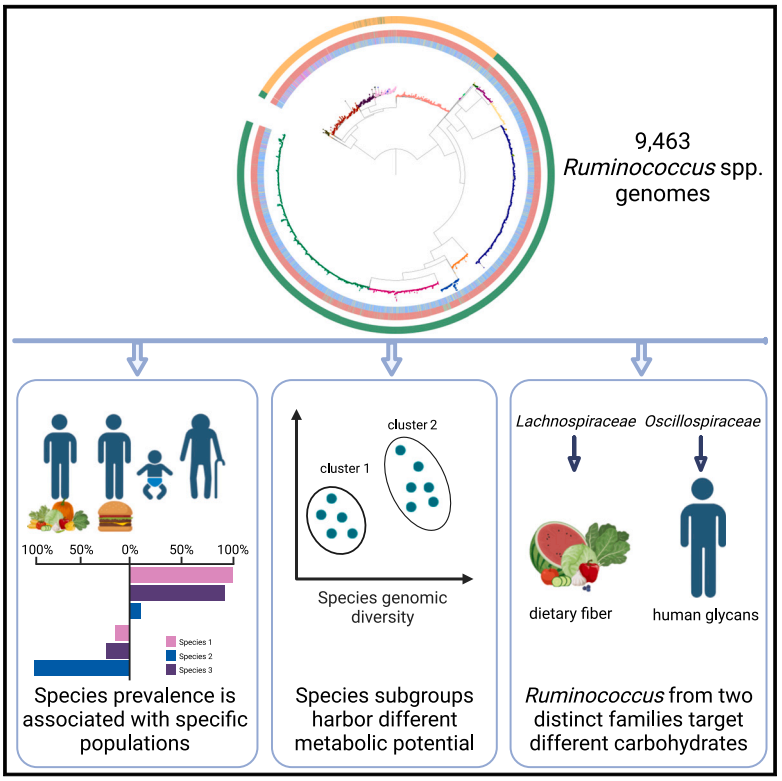
Cell Reports

Article

Genomic features and prevalence of *Ruminococcus* species in humans are associated with age, lifestyle, and disease

Graphical abstract



Highlights

- Prevalence of *Ruminococcus* species link with health, lifestyle, and age
- *Ruminococcus* spp. encode family- and species-specific CAZymes
- Subgroups of the most relevant species have specific functional traits

Authors

Vincenzo Valentino, Francesca De Filippis, Roberto Marotta, Edoardo Pasolli, Danilo Ercolini

Correspondence

ercolini@unina.it

In brief

In this study, Valentino et al. explored the diversity of the genus *Ruminococcus*, highlighting associations between species/subspecies prevalence and geographic origin, lifestyle, age, and disease. Furthermore, the authors showed that *Ruminococcus* species from families Lachnospiraceae and Oscillospiraceae harbor distinctive carbohydrate-active enzymes and engraftment potential.

Valentino et al., 2024, Cell Reports *43*, 115018
December 24, 2024
https://doi.org/10.1016/j.celrep.2024.115018

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Genomic features and prevalence of *Ruminococcus* species in humans are associated with age, lifestyle, and disease

Vincenzo Valentino,[1] Francesca De Filippis,[1,2] Roberto Marotta,[1] Edoardo Pasolli,[1] and Danilo Ercolini[1,2,3,*]
[1]Department of Agricultural Sciences, University of Naples Federico II, Piazza Carlo di Borbone 1, Portici, 80055 Naples, Italy
[2]Task Force on Microbiome Studies, University of Naples Federico II, Naples, Italy
[3]Lead contact
*Correspondence: ercolini@unina.it
https://doi.org/10.1016/j.celrep.2024.115018

SUMMARY

The genus *Ruminococcus* is dominant in the human gut, but higher levels of some species, such as *R. gnavus*, *R. torques*, and *R. bromii*, have been linked to health or disease. In this study, we analyzed >9,000 *Ruminococcus* metagenome-assembled genomes (MAGs) reconstructed from >5,000 subjects and revealed significant links between the prevalence of some species/subspecies and geographic origin, age, lifestyle, and disease, with subspecies prevalent in specific subpopulations showing divergent metabolic potential. Furthermore, *Ruminococcus* species from Lachnospiraceae encoded for carbohydrate-active enzymes (CAZy) potentially involved in the metabolism of human N- and O-glycans, whereas those from Oscillospiraceae appear to be more adapted toward fiber metabolism. These new findings contribute to elucidating the potential functional role of *Ruminococcus* in specific lifestyles and diseases and to decipher the diversity and the adaptation of members of this genus to the human gut.

INTRODUCTION

The *Ruminococcus* genus was discovered in 1948[1] with the first isolation of the cellulose-degrading *R. flavefaciens* from cattle rumen. Since then, this taxonomic group has been studied further, and isolation of novel *Ruminococcus* spp. from the human gut has demonstrated a strong association with the human host. However, despite the high prevalence of these microbes in the gut, their role in health and disease and their taxonomic classification have always been intricate and remain controversial. Indeed, the advent of phylogenetics-based analysis has led to the description of *Ruminococcus* as a polyphyletic genus[2] including two distinct clades assigned to two different families: Lachnospiraceae and Oscillospiraceae. Furthermore, a recent reclassification of several Lachnospiraceae ruminococci (LRs) into the genus *Mediterraneibacter* has been proposed.[3]

Beside the taxonomic uncertainty, both LRs and Oscillospiraceae ruminococci (ORs) are described as gram-positive, nonmotile, and strictly anaerobic cocci[4] requiring fermentable carbohydrates for growth.[5] Also, a large-scale survey of human gut microbiomes has defined the genus as keystone member of the most widespread enterotype.[6] Subsequently, multiple research efforts have been focused on this group, and studies have attempted to outline the role of some *Ruminococcus* species in health and disease. Indeed, high levels of *R. torques* and *R. gnavus* have been linked with gut inflammation,[7] inflammatory bowel disease (IBD), and irritable bowel syndrome,[8] allergy,[9] and early stages of colorectal cancer (CRC),[10] whereas *Ruminococcus bromii* has been identified as a promising next-generation probiotic thanks to its ability to hydrolyze resistant starch (RS).[11]

However, several other species are included in the *Ruminococcus* genus and detected in the human gut microbiome, although their role in health and disease has been poorly investigated. This is mainly due to their strict nutritional and environmental requirements for growth (e.g., presence of fiber and an anoxic atmosphere), which are difficult to reproduce in the laboratory.

In this work, we take advantage of a large curated collection of metagenome-assembled genomes (MAGs) reconstructed from more than 9,000 metagenomes from all over the world[12] to perform a large-scale analysis of *Ruminococcus* species in order to delve deeper into their phylogenetic diversity, metabolic potential, and distribution in the population.

RESULTS

Ruminococci are highly prevalent in the human gut microbiome

Overall, the 9,463 medium/high-quality MAGs (completeness >50% and contamination <5%) taxonomically assigned to *Ruminococcus*, *Mediterraneibacter*, or *Pseudoruminococcus* clustered into 29 clusters of genomes showing a whole-genome nucleotide similarity ≥95%, defined as species-level genome bins (SGBs; Table S1). These MAGs were reconstructed from 5,952 fecal samples donated by 5,132 subjects. Six of the 29 SGBs were assigned to Lachnospiraceae, accounting for 2,570 genomes (27.2% of the total), whereas the remaining 23 SGBs were taxonomically labeled as Oscillospiraceae and included

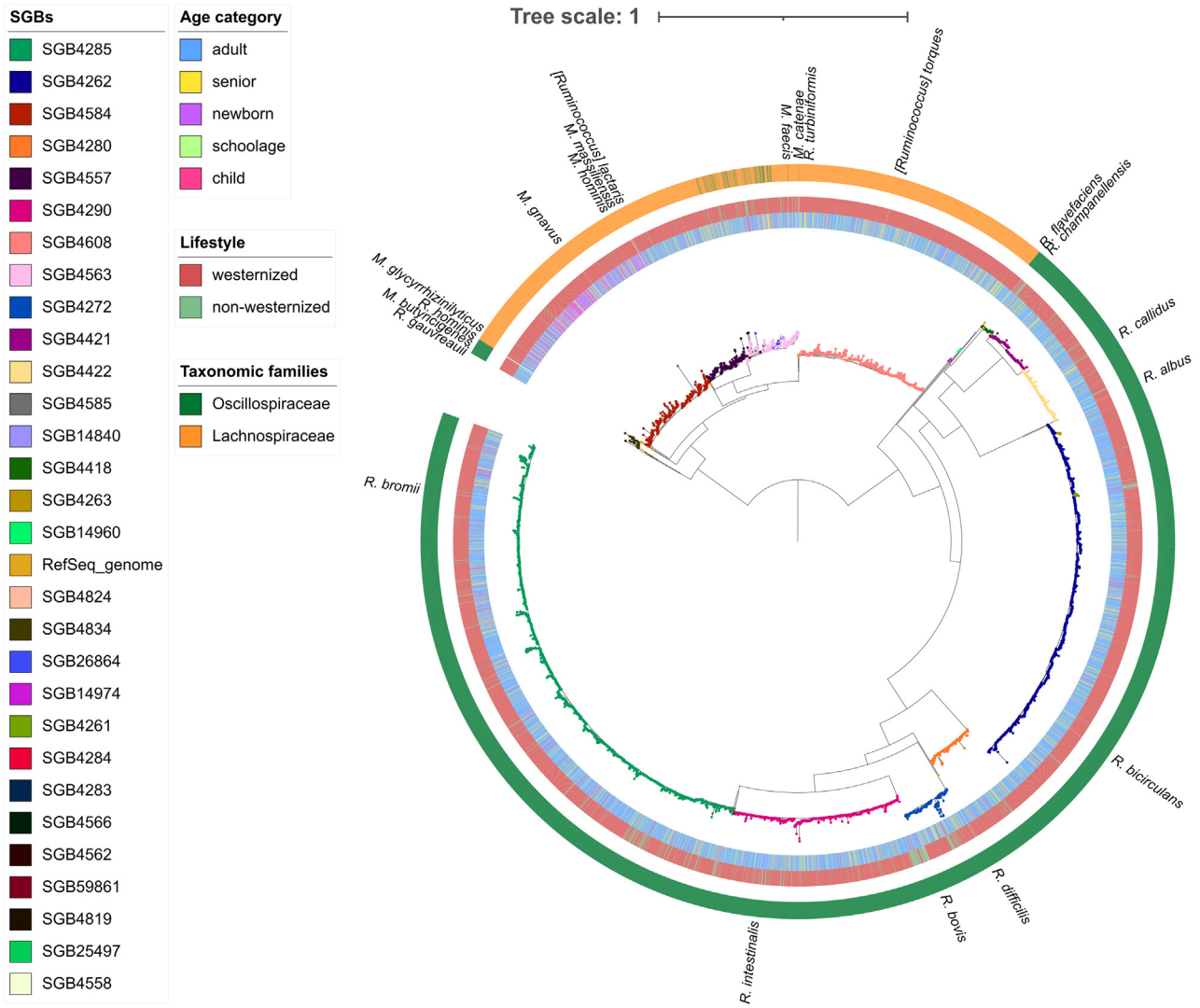


Figure 1. Phylogenetic tree of the 9,463 *Ruminococcus* and *Mediterraneibacter* spp.

