npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-024-00557-x

Host specificity and cophylogeny in the "animal-gut bacteria-phage" tripartite system

Check for updates

Ye Feng[1,2] ✉, Ruike Wei[3], Qiuli Chen[2], Tongyao Shang[2], Nihong Zhou[3], Zeyu Wang[2], Yanping Chen[4], Gongwen Chen[3], Guozhi Zhang[3], Kun Dong[5], Yihai Zhong[6], Hongxia Zhao[7], Fuliang Hu[3] & Huoqing Zheng[3] ✉

Cophylogeny has been identified between gut bacteria and their animal host and is highly relevant to host health, but little research has extended to gut bacteriophages. Here we use bee model to investigate host specificity and cophylogeny in the "animal-gut bacteria-phage" tripartite system. Through metagenomic sequencing upon different bee species, the gut phageome revealed a more variable composition than the gut bacteriome. Nevertheless, the bacteriome and the phageome showed a significant association of their dissimilarity matrices, indicating a reciprocal interaction between the two kinds of communities. Most of the gut phages were host generalist at the viral cluster level but host specialist at the viral OTU level. While the dominant gut bacteria *Gilliamella* and *Snodgrassella* exhibited matched phylogeny with bee hosts, most of their phages showed a diminished level of cophylogeny. The evolutionary rates of the bee, the gut bacteria and the gut phages showed a remarkably increasing trend, including synonymous and non-synonymous substitution and gene content variation. For all of the three codiversified tripartite members, however, their genes under positive selection and genes involving gain/loss during evolution simultaneously enriched the functions into metabolism of nutrients, therefore highlighting the tripartite coevolution that results in an enhanced ecological fitness for the whole holobiont.

Animal gut harbors a myriad of gut bacteria that play important roles in modulating the immune functions during development, synthesizing vital chemical compounds like hormones, and aiding in the process of digestion while competitively excluding pathogens[1]. In return, gut bacteria depend on their host to maintain a stable ecosystem and access nutrients. The prevalence of these intimate interactions is often cited as evidence supporting the notion that animals and their gut bacteria have mutually evolved into a symbiosis relationship[2]. Typical symbiotic relationships frequently exhibit a fascinating trend of congruent phylogenies[3]. However, because gut bacteria may undergo frequent horizontal transmission between individuals that may disrupt the cophylogenetic signal, strong host-bacteria fidelity over evolutionary timescales was only seen in a few bacterial taxa, e.g., less than 1/3 of gut bacteria in humans[4,5]. Given that many of these bacteria are tightly and specifically linked to host development and immune functions, identification of bacteria that have coevolved with their hosts carries essential implications for understanding how gut bacteria impact host physiology and how host selection influences the composition of the gut microbiome.

In comparison to the bacterial components of the gut microbiome, the viral fraction, known as phageome, is characterized by a greater genetic diversity and a more labile composition between individuals[6]. Although a majority of gut bacteriophages remain uncultured and unclassified[7], dysbiotic gut phageomes have been linked to diseases such as colitis and type II diabetes[8,9]. This underscores the potential impact of phageome on the overall health of their superhost, a term referring to the host of bacteria

[1]Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China. [2]Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, China. [3]College of Animal Sciences, Zhejiang University, Hangzhou, China. [4]USDA-ARS Bee Research Laboratory, Beltsville, MD, USA. [5]Eastern Bee Research Institute, College of Animal Science and Technology, Yunnan Agricultural University, Kunming, Yunnan, China. [6]Environment and Plant Protection Institute, Chinese Academy of Tropical Agriculture Sciences, Haikou, Hainan, China. [7]Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Guangdong Public Laboratory of Wild Animal Conservation and Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou, China.
✉e-mail: pandafengye@zju.edu.cn; hqzheng@zju.edu.cn

which in turn host the phages. The impact of phageome is primarily believed to occur through the capacity of phages to shape the ecology of their bacterial hosts via top-down pressure. However, it has also been observed that phages can directly modulate the superhost's metabolism or immune system[10].

Given that coevolution could occur between animals and their gut bacteria, as well as between gut bacteria and their phages[4,11], these two pairs can theoretically combine to create a coevolving "animal-gut bacteria-phage" system. However, whether and how this tripartite interplay adapts to the external environment and influence the animal's development, health and even evolution remain largely unclear. When focusing on a narrower point, it is generally believed that host tropism of bacteriophage is constrained by phylogenetic barriers, typically limiting one viral cluster (VC) to infecting one bacterial species[6]. Some gut bacteria with particularly close link to the host's functions may also exhibit specific host range[5]. Within this context, the question arises as to whether host specificity and cophylogeny are likely to manifest across the entire tripartite system. Notably, Gogarten et al. found that 22.1% of the gut phages showed cophylogeny with their primate superhost, in which, however, the intermediate bacteria were not included for analysis[12]. The answers to the aforementioned questions undoubtedly hold important ecological and evolutionary implications. Moreover, they also have practical relevance in clinical settings, such as in fecal microbiