized into four sample types. Each point represents the proportion of *Ruminococcus* sequences in a given dataset. Box and whisker plots are overlaid, showing the median (black bar), 75th percentile (upper hinge) and 1.5 × interquartile range (upper whisker).

Ruminococcus sequences, although at lower percentages relative to herbivores and omnivores (11/32 animals examined; Fig. 3). One exception to this was for insectivorous/myrmecophagous mammals, some of which had sequence proportions similar to herbivores and omnivores, including the aardvark (0.42 %), the nine-banded armadillo (1.21 %) and the southern tamandua (0.61 %). Finally, we found that very few of the environmental sources contained sequences that classified to *Ruminococcus* (Fig. 3). The highest percentage of *Ruminococcus* sequences observed was 0.16 % for snow samples from a Greenland Ice Sheet (Table S7). In general, *Ruminococcus* sequence proportions were found to be significantly enriched in host-associated samples, relative to the environmental samples (Fig. 3; Fisher's exact test, *P* value $<2.2\times10^{-16}$).

Discussion

In this study, we used detailed single-gene-sequence-based and whole-genome-sequence-based phylogenetic analyses to assess the diversity, host specificity and environmental distribution of *Ruminococcus* spp. Having first been isolated from the bovine rumen, this genus now consists of several isolates originating from other herbivorous and omnivorous sources, leading to the hypothesis that members of this genus are strictly host associated. Although much work has been performed on the cellulolytic members of this genus to elucidate their potential roles in their given hosts, little is known about the broad distribution of this genus and the phenotypic diversity that may be present between its members.

Taken together, our phylogenetic analyses resulted in a highly resolved phylogeny that confirmed the polyphyletic nature of the genus (Fig. 1a–d). Given that several former *Ruminococcus* spp. have now been reclassified to other genera such as *Blautia* (Liu *et al.*, 2008; Lawson & Finegold, 2015), these data reinforce the need for future reclassifications of many species members (i.e. the *Ruminococcus* species within the *Lachnospiraceae*) to avoid confusion in the literature. The topology of our phylogenetic analyses also matches partial phylogenetic analyses performed for recently described species (Chassard *et al.*, 2012; Wegmann *et al.*, 2014) and offers some insights. For example, *R. bromii* appears less related to all other true ruminococci, as determined in our pairwise 16S rRNA similarity comparisons (88–89 % 16S identity; Table S8). Furthermore, it is more closely related to *Clostridium* spp., providing further evidence for the hypothesis that these organisms represent an entirely separate genus of their own (Rainey & Janssen, 1995). One possibility for this finding is that *R. bromii* may have diverged from other *Ruminococcaceae* early on, and has since become specialized. This is supported by its affinity for resistant starch substrates (Ze *et al.*, 2012) and the reduced number of CAZymes encoded by its genome, relative to other *Ruminococcus* spp., with apparent exclusivity for amylases (Ze *et al.*, 2015). Moreover, the only CAZymes shared amongst all genome sequences in our phylogenomic analysis were two genes involved in starch phosphorylation (Table S3), suggesting that the ability to utilize non-resistant dietary starches may be an ancient trait possessed by an ancestor of both *Lachnospiraceae* and *Ruminococcaceae*.

Among the other five species, two distinct clades exist: one containing *R. albus* and '*R. bicirculans*', and the other containing *R. callidus*, *R. champanellensis* and *R. flavefaciens* (Fig. 1). *R. albus* and *R. flavefaciens* are distantly related (91 % 16S rRNA sequence similarity; Table S8) and coexist as important members of the rumen cellulolytic community. Previous work has shown that they possess markedly different cellulolytic strategies (Christopherson *et al.*, 2014; Dassa *et al.*, 2014) and likely compete for access to cellulosic biomass in the same niche, as *R. albus* is known to use bacteriocins to inhibit *R. flavefaciens* (Chen *et al.*, 2004). This may indicate that both species were independently acquired by ruminants, thereby reflecting their phylogenetic placement, and that the inhibitory use of bacteriocins by *R. albus* (Shi *et al.*, 1997) may serve to reinforce their distinct lineages by providing a competitive advantage within the rumen ecosystem.

