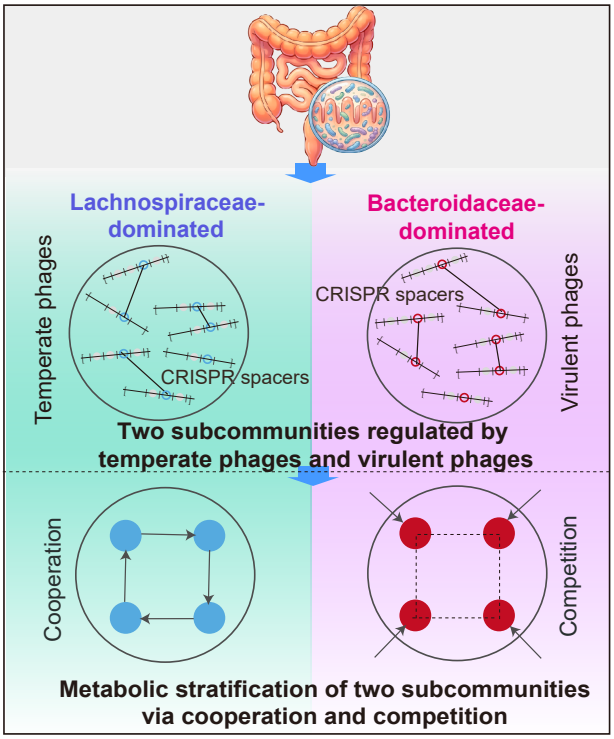
Cell Reports

Resource

Phage-related symbiosis and antagonism shape gut ecosystem dynamics in Lachnospiraceae and Bacteroidaceae

Graphical abstract



Highlights

- cMECOS recovers high- or medium-quality (HMQ) genomes from low-abundance gut bacteria
- HMQ genomes enable the construction of a comprehensive bacteria-phage interaction network
- Lachnospiraceae and Bacteroidaceae balance cooperative and competitive networks
- Phage-associated ecological networks are consistently disrupted in IBD and obesity

Authors

Yiming Zhao, Jiarui Li, Kai Han, ..., Hui Zeng, Zheng Yu, Chen Chen

Correspondence

zenghui@ccmu.edu.cn (H.Z.), yuzheng@csu.edu.cn (Z.Y.), chenchen1@ccmu.edu.cn (C.C.)

In brief

Zhao et al. develop cMECOS to recover 5,006 HMQ MAGs and reveal two gut bacterial networks linked to temperate/virulent phages, which are disrupted in IBD and obesity. These phage-related networks regulate gut homeostasis, offering insights for microbiome-targeted therapies.

Zhao et al., 2026, Cell Reports *45*, 117166
April 28, 2026 © 2026 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.celrep.2026.117166

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Phage-related symbiosis and antagonism shape gut ecosystem dynamics in Lachnospiraceae and Bacteroidaceae

Yiming Zhao,[1,6] Jiarui Li,[2,6] Kai Han,[2,6] Liang Chen,[2] Qinghui Zhuang,[2] Sha Li,[2] Mingxi Hua,[2] Na Li,[2] Jinglin Yue,[2] Chaoyang Gu,[2] Chengbo Rong,[2] Duo Yang,[2] Ziqing Deng,[3] Jing Huang,[4] Lian He,[5] Hui Zeng,[2,*] Zheng Yu,[1,*] and Chen Chen[2,7,*]

[1]Human Microbiome and Health Group, Department of Microbiology, Xiangya School of Basic Medical Sciences, Central South University, Changsha, China
[2]Biomedical innovation center and Beijing Key Laboratory for Therapeutic Cancer Vaccines, Beijing Shijitan Hospital, Capital Medical University, Beijing, China
[3]BGI-Beijing, Beijing 102601, China
[4]Department of Parasitology, Xiangya School of Basic Medical Sciences, Central South University, Changsha, China
[5]Department of Chemical Engineering, University of Washington, Seattle, WA, USA
[6]These authors contributed equally
[7]Lead contact
*Correspondence: zenghui@ccmu.edu.cn (H.Z.), yuzheng@csu.edu.cn (Z.Y.), chenchen1@ccmu.edu.cn (C.C.)
https://doi.org/10.1016/j.celrep.2026.117166

SUMMARY

The human gut microbiota is shaped by intricate, yet poorly resolved interactions among bacteria, as well as their relationship to bacteriophages. However, resolving this complex interaction and dynamics has been limited by the challenges in genome recovery and functional characterization. We develop culture-enriched metagenomic co-barcoding sequencing (cMECOS), obtain 5,006 high- or medium-quality (HMQ) metagenome-assembled genomes (MAGs) and reconstruct bacteria-phage interaction networks via CRISPR spacer mapping. This framework uncovers two ecologically distinct, inter-specific bacterial networks: a Lachnospiraceae-dominated community associates with temperate phages and is characterized by metabolic cross-feeding and a Bacteroidaceae-dominated community linked to virulent phages and marked by resource competition. Both network architectures are disrupted in both inflammatory bowel disease (IBD) and obesity (OB), underscoring their role in ecosystem stability. Our work establishes cMECOS as a powerful platform for deciphering complex microbiome interactions and identifies phage-related bacterial networks as critical regulators of gut homeostasis, providing a foundation for phage-informed therapeutic development.

INTRODUCTION

The human gut microbiota, often termed as a hidden "organ," forms an extraordinarily complex and dynamic ecosystem,[1] with a collective genetic repertoire that exceeds the human genome by more than 150-fold.[2] This microbial community plays a central role in host physiology, influencing metabolism, immune regulation, and disease susceptibility.[3–5] Pioneering metagenomic initiatives Human Microbiome Project (HMP, including two phases HMP1 and HMP2) have generated extensive genomic resources, including the Integrated Gene Catalog (IGC), the Unified Human Gastrointestinal Genome (UHGG) and Protein (UHGP) catalogs, and 204,938 genomes spanning from 4,644 bacterial species,[6] substantially advancing our understanding of gut microbial diversity and its implications for health.[7–9] Despite these advances, an estimated 60%–80% of gut bacteria remain uncultured due to either fastidious growth requirements or low abundance.[10] This limitation hinders our ability to build a complete ecological framework and to decipher the complex interactions that govern gut microbial communities.

Bacteriophages (phages) play a crucial role in shaping microbial ecosystems through predation, horizontal gene transfer, and metabolic reprogramming.[11–13] However, the diversity and functional roles of gut phages remain incompletely unresolved.[14] Many phages remain uncharacterized, and bacterial genomes retain records of prior phage encounters within CRISPR spacer arrays that serve as a molecular signature of potential historical infections.[15,16] Recent investigations have cataloged more than 142,000 non-redundant gut viral genomes, revealing complex phage-bacteria networks and offering new perspectives on microbial community organization.[17] Nevertheless, fully mapping phage-bacteria interaction networks in the human gut continues to pose significant technical challenges.

Most CRISPR arrays are embedded in short tandem repeats, which are notoriously difficult to assemble using short-read sequencing technologies. As a result, these regions are often

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

fragmented or misassembled, limiting our ability to extract key insights into phage-host interactions.[18] High-quality genomic data are therefore critical to resolving these challenges and reconstructing the intricate networks that shape microbial communities. To overcome limitations in characterizing low-abundance and unculturable bacteria taxa, several complementary strategies have been developed: (1) cultivation-based approaches that optimize growth conditions to isolate previously unculturable bacteria, followed by whole-genome sequencing[19,20]; (2) sequencing-based methods that leverage long-read technologies, such as PacBio SMRT and Oxford Nanopore[21]; and (3) improvements in assembly algorithms for reconstructing genomes from short-read data.[22] These methods have collectively enabled the recovery of thousands of high-quality genomes and have significantly expanded reference databases, thereby improving metagenomic analyses and enhancing our understanding of phage-bacteria interactions.

In this study, we developed culture-enriched metagenomic co-barcoding sequencing (cMECOS), a high-throughput approach that integrates community-level cultivation with single-molecule long-read sequencing.[23] Using cMECOS, we recovered 5,006 high- or medium-quality (HMQ) genomes and identified 38 previously not reported species-level and 105 previously not reported strain-level genome clusters, the majority of which belong to previously uncultured lineages. Through analyses of CRISPR spacer profiles in bacterial genomes, reconstructed genome-scale metabolic models, and computed metabolic interaction potential (MIP) and metabolic resource overlap (MRO) scores, we identified two distinct inter-specific relationships in the gut microbiota: symbiotic interactions (cooperation) in Lachnospiraceae with temperate phages and antagonistic interactions (competition) in Bacteroidaceae that are related to virulent phages. These dual interaction modes exhibited similar disease-specific perturbations, underscoring their potential clinical relevance and highlighting opportunities for phage-guided therapeutic intervention.

RESULTS

Expanding the phylogenomic diversity of human gut bacterial species

To comprehensively resolve the genomic landscape and phage-bacteria interactions within the human gut microbiota, we applied the cMECOS approach to fecal samples from 84 healthy Chinese volunteers (Table S1). This high-throughput approach generated 2.21 Tb of sequencing data across 252 libraries, comprising both untreated fecal samples and short-term liquid-cultured samples grown under aerobic and anaerobic conditions (Figure 1A). Compared to MECOS (i.e., the uncultured component of cMECOS), short-term cultivation in cMECOS reduced host DNA in 73.8% (62/84) of samples (Table S2 and Dataset S1), and preserved longer fragments of bacterial DNA (Figure S1A), which improved the efficiency of co-barcoding sequencing, ultimately producing longer contigs for genome assembly (Figure S1B and Dataset S2).

The further *de novo* assembly of the metagenomic reads yielded 1,446,896 contigs, which were clustered into 15,993 metagenome-assembled genomes (MAGs) (Figure 1A). Of these, 5,006 genomes met the UHGG HMQ criteria (genome completeness > 50%, contamination < 5%, and combined with an estimated quality score > 50) (Figure S1C and Dataset S3) and formed a comprehensive gut genome (CGG) dataset. Taxonomic profiling of the CGG dataset revealed 456 species-level clusters (95% average nucleotide identity [ANI]) and 2,439 strain-level (99% ANI) clusters (Figure 1A), with Firmicutes dominating the dataset (66.0%) followed by Bacteroidota (17.3%) (Figure 1B).

Because most "uncultured" gut microbes exist in low abundance (relative abundance < 0.1%), recovering HMQ MAGs from these taxa remains challenging.[10] To evaluate whether cMECOS improves the genome quality for such lineages, we compared assembled genomes generated from cMECOS and standard MECOS (uncultured fraction in cMECOS): cMECOS revealed enhanced relative abundance of low-abundance bacteria (Figure 1C and Dataset S2). Aerobic and anaerobic cultivation further provided complementary advantages, increasing the recovery of HMQ genomes (Figure S1D). Integrating these data enhanced the ability of recovering HMQ genomes in 254 species, of which 70.9% (180/254) were low-abundance bacteria (Table S3 and Dataset S2), and cMECOS has more than doubled the relative abundance of 68.9% (124/180) low-abundance species (Table S4 and Dataset S2). For example, the completeness and continuity of MAGs (stool sample) showed significant improvement in P37. The abundance of *Dorea formicigenerans* (Lachnospiraceae, Firmicutes) rose from 0.53% in uncultured conditions to 1.32% in cultured conditions, which subsequently increased the completeness of *Dorea formicigenerans* (Lachnospiraceae, Firmicutes) from 92.7% to 98.8%, with its N50 rising from 30.8 to 282.2 kb. Similarly, *Bacteroides faecis* (Bacteroidaceae, Bacteroidota) showed an increase in abundance from 0.16% to 0.25%, ts completeness improved from 64.1% to 96.0%, and N50 improved from 19.6 to 115.9 kb (Figure 1D).

We next assessed the quality of all genomes in each species- and strain-level clusters based on completeness, contamination, N50 length, and genome quality score. The CGG dataset provided superior representative genomes for 933 strain-level and 78 species-level clusters with UHGG data (Figures 1E and S1E). These representative genomes showed substantial improvement in completeness, contamination, N50 length, and the genome quality score (Figures S1F and S1G; Datasets S4 and S5), further confirming the effectiveness of cMECOS in reducing contamination and improving completeness and assembly contiguity.

Geographic tracing of the 5,006 HMQ genomes revealed 1,245 strain-level and 50 species-level clusters previously unobserved in Asian (Figures 1F and S1H), although MAGs for these clusters were present in European and American datasets (Figure S1H and Table S5). Prior work attributed such absences to geographic drivers of gut microbiota diversification. The recovery of these lineages from an Asian population highlights how improved methodologies overcome sampling biases and refine the global map of human gut microbial diversity. Additionally, cMECOS recovered 105 previously not reported strains (average nucleotide identity, ANI > 99%) and 38 previously not reported species (ANI > 95%) are absent from the UHGG dataset[6] (Figure 1G; Tables S6 and S7). Among 105 previously not reported strains, 70.4% (74/105) belong to Firmicutes and 19.0%

CellPress
OPEN ACCESS

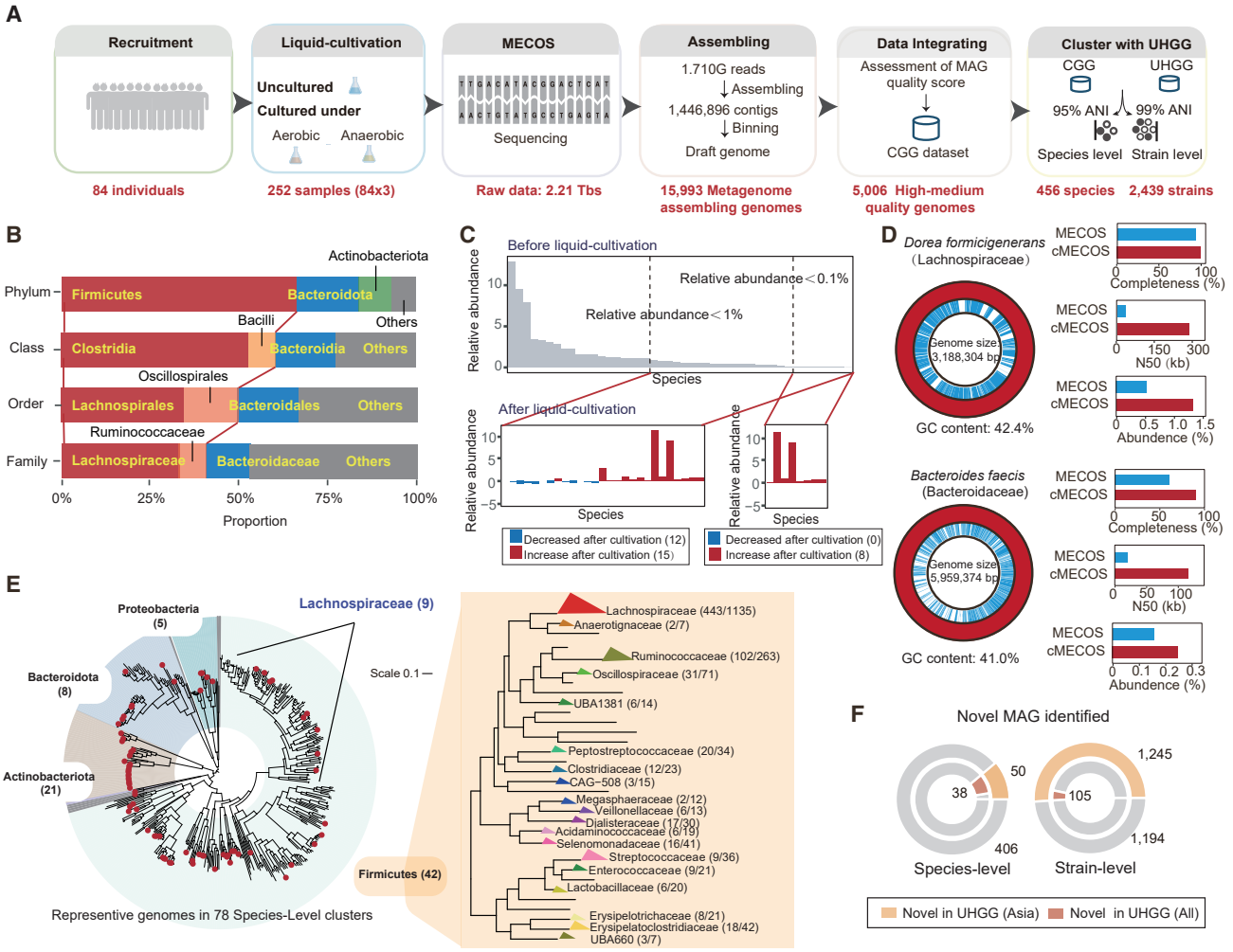


Figure 1. Culture-enriched metagenomic cMECOS enhances gut microbial genome quality and diversity

(A) Overview of the culture-enriched metagenomic co-barcoding sequencing (cMECOS). The cMECOS process integrates culture-enriched metagenomic sequencing and bioinformatics analysis, including the following six parts: (1) individual recruitment and fecal samples collection; (2) aerobic and anaerobic liquid culture of fecal samples; (3) standard MECOS sequencing for uncultured fecal samples, aerobic cultured samples, and anaerobic cultured samples; (4) reads filtering and assembly; (5) construction of HMQ genome; and (6) cluster with UHGG.

(B) Taxonomic distribution of the CGG dataset at phylum, class, order, and family levels. Only the top three taxa are shown at each taxonomic level, and the remaining ones are grouped as "others."

(C) Relative abundance of species derived from before liquid-cultivation and after liquid-cultivation samples. Comparison of abundance changes after cultivation was conducted for species with low abundance (relative abundance below 0.1%) and relative abundance below 1% and in the uncultured sample, respectively. Positive relative abundance values indicate an increase in relative abundance after cultivation (see also STAR Methods), which are colored red, while negative values indicate a decrease, which are colored dark blue.

(D) Schematic genome maps of cMECOS and MECOS. The genomic completeness, continuity improvements, and relative abundance for *Dorea formicigenerans* (Lachnospiraceae) and *Bacteroides faecis* (Bacteroidaceae) via cMECOS. The bar plots compare the completeness, the N50 value, and the relative abundance for each genome before and after the cMECOS process, suggesting an increase in completeness and N50.

(E) A maximum-likelihood phylogenetic tree constructed from 456 representative species-level CGG assembled in this study. Left: 456 species-level genomes from the CGG database, colored by phylum. Red circles indicate representative genomes (the genomes with higher quality scores [completeness − 5 × contamination + log10(N50)] compared to the UHGG database) from the CGG database; the number in parentheses represents the number of representative genomes provided by the CGG database. Right: expanded Firmicutes clade from the left tree, collapsed to family level; triangles annotate strain counts (the number of strains from CGG contributed representative genomes/total strain number). The triangle size corresponds to the proportion of strains contributed, and the color scale indicates the total number of strains.

(F) Previously not reported MAG identified. Compared with UHGG, the previously not reported microbes in the CGG database at the species level (ANI 95%) and strain level (ANI 99%). UHGG (Asia) is the sub-dataset extracted from UHGG, which contains all the microbial genomes obtained from the Asian continent.

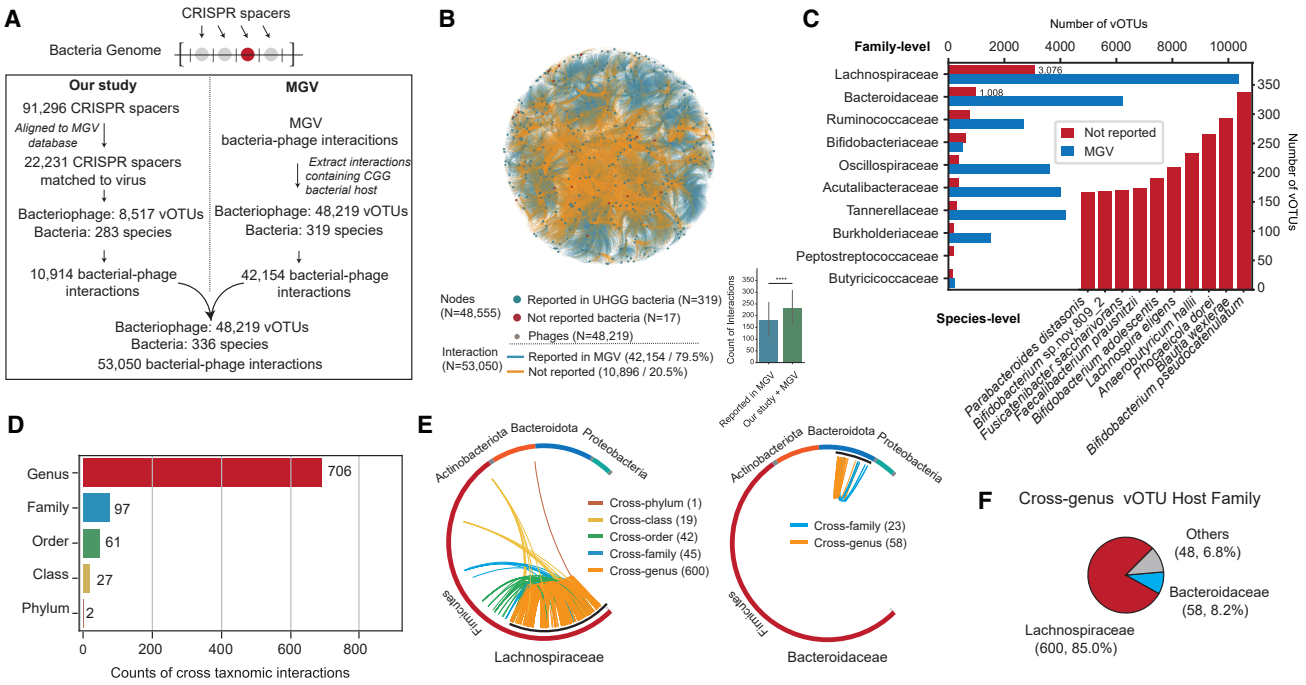


Figure 2. More comprehensive bacteria-phage interactions using cMECOS

(A) Schematic of the pipeline for identifying phages and bacterial host interactions. The left and right sides of the schematic diagram respectively show the bacterial phage relationships obtained using the CRISPR spacers in the bacterial genomes of this study and the bacterial phage relationships recorded in the MGV database.

(B) Co-occurring network showed the bacteria-phage interactions. The known interactions from MGV are colored blue and the not reported interactions identified in this study are colored orange. The reported bacterial species are colored dark blue, previously not reported bacterial species are colored red, and phages are colored gray. The bar chart at the lower right shows the number of phage-host infection relationships that only contain MGV records and the number of relationships after being included in this study. Data are presented as means ± standard of mean and two-tailed Mann-Whitney-Wilcoxon test is used. ****$p < 0.0001$.

(C) The bar plots showed the vOTU counts of top 10 bacterial families and bacterial species with most not reported bacterial host-phage interactions. The known interactions in MGV are colored in blue, and not reported are colored with red. The number of the plot represents the not reported vOTU counts targeting Lachnospiraceae and Bacteroidaceae.

(D) Distribution of cross-taxonomic phage interactions across taxonomic levels.

(E) Circular plots of bacterial hosts. Left: Lachnospiraceae and right: Bacteroidaceae with phage infection relationships. The outer rings indicate bacterial phyla (colored by phylum), and connecting lines represent cross-taxonomic infection relationships.

(F) Proportion of bacterial families of cross-genus phage interactions. Lachnospiraceae (red) and Bacteroidaceae (blue) are the most dominant bacterial families.

(20/105) to Actinobacteria (Figure S1I). Actinobacteria are often underrepresented in gut microbiome surveys due to their typically low abundance or high GC content (Table S7). Strikingly, we identified nine previously not reported species from the Coriobacteriaceae family (Actinobacteria), a taxonomic group rarely detected in adult guts but frequently found in early-life microbiota (Figure S1I and Table S6).

Together, these HMQ genomes offer valuable insights into the genomic diversity and ecological functions of previously understudied members of the gut microbiome.

Dense bacteria-phage interaction networks in the gut microbiome

Following bacterial genome construction via cMECOS, we identified 6,130 prophage regions, and 91,296 CRISPR spacers of which 22,231 matched to the viral sequences in the Metagenomic Gut Virus (MGV) database[1] (Figure 2A; Dataset S6). Then, bacteria-phage interactions were determined by directly comparing spacers and prophage sequences to establish homology. These interactions encompassed 1,695 prophage-derived (Dataset S7) and 8,517 spacer-associated viral operational taxonomic units (vOTUs) (Figure 2A and Dataset S8).

Mapping 22,231 CRISPR spacers to their bacterial hosts revealed 53,050 bacteria-phage interactions (Figure 2B). Among them, 20.5% (10,896/53,050) of these interactions were previously unobserved in the MGV database (Figure 2B). These interactions spanned 84.2% (283/336) of the bacterial species in our dataset (Figure 2B) and were concentrated in the Lachnospiraceae and Bacteroidaceae families, with 3,076 vOTUs targeting Lachnospiraceae and 1,008 vOTUs targeting Bacteroidaceae (Figure 2C).

Consistently, prophage analysis revealed 2,050 integration-based bacteria-phage interactions, which were also predominantly located within Lachnospiraceae (491 interactions) and Bacteroidaceae (742 interactions) (Figure S2A). Together, CRISPR and prophage evidences demonstrate that Lachnospiraceae and Bacteroidaceae are major hubs for dense and diverse bacteria-phage interactions in the human gut.

CellPress
OPEN ACCESS

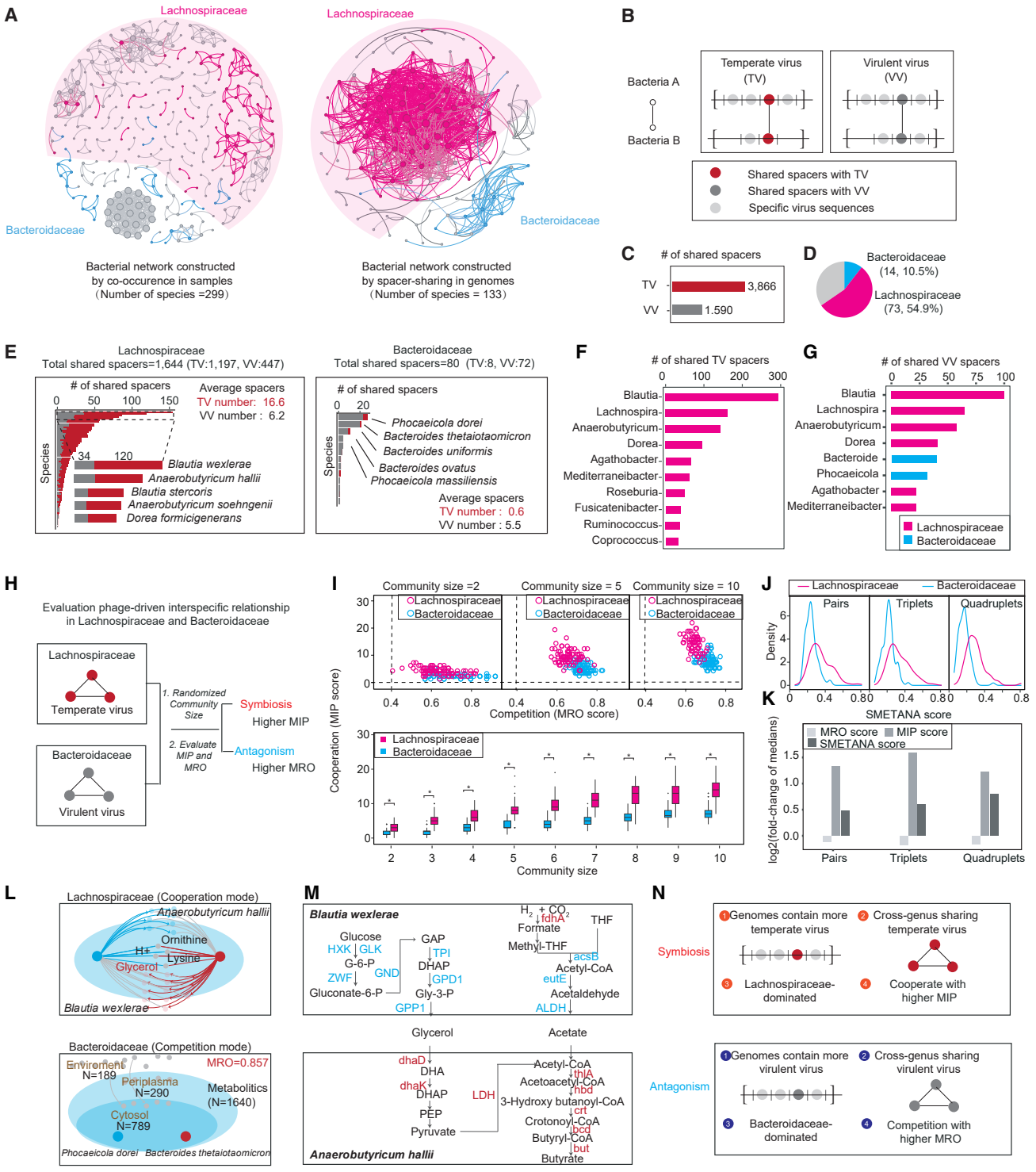


Figure 3. Dynamic balance between cooperation and competition in Lachnospiraceae- and Bacteroidaceae-dominated subcommunities

(A) Bacterial interaction networks constructed by two methods: co-occurrence network. Left: based on bacterial co-occurrence frequencies across 84 individuals, with edges weighted by correlation and CRISPR spacer-sharing network; right: based on shared CRISPR spacers. The co-occurrence network contains 299 nodes (bacterial species) and 650 edges; the spacer-sharing network contains 133 nodes and 913 edges. In both networks, nodes are colored by family: Lachnospiraceae (magenta, $n = 81$ species in co-occurrence network, $n = 73$ in spacer-sharing network) and Bacteroidaceae (blue, $n = 22$ species in co-occurrence network, $n = 14$ in spacer-sharing network); all other families are shown in light gray.

(B) Model of temperate phage (TV, red) and virulent phage (VV, dark gray) interactions with bacterial hosts.

(legend continued on next page)

Cross-taxonomic bacteria-phage interaction

To assess the taxonomic breadth of bacteria-phage interactions, we constructed ecological networks by connecting bacterial taxa that shared identical phage-derived CRISPR spacers, indicating encounters by the same phage strain. In total, we identified 706 cross-genus, 97 cross-family, 61 cross-order, 27 cross-class, and 2 cross-phylum interactions (Figure 2D; Dataset S9). Notably, *Bifidobacterium pseudocatenulatum* (Bifidobacteriaceae) and *Blautia wexlerae* (Lachnospiraceae) carried extensive CRISPR spacer arrays linked to prophages at the species level, revealing rich phage interaction histories (Figure 2C). However, their cross-taxonomic patterns sharply diverged: *B. pseudocatenulatum* showed few cross-taxonomic links and retained strict host boundaries largely confined to species or genus levels, whereas *B. wexlerae* displayed broad, promiscuous cross-lineage associations (Figure S3A).

Cross-taxonomic interactions mapping further revealed that Bacteroidaceae appeared to be targeted by a more specialized phage community. We observed 58 cross-genus and 23 cross-family interactions, but none of the connections extending beyond the order level (Figure 2E) suggested a relative restricted and host-constrained encounters pattern. In contrast, Lachnospiraceae exhibited a remarkably broad interaction spectrum, encompassing 600 cross-genus, 45 cross-family, 42 cross-order, 19 cross-class, and one cross-phylum associations (Figures 2E and 2F). Prophage-based analyses independently supported this pattern, revealing equally expansive cross-taxonomic linkages in Lachnospiraceae (Figures S2B and S2C).

Together, these results underscored extensive bacteriophage diversity and uncovered two contrasting encounter strategies in the gut microbiome: a specialized, constrained mode in Bacteroidaceae and a generalized, expansive mode in Lachnospiraceae (Figure S3B).

Lachnospiraceae and Bacteroidaceae dominate gut bacteria-bacteria networks

To characterize bacterial interaction architecture in the gut microbiome, we constructed two complementary bacteria-bacteria networks: a co-occurrence network based on bacterial species abundance correlations across individuals and a cross-taxonomic bacteria-phage interaction network (Figure 3A).

In the co-occurrence network, we identified 299 species as nodes and 650 edges representing significant co-occurrence relationships (correlation of variation [COV] > 0.6) (Figure S4A). Lachnospiraceae exhibited dense connections with other taxa, underscoring their central ecological role.

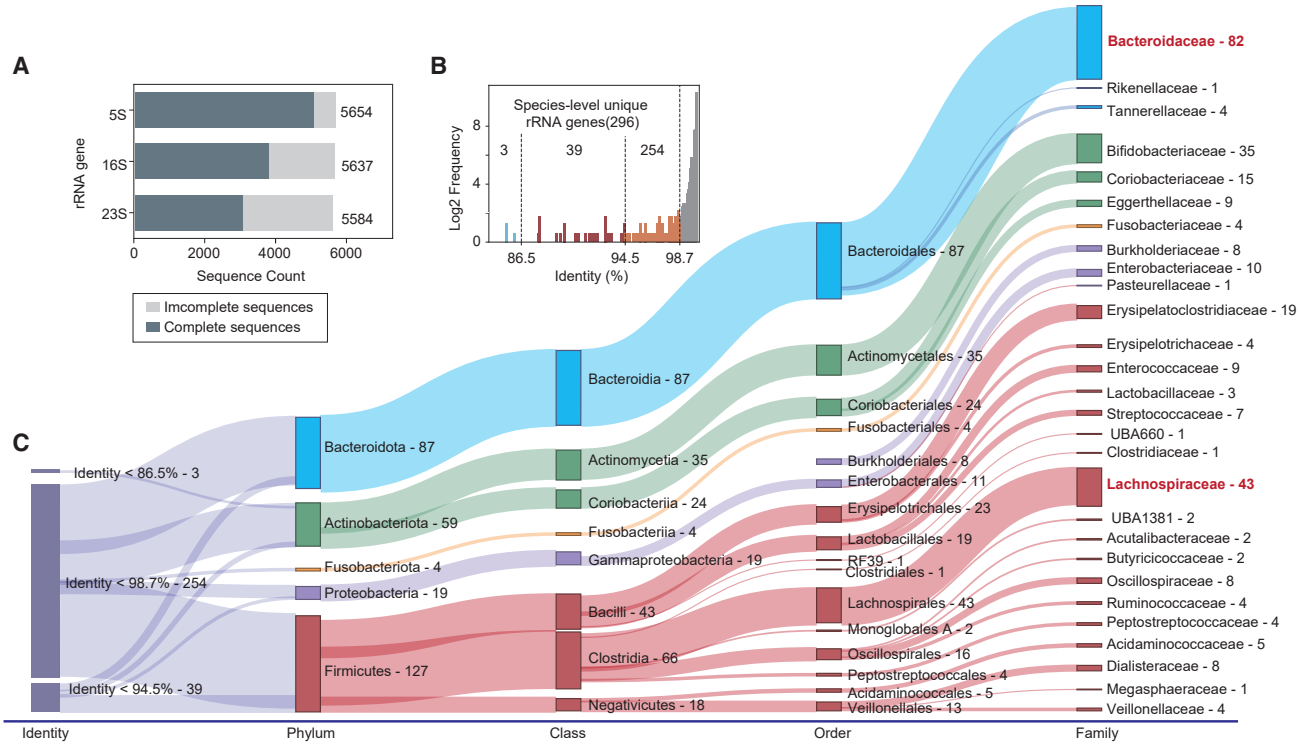
Given the limited number of cross-taxonomic interactions detected from prophages and the temperate-phage biases inherent to prophage-based inference, we constructed a spacer-sharing network by linking species from different genera that shared phage-derived CRISPR spacers (Figure 3A). This network comprised 133 nodes and 913 edges, demonstrating extensive cross-genus connectivity mediated by shared phage exposure (Figure S4B).

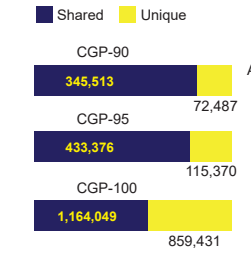
Collectively, these networks illustrate the dense and intricate architecture of bacterial interactions in the human gut, while Lachnospiraceae and Bacteroidaceae form two distinct subcommunities, each characterized by different bacteria-bacteria interaction patterns, associated with the underlying bacteria-phage connectivity.

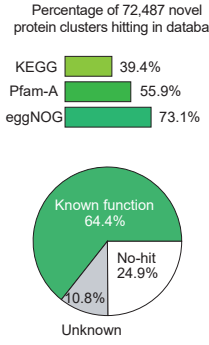
Two subcommunities regulated by temperate phages and virulent phages

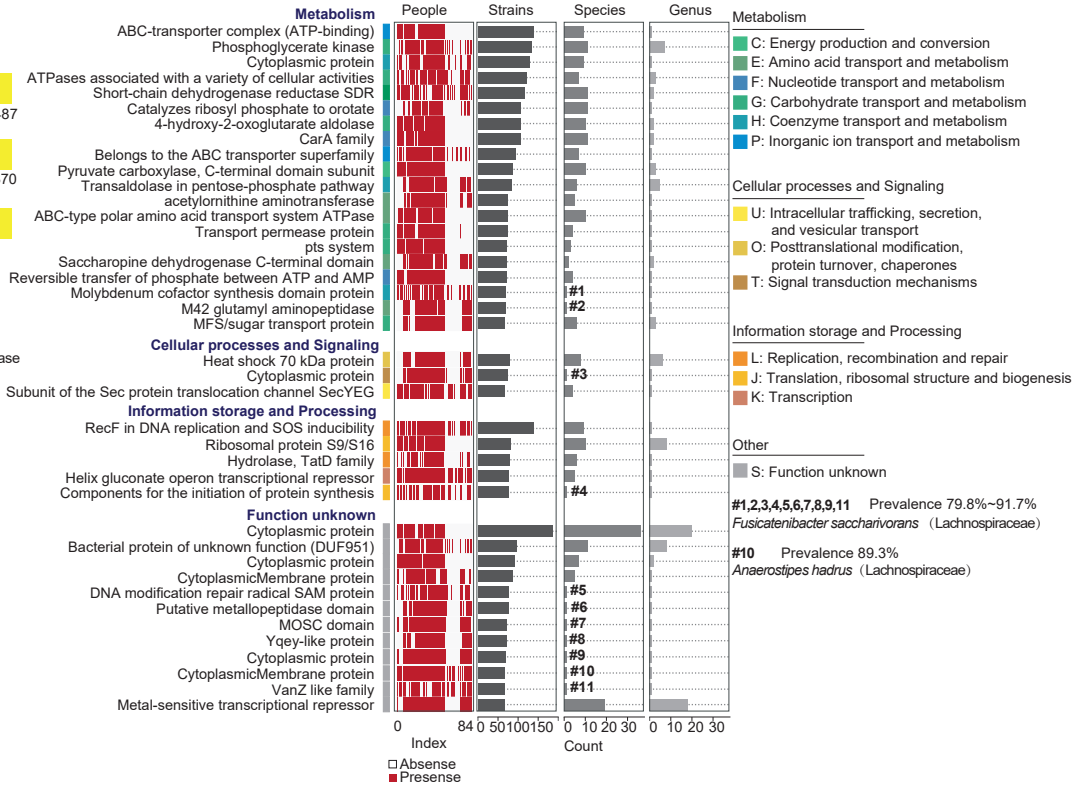
Using temperate scores calculated by BACPHLIP, we classified CRISPR spacers in each genome from Lachnospiraceae and Bacteroidaceae into two categories: those derived from temperate phages (TV) and those from virulent phages (VV)

(C) Distribution of temperate phages (TV, red) and virulent phages (VV, dark gray) in this study. The bar chart quantifies the number of identified phage vOTUs, with 3,866 classified as temperate (TV, red) and 1,590 as virulent (VV, dark gray).
(D) Proportion of bacterial families involved in TV and VV interactions. Lachnospiraceae (magenta) and Bacteroidaceae (blue) are highlighted.
(E) Shared TV and VV spacers in Lachnospiraceae (total shared spacers = 1,644) and Bacteroidaceae (total shared spacers = 80). The length of the red segment indicates the number of spacers targeting temperate phages, and the length of the dark gray segment indicates the number of spacers targeting virulent phages. (TV: red; VV: dark gray). The top 5 bacterial species, including *Blautia wexlerae* (120 spacers) in Lachnospiraceae and *Phocaeicola dorei* in Bacteroidaceae, with the highest number of associated phages are highlighted with a magnified inset, and their corresponding phage counts are annotated numerically above the respective bars.
(F) Genus-level distribution of shared TV spacers. Genera from Lachnospiraceae (magenta).
(G) Genus-level distribution of shared VV spacers. Genera from Lachnospiraceae (magenta) and Bacteroidaceae (blue) are labeled.
(H) Schematic of evaluating phage-driven interspecific relationship in Lachnospiraceae and Bacteroidaceae.The diagram outlines the workflow from defining species pairs/groups to simulating their metabolic networks and calculating cooperation (MIP, SMETANA) and competition (MRO) metrics.
(I) Metabolic interaction potential (MIP) and metabolic resource overlap (MRO) scores across community size categories. Up: the trade-off between the competition (MRO) and cooperation (MIP) scores for different community sizes. Magenta and blue dots show Lachnospiraceae and Bacteroidaceae communities, with a total of 100 simulated communities per families at each community size. Down: boxplot of MIP scores for different community size and families were shown. In the boxplots, the center line represents the median, the box limits indicate the upper and lower quartiles, and the whiskers extend to the 1.5 × interquartile range. *p* values were computed using the Wilcoxon rank-sum test.
(J) SMETANA scores for metabolic coupling strength in 100 simulated Lachnospiraceae and Bacteroidaceae subcommunities (paired, triplet, and quadruplet groups).
(K) Distinction of co-occurring subcommunities in various cooperation (MIP and SMETANA score) and competition (resource overlap, MRO scores) metrics. The effect size was calculated as the ratio of Lachnospiraceae to Bacteroidaceae. Data were represented as the mean of the 100 simulated subcommunities.
(L) Cooperative and competitive interaction models in Lachnospiraceae and Bacteroidaceae. The schematic contrasts the dominant interaction mode: Lachnospiraceae subcommunities primarily exhibit cooperative, cross-feeding interactions (magenta arrows), whereas Bacteroidaceae subcommunities exhibit more competitive, resource-overlap interactions (blue blocked arrows).
(M) Cross-feeding pathways for glycerol and acetate in *Blautia wexlerae* and *Anaerobutyricum hallii*.
(N) Schematic of symbiosis and antagonism driven by temperate phages and virulent phages. The model illustrates how shared temperate phages (red) can mediate gene transfer and stability between species (symbiosis), while virulent phages (dark gray) can selectively lyse hosts, creating a competitive advantage for resistant species (antagonism).









(legend on next page)

(Figure 3B). Lachnospiraceae species predominantly harbored temperate-phage signatures, averaging 16.6 TV spacers per species, whereas Bacteroidaceae contained comparatively few, averaging 0.6 per species (Figure 3E).

To explore how phage lifestyle influences bacteria-bacteria interactions in two subcommunities, we constructed separate bacteria-bacteria networks for TV- and VV-associated interactions (Figure S4C). Across all species, we identified 3,866 TV interactions (70.9%), compared to 1,590 VV interactions (29.1%), underscoring the predominance of temperate phages within the gut microbiota (Figure 3C). Within the TV network, 73 species from the Lachnospiraceae family contributed to 1,197 TV interactions (Figures 3D, 3E, and 3F). In contrast, bacteria-bacteria networks constructed by VV-associated interaction mostly occurred in the Bacteroidaceae family (Figures 3D, 3E, and 3G).

Together, these results indicate that the two subcommunities differ in both phage exposure profiles and phage lifestyle signatures, which likely generate distinct bacteria-bacteria interaction regimes that shape microbial ecology and stability in the gut.

Metabolic stratification of two subcommunities via cooperation and competition

To investigate the metabolic features underlying these two subcommunities, we reconstructed genome-scale metabolic models for representative microbial genomes and quantified metabolic interaction potential (MIP; cooperation) and metabolic resource overlap (MRO; competition) with bioinformatics analysis (Figure 3H; Dataset S10). Lachnospiraceae exhibited significantly higher MIP scores and lower MRO scores than Bacteroidaceae, and these differences remained robust in downsampling analyses (Figure 3I). In addition, the disparity in MIP scores increased with community size, indicating that the extensive metabolic cross-feeding became more pronounced with increasing community size (Figures 3I and S4D). Given the high prevalence of Lachnospiraceae species in our dataset, these cooperative metabolic networks likely constitute a cooperative fundamental feature in the healthy human gut. In contrast, Bacteroidaceae-dominated communities showed higher MRO scores, reflecting greater resource competition (Figures 3I and S3D).

To validate these observations, we quantified metabolic coupling strength in Lachnospiraceae and Bacteroidaceae using SMETANA through *in silico* predictions. Lachnospiraceae exhibited a 1.5-fold higher metabolic coupling strength compared to Bacteroidaceae (Figures 3J and 3K). Analysis of exchanged metabolites among co-occurring Lachnospiraceae identified nucleosides and sugars as the most frequent interaction currencies (Figure S3E), consistent with core metabolic dependencies described previously.[24]

We then examined the exchanged or competitive metabolites calculated by SMETANA within representative coupling pairs in Lachnospiraceae (*Blautia wexlerae* and *Anaerobutyricum hallii*) and Bacteroidaceae (*Bacteroides thetaiotaomicron* and *Phocaeicola dorei*). In Lachnospiraceae pairs, *Blautia wexlerae* and *Anaerobutyricum hallii* frequently exchanged glycerol, lysine, ornithine, and protons (H^+) (Figure 3L). Then, we mapped the annotated genes from MAGs of these representative coupling pairs to the map of metabolic pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) to estimate their potential metabolism pathway. The analysis revealed that glycerol transport reactions contributed to butyrate biosynthesis, a major short-chain fatty acid (SCFA), supporting gut health (Figure 3M). In contrast, *Bacteroides thetaiotaomicron* and *Phocaeicola dorei* exhibited significant overlap in metabolic products (Figure 3L), particularly for nutrients essential for cytosolic processes, consistent with their competitive resource interaction.

Together, these findings indicate that the two phage-related subcommunities, Lachnospiraceae enriched for temperate-phage exposure and Bacteroidaceae enriched for virulent-phage exposure, governed by distinct interaction regimes balancing cooperation and competition (Figure 3N). These contrasting modes likely play fundamental roles in shaping gut ecosystem structure and stability.

Exploration of unknown genes in two subcommunities

To characterize the genetic potential of the Lachnospiraceae and Bacteroidaceae subcommunities, we analyzed their gene content through comparative genomics and functional prediction approaches. We assessed genome quality within our set of 5,006 HMQ MAGs, identifying 5,654, 5,637, and 5,584 sequences corresponding to 5S, 16S, and 23S rRNA genes, respectively. Completeness analysis revealed that 89.3% (5,046) of 5S, 66.7% (3,760) of 16S, and 54.6% (3,046) of 23S rRNA genes were complete. Further analysis of the 3,760 complete 16S rRNA genes revealed 296 species-level unique rRNA genes (identity < 98.7%) and 42 genus-level unique sequences (identity < 94.5%) (Figures 4A and 4B). Notably, 125 of 296 rRNA genes originated from the Lachnospiraceae and Bacteroidaceae families (Figure 4C), reflecting their substantial genetic diversity and indicating a reservoir of unexplored genomic elements within these subcommunities.

We predicted 10,679,102 protein-coding genes in 5,006 MAGs. These genes were clustered into 1,114,389, 1,302,162,

Figure 4. Bridging the rRNA gene knowledge gap and expanding the gut protein landscape

(A) The number of complete and partial rRNA genes (5S, 16S, and 23S) in the CGG database. Partial genes were annotated by Prokka as "incomplete."

(B) Mapping of complete 16S rRNA genes from the CGG database to integrated databases (rrnDB, SILVA, and Greengenes2). Previously not reported genus- and species-level rRNA genes are stratified by identity thresholds (88.5% and 94.5%), and the total gene number is also marked on each threshold.

(C) Taxonomic distribution of unmapped 16S rRNA genes (Sankey diagram). Phylum, class, order, and family levels are shown; gene counts are labeled.

(D) The number of shared and unique protein clusters between UHGP and CGP databases at 90%, 95%, and 100% amino acid identity (AAI) thresholds.

(E) Functional annotation of 72,487 previously not reported protein clusters (CGP-90 vs. KEGG, Pfam-A, and eggNOG). Clusters with no hits or "unknown function" in all databases are labeled "Unknown."

(F) The prevalence of previously not reported protein clusters (AAI 90%) across individuals, strains, species, and genera. Only clusters detected in >60 strains are shown (shared *y* axis). Species-specific clusters (detected in one species) are marked with "#."

and 3,248,837 gene clusters at 90%, 95%, and 100% amino acid identity (AAI) thresholds in the UHGG database. Remarkably, 6.5%–26.5% of these genes (72,487–859,431) were unannotated, with the proportion of previously not reported genes increasing as the AAI threshold became more stringent (Figure 4D). Functional annotation using KEGG, eggNOG, and Pfam-A databases revealed that 35.6% (25,827 of 72,487) of these genes could not be assigned to any known function (Figure 4E), and another 64.4% (46,660 of 72,487) of the genes had putative functional annotations. To investigate further, we traced the source genomes of these previously not reported genes through binning analysis. Among them, 66% (47,846/72,487) were derived from uncultured taxa, which are indexed only as MAGs in the UHGG database. (Figure S5A), emphasizing their association with previously unidentified gut microbiota. To exclude assembly artifacts, only genes shared by $\geq$2 strains were considered. Functional analysis of these shared genes revealed a strong enrichment in information storage and processing categories—such as replication, recombination, and repair (Figure S5B)—likely reflecting the inherent difficulty of assembling repetitive sequences in MAGs.

Among these, we identified 40 high-prevalence genes present in over 50% of individuals and more than 60 strains (Figure 4F). Based on their distribution across species, we classified these genes into two categories: core genes, shared across multiple species or genera, and species-specific genes, predominantly found in a single species. Functional profiling revealed that the core genes primarily encode proteins involved in essential metabolic processes, including ABC transporters and phosphoglycerate kinase (Figure 4E). In contrast, species-specific genes were commonly concentrated in Lachnospiraceae, notably in *Fusicatenibacter saccharivorans* (genes #1–9 and #11) and *Anaerostipes hadrus* (gene #10). Functional annotation suggested that these genes are rapidly evolving and involved in DNA modification and cell membrane processes (Figure S5C), potentially contributing to phage defense mechanisms via modification of restriction-modification systems.

Collectively, these uncharacterized genes represent a substantial reservoir of molecular diversity in Lachnospiraceae, with potential roles in phage defense, metabolic specialization, and interbacterial competition, meriting targeted experimental validation.

Disease-associated dysregulation of cross-feeding and competition in subcommunities

To determine whether these subcommunities are perturbed in disease, we performed cross-sectional analysis of cross-feeding and competition pairs (CPCCs) in two public cohorts: inflammatory bowel disease (IBD, 505 patients[25]) and obesity (OB, 73 patients[26]) (Figure 5A; Dataset S11). Among 86 species, 73 from Lachnospiraceae and 13 from Bacteroidaceae, that shared inter-genus spacers, all species were present in either IBD or OB. Notably, 659 and 660 of 665 co-occurrences were preserved in IBD and OB, underscoring the functional relevance of these subcommunities.

We further assessed symbiosis (via Lachnospiraceae-subpopulations) and antagonism (via Bacteroidaceae-dominated subcommunities) by quantifying CPCC occupancy and weighted abundance and average bacterial abundance. Both diseases exhibited disrupted symbiosis networks while enhancing antagonistic interactions. Specifically, IBD exhibited reduced CPCC occupancy and OB showed reduced total bacterial abundance in coupling pairs (Figure 5B). Analysis of top CPCC connections confirmed these patterns (Figures S6A and S6B). The consistency of these network-level changes across taxonomic differences suggests that functional ecological relationships follow conserved principles in dysbiosis.

The findings establish an ecological framework that links distinct phage lifestyles to divergent bacterial network states in disease. The collapse of Lachnospiraceae symbiosis and the dominance of Bacteroidaceae antagonism are strongly associated with temperate and virulent phages, as evidenced by their respective CRISPR spacer profiles, suggesting that phages may play an important role in shaping dysbiotic community structures.

DISCUSSION

In this study, we developed a high-throughput platform (cMECOS) that integrates culturomics with single-molecule metagenomic sequencing to generate high-resolution genomic data and recover previously not reported representative genomes. This approach enabled the reconstruction of HMQ genomes from previously underrepresented bacterial taxa and revealed two distinct subcommunities, Lachnospiraceae and Bacteroidaceae, that maintain ecologically significant roles despite their low relative abundances. Our analyses demonstrate that these bacterial subcommunities exhibit fundamentally different ecological strategies: Lachnospiraceae engage in cooperative metabolic networks, while Bacteroidaceae employ competitive resource utilization patterns. These distinct bacterial interaction styles are associated with and potentially shaped by their respective phage associations: temperate phages with Lachnospiraceae and virulent phages with Bacteroidaceae.

Members of the Bacteroidaceae family, particularly Bacteroides spp., exemplify exploitative competition, preferentially catabolizing polysaccharides for individual metabolic benefit rather than participating in communal cross-feeding. This competitive behavior is reinforced by bacteriocin production (e.g., BfUbb, BSAP-1, and BSAP-2), mediating direct interbacterial antagonism.[27] Our study extends this framework by revealing dense ecological networks among Bacteroides species, interconnected through shared genomic traces of wide-range virulent phages. This finding, supported by both CRISPR-spacer and prophage analyses, underscores a layer of phage-mediated connectivity. Subsequent metabolic network modeling delineated potential inter-genus competitive relationships among Bacteroides, although the specific characteristics of these interactions will require validation through targeted cross-feeding and competition experiments. Notably, in dysbiotic states such as OB and IBD, keystone taxa governing these antagonistic networks shifted, indicating that phage-linked competition networks are sensitive to disease-driven pressures and can reshape broader community structures.

Lachnospiraceae contribute to host physiology through carbohydrate fermentation to SCFAs, such as butyrate, which

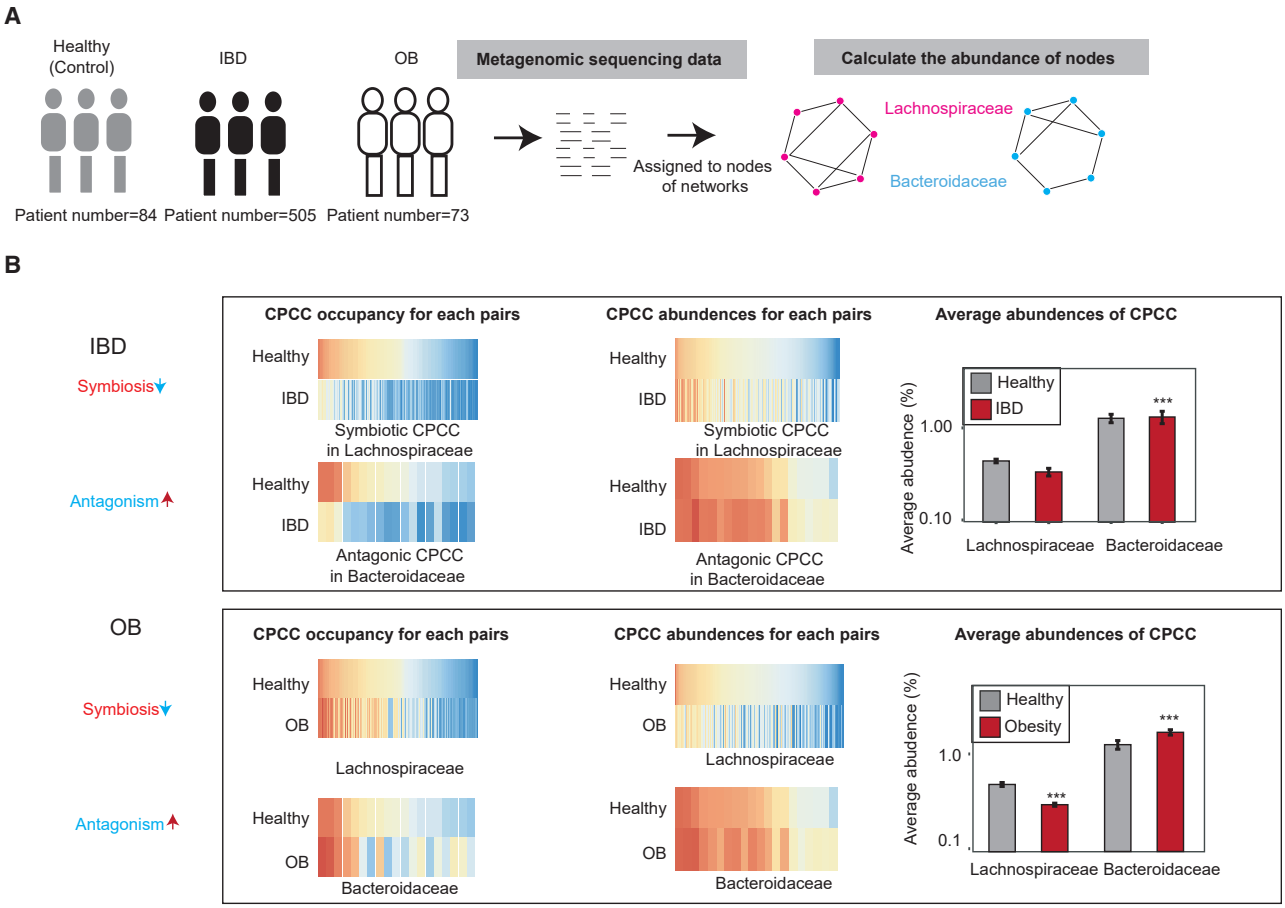


Figure 5. CPCCs in IBD and OB cohorts

(A) Schematic of Cross-feeding and competition pair (CPCC) analysis in IBD (n = 505) and OB (n = 73). CPCCs were obtained from the previously defined phage-linked subcommunities of 456 species, and their co-occurrence patterns were assessed in IBD and OB cohorts using co-occurrence proportion and weighted abundance metrics. The co-occurrence proportion, defined as the fraction of samples where both taxa were present at a relative abundance $> 1 \times 10^{-5}$, and the weighted abundance, defined as the mean relative abundance of the co-occurring taxa across all CPCCs within a subcommunity (see also STAR Methods).

(B) CPCC occupancy and weighted abundance distributions between IBD and OB cohorts. Bootstrap validation (100 iterations, n = 50 individuals per group) confirmed metric robustness. Median co-occurrence proportions and weighted abundances (normalized per group) were compared using Wilcoxon rank-sum tests with Benjamini-Hochberg false discovery rate correction (α = 0.01). CPCC abundance distributions were log-scaled. Bar graphs show the mean weighted abundance across 100 bootstrap iterations, with error bars representing the standard deviation.***$p < 0.001$.

support intestinal barrier integrity and metabolic health.[28] However, SCFAs can also enhance energy harvest, potentially contributing to OB.[29] For instance, *Blautia wexlerae* participates in symbiotic cross-feeding with butyrate-producing species within Lachnospiraceae and exhibits extensive viral spacer sharing, which promotes phage resistance and cooperative metabolic interactions.[30] Our genomic analysis confirms that Lachnospiraceae are predominantly associated with temperate phages, supported by both CRISPR-spacer and prophage data. Metabolic network reconstruction highlights their high cooperative potentials. In addition, we also identified 236 auxiliary metabolic genes (AMGs) in 143 of 429 prophages in Lachnospiraceae. Further analysis of the AMGs of related prophages suggests links to bacterial metabolism, notably in vitamin B3 pathways that may facilitate broader interactions (Figure S7; Dataset S7). Also, this prevalence of AMGs aligns with recent studies of human-associated prophages,[31] suggesting that the potential roles of bacteriophages in modulating bacteria-bacteria interactions for metabolism and conferring fitness advantages.

Importantly, Lachnospiraceae-dominated cooperative networks were consistently disrupted in both OB and IBD, indicating that the collapse of these bacterial consortia represents a common pathway in dysbiosis, irrespective of disease etiology. However, the manifestation of this collapse differed mechanistically between conditions. In IBD, network simplification involved reduced cooperating taxa, whereas in OB, perturbations manifested functionally through altered metabolite abundance. While these findings derive from bioinformatic inference, the precise ecological interactions—both bacterial and phage-mediated—require experimental validation to confirm their roles in gut ecosystem stability.

Collectively, our results highlight the evolutionary trade-offs between cooperative and competitive strategies in gut microbial ecosystems. Although the interaction manner, contact-dependent or

independent, remains to be explored, these findings provide a conceptual framework for understanding how phages simultaneously drive diversity and enforce stability. This dual role underscores the potential of phage-related interventions to restore dysbiosis states and suggests that therapeutic strategies combining phages with probiotics or prebiotics could target the gut microbiome more effectively.

Limitations of the study

This study has several limitations. First, although cMECOS enhances genome recovery from low-abundance species, cultivation biases may influence inferred ecological relationships. Future work should incorporate broader cultivation conditions to better recapitulate *in vivo* microbial interactions. Second, both the bacteria-phage and bacteria-bacteria interactions were inferred computationally and require experimental validation to establish their biological relevance. Finally, although we demonstrate that bacterial families are preferentially targeted by phages with distinct lifestyles, and that these ecological networks shift consistently in disease states, the precise mechanisms underlying phage-mediated modulation of gut ecosystem homeostasis, including whether AMGs contribute to this process, remain to be fully elucidated.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Chen Chen (chenchen1@ccmu.edu.cn).

Materials availability

All materials are available upon reasonable request to the lead contact.

Data and code availability

- The sequencing data have been deposited at the National Genomics Data Center, China. Databases: CRA021198 and CRA008331 (https://ngdc.cncb.ac.cn/gsa/browse/CRA021198 and https://ngdc.cncb.ac.cn/gsa/browse/CRA008331, respectively).
- The original code supporting this study has been deposited at the GitHub repository ChenChen-Lab/cMECOSand and is publicly available via Zenodo under the permanent identifier https://doi.org/10.5281/zenodo.18637575.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank the members of the Biomedical Innovation Center for general discussion and comments on the manuscript. We thank L. Zhang from the Chinese Institute For Brain Research, Beijing, and M. Ni from the Academy of Military Medical Sciences for their technical assistance and/or discussions. We thank all volunteers for providing stool samples. We thank the BGI for providing sequencing services. This work was supported by the following: Beijing Municipal Health Commission "high-level public health talents" project, China (xuekedaitouren-03-13 to C.C.), Beijing Municipal Health Commission "high-level public health talents" project, China (lingjunrencai-02-06 to H.Z.); Beijing Hospital Authority, China (no. DFL20240701); National Natural Science Foundation of China, China (no. 82272339); National Natural Science Foundation of China, China (no. 82102447), and Beijing Natural Science Foundation , China (no. 7252073).

AUTHOR CONTRIBUTIONS

C.C., Z.Y., and H.Z. conceived of the project. K.H., L.C., M.H., and Q.Z. planned the experiments and conducted data analyses. N.L., D.Y., C.R., S.L., J.H., and Z.D. performed the experiments. J.L., Y.Z., J.Y., and C.G. analyzed the experimental data and carried out bioinformatic analyses. L.H., K.H., J.L., Y.Z., C.C., Z.Y., and H.Z. wrote the manuscript. All authors commented on the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT in order to generate the intestinal illustration used in the graphical abstract. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Culture-enriched metagenomics co-barcoding sequencing (cMECOS) pipeline
  - Feces collection and multiple lipid culture enrichment approaches
  - High-molecular-weight DNA extraction
  - Metagenomics co-barcoding library preparation and sequencing
  - Reads processing, co-barcode assembling, and genome binning
  - Completeness, contamination, genome quality score calculation and HMQ genome definition
  - Cluster analysis from genomes to strains and species
  - Previously not reported species-level clusters and strains-level cluster definition
  - rRNA gene analysis in microbial genomes
  - Gene prediction, clustering, and annotation
  - Bacteria-phage interactions establishment
  - Prophage region identification
  - Determine the temperate and virulent phages for CRISPR spacer associated bacteriophages
  - Bacteria-bacteria interaction network for two type phages
  - Metabolic relationship profiling of bacteria-bacteria interaction by their metabolic scores
  - Ecological validation of cross-feeding and competition pairs (CPCCs)
  - AMG gene detection
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2026.117166.

Received: July 17, 2025
Revised: February 10, 2026
Accepted: March 3, 2026

REFERENCES

1. Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., Hugenholtz, P., and Kyrpides, N.C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat. Microbiol. *6*, 960–970. https://doi.org/10.1038/s41564-021-00928-6.

2. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65. https://doi.org/10.1038/nature08821.

3. Chu, H., Khosravi, A., Kusumawardhani, I.P., Kwon, A.H.K., Vasconcelos, A.C., Cunha, L.D., Mayer, A.E., Shen, Y., Wu, W.L., Kambal, A., et al. (2016). Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. Science *352*, 1116–1120. https://doi.org/10.1126/science.aad9948.

4. Fujimura, K.E., Sitarik, A.R., Havstad, S., Lin, D.L., Levan, S., Fadrosh, D., Panzer, A.R., LaMere, B., Rackaityte, E., Lukacs, N.W., et al. (2016). Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. Nat. Med. *22*, 1187–1191. https://doi.org/10.1038/nm.4176.

5. Hou, K., Wu, Z.X., Chen, X.Y., Wang, J.Q., Zhang, D., Xiao, C., Zhu, D., Koya, J.B., Wei, L., Li, J., and Chen, Z.S. (2022). Microbiota in health and diseases. Signal Transduct. Target. Ther. *7*, 135. https://doi.org/10.1038/s41392-022-00974-4.

6. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. *39*, 105–114. https://doi.org/10.1038/s41587-020-0603-3.

7. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

8. Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192. https://doi.org/10.1038/s41587-018-0009-7.

9. 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.

10. Lagier, J.C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.M., Fournier, P.E., and Raoult, D. (2018). Culturing the human microbiota and culturomics. Nat. Rev. Microbiol. *16*, 540–550. https://doi.org/10.1038/s41579-018-0041-0.

11. Hsu, B.B., Gibson, T.E., Yeliseyev, V., Liu, Q., Lyon, L., Bry, L., Silver, P.A., and Gerber, G.K. (2019). Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. Cell Host Microbe *25*, 803–814.e5. https://doi.org/10.1016/j.chom.2019.05.001.

12. Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA *96*, 3801–3806. https://doi.org/10.1073/pnas.96.7.3801.

13. Reyes, A., Wu, M., McNulty, N.P., Rohwer, F.L., and Gordon, J.I. (2013). Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc. Natl. Acad. Sci. USA *110*, 20236–20241. https://doi.org/10.1073/pnas.1319470110.

14. Shen, J., Zhang, J., Mo, L., Li, Y., Li, Y., Li, C., Kuang, X., Tao, Z., Qu, Z., Wu, L., et al. (2023). Large-scale phage cultivation for commensal human gut bacteria. Cell Host Microbe *31*, 665–677.e7. https://doi.org/10.1016/j.chom.2023.03.013.

15. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712. https://doi.org/10.1126/science.1138140.

16. Canuti, M., and van der Hoek, L. (2014). Virus discovery: are we scientists or genome collectors. Trends Microbiol. *22*, 229–231. https://doi.org/10.1016/j.tim.2014.02.004.

17. Camarillo-Guerrero, L.F., Almeida, A., Rangel-Pineros, G., Finn, R.D., and Lawley, T.D. (2021). Massive expansion of human gut bacteriophage diversity. Cell *184*, 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.

18. Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016). Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol. Rev. *40*, 258–272. https://doi.org/10.1093/femsre/fuv048.

19. Lagier, J.C., Khelaifia, S., Alou, M.T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S.I., Seck, E.H., et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. Nat. Microbiol. *1*, 16203. https://doi.org/10.1038/nmicrobiol.2016.203.

20. Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185. https://doi.org/10.1038/s41587-018-0008-8.

21. Jin, H., Quan, K., He, Q., Kwok, L.Y., Ma, T., Li, Y., Zhao, F., You, L., Zhang, H., and Sun, Z. (2023). A high-quality genome compendium of the human gut microbiome of Inner Mongolians. Nat. Microbiol. *8*, 150–161. https://doi.org/10.1038/s41564-022-01270-1.

22. Cleary, B., Brito, I.L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E.J. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat. Biotechnol. *33*, 1053–1060. https://doi.org/10.1038/nbt.3329.

23. Han, K., Li, J., Yang, D., Zhuang, Q., Zeng, H., Rong, C., Yue, J., Li, N., Gu, C., Chen, L., and Chen, C. (2024). Detecting horizontal gene transfer with metagenomics co-barcoding sequencing. Microbiol. Spectr. *12*, e0360223. https://doi.org/10.1128/spectrum.03602-23.

24. Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D.R., Bork, P., and Patil, K.R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc. Natl. Acad. Sci. USA *112*, 6449–6454. https://doi.org/10.1073/pnas.1421834112.

25. Kumbhari, A., Cheng, T.N.H., Ananthakrishnan, A.N., Kochar, B., Burke, K.E., Shannon, K., Lau, H., Xavier, R.J., and Smillie, C.S. (2024). Discovery of disease-adapted bacterial lineages in inflammatory bowel diseases. Cell Host Microbe *32*, 1147–1162.e12. https://doi.org/10.1016/j.chom.2024.05.022.

26. Hoyles, L., Fernández-Real, J.M., Federici, M., Serino, M., Abbott, J., Charpentier, J., Heymes, C., Luque, J.L., Anthony, E., Barton, R.H., et al. (2018). Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. Nat. Med. *24*, 1070–1080. https://doi.org/10.1038/s41591-018-0061-3.

27. Tong, M., Xu, J., Li, W., Jiang, K., Yang, Y., Chen, Z., Jiao, X., Meng, X., Wang, M., Hong, J., et al. (2024). A highly conserved SusCD transporter determines the import and species-specific antagonism of Bacteroides ubiquitin homologues. Nat. Commun. *15*, 8794. https://doi.org/10.1038/s41467-024-53149-w.

28. Hosomi, K., Saito, M., Park, J., Murakami, H., Shibata, N., Ando, M., Nagatake, T., Konishi, K., Ohno, H., Tanisawa, K., et al. (2022). Oral administration of Blautia wexlerae ameliorates obesity and type 2 diabetes via metabolic remodeling of the gut microbiota. Nat. Commun. *13*, 4477. https://doi.org/10.1038/s41467-022-32015-7.

29. Takeuchi, T., Kameyama, K., Miyauchi, E., Nakanishi, Y., Kanaya, T., Fujii, T., Kato, T., Sasaki, T., Tachibana, N., Negishi, H., et al. (2023). Fatty acid overproduction by gut commensal microbiota exacerbates obesity. Cell Metab. *35*, 361–375.e9. https://doi.org/10.1016/j.cmet.2022.12.013.

30. Liu, X., Mao, B., Gu, J., Wu, J., Cui, S., Wang, G., Zhao, J., Zhang, H., and Chen, W. (2021). Blautia-a new functional genus with potential probiotic properties. Gut Microbes *13*, 1–21. https://doi.org/10.1080/19490976.2021.1875796.

31. Dikareva, E., Matharu, D., Lahtinen, E., Kolho, K.L., De Vos, W.M., Salonen, A., and Ponsero, A.J. (2023). An extended catalog of integrated prophages in the infant and adult fecal microbiome shows high prevalence of

lysogeny. Front. Microbiol. *14*, 1254535. https://doi.org/10.3389/fmicb.2023.1254535.

32. Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience *7*, 1–6. https://doi.org/10.1093/gigascience/gix120.
33. Langdon, W.B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min. *8*, 1. https://doi.org/10.1186/s13040-014-0034-0.
34. 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.
35. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.
36. Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat. Biotechnol. *36*, 1067–1075. https://doi.org/10.1038/nbt.4266.
37. 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.
38. 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.
39. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868. https://doi.org/10.1038/ismej.2017.126.
40. 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.
41. 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.
42. Aroney, S.T.N., Newell, R.J.P., Nissen, J.N., Camargo, A.P., Tyson, G.W., and Woodcroft, B.J. (2025). CoverM: read alignment statistics for metagenomics. Bioinformatics *41*, btaf147. https://doi.org/10.1093/bioinformatics/btaf147.
43. 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.
44. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.
45. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.
46. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.
47. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
48. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121. https://doi.org/10.1093/nar/gkt263.
49. 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.
50. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251. https://doi.org/10.1093/nar/gky425.
51. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.
52. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.
53. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 69. https://doi.org/10.1186/s40168-017-0283-5.
54. Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S.G., Nayfach, S., and Kyrpides, N.C. (2024). Identification of mobile genetic elements with geNomad. Nat. Biotechnol. *42*, 1303–1312. https://doi.org/10.1038/s41587-023-01953-y.
55. Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome *8*, 90. https://doi.org/10.1186/s40168-020-00867-0.
56. Machado, D., Andrejev, S., Tramontano, M., and Patil, K.R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. *46*, 7542–7553. https://doi.org/10.1093/nar/gky537.
57. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
58. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.
59. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.
60. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.
61. Hockenberry, A.J., and Wilke, C.O. (2021). BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ *9*, e11396. https://doi.org/10.7717/peerj.11396.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
GAM anaerobic bottles	GermountX Health Technology	Cat#KC001
GAM aerobic bottles	GermountX Health Technology	Cat#KC002
MagAttract HMW DNA Kit	QIAGEN	Cat#67563
dsDNA Assay Kit	ExCell Bio	Cat#NGS00-3022
MGIEasy stLFR	MGI	Cat#1000021745
Standard cartridge	Houze Bio-tech	Cat#C105101
Deposited data		
cMECOS data for 84 healthy Chinese volunteers	This paper	NGDC: CRA021198, CRA008331
Shotgun metagenomic sequences for IBD cohorts	Kumbhari et al.[25]	NCBI: PRJNA993675
Shotgun metagenomic sequences for obesity cohorts	Hoyles et al.[26]	NCBI: PRJEB14215
Software and algorithms		
SOAPfilter v2.2	Chen et al.[32]	https://github.com/BGI-Qingdao/stlfr2supernova_pipeline/tree/master/bin
Bowtie2 v2.2.5	Langdon et al.[33]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
metaSPAdes v3.13.1	Nurk et al.[34]	https://github.com/ablab/spades
BWA v0.7.15	Li et al.[35]	https://github.com/lh3/bwa
Athena v1.2	Bishara et al.[36]	https://github.com/abishara/athena_meta
SAMtools v 1.9	Li et al., 2009	http://samtools.sourceforge.net/
MetaBAT2 v2.12.1	Kang et al.[37]	https://anaconda.org/bioconda/metabat2
CheckM v1.2.0	Parks et al.[38]	https://github.com/Ecogenomics/CheckM
dRep v3.4.0	Olm et al.[39]	https://github.com/MrOlm/drep
Mash v2.3	Ondov et al.[40]	https://github.com/marbl/Mash
FastANI v1.331	Jain et al.[41]	https://github.com/ParBLiSS/FastANI
CoverM toolkit v0.6.1	Aroney et al.[42]	https://github.com/wwood/CoverM
GTDB-tk v2.4.1	Chaumeil et al.[43]	https://github.com/Ecogenomics/GTDBTk
FastTree v2.1.10	Price et al.[44]	https://morgannprice.github.io/fasttree/
MMseqs v13.45111	Steinegger et al.[45]	https://github.com/soedinglab/MMseqs2
Blast+ v2.9.0	Camacho et al.[46]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Prokka v1.13.3	Seemann et al.[47]	https://github.com/tseemann/prokka
HMMER v3.1b2	Mistry et al.[48]	https://github.com/EddyRivasLab/hmmer
Prodigal v2.6.3	Hyatt et al.[49]	https://github.com/hyattpd/Prodigal
CRISPRCasFinder v2.0.2	Couvin et al.[50]	https://github.com/dcouvin/CRISPRCasFinder
eggnog-mapper v2.0.1	Cantalapiedra et al.[51]	http://eggnog-mapper.embl.de/
MGV viral detection pipeline	Nayfach et al.[1,52]	https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline
VirFinder v1.1	Ren et al.[53]	https://jessieren.github.io/VirFinder/
CheckV v1.0.1	Nayfach et al.[52]	https://pypi.org/project/checkv/
geNomad v1.11.2	Camargo et al.[54]	https://github.com/apcamargo/genomad/
VIBRANT v1.2.1	Kieft et al.[55]	https://github.com/AnantharamanLab/VIBRANT
CarveMe v1.5.1	Machado et al.[56]	https://github.com/cdanielmachado/carveme
SMETANA v1.0	Zelezniak et al.[24]	https://github.com/cdanielmachado/smetana
R v4.4.2	N/A	https://www.R-project.org/
Gephi v0.9.2	N/A	https://gephi.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Analysis code	This paper	https://doi.org/10.5281/zenodo.18637575
Other		
UHGG database	Nayfach et al.[1,52]	https://www.ebi.ac.uk/metagenomics/genome-catalogues/human-gut-v2-0.
MGV database	Almeida et al.[6]	https://portal.nersc.gov/MGV

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We recruited 84 healthy volunteers residing in Beijing through Beijing Shijitan Hospital. All participants provided written informed consent prior to enrollment. None of the participants had taken antibiotics within the six months preceding sample collection. Fecal samples were collected from all 84 individuals and subjected to metagenomic sequencing. The cohort consisted of 22 males and 62 females, aged between 18 and 40 years (mean ± s.d.: 35.84 ± 11.27 years). Detailed demographic and clinical characteristics, including sex, age, BMI, smoking status, alcohol consumption, prior antibiotic use, and comorbidities, are listed in Table S1.

This study was approved by the ethics committee of Beijing Shijitan Hospital (Project Number: 2023-KY-0007-002).

METHOD DETAILS

Culture-enriched metagenomics co-barcoding sequencing (cMECOS) pipeline

We implemented the cMECOS pipeline, which includes the following key steps: 1) Multiple lipid culture enrichment approaches to enrich microbial abundance from fecal samples, 2) Construction and sequencing of co-barcoding metagenomic libraries, 3) Co-barcoding assembly and genome binning, 4) Integration and redundancy reduction of genome bins across multiple samples. These four key steps were described in detail in the Methods that follows.

Feces collection and multiple lipid culture enrichment approaches

Participants collected fecal samples using sterile collection boxes. Within 2 h of collection, we prepared fecal suspensions for DNA extraction and culture enrichment testing. Briefly, we dissolved 400–500 mg of each fecal sample in 2 mL of phosphate-buffered saline (PBS) and mixed thoroughly. We then allocated 400 μL of the suspension for DNA extraction and 200 μL for two culture enrichment tests: 1) Anaerobic culture using Gifu Anaerobic Broth Medium (GAM), and 2) Aerobic culture using GAM.

For anaerobic cultures, we used GAM anaerobic bottles (Beijing GermountX Health Technology Co., LTD, KC001), which contain GAM medium optimized for the growth of most intestinal anaerobic microorganisms. For aerobic cultures, we used GAM aerobic bottles (Beijing GermountX Health Technology Co., LTD, KC002), which also utilize GAM medium but are designed for cultivating aerobic intestinal microorganisms. Both media were supplemented with essential growth factors, including amino acids, vitamins, and volatile fatty acids. All consumables were sterile, and we performed all procedures in a biosafety cabinet to maintain aseptic conditions. Following inoculation, we incubated the cultures at 37°C with continuous shaking at 180 rpm for 24 h. After incubation, we centrifuged the cultured media at 8,000 rpm for 5 min to pellet the microbiota, which we then used for DNA extraction.

High-molecular-weight DNA extraction

For DNA extraction, we treated fecal or microbiota samples with 200 μL of lysozyme (200 mg/mL) at 37°C for 1 h to lyse bacterial cell walls. We then extracted high-molecular-weight (HMW) DNA using a MagAttract HMW DNA Kit (QIAGEN, Cat. No. 67563), following the manufacturer's instructions. We quantified DNA concentration using a dsDNA Assay Kit (ExCell Bio, NGS00-3022), and assessed genomic integrity via agarose gel electrophoresis. Samples that yielded HMW DNA with sufficient concentration and integrity were used for library preparation. Using a MagAttract HMW DNA Kit, we consistently obtained DNA fragments exceeding 10 kb in length.

Metagenomics co-barcoding library preparation and sequencing

After extracting high-molecular-weight DNA, we constructed metagenomic co-barcoding libraries for MECOS using a customized protocol that included the following key steps[23]: (1) Transposome insertion, (2) Co-barcoding of long DNA fragments, (3) Adaptor ligation followed by polymerase chain reaction (PCR) amplification, and (4) Formation of DNA nanoballs (DNBs) and subsequent high-throughput sequencing.

We quantified the MECOS libraries using a dsDNA Assay Kit, ensuring DNA concentrations were ≥2.6 ng/μL. To assess library size distribution, we used the Qsep system. The resulting metagenomic co-barcoding libraries typically ranged from 200 to 2,000 base pairs in length. We then followed the standard protocol for MGI platforms, which includes circularization and DNB preparation. Sequencing was performed on the MGISEQ-2000 platform, using paired-end reads of 200 base pairs for all datasets. This process generated approximately 1.88 terabases (Tbps) of high-quality, clean sequencing data.

Reads processing, co-barcode assembling, and genome binning

We used SOAPfilter[32] (version 2.2) to remove low-quality reads and adapter sequences. To eliminate host contamination, we mapped the filtered reads to the human reference genome (hg38) using Bowtie2[33] (version 2.2.5) in "very sensitive" mode. For assembly, we first generated preliminary contigs using metaSPAdes[34] (version 3.13.1) with default parameters. We then mapped the raw reads back to the assembled contigs using BWA[35] (version 0.7.15, 'mem' mode) and sorted the resulting alignments into BAM files with Samtools[57] (version 1.9) to prepare input for Athena[36] (version 1.2, default settings). Athena was subsequently used to refine contig assemblies by leveraging co-barcoded sequences. We binned raw contigs from each sample into Metagenome-Assembled Genomes (MAGs)using MetaBAT2[37] (version 2.12.1). We assessed the quality of each MAG using, evaluating genome completeness and contamination.