Shown are MAGs and the NCBI RefSeq reference genomes obtained through the alignment of genomes against a set of 120 bacterial markers performed through the Genome Taxonomy Database Toolkit (GTDB-Tk). Nodes are color coded according to the SGB, and annotations regarding the subjects' age category (inner circle), lifestyle (middle circle), and genome family (outer circle) are shown. The labels report the approximative position where the NCBI RefSeq reference genomes fall.

6,893 genomes (Figure 1). Also, 9 Oscillospiraceae and 2 Lachnospiraceae SGBs were unknown (uSGBs; lacking genomes from public repositories[12]), comprising 402 genomes (4.25% of the total genomes; Table S1).

According to a phylogenetic analysis based on 120 bacterial marker genes, genomes from Lachnospiraceae and Oscillospiraceae formed two distinct phylogenetic clusters (except for uSGBs SGB26864 and SGB4566, taxonomically labeled as Oscillospiraceae but clustering within the Lachnospiraceae clade), with several reference genomes currently named as *Mediterraneibacter* clustering within the Lachnospiraceae clade (Figure 1), consistent with the last taxonomic classification.

About 48% of the subjects ($n = 2,462$) were European, whereas 34% ($n = 1,788$) came from Asia. The other continents were underrepresented, with 620 Americans, 182 Africans, and 80 people from Fiji islands. About 76% of the subjects ($n = 3,879$) were adults (age ranging from 19 to 65 years), 693 senior (13.5%, ≥65 years), 274 newborns (5%, <1 year of age), 202 children (~4%, age between 1 and 12 years), and 79 school aged (1.5%, age ≥12 and <19 years; Table S2). For 5 subjects, the age was not reported.

The distribution of age categories and health status between the continents was not homogeneous (χ^2 test false discovery rate [FDR] < 0.001; Table S3). In particular, a lower number of newborns and children was observed than expected in the Asian population (newborn, 0 observed and 95 expected; children, 5 observed and 70 expected). Furthermore, no people diagnosed with CRC were observed among non-Westernized (expected, 13, χ^2 test FDR = 0.011; Table S3).

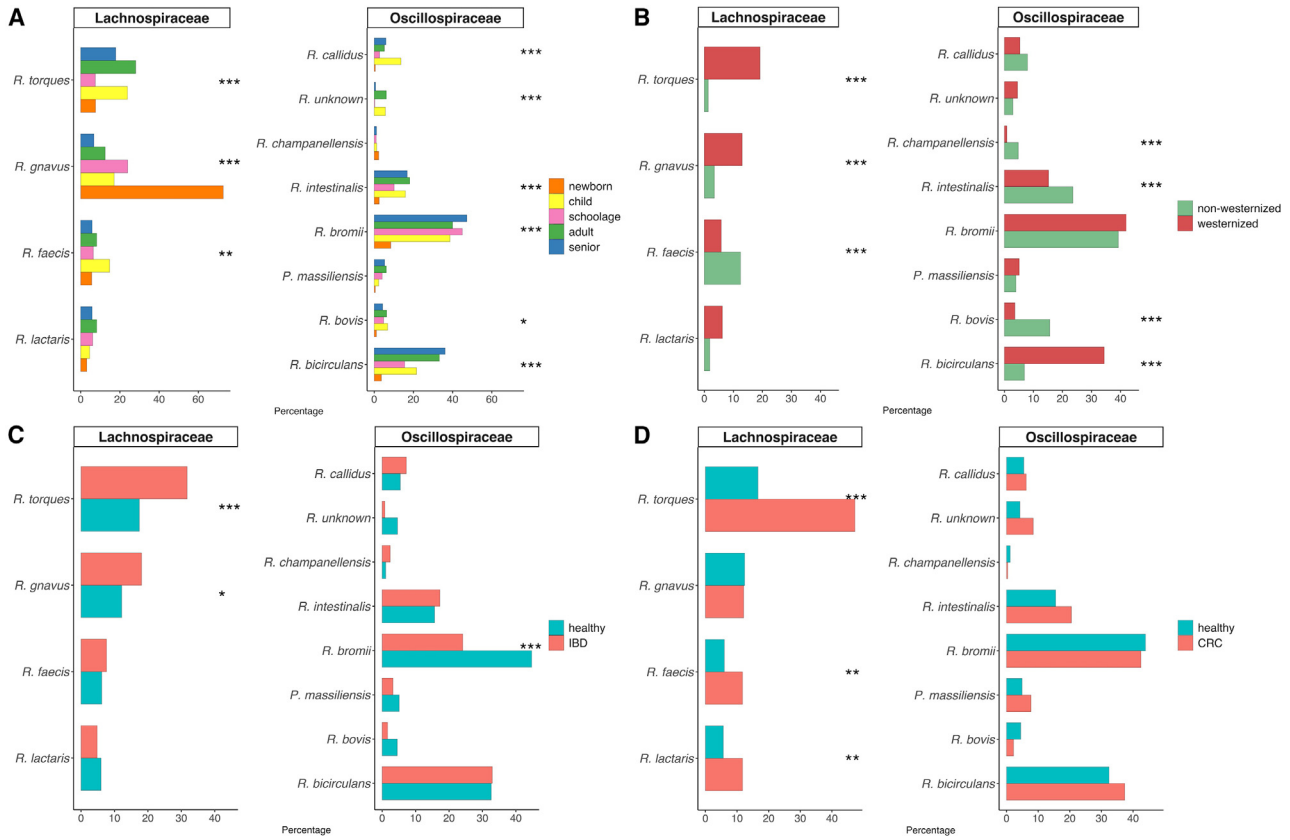


Figure 2. Prevalence plot

(A–D) The height of the bar corresponds to the percentage of subjects from each age category (A), lifestyle (B), diagnosis of inflammatory bowel disease (IBD) (C), and diagnosis of colorectal cancer (CRC) (D) showing at least 1 genome of a species. Asterisks highlight statistically significant differences as computed by Pearson's χ^2 test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Only SGBs with >60 genomes are shown.

Interestingly, all subjects from Fiji islands (hereafter reported as Oceania) and Africa were associated with a non-Westernized lifestyle (NWL), whereas populations from America, Asia, and Europe mainly encompassed people living with a Westernized lifestyle (97.9%, 91.9%, and 100% of the groups, respectively). Overall, an NWL accounted for 6.8% of total subjects (n = 349).

Also, 273 subjects were reported as affected by CRC. Of these, 149 (55%) were adults from America (n = 32), Asia (n = 28), and Europe (n = 89), whereas 124 (45%) were elderly (23 Americans, 31 Asian, and 70 European subjects).

Similarly, 169 people suffered from IBD, whereas 290 suffered from type 2 diabetes (Table S2).

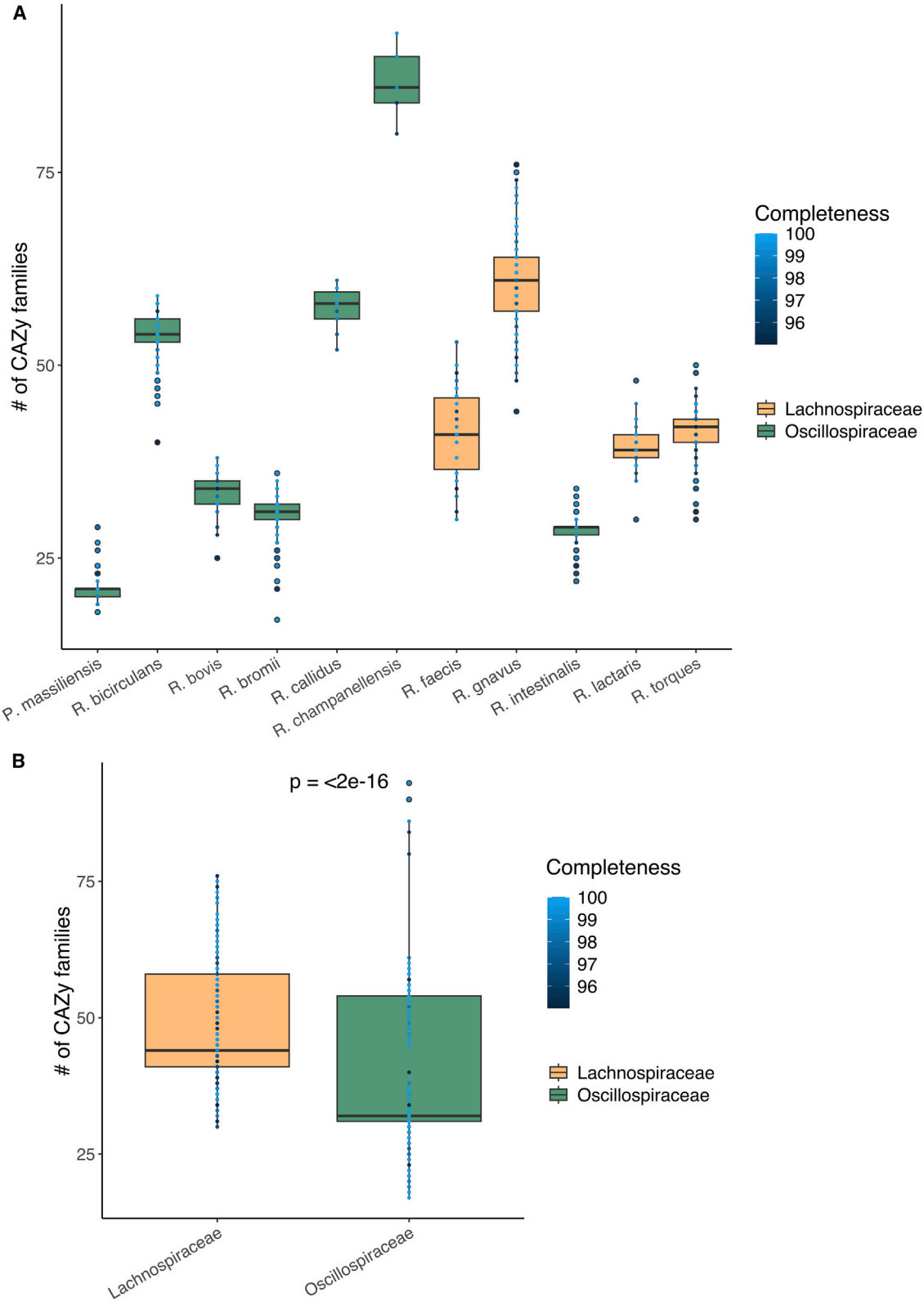
The prevalence of *Ruminococcus* SGBs is linked to age, lifestyle, and disease

