A Phylogenomic View of Ecological Specialization in the Lachnospiraceae, a Family of Digestive Tract-associated Bacteria

Conor J Meehan[1,2], Robert G Beiko [2*]

[1]Department of Biochemistry and Molecular Biology, Dalhousie University, 5080 College Street, Halifax, NS, B3H 4R2, Canada

[2]Faculty of Computer Science, 6050 University Avenue, Halifax, NS, B3H 1W5, Canada

[*]Corresponding author (beiko@cs.dal.ca)

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

ABSTRACT

Several bacterial families are known to be highly abundant within the human microbiome, but their ecological roles and evolutionary histories have yet to be investigated in depth. One such family, Lachnospiraceae (phylum Firmicutes, class Clostridia) is abundant in the digestive tracts of many mammals and relatively rare elsewhere. Members of this family have been linked to obesity and protection from colon cancer in humans, mainly due to the association of this group with the production of butyric acid, a substance that is important for both microbial and host epithelial cell growth. We examined the genomes of 30 Lachnospiraceae isolates to better understand the phylogenetic relationships and basis of ecological differentiation within this group. Although this family is often used as an indicator of butyric acid production, fewer than half of the examined genomes contained genes from either of the known pathways that produce butyrate, with the distribution of this function likely arising in part from lateral gene transfer. An investigation of environment-specific functional signatures indicated that human gut-associated Lachnospiraceae possessed genes for endospore formation while other members of this family lacked key sporulation-associated genes, an observation supported by analysis of metagenomes from the human gut, oral cavity and bovine rumen. Our analysis demonstrates that despite a lack of agreement between Lachnospiraceae phylogeny and assigned habitat there are several examples of genetic signatures of habitat preference derived from both lateral gene transfer and gene loss.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Keywords: lateral gene transfer, microbial genomes, metagenomics, phylogenomics, butyric acid, sporulation

Mammal-associated microbiomes have been shown to influence host health and behavior (Cryan and O'Mahony 2011; Kinross et al. 2011; Muegge et al. 2011) and appear to be hotbeds for lateral gene transfer (LGT) (Smillie et al. 2011; Meehan and Beiko 2012). Lachnospiraceae is a family of clostridia that includes major constituents of mammalian gastrointestinal (GI) tract microbiomes, especially in ruminants (Kittelmann et al. 2013) and humans (Gosalbes et al. 2011). The family is currently described in the NCBI taxonomy as comprised of 24 named genera and several unclassified strains (Sayers et al. 2010) and is defined solely based upon 16S ribosomal RNA gene (henceforth referred to as 16S) similarity (Bryant 1986; Dworkin and Falkow 2006). Family members are strictly anaerobic (Dworkin and Falkow 2006), reside mainly within the digestive tracts of mammals (Bryant 1986; Downes et al. 2002; Carlier et al. 2004; Moon et al. 2008), and are thought to be primarily non-spore-forming (Dworkin and Falkow 2006). They play key roles within the human GI microbiome, demonstrated by their inclusion in an artificial bacterial community that has been used to repopulate a gut microbiome and remedy *Clostridium difficile* infections (Petrof et al. 2013). Early blooms of Lachnospiraceae may be linked with obesity (Cho et al. 2012), most likely due to their short chain fatty acid production (Duncan et al. 2002). However, despite their apparent importance,

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

little is known about the group as a whole outside of its use as an indicator of fecal contamination in water and sewage (Newton et al. 2011; McLellan et al. 2013) and the abundance of butyric acid-producing species within the group (Bryant 1986; Duncan et al. 2002; Louis et al. 2004, 2010; Charrier et al. 2006).

Butyric acid (also known as butanoic acid, butanoate and butyrate) is a short chain fatty acid whose production prevents the growth of some microbes within the digestive tract (Zeng et al. 1994; Sun et al. 1998), and provides a source of energy for other microbes (Liu et al. 1999) and host epithelial cells (Roediger 1980; McIntyre et al. 1993; Hague et al. 1996; Pryde et al. 2002). Butyrate also regulates expression of the AP-1 signaling pathway in key components of human physiology (Nepelska et al. 2012). These functions link butyric acid to protection against colon cancer (Hague et al. 1996; Mandal et al. 2001) and a potential influence on obesity levels (Duncan et al. 2008; Turnbaugh et al. 2008). Two pathways are responsible for fermentation of this short chain fatty acid: through butyrate kinase or through butyryl-CoA:acetate CoA-transferase (BCoAT) (Walter et al. 1993; Duncan et al. 2002). This production appears to be restricted mainly to organisms within the class Clostridia (Louis et al. 2010), and has been demonstrated in many strains of Lachnospiraceae (Attwood et al. 1996; Duncan et al. 2002; Charrier et al. 2006; Kelly et al. 2010; Louis et al. 2010).

Outside of butyrate production, the understanding of how the Lachnospiraceae adapted to their environments, and potential ecological

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

differences within the group, is lacking. Here we investigate the relationship between phylogeny, ecology and biochemistry in this group by examining a set of 30 sequenced genomes, combined with marker gene surveys from a wide range of habitats and metagenomic samples collected from the habitats with high numbers of Lachnospiraceae. Endospore formation distinguished Lachnospiraceae from different habitats, with complete or near-complete sporulation pathways in human gut-associated microorganisms, and many key pathways absent from other members of the group. Butyrate production appears to be sporadically distributed within this family, with strong evidence that some steps of the pathway have undergone lateral gene transfer (LGT). The fluidity of butyrate production and other properties highlights a range of evolutionary processes that impact on adaptation and host interactions.

MATERIALS AND METHODS

Assessing the Habitat of Lachnospiraceae Members

A determination of the environmental range of members of the Lachnospiraceae was undertaken using a phylogenetic assignment method. All 16S sequences from completed genomes and all Clostridiales type strains in the Ribosomal Database Project (Cole et al. 2009) were aligned to the Greengenes reference alignment template using PyNAST (Caporaso et al. 2010) and masked to include only the phylogenetically informative sites, resulting in an alignment of 2,217 sequences and 1,287 sites. A reference tree was then created from these sequences using RaxML version 7.2.5 (Stamatakis 2006). Presence within a

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

habitat was assessed by aligning reads from 1,697 environmental samples of 16S sequences from MG-RAST (Meyer et al. 2008), sorted into 17 habitat types (Supplementary table 1), added to the reference alignment using PyNAST and placed on the reference tree using pplacer version 1.1.alpha13 (Matsen et al. 2010). Taxonomic classification of reads was then undertaken using the classify function of guppy, a part of the pplacer package. Reads were classified as a given taxonomic rank if the posterior probability of that assignment was above 0.7. The percentage of classified reads assigned to Lachnospiraceae was calculated per sample and then aggregated between samples into broad habitat definitions.

Butyric Acid Production

Sequenced Lachnospiraceae genomes were retrieved from NCBI on April 18, 2012 (Supplementary table 2). This resulted in 30 genomes (2 completed and 28 permanent draft) from four primary habitats: the human digestive tract, cow rumen, human oral cavity and sediment containing paper-mill and domestic waste. The potential for butyric acid production was then assessed within each Lachnospiraceae sequenced genome. Sequences annotated as butyrate kinase were retrieved from the KEGG database, version 58.1 (Kanehisa et al. 2004), as this encodes one of the final steps of the two butyric acid pathways. The other path to butyric acid production is through utilization of butyryl CoA:acetate CoA-transferase (BCoAT) (Louis et al. 2010). The sequences derived from (Louis et al. 2010) constituted the reference database for our search. These two datasets were used to mine the protein sets of each sequenced Lachnospiraceae genome

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

using USEARCH 4.0.38 (Edgar 2010) with an e-value cut-off of 10^{-30} and a minimum identity cut-off of 70%. The origin of the butyrate-related genes was assessed using a phylogenetic approach. Protein sequences encoded by 3,500 bacterial and archaeal genomes were retrieved from NCBI and USEARCH was used in same manner as above to search for the two butyric acid-related genes, with the Lachnospiraceae sequences identified above as queries. Sequences were aligned using MUSCLE version 3.8.31 (Edgar 2004a, 2004b) and trimmed using BMGE version 1.1 (Criscuolo and Gribaldo 2010) with a BLOSUM30 matrix and a 0.7 entropy cut-off. A phylogenetic tree was created using FastTree version 2.1.4 (Price et al. 2010) with a gamma parameter to model rate variation across sites.

In order to test whether lateral gene transfer occurred within the history of these genes, a comparison of the resulting topologies to the 16S tree (as a proxy for implied vertical inheritance) was undertaken. The longest 16S sequence from each genome found to have a predicted butyrate kinase was extracted and an alignment and tree were built as above. The per-site likelihoods of the 16S topology and the topology based upon the butyrate kinase alignment were calculated using FastTree with the butyrate kinase alignment as the dataset and an AU test was performed using CONSEL (Shimodaira and Hasegawa 2001). This procedure was repeated using the BCoAT-containing species.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Clustering of Genomes Based on Homologous Gene Groups

A comparative genomics approach was undertaken to understand the shared functional repertoires of members of the Lachnospiraceae. To construct a set of shared homologous protein-coding genes, BlastClust (Altschul et al. 1990) was employed with a minimum match criterion of 40% identity and 70% length on all genes. Functional assignment to each cluster was performed using the Clusters of Orthologous Groups (COG) database (Tatusov et al. 2000). BLASTP (Altschul et al. 1990) with a 10^{-3} e-value cut-off was employed for each gene cluster using representative protein sequences for each of the 18 COG functional categories as a database. Lachnospiraceae genomes were then clustered based upon pair-wise counts of shared homologous gene clusters to look for associations between shared genome content and habitat. These pair-wise counts were calculated using a normalized Hamming distance such that the distance between genomes *x* and *y* is (A+B-2S)/(A+B) where A and B are the total gene counts of *x* and *y* respectively and S is the number of shared genes between *x* and *y* (Lin and Gerstein 2000). If a cluster contained more than one gene in a given genome (e.g. in-paralogs) S equals the smaller gene count per genome. Counts were then clustered and displayed using the R package gplots (Warnes et al. 2012). Groups of interest were further analyzed using Interproscan version 4.8 (Zdobnov and Apweiler 2001) to determine what functions may define such groups.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Distribution of Sporulation Capabilities in Sequenced Genomes and Metagenomes

Each Lachnospiraceae genome was compared to the sporulation-associated proteins as found within *Bacillus subtilis* strain 168 (Kunst et al. 1997). The *B. subtilis* proteins labeled as within the main sporulation-associated families *cot*, *spol*, *sps*, and *ssp* were used as a database for a BLASTP search with a 10^{-30} e-value cut-off and all Lachnospiraceae proteins as queries. The putative history of each sporulation protein was assessed with the same phylogenetic method as was used for the butyric acid-related proteins.

Metagenomes for the human digestive tract ((Yatsunenko et al. 2012);MG-RAST project 401; 107 samples), human oral cavity (Human Microbiome Project; MG-RAST project 385; 12 samples) and cow rumen ((Brulc et al. 2009); MG-RAST project 24; 4 samples; (Hess et al. 2011); SRA023560; 1 sample) were used to assess the distribution of Lachnospiraceae-derived sporulation proteins in culture-independent data sets. The Lachnospiraceae-associated sporulation genes were used as a database with a metagenome sample as a query input to USEARCH with a 10^{-10} e-value cut-off. From this set of results we removed all reads whose best match was to a non-Lachnospiraceae genome in the set of 3,500 NCBI genomes. Final counts of reads designated as sporulation-associated were compared between habitats using STAMP version 2 (Parks and Beiko 2010) with a two-sided Welch's t-test and Bonferroni multiple test correction.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Phylogenomic Analysis of the Lachnospiraceae

Assessment of intra-family relationships was undertaken using three different methods: phylogenetic tree inference using 16S, tree inference using a concatenated alignment of 91 shared protein-coding genes and a consensus network of relationships (Holland et al. 2004) based on the same set of shared genes.

