Genomics Proteomics Bioinformatics 20 (2022) 334–349

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Panoramic Insights into Microevolution and Macroevolution of A *Prevotella copri*-containing Lineage in Primate Guts

Hao Li [1], Jan P. Meier-Kolthoff [2], Canxin Hu [1], Zhongjie Wang [1], Jun Zhu [1], Wei Zheng [1], Yun Tian [1,3], Feng Guo [1,3,*]

[1] *School of Life Sciences, Xiamen University, Xiamen 361102, China*
[2] *Department of Bioinformatics, Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, D-38124 Braunschweig, Germany*
[3] *Fujian Provincial Universities Key Laboratory of Microbial Resource, Xiamen University, Xiamen 361102, China*

Received 28 June 2021; revised 23 October 2021; accepted 1 November 2021
Available online 3 February 2022

Handled by Fangqing Zhao

KEYWORDS

Prevotella copri;
Co-speciation;
Gut microbiome;
Host;
Biogeography

Abstract ***Prevotella copri*** and its related taxa are widely detected in mammalian **gut microbiomes** and have been linked with an enterotype in humans. However, their microevolution and macroevolution among **hosts** are poorly characterized. In this study, extensively collected marker genes and genomes were analyzed to trace their evolutionary history, host specificity, and biogeographic distribution. Investigations based on marker genes and genomes suggest that a *P. copri*-containing lineage (PCL) harbors diverse species in higher primates. Firstly, *P. copri* in the human gut consisted of multiple groups exhibiting high genomic divergence and conspicuous but non-strict biogeographic patterns. Most African strains with high genomic divergence from other strains were phylogenetically located at the root of the species, indicating the co-evolutionary history of *P. copri* and *Homo sapiens*. Secondly, although long-term co-evolution between PCL and higher primates was revealed, sporadic signals of **co-speciation** and extensive host jumping of PCL members were suggested among higher primates. Metagenomic and phylogenetic analyses indicated that *P. copri* and other PCL species found in domesticated mammals had been recently transmitted from humans. Thirdly, strong evidence was found on the extensively horizontal transfer of genes (*e.g.*, genes encoding carbohydrate-active enzymes) among sympatric *P. copri* groups and PCL species in the same primate host. Our study provides panoramic insights into the combined effects of vertical and horizontal transmission, as well as potential niche

* Corresponding author.
E-mail: fguo.bio@xmu.edu.cn (Guo F).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China

https://doi.org/10.1016/j.gpb.2021.10.006

adaptation, on the microevolutionary and macroevolutionary history for an enterotype-representative lineage.

Introduction

Animal guts harbor complex microbial assemblies that play key roles in host development, metabolism, and immunity [1]. Phylosymbiosis between host and gut microbiome has been widely investigated at the community level, and many microbial assemblies show a congruence with their host phylogeny [2,3]. However, such a congruence may not a necessarily result from the long-term co-evolution (*i.e.*, continuous co-adaptation) between hosts and symbionts; other factors, such as diet, host physiology, and immunology, may play uncharacterized roles in shaping gut microbiomes along with host phylogeny [3]. Alternatively, focusing on certain microbial lineages is a reliable and direct way to trace the history of vertical transfer [4–6]. The evolution of a bacterial lineage can be viewed at two levels, namely, macroevolution (among multiple bacterial species) and microevolution (within one bacterial species). The former is often discovered among remotely related host species [4,5,7], while the latter is observed on hosts belonging to either the same or different host species [6,8]. In some cases, co-evolving microorganisms vertically transferred along a host lineage can exhibit the co-speciation feature, possibly driven by host allopatric speciation, limited inter-host dispersal, and genomic recombination [3,5].

In addition to vertical transfer, the biogeographical and host-specific distribution of certain gut bacterial lineages could be largely influenced by the potential horizontal transfer among heterospecific hosts and ecological selection [9,10]. Allochthonous taxa may switch to new hosts and initiate new evolutionary branches, potentially causing promiscuous generalists in many host species [11,12]. Novel host–microbe interactions and adaptations can be introduced in such a scenario [13]. So far, only a few gut bacterial lineages have been comprehensively studied at the microevolutionary and macroevolutionary levels to understand the potential effects of vertical and horizontal transfers on their biogeography and host specificity. This limitation is possibly caused by the lack of comprehensive information for a certain microbial lineage from a wide range of host species and geographic regions, as well as the frequent extinction of hosts and microbes. In addition, samples must be collected from wild animals barely affected by humans as a prerequisite to minimize anthropogenic interferences [10,14]. However, data from wild animals are usually less available than those from domesticated animals.

Prevotella is a representative genus of human enterotypes [15]. Multiple species in this genus have been detected in human feces [16]. *Prevotella copri* and a few closely related species have the highest frequency and abundance [16,17], are thought to be positively selected by non-westernized diet with high plant-sourced polysaccharides [18,19], and link with a few human diseases as potential disadvantageous factors [20,21]. Two studies based on single nucleotide polymorphisms (SNPs) have preliminarily reported the intra-species diversification and phylogeography [22,23]. A more recent work collected over 1000 *P. copri*-related genomes, most using a reference-based metagenomic binning strategy, and reported four species-level clades occurring in human guts, thereby validating the species-level diversification in this lineage [18]. However, dynamic genetic repertoire within and among bacterial species, as well as the high frequency of co-occurred conspecific strains in one metagenomic sample, resulted in the necessity of a *de novo* binning strategy [22,24]. Moreover, phylogeny conducted solely based on genome bins may lose the overall diversity because genomes of minor organisms or those with microdiversity can hardly be retrieved from metagenomes [25]. Thus, to obtain more extensive information, phylogenetic marker genes are still useful.

In this study, we focused on two major questions about the co-evolution of *P. copri* and its related taxa with hosts. On one hand, as a highly concerned bacterial species in human gut, although the phylogeographical pattern of *P. copri* has been reported, its microevolutionary history with human being has not been fully addressed. However, previous studies referred to less informative data, such as SNPs and genomes retrieved by reference-based binning strategy, and data with poor representativity, especially for lacking genomes from African (although Tett et al. [18] contained metagenome-derived genomes from African, they did not include these genomes in phylogeographical analysis). On the other hand, *P. copri* and its related taxa have been frequently discovered in the gut microbiomes of nonhuman primates and other mammals [16,26,27]. How they transferred and potentially evolved among host species at the macroevolutionary level remains unknown. To address the two questions, we reconstructed the robust phylogeny of *P. copri* and its related taxa by comprehensively collecting their phylogenetic markers from multiple hosts and various geographic regions. The 16S rRNA gene, a gene encoding DNA gyrase subunit B (*gyrB*), a selected marker gene with the intra-specific resolution, genomes from 7 isolates, and 135 *de novo* binned genomes from metagenomes were used as references to uncover the inter-species and intra-species phylogenies, host distribution, and co-evolutionary history. The results showed the existence of a *P. copri*-containing lineage (PCL) containing much more species than previous reports. Regarding the phylogenies at microevolutionary and macroevolutionary levels, the history of vertical and horizontal transmission and genomic evolution of *P. copri* and its related species can be deduced.

Results

Phylogeny of the 16S rRNA gene suggests the existence of a PCL in multiple hosts

Among a total collection of 534 16S rRNA gene sequences from *Prevotella*, 43 representative 16S rRNA gene sequences, including six isolates (five from this study), the type strain of *P. copri* DSM18205, and 36 downloaded clone sequences, formed a monophyletic branch in *Prevotella* with moderately supportive bootstrapping values (**Figure 1**A). All 43 sequences were obtained from the fecal samples of four hosts,

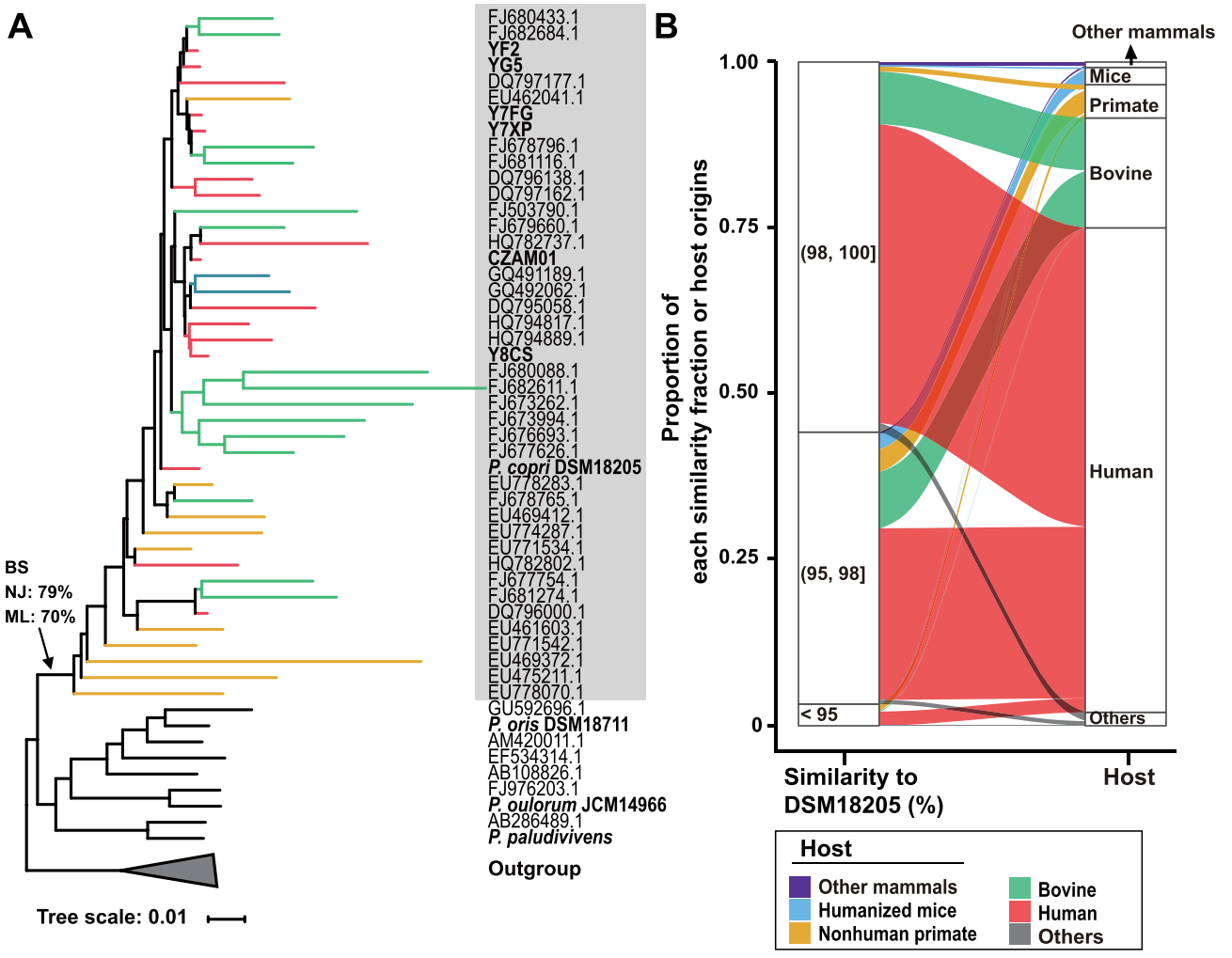

Figure 1 16S rRNA gene-based analysis on the distribution and phylogeny of the PCL