The most frequently detected species was *R. bromii* (SGB 4285), with a total of 2,610 genomes (27%) from 2,546 samples, followed by *R. bicirculans* (SGB 4262, n = 1,942 from 1,877 samples, 20.1% of the total), *R. torques* (SGB 4608, n = 1,073 genomes from 1,055 samples, 11.1%), *R. intestinalis* (SGB 4290, n = 937 from 900 samples, 9.7%), and *R. gnavus* (SGB 4584, n = 738, from 732 samples, 7.7%). However, differences in the distribution of these SGBs in the population were highlighted. Indeed, *R. torques* (SGB 4608, n = 1,073 genomes) was overrepresented in Westernized subjects (χ^2 test FDR < 0.001, prevalence of 7.6% and 0.4% in Western and non-Western subjects, respectively), and it was also significatively linked with diagnosis of CRC and IBD (FDR < 0.001, prevalence of 44% and 16.4% in Westernized CRC-affected and healthy subjects, respectively; Figures 2B–2D). In addition, *R. faecis* was significatively more prevalent in Americans diagnosed with CRC (Table S3).

Furthermore, ~80% of newborns showed at least one genome taxonomically assigned to *R. gnavus* (χ^2 test FDR < 0.001), whereas it was less represented in adults and seniors (Figures 1 and 2A; Table S3).

Interestingly, *R. champanellensis* was overrepresented in non-Westernized subjects (FDR < 0.001, 4.8% in non-Westernized vs. 0.8% in Westernized), and a similar result was observed for *R. intestinalis*, albeit less significant (Figure 2B). In addition, we observed a significant depletion of *R. bromii* in subjects with IBD (χ^2 test FDR < 0.001; Table S3; Figure 2C).

At the family level, 85.4% of all genomes reconstructed from non-Westernized subjects were from the Oscillospiraceae family, significantly higher (χ^2 test FDR < 0.001) than the prevalence of the same family in Westernized subjects (prevalence, 72.1%). Also, the Lachnospiraceae family was significatively linked with



(legend on next page)

CRC (227 genomes observed in CRC-affected subjects, 157 expected; χ^2 test FDR <0.001).

Unfortunately, we were not able to explore co-occurrence and co-exclusion patterns between the species, since 3,433 samples (~58% of the total) showed only one genome identified as *Ruminococcus* spp.

Species from Oscillospiraceae and Lachnospiraceae show a different pattern of genes coding for carbohydrate-active enzymes

A total of 23,111 carbohydrate-active enzymes (CAZy) were found in 9,463 genomes belonging to 332 CAZy families. The most widespread family was GT2, observed in 9,455 genomes (99.9% of the total), followed by the GH13 subfamily 31 (n = 9,033, 95.5%) and GH1 (n = 8949, 94.6%). Interestingly, these families, together with GH13 subfamily 11, GH23, and GT4, were the most prevalent in all SGBs, as they were present in at least 75% of the genomes of each species (Figure S1A; Table S4).

As expected, the number of CAZy found in each genome was significantly and positively correlated with the genome completeness (Figure S1B); therefore, only genomes showing completeness ≥95% (n = 3,204; 33.9%) were retrieved to estimate the average CAZy diversity harbored by each SGB.

The results shed light on the different potential adaptation fitness to carbohydrate hydrolysis by each species (Figure 3). Indeed, *R. champanellensis* was the SGB showing the highest number of CAZy families (average 86 ± 5), followed by *R. gnavus*, *R. callidus*, and *R. bicirculans*, which harbored a similar number of different CAZy families on average (61 ± 6, 57 ± 3, and 54 ± 2, respectively). On the contrary, *R. bromii* and *R. bovis* showed a limited group of CAZy families (31 ± 2 and 34 ± 2, respectively; Figure 3A).

Within each SGB, no differences were observed in the number of CAZy families between genomes from non-Westernized and Westernized populations (Wilcoxon rank-sum test $p > 0.05$) as well as from CRC-, IBD-, or TD2-affected and healthy subjects. However, high-quality *R. gnavus* genomes from children had a significatively higher number of CAZy families compared with all the other age categories (Wilcoxon rank-sum test $p < 0.05$).

Interestingly, genomes belonging to the Lachnospiraceae family had an average of 49 ± 11 CAZy, significatively more than those harbored by Oscillospiraceae (40 ± 12, Wilcoxon's rank-sum test $p < 2e{-}16$), regardless of the species (Figure 3B). Further differences in the composition of CAZy families harbored by the Lachnospiraceae and Oscillospiraceae and by species were analyzed more deeply. Indeed, the two taxonomic families showed a significantly different pattern in CAZy family presence-absence (Figure 4; Table S5; permutation multivariate ANOVA $p < 0.05$). Interestingly, Lachnospiraceae was enriched in members from families GH18 (chitosanases, mannosyl-glycoprotein endo-β-N-acetylglucosaminidase), GH112 (β-galactoside phosphorylase, including lacto-N-biosidases and galacto-N-biosidases, represented in the genomes by homologs of GenBank: ASM70203.1, AXA82973.1, QEI30710.1, QHB23220.1, and QRT29560.1), GH42 (β-galactosidases), GH29 (α-1,3-L-fucosidases and α-1,6-L-fucosidases with genes from genomes matching GenBank: ALS29055.1 and QNK57166.1), GH101 (endo-α-N-acetylgalactosaminidase), GH95 (α-L-galactosidases and α-L-fucosidases), and GT0 (i.e., glycosyltransferases not yet assigned to a family due to low similarity with previously established families), most of which could potentially be involved in the hydrolysis of human N- and O-glycans[13–16] (Figure S2). On the contrary, Oscillospiraceae showed higher prevalence of several GH13 subfamilies (including α-glucosidases, α-amylases, and pullulanases from subfamilies GH13_44, GH13_39, GH13_42, and GH13_28, some of which also linked with CBM26) (Figure 5).

In addition, some CAZy were exclusively found in one species. Indeed, GH5 subfamily 2 was observed in ~88% of *R. bicirculans* genomes, whereas GH88 and GH98 were exclusively found in ~59% and 50% of *R. gnavus* genomes, respectively (Figure 5; Table S6). More information about the activities and the target carbohydrates of the CAZy families differentially prevalent between families and species is reported in Tables S5 and S6, respectively.

Lachnospiraceae and Oscillospiraceae show different engraftment potential

We mapped predicted genes against a database including 230 genes linked to adherence to the gut epithelium from a previous study[17] and all genes from the Virulence Factor Database (VFDB) involved in adhesion (i.e., VFC0001, some of which linked to both biotic and abiotic surfaces). Collectively, 6,344 genes potentially linked with adhesion processes were found in 4,052 genomes (42% of the total). Among these, only 1 genome of *R. intestinalis* (SGB 4290) harbored 5 genes involved in adherence. In addition, 2,369 genomes harbored only 1 gene linked with adherence, whereas 1,182 genomes had >2 genes.

Most of the genes (n = 2,423, 38.2%) matched with a sequence included in the VFDB encoding for a fibronectin-binding chaperonine (*GroEL* protein) involved in intestinal colonization,[18] whereas 1,777 (28%) showed a high similarity with an elongation factor (EF) interacting with host cell nucleoin.[19] In addition, 1,528 and 612 genes, accounting for 24.1% and 9.6% of the matching genes, mapped against the proteins *EpsF* and *EpsJ* (involved in the synthesis of exopolysaccharides, playing a role in adhesion to biotic and abiotic surfaces[17]), respectively.

Figure 3. Box plots reporting the number of carbohydrate-active enzyme (CAZy) families found in each genome

(A) Points represent single genomes grouped by species and are color coded according to the percentage of completeness as computed by CheckM2. Only genomes with a percentage of completeness >95% from species-level genome bins (SGBs) with >60 metagenome-assembled genomes (MAGs) are included in this plot.

(B) Box plot comparing the number of different CAZy families harbored by genomes from Lachnospiraceae and Oscillospiraceae families. The means are compared through Wilcoxon rank-sum test.