The longest 16S rRNA gene sequence from each Lachnospiraceae genome along with those of two species from the family Ruminococcaceae as an outgroup (*Ruminococcus albus* 7 and *Ethanoligenens harbinense* YUAN-3) were aligned using PyNAST and trimmed to include only the phylogenetically informative sites used by Greengenes (DeSantis et al. 2006). A reference tree was created using RaxML version 7.2.5 with the evolutionary model GTR + Γ + I as selected using the Bayesian information criterion in PartitionFinder (Lanfear et al. 2012). A set of family-wide shared genes was created from the homologous gene clusters output from BlastClust. Those gene sets that were present as a single copy in each genome were selected and USEARCH was used with an e-value cut-off of 10^{-30} to find genes in the completed genomes of members of the Ruminococcaceae that would serve as an outgroup. Alignments were constructed using MUSCLE and trimmed using BMGE as above. Resulting alignments were then concatenated and a tree inferred using FastTree with a gamma parameter.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

SEQBOOT (Felsenstein 1989) was used to generate 100 randomizations each of the 16S and concatenated alignments, which were then subjected to phylogenetic analysis as above to establish bootstrap support. Concordance between the 16S tree and concatenated alignment tree was tested using the subtree prune-and-regraft (SPR) distance with rSPR version 1.0.2 (Whidden et al. 2010) and the approximately unbiased (AU) test in CONSEL version 0.20 (Shimodaira and Hasegawa 2001). Individual gene alignments were also tested for concordance with the concatenated sequence tree using the AU test in CONSEL. The set of shared Lachnospiraceae protein-coding genes was used to create a consensus network using SplitsTree4 (Huson and Bryant 2006) with a 0.7 similarity cut-off and edges weighted by counts.

RESULTS

Lachnospiraceae are Common only in Host-associated and Sewage Effluent Samples

We examined a total of 1,697 published marker-gene surveys from different environments to determine the primary habitats of the Lachnospiraceae. Sequences associated with the group were more abundant in the GI tracts of mammals compared to other environments, including other mammal-associated body sites (Fig. 1). While mammalian GI samples tended to have a relative abundance of Lachnospiraceae in excess of 10%, in others the relative abundance was often less than 1%. Variation was found between different human life stages with abundance of Lachnospiraceae highest in the adult GI

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

tract, moderate in infants and approximately 1% in newborns. Smaller proportions were found in other animals such as fleas and snakes, whose numbers were higher than those of newborn humans and all non-animal-associated habitats. Only the cow rumen, human digestive tract, human oral cavity and sewage effluent microbiomes were predicted to have Lachnospiraceae in every sample. As Lachnospiraceae genomes from similar environments were available, extensive functional and phylogenomic analysis of the group was undertaken using 30 representative genomes (Supplementary table 2).

Butyric Acid Production is not a Defining Trait of the Lachnospiraceae

Lachnospiraceae members have been implicated in butyric acid production in the human GI tract (Duncan et al. 2008; Louis et al. 2010; Van-den-Abbeele et al. 2012). Here the capability to produce butyric acid along with its evolutionary history was investigated to determine whether it is a defining characteristic of the family as a whole. Two enzymes allow for the production of butyric acid: butyrate kinase (from Butanoyl-P) (Walter et al. 1993) and butyryl-CoA:acetate CoA-transferase (BCoAT) (from Butanoyl-CoA) (Duncan et al. 2002). Only twelve of the thirty sequenced organisms contained genes annotated from at least one of these two pathways (Table 1). Pathways appeared to be genus-specific as all *Shuttleworthia, Butyrivibrio* and *Coprococcus* genomes encode butyrate kinase and both *Roseburia* strains, both *Anaerostipes* strains and *Lachnospiraceae* bacterium 5_1_63FAA encode BCoAT. Analysis with TBLASTN did not reveal any additional hits within the Lachnospiraceae genomes.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Phylogenetic examination of the two genes revealed potential LGT within their histories. The topology of each gene tree was tested against a 16S tree derived from the same genomes (Supplementary fig. 1). Use of the AU test showed that the gene trees for butyrate kinase and BCoAT in these species were significantly different from the companion 16S tree ($p < 0.001$). This indicates that rearrangements away from a proxy for vertical inheritance occurred within the gene trees, indicative of LGT of both butyrate kinase and BCoAT. Additionally, the 16S tree placed many species that are not currently classed in the NCBI taxonomy as Lachnospiraceae (e.g. *Eubacterium rectale*) proximal to recognized members of this family, suggesting the need for taxonomic revision of the group.

Shared Gene Clusters Reveal Functional Signatures of Habitat Specialization

A thorough investigation of the family was undertaken to look for defining features of the Lachnospiraceae using sets of homologous gene clusters shared between members of this bacterial family. A total of 167 gene clusters were shared by all sequenced Lachnospiraceae with predicted functions spanning information processing (46%), metabolism (15%, primarily glycolysis and fructose metabolism; COG category G) and cellular processes/ signaling (9%), including two multi-drug resistance mechanisms and several sigma factors. Thus only 16S similarity and a handful of metabolic and cellular processes appear to be shared by all members of the Lachnospiraceae family.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Pair-wise gene cluster counts between sequenced genomes were computed in order to observe whether habitat correlated with the presence of specific groups of genes (Fig. 2). Some association between habitat and clustering was observed, including a basal split into a group consisting exclusively of twelve human gut-associated family members (referred to as the gut-restricted group) and another group containing genomes from all represented habitats, which contained a smaller cluster of eight gut-associated genomes (Fig. 2). The average genome size was 3,539 genes (range: 1,950 to 6,887) for the mixed habitat group and 2,920 genes (range: 2,081 to 3,534) for the gut-restricted group. The average genome size within this dataset, regardless of clustering, is 3,291, suggesting that group associations are not biased by genome size.

Gene clusters that defined certain groups were investigated further in order to observe functional patterns. A gene cluster was classed as group-specific if it was present in at least 90% of the genomes in one group and absent from 90% of the complementary group. Comparison of the gut-restricted group and all other Lachnospiraceae revealed 41 shared gene clusters that were indicative of this group (i.e. present in at least 11 gut-restricted genomes and absent from at least 16 of the other genomes). Functionally these genes encompassed mostly protein binding (primarily tetratricopeptide repeat motifs), signal transduction and sporulation, with almost a third of the homologous gene clusters having no annotated function (Supplementary table 3a). Only one gene cluster, annotated as an inner membrane component of a transporter complex, was classed as a

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

defining gene cluster for the multi-habitat group when compared to the gut-restricted group.

The gut-restricted group was also found to have several gene clusters that distinguish them from the 8 genomes of the other gut-associated Lachnospiraceae (Supplementary table 3b). Several tetratricopeptide repeat protein-binding motifs were present in the gut-restricted group and absent from many of the other gut-associated genomes. Most other potentially defining functions encompassed transporters and signaling pathways with 30% of clusters having no known function. The reverse comparison (clusters absent from the majority of the gut-restricted group but present in the other gut-associated members) revealed several catalytic and transportation-related functions without any discernible pattern.

Almost all of the organisms in the gut-restricted group were also those predicted to be incapable of producing butyric acid (Supplementary table 4; Fig. 2). This indicates a split in the human gut-associated Lachnospiraceae between those capable of producing butyric acid by either of the known pathways and those who, while lacking this capability, have genomes that are more closely related to each other (the gut-restricted group). Several gene clusters that correlated with the presence or absence of butyric acid production within the human GI-tract-associated Lachnospiraceae (Supplementary table 3c) also distinguished the gut-restricted group from the other gut-associated family members (Supplementary table 3b, Supplementary table 4). Thus, even though

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

multiple pathways can result in butyric acid production, the presence or absence of this function appears to have an influence on the specialization of certain organisms within the human gut microbiome.

In order to observe whether similar patterns of distinguishing functions existed between all gut-associated family members (22 strains) and non-gut associated members (8 strains), a similar analysis of gene group presence/absence was performed. Fifty-seven functions present in 20 or more gut-associated strains and absent from 7 or 8 non-gut-associated strains were identified (Supplementary table 3d). Only one protein was of unknown function with the remaining spread across designations such as DNA binding, repair and transcription. Several serine-type endopeptidases or associated proteins were present within this group and lacking from the others, suggesting potential involvement of protein modification in adaption to the human GI tract environment. As was observed with the gut-restricted group, sporulation-related proteins comprised a large fraction of these functions (28%), although different sporulation proteins distinguished these two groups.

Key Sporulation Proteins are Detected only in Human Digestive Tract-associated Family Members

We further examined the distribution of four types of sporulation genes: *cot* genes, which encode protein components of the coat; *spo* genes, which perform functions across all six stages of sporulation; *sps* genes, involved in spore coat polysaccharide synthesis; and *ssp* genes, which create small acid-soluble spore

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

proteins. Homology searches against related sequences from *Bacillus subtilis* (84 genes) revealed that 27 of these genes had no known homolog in any Lachnospiraceae sequenced genome. Of the remaining 57 genes, 29 were present in the majority of gut-associated Lachnospiraceae and completely absent from the rumen and oral-associated family members (Fig. 3). These genes were not restricted to any one class or stage of sporulation protein. *Cellulosilyticum lentocellum* DSM 5427, isolated from sediment containing domestic waste, grouped with the gut-associated members suggesting that it too may be adapted to the human digestive tract.

All sporulation-controlling sigma factors (σ^A, σ^{E-H} and σ^K) were detected in all Lachnospiraceae genomes, which suggests this function was present in the ancestor the group. Phylogenetic analysis also suggested vertical transmission of this function, although uncertain taxonomic assignments make such conclusions difficult to confirm. These analyses suggest that gene loss rather than LGT is responsible for the observed habitat-associated pattern of sporulation genes. In order to confirm a differential presence of sporulation capability in the three habitats (human gut, human oral cavity and cow rumen) metagenomic samples from each microbiome were mined to find sequences related to each Lachnospiraceae-associated sporulation protein. Lachnospiraceae-derived sporulation-associated reads were found to be more abundant within the human GI tract compared to the cow rumen or human oral cavity ($p < 0.001$) (Supplementary fig. 2). The difference in abundance between the rumen and oral cavity was less well supported ($p = 0.022$; difference in relative means = 0.013).

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Thus it is likely that sporulation capabilities within this family are restricted to those found in the human GI tract.

Candidate Phylogenies do not Reflect Habitat Diversification

Functional analysis of the Lachnospiraceae-associated genomes revealed both vertical and lateral acquisition of genes that were indicative of sub-groups within the family. Although species tree reconstruction can be undertaken in several ways, we chose two popular methods for comparison: 16S phylogeny and a shared ortholog concatenated alignment phylogeny. The 16S rRNA gene tree yielded little support for the majority of clades (38% of clades with >75% bootstrap support) (Fig. 4a), likely due to short internal branches in the tree (Wiens et al. 2008). This poor support contrasted with strong support across the tree derived from the concatenated alignment from 91 ubiquitous, single-copy orthologous genes (88% of clades with >75% bootstrap support) (Fig. 4b). However, this tree was not in strong agreement with those of the 91 constituent genes according to the AU test (82% rejected with $p < 0.001$). Even within this restricted 'core' set shared by all family members, significant phylogenetic discordance is observed. Comparison of phylogenetic relationships derived from 16S sequences and concatenated shared orthologs revealed substantial topological differences, as demonstrated by an SPR distance of 12 between trees with only 30 leaves (Supplementary fig. 3). Additionally, each tree was rejected under the AU test ($p < 0.001$) when compared to the alignment of the other (i.e. 16S topology derived from the concatenated alignment and vice-versa), demonstrating that neither the 16S tree nor the concatenated alignment

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

tree is a convincing proxy for the evolutionary history of the full genomes. The consensus network based upon the 91 shared orthologs demonstrated that no clear signal could differentiate the majority of individual strains into a hierarchical structure with little grouping at the genus level, despite high bootstrap support for groupings in the concatenated sequence tree (Supplementary fig. 4).