In contrast, *R. callidus* and *R. champanellensis* cluster together, are more closely related (95 % 16S rRNA sequence similarity; Table S8) and are both found in the human colon. These species have different substrate preferences (Lay *et al.*, 2005; Chassard *et al.*, 2012), likely reflecting divergence and differential specialization from a common ancestor within the same host environment. One explanation for their close phylogenetic relatedness is that their divergence is more recent, and that the diversity of the human diet may have rapidly contributed to shifts in their substrate preference. Future growth experiments should be performed to assess the potential cooperation and competition between these two species.

Finally, '*R. bicirculans*' appears most closely related to *R. albus* at 94 % 16S rRNA sequence similarity (Table S8). Both species inhabit different hosts(Rainey, 2009b; Wegmann *et al.*, 2014), and '*R. bicirculans*' is unique from other ruminococci in that its genome consists of two chromosomes thought to be the result of a recombination event between two rRNA operons (Wegmann *et al.*, 2014). Moreover, both species prefer different but overlapping substrates, with '*R. bicirculans*' selectively utilizing certain hemicelluloses but not cellulose or arabinoxylan (Wegmann *et al.*, 2014), and *R. albus* capable of utilizing a wide range of substrates, including cellulose and xylan (Christopherson *et al.*, 2014). Taken together, these observations may explain why these two species form a sister clade to the other ruminococci, as they likely diverged and became specialized within their given hosts (i.e. ruminants vs humans).

The wealth of publicly available genomes and 16S rRNA gene sequences allowed us to further explore the diversity of this genus beyond its formally described species members. Both our GBDP and OTU analyses revealed unrealized diversity at the whole-genome level as demonstrated by the dDDH values and 97 % 16S sequence similarity level, respectively. In particular, both provide evidence that many novel species of *Ruminococcus* (both cultured and uncultured) likely exist beyond those that have been described to date. For example, *R. flavefaciens* strains alone accounted for 5 OTUs at 97 % 16S rRNA sequence similarity (Fig. 2, Table 2) and 13 potential separate species using dDDH (Fig. 1d, Table S4), suggesting that *R. flavefaciens* is an extraordinarily diverse species, and that some strains may even represent separate species-level lineages. Indeed, a similar result was reported in a previous study comparing the diversity of *R. flavefaciens* strains based on *scaC* gene sequences (Jindou *et al.*, 2008). Similarly, *R. albus* accounted for three separate OTUs at 97 % sequence similarity (Fig. 2, Table 2) and five potential species using dDDH (Fig. 1d, Table S4). *R. flavefaciens* OTUs did not collapse into a single OTU until 95 % 16S rRNA sequence similarity, and *R. albus* did not do so until 92 % similarity. Furthermore, the entire genus did not fall into a single OTU until 86 % sequence similarity. Given the nature of the sequence data in our OTU analysis (i.e. a number of non-full-length 16S sequences), we chose a more conservative 97 % 16S similarity threshold for species delineation (Stackebrandt & Goebel, 1994). However, recent work suggests that even higher thresholds (e.g. 98.2–99.0 %) can be used safely depending on the particular taxonomic group (Meier-Kolthoff *et al.*, 2013b), and so we expect that the number of novel OTUs is likely to be even higher. Nevertheless, at the 97 % similarity level, the 20 'species-level' OTUs in the genus *Ruminococcus* and the 27 separate species estimated using dDDH hint at the wealth of unexplored diversity that remains to be uncovered in this genus.

Table 2. Number of *Ruminococcus* OTUs by percentage of 16S rRNA gene sequence similarity

Ruminococcus OTUs (*Ruminococcaceae*)	
16S rRNA similarity (%)	**No. of OTUs**
99	39
98	26
97	21
96	17
95*	11
94	9
93	8
92†	5
91	5
90‡	4
89	3
88	2
86§	1

*All *R. flavefaciens* OTUs cluster together.

†All *R. albus* OTUs cluster together.

‡All ruminococci except *R. bromii* cluster together.

§All ruminococci cluster together.

Another intriguing finding is that *R. bromii*, *R. callidus* and *R. champanellensis* fall within the *Lachnospiraceae* clade (as opposed to their usual positions in the *Ruminococcaceae*) in our GBDP nucleotide analysis (Fig. S2); however, this observation did not hold for our phylogenetic tree generated using protein-encoding genes (Fig. 1d). Although we cannot make strong conclusions due to the low branch support in the former tree, one possibility for this clading is that these particular genomes share features in common with the *Lachnospiraceae* that are not found in the *Ruminococcaceae*. This model is somewhat supported given that these particular ruminococci are all human isolates and share this feature with many of the genera within the *Lachnospiraceae* (Eren *et al.*, 2015), such as *Blautia*.

The deep sequencing of various microbiota has now become routine, and the abundance of publicly available datasets allows for the determination of *Ruminococcus* spp. distribution in various host and non-host environments. We found very few *Ruminococcus* sequences in any of the non-host datasets we examined. Indeed, no environmental dataset contained more than 0.16 % *Ruminococcus* sequences and 15 of 18 environmental datasets contained no *Ruminococcus* sequences whatsoever. The low levels of *Ruminococcus* sequences in these datasets suggests that these may belong to transient bacteria, bacteria that are not alive or may have resulted due to sequence contamination. This finding is further supported by the fact that all described ruminococci have been isolated from animal host sources, although we recognize that exhaustive culturing efforts have not been undertaken for many of the environmental samples in our dataset. Moreover, genome analysis of ruminococci like '*R. bicirculans*' suggests that it has lost the ability to synthesize many essential vitamins and cofactors, while gaining other traits (e.g. bile salt hydrolases and ureases) that allow it to survive in the human gut (Wegmann *et al.*, 2014). A survey of the ruminococci in the KEGG database also confirms that pathways responsible for many essential vitamin and cofactor biosynthesis are missing or incomplete in most of the genomes we examined, including those for biotin, lipoic acid, thiamine and pyridoxal phosphate (data not shown). Based on these lines of evidence, we posit that ruminococci may have evolved to survive within a host environment and have lost many of the traits required to proliferate effectively in non-host environments.

Our analysis of the deep sequence microbiota datasets also revealed that herbivores and omnivores contained the highest proportions of *Ruminococcus* sequences. Moreover, we found that the genus *Ruminococcus* has a far greater host range than has been previously reported from described isolates (Table 1). In addition to known sources like humans and ruminants, we found ruminococci in a large number of non-human primates, non-ruminant herbivores and carnivores. We also found an unexpectedly large presence of ruminococci in some insectivorous/myrmecophagous mammals. This may be due to the high abundance of chitin, the major constituent of the insect exoskeleton, which is structurally similar to cellulose. This raises the possibility that these *Ruminococcus* spp. have developed chitin degrading and fermentative abilities; thus, allowing their hosts to capitalize on this unique diet. This hypothesis is supported by the report of a *R. flavefaciens* strain with weak chitinolytic capacity (Kopecný *et al.*, 1996) and a recent metagenomic characterization of Baleen whales, which primarily eat chitin-rich crustaceans (Sanders *et al.*, 2015), that revealed an abundance of the carbohydrate-binding module family 37 enzyme that is unique to *R. albus* (Ezer *et al.*, 2008). Attempts to culture *Ruminococcus* from hosts with chitin-rich diets should be explored in the future.