Completeness, contamination, genome quality score calculation and HMQ genome definition

To ensure the reliability of downstream genomic analyses, a rigorous quality assessment and filtering step was implemented for all MAGs. The quality of each MAG was evaluated based on estimates of completeness and contamination, which were calculated using CheckM[38] (version 1.2.0). A composite quality score was then defined for each MAG using the formula: Quality Score = Completeness $-$ (5 $\times$ Contamination) + log10(N50). This scoring system prioritizes high completeness while heavily penalizing contamination. For subsequent analyses, we retained only those MAGs classified as high- or medium-quality (HMQ) MAGs. The HMQ threshold was defined by the following stringent criteria: a minimum of 50% completeness, a maximum of 5% contamination, and an overall quality score greater than 50. This multi-faceted filtering strategy ensures that the genomic dataset used in this study is of high integrity and suitable for robust comparative genomic and phylogenetic inferences. Finally, we constructed a CGG dataset from the HMQ MAGs.

Cluster analysis from genomes to strains and species

We first downloaded all representative genomes from the UHGG database—namely, the HMQ genomes that had the highest genome quality score for each species in the UHGG database, https://www.ebi.ac.uk/metagenomics/genome-catalogues/human-gut-v2-0.

And then all HMQ genomes in our study and UHGG representative genome were clustered to species-level clusters at 95% average Nucleotide Identity (ANI) using dRep (version 3.4.0).[39] Average Nucleotide Identity (ANI) were calculated by fastANI v1.331[41] with default parameters, which estimate ANI through the identification of bidirectional best mapping fragments between genome pairs. Next, the HMQ genomes from each same species-level cluster were clustered to strains-level clusters at 99% ANI using Mash (version 2.3)[40] with sketch size of 10k and k-mer size of 21 bp, afterward the agglomerative clustering from python package sklearn with single linkage option was used. Within each species-level cluster and strains-level cluster, the genome with highest genome quality score was selected as the representative genome. For species-level clusters and strains-level cluster containing UHGG genomes, we annotated all member genomes based on the corresponding UHGG taxonomy. The relative abundance of bacterial species was calculated using the CoverM toolkit (version 0.6.1)[42] with the species-level representative genomes from the CGG database. Species constituting less than 0.1% of the community were classified as low-abundance.

Previously not reported species-level clusters and strains-level cluster definition

For species-level cluster and strains-level clusters that initially did not contain representative genomes from the UHGG database, they were defined as candidate previous not reported species-level clusters or strains-level cluster. To robustly validate the novelty of the candidate previous not reported clusters, we performed a rigorous comparative genomic analysis. Genomes that did not cluster with any others were processed independently for taxonomic assignment using GTDB-tk[43] (version 2.4.1) with Genome Taxonomy Database (r206). Phylogenetic analysis for the entire dataset was performed based on the concatenation of 120 ubiquitous single-copy marker genes using FastTree[44] (version 2.1.10). The maximum-likelihood placement of each genome in the GTDB-Tk reference tree was found by a "classify" step using pplacer.

We combined these candidate species-level clusters with comprehensive public genomic resources, including the UHGG and the standardized genomes from the GTDB encompassing both RefSeq and GenBank. For this combined set, we reconstructed a detailed phylogenetic tree from the marker gene alignment and calculated pairwise ANI values. The concordant results from these phylogenomic and ANI analyses served to conclusively verify that the clusters as final previous not reported species-level clusters.

rRNA gene analysis in microbial genomes

We predicted rRNA genes using barnnap (version 0.9) on all HMQ genomes in our study and 289,232 genomes from the UHGG database. To enhance the rRNA gene database, we integrated three databases—Silva2, Greengene2, and rrnDB—using MMseq2[45] (version 13.45111). The integration applied a minimum coverage threshold of 80% of the shortest sequence length with 100% sequence identity. Finally, we mapped the full-length 16S rRNA gene sequences onto the integrated database using Blast+[46](version 2.9.0).

Gene prediction, clustering, and annotation

We predicted and annotated protein-coding sequences (CDS) for 5,006 genomes using Prokka[47] (version 1.13.3), with Prodigal[49] (version 2.6.3) in single-mode ('-p single') for genome assemblies containing a single species. We applied the following options: '-c' to predict proteins with closed ends only, '-m' to prevent genes from being predicted across stretches marked as Ns, and

'-p single' for single-species assemblies. From the 10,679,102 CDS predictions, we generated the Comprehensive Gut Protein (CGP) catalog. Protein clustering for both the CGP and UHGP was performed using the 'linclust' function in MMseq2[45] (version 13.45111), applying a minimum coverage threshold of 80% of the shortest sequence length. We set the '–min-seq-id' option to 1, 0.95, and 0.9 to generate catalogs at 100%, 95%, and 90% protein identity, respectively. The resulting clustered genes at 90% protein identity (HC-CGP-90) were annotated against the KEGG,[58] Pfam-A,[59] and eggNOG[60] databases. KEGG and Pfam-A annotations were performed using the hmmsearch utility in the HMMER package v3.1b2.[48] Genes were annotated with KEGG based on the top-scoring alignment with a bit-score $\geq$50, while for Pfam-A, the trusted cutoff was applied. Gene annotations based on the eggNOG database were performed using eggnog-mapper[51] (version 2.0.1), considering only hits with an e-value <0.00001.

Bacteria-phage interactions establishment

We predicted CRISPR spacer sequences from all HMQ MAGs in our study using CRISPRCasFinder[50] (version 2.0.2). To remove the potential false positive, only spacers found in CRISPR arrays with Evidence Levels 3 and 4 were retained. We then searched these spacers against the MGV database[1] (https://portal.nersc.gov/MGV) using Blast^{+46} (version 2.9.0). Bacteria-phage interactions were constructed based on hits with a maximum of zero mismatch or gap over at least 100% of the spacer length. Finally, the network of bacteria-phage network was constructed by igraph package (version 2.1.4) in R, and visualized using Gephi (version 0.9.2) with 'Fruchterman-Reingold' layout.

Prophage region identification

We detected prophage regions following the MGV viral detection pipeline,[1] which consists of two main steps aimed at comprehensively identifying integrated prophages. First, to recover high-quality viral genomes from metagenomic assemblies, viral contigs were identified using four complementary signatures: (1) the presence of viral protein families, (2) the absence of microbial protein families, (3) viral-like nucleotide composition, and (4) a low strand-switch rate indicating co-oriented genes. Viral and microbial protein families were classified using hmmsearch (HMMER v.3.1b2[48]; E-value $<1 \times 10^{-10}$) against HMM profiles from IMG/VR and Pfam-A, assigning each protein based on the database of its top hit. Viral nucleotide composition was assessed using VirFinder v.1.1,[53] and the strand-switch rate was calculated as the number of strand switches divided by the total gene count per contig. Following identification, viral genome completeness and quality were assessed using CheckV (version 1.0.1).[52] Taxonomic classification was performed using the genome-based tool geNomad (version 1.11.2)[54] against the geNomad database (version 1.9). Second, to recover prophage regions that may have been missed during initial viral contig identification-particularly those integrated into host genomes—we screened contigs derived from MAGs. To maximize detection sensitivity, we combined the newly identified viral genomes and MGV genomes. To minimize false positives from mis-binned sequences, contigs with less than 50% host-derived sequence were excluded. The remaining host-associated contigs were aligned to viral genomes using blastn (Blast+ version .2.9.0),[46] and prophage regions were defined as matches $\geq$10 kb in length with $\geq$96% nucleotide identity. Regions within 95% identity were annotated to the same phage species.

Determine the temperate and virulent phages for CRISPR spacer associated bacteriophages

To predict the lifestyle of the CRISPR spacer associated bacteriophages, we used the annotation of temperate score which calculated by BACPHLIP for each bacteriophages, to divided the bacteriophages into the temperate or virulent phage.[61] Bacteriophages with a temperate score greater than 0.5 were categorized as temperate phage, while those with a score lower than 0.5 were identified as virulent phage.

Bacteria-bacteria interaction network for two type phages

Bacteria-bacteria interactions were established when two bacterial hosts shared by the same phage. According to shared phage characters, each Bacteria-bacteria interaction was classified as either 'TV' (temperate phage) or 'VV' (virulent phage). The strength of the interaction was defined with number of sharing phages. Bacteria-bacteria network was constructed by igraph (version 2.1.4) in R, and visualized by Gephi (version 0.9.2). For the comparison of the networks associated to TV and VV, we fixed the position of nodes and connected the node with shared TV and VV, and the edges represent the number of shared phages.

Metabolic relationship profiling of bacteria-bacteria interaction by their metabolic scores

To investigate the metabolic interactions underlying bacteria-bacteria relationships shaped by phage-mediated co-infection, we reconstructed genome-scale metabolic models (GEMs) for representative microbial genomes from each subcommunity using CarveMe[56] (version 1.5.1) with default parameters. We then computed the pairwise MIP, MRO, and SMETANA scores for all host pairs linked by shared phages using SMETANA[24] (version 1.0). These metrics quantified the cooperative or competitive tendencies between the microbial hosts.

We further examined subcommunities dominated by Lachnospiraceae or Bacteroidaceae linkagesto assess family-specific metabolic interaction patterns. To ensure robustness, we randomly sampled nested communities of increasing complexity (ranging from 2 to 10 species) from each subcommunity, performing 100 iterations for each size. We recalculated the metabolic interaction scores for these subcommunities and tested their consistency against the full subcommunity profiles using Wilcoxon rank-sum tests

(FDR-corrected $p < 0.05$). All community simulations and the calculation of minimal media composition were carried out using SMETANA (version 1.0). Based on these analyses, we identified cross-feeding and competition pairs (CPCCs).

Ecological validation of cross-feeding and competition pairs (CPCCs)

To ecologically validate the inferred metabolic interactions, we identified coupling pairs of cross-feeding and competition (CPCCs) from the 456 species comprising the phage-linked subcommunities. The existence of one of the following two pieces of evidence can be defined a sCPCCs: (1) evidence of potential cross-feeding (i.e., positive scores in MIP and/or SMETANA), and (2) evidence of potential competition (i.e., positive scores in MRO). Species pairs satisfying both conditions were considered CPCCs. We then assessed the co-occurrence patterns of these CPCCs in two human cohorts. For each CPCC, we calculated two metrics: (1) the co-occurrence proportion, defined as the fraction of samples where both taxa were present at a relative abundance $>1 \times 10^{-5}$, and (2) the weighted abundance, defined as the mean relative abundance of the co-occurring taxa across all CPCCs within a subcommunity.

To validate the potential role of the identified CPCCs in human disease contexts, we applied our analysis to two independent cohorts representing distinct disease states: an inflammatory bowel disease (IBD) cohort (NCBI SRA: PRJNA993675) and an obesity cohort (EBI: PRJEB14215) (Dataset S10). To ensure robust statistical comparisons between disease and healthy control groups, we employed a bootstrap resampling procedure. Specifically, we performed 100 iterations where we randomly subsampled 50 individuals from each group (disease and healthy control) within a cohort. For each iteration, we computed the median co-occurrence proportion and weighted abundance for both groups. Significant differences between cases and controls were determined using Wilcoxon rank-sum tests, with a significance threshold of FDR-corrected $p < 0.05$ across all iterations.

AMG gene detection

We utilized the VIBRANT (version 1.2.1)[55] tool to identify auxiliary metabolic genes (AMGs) from viral contigs. Candidate AMGs were initially screened based on KEGG annotations, with a focus on genes involved in "metabolic pathways" and the "sulfur relay system". Manual curation was subsequently performed to exclude non-AMG annotations, such as nrdAB and thyAX, as well as genes associated with direct nucleotide conversions. All identified AMGs were mapped to respective KEGG metabolic pathways.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed with various packages within the statistical program R v4.4.2. For the statistical analyses, unless otherwise indicated, statistical comparisons between groups were performed using the Wilcoxon rank-sum test for continuous data and Fisher's exact test for count data. Other details on statistical analysis can be found in figure legends.