The estimated gain and loss of both butyric acid production and sporulation functionality was mapped onto both the 16S and the concatenated sequence trees (Fig. 4). Multiple acquisition points of each type of butyric acid production can be observed in both trees, supporting the case for LGT of this function into this group. However, if the 16S tree does map the true history of this group, the butyrate kinase gene (Fig. 4a) may have been acquired through LGT by an ancestor of many of the family members and lost in three subsequent lineages, as opposed to five independent gains. This is supported by the phylogenetic analysis of this gene, although directionality cannot be determined due to an unresolved species tree (Supplementary fig. 1). The observed pattern of sporulation capabilities (Fig. 3) could be explained by four gene loss events, no matter the representative tree. This supports a model of vertical inheritance with subsequent gene loss in a habitat-specific manner. Additionally, within the 16S tree, most of the gut-restricted group (Fig. 2) formed a near-clade with one non-group intruder (*Coprococcus comes* ATCC 27758) and one member absent (*Lachnospiraceae* bacterium 6_1_63FAA) (Fig. 4).

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

DISCUSSION

Lachnospiraceae were found to be present primarily within the mammalian GI tract (Fig. 1), as has been suggested previously (Gosalbes et al. 2011; Kittelmann et al. 2013), although low-abundance populations are present in a wider range of environments including non-host-associated microbiomes. The capacity for butyric acid production was found in fewer than half of the Lachnospiraceae genomes and present in genomes associated with three of the four sampled habitats. Both pathways for producing butyric acid (butyrate kinase and BCoAT) were present in Lachnospiraceae members, with no genome containing both (Table 1). Although seven genomes contained butyrate kinase they appear to have potentially acquired the corresponding gene laterally from other members of class Clostridia (Supplementary fig. 1), a group associated with frequent LGT events (Beiko et al. 2005; Sebaihia et al. 2006; Nelson et al. 2010), especially within GI tracts (Meehan and Beiko 2012). LGT has also contributed to the distribution of the BCoAT-mediated pathway, the main route for butyric acid production within the human GI tract (Louis et al. 2004; Louis and Flint 2009). However, several species not designated as Lachnospiraceae in the NCBI taxonomy were found in close proximity to organisms such as Roseburia in the 16S phylogeny. An example is *E. rectale*, which Mannarelli et al. (Mannarelli et al. 1990) also placed in family Lachnospiraceae. Such discrepancies between published work and taxonomic databases make determination of directionality and evolutionary history difficult. Reconciling taxonomy and phylogeny is no

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

trivial task given LGT and other challenges, but would clarify the origin of butyrate production and other capabilities in the Lachnospiraceae.

Although butyric acid production was not found to segregate the Lachnospiraceae by habitat, several other functions were correlated with specific habitat-associated groups. Tetratricopeptide repeat motif-containing proteins were present in a subset of human GI tract-associated strains and absent from other members in the same environment (Supplementary table 3a, 3b). These motifs play a role in protein-protein interactions and have been associated previously with bacterial pathogens and virulence (Cerveny et al. 2013). As no Lachnospiraceae pathogens have been found before, further investigations into this group, which also lack butyric-acid production capabilities (Supplementary table 3), are needed to clarify their role or roles within the human gut.

Genome-wide investigation into the 22 Lachnospiraceae associated with the human GI tract revealed an almost full complement of sporulation proteins while those residing in the human oral cavity or cow rumen were lacking such functions (Fig. 3, Supplementary table 3d). *C. lentocellum*, the only Lachnospiraceae with confirmed endospore formation capabilities (Attwood et al. 1996; Kelly et al. 2010), grouped with the GI tract-associated genomes. This strain was isolated from a sediment bank receiving domestic waste (Murray et al. 1986), and thus may actually be human-associated with endospore formation as a habitat adaptation for passage through the human stomach as is observed in *C. difficile* (Wilson 1983) and cyst formation in several protist species (Bingham

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

and Meyer 1979; Lujan et al. 1997). As analysis of these proteins suggested primarily vertical inheritance of the associated genes, it is likely that this capability was present in a common ancestor and subsequently lost in a habitat-specific fashion.

Our approach to understanding the Lachnospiraceae combined reference genomes of known provenance with marker-gene and metagenome samples from a range of habitats. No phylogenomic approach we used produced a separation of lineages based on habitat, raising the question of how lineages can change their habitat preference through time. We found little support for many genera within this family, and 16S trees placed several other organisms within this group (Supplementary fig. 1), suggesting taxonomic revisions may be required as has been done previously (Moon et al. 2008; Cai and Dong 2010). Despite the inconsistencies observed with regards to taxonomic classifications, some genes clearly separated lineages based on habitat. These genes shed light on how important habitat-specific transitions in the Lachnospiraceae have occurred and how within-habitat divisions, such as the ability to produce butyric acid, can influence the evolution of closely related organisms. As more Lachnospiraceae genomes become available covering important genera such as *Blautia* and likely mislabeled members such as *Eubacterium rectale*, similar analysis may reveal this pattern to extend to these genera and also potentially to other GI tract-associated microorganisms, revealing how such microbes adapt to the host environment.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

ACKNOWLEDGEMENTS

We thank Dr. Petra Louis for supplying the BCoAT database sequences and for discussion of butyrate production. We also thank Morgan Langille, Eva Boon, Katherine Dunn and W. Ford Doolittle for input and comments.

REFERENCES

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Attwood G.T., Reilly K., Patel B.K. 1996. Clostridium proteoclasticum sp. nov., a novel proteolytic bacterium from the bovine rumen. Int J Syst Bacteriol. 46:753–758.

Beiko R.G., Harlow T.J., Ragan M.A. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A. 102:14332–14337.

Bingham A.K., Meyer E.A. 1979. Giardia excystation can be induced in vitro in acidic solutions. Nature. 277:301–302.

Brulc J.M., Antonopoulos D.A., Miller M.E., Wilson M.K., Yannarell A.C., Dinsdale E.A., Edwards R.E., Frank E.D., Emerson J.B., Wacklin P., Coutinho P.M., Henrissat B., Nelson K.E., White B.A. 2009. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc Natl Acad Sci U S A. 106:1948–1953.

Bryant M.P. 1986. Genus IV. Butryrivibrio . In: Sneath P.H.A., Mair N.S., Sharpe M.E., Holt J.G. Bergey's Manual of Systematic Bacteriology.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Cai S., Dong X. 2010. Cellulosilyticum ruminicola gen. nov., sp. nov., isolated from the rumen of yak, and reclassification of Clostridium lentocellum as Cellulosilyticum lentocellum comb. nov. Int J Syst Evol Microbiol. 60:845–849.

Caporaso J.G., Bittinger K., Bushman F.D., DeSantis T.Z., Andersen G.L., Knight R. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 26:266–267.

Carlier J.P., K'Ouas G., Bonne I., Lozniewski A., Mory F. 2004. Oribacterium sinus gen. nov., sp. nov., within the family "Lachnospiraceae" (phylum Firmicutes). Int J Syst Evol Microbiol. 54:1611–1615.

Cerveny L., Straskova A., Dankova V., Hartlova A., Ceckova M., Staud F., Stulik J. 2013. Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. Infect Immun. 81:629–635.

Charrier C., Duncan G.J., Reid M.D., Rucklidge G.J., Henderson D., Young P., Russell V.J., Aminov R.I., Flint H.J., Louis P. 2006. A novel class of CoA-transferase involved in short-chain fatty acid metabolism in butyrate-producing human colonic bacteria. Microbiology. 152:179–185.

Cho I., Yamanishi S., Cox L., Methe B.A., Zavadil J., Li K., Gao Z., Mahana D., Raju K., Teitler I., Li H., Alekseyenko A. V, Blaser M.J. 2012. Antibiotics in early life alter the murine colonic microbiome and adiposity. Nature. 488:621–626.

Cole J.R., Wang Q., Cardenas E., Fish J., Chai B., Farris R.J., Kulam-Syed-Mohideen A.S., McGarrell D.M., Marsh T., Garrity G.M., Tiedje J.M. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37:D141–5.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Criscuolo A., Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10:210.

Cryan J.F., O'Mahony S.M. 2011. The microbiome-gut-brain axis: from bowel to behavior. Neurogastroenterol. Motil. 23:187–92.

DeSantis T.Z., Hugenholtz P., Larsen N., Rojas M., Brodie E.L., Keller K., Huber T., Dalevi D., Hu P., Andersen G.L. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Env. Microbiol. 72:5069–5072.

Downes J., Munson M.A., Radford D.R., Spratt D.A., Wade W.G. 2002. Shuttleworthia satelles gen. nov., sp. nov., isolated from the human oral cavity. Int J Syst Evol Microbiol. 52:1469–1475.

Duncan S.H., Barcenilla A., Stewart C.S., Pryde S.E., Flint H.J. 2002. Acetate Utilization and Butyryl Coenzyme A (CoA):Acetate-CoA Transferase in Butyrate-Producing Bacteria from the Human Large Intestine. Appl. Environ. Microbiol. 68:5186–5190.

Duncan S.H., Lobley G.E., Holtrop G., Ince J., Johnstone a M., Louis P., Flint H.J. 2008. Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes. 32:1720–1724.

Dworkin M., Falkow S. 2006. The prokaryotes. Vol. 4, Bacteria : firmicutes, cyanobacteria a handbook on the biology of bacteria. 1 v.

Edgar R.C. 2004a. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edgar R.C. 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Edgar R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 26:2460–2461.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 5:164–166.

Gosalbes M.J., Durbán A., Pignatelli M., Abellan J.J., Jiménez-Hernández N., Pérez-Cobas A.E., Latorre A., Moya A. 2011. Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One. 6:e17447.

Hague A., Butt A.J., Paraskeva C. 1996. The role of butyrate in human colonic epithelial cells: an energy source or inducer of differentiation and apoptosis? Proc Nutr Soc. 55:937–943.

Hess M., Sczyrba A., Egan R., Kim T.-W.W., Chokhawala H., Schroth G., Luo S., Clark D.S., Chen F., Zhang T., Mackie R.I., Pennacchio L.A., Tringe S.G., Visel A., Woyke T., Wang Z., Rubin E.M. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science (80-.). 331:463–7.

Holland B.R., Huber K.T., Moulton V., Lockhart P.J. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol Biol Evol. 21:1459–1461.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Kanehisa M., Goto S., Kawashima S., Okuno Y., Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32:D277–80.

Kelly W.J., Leahy S.C., Altermann E., Yeoman C.J., Dunne J.C., Kong Z., Pacheco D.M., Li D., Noel S.J., Moon C.D., Cookson A.L., Attwood G.T. 2010. The

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

glycobiome of the rumen bacterium Butyrivibrio proteoclasticus B316(T) highlights adaptation to a polysaccharide-rich environment. PLoS One. 5:e11942.

Kinross J.M., Darzi A.W., Nicholson J.K. 2011. Gut microbiome-host interactions in health and disease. Genome Med. 3:14.

Kittelmann S., Seedorf H., Walters W.A., Clemente J.C., Knight R., Gordon J.I., Janssen P.H. 2013. Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. PLoS One. 8:e47879.

Kunst F., Ogasawara N., Moszer I., Albertini A.M., Alloni G., Azevedo V., Bertero M.G., Bessieres P., Bolotin A., Borchert S., Borriss R., Boursier L., Brans A., Braun M., Brignell S.C., Bron S., Brouillet S., Bruschi C. V, Caldwell B., Capuano V., Carter N.M., Choi S.K., Codani J.J., Connerton I.F., Danchin A., et al. 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature. 390:249–256.

Lanfear R., Calcott B., Ho S.Y., Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 29:1695–1701.

Lin J., Gerstein M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10:808–818.

Liu Y., Balkwill D.L., Aldrich H.C., Drake G.R., Boone D.R. 1999. Characterization of the anaerobic propionate-degrading syntrophs Smithella propionica gen. nov., sp. nov. and Syntrophobacter wolinii. Int J Syst Bacteriol. 49 Pt 2:545–556.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Louis P., Duncan S.H., McCrae S.I., Millar J., Jackson M.S., Flint H.J. 2004. Restricted distribution of the butyrate kinase pathway among butyrate-producing bacteria from the human colon. J Bacteriol. 186:2099–2106.