A. Neighbor-joining tree based on 534 16S rRNA gene sequences of the genus *Prevotella* with 100 iterations of bootstrapping. The 43 sequences affiliated with the PCL are shaded in gray. Bold labels indicate sequences from isolates. **B.** Sankey diagram showing the distribution of host origins and similarity fraction (to DSM18205) for 16S rRNA gene sequences from the SILVA database identified as the PCL (n = 5316). BS, bootstrap; NJ, neighbor-joining; ML, maximum-likelihood; PCL, *Prevotella copri*-containing lineage.

i.e., human, nonhuman primates, bovines, and humanized mice. Several clone sequences from the fecal samples of nonhuman primates were located at the root of the clade and exhibited a deeply branching feature.

A total of 5316 16S rRNA gene sequences putative affiliated with the PCL were retrieved from the SILVA database to comprehensively profile the source of this lineage. As shown in Figure 1B, most of the sequences were obtained from the fecal samples of the four hosts, and the rest (< 3%) were obtained from the fecal samples of other mammals (mostly domesticated ones such as pig, dog, and mammals from zoos) or from human-related environments (*e.g.*, human skin and wastewater). Nonhuman primate-derived sequences contributed over 25% of the remotely related fraction (< 95% similarity to the 16S rRNA gene of DSM18205; Figure 1B) but only accounted for 5% of the total sequences. This analysis provides preliminary evidence of a multi-specific PCL, and the results point to its potential macroevolutionary history with primates and the occurrence of PCL members in the guts of domesticated mammals.

The phylogeographical pattern of *P. copri* suggests its co-evolutionary history with *Homo sapiens*

A total of 123 *P. copri* genomes (all with > 70% completeness and < 10% contamination; 5 isolates and 118 metagenomic bins; Table S1) were used in the phylogenomic reconstruction to investigate the intra-species phylogeny of the most dominant human gut PCL member, *P. copri*. Therein, 116 genomes reached the high-quality criterion (> 90% completeness and < 5% contamination), according to Bowers and colleagues [28]. NMPZ01, Y7XP, and Y7FG are not *P. copri* according to the average nucleotide identity (ANI) and digital DNA–DNA hybridization (dDDH) values (Table S2) and thus were set as the outgroup. The phylogenetic tree reconstructed by concatenating the sequences of determined 1095 core single-copy genes in *P. copri* (**Figure 2**A, left) and that reconstructed based on dDDH values (Figure 2, right) were highly consistent in terms of topology with a few exceptions (T2D-8A and Y8CS). In the core gene-based phylogenetic tree (Figure 2A), the genomes were divided into nine groups (non-monophyletic

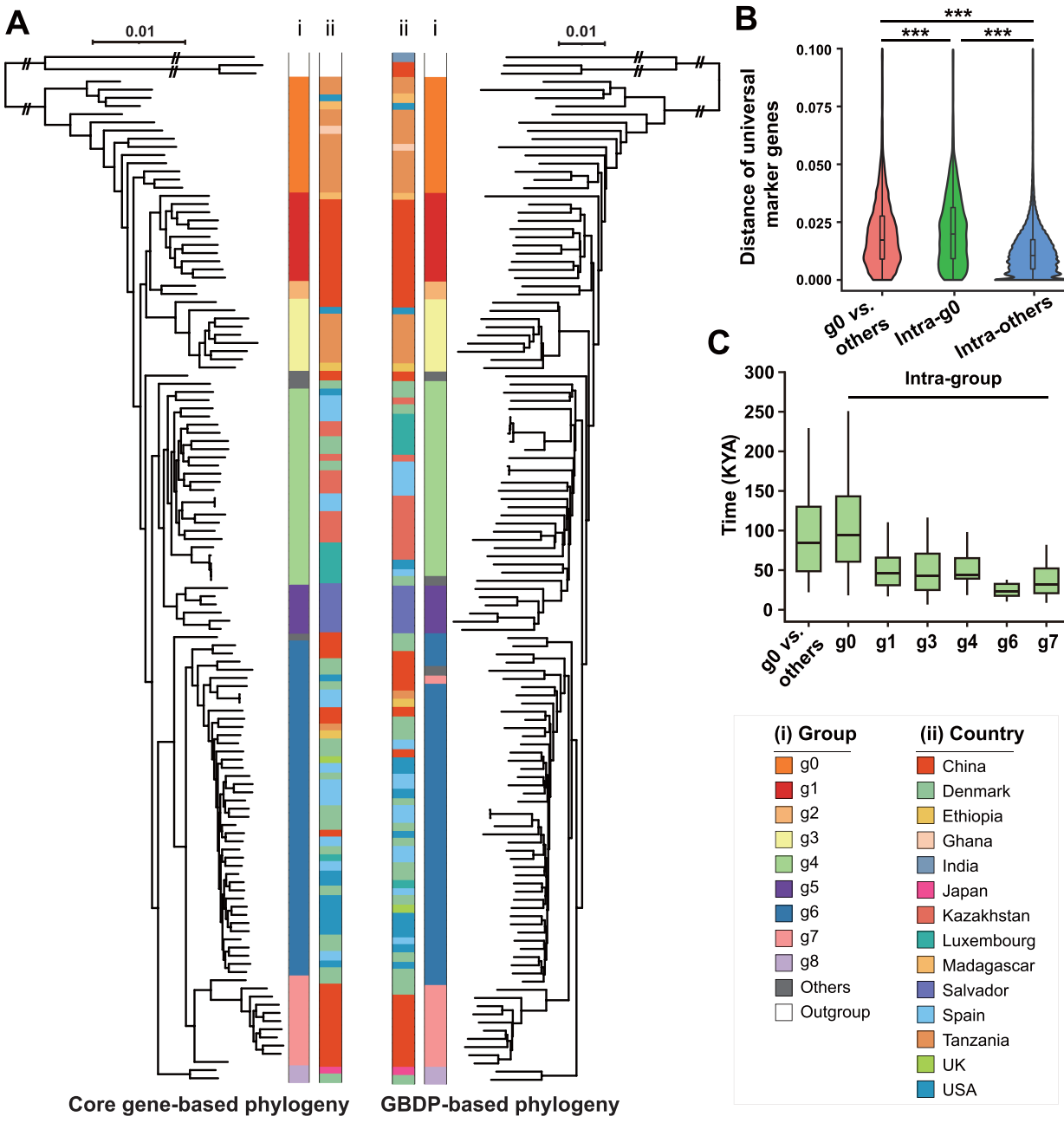

Figure 2 Phylogenomic analyses and molecular dating of 123 globally collected *P. copri* genomes

A. Maximum-likelihood tree of 1095 concatenated core single-copy orthologous genes (left) and GBDP-based phylogenomic analysis of the nucleotide sequences restricted to coding regions (right). Shared annotations include (i) clustered groups and (ii) geographical origin. **B.** Distance of 120 universal marker genes between g0 strains and stirans in other groups (g0 *vs.* others), among g0 strains (intra-g0), or among strains in other groups (intra-others). ***, $P < 0.001$ (Wilcoxon test). **C.** Split time of inter-group (g0 *vs.* others) and intra-group strains. KYA, kilo years ago; GBDP, Genome BLAST Distance Phylogeny.

g0 and monophyletic g1–g8) and three single branches. The seven major groups were g0 (n = 14), g1 (n = 11), g3 (n = 9), g4 (n = 24), g5 (n = 6), g6 (n = 41), and g7 (n = 11) and exhibited a well-supported phylogeographical pattern to a great extent. Group g0 was mostly contributed by Africans (only one from USA), g1 and g7 almost exclusively occurred in China (two g1 strains from Africa), and g5 was only covered by the strains found in Salvador. Groups g4 and g6 consisted of strains from multiple continents, mostly from European countries, USA, and Kazakhstan, and a few from China and African countries. Type strain DSM18205 formed g8 together with one strain from Denmark. Under the dDDH-based species clustering [29], these 123 *P. copri* genomes were assigned into nine maximally supported, monophyletic species-level clusters or single branches (not shown in Figure 2A). Therein, g0 comprised eight clusters, and all the other groups

constituted a single cluster. According to both approaches, no close relatives were observed for any of the 123 genomes, except for those obtained from the sequential samples of one person (O2.UC17-0 and O2.UC17-1 in g4 and O2.UC38-0 and O2.UC38-2 in g6) and four from a family (M04.1-V3, M04.3-V1, M04.4-V3, and M04.5-V3 in g4) (Figure S1).

The co-evolutionary history between the species and *Homo sapiens* was supported by the finding that most strains from Africa were located at the root of the tree and remotely separated from other groups. The large phylogenetic distance between g0 strains and strains in other groups or among g0 strains was further confirmed for single housekeeping genes (Figure 2B) [30]. The median synonymous mutation rate of 34 housekeeping genes [without significant intragenic recombination in pairwise-homoplasy index (PHI) test] from African-derived strains and other strains was 0.044 (Table S3) [31]. On the basis of the mutation rate of 2.6×10^{-7} per site per year for housekeeping genes of another human symbiont, *Helicobacter pylori* [32], the split of g0 strains from other strains was dated at approximately 84 kilo years ago (KYA) (Figure 2C). This period roughly coincided with the time for modern humans outside Africa and was also supported by *H. pylori* results [33,34]. The median split time for strains in each major group g0, g1, g3, g4, g6, and g7 was 95 KYA, 47 KYA, 43 KYA, 45 KYA, 23 KYA, and 32 KYA, respectively (Figure 2C).

Divergence and potential sympatric inter-group gene transfer for carbohydrate-active enzymes among *P. copri* groups

As mentioned, *P. copri* was thought to be enriched by diet with high plant-sourced polysaccharides [18,19]. To test whether the carbohydrate-active enzymes (CAZys) have a group-specific pattern, the CAZy modules in the major groups were examined. Pan genomes of the six major groups contained 143 CAZys modules, nearly half of which were generally distributed in all genomes without significant difference between any two groups [Fisher's exact test, false discovery rate (FDR)-corrected $P > 0.05$]. However, 32 group-specific and 42 sporadic CAZy modules were determined (**Figure 3**A). High group specificity was found in several putative alginase-encoding genes (genes containing PL6, PL6_1, and PL17) in g1 and a putative hyaluronidase-encoding gene (a gene containing GH84) in g6. A gene almost exclusively detected in g1 was identified as a putative alginase-encoding gene by Pfam annotation (ID: PF05426, not annotated against the CAZy database) and functionally verified via the heterologous

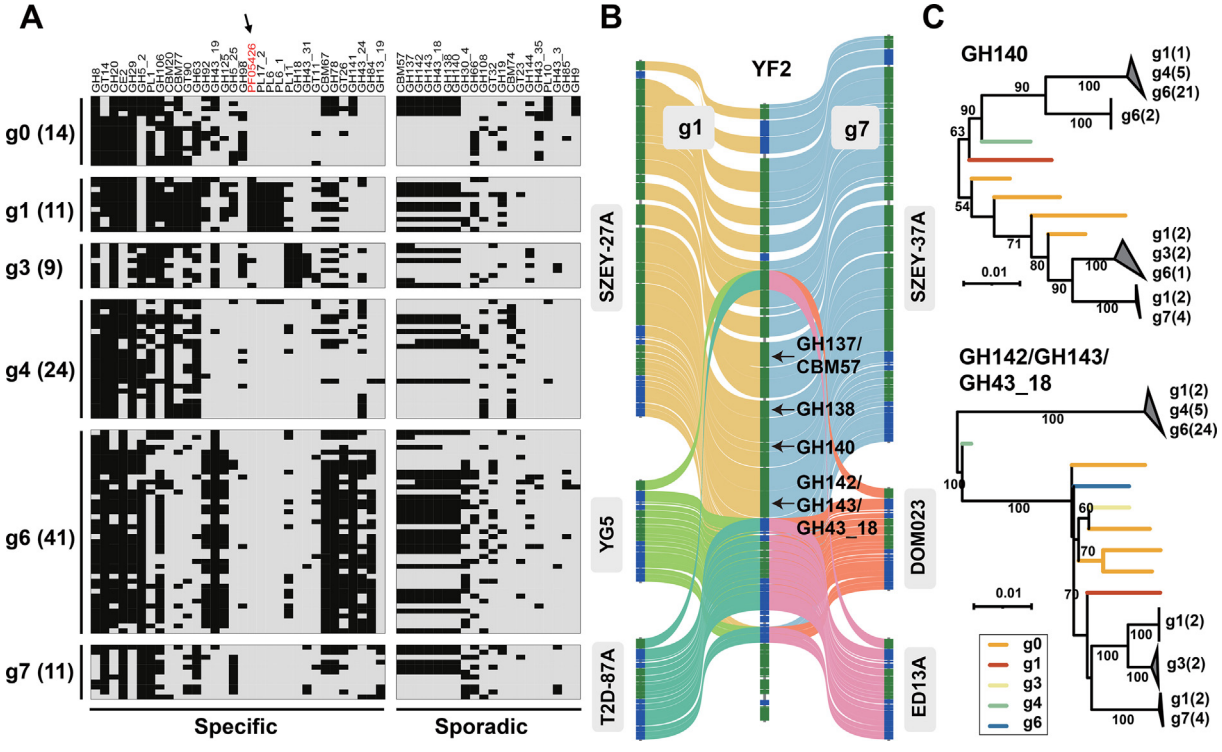

Figure 3 CAZys of the six major *P. copri* groups and their biogeography-related microevolution

A. Group-specific (FDR-corrected $P < 0.05$ for Fisher's exact test on the frequency of any groups) or sporadic (detected in less than half of the genomes and no significant difference between any two groups) CAZy modules in six major groups. The modules only present in at least 3 genomes are shown in the heatmap (black: present; gray: absent). The gene indicated by the arrow was identified by Pfam annotation using HMMER and functionally verified (Figure S2). **B.** Genomic synteny of the fragment containing four sporadic CAZy-encoding genes with flanking genes is displayed for six genomes belonging to g1 ($n = 3$) and g7 ($n = 3$) using the complete genome of YF2 for mapping. Genes in sense strand (green) or reverse strand (blue) are shown by block colors. **C.** Maximum-likelihood trees of two sporadic CAZy-encoding genes (in nucleotide) with 100 bootstrap iterations. Sequences exhibiting extremely high similarity (> 99%) are collapsed, and their composition is shown. CAZy, carbohydrate-active enzyme; FDR, false discovery rate.

expression in *Escherichia coli* (cloned from YF2) and biochemical assays (Figure S2; File S1).

A sporadic CAZy (encoded by a gene containing the three modules of GH142, GH143, and GH43_18) was related with a novel depolymerase from *Bacteroides thetaiotaomicron* targeting on complex glycans (> 60% amino acid similarity with protein BT1020 [35]). Genomic synteny showed that the absence or presence of the gene-related cluster (containing four CAZy-encoding genes) was not due to incorrect assembly or binning (Figure 3B). The fragment (approximately 40 kb in length, containing 15 genes in the complete genome of YF2) was colocalized on the same genomic region in all positive strains but was clearly deleted in negative strains. Phylogenies of the two sporadic CAZy-encoding genes located in the cluster revealed inter-group horizontal gene transfer (HGT) for sympatric groups (Figure 3C, Figure S2); the other two CAZy-encoding genes had similar signals (data not shown). For example, a gene containing GH140 had three phylogenetic clusters (consisting of sequences with > 99.5% similarity), each of which contained strains from geographically co-occurring groups (*e.g.*, g1/g7 in China and g4/g6 mostly in European countries and USA). Further investigation on the *P. copri* genomes of g1 ($n = 8$), g6 ($n = 1$), and g7 ($n = 3$) isolates provided by a recent study on the Chinese population [36] excluded the possibility of genomic contamination for metagenomic bins. The findings further supported that the aforementioned phylogenetic crosslinks were not derived from genomic contamination (Figure S3).

Non-strict phylogeographic pattern for *P. copri* groups

Although phylogenomics revealed the biogeographic distribution of groups, a few exceptions could be found in Figure 2A (*e.g.*, g6 strains from China). The results based on genome bins may not sufficiently represent the population-level composition in each fecal sample, because some strains could be missed during genome binning due to their low abundance or microdiversity [25]. Two analyses based on a selected intra-species marker gene, *orth10* (see File S1 and Figure S4 for the reason to use this gene), were conducted to further investigate the strictness of the phylogeography.

We first quantitatively assigned the metagenomic reads of *orth10* into groups. The analysis was conducted for 47, 70, 139, and 14 metagenomic datasets selected from Africa, China, Europe, and USA, respectively, in accordance with *P. copri* abundance determined by its *gyrB* abundance ($> 1 \times 10^{-6}$) in integrated gene catalog (IGC) of 1267 human gut metagenomes and 67 Hadza hunter-gatherers of Africa [37–39]. All the four datasets showed a non-strict group-level distribution pattern, while the dominant groups were consistent with the aforementioned biogeographic pattern (**Figure 4**A). Noticeably, the presence of g0, g1, and g7 in Europeans and the detection of g0 in Chinese were revealed by this approach, which were completely missing based on genomic information.

Next, we conducted high-throughput sequencing for the *orth10* amplicons of sewage samples collected from five cities in China. As shown in Figure 4B, each sample contained 27–142 unique *orth10* phylotypes. Although sequences affiliated with g1 and g7 usually exhibited dominance, other groups were detected in the sewage samples, especially g0 and g6 (the former can be supported by the aforementioned metagenomic survey). Although foreigners live in the cities (< 0.5% in total population), their contribution to the sewage was improbably high enough to change the main profile. Moreover, this result also suggested that the collected genome bins were a good representation of group-level diversity, because all detected phylotypes are highly similar with the references (all with > 95% identity and mostly with 100% identity in Figure 4B, all intra-group *orth10* with > 95% similarity except group g0 in Figure S4). Based on the aforementioned results, we conclude that the group-level distribution in *P. copri* is not geographically strict, at least for the major groups.

Evidence of long-term co-evolution and sporadic co-speciation of PCL with higher primates

Phylogeny based on the 16S rRNA gene suggested that multiple PCL species were associated with primates (Figure 1). Thus, the PCL abundance in fecal metagenomes ($n = 168$) from 20 species of wild primates (Table S4) was analyzed based on the abundance of *gyrB* affiliated with the PCL in these metagenomes (**Figure 5**A). Metagenomic assembly initially generated 82 PCL *gyrB* sequences, which represented 39 species-level clusters under 98% similarity cut-off (Figure S5). All these sequences were retrieved from eight species of higher wild primates (all from Cercopithecidae and Hominidae). The 39 representatives and 9 de-replicated (under 98% similarity cut-off) human gut *gyrB* sequences (extracted from the IGC database and genomes of isolates) were verified as PCL members, because they formed a well-supported clade within *Prevotella* (Figure 5B, Figure S6). Read-based quantification confirmed the absence of PCL in the guts of all lower wild primates and *Colobus guereza* (Figure 5A). The high diversity of previously unrecognized PCL species in higher primates strongly supported a long-term co-evolutionary history between the lineage and the hosts.