In each box plot, the box indicates the interquartile range (IQR), with the lower and upper limit of the box representing the first and third quartiles, respectively. The horizontal line denotes the median, while whiskers indicate the lowest and highest values within 1.5 × IQR from lower and upper quartiles, respectively

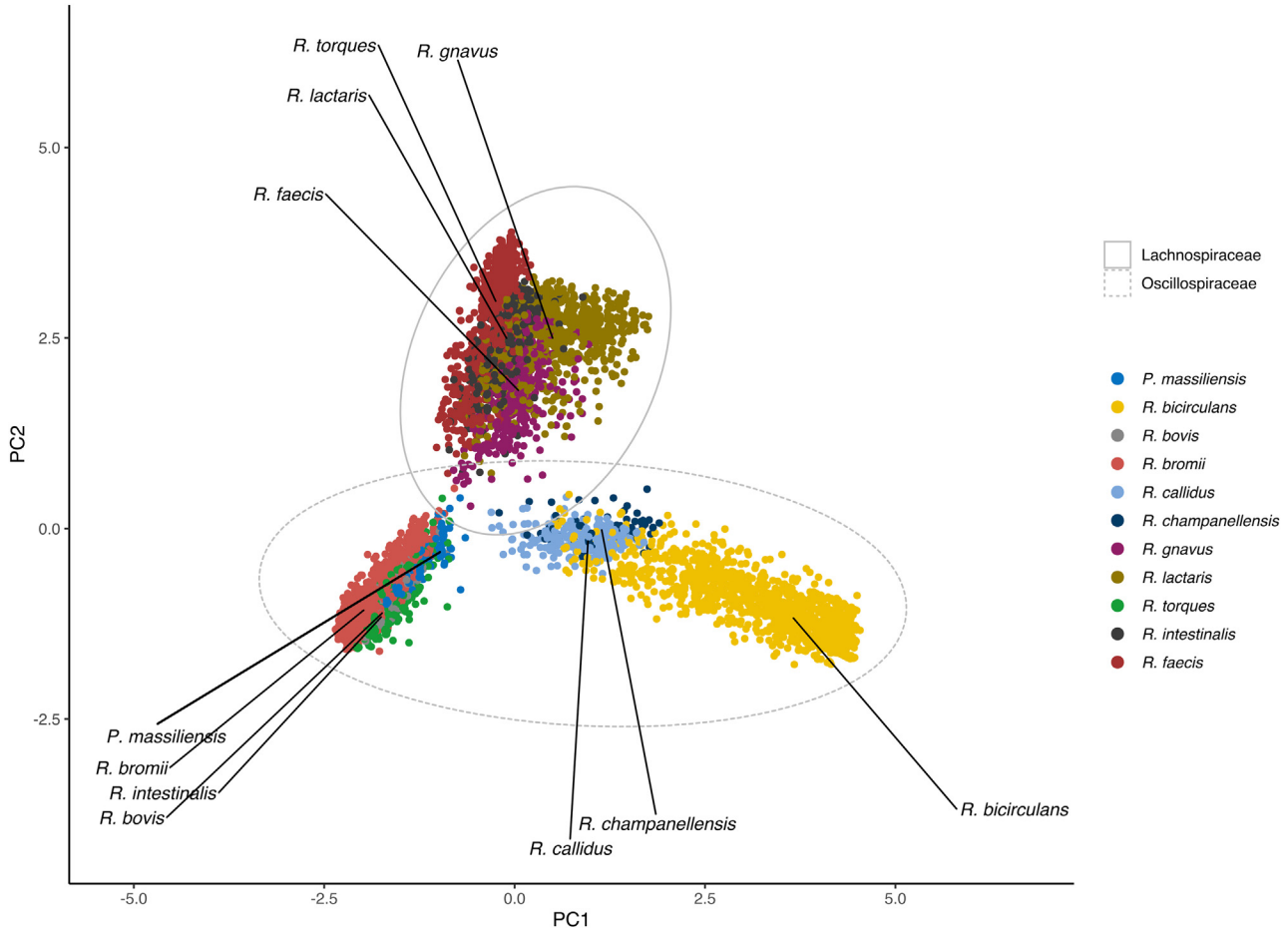


Figure 4. PCA of the CAZy family presence-absence profiles

Each point represents an MAG, color-coded according to the species-level taxonomic assignment. Genomes are further grouped according to the family. Only SGBs with >60 genomes are shown.

Interestingly, a heatmap based on the presence-absence of adherence-linked genes in the genomes clearly showed a separate clustering of Lachnospiraceae and Oscillospiraceae genomes, with the former harboring the *GroEL* and/or the EF proteins and the latter mainly harboring *EpsF* and/or *EpsJ* (Figure S3A). Also, 84.7% of all Lachnospiraceae genomes (n = 2,176) harbored at least 1 gene linked with adherence, in contrast with 27.2% of the Oscillospiraceae genomes (n = 1,876). The higher prevalence of the adherence traits in Lachnospiraceae was further confirmed by χ^2 test ($p < 0.05$; Figure S3B).

Ruminococcus and *Mediterraneibacter* spp. subgroups are linked with specific health status and geographical origin

In order to dive more in depth into the diversity within each species, average nucleotide identity (ANI) distances between each couple of genomes were computed within each SGB. For some of the species, we found sub-clusters significatively linked with populations subgroups.

Indeed, the partition around medoids (PAM) algorithm highlighted three distinct subclades within *R. faecis* (SGB 4563; Figure 6). Interestingly, PAM 3, comprising 58 genomes, was mainly detected in Asian people (χ^2 $p < 0.001$; Figures 6A and 6B), whereas 18% of the MAGs included in PAM 2 (10 of 53) were reconstructed from subjects diagnosed with CRC (Figures 6C and 6D; χ^2 $p < 0.05$). Furthermore, the functional analysis supported the separation between the subgroups. Indeed, a principal-component analysis (PCA) performed on CAZy presence-absence profiles in PAM clusters highlighted that families GH136 (lacto-*n*-biosidases), GH112, GH29, and GH106 (α-L-rhamnosidases) were the most discriminating between the subgroups, being more prevalent in the CRC-enriched PAM cluster 2 (Figures S4A and S4B). Similarly, this subgroup was enriched in genes such as O-antigen polysaccharide polymerase and others involved in pilus synthesis (Figure S4C).

Similarly, subgroups of the species *R. champanellensis* (SGB 4418) and *R. bovis* (SGB 4272) were significatively enriched in non-Westernized subjects (permutational multivariate ANOVA $p < 0.05$; Figures S5A and S5B). For *R. bovis*, the pangenome analysis revealed a higher prevalence of genes from the *glg* operon, k-carrageenases, and bile acid dehydroxylases in genomes reconstructed from Westernized subjects, whereas

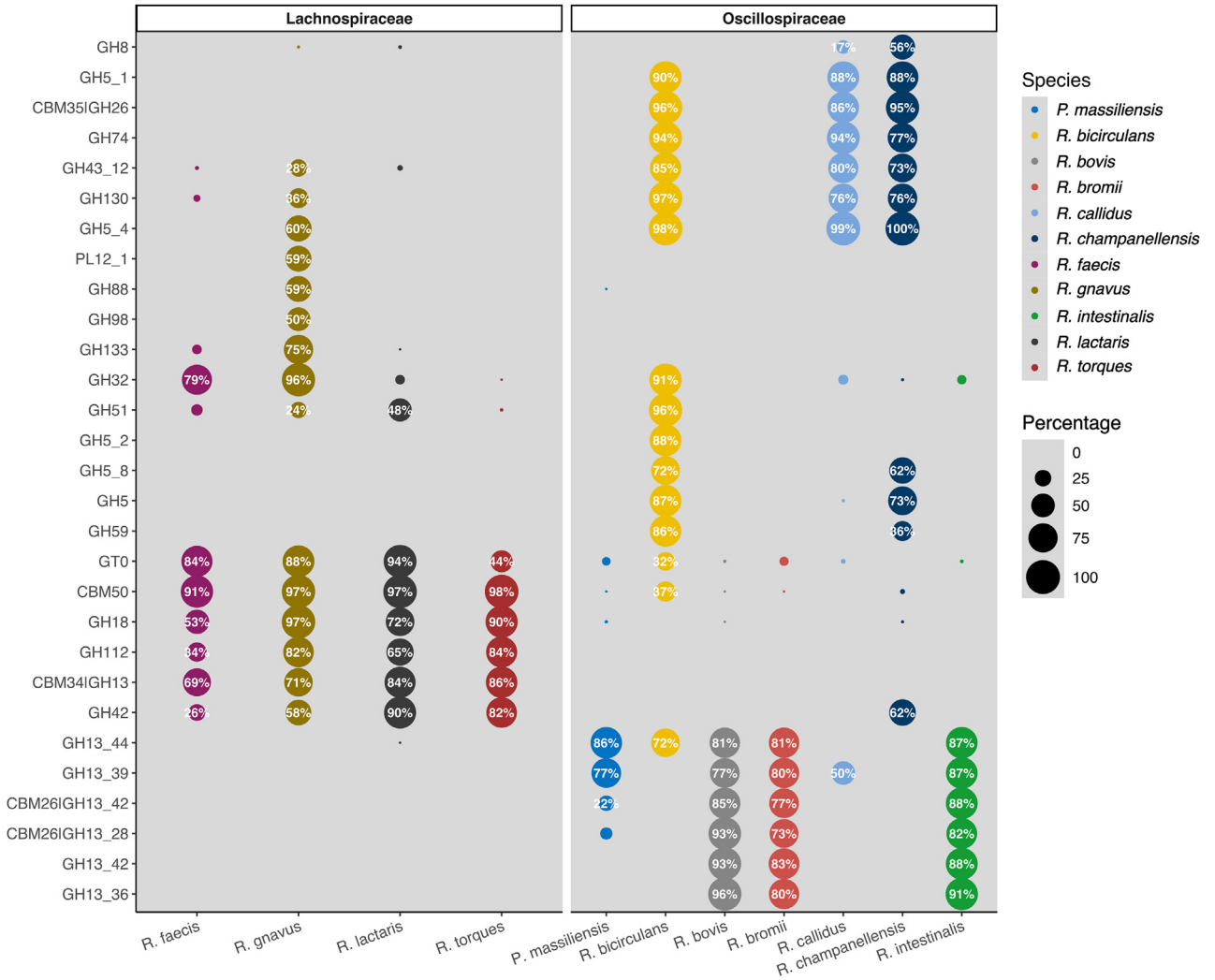


Figure 5. Bubble plot showing the percentage of MAGs from a species harboring at least 1 CAZy from a family/subfamily

Only a subset of all CAZy families/subfamilies showing a significantly different prevalence (FDR-corrected p <0.05) between the species according to Pearson's χ^2 test is shown. Bubbles are color coded according to the species.

those deriving from non-Westernized were enriched in genes involved in biosynthesis of butanol (Figure S5C).

Interestingly, geographical origin was the main factor driving the separation of *R. gnavus* genomes, with sialate O-acetylesterase domain-containing proteins and maltose-maltodextrin-binding proteins occurring more frequently in genomes detected in Asian subjects, whereas proteins harboring peptidase M56 domains and L(+)-tartrate dehydratases overrepresented in genomes from Europeans (Figures S6A and S6B). Furthermore, the operon involved in the synthesis and export of the proinflammatory polysaccharide produced by some *R. gnavus* strains[20] was significantly more complete in European than Asian genomes (Figure S6C), whereas no differences were observed when comparing genomes from different age categories or diseases (Wilcoxon rank-sum test $p > 0.05$).

An underlying subgroup diversity in *R. bromii* was also explored further (Figure S7A). Within this species, two PAM clusters were observed, with one encompassing 2,369 genomes of 2,575 (92% of the total). PAM subgroups were not linked with any specific metadata (χ^2 test $p > 0.05$) except for geographical origin, since PAM subgroup 1 was slightly overrepresented in Asians (Figure S7B). However, genomes from the two PAMs groups had a diverse pangenome. Indeed, PAM subcluster 1 was enriched in uridine diphosphate glucose:undecaprenyl-phosphate glucose-1-phosphate transferases, neopullunanases, and glucose resistance amylase regulators, while lipid A core O-antigen ligases, fibronectin type III, and ComEC/Rec2 genes were overrepresented in PAM 2 (Figure S7C).

DISCUSSION

The main purpose of this work was to elucidate the taxonomic and functional diversity of *Ruminococcus* and *Mediterraneibacter* (previously known only as *Ruminococcus*). Mounting

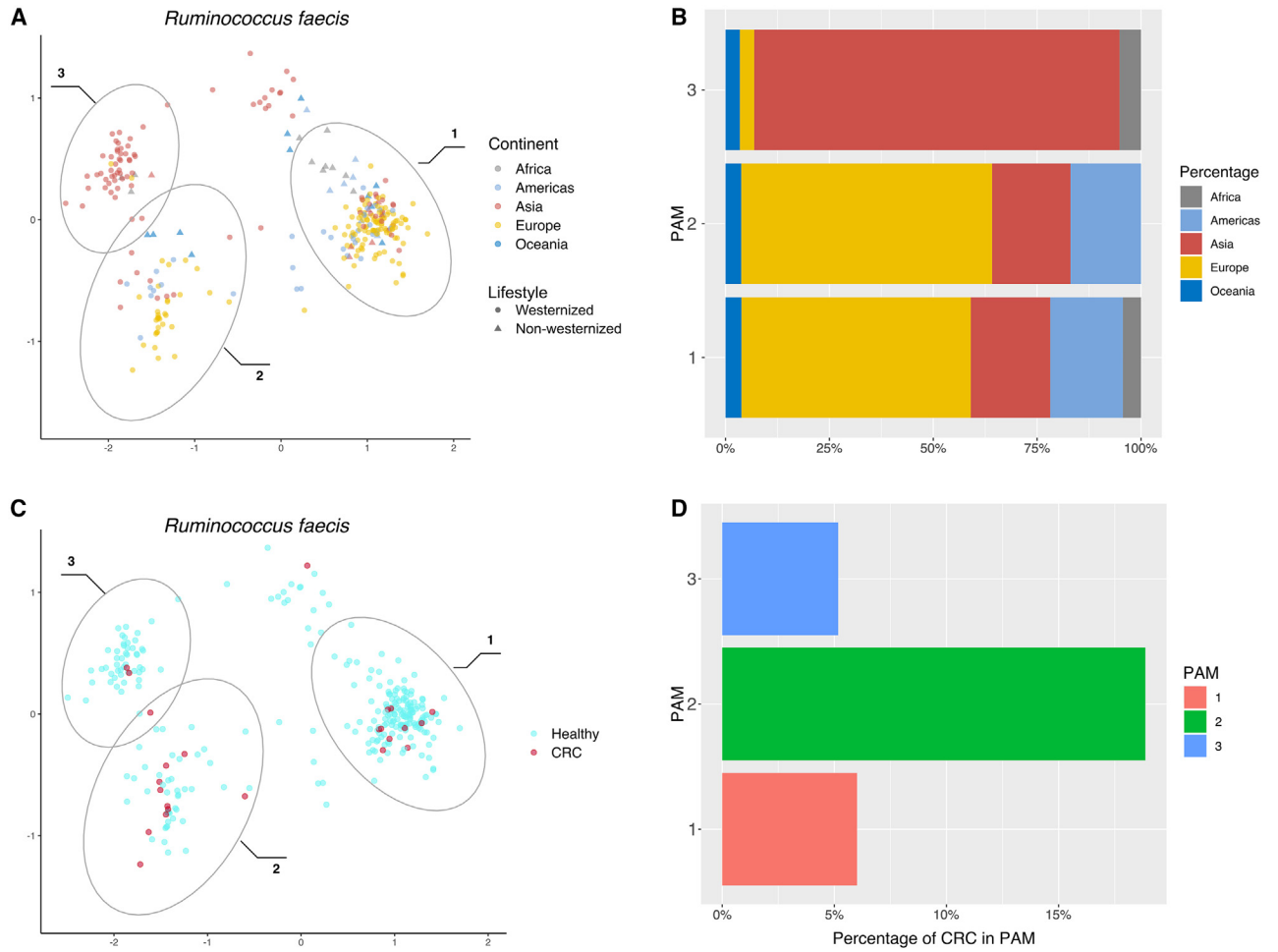


Figure 6. *R. faecis* subspecies

(A–D) Three partition around medoids (PAM) clusters are highlighted in the principal-coordinates analysis performed on the pairwise average nucleotide identity (ANI) distance matrix (A and C). PAM cluster 3 is significatively more prevalent in Asian subjects (B; Pearson's χ^2 test $p < 0.001$), whereas PAM cluster 2 shows a high proportion of subjects diagnosed with CRC (D; Pearson's χ^2 test $p < 0.001$).

evidence has highlighted the relevance of this taxonomic group in gut microbiome composition, which has not only been identified as a keystone for the most frequently detected enterotype of the human gut microbiome[6] but also acknowledged as part of the "core" microbiome,[1] thus being highly prevalent but with a lower relative abundance in most humans.

The role of several *Ruminococcus* and *Mediterraneibacter* species in human health and disease is still ambiguous. Several studies have linked the increase in *R. torques* (family Lachnospiraceae) with adverse conditions, such as CRC[21,22] and IBD,[23] and Kwak and colleagues have proposed this species as a biomarker for predicting Crohn's disease.[24] The results of our large-scale analysis confirm these observations, as we found a strongly significant link between the occurrence of *R. torques* and diagnosis of CRC and IBD, whereas the opposite was observed for *R. bromii*, which was significantly depleted in subjects affected by IBD, as reported previously.[23] Similarly, the higher prevalence of *R. gnavus* in infants is in line with findings of Sagheddu and colleagues,[25] who identified the species in all 25 babies aged between 1 month and 2 years.

To the best of our knowledge, no other study has reported the potential link of *R. bovis* and *R. champanellensis* (both members of the family Oscillospiraceae) with an NWL, although *R. champanellensis* has been reported to dominate the gut microbiome of ancient humans compared to modern Westernized/non-Westernized populations.[26] Previous findings indicate that *R. champanellensis* harbors a complex group of genes involved in hydrolysis of cellulose,[27,28] abundant in plant-origin foods. This supports the high diversity of CAZy families found within this species as well as the consistent prevalence within the species of the family GH8, involved in cellulose catabolism.[29]

Strain-level analyses can help to define the potential effects of metabolism of specific strains on host health and to highlight the extent and the conditions under which the adverse effects might occur, as either probiotic or pathogenetic traits are strain and dose specific.[30,31] Interestingly, our genomic and functional

analysis of *R. faecis* (a species first isolated from the human gut and phylogenetically close to *R. torques*[32]) highlighted the presence of three distinct clades harboring different metabolic potentials (Figure 6). A high proportion of genomes in clusters 2 and 3 was reconstructed from CRC-affected individuals and also showed a significantly higher prevalence of CAZy families potentially involved in intestinal gut barrier disruption, such as glycosyl hydrolase GH29 (fucosidases) and GH136 and GH112 (lacto-N-biosidases and lacto-N-biose phosphorylases), as well as carbohydrate binding modules from family CBM32, often associated with sialidases.[33] It is known that *R. torques* can disrupt gut barrier integrity by degrading mucin,[34] thus promoting pathogen invasion and triggering inflammation and disease.[35] However, it cannot be excluded that this trait is shared with phylogenetically similar microbial species adapted to the same ecological niches, such as *R. faecis*. In addition, genomes from groups 2 and 3 showed a higher prevalence of Wzy-dependent O-antigen polysaccharide polymerase, which is involved in a virulence factor pathway strongly linked with all CRC stages.[36]

Different potential sub-species were also observed within *R. gnavus* and *R. bromii*. In *R. gnavus*, we observed a separation between genomes from Asian adults and European newborns. The dominance of sialate O-acetylesterases in the first group suggests its contribution to removing esterification from N-acetylneuraminic acid residues in mucin,[37] thus enhancing the activity of sialidases from other microorganisms[38–40] and leading to the hypothesis of cooperative and complementary mucin-foraging mechanisms potentially involving *R. gnavus* and sialidase-encoding microorganisms. On the contrary, a high percentage of *R. gnavus* genomes from European newborns harbor type II toxin-antitoxin system HicAB genes providing resistance to environmental stress and to potentially competing microorganisms,[41] which might explain their ability to colonize the gut microbiome from the first months of life independent of diet and delivery method.[25,42] Furthermore, the operon involved in proinflammatory polysaccharide synthesis was more complete in genomes of European subjects, while it was independent of age, lifestyle, and diseases. However, the mechanisms leading to this result need to be elucidated.