Louis P., Flint H.J. 2009. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS Microbiol Lett. 294:1–8.

Louis P., Young P., Holtrop G., Flint H.J. 2010. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. Env. Microbiol. 12:304–314.

Lujan H.D., Mowatt M.R., Nash T.E. 1997. Mechanisms of Giardia lamblia differentiation into cysts. Microbiol Mol Biol Rev. 61:294–304.

Mandal M., Olson D.J., Sharma T., Vadlamudi R.K., Kumar R. 2001. Butyric acid induces apoptosis by up-regulating Bax expression via stimulation of the c-Jun N-terminal kinase/activation protein-1 pathway in human colon cancer cells. Gastroenterology. 120:71–78.

Mannarelli B.M., Stack R.J., Lee D., Ericsson L. 1990. Taxonomic Relatedness of Butyrivibrio, Lachnospira, Roseburia, and Eubacterium Species as Determined by DNA Hybridization and Extracellular-Polysaccharide Analysis. Int. J. Syst. Bacteriol. 40:370–378.

Matsen F. a, Kodner R.B., Armbrust E.V. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 11:538.

McIntyre A., Gibson P.R., Young G.P. 1993. Butyrate production from dietary fibre and protection against large bowel cancer in a rat model. Gut. 34:386–391.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

McLellan S.L., Newton R.J., Vandewalle J.L., Shanks O.C., Huse S.M., Eren A.M., Sogin M.L. 2013. Sewage reflects the distribution of human faecal Lachnospiraceae. Env. Microbiol. .

Meehan C.J., Beiko R.G. 2012. Lateral gene transfer of an ABC transporter complex between major constituents of the human gut microbiome. BMC Microbiol. 12:248.

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E.M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., Edwards R.A. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 9:386.

Moon C.D., Pacheco D.M., Kelly W.J., Leahy S.C., Li D., Kopecny J., Attwood G.T. 2008. Reclassification of Clostridium proteoclasticum as Butyrivibrio proteoclasticus comb. nov., a butyrate-producing ruminal bacterium. Int. J. Syst. Evol. Microbiol. 58:2041–5.

Muegge B.D., Kuczynski J., Knights D., Clemente J.C., Gonzalez A., Fontana L., Henrissat B., Knight R., Gordon J.I., González A. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science. 332:970–4.

Murray W.D., Hofmann L., Campbell N.L., Madden R.H. 1986. Clostridium lentocellum sp. nov., a cellulolytic species from river sediment containing paper-mill waste. Syst. Appl. Microbiol. 8:181–184.

Nelson D.M., Cann I.K., Altermann E., Mackie R.I. 2010. Phylogenetic evidence for lateral gene transfer in the intestine of marine iguanas. PLoS One. 5:e10785.

Nepelska M., Cultrone A., Beguet-Crespel F., Le Roux K., Dore J., Arulampalam V., Blottiere H.M. 2012. Butyrate Produced by Commensal Bacteria Potentiates

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Phorbol Esters Induced AP-1 Response in Human Intestinal Epithelial Cells. PLoS One. 7:e52869.

Newton R.J., Vandewalle J.L., Borchardt M.A., Gorelick M.H., McLellan S.L. 2011. Lachnospiraceae and Bacteroidales alternative fecal indicators reveal chronic human sewage contamination in an urban harbor. Appl Env. Microbiol. 77:6972–6981.

Parks D.H., Beiko R.G. 2010. Identifying biologically relevant differences between metagenomic communities. Bioinformatics. 26:715–21.

Petrof E.O., Gloor G.B., Vanner S.J., Weese S.J., Carter D., Daigneault M.C., Brown E.M., Schroeter K., Allen-Vercoe E. 2013. Stool substitute transplant therapy for the eradication of Clostridium difficile infection: “RePOOPulating” the gut. Microbiome. 1:3.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 5:e9490.

Pryde S.E., Duncan S.H., Hold G.L., Stewart C.S., Flint H.J. 2002. The microbiology of butyrate formation in the human colon. FEMS Microbiol Lett. 217:133–139.

Roediger W.E. 1980. Role of anaerobic bacteria in the metabolic welfare of the colonic mucosa in man. Gut. 21:793–798.

Sayers E.W., Barrett T., Benson D.A., Bolton E., Bryant S.H., Canese K., Chetvernin V., Church D.M., Dicuccio M., Federhen S., Feolo M., Geer L.Y., Helmberg W., Kapustin Y., Landsman D., Lipman D.J., Lu Z., Madden T.L., Madej T., Maglott D.R., Marchler-Bauer A., Miller V., Mizrachi I., Ostell J., Panchenko A., Pruitt K.D., Schuler G.D., Sequeira E., Sherry S.T., Shumway M., Sirotkin K., Slotta D., Souvorov A., Starchenko G., Tatusova T.A., Wagner L., Wang Y., John Wilbur W.,

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Yaschenko E., Ye J. 2010. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 38:D5–16.

Sebaihia M., Wren B.W., Mullany P., Fairweather N.F., Minton N., Stabler R., Thomson N.R., Roberts A.P., Cerdeno-Tarraga A.M., Wang H., Holden M.T., Wright A., Churcher C., Quail M.A., Baker S., Bason N., Brooks K., Chillingworth T., Cronin A., Davis P., Dowd L., Fraser A., Feltwell T., Hance Z., Holroyd S., Jagels K., Moule S., Mungall K., Price C., Rabbinowitsch E., Sharp S., Simmonds M., Stevens K., Unwin L., Whithead S., Dupuy B., Dougan G., Barrell B., Parkhill J. 2006. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nat Genet. 38:779–786.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.

Smillie C.S., Smith M.B., Friedman J., Cordero O.X., David L. a., Alm E.J. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 480:241–244.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Sun C.Q., O'Connor C.J., Turner S.J., Lewis G.D., Stanley R.A., Roberton A.M. 1998. The effect of pH on the inhibition of bacterial growth by physiological concentrations of butyric acid: implications for neonates fed on suckled milk. Chem Biol Interact. 113:117–131.

Tatusov R.L., Galperin M.Y., Natale D.A., Koonin E. V 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Turnbaugh P.J., Hamady M., Yatsunenko T., Cantarel B.L., Duncan A., Ley R.E., Sogin M.L., Jones W.J., Roe B.A., Affourtit J.P., Egholm M., Henrissat B., Heath A.C., Knight R., Gordon J.I. 2008. A core gut microbiome in obese and lean twins. Nature. 457:480–484.

Van-den-Abbeele P., Belzer C., Goossens M., Kleerebezem M., De Vos W.M., Thas O., De Weirdt R., Kerckhof F.M., Van de Wiele T. 2012. Butyrate-producing Clostridium cluster XIVa species specifically colonize mucins in an in vitro gut model. ISME J. .

Walter K.A., Nair R. V, Cary J.W., Bennett G.N., Papoutsakis E.T. 1993. Sequence and arrangement of two genes of the butyrate-synthesis pathway of Clostridium acetobutylicum ATCC 824. Gene. 134:107–111.

Warnes G.R., Bolker B., Lumley T. 2012. gplots: Various R programming tools for plotting data. R package version 2.6.0. Available from http://cran.r-project.org/web/packages/gplots/.

Whidden C., Beiko R.G., Zeh N. 2010. Fast FPT algorithms for computing rooted agreement forests: theory and experiments. Exp. Algorithms, 9th Int. Symp. .

Wiens J.J., Kuczynski C.A., Smith S.A., Mulcahy D.G., Sites Jr. J.W., Townsend T.M., Reeder T.W. 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst Biol. 57:420–431.

Wilson K.H. 1983. Efficiency of various bile salt preparations for stimulation of Clostridium difficile spore germination. J Clin Microbiol. 18:1017–1019.

Yatsunenko T., Rey F.E., Manary M.J., Trehan I., Dominguez-Bello M.G., Contreras M., Magris M., Hidalgo G., Baldassano R.N., Anokhin A.P., Heath A.C., Warner B., Reeder J., Kuczynski J., Caporaso J.G., Lozupone C.A., Lauber C., Clemente J.C.,

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Knights D., Knight R., Gordon J.I. 2012. Human gut microbiome viewed across age and geography. Nature. 486:222–227.

Zdobnov E.M., Apweiler R. 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17:847–848.

Zeng A.P., Ross A., Biebl H., Tag C., Gunzel B., Deckwer W.D. 1994. Multiple product inhibition and growth modeling of clostridium butyricum and klebsiella pneumoniae in glycerol fermentation. Biotechnol Bioeng. 44:902–911.

Figure Captions

Figure 1 - Environmental distribution of the Lachnospiraceae.

A total of 25 16S rRNA gene surveys containing a total of 1,697 samples covering 17 different habitat classes were taxonomically profiled to identify the overall percentage of Lachnospiraceae. Boxplots outline the 25th, 50th and 75th percentiles of the data. The minimum, maximum and average (red box) percent abundance per sample of this family are also indicated. The number of samples per environment is listed beside habitat type and in Supplementary Table 1. Each GI tract-associated habitat is highlighted in bold.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Figure 2 - Grouping of genomes based upon counts of shared gene clusters.

Heatmap showing the number of gene clusters shared between genomes, inversely weighted by genome size. Genomes are clustered with intersecting cells between two genomes colored based on similarity ranging from low (red) to high (blue). The hierarchy of clustering is displayed along the side and top of the heat map with branches colored according to habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract). Names of gut-associated members predicted to be lacking butyric acid production are highlighted by an asterisk.

Figure 3. Distribution of sporulation-associated genes within Lachnospiraceae genomes.

A range of sporulation genes was examined for each genome to assess the capabilities of producing endospores within each strain. Each gene is displayed as present (green) or absent (white) from each Lachnospiraceae genome. Organisms are clustered based upon their distribution of sporulation genes. Hierarchical clustering of genomes is displayed at the top of the grid with branches colored according to habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract). Gray lines separate sporulation genes into the broad categories listed on the right-hand side.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Figure 4. Relationships of 30 Lachnospiraceae genomes based on marker-gene and concatenated alignments.

Phylogenetic trees based upon the 16S ribosomal RNA gene (A) and the family-wide shared orthologs (B). Trees are rooted using two Ruminococcaceae as outgroup. Branches are colored based upon listed habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract). Bootstrap support values greater than 0.5 are displayed. Locations of putative gain and loss of functions are also shown on the trees. Stars mark the gain of butyric acid production capabilities (pink = butyrate kinase; orange = butyryl-CoA:acetate CoA-transferase). An alternative gain of butyrate kinase is marked with a pink X on the 16S tree (part A). Putative loss of sporulation capabilities is marked with a black bar. Strains classified as gut-restricted based upon shared gene clusters are underlined.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Table 1. The distribution of butyric acid production genes.

The final stage of butyric acid production can be undertaken by 2 gene groups: butyrate kinase or butyryl-CoA:acetate CoA-transferase. The presence of each gene within a Lachnospiraceae genome is marked with a +.