Most primate host species were inhabited by multiple PCL species (Figure 5B). The PCL profile exhibited a conspicuous host-specific pattern, and no strong signals of phylosymbiosis were observed (clustering at the bottom in Figure 5B). Four *gyrB* representatives contained assembled sequences from multiple hosts, indicating their non-strict host specificity (Figure 5B). Intriguingly, the phylogeny of *gyrB* hinted co-speciation events among four host species (*Papio anubis*, *Papio cynocephalus*, *Papio kindae*, and *Cercopithecus Ascanius*) with the furthest split time of 16.2–22.4 million years ago (MYA), but little signal could be detected across all higher primate hosts (Figure 5B). The split time of the four corresponding *gyrB* clusters (determined as the split time between hosts) was used as a reference to calculate the molecular clock rate and perform dating for the whole phylogenetic tree (Figure 5B). The initial separation of PCL from other *Prevotella* species was deduced to occur duing 8.7–43.8 MYA (Figure S6), which was highly variable but covered the split time of higher primates from ancestor (28.0–31.4 MYA). Consistency was compared between the split time of host pairs and bacterial pairs (Figure 5C). Although the molecular clock rate for bacteria was highly variable, the split time of host pairs and that of bacterial pairs were still inconsistent in most pairs, except for some closely related host species such as between *Papio* spp. and between *Homo sapiens* and *Pan troglodytes*.

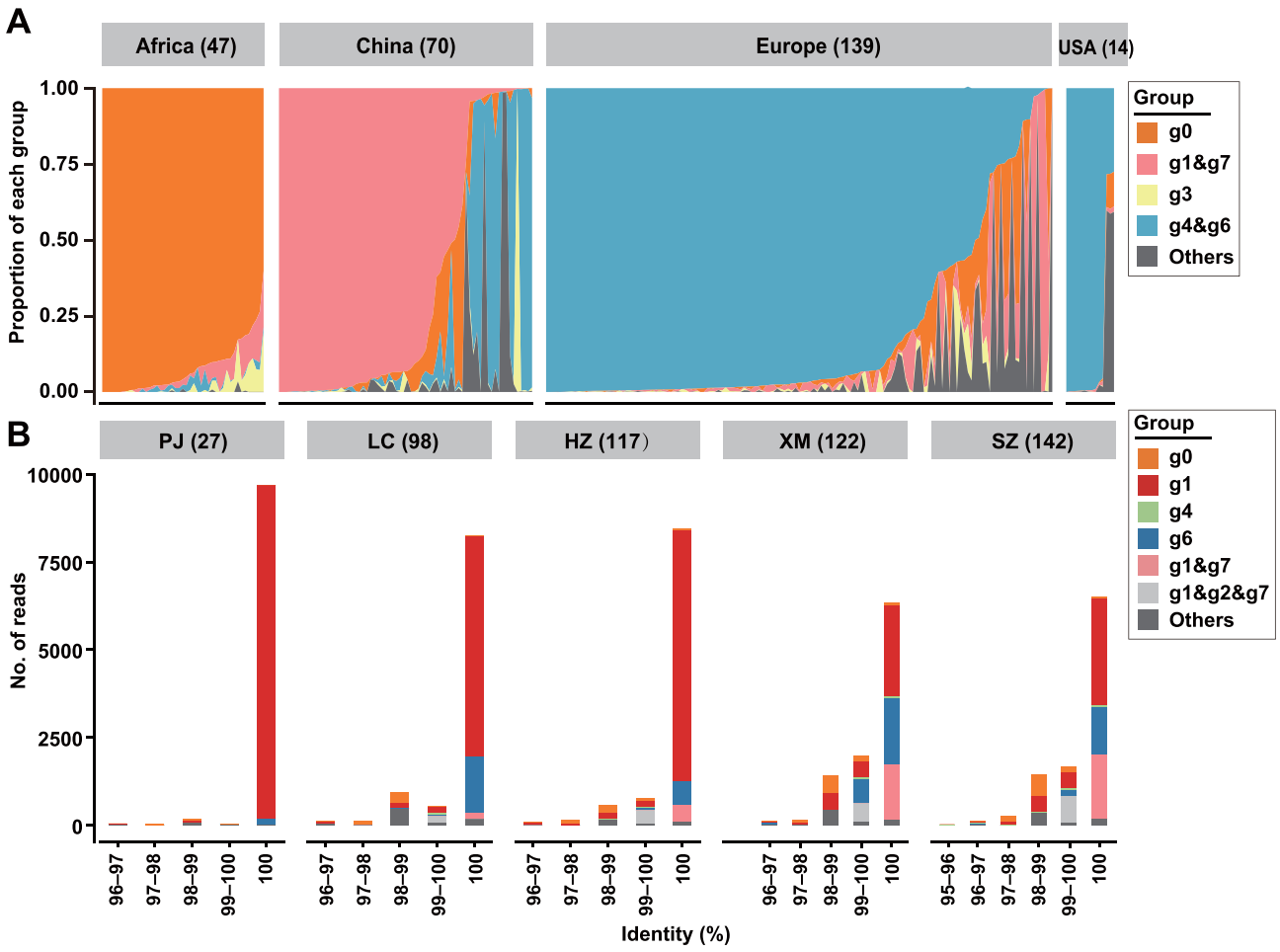

Figure 4 Non-strict biogeographical distribution for *P. copri* groups

A. Group-level profile of *P. copri* based on the full-length *orth10* gene in human gut metagenomes. The number of samples with *P. copri* abundance $> 1 \times 10^{-6}$ are shown in the brackets. Groups (*i.e.*, g1&g7 and g4&g6) from the same geographical origin were merged to display. **B.** *P. copri* group-level profile in sewages of five cities in China based on the amplicon sequencing of the *orth10* gene. All samples are normalized to 10,000 sequences. In some cases, the short amplicon could not be clearly classified because of multiple top hits belonging to different groups (*e.g.*, g1&g7). The number in brackets indicates the number of unique phylotypes detected in the sewage sample. PJ, Panjin; LC, Liaocheng; HZ, Hangzhou; XM, Xiamen; SZ, Shenzhen.

Evidence of gene HGT among PCL members detected in the same primate host

In addition to *gyrB*, 16 PCL genomes were retrieved from the fecal metagenomes of nonhuman primates (only three host species, *i.e.*, *Papio cynocephalus, Pan troglodytes*, and *Gorilla gorilla*). Phylogeny of these strains representing seven uncultured species (designated as s1–s7 according to ANI values), *P. copri*, and Y7XP/Y7FG based on concatenated universal genes was generally consistent with that based on *gyrB* (**Figure 6**A, Figure S7). The CAZy modules of the seven uncultured species highly overlapped with those of *P. copri*. The few absent CAZy modules in the genomes of *P. copri* include CE3, GH30_2, GH39, and GH76 that putatively target xylan or mannan (Figure 6A), which are important plant cell wall components [40]. These polysaccharides are reasonably less abundant in the diet of modern human beings than in the diet of wild primates.

Since the HGT signals for CAZy-encoding genes have been detected among sympatric *P. copri* groups (Figure 3C), the potential HGT events among PCL members were examined. The upper boundary of 95% confidence interval of the similarity of universal marker genes was set as the threshold for recognizing the HGT genes (Figure S8). The bacterial species detected in different hosts only shared 0.6% genes with HGT signals, while the value was 4.0% for species from the same host (median value, Wilcoxon test, $P < 2.2 \times 10^{-16}$; Figure 6B). Although this phenomenon may be partially attributed to the potential genomic contamination for metagenome-derived genomes, analysis based on isolates still showed the high proportion of HGT signals for the species from the same host (the median proportion between 24 *P. copri* isolates and Y7XP/Y7FG was 4.3%). Figure 6C shows the gene synteny