In conclusion, we have presented a highly detailed view of the phylogeny, diversity and environmental distribution of the genus *Ruminococcus* using publicly available sequence datasets. In particular, we have provided evidence for the presence of potentially novel species in disparate host environments. Given the cellulolytic capability of some ruminococci, these host sources could harbour unique cellulolytic members with potential to inform advances in the production of valuable bioproducts (e.g. biofuels). Similarly, given their documented abundance in bovines and humans, a better understanding of the biology of individual species members could have important implications for animal health in general, and human health in particular. Overall, we show that this genus is extremely widespread in the animal world and more work should be undertaken to assess the roles of *Ruminococcus* spp. in their unique host environments.

Acknowledgements

The authors would like to thank all members of the Suen lab for their support, insightful discussions and careful reading of the manuscript. This work was supported by a US Department of Energy Biological and Environmental Research Early Career Research Program Award DE–SC0008104 to G. S.

References

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Chassard, C., Delmas, E., Robert, C., Lawson, P. A. & Bernalier-Donadille, A. (2012). *Ruminococcus champanellensis* sp. nov., a cellulose-degrading bacterium from human gut microbiota. *Int J Syst Evol Microbiol* **62**, 138–143.

Chen, J., Stevenson, D. M. & Weimer, P. J. (2004). Albusin B, a bacteriocin from the ruminal bacterium *Ruminococcus albus* 7 that inhibits growth of *Ruminococcus flavefaciens*. *Appl Environ Microbiol* **70**, 3167–3170.

Christopherson, M. R., Dawson, J. A., Stevenson, D. M., Cunningham, A. C., Bramhacharya, S., Weimer, P. J., Kendziorski, C. & Suen, G. (2014). Unique aspects of fiber degradation by the ruminal ethanologen *Ruminococcus albus* 7 revealed by physiological and transcriptomic analysis. *BMC Genomics* **15**, 1066.

Cornick, N. & Stanton, T. (2009). Genus I. *Lachnospira*. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 3, pp. 921–922. Dordrecht, Heidelberg, London and New York: Springer.

Dassa, B., Borovok, I., Ruimy-Israeli, V., Lamed, R., Flint, H. J., Duncan, S. H., Henrissat, B., Coutinho, P., Morrison, M. & other authors (2014). Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains. *PLoS One* **9**, e99221.

Eren, A. M., Sogin, M. L., Morrison, H. G., Vineis, J. H., Fisher, J. C., Newton, R. J. & McLellan, S. L. (2015). A single genus in the gut microbiome reflects host preference and specificity. *ISME J* **9**, 90–100.

Ezer, A., Matalon, E., Jindou, S., Borovok, I., Atamna, N., Yu, Z., Morrison, M., Bayer, E. A. & Lamed, R. (2008). Cell surface enzyme attachment is mediated by family 37 carbohydrate-binding modules, unique to *Ruminococcus albus*. *J Bacteriol* **190**, 8220–8222.

Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol* **6**, 121–131.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K. & Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* **21**, 2329–2335.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.

Hungate, R. E. (1957). Microorganisms in the rumen of cattle fed a constant ration. *Can J Microbiol* **3**, 289–311.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Jindou,S., Brulc, J. M., Levy-Assaraf, M., Rincon, M. T., Flint, H. J., Berg, M. E., Wilson, M. K., White, B. A., Bayer, E. A. & other authors (2008). Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens*. *FEMS Microbiol Lett* **285**, 188–194.

Julliand, V., de Vaux, A., Millet, L. & Fonty, G. (1999). Identification of *Ruminococcus flavefaciens* as the predominant cellulolytic bacterial species of the equine cecum. *Appl Environ Microbiol* **65**, 3738–3741.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066.

Kembel, S. W., O'Connor, T. K., Arnold, H. K., Hubbell, S. P., Wright, S. J. & Green, J. L. (2014). Relationships between phyllosphere bacterial communities and plant functional traits in a neotropical forest. *Proc Natl Acad Sci U S A* **111**, 13715–13720.