Interestingly, dietary lifestyle showed a strong effect on strain-level stratification of *R. champanellensis* and *R. bovis* (family Oscillospiraceae), previously highlighted as members of the non-Westernized gut microbiome (Figures 2B; Table S3). Consistently, *R. bovis* strains from NW subjects exhibited an enrichment in butanol-producing genes, which has been associated previously with higher fiber intake.[43,44] On the contrary, the significatively higher prevalence of genes linked with k-carrageenan metabolism in *R. bovis* genomes from Westernized compared with those from non-Westernized subjects suggests a possible genomic adaptation of the species to the consumption of this food additive in Western populations.[45] Interestingly, *R. bovis* from Westernized subjects was also were enriched in bile acid 7-dehydrolase genes, involved in conversion of primary bile salts into secondary toxic compounds.[46]

In addition to strain-level diversity, we also highlighted family-level functional stratification. To date, the grouping of former *Ruminococcus* into the new *Mediterraneibacter* species was mainly based on phylogenetic analyses,[3] but differences in the functional potential between the two clades were never explored. Here, we highlighted a clustering of CAZy profiles according to the families, showing that all Oscillospiraceae species (including *R. bromii*) were rich in families involved in (resistant) starch binding and hydrolysis (such as GH13_39, GH13_36, and CBM26, encoding pullulanases, neopullunanases, and specific starch binding proteins, respectively), whereas those belonging to Lachnospiraceae might contribute to hydrolyze human N- and O-glycan utilization through expression of, for example, CAZy from families GH112, GH42, GH18, and GH29.[16] This suggests that the Lachnospiraceae family might encompass ruminococci that are better adapted to colonize the human mucosa, thus potentially inhabiting the gut barrier and behaving as opportunistic pathogens. Indeed, *R. torques* and *R. gnavus* are part of the Lachnospiraceae family and have been reported as colonizers of gut mucin potentially involved in pro-inflammatory phenomena.[47–49] Therefore, the identification of novel putative species showing high similarity to LRs may ease the exploration of their metabolic fitness, as some traits are shared between all members of the family. The better ability of Lachnospiraceae to adhere to the gut mucosa compared to Oscillospiraceae is also supported by the higher prevalence of virulence factors involved in adherence that we observed here.

Conversely, some of the Ors, such as *R. bromii*, are reportedly able to digest fiber, such as RS,[50] and their prevalence in the human gut is usually linked to beneficial effects, such as lowering of insulin resistance,[51] reduction of cardiovascular risk,[52] and cross-feeding dynamics, supporting the growth of other health-related taxa such as *Faecalibacterium prausnitzii*.[53,54] However, while most studies have emphasized the role of *R. bromii* in degrading RS, our results suggest that some of the CAZy encoded by this species and involved in RS hydrolysis might be shared with other ORs, suggesting their potential role in degrading RS as well.

Overall, these results show that both OR and LR species harbor different subgroups, some of which directly link to specific outcomes, while others are independent of the health conditions considered in this analysis or lifestyle. This further underlines the need of investigating phylogenetic and pangenomic traits on isolates to better describe the functional potential of such strains on human health.

Consistent with previous reports, the link between some *Ruminococcus* species and specific diseases is clear. Indeed, fiber consumption represents the factor contributing most in the selection of species, which further indicates the importance of dietary choices as key factors for gut microbiome composition and function. Adequate fiber intake can contribute to the selection of fiber-degrading *Ruminococcus* potentially providing health-related activities.

Limitations of the study

Our study focuses on the distribution of the genus *Ruminococcus* in the population and on the description of genomic features potentially discriminating between subgroups of the same species. We highlighted that the prevalence of some *Ruminococcus* species is linked with specific conditions and that species from two distinct families are mainly distinguished by the targeted carbohydrates. Notably, all of these results are based on the description of the metabolic potential of *Ruminococcus* spp. Understanding the conditions that might lead to the expression of specific

Table 1. Number of genomes from isolates and reconstructed from MAGs

	MetaRefSGB genomes	NCBI RefSeq genomes
Isolate	0	21
MAG	9,463	1

genomic features discussed in this work might help to more reliably relate the presence and the activity of *Ruminococcus* genomes within the human gut to specific conditions or populations. Future large-scale experiments involving transcriptomics analysis of *Ruminococcus* spp. isolates might help to fill this knowledge gap.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Danilo Ercolini (ercolini@unina.it).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The MAGs from human gut metagenomes used in this study are available in a public repository (see key resources table and Pasolli et al.[12]). The reference genomes of *Ruminococcus* spp. and *Mediterraneibacter* spp. are available in the NCBI database (see key resources table and Table S7 for accession numbers).
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This project was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 – Call for Tender No. 341 of 15 March 2022 of the Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, Project title "ON Foods – Research and Innovation Network on Food and Nutrition Sustainability, Safety and Security – Working ON Foods."

AUTHOR CONTRIBUTIONS

V.V., investigation, formal analysis, visualization, and writing – original draft; F.D.F., conceptualization and writing – review & editing; R.M., formal analysis; E.P., data curation; D.E., conceptualization, resources, and writing – review & editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - MAG and reference genome retrieval
  - Bioinformatics processing of MAGs and reference genomes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.115018.

Received: May 15, 2024
Revised: September 23, 2024
Accepted: November 11, 2024

REFERENCES

1. La Reau, A.J., and Suen, G. (2018). The Ruminococci: key symbionts of the gut ecosystem. J. Microbiol. *56*, 199–208. https://doi.org/10.1007/s12275-018-8024-4.
2. Abell, G.C.J., Cooke, C.M., Bennett, C.N., Conlon, M.A., and McOrist, A.L. (2008). Phylotypes related to Ruminococcus bromii are abundant in the large bowel of humans and increase in response to a diet high in resistant starch. FEMS Microbiol. Ecol. *66*, 505–515. https://doi.org/10.1111/j.1574-6941.2008.00527.x.
3. Togo, A.H., Diop, A., Bittar, F., Maraninchi, M., Valero, R., Armstrong, N., Dubourg, G., Labas, N., Richez, M., Delerce, J., et al. (2018). Description of Mediterraneibacter massiliensis, gen. nov., sp. nov., a new genus isolated from the gut microbiota of an obese patient and reclassification of Ruminococcus faecis, Ruminococcus lactaris, Ruminococcus torques, Ruminococcus gnavus and Clostridium glycyrrhizinilyticum as Mediterraneibacter faecis comb. nov., Mediterraneibacter lactaris comb. nov., Mediterraneibacter torques comb. nov., Mediterraneibacter gnavus comb. nov. and Mediterraneibacter glycyrrhizinilyticus comb. nov. Antonie Leeuwenhoek *111*, 2107–2128. https://doi.org/10.1007/s10482-018-1104-y.
4. La Reau, A.J., Meier-Kolthoff, J.P., and Suen, G. (2016). Sequence-based analysis of the genus Ruminococcus resolves its phylogeny and reveals strong host association. Microb. Genom. *2*, e000099. https://doi.org/10.1099/mgen.0.000099.
5. Vos, P., Garrity, G., Jones, D., Krieg, N.R., Ludwig, W., Rainey, F.A., Schleifer, K.-H., and Whitman, W.B. (2011). Bergey's manual of systematic bacteriology*Volume 3* (The Firmicutes: Springer Science & Business Media).
6. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., et al. (2011). Enterotypes of the human gut microbiome. Nature *473*, 174–180. https://doi.org/10.1038/nature09944.
7. Bhattarai, Y., Muniz Pedrogo, D.A., and Kashyap, P.C. (2017). Irritable bowel syndrome: a gut microbiota-related disorder? Am. J. Physiol. Gastrointest. Liver Physiol. *312*, G52–G62. https://doi.org/10.1152/ajpgi.00338.2016.
8. Kandasamy, S., Letchumanan, V., Hong, K.W., Chua, K.-O., Ab Mutalib, N.S., Ng, A.L.O., Ming, L.C., Lim, H.X., Thurairajasingam, S., Law, J.W.-F., and Tan, L.T.H. (2023). The role of human gut microbe Ruminococcus gnavus in inflammatory diseases. Prog. Micobes. Mol. Biol. *6*. https://doi.org/10.36877/pmmb.a0000396.
9. De Filippis, F., Paparo, L., Nocerino, R., Della Gatta, G., Carucci, L., Russo, R., Pasolli, E., Ercolini, D., and Berni Canani, R. (2021). Specific gut microbiome signatures and the associated pro-inflamatory functions are linked to pediatric allergy and acquisition of immune tolerance. Nat. Commun. *12*, 5958. https://doi.org/10.1038/s41467-021-26266-z.
10. Ruiz-Saavedra, S., Arboleya, S., Nogacka, A.M., González del Rey, C., Suárez, A., Diaz, Y., Gueimonde, M., Salazar, N., González, S., and de los Reyes-Gavilán, C.G. (2023). Commensal fecal Microbiota profiles associated with initial stages of intestinal mucosa damage: A pilot study. Cancers *16*, 104. https://doi.org/10.3390/cancers16010104.
11. Kumari, M., Singh, P., Nataraj, B.H., Kokkiligadda, A., Naithani, H., Azmal Ali, S., Behare, P.V., and Nagpal, R. (2021). Fostering next-generation probiotics in human gut by targeted dietary modulation: An emerging

perspective. Food Res. Int. *150*, 110716. https://doi.org/10.1016/j.foodres.2021.110716.

12. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*, 649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.
13. Luis, A.S., and Hansson, G.C. (2023). Intestinal mucus and their glycans: A habitat for thriving microbiota. Cell Host Microbe *31*, 1087–1100. https://doi.org/10.1016/j.chom.2023.05.026.
14. Crouch, L.I. (2023). N-glycan breakdown by bacterial CAZymes. Essays Biochem. *67*, 373–385. https://doi.org/10.1042/ebc20220256.
15. Tailford, L.E., Crost, E.H., Kavanaugh, D., and Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. Front. Genet. *6*, 81. https://doi.org/10.3389/fgene.2015.00081.
16. Labourel, A., Parrou, J.-L., Deraison, C., Mercier-Bonin, M., Lajus, S., and Potocki-Veronese, G. (2023). *O*-Mucin-degrading carbohydrate-active enzymes and their possible implication in inflammatory bowel diseases. Essays Biochem. *67*, 331–344. https://doi.org/10.1042/ebc20220153.
17. Lebeer, S., Vanderleyden, J., and De Keersmaecker, S.C.J. (2008). Genes and molecules of lactobacilli supporting probiotic action. Microbiol. Mol. Biol. Rev. *72*, 728–764. https://doi.org/10.1128/mmbr.00017-08.
18. Hennequin, C., Porcheray, F., Waligora-Dupriet, A., Collignon, A., Barc, M., Bourlioux, P., and Karjalainen, T. (2001). GroEL (Hsp60) of Clostridium difficile is involved in cell adherence. Microbiology *147*, 87–96. https://doi.org/10.1099/00221287-147-1-87.
19. Harvey, K.L., Jarocki, V.M., Charles, I.G., and Djordjevic, S.P. (2019). The diverse functional roles of elongation factor Tu (EF-Tu) in microbial pathogenesis. Front. Microbiol. *10*, 2351. https://doi.org/10.3389/fmicb.2019.02351.
20. Henke, M.T., Kenny, D.J., Cassilly, C.D., Vlamakis, H., Xavier, R.J., and Clardy, J. (2019). Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. Proc. Natl. Acad. Sci. USA *116*, 12672–12677. https://doi.org/10.1073/pnas.1904099116.
21. Wu, Y., Zhuang, J., Zhang, Q., Zhao, X., Chen, G., Han, S., Hu, B., Wu, W., and Han, S. (2023). Aging characteristics of colorectal cancer based on gut microbiota. Cancer Med. *12*, 17822–17834. https://doi.org/10.1002/cam4.6414.
22. Liu, C., Li, Z., Ding, J., Zhen, H., Fang, M., and Nie, C. (2021). Species-level analysis of the human gut microbiome shows antibiotic resistance genes associated with colorectal cancer. Front. Microbiol. *12*, 765291. https://doi.org/10.3389/fmicb.2021.765291.
23. Wiredu Ocansey, D.K., Hang, S., Yuan, X., Qian, H., Zhou, M., Valerie Olovo, C., Zhang, X., and Mao, F. (2023). The diagnostic and prognostic potential of gut bacteria in inflammatory bowel disease. Gut Microb. *15*, 2176118. https://doi.org/10.1080/19490976.2023.2176118.
24. Kwak, M.S., Cha, J.M., Shin, H.P., Jeon, J.W., and Yoon, J.Y. (2020). Development of a novel metagenomic biomarker for prediction of upper gastrointestinal tract involvement in patients with Crohn's disease. Front. Microbiol. *11*, 1162. https://doi.org/10.3389/fmicb.2020.01162.
25. Sagheddu, V., Patrone, V., Miragoli, F., Puglisi, E., and Morelli, L. (2016). Infant early gut colonization by Lachnospiraceae: High frequency of Ruminococcus gnavus. Front. Pediatr. *4*, 57. https://doi.org/10.3389/fped.2016.00057.
26. Wibowo, M.C., Yang, Z., Borry, M., Hübner, A., Huang, K.D., Tierney, B.T., Zimmerman, S., Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., et al. (2021). Reconstruction of ancient microbial genomes from the human gut. Nature *594*, 234–239. https://doi.org/10.1038/s41586-021-03532-0.
27. Cann, I., Bernardi, R.C., and Mackie, R.I. (2016). Cellulose degradation in the human gut: *Ruminococcus champanellensis* expands the cellulosome paradigm. Environ. Microbiol. *18*, 307–310. https://doi.org/10.1111/1462-2920.13152.
28. Moraïs, S., Winkler, S., Zorea, A., Levin, L., Nagies, F.S.P., Kapust, N., Lamed, E., Artan-Furman, A., Bolam, D.N., Yadav, M.P., et al. (2024). Cryptic diversity of cellulose-degrading gut bacteria in industrialized humans. Science *383*, .eadj9223. https://doi.org/10.1126/science.adj9223.
29. Ben David, Y., Dassa, B., Borovok, I., Lamed, R., Koropatkin, N.M., Martens, E.C., White, B.A., Bernalier-Donadille, A., Duncan, S.H., Flint, H.J., et al. (2015). Ruminococcal cellulosome systems from rumen to human. Environ. Microbiol. *17*, 3407–3426. https://doi.org/10.1111/1462-2920.12868.
30. Campana, R., van Hemert, S., and Baffone, W. (2017). Strain-specific probiotic properties of lactic acid bacteria and their interference with human intestinal pathogens invasion. Gut Pathog. *9*, 12. https://doi.org/10.1186/s13099-017-0162-4.
31. Stromberg, Z.R., Van Goor, A., Redweik, G.A.J., Wymore Brand, M.J., Wannemuehler, M.J., and Mellata, M. (2018). Pathogenic and non-pathogenic *Escherichia coli* colonization and host inflammatory response in a defined microbiota mouse model. Dis. Model. Mech. *11*, dmm035063. https://doi.org/10.1242/dmm.035063.
32. Kim, M.-S., Roh, S.W., and Bae, J.-W. (2011). Ruminococcus faecis sp. nov., isolated from human faeces. J. Microbiol. *49*, 487–491. https://doi.org/10.1007/s12275-011-0505-7.
33. Boraston, A.B., Ficko-Blean, E., and Healey, M. (2007). Carbohydrate Recognition by a Large Sialidase Toxin from *Clostridium perfringens*. Biochemistry *46*, 11352–11360. https://doi.org/10.1021/bi701317g.
34. Deaver, J.A., Eum, S.Y., and Toborek, M. (2018). Circadian disruption changes gut microbiome taxa and functional gene composition. Front. Microbiol. *9*, 737. https://doi.org/10.3389/fmicb.2018.00737.
35. Pothuraju, R., Chaudhary, S., Rachagani, S., Kaur, S., Roy, H.K., Bouvet, M., and Batra, S.K. (2021). Mucins, gut microbiota, and postbiotics role in colorectal cancer. Gut Microb. *13*, 1974795. https://doi.org/10.1080/19490976.2021.1974795.
36. Mathlouthi, N.E.H., Kriaa, A., Keskes, L.A., Rhimi, M., and Gdoura, R. (2022). Virulence factors in colorectal cancer metagenomes and association of microbial siderophores with advanced stages. Microorganisms *10*, 2365. https://doi.org/10.3390/microorganisms10122365.
37. Bell, A., Brunt, J., Crost, E., Vaux, L., Nepravishta, R., Owen, C.D., Latousakis, D., Xiao, A., Li, W., Chen, X., et al. (2019). Elucidation of a sialic acid metabolism pathway in mucus-foraging Ruminococcus gnavus unravels mechanisms of bacterial adaptation to the gut. Nat. Microbiol. *4*, 2393–2404. https://doi.org/10.1038/s41564-019-0590-7.
38. Bell, A., Severi, E., Owen, C.D., Latousakis, D., and Juge, N. (2023). Biochemical and structural basis of sialic acid utilization by gut microbes. J. Biol. Chem. *299*, 102989. https://doi.org/10.1016/j.jbc.2023.102989.
39. Ashida, H., Tanigawa, K., Kiyohara, M., Katoh, T., Katayama, T., and Yamamoto, K. (2018). Bifunctional properties and characterization of a novel sialidase with esterase activity from *Bifidobacterium bifidum*. Biosci. Biotechnol. Biochem. *82*, 2030–2039. https://doi.org/10.1080/09168451.2018.1497944.
40. Phansopa, C., Kozak, R.P., Liew, L.P., Frey, A.M., Farmilo, T., Parker, J.L., Kelly, D.J., Emery, R.J., Thomson, R.I., Royle, L., et al. (2015). Characterization of a sialate-*O*-acetylesterase (NanS) from the oral pathogen *Tannerella forsythia* that enhances sialic acid release by NanH, its cognate sialidase. Biochem. J. *472*, 157–167. https://doi.org/10.1042/bj20150388.
41. Turco, S., Zuppante, L., Drais, M.I., and Mazzaglia, A. (2022). Dressing like a pathogen: Comparative analysis of different Pseudomonas genomospecies wearing different features to infect Corylus avellana. J. Phytopathol. *170*, 504–516. https://doi.org/10.1111/jph.13101.
42. Crost, E.H., Coletto, E., Bell, A., and Juge, N. (2023). *Ruminococcus gnavus*: friend or foe for human health. FEMS Microbiol. Rev. *47*, fuad014. https://doi.org/10.1093/femsre/fuad014.
43. Yoo, M., Croux, C., Meynial-Salles, I., and Soucaille, P. (2016). Elucidation of the roles of adhE1 and adhE2 in the primary metabolism of Clostridium acetobutylicum by combining in-frame gene deletion and a quantitative

system-scale approach. Biotechnol. Biofuels *9*, 92. https://doi.org/10.1186/s13068-016-0507-0.