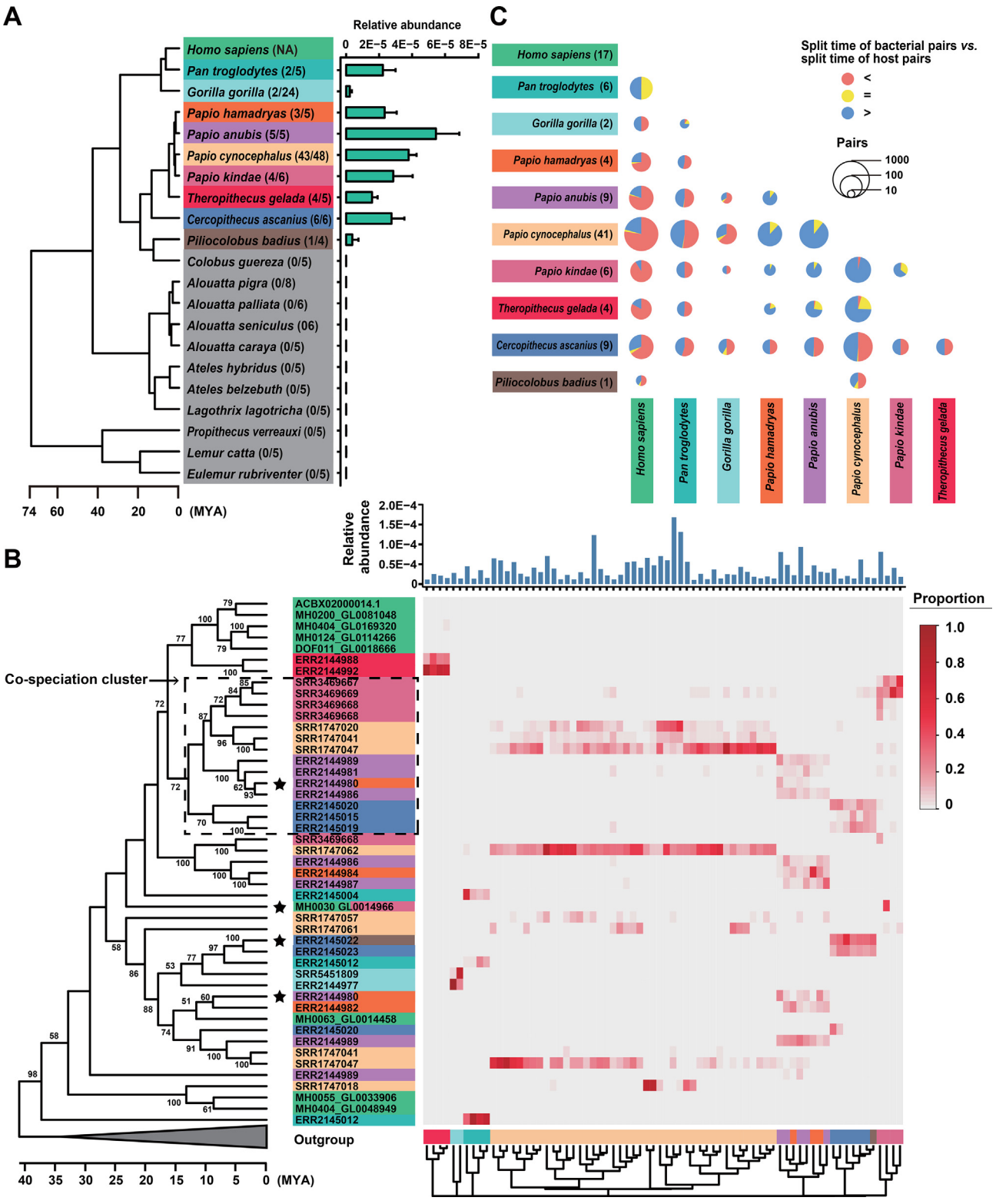

Figure 5 Diverse PCL members detected in the gut microbiome of nonhuman primates

A. Time-tree of hosts based on the evolutionary timescale. The relative abundance of total PCL in each host is shown in the barplot. **B.** Time constraint phylogenetic tree based on the 47 PCL *gyrB* representatives retrieved from wild nonhuman primates and human. The timescale is estimated by calibration based on the co-speciation cluster. Four representative sequences recovered from multiple host origins are marked by stars. Read-based abundances are shown in the heatmap. Only samples with over 1×10^{-5} relative abundance of PCL in the metagenomes are listed with clustering according to Bray-Curtis dissimilarity (bottom). **C.** Pie chart showing the consistency between the split time of bacterial pairs and the split time of host pairs. Only bacterial pairs more than 10 are shown. The number of *gyrB* sequences retrieved from the corresponding host is shown in the brackets. *gyrB*, gyrase subunit B; MYA, million years ago.

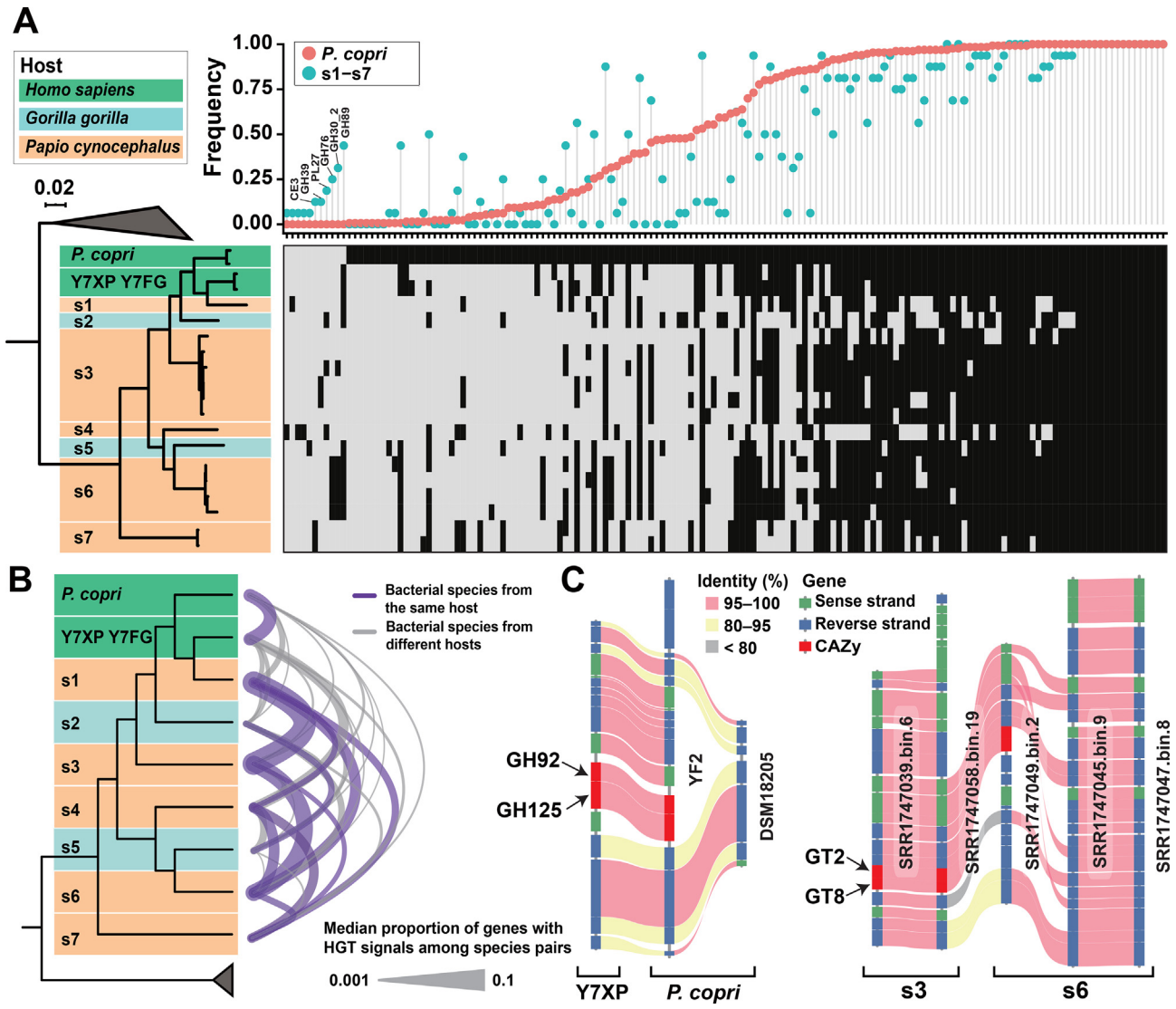

Figure 6 CAZys and HGT events of genomes affiliated with the PCL

A. Phylogenetic tree of genomes affiliated with the PCL inferred using 120 universal marker genes under the JTT model is shown on the left. The background color indicates the host origin. CAZy modules are shown in the heatmap (black, present; gray, absent). The dropline plot shows the frequency of the CAZy modules in 123 *P. copri* genomes (red) or 16 nonhuman primate-derived PCL genomes (green). **B.** HGT events among species of PCL. **C.** Two examples of selected CAZy-encoding genes with HGT signals shown by genomic synteny. HGT, horizontal gene transfer.

of representative CAZy-encoding genes with inter-species HGT signals, which were putatively occurred in homologous genomic regions.

PCL members in domesticated mammalian hosts were recently gained from humans

On the basis of the aforementioned results, PCL was thought to have co-evolved with higher primates for a long period. However, *P. copri* and its related taxa were widely detected in diverse non-primate domesticated mammals. Whether these taxa have evolved vertically in or horizontally transferred to non-primate mammalian hosts remains unknown. Therefore, potential PCL *gyrB* sequences were extracted from the gut microbiome gene catalog of pigs and mice [41,42]. Six pig-derived PCL *gyrB* sequences were phylogenetically affiliated with the PCL, but no mouse-derived PCL *gyrB* was found (Figure 7A). However, three of the six pig-derived sequences were clustered (> 98% similarity) with the human-derived *gyrB* representatives extracted from the IGC database. The other three sequences still shared > 95% identity to human-derived *gyrB* representatives. Noticeably, de-redundancy at a cut-off of 95% was conducted for the genes in the IGC database [37]. Hence, every pig-derived PCL member has close relatives in human-derived members, but not *vice versa*. This suggests that these PCL species are horizontally transferred from humans.

In reference to the aforementioned *gyrB* sequences, PCL was detected in fecal metagenomes from cats (n = 36), dogs (n = 125), pigs (n = 533), and bovines (n = 52) (Table S4). Figure 7B shows the relative abundance of total PCL in the samples with an abundance over 1×10^{-6} (one assigned as