Name	Butyrate Kinase	BCoAT
Anaerostipes caccae DSM 14662		+
Anaerostipes sp. 3_2_56FAA		+
Butyrivibrio crossotus DSM 2876	+	
Butyrivibrio proteoclasticus B316	+	
Catonella morbi ATCC 51271		
Cellulosilyticum lentocellum DSM 5427		
Coprococcus comes ATCC 27758	+	
Coprococcus eutactus ATCC 27759	+	
Dorea formicigenerans ATCC 27755		
Dorea longicatena DSM 13814		
Lachnospiraceae bacterium 1_1_57FAA		
Lachnospiraceae bacterium 1_4_56FAA	+	
Lachnospiraceae bacterium 2_1_46FAA		
Lachnospiraceae bacterium 2_1_58FAA		
Lachnospiraceae bacterium 3_1_46FAA		
Lachnospiraceae bacterium 3_1_57FAA_CT1	+	
Lachnospiraceae bacterium 4_1_37FAA		
Lachnospiraceae bacterium 5_1_57FAA		
Lachnospiraceae bacterium 5_1_63FAA		+
Lachnospiraceae bacterium 6_1_63FAA		
Lachnospiraceae bacterium 8_1_57FAA		
Lachnospiraceae bacterium 9_1_43BFAA		
Lachnospiraceae oral taxon 107 str. F0167		
Marvinbryantia formatexigens DSM 14469		
Oribacterium sinus F0268		
Oribacterium sp. oral taxon 078 str. F0262		
Oribacterium sp. oral taxon 108 str. F0425		
Roseburia intestinalis L1-82		+
Roseburia inulinivorans DSM 16841		+
Shuttleworthia satelles DSM 14600	+	

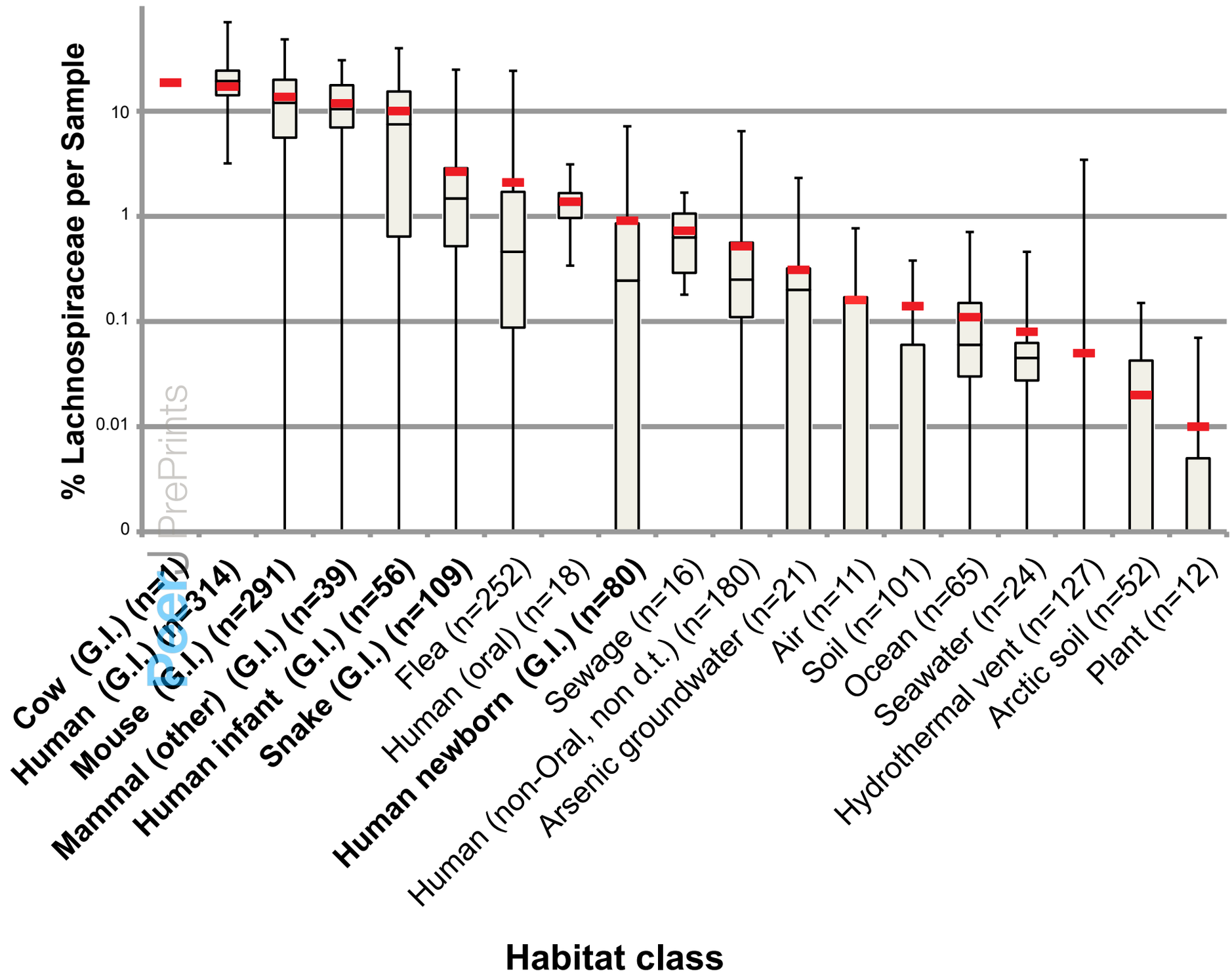


Figure 1 - Environmental distribution of the Lachnospiraceae.
A total of 25 16S rRNA gene surveys containing a total of 1,697 samples covering 17 different habitat classes were taxonomically profiled to identify the overall percentage of Lachnospiraceae. Boxplots outline the 25th, 50th and 75th percentiles of the data. The minimum, maximum and average (red box) percent abundance per sample of this family are also indicated. The number of samples per environment is listed beside habitat type and in Supplementary Table 1. Each GI tract-associated habitat is highlighted in bold.

PeerJ PrePrints | https://peerj.com/preprints/685v1 | Received: 21 Dec 2012, Published: 28 Dec 2012, doi: 10.7287/peerj.preprints.685v1

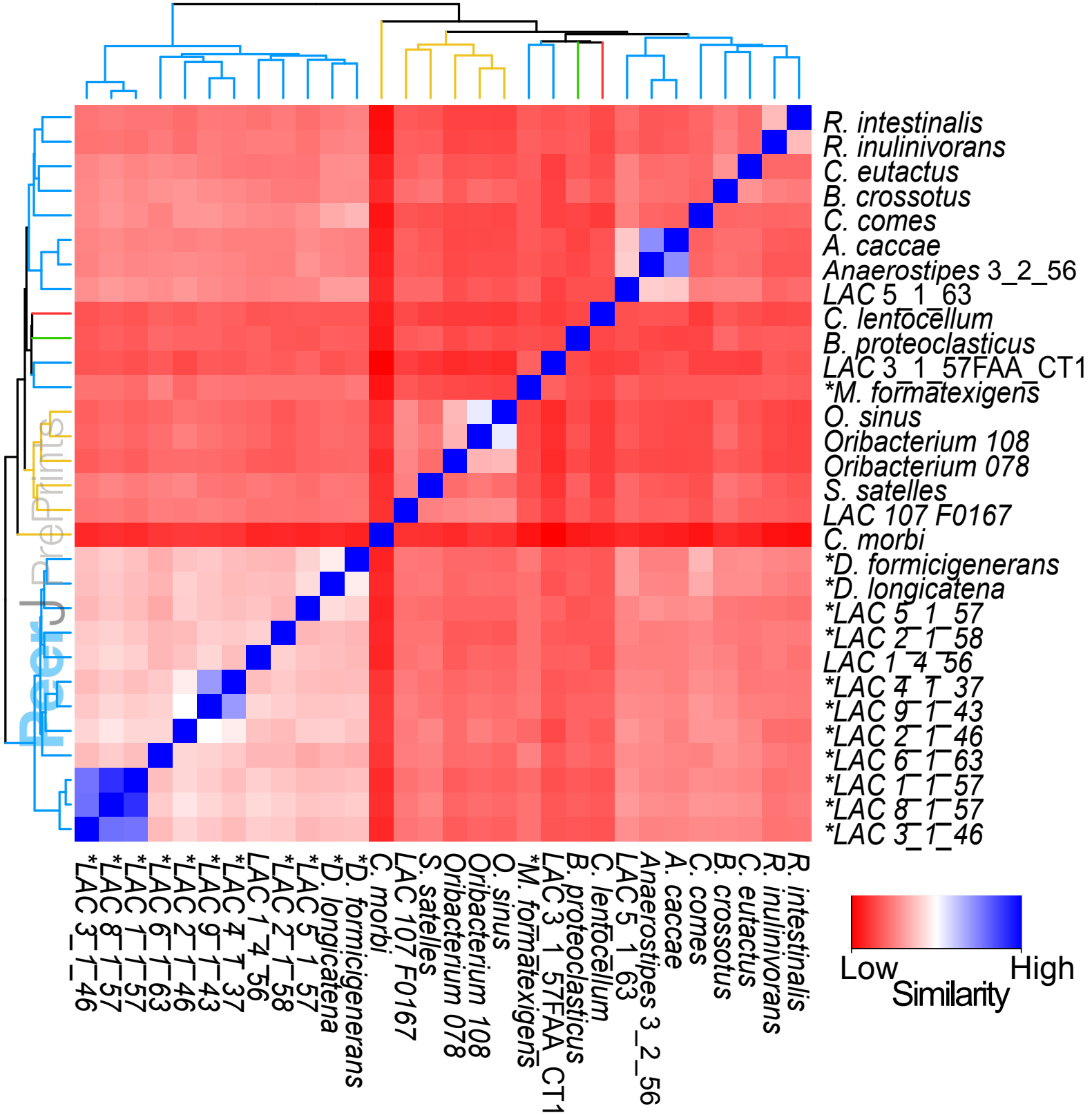


Figure 2 - Grouping of genomes based upon counts of shared gene clusters. Heatmap showing the number of gene clusters shared between genomes, inversely weighted by genome size. Genomes are clustered with intersecting cells between two genomes colored based on similarity ranging from low (red) to high (blue). The hierarchy of clustering is displayed along the side and top of the heat map with branches colored according to habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract). Names of gut-associated members predicted to be lacking butyric acid production are highlighted by an asterisk.

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 | received: 23 Dec 2013 | published: 23 Dec 2013 | doi: 10.7287/peerj.preprints.168v1

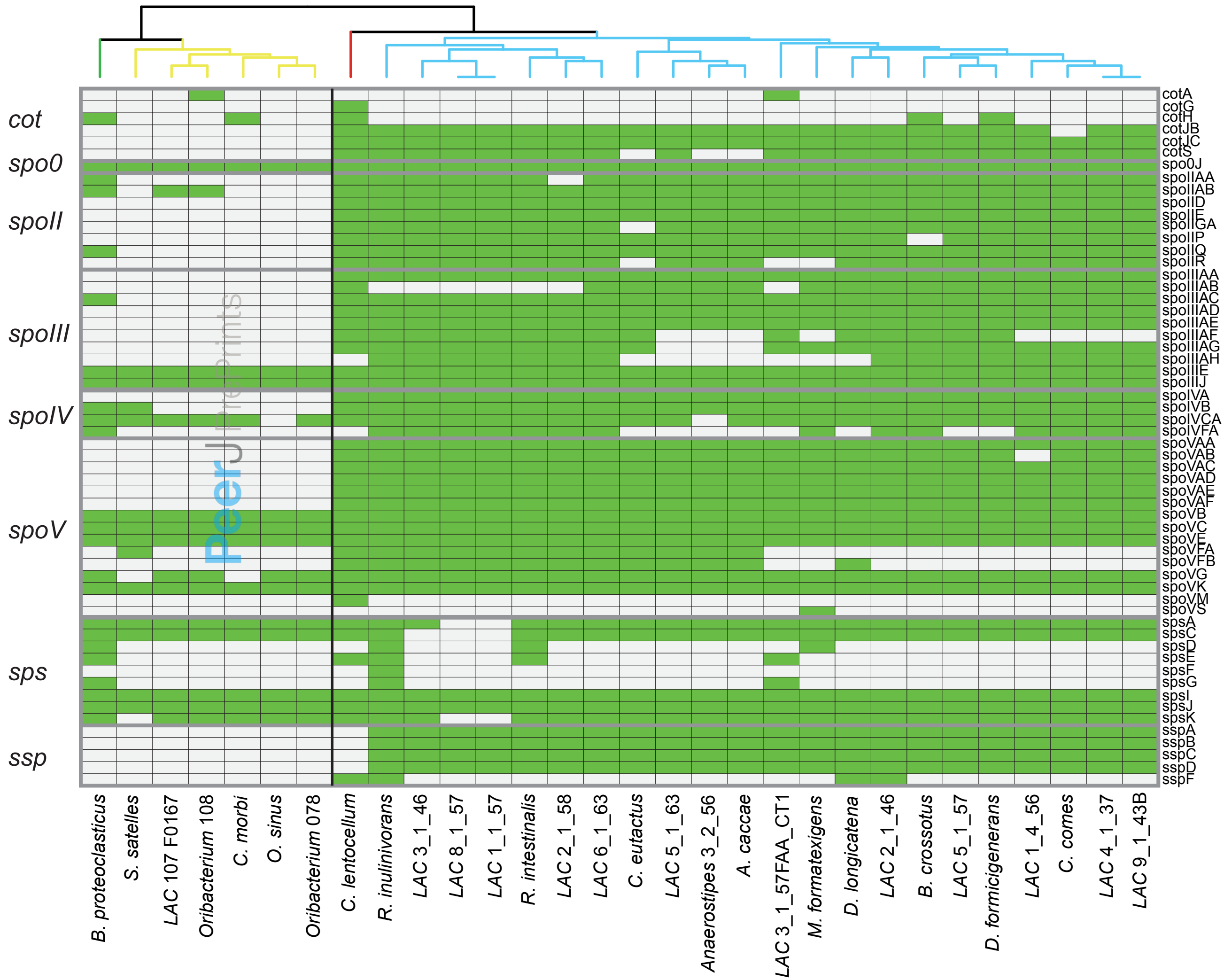


Figure 3. Distribution of sporulation-associated genes within Lachnospiraceae genomes.
A range of sporulation genes was examined for each genome to assess the capabilities of producing endospores within each strain. Each gene is displayed as present (green) or absent (white) from each Lachnospiraceae genome. Organisms are clustered based upon their distribution of sporulation genes. Hierarchical clustering of genomes is displayed at the top of the grid with branches colored according to habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract). Gray lines separate sporulation genes into the broad categories listed on the right-hand side.