44. Neyrinck, A.M., Rodriguez, J., Zhang, Z., Nazare, J.-A., Bindels, L.B., Cani, P.D., Maquet, V., Laville, M., Bischoff, S.C., Walter, J., and Delzenne, N.M. (2022). Breath volatile metabolome reveals the impact of dietary fibres on the gut microbiota: Proof of concept in healthy volunteers. EBioMedicine *80*, 104051. https://doi.org/10.1016/j.ebiom.2022.104051.
45. Borsani, B., De Santis, R., Perico, V., Penagini, F., Pendezza, E., Dilillo, D., Bosetti, A., Zuccotti, G.V., and D'Auria, E. (2021). The role of carrageenan in inflammatory bowel diseases and allergic reactions: Where do we stand? Nutrients *13*, 3402. https://doi.org/10.3390/nu13103402.
46. Zeng, H., Umar, S., Rust, B., Lazarova, D., and Bordonaro, M. (2019). Secondary bile acids and short chain fatty acids in the colon: A focus on colonic microbiome, cell proliferation, inflammation, and cancer. Int. J. Mol. Sci. *20*, 1214. https://doi.org/10.3390/ijms20051214.
47. Owen, C.D., Tailford, L.E., Monaco, S., Šuligoj, T., Vaux, L., Lallement, R., Khedri, Z., Yu, H., Lecointe, K., Walshaw, J., et al. (2017). Unravelling the specificity and mechanism of sialic acid recognition by the gut symbiont Ruminococcus gnavus. Nat. Commun. *8*, 2196. https://doi.org/10.1038/s41467-017-02109-8.
48. Schaus, S.R., Vasconcelos Periera, G., Luis, A.S., Madlambayan, E., Terrapon, N., Ostrowski, M.P., Jin, C., Hansson, G.C., and Martens, E.C. (2024). Ruminococcus torques is a keystone degrader of intestinal mucin glycoprotein, releasing oligosaccharides used by Bacteroides thetaiotaomicron. Preprint at bioRxiv, 2024.01.15.575725. https://doi.org/10.1101/2024.01.15.575725.
49. Paone, P., and Cani, P.D. (2020). Mucus barrier, mucins and gut microbiota: the expected slimy partners? Gut *69*, 2232–2243. https://doi.org/10.1136/gutjnl-2020-322260.
50. Ze, X., Duncan, S.H., Louis, P., and 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. https://doi.org/10.1038/ismej.2012.4.
51. Li, H., Zhang, L., Li, J., Wu, Q., Qian, L., He, J., Ni, Y., Kovatcheva-Datchary, P., Yuan, R., Liu, S., et al. (2024). Resistant starch intake facilitates weight loss in humans by reshaping the gut microbiota. Nat. Metab. *6*, 578–597. https://doi.org/10.1038/s42255-024-00988-y.
52. Lakshmanan, A.P., Al Zaidan, S., Bangarusamy, D.K., Al-Shamari, S., Elhag, W., and Terranegra, A. (2022). Increased relative abundance of ruminoccocus is associated with reduced cardiovascular risk in an obese population. Front. Nutr. *9*, 849005. https://doi.org/10.3389/fnut.2022.849005.
53. Cheng, C.C., Duar, R.M., Lin, X., Perez-Munoz, M.E., Tollenaar, S., Oh, J.-H., van Pijkeren, J.-P., Li, F., van Sinderen, D., Gänzle, M.G., and Walter, J. (2020). Ecological importance of cross-feeding of the intermediate metabolite 1,2-propanediol between bacterial gut symbionts. Appl. Environ. Microbiol. *86*, e00190-20. https://doi.org/10.1128/aem.00190-20.
54. Teichmann, J., and Cockburn, D.W. (2021). In vitro fermentation reveals changes in butyrate production dependent on resistant starch source and microbiome composition. Front. Microbiol. *12*, 640253. https://doi.org/10.3389/fmicb.2021.640253.
55. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics *38*, 5315–5316. https://doi.org/10.1093/bioinformatics/btac672.
56. Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. *37*, D233–D238. https://doi.org/10.1093/nar/gkn663.
57. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932. https://doi.org/10.1093/bioinformatics/btu739.
58. Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. *50*, D912–D917. https://doi.org/10.1093/nar/gkab1107.
59. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat. Methods *20*, 1203–1212. https://doi.org/10.1038/s41592-023-01940-w.
60. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.
61. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114. https://doi.org/10.1038/s41467-018-07641-9.
62. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.
63. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.
64. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
65. Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47*, W256–W259. https://doi.org/10.1093/nar/gkz239.
66. Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., et al. (2013). BLAST: a more efficient report with usability improvements. Nucleic Acids Res. *41*, W29–W33. https://doi.org/10.1093/nar/gkt282.
67. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
68. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834. https://doi.org/10.1101/gr.213959.116.
69. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.
70. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.
71. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.
72. (2023). Skani enables accurate and efficient genome comparison for modern metagenomic datasets. Nat. Methods *20*, 1633–1634. https://doi.org/10.1038/s41592-023-02019-2.
73. Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinf. *11*, 538. https://doi.org/10.1186/1471-2105-11-538.
74. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Ruminococcus ssp. and *Mediterraneibacter* ssp. MAGs from human gut metagenomes	Pasolli et al.[12]	http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html
Ruminococcus spp. and *Mediterraneibacter* spp. reference genomes	NCBI RefSeq database	Accession numbers reported in Table S7
GTDB Release 95	Chaumeil et al.[55]	https://github.com/ecogenomics/gtdbtk
CAZy (Release 07262023)	Cantarel et al.[56]	https://bcb.unl.edu/dbCAN2/download/
UniRef90 (Release 2023_04)	Suzek et al.[57]	https://www.uniprot.org
Virulence Factor Database (VFDB)	Liu et al.[58]	http://www.mgc.ac.cn/VFs/main.htm
Software and algorithms		
CheckM2 1.0.1	Chklovski et al.[59]	https://github.com/chklovski/CheckM2
GTDB-Tk 1.3.0	Chaumeil et al.[55], Chaumeil et al.[60]	https://github.com/ecogenomics/gtdbtk
fastANI 1.0	Jain et al.[61]	https://github.com/ParBLiSS/FastANI
Roary 3.12.0	Page et al.[62]	https://github.com/sanger-pathogens/Roary
Diamond 2.0.4	Buchfink et al.[63]	https://github.com/bbuchfink/diamond
Prokka 1.11	Seemann et al.[64]	https://github.com/tseemann/prokka
iTOL 5.5.1	Letunic et al.[65]	https://itol.embl.de
Blast 2.2.30	Boratyn et al.[66]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=GeoBlast&PAGE_TYPE=BlastSearch

METHOD DETAILS

MAG and reference genome retrieval

Metagenome-Assembled Genomes (MAGs) belonging to Species-level Genome Bins (SGBs) taxonomically assigned to the genera *Ruminococcus*, *Pseudoruminococcus* and *Mediterraneibacter* were downloaded from the database curated in ref.[12]. Briefly, the MAGs were obtained by independently assembly the short reads of >9,000 human metagenomes using either MEGAHIT[67] or metaSPAdes.[68] The resulting contigs were therefore binned through metaBAT2,[69] taking into account both the contigs' tetranucleotide frequency and abundance. The completeness and contamination of each bin was estimated through CheckM,[70] and those showing completeness $\geq$50% and contamination $\leq$5% were considered MAGs and retained for further analyses. The SGBs were then obtained by clustering the MAGs according to their whole-genome nucleotide similarity using Mash.[71] Further details about MAGs and SGBs reconstruction and taxonomic assignment are available in the original paper.[12] Since quality estimates of MAGs in terms of completeness and contamination were computed through CheckM, we applied a refined filtering of MAGs based on CheckM2[59] estimates: high quality (HQ; completeness >90% and contamination <5%) and medium quality (MQ; completeness >50% and contamination <5%) MAGs. This resulted in the final set of 9,643 MAGs.

All the MAGs resulted to be reconstructed from human gut metagenomes. Metadata information about host characteristics in terms of geographical origin, lifestyle (Westernized vs. Non-Westernized), and health status (classified as healthy, diagnosis of colorectal cancer, diagnosis of Inflammatory Bowel Diseases, other disease).

Taxonomic assignment of each MAG was re-inferred using the GTDB-Tk tool (release 95; version v1.3.0; ref.[55]) through the 'classify_wf' pipeline, which firstly compares the provided genomes to a set of reference genomes through skani[72] then, if the previous step does not match minimum filtering thresholds, computes the maximum-likelihood placement of each genome in the GTDB-Tk reference genome tree through pplacer.[73]

The set of MAGs was integrated with reference genomes from isolate sources (''reference genomes'') for all the species belonging to the genera *Ruminococcus* (*n* = 12) and *Mediterraneibacter* (*n* = 10) available in the RefSeq NCBI database (Figure 1). This resulted in the final set of 9,644 MAGs and 21 genomes from isolate that were the basis for the downstream analyses (Table 1).

Bioinformatics processing of MAGs and reference genomes

We computed pairwise Average Nucleotide Identity (ANI) within each SGB through fastANI (version 1.0; ref.[61]). The resulting distance matrix (expressed as 1 - ANI) was then used to unveil putative subgroups through the Partition Around Medoids (PAM) method using the function 'pam' from the R 'vegan' package.

We used the GTDB-Tk tool to build a phylogenetic tree of the entire set of genomes (Figure 1). Briefly, the command 'identify' predicted coding sequences (CDS) from each MAG using Prodigal[74] and aligned the genes to a Hidden Markov Model (HMM) built on a set of 120 bacterial marker genes described in ref.[60]. For each genome, the 'align' step concatenated the alignment of the predicted genes to the set of 120 bacterial marker genes, and a phylogenetic tree is inferred from the Multiple Sequence Alignment obtained from the 'align' command through the 'infer' pipeline. Further details are available in the original papers[55,60] and in the tool's documentation. The phylogenetic tree was visualized and annotated through iTol.[65]

Coding sequences (CDS) of each genome were predicted using Prokka (version v1.11; ref.[64]). Pangenomes for each SGB were predicted using Roary (options ''-e -n''; version v3.12.0; ref.[62]). The reference sequences from each group of orthologs were mapped against the bacterial proteins from the UniRef90 database (Release 2023_04; ref.[57]) using Diamond blastx (options ''–evalue 0.00001'' and ''–very-sensitive''; version v2.0.4; ref.[63]).

We screened the genomes for adherence abilities by pooling the set of previously described genes (ref.[17]; Table S8) with all the genes involved in adhesion to the gut epithelium reported in the Virulence Factor Database (VFC0001; some of the genes from this class also confer the ability to adhere to abiotic surfaces, ref.[58]). Similarly, invasion abilities were screened by considering the genes encoding for invasion traits available in the VFDB. The mapping of the CDS of our genomes against the so defined adherence and invasion databases was done through Diamond blastx. Only matches with >80% of identity over at least 80% of the database sequences were considered.

The ability of genomes to degrade carbohydrates was estimated by mapping through blast the CDS of our genomes against the Carbohydrate-Active Enzyme database (CAZy; ref.[56], release 26 July 2023), Only queries aligning with >80% of identity over at least 80% of the database sequences were retained. The genes were functionally assigned at CAZy family level, and a presence-absence matrix was generated for all MAGs.

Finally, the genomes belonging to *Ruminococcus gnavus* (SGB 4584; *n* = 738) were screened for genes involved in the synthesis of a proinflammatory polysaccharide by considering the set 23 genes reported in ref.[16]. Mapping was performed through blastn (v2.2.30; ref.[66]; option ''-word_size 7''), and the completeness of the operon was calculated for each genome by dividing the number of genes linked with synthesis/regulation of the proinflammatory polysaccharide by 23 (i.e., length of the operon).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed by custom scripts developed in the R environment (https://www.r-project.org). Differences in the prevalence of genomes and genes across categories were assessed using the Pearson's χ^2 test (function 'chisq.test' from the R package 'base'). The normality distribution of numerical variables was tested with the Shapiro–Wilk test (function 'shapiro.test' from the R package 'base'), and variables non-normally distributed were compared between groups using the Wilcoxon rank-sum test (function 'wilcox.test' from the R package 'base'). When needed, P-values were corrected through FDR to account for multiple comparisons.

Principal Component Analysis (PCA) of the CAZy families presence-absence profiles was computed with the function 'prcomp' (R package 'base'). Furthermore, Principal Coordinate Analysis (PCoA) on the species-level ANI distance matrices was performed by the function 'cmdscale' from the 'base' R package.

Heatmaps were obtained using the R package 'pheatmap', and remaining plots were generated through the R package 'ggplot2'.