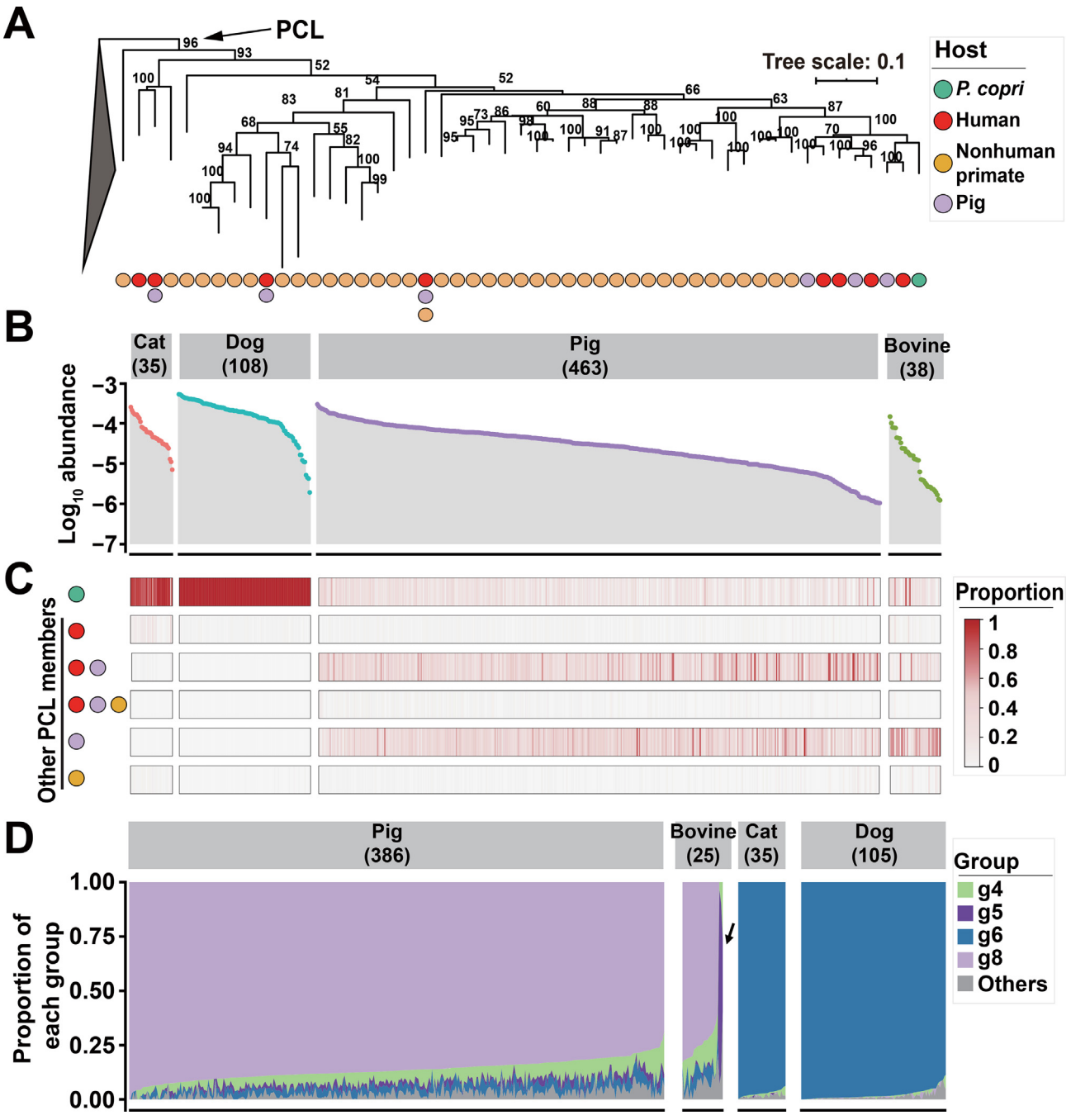

Figure 7 Distribution of PCL members and group-level profile of *P. copri* in the domesticated mammals

A. Phylogenetic tree based on representative *gyrB* sequences affiliated with the PCL retrieved from humans, nonhuman primates, and pigs. The circle color shows the host origin. **B.** Relative abundance of total PCL in mammalian gut metagenomes according to their hits against the *gyrB* database. Only samples with the abundance of total PCL *gyrB* $> 1 \times 10^{-6}$ are displayed, and the numbers of samples are presented in the brackets. **C.** Heatmap showing the proportion of PCL members retrieved from different host origins in the domesticated mammals. **D.** Group-level profile of *P. copri* in the mammalian gut metagenomes according to their hits against the *orth10* database. The numbers shown in the brackets represent the sample numbers of animals with *orth10* abundance $> 1 \times 10^{-6}$.

PCL *gyrB* per million reads, requiring > 95% similarity and 90% coverage; Figure S9). In particular, 35, 108, 463, and 38 samples passed the abundance threshold for cats, dogs, pigs, and bovines, respectively. All these hits were profiled into six catalogues, including *P. copri*, human-derived, human–pig shared, human–pig–nonhuman primate shared, pig-derived, and nonhuman primate-derived PCL members other than *P. copri* (Figure 7C). The PCL species detected in cat and dog samples were dominated by *P. copri*, and pig and bovine samples were mainly inhabited by *P. copri*, human–pig shared species, and pig-derived species.

Considering that *P. copri* was widely detected in these samples, group-level profiles of *P. copri* were established in the pig ($n = 386$), bovine ($n = 25$), cat ($n = 35$), and dog ($n = 105$)

samples with high *orth10* abundance ($> 1 \times 10^{-6}$). As shown in Figure 7D, *orth10*-based quantification showed that cats and dogs (sampled from Europe and North America, respectively) almost exclusively harbored g6 strains, which geographically co-occurred in Europe and North America. Pigs and bovines from Asia and Europe and bovines from North America were all dominated by g8. Bovines from Salvador were dominated by g5, a group also dominating the gut of Salvadorians.

Discussion

Co-evolutionary history for PCL in higher primates

Microbial samples from wild animals instead of domesticated ones are fundamental in determining their phylosymbiosis and co-evolutionary relationships to minimize the artificial influence from humans [5,10,14]. Besides that, our work emphasizes the need to use a comprehensive data source from multiple hosts in diverse geographical regions to obtain panoramic information [8,18,23]. *De novo* retrieval of PCL genomes from the metagenomes of most human and animal samples must be conducted due to the high genomic divergence and microdiversity [22,25]. In addition to genomes, the most comprehensive and accessible biomarkers, such as rRNA and *gyrB* gene sequences (although not the most precise), can also be used as references to bypass the limitations of genome binning (*e.g.*, low abundance and microdiversity). A recent study found that *P. copri* complex in human gut comprises four species-level clades based on the genomes retrieved from metagenomes via referring to a few core genomes [18]. The current work discovered that PCL members in the human gut and higher primates are far more diverse than only four species according to *gyrB* sequences and genomes from the expanded host spectrum. Similar to the co-speciation of *Bacteroide* spp. detected in extant hominid species [5], a signal of a few PCL members was found in four higher primates. However, the overall phylogenetic inconsistency suggested extensive horizontal transfer and extinction for the PCL members. A recent study showed the strong influence of environmental microbes on the gut microbiome of baboons [43], thus representing a possible pathway of host jumping. Our data of sharing the same PCL species among different wild hosts provided additional evidence for recent host jumping (Figure 5A). In addition, the extinction of certain species, which may be related to diet and behavior changes as observed in experimental animals over several generations [44], may also play important roles in the distribution of PCL members.

Are gut bacterial species shared by remotely related hosts evolved independently?

The genomic analysis further confirmed the intra-specific diversity and biogeographical pattern of *P. copri* [22,23]. Different *P. copri* groups shared critical or lower values to the species-level ANI and dDDH, suggesting their rapid evolutionary rates, which has also been proposed for endosymbionts [45], and experiencing allopatric speciation, a major mechanism for bacterial speciation [46]. Coincident with *H. pylori* [34] and *Eubacterium rectale* [47], our results indicate that *P. copri* has a consistent phylogeographical pattern with human migrating out of Africa, thus allowing calibration for its genomic evolutionary rate. As the low ANI values between conspecific strains in *H. pylori* have also been reported [48], the high evolutionary rates in *H. pylori* and *P. copri* raise a concern on whether a certain symbiotic bacterial species or clade shared by remotely related hosts has thoroughly evolved across the host phylogeny. Several previously reported gut bacteria, such as *Lactobacillus reuteri*, *Enterococcus faecalis*, and *Enterococcus faecium*, have experienced host-driven evolution across a wide range of mammalians or even vertebrates [6,7]. Oh et al. [6] deduced the split time of *L. reuteri* strains from multiple hosts by referring to a low mutation rate [49], resulting in molecular dating approximately 10 MYA. However, the low mutation rate has been suspected to be caused by the outdated methodology [50]. If these species have evolved at rates comparable with those of *H. pylori* and *P. copri*, then they are not likely to continuously co-evolve with the hosts for such a long period without speciation. The conspecific ancestral strains were possibly incorporated into the gut microbiome of various hosts recently and have initiated a host-driven adaption [13]. A very high frequency of genomic recombination among related taxa has been recently validated in the gut microbiome [51]. This undoubtedly can accelerate the genomic diversification process in addition to mutation. Comprehensive surveys on various symbiotic bacterial species across hosts will provide convincing evidence on clocking their diversifying processes.

Non-strict biogeography and host for PCL members suggest limited transmission barrier and potential niche selection

Poor host and geographical barriers have been observed in animal-associated microbial transmission [13,52]. Non-strict biogeographic distribution for the subspecies-level profile of *P. copri* in human gut was proven by our study as well as by Tett and colleagues [18]. The current study also revealed the putative extensive transmission of PCL members within different higher primates. Different (or at least for some) groups of *P. copri* and PCL members are distributed more ubiquitously than expected to a large extent following the microbiological tenet "everything is everywhere, but, the environment selects" [53]. Factors other than geographical and host isolation, such as host diets and behaviors and environmental characteristics, can favor their occurrence and dissemination in the local population and certain host species. This explanation could be supported by the strain-level profile of *P. copri* being associated with different habitual diets [54]; and some group-specific CAZy-encoding genes (*e.g.*, several alginase-encoding genes in g1 and a hyaluronidase-encoding gene in g6) were detected in our study.

Although no PCL *gyrB* sequence was detected in mouse gut bacterial gene catalog, which was generated from common laboratory mice (non-humanized) [42], a recent study verified the reliable transmission of Prevotellaceae from human feces to germ-free mice [55]. Our investigation on the PCL members in domesticated mammals suggested that these bacteria recently originated from sympatric human hosts and potentially experienced niche selection, *e.g.*, g8 of *P. copri* in bovines and pigs. Group g8 might be selected by the farming mode (*e.g.*, diet) for the bovines and pigs. The bovines in Salvador, which were dominated by g5, were domesticated in different

ways (*e.g.*, the animals may be closer to humans than the industrial farming mode and fed with different diets) [56]. Further comparison of the metabolic features of these animal-derived strains and human-derived strains may illustrate the evolutionary shaping of host adaptation within a limited period. However, obtaining high-quality genomes from domesticated animals is difficult, possibly due to the high microdiversity of PCL in their guts. Isolation of strains can be a more reliable way to obtain genomic representatives.

The potential relevance of intra-lineage horizontal transfer of functional genes

Each host provides a distinct niche (or a collection of subniches) that can be colonized by bacteria [13,57]. HGT and recombination are the main drivers of genomic divergence of the human gut microbiome [58,59] and play key roles in ecological adaption to new niches [51]. HGT is facilitated by the closely related phylogeny of donor and acceptor [51,60]. In addition, a recent study has reported that HGT events occur with a very high and extensive frequency in the human gut microbiome [51]. Despite focusing on the genes encoding CAZys and the uncharacterized HGT mechanisms, our results provided evidence that HGT events have widely occurred among sympatric *P. copri* strains and closely related intra-lineage PCL members from the same host. The unique glycan degradation capability is important for gut colonization and sustention in human gut bacteria [61]. Niche-driven rapid gain and loss of these genes within a large exchangeable pool may render the PCL members to be highly superior in the source competition. Moreover, extensive intra-lineage HGT events may result in the unreliable determination of specific phenotypes by group-level (or subspecies-level) and species-level identification. Caution must be adopted when linking phenotypes with taxa, because the novel gene function may be rendered by recent HGT. A direct comparison among strains with different phenotypes is preferred [20].

In summary, our study focused on the macroevolution and microevolution of PCL in the guts of higher primates and humans. The results provided panoramic insights into the multiple effects of vertical and horizontal transmission and niche selection on the host and biogeographical distribution and genomic evolution of a certain gut bacterial lineage. Studying the effects of PCL or other co-evolutionary lineages in animal guts on host phenotypes (*e.g.*, health or disease) from the co-evolution perspective can aid the comprehensive understanding of the interactions between host and gut microbes.

Materials and methods

Data collection for the 16S rRNA gene, *gyrB*, metagenomes, and genomes

16S rRNA gene sequences from type strains and clones classified as *Prevotella* (n = 534) were downloaded from EzBioCloud [62]. Seven additional sequences were obtained from *P. copri*-like isolates (GCF_002224675.1, GCA_001405915.1, and five contributed by this study). SILVA SSU reference database (v132) was used to track the host origins of *P. copri*-related sequences [63] (File S1).

As a species-level marker, the *gyrB* sequences of Prevotellaceae members in the human gut were retrieved from the IGC database of human, pig, and mouse gut microbiomes [37,41,42], the metagenomic assemblies of wild nonhuman primates, and 50 reference genomes (File S1). The *gyrB* sequences affiliated with the PCL were included in the database to profile PCL members in the gut metagenomes of humans, nonhuman primates, and domesticated mammalian hosts (File S1).

A total of 2811 publicly available gut metagenomes of humans, nonhuman primates, and other mammals were collected from 22 studies involving 26 host species from 37 countries (Table S4). We also collected 21 published *P. copri*-like genomes [36]. The present study contributed 139 new genomes (5 from isolates and 134 from metagenomes). Information for all genomes is listed in Table S1.

Isolates and genome sequencing

Fresh stool samples were collected from four healthy Chinese volunteers (the previous investigation on their gut microbiota suggested a high abundance of *P. copri*-like taxa) and immediately transferred to an anaerobic glovebox (N_2:CO_2:H_2 = 80:15:5) for isolation on YCFA medium [64]. Colonies were picked after cultivation at 37 °C for 96 h. Full-length 16S rRNA gene sequences of the isolates were used to identify *P. copri* and its related strains on the EzBioCloud platform [62].

Genomic DNA of *P. copri* and its related species isolated in this study was extracted and sequenced with PE150 strategy on Illumina Hiseq 4000 platform (Novogene, Beijing, China). *De novo* assembly was performed by SPAdes (v3.9.0) [65]. Only scaffolds longer than 1000 bp were included in the downstream analysis. The whole genome of the YF2 strain was achieved by combining Illumina and PacBio RSII platform sequencing (Novogene, Beijing, China).

Genome binning, quality assessment, and annotation

Genomic binning using mmgenome was manually performed to obtain high-quality *de novo* assembled genomes of *P. copri* and related taxa from humans and nonhuman primates, respectively [66]. Prescreening of the 1679 gut metagenomes from humans (Table S4) was conducted using the relative abundance of *P. copri* as estimated by the relative abundance of *gyrB* (usually $> 1 \times 10^{-5}$ for IGC data) or MetaPhlAn v2.0 (> 10% relative abundance for non-IGC data) to improve the efficiency [67]. For the 168 gut metagenomes from nonhuman primates (Table S4), raw reads were quality filtered with Trimmomatic (v.0.36) [68]. The raw reads of selected human and nonhuman primate samples were first assembled using SPAdes (v3.9.0) [65]. Only scaffolds longer than 1000 bp were retained for genome binning. The necessary files were generated using script data.generation.2.1.0.sh [66].

Completeness and contamination of all draft genomes were assessed by CheckM (v1.0.7) [69]. Pairwise ANI and dDDH values among *P. copri* genomes were calculated by FastANI (v1.3) [70] and Genome-to-Genome Distance Calculator 2.1 [29], respectively. Genes encoding CAZy families were annotated using dbCAN HMMs (v6) [71], and the results were filtered according to the recommended threshold.

Defining core protein orthologs of *P. copri*

Core orthologous gene clusters of *P. copri* genomes were defined using the method of Oyserman et al. [72] with modifications. All incomplete open reading frames (ORFs) with potential redundancy (*i.e.*, multiple fragments from one ORF) were cleaned prior to downstream analysis (File S1). All-against-all BLASTP was performed for cleaned ORFs from *P. copri* genomes [73]. Identity and inflation values were determined according to McCill et al. [74] by maximizing the maintenance of genes with the same function in a cluster (File S1; Table S5). A total of 1095 single-copy core orthologs that appeared in more than 90% of the *P. copri* genomes were determined (File S1).

Phylogenetic analysis

Phylogenetic analyses were performed for single genes and concatenated alignments of single-copy core ORFs. The trees based on single genes were reconstructed using MEGA (v6.06) with 100 bootstrap iterations [75], and those based on concatenated genes were reconstructed by maximum-likelihood analysis using RAxML (v8.2.4) [76] or FastTree (v2.1) [77]. A truly whole genome-based phylogenetic analysis of the coding sequences was conducted at the nucleotide level using the latest version of the Genome BLAST Distance Phylogeny (GBDP) method under recommended settings [29,78]. All phylogenetic trees were visualized via the iTOL web server [79]. Further details are described in File S1.

Determination and application of a quantitative gene with the intra-specific resolution for metagenomes

A quantitative marker gene with intra-specific resolution must be selected because *P. copri* has high genomic diversity that may cause quantitative biases at the subspecies level in metagenomes. For *P. copri* genomic pairs, the best candidate was determined by calculating the Spearman correlation of distances between 1095 concatenated single-copy core orthologs and each single core ortholog (Table S6; File S1). The optimized gene was designated as *orth10* (the corresponding gene in the type strain DSM18205 was EFB36125.1, encoding a response regulator receiver domain protein), and was used as the basis for the group-level profiling of *P. copri* in human and mammalian gut metagenomes and the investigation on *P. copri* populations in raw sewages collected from five cities [Panjin (PJ), Liaocheng (LC), Hangzhou (HZ), Xiamen (XM), and Shenzhen (SZ)] of China (Tables S7 and S8; File S1).

Molecular dating for the split time between *P. copri* groups and between PCL members

Molecular dating was conducted as previously described [6]. PHI test was used to identify the intragenic recombination of 120 universal genes as proposed by previous studies [30,31]. The dN/dS ratios for genes without significant intragenic recombination (PHI test, $P > 0.05$) were calculated by using KaKs_Calculator [80]. Split time was estimated by the synonymous mutation rate among various groups and the long-term mutation rate of housekeeping genes of another human gut symbiont, *H. pylori* (2.6×10^{-7} per site per year) [32].

The divergence time of the *gyrB* sequences retrieved from IGC and nonhuman primates was estimated by Bayesian MCMC analysis implemented in BEAST2 (v2.5.2) [81]. The bacterial lineages showing signals of co-speciation with primate hosts were used as calibration, and the maximum-likelihood tree inferred by MEGA was employed as the starting tree. The analysis was run 50 million generations and sampled every 1000 steps under the GTR + G + I substitution model with a lognormal relaxed molecular clock [5]. Tracer (v1.7.1; http://tree.bio.ed.ac.uk/software/tracer/) was utilized to ensure that the effective sample size was larger than 200 for all parameters. The tree files were summarized in TreeAnnotator with the first 25% discarded as burn-in [82].

Determination of HGT events among PCL species

All-against-all BLSATN was performed between heterospecific genome pairs to define the HGT events among PCL species [73]. Shared genes with a high similarity between any two heterospecific genomes were classified as HGT due to the lack of available tools to identify HGT events among closely related species. For a given species pair, the HGT signal threshold was set as the upper boundary of the 95% confidence interval of similarity between complete universal genes [30]. Gene pairs with similarity higher than the threshold were recognized as HGT-positive, and the proportion of genes with HGT signals was calculated for each genome pair ($\frac{\text{Number of genes with HGT signals}}{\text{Total number of ORFs in one genome}}$).

Statistical analysis and visualization

Statistical analysis was conducted in R (v3.5.1). The *rcompanion* (v2.0.0; https://CRAN.R-project.org/package=rcompanion) package was used for Fisher exact tests, and *ggplot2* [83], *pheatmap* (v1.0.10; https://CRAN.R-project.org/package=pheatmap), and *ggalluvial* (v0.9.1; https://doi.org/10.21105/joss.02017) packages were applied for data visualization. The genomic synteny of the fragments containing CAZy-encdoing genes was visualized by MCscan (Python version) [84].

Ethical statement

All participants provided signed informed consent.

Data availability

The 16S rRNA gene sequences contributed by this study have been deposited in GenBank (GenBank: MN658562–MN658566). The raw genome sequencing data of isolates in this study and the *orth10* amplicon sequencing data have been deposited in the Sequence Read Archive database (SRA: PRJNA555745, PRJNA565808, and PRJNA557417). The raw sequencing data have also been deposited in the Genome Sequence Archive [85] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GSA: CRA004122), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The

genomes of isolates have been deposited in the Genome Warehouse database [86] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (BioProject: PRJCA004805), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. All genome sequences have been deposited in Figshare at https://doi.org/10.6084/m9.figshare.14626872.

CRediT author statement

Hao Li: Investigation, Data curation, Formal analysis, Validation, Visualization, Writing - original draft, Writing - review & editing. **Jan P. Meier-Kolthoff:** Data curation, Writing - review & editing. **Canxin Hu:** Investigation. **Zhongjie Wang:** Data curation. **Jun Zhu:** Investigation. **Wei Zheng:** Writing - review & editing. **Yun Tian:** Writing - review & editing. **Feng Guo:** Conceptualization, Supervision, Writing - original draft, Writing - review & editing, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant Nos. 31670492 and 31500100). We thank Prof. Xin Yu and Mr. Chengsong Ye for their help in obtaining the sewage samples.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.10.006.

ORCID

ORCID 0000-0002-8151-3830 (Hao Li)
ORCID 0000-0001-9105-9814 (Jan P. Meier-Kolthoff)
ORCID 0000-0003-0577-0891 (Canxin Hu)
ORCID 0000-0001-6498-6532 (Zhongjie Wang)
ORCID 0000-0003-4934-5889 (Jun Zhu)
ORCID 0000-0002-1999-6376 (Wei Zheng)
ORCID 0000-0003-3233-4470 (Yun Tian)
ORCID 0000-0001-7785-8388 (Feng Guo)

References

[1] Rook G, Bäckhed F, Levin BR, McFall-Ngai MJ, McLean AR. Evolution, human-microbe interactions, and life history plasticity. Lancet 2017;390:521–30.
[2] Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR. Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. PLoS Biol 2016;14:e2000225.
[3] Groussin M, Mazel F, Alm EJ. Co-evolution and co-speciation of host-gut bacteria systems. Cell Host Microbe 2020;28:12–22.
[4] Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'Dwyer JP, Sharpton TJ. Ecophylogenetics clarifies the evolutionary association between mammals and their gut microbiota. mBio 2018;9:e01348-18.
[5] Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, et al. Cospeciation of gut microbiota with hominids. Science 2016;353:380–2.
[6] Oh PL, Benson AK, Peterson DA, Patil PB, Moriyama EN, Roos S, et al. Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. ISME J 2010;4:377–87.
[7] Lebreton F, Manson AL, Saavedra JT, Straub TJ, Earl AM, Gilmore MS. Tracing the enterococci from Paleozoic origins to the hospital. Cell 2017;169:849–61.
[8] Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res 2016;26:1612–25.
[9] Shapira M. Gut microbiotas and host evolution: scaling up symbiosis. Trends Ecol Evol 2016;31:539–49.
[10] Perofsky AC, Lewis RJ, Meyers LA. Terrestriality and bacterial transfer: a comparative study of gut microbiomes in sympatric Malagasy mammals. ISME J 2019;13:50–63.
[11] Toju H, Tanabe AS, Notsu Y, Sota T, Fukatsu T. Diversification of endosymbiosis: replacements, co-speciation and promiscuity of bacteriocyte symbionts in weevils. ISME J 2013;7:1378–90.
[12] Kikuchi Y, Hosokawa T, Fukatsu T. An ancient but promiscuous host–symbiont association between *Burkholderia* gut symbionts and their heteropteran hosts. ISME J 2011;5:446–60.
[13] Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. Nat Rev Genet 2018;19:549–65.
[14] Amato KR. Co-evolution in context: the importance of studying gut microbiomes in wild animals. Microbiome Sci Med 2013;1:10–29.
[15] Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 2018;3:8–16.
[16] Tett A, Pasolli E, Masetti G, Ercolini D, Segata N. *Prevotella* diversity, niches and interactions with the human host. Nat Rev Microbiol 2021:1–15.
[17] Ley RE. *Prevotella* in the gut: choose carefully. Nat Rev Gastroenterol Hepatol 2016;13:69–70.
[18] Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, et al. The *Prevotella copri* complex comprises four distinct clades underrepresented in westernized populations. Cell Host Microbe 2019;26:666–79.e7.
[19] Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee Y, De Vadder F, Arora T, et al. Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. Cell Metab 2015;22:971–82.
[20] Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. Elife 2013;2:e01202.
[21] Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BAH, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. Nature 2016;535:376–81.
[22] Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res 2017;27:626–38.
[23] Costea PI, Coelho LP, Sunagawa S, Munch R, Huerta-Cepas J, Forslund K, et al. Subspecies in the global human gut microbiome. Mol Syst Biol 2017;13:960.
[24] Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, et al. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 2015;3:1.

[25] Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 2013;31:533–8.
[26] Barr T, Sureshchandra S, Ruegger P, Zhang J, Ma W, Borneman J, et al. Concurrent gut transcriptome and microbiota profiling following chronic ethanol consumption in nonhuman primates. Gut Microbes 2018;9:338–56.
[27] Brooke CG, Najafi N, Dykier KC, Hess M. *Prevotella copri*, a potential indicator for high feed efficiency in western steers. Anim Sci J 2019;90:696–701.
[28] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 2017;35:725–31.
[29] Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 2013;14:60.
[30] Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018;36:996–1004.
[31] Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics 2006;172:2665–81.
[32] Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, et al. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. PLoS Genet 2010;6:e1001036.
[33] Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of human migrations in *Helicobacter pylori* populations. Science 2003;299:1582–5.
[34] Hershkovitz I, Weber GW, Quam R, Duval M, Grün R, Kinsley L, et al. The earliest modern humans outside Africa. Science 2018;359:456–9.
[35] Ndeh D, Rogowski A, Cartmell A, Luis AS, Baslé A, Gray J, et al. Complex pectin metabolism by gut bacteria reveals novel catalytic functions. Nature 2017;544:65–70.
[36] Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol 2019;37:179–85.
[37] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 2014;32:834–41.
[38] Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 2017;357:802–6.
[39] Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun 2014;5:3654.
[40] Selvendran RR. The plant cell wall as a source of dietary fiber: chemistry and structure. Am J Clin Nutr 1984;39:320–37.
[41] Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, et al. A reference gene catalogue of the pig gut microbiome. Nat Microbiol 2016;1:16161.
[42] Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, et al. A catalog of the mouse gut metagenome. Nat Biotechnol 2015;33:1103–8.
[43] Grieneisen LE, Charpentier MJE, Alberts SC, Blekhman R, Bradburd G, Tung J, et al. Genes, geology and germs: gut microbiota across a primate hybrid zone are explained by site soil properties, not host species. Proc Biol Sci 2019;286:20190431.
[44] Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. Nature 2016;529:212–5.
[45] Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet 2008;42:165–90.
[46] Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science 2007;315:476–80.
[47] Karcher N, Pasolli E, Asnicar F, Huang KD, Tett A, Manara S, et al. Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. Genome Biol 2020;21:138.
[48] On SLW, Miller WG, Houf K, Fox JG, Vandamme P. Minimal standards for describing new species belonging to the families Campylobacteraceae and Helicobacteraceae: *Campylobacter*, *Arcobacter*, *Helicobacter* and *Wolinella* spp. Int J Syst Evol Microbiol 2017;67:5296–311.
[49] Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. Proc Natl Acad Sci U S A 1999;96:12638–43.
[50] Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 2008;6:431–40.
[51] Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell 2021;184:2053–67.e18.
[52] Livermore JA, Jones SE. Local–global overlap in diversity informs mechanisms of bacterial biogeography. ISME J 2015;9:2413–22.
[53] de Wit R, Bouvier T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? Environ Microbiol 2006;8:755–8.
[54] De Filippis F, Pasolli E, Tett A, Tarallo S, Naccarati A, De Angelis M, et al. Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. Cell Host Microbe 2019;25:444–53.e3.
[55] Fouladi F, Glenny EM, Bulik-Sullivan EC, Tsilimigras MCB, Sioda M, Thomas SA, et al. Sequence variant analysis reveals poor correlations in microbial taxonomic abundance between humans and mice after gnotobiotic transfer. ISME J 2020;14:1809–12.
[56] Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, Navarrete KM, et al. Interconnected microbiomes and resistomes in low-income human habitats. Nature 2016;533:212–6.
[57] Toft C, Andersson SGE. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat Rev Genet 2010;11:465–75.
[58] Bailly X, Olivieri I, Brunel B, Cleyet-Marel JC, Béna G. Horizontal gene transfer and homologous recombination drive the evolution of the nitrogen-fixing symbionts of Medicago species. J Bacteriol 2007;189:5223–36.
[59] Potnis N, Kandel PP, Merfa MV, Retchless AC, Parker JK, Stenger DC, et al. Patterns of inter-and intrasubspecific homologous recombination inform eco-evolutionary dynamics of *Xylella fastidiosa*. ISME J 2019;13:2319–33.
[60] Popa O, Dagan T. Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol 2011;14:615–23.
[61] Shepherd ES, DeLoache WC, Pruss KM, Whitaker WR, Sonnenburg JL. An exclusive metabolic niche enables strain engraftment in the gut microbiota. Nature 2018;557:434–8.
[62] Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 2017;67:1613–7.
[63] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2012;41:D590–6.
[64] Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 2016;533:543–6.
[65] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and

its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.
[66] Karst SM, Kirkegaard RH, Albertsen M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. bioRxiv 2016;059121.
[67] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9:811–4.
[68] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[69] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25:1043–55.
[70] Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018;9:5114.
[71] Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 2012;40:W445–51.
[72] Oyserman BO, Moya F, Lawson CE, Garcia AL, Vogt M, Heffernen M, et al. Ancestral genome reconstruction identifies the evolutionary basis for trait acquisition in polyphosphate accumulating bacteria. ISME J 2016;10:2931–45.
[73] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[74] McGill S, Barker D. Comparison of the protein-coding genomes of three deep-sea, sulfur-oxidising bacteria: "*Candidatus* Ruthia magnifica", "*Candidatus* Vesicomyosocius okutanii" and *Thiomicrospira crunogena*. BMC Res Notes 2017;10:296.
[75] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013;30:2725–9.
[76] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.
[77] Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.
[78] Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun 2019;10:2182.
[79] Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 2016;44:W242–5.
[80] Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 2006;4:259–63.
[81] Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2019;15:e1006650.
[82] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2014;10:e1003537.
[83] Wickham H. ggplot2: elegant graphics for data analysis. Berlin: Springer Publishing Company; 2016.
[84] Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science 2008;320:486–8.
[85] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.
[86] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics 2021;19:584–9.