Kopecný, J., Hodrová, B. & Stewart, C. (1996). The effect of rumen chitinolytic bacteria on cellulolytic anaerobic fungi. *Lett Appl Microbiol* **3**, 199–202.

Krause, D. O., Dalrymple, B. P., Smith, W. J., Mackie, R. I. & McSweeney, C. S. (1999). 16S rDNA sequencing of *Ruminococcus albus* and *Ruminococcus flavefaciens*: design of a signature probe and its application in adult sheep. *Microbiology* **145**, 1797–1807.

Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B. U., Meier-Kolthoff, J. P., Kumar, N., Bresciani, A., Martínez, I., Just, S. & other authors (2016). The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol* **1**, 16131.

Lawson, P. A. & Finegold, S. M. (2015). Reclassification of *Ruminococcus obeum* as *Blautia obeum* comb. nov. *Int J Syst Evol Microbiol* **65**, 789–793.

Lay, C., Sutren, M., Rochet, V., Saunier, K., Doré, J. & Rigottier-Gois, L. (2005). Design and validation of 16S rRNA probes to enumerate members of the *Clostridium leptum* subgroup in human faecal microbiota. *Environ Microbiol* **7**, 933–946.

Lefort, V., Desper, R. & Gascuel, O. (2015). FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol* **32**, 2798–2800.

Leschine, S. B. (1995). Cellulose degradation in anaerobic environments. *Annu Rev Microbiol* **49**, 399–426.

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D. & other authors (2008a). Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651.

Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. (2008b). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**, 776–788.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen, J., Pang, X. & other authors (2008). Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci U S A* **105**, 2117–2122.

Liu, C., Finegold, S. M., Song, Y. & Lawson, P. A. (2008). Reclassification of *Clostridiumcoccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus* and *Ruminococcus schinkii*as *Blautia coccoides*gen. nov., comb. nov., *Blautia hansenii*comb. nov., *Blautia hydroge. Int J Syst Evol Microbiol* **58**, 1896–1902.

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J. & other authors (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* **40**, D115–D122.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P. & Göker, M. (2013a). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.

Meier-Kolthoff, J. P., Göker, M., Spröer, C. & Klenk, H. P. (2013b). When should a DDH experiment be mandatory in microbial taxonomy? *Arch Microbiol* **195**, 413–418.

Meier-Kolthoff, J., Auch, A. F., Klenk, H. P. & Göker, M. (2014). Highly parallelized inference of large genome-based phylogenies. *Concurr Comput: Pract Exp* **26**, 1715–1729.

Moraïs, S., Ben David, Y., Bensoussan, L., Duncan, S. H., Koropatkin, N. M., Martens, E. C., Flint, H. J. & Bayer, E. A. (2016). Enzymatic profiling of cellulosomal enzymes from the human gut bacterium, *Ruminococcus champanellensis*, reveals a fine-tuned system for cohesin-dockerin recognition. *Environ Microbiol* **18**, 542–556.

Orpin, C. G., Mathiesen, S. D., Greenwood, Y. & Blix, A. S. (1985). Seasonal changes in the ruminal microflora of the high-arctic Svalbard reindeer (*Rangifer tarandus platyrhynchus*). *Appl Environ Microbiol* **50**, 144–151.

Pavlostathis, S. G., Miller, T. L. & Wolin, M. J. (1988). Kinetics of insoluble cellulose fermentation by continuous cultures of *Ruminococcus albus*. *Appl Environ Microbiol* **54**, 2660–2663.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F. & other authors (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65.

Rainey, F. (2009a). Family V. *Lachnospiraceae* fam. nov. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 3, pp. 921. Dordrecht, Heidelberg, London and New York: Springer.

Rainey, F. (2009b). Family VIII. *Ruminococcaceae* fam. nov. In *Bergey's Manual of Systematic Bacteriology*, 2nd edn, vol. 3, pp. 1016–1043. Dordrecht, Heidelberg, London and New York: Springer.