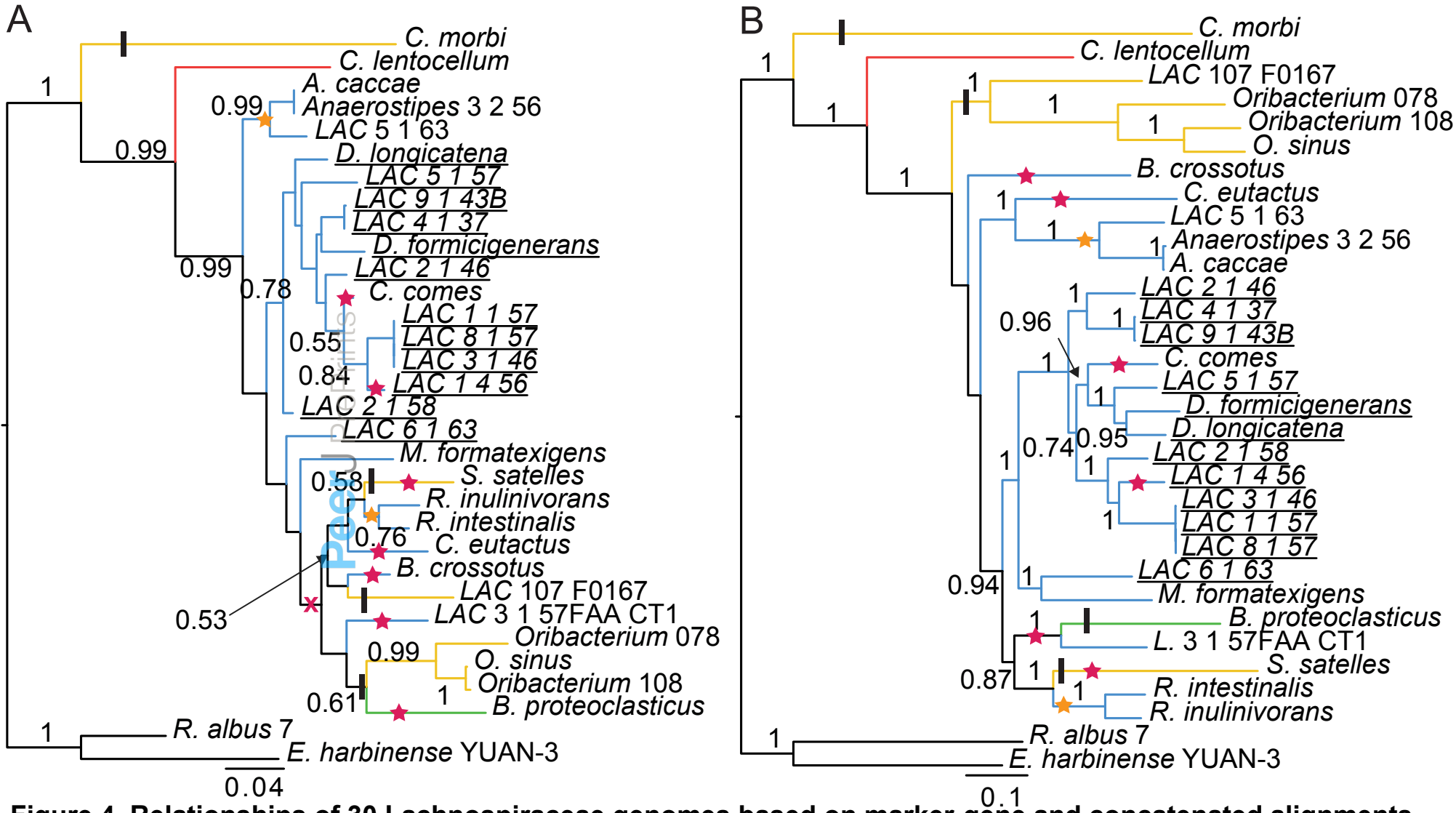
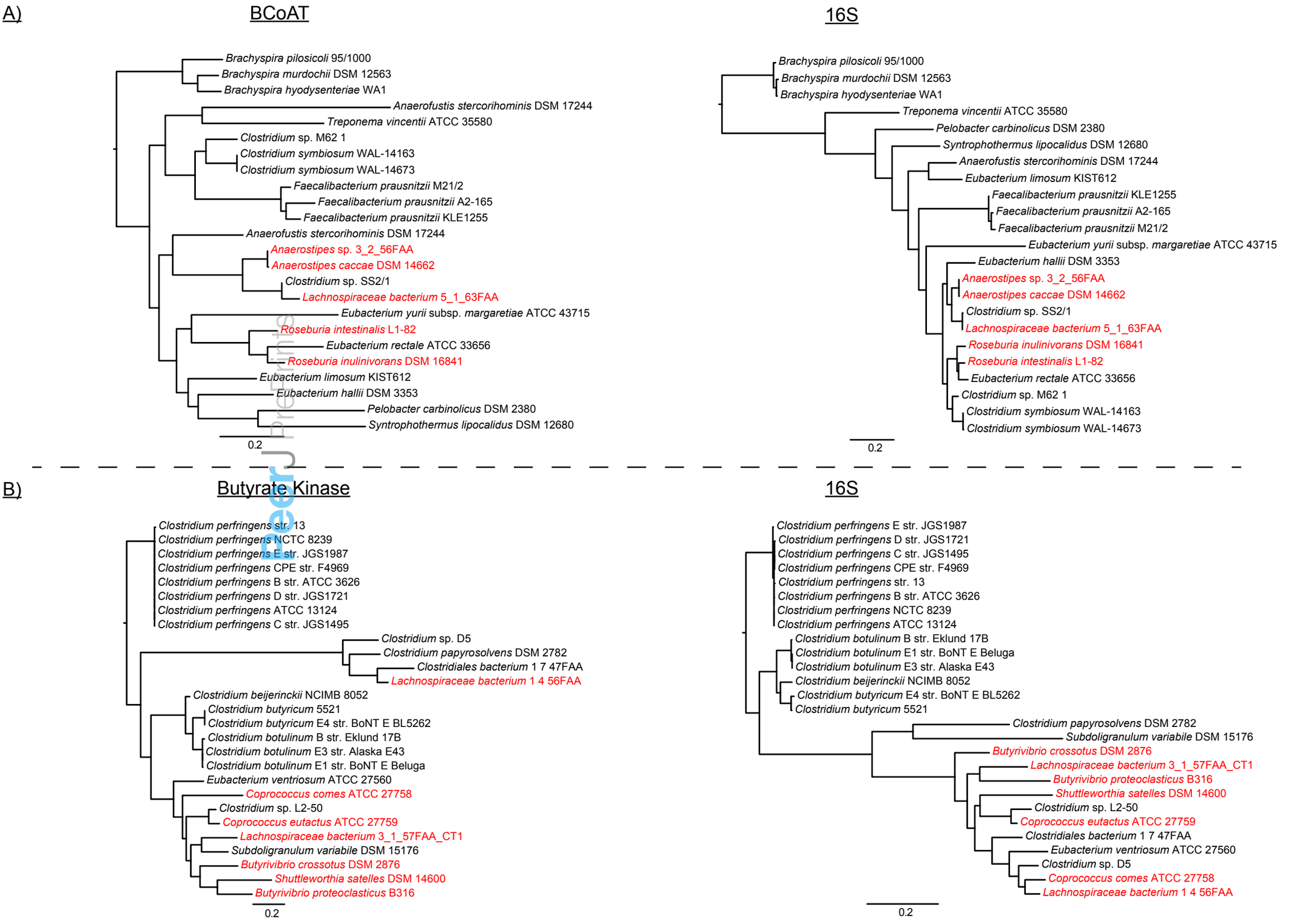
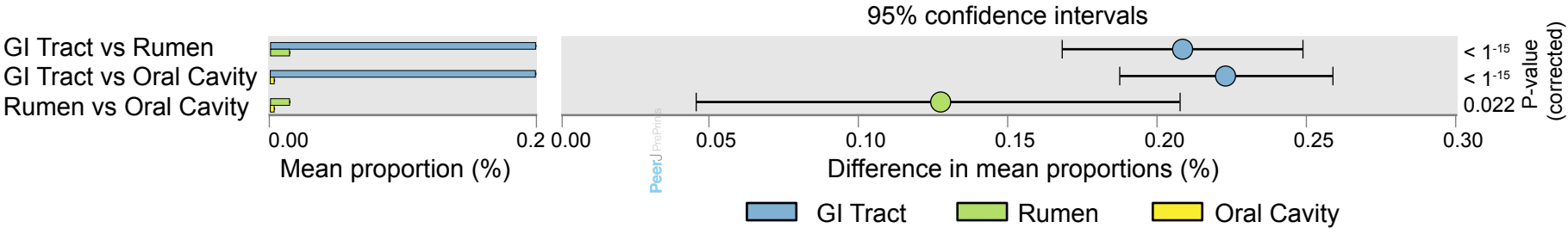


Figure 4. Relationships of 30 Lachnospiraceae genomes based on marker-gene and concatenated alignments. Phylogenetic trees based upon the 16S ribosomal RNA gene (A) and the family-wide shared orthologs (B). Trees are rooted using two Ruminococcaceae as outgroup. Branches are colored based upon listed habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract). Bootstrap support values greater than 0.5 are displayed. Locations of putative gain and loss of functions are also shown on the trees. Stars mark the gain of butyric acid production capabilities (pink = butyrate kinase; orange = BCoAT). An alternative gain of butyrate kinase is marked with a pink X on the 16S tree (part A). Putative loss of sporulation capabilities is marked with a black bar. Strains classified as gut-restricted based upon shared gene clusters are underlined.