Rainey, F. A. & Janssen, P. H. (1995). Phylogenetic analysis by 16S ribosomal DNA sequence comparison reveals two unrelated groups of species within the genus *Ruminococcus*. *FEMS Microbiol Lett* **129**, 69–73.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Round, J. L. & Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol* **9**, 313–323.

Sanders, J. G., Beichman, A. C., Roman, J., Scott, J. J., Emerson, D., McCarthy, J. J. & Girguis, P. R. (2015). Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores. *Nat Commun* **6**, 8285–8288.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H. & other authors (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537–7541.

Shi, Y., Odt, C. L. & Weimer, P. J. (1997). Competition for cellulose among three predominant ruminal cellulolytic bacteria under substrate-excess and substrate-limited conditions. *Appl Environ Microbiol* **63**, 734–742.

SijpesteijnA. K. (1948). Cellulose-decomposing bacteria from the rumen of the cattle. *PhD thesis*, Leiden University, The Netherlands.

Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Evol Microbiol* **44**, 846–849.

Suen, G., Stevenson, D. M., Bruce, D. C., Chertkov, O., Copeland, A., Cheng, J. F., Detter, C., Detter, J. C., Goodwin, L. A. & other authors (2011). Complete genome of the cellulolytic ruminal bacterium *Ruminococcus albus* 7. *J Bacteriol* **193**, 5574–5575.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729.

Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., Brown, D., Stares, M. D., Scott, P. & other authors (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J* **5**, 220–230.

Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261–5267.

Wegmann, U., Louis, P., Goesmann, A., Henrissat, B., Duncan, S. H. & Flint, H. J. (2014). Complete genome of a new Firmicutes species belonging to the dominant human colonic microbiota ('*Ruminococcus bicirculans*') reveals two chromosomes and a selective capacity to utilize plant glucans. *Environ Microbiol* **16**, 2879–2890.

Yu, Y., Li, H., Zeng, Y. & Chen, B. (2010). Phylogenetic diversity of culturable bacteria from Antarctic sandy intertidal sediments. *Polar Biol* **33**, 869–875.

Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. (2012). *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *ISME J* **6**, 1535–1543.

Ze, X., Ben David, Y., Laverde-Gomez, J. A., Dassa, B., Sheridan, P. O., Duncan, S. H., Louis, P., Henrissat, B., Juge, N. & other authors (2015). Unique organization of extracellular amylases into amylosomes in the resistant starch-utilizing human colonic *Firmicutes* bacterium *Ruminococcus bromii*. *MBio* **6**, e01058-15.

Data Bibliography

Studies used in *Ruminococcus* distribution analysis (detailed in Table S4):

1. Ley *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore?term=(EU458114%3AEU475873%5Baccn%5D)%20OR%20EU771093%3AEU779492%5Baccn%5D (2008).
2. Kautz *et al.* NCBI – SRA https://www.ncbi.nlm.nih.gov/bioproject/PRJNA177884 (2013).
3. Warnecke *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore?term=(EF453758%3AEF455009%5Baccn%5D)%20OR%20EU024891%3AEU024927%5Baccn%5D (2007).
4. Hongoh *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=AB088857:AB089126%5Baccn%5D (2003).
5. Waite *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore?term=((JQ283115%3AJQ283245%5Baccn%5D)%20OR%20JQ302756%5BAccession%5D)%20OR%20JQ302757%5Baccn%5D (2012).
6. Godoy-Vitorino *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore?term=(EU344158%3AEU344737%5Baccn%5D)%20OR%20EU747884%3AEU748538%5Baccn%5D (2008).
7. Hong *et al.* ENA – SRA http://www.ebi.ac.uk/ena/data/view/ERX366920 (2015).
8. Yuan *et al.* Dryad http://datadryad.org/resource/doi:10.5061/dryad.54bm8 (2015).
9. Schauer *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=JN680560:JN68068%5Baccn%5D (2012).
10. Bennett *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRA071216 (2013).
11. Matsui *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=AB385881:AB386190%5Baccn%5D (2010).
12. Wienemann *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=DQ071433%3ADQ071546%5Baccn%5D (2011).
13. Sergeant *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/SRX253206[accn] (2014).
14. Roeselers *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=HM778178:HM780469%5Baccn%5D (2011).
15. King *et al.* MG-RAST http://metagenomics.anl.gov/mgmain.html?mgpage=project&project=mgp1994 (2012).
16. Menke *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRP044660 (2014).
17. Handl *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRA012231 (2011).
18. Zhang and Chen. GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=FJ978479:FJ978785%5Baccn%5D (2010).
19. Delsuc *et al.* EMBL – EBI http://www.ebi.ac.uk/biosamples/group/SAMEG163130 (2014).
20. Keenan *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRA023831 and http://www.ncbi.nlm.nih.gov/sra/?term=SRA062824 (2013).
21. Xia *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=KC599554:KC601631%5Baccn%5D (2014).
22. Miyake *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRP040147 (2015).
23. Weldon *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB9715 (2015).
24. Merson *et al.* MG-RAST http://metagenomics.anl.gov/mgmain.html?mgpage=search&search=Manatee%20gut%20microbiology (2014).
25. Baldo *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRP045780 (2015).
26. Sapountzis *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=280270 (2015).
27. Bolaños *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=KM978218:KM978331%5Baccn%5D (2015).
28. Dewar *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB3083 (2013).
29. Fierer *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB4334 (2013).
30. Navarrete *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB8643 (2015).
31. Humbert *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore?term=((((AJ965761%3AAJ966243%5Baccn%5D)%20OR%20FJ185706%3AFJ185780%5Baccn%5D)%20OR%20FJ207177%3AFJ207344%5Baccn%5D)%20OR%20FJ208378%3AFJ208589%5Baccn%5D)%20OR%20FJ262737%3AFJ262952%5Baccn%5D (2009).
32. Llirós *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=PRJNA241494 (2014).
33. Tang *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore?term=(EU272936%3AEU273248%5Baccn%5D)%20OR%20EU373097%3AEU373211%5Baccn%5D (2009).

34. Sjöstedt *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=KC197360:KC197707%5Baccn%5D (2014).
35. Huber *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRA000195 and http://www.ncbi.nlm.nih.gov/sra/?term=SRA000196 (2007).
36. Andersson *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRA009836%20 (2010).
37. Cameron *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB4904 (2014).
38. França *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB1538 (2014).
39. Bottos *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=EU044919:EU045213%5Baccn%5D (2008).
40. Udikovic-Kolic *et al.* MG-RAST http://metagenomics.anl.gov/mgmain.html?mgpage=search&search=8945 (2014).
41. VandeWalle *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=189333 (2012).
42. McKlellan *et al.* NCBI – SRA http://www.ncbi.nlm.nih.gov/sra/?term=SRP000905 (2010).
43. Knief *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=HE589809:HE589931%5Baccn%5D (2012).
44. Delmotte *et al.* GenBank http://www.ncbi.nlm.nih.gov/nuccore/?term=FN421480:FN421999%5Baccn%5D (2009).
45. Kembel *et al.* Figshare https://figshare.com/articles/Panama_bacterial_phyllosphere_sequences/928573 (2014).
46. Cardinale *et al.* EMBL – EBI http://www.ebi.ac.uk/ena/data/view/PRJEB5101 (2014).

Genomes used in phylogenomic analysis (detailed in Table S6):

Joint Genome Institute – IMG/M database samples

- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=GenomeList&page=phylumGenomeList&domain=Bacteria&type=species&genus=*Ruminococcus*
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2562617090
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2562617038
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=643886206
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=641380422
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2562617078
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=641380427
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2562617159
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=650377941
- https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2519899647

PATRIC database

- https://www.patricbrc.org/portal/portal/patric/Genome?cType=genome&cId=39482.3 (accessed August 2016).