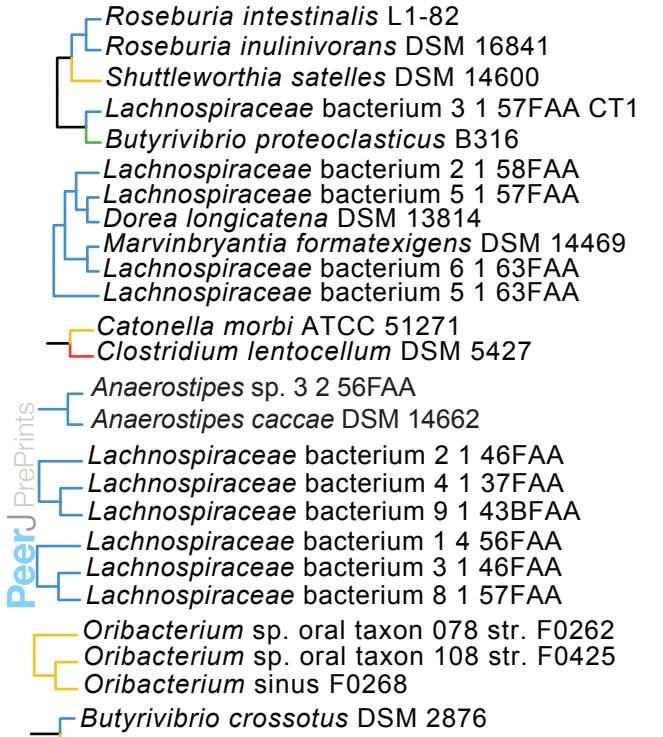


Supplementary figure 1 - Phylogenetic analysis of Lachnospiraceae-associated genes involved in the production of butyric acid and their associated 16S phylogenies.

The genes for butyryl-CoA:acetate CoA-transferase (A) and butyrate kinase (B) within Lachnospiraceae genomes were compared to 3,500 other prokaryotic genomes to find sources of potential LGT of these functions. Individual phylogenies were built using 16S sequences from genomes found to have the relevant butyrate-related gene and are displayed beside the BCoAT (A) and butyrate kinase (B) phylogenies. Lachnospiraceae members are highlighted in red.



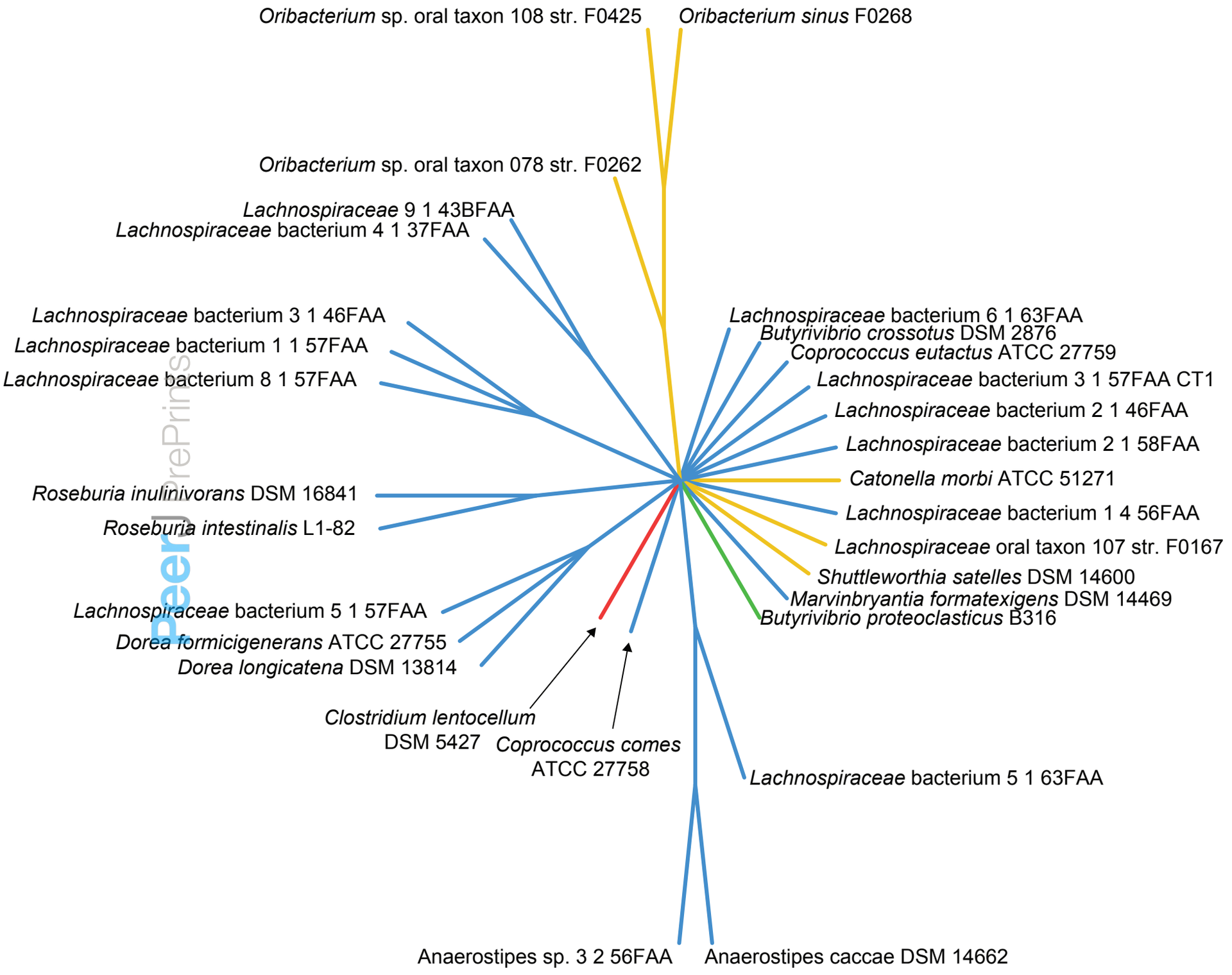
Supplementary figure 2- Sporulation-related sequences that differ in abundance between the human GI tract microbiome, and the cow rumen and human oral cavity. The abundance of reads assigned to Lachnospiraceae-associated sporulation genes within metagenomic samples from the human gut microbiome were compared to those within the cow rumen and the human oral cavity using STAMP. This revealed several genes that were more abundant in the human GI tract (blue) compared to the rumen (green) or oral cavity (yellow). The mean proportions of assigned reads within each dataset are shown in addition to the difference of these proportions between datasets. The p-value from the Bonferroni-corrected two-sided Welch's t-test is shown for each comparison.



PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Supplementary figure 3 - Maximum agreement forest between the 16S and shared gene cluster phylogenetic trees.

SPR operations were used to assess the congruence of phylogenetic trees based upon the 16S gene and the shared gene clusters of all analyzed genomes. The maximum agreement forest displays components that are in present in both trees. Branches are colored based upon listed habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract).



Supplementary figure 4- Phylogenetic network of shared gene clusters based upon individual gene tree topologies.

The gene trees of 91 family-wide shared gene clusters were input to SplitsTree4 to construct an unrooted phylogenetic network that best represented all the individual relationships. Most gene trees were found to disagree, resulting in a star-like topology. Branch coloring is based upon listed habitat (yellow = oral; red = sediment; green = rumen; blue = human GI tract).

Supplementary table 1 - Metagenomic samples utilized for environmental distribution analysis.
Multiple habitat types were tested for the presence of Lachnospiraceae. Each habitat type is listed along with the MG-RAST ID of the project samples were retrieved from and the number of sample obtained from each project.

Habitat	Project Id	Sample Count
Air	74	11
Arsenic groundwater	70	21
Cow (G.I.)	504	1
Flea	75	252
Human (G.I.)	66	281
	133	18
	81	17
Human (non-Oral, non G.I.)	81	180
Human (oral)	81	18
Human infant (G.I.)	65	56
Human newborn (G.I.)	79	80
Hydrothermal vent	327	127
Mammal (other) (G.I.)	114	39
Mouse (G.I.)	245	181
	157	172
	83	38
Ocean	56	12
	57	12
	189	72
Plant	72	12
Sewage	122	16
Snake (G.I.)	77	109
Soil	67	26
	69	52
	71	48
	80	27

PeerJ PrePrints | https://peerj.com/preprints/168v1/ | v1 received: 23 Dec 2013, published: 23 Dec 2013, doi: 10.7287/peerj.preprints.168v1

Supplementary table 2 - Lachnospiraceae genomes.
The designation, abbreviation used in this manuscript, NCBI taxon identification number, associated habitat according to IMG and source of for each genome utilized in this study is listed.

Name	Abbreviated name	NCBI ID	Habitat	Reference
Anaerostipes caccae DSM 14662	*A. caccae*	411490	Human Digestive Tract	(S1)
Anaerostipes sp. 3_2_56FAA	*Anaerostipes 3_2_56*	665937	Human Digestive Tract	(S2)
Butyrivibrio crossotus DSM 2876	*B. crossotus*	511680	Human Digestive Tract	(S3)
Butyrivibrio proteoclasticus B316	*B. proteoclasticus*	515622	Cow Rumen	(S4)
Catonella morbi ATCC 51271	*C. morbi*	592026	Human Oral Cavity	(S5)
Cellulosilyticum lentocellum DSM 5427	*C. lentocellum*	642492	Estuarine mud bank	(S6)
Coprococcus comes ATCC 27758	*C. comes*	470146	Human Digestive Tract	(S7)
Coprococcus eutactus ATCC 27759	*C. eutactus*	411474	Human Digestive Tract	(S8)
Dorea formicigenerans ATCC 27755	*D. formicigenerans*	411461	Human Digestive Tract	(S9)
Dorea longicatena DSM 13814	*D. longicatena*	411462	Human Digestive Tract	(S10)
Lachnospiraceae bacterium 1_1_57FAA	*LAC* 1_1_57	658081	Human Digestive Tract	(S11)
Lachnospiraceae bacterium 1_4_56FAA	*LAC* 1_4_56	658655	Human Digestive Tract	(S12)
Lachnospiraceae bacterium 2_1_46FAA	*LAC* 2_1_46	742723	Human Digestive Tract	(S13)
Lachnospiraceae bacterium 2_1_58FAA	*LAC* 2_1_58	658082	Human Digestive Tract	(S14)
Lachnospiraceae bacterium 3_1_46FAA	*LAC* 3_1_46	665950	Human Digestive Tract	(S15)
Lachnospiraceae bacterium 3_1_57FAA_CT1	*LAC* 3_1_57FAA_CT1	658086	Human Digestive Tract	(S16)
Lachnospiraceae bacterium 4_1_37FAA	*LAC* 4_1_37	552395	Human Digestive Tract	(S17)
Lachnospiraceae bacterium 5_1_57FAA	*LAC* 5_1_57	658085	Human Digestive Tract	(S18)
Lachnospiraceae bacterium 5_1_63FAA	*LAC* 5_1_63	658089	Human Digestive Tract	(S19)
Lachnospiraceae bacterium 6_1_63FAA	*LAC* 6_1_63	658083	Human Digestive Tract	(S20)
Lachnospiraceae bacterium 8_1_57FAA	*LAC* 8_1_57	665951	Human Digestive Tract	(S21)
Lachnospiraceae bacterium 9_1_43BFAA	*LAC* 9_1_43B	658088	Human Digestive Tract	(S22)
Lachnospiraceae oral taxon 107 str. F0167	*LAC 107 F0167*	575593	Human Oral Cavity	(S23)
Marvinbryantia formatexigens DSM 14469	*M. formatexigens*	478749	Human Digestive Tract	(S24)
Oribacterium sinus F0268	*O. sinus*	585501	Human Oral Cavity	(S25)
Oribacterium sp. oral taxon 078 str. F0262	*Oribacterium 078*	608534	Human Oral Cavity	(S26)
Oribacterium sp. oral taxon 108 str. F0425	*Oribacterium 108*	904296	Human Oral Cavity	(S27)
Roseburia intestinalis L1-82	*R. intestinalis*	536231	Human Digestive Tract	(S28)
Roseburia inulinivorans DSM 16841	*R. inulinirans*	622312	Human Digestive Tract	(S29)
Shuttleworthia satelles DSM 14600	*S. satelles*	626523	Human Oral Cavity	(S30)

References for Supplementary table 1

(S1) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/?term=txid411490)
(S2) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/13727?project_id=61867)
(S3) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2083?project_id=55091)
(S4) Kelly WJ, et al. (2010) The glycobiome of the rumen bacterium Butyrivibrio proteoclasticus B316(T) highlights adaptation to a polysaccharide-rich environment. PloS one 5(8):e11942.
(S5) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/1946?project_id=55757)
(S6) Miller DA, et al. (2011) Complete genome sequence of the cellulose-degrading bacterium Cellulosilyticum lentocellum. J Bacteriol 193(9):2357-2358
(S7) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/967?project_id=54883)
(S8) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/996?project_id=54541)
(S9) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2064?project_id=54513
(S10) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/985?project_id=54515)
(S11) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2341?project_id=68209)
(S12) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2342?project_id=68205)
(S13) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2855?project_id=66429)
(S14) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2343?project_id=68203)
(S15) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2371?project_id=66427
(S16) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2344?project_id=68201)
(S17) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2526?project_id=63581)
(S18) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2346?project_id=68199)
(S19) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2347?project_id=61883)
(S20) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2350?project_id=66423)
(S21) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2372?project_id=61885)
(S22) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2352?project_id=66425)
(S23) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2556?project_id=66385)
(S24) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/957?project_id=54943)
(S25) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/13439?project_id=55891)
(S26) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/13438?project_id=55773)
(S27) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/13438?project_id=67819)
(S28) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2047?project_id=55267)
(S29) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/2081?project_id=55375)
(S30) Unpublished (see http://www.ncbi.nlm.nih.gov/genome/1952?project_id=55775)

Supplementary table 3 - Functions characterizing sub-groups of Lachnospiracoae.
Gene clusters present in over 90% of one group of Lachnospiracoae genomes and absent in over 90% of another were analyzed using Interproscan to determine their functions, as was the reverse. The general InterPro functional categories and GO

Supplementary table 3a. Gut-restricted Lachnospiracoae compared to all other Lachnospiracoae

Gut-restricted-associated genome count	Non-Gut-restricted genome count	InterPro	GO
11	0	Tetratricopeptide-like helical	Protein binding
		none	None
		ToiB-like	Protein binding
		Tetratricopeptide-like helical	Protein binding
		None	None
11	1	Signal transduction histidine kinase	Signal Transduction
		None	none
		Peptidoglycan-binding Lysin subgroup	Cell wall macromolecule catabolic process
		Protein phosphatase 2C-like	Catalytic activity
		Periplasmic binding protein domain	None
		YbbR-like	None
		Type II secretion system F domain	None
		None	None
		None	None
		Colicin V production, CvpA	Toxin biosynthetic process
		Tetratricopeptide-like helical	Protein binding
		Aminoglycoside phosphotransferase	Transferring phosphorus-containing groups
		Tetratricopeptide-like helical	Protein binding
		None	None
		None	None
		None	None
		None	None
		Haemerythrin/HHE cation-binding motif	Metal ion binding
		none	None
		Spore cortex biosynthesis protein, YabQ-like	None
		none	None
11	2	Periplasmic binding protein domain	None
		Permease FtsX-like	None
		Sporulation stage III, protein AE	None
		Sporulation stage II protein D, amidase enhancer LytB	Sporulation resulting in formation of a cellular spore
		Integral membrane protein 1906	None
		Vitamin B12-dependent methionine synthase, activation domain	Methionine synthase activity
12	0	None	None
12	1	Bacterial periplasmic spermidine/putrescine-binding protein	Transporter activity
		Prokaryotic chromosome segregation/condensation protein MukB, N-terminal	Chromosome segregation
		Spore coat protein CotS	Transferase activity, transferring phosphorus-containing groups
		Nucleoside recognition Gate	Nucleoside binding
12	2	Peptidyl-prolyl cis-trans isomerase, PpiC-type	Isomerase activity
		Signal transduction histidine kinase	Signal transduction
		None	None
0	16	No clusters	No clusters
0	17	No clusters	No clusters
0	18	No clusters	No clusters
1	16	Binding-protein-dependent transport systems inner membrane component	Transport activity
1	17	No clusters	No clusters
1	18	No clusters	No clusters

Supplementary table 3b. Gut-restricted Lachnospiracoae compared to all other gut-associated Lachnospiracoae

Gut-restricted genome count	Other gut-associated genome count	InterPro	GO
11	0	Alcohol dehydrogenase, iron-type	Oxidoreductase activity
		Aminoglycoside phosphotransferase	Transferase activity, transferring phosphorus-containing groups
		Tetratricopeptide repeat	Protein binding
		Six-bladed beta-propeller, ToiB-like	None
		Tetratricopeptide repeat	Protein binding
		None	None
		None	None
		None	None
11	1	Periplasmic binding protein domain	None
		Vacuolating cytotoxin	Pathogenesis
		Integral membrane protein 1906	None
		Signal transduction histidine kinase	Phosphorelay sensor kinase activity
		Peptidoglycan-binding lysin domain	Cell wall macromolecule catabolic process
		Protein phosphatase 2C (PP2C)-like	Catalytic activity
		Periplasmic binding protein domain	None
		YbbR-like	None
		Type II secretion system F domain	None
		Colicin V production, CvpA	Toxin biosynthetic process
		Tetratricopeptide repeat	Protein binding
		Tetratricopeptide repeat	Protein binding
		Haemerythrin-like, metal-binding domain	Metal ion binding
		Spore cortex biosynthesis protein, YabQ-like	None
		None	None
		None	None
		None	None
		None	None
		None	None
		None	None
		None	None
12	0	None	None
12	1	Bacterial periplasmic spermidine/putrescine-binding protein	Polyamine transport
		Spore coat protein CotS	None
		Nucleoside recognition Gate	Nucleoside binding
		None	None
0	9	Electron transfer flavoprotein, alpha subunit	Electron carrier activity
		Acyl-CoA oxidase/dehydrogenase	Acyl-CoA dehydrogenase activity
		Thiolase	Transferase activity
		Enoyl-CoA hydratase/isomerase, conserved site	Catalytic activity
0	10	No clusters	No clusters
1	9	Nitrogen regulatory protein PII	Regulation of nitrogen utilization
		Binding-protein-dependent transport systems inner membrane component	Transporter activity
		NUDIX hydrolase domain	Hydrolase activity
		Nitroreductase-like	Oxidoreductase activity
1	10	No clusters	No clusters

Supplementary table 3c. Functions associated with Lachnospiracoae within the human GI tract that can produce butyric acid compared to those lacking this capability

Gut-associated butyric acid producing genome count	Gut-associated non-butyric acid producing genome count	InterPro	GO
9	0	Electron transfer flavoprotein, alpha subunit	Flavin adenine dinucleotide binding
		Acyl-CoA dehydrogenase, conserved site	Oxidation-reduction process
		Thiolase	Transferase activity
		Crotonase superfamily	Metabolic process
9	1	Nitrogen regulatory protein PII	Enzyme regulator activity
10	0	No clusters	No clusters
10	1	No clusters	No clusters
0	11	No clusters	No clusters
0	12	No clusters	No clusters
1	11	Vacuolating cytotoxin	Pathogenesis
		Protein phosphatase 2C (PP2C)-like	Catalytic activity
		Haemerythrin-like, metal-binding domain	Metal ion binding
		None	None
1	12	Protein kinase-like domain	Transferase activity

Supplementary table 3d. All gut associated Lachnospiracoae compared to Lachnospiracoae from other habitats

Gut-associated genome count	Non-gut genome count	InterPro	GO
20	0	Sulfatase	Sulfuric ester hydrolase activity
		Replication protein, DnaD/DnaB-domain	None
		Phosphoglycerate/bisphosphoglycerate mutase	Catalytic activity
		Calycin-like	None
		Sporulation protein YtmC/YmaH	None
20	1	Signal transduction histidine kinase	Signal transduction
		Signal transduction response regulator, receiver domain	Two-component signal transduction system (phosphorelay)
		Peptidase S8/S53, subtilisin/kexin/sedolisin	Serine-type endopeptidase activity
		Phospholipid/glycerol acyltransferase	Phospholipid biosynthetic process
		Spore coat assembly protein CotJB	None
		Folylpolyglutamate synthetase	Folic acid-containing compound biosynthetic process
		Phosphoribosyl-ATP pyrophosphohydrolase	Phosphoribosyl-AMP cyclohydrolase activity
		Uncharacterised protein family UPF0348	Catalytic activity
		PBP domain	None
		Nucleoside recognition Gate	Nucleoside binding
21	0	Heat shock protein DnaJ	Heat shock protein binding
		None	None
		Peptidase S11, D-alanyl-D-alanine carboxypeptidase A	Serine-type D-Ala-D-Ala carboxypeptidase activity
		Stage III sporulation protein AH	None
		AH-like	None
		FMN-binding	FMN binding
		Protein of unknown function DUF3792, transmembrane	None
		Stage III sporulation protein AC/AD family	None
		Sporulation protein YabP/YqfC	None
		Stage III sporulation protein AC	None
21	1	Multi antimicrobial extrusion protein	Drug transmembrane transporter activity
		PemK-like protein	DNA binding
		Transcription regulator HTH, GntR	Sequence-specific DNA binding transcription factor activity
		Catalase, manganese	Transition metal ion binding
		Small acid-soluble spore protein, alpha/beta-type	DNA topological change
		DNA helicase, UvrD/REP type	ATP-dependent DNA helicase activity
		Penicillin-binding protein, transpeptidase	Peptidoglycan-based cell wall biogenesis
		Peptidase S11, D-alanyl-D-alanine carboxypeptidase A	Serine-type D-Ala-D-Ala carboxypeptidase activity
		Sporulation stage III, protein AA	Nucleoside-triphosphatase activity
		Peptidase A25, germination protease	Spore germination
		Endodeoxyribonuclease IV	DNA repair
		ClpP/TepA	Serine-type endopeptidase activity
		Nucleoside recognition Gate	Nucleoside binding
		NIF system FeS cluster assembly, NifU, N-terminal	Iron-sulfur cluster assembly
		STAS domain	Regulation of transcription, DNA-dependent
		Sporulation protein YabP	None
22	0	Transposon-encoded protein TnpV	None
		Sporulation stage II, protein P	Protein binding
		RNA-binding S4 domain	RNA binding
		PhoU	None
		Stage V sporulation protein AA	None
22	1	RNA polymerase sigma-70 factor	Regulation of transcription, DNA-dependent
		RNA polymerase sigma-70 factor	Regulation of transcription, DNA-dependent
		Primosome PriB/single-strand DNA-binding	Single-stranded DNA binding
		Peptidase S55, sporulation stage IV, protein B	Protein binding
		Sporulation stage II protein D, amidase enhancer LytB	Metabolic process
		Stage V sporulation AD	Catalytic activity
		Cell wall hydrolase/autolysin, catalytic	Peptidoglycan catabolic process
		Sporulation stage V, protein T	None
		Protein of unknown function DUF177	None
		Sporulation stage V, protein AC	None
		Anti-sigma F factor	Protein serine/threonine kinase activity
		Sporulation stage V, protein AE	None

Supplementary table 4 - Shared features between members of the human gut Lachnospiraceae listed as gut-restricted/non-gut-restricted and those either possessing or lacking the capability to produce butyric acid.

Lachnospiraceae residing in the human GI tract were classified in 2 ways: those classed as gut-restricted or not based upon shared gene clusters (Fig. 2) and those classed based upon their capability to produce butyric acid or not (Table 1). Overlap of species assigned as one or the other within each classification was identified, as were functions associated with each classification.

Gut-restricted	Butyric acid-producing	Species	Associated functions
No	Yes	*Anaerostipes 3_2_56* *A. caccae* *B. crossotus* *C. comes* *C. eutactus* *LAC 3_1_57FAA_CT1* *LAC 5_1_63* *R. intestinalis* *R. inulinivorans*	Acyl-CoA oxidase/dehydrogenase Electron transfer flavoprotein, alpha subunit Nitrogen regulatory protein PII Thiolase
No	No	*M. formatexigens*	
Yes	Yes	*LAC 1_4_56*	
Yes	No	*D. formicigenerans* *D. longicatena* *LAC 1_1_57* *LAC 2_1_46* *LAC 2_1_58* *LAC 3_1_46* *LAC 4_1_37* *LAC 5_1_57* *LAC 6_1_63* *LAC 8_1_57* *LAC 9_1_43B*	Vacuolating cytotoxin Protein phosphatase 2C (PP2C)-like Haemerythrin-like, metal-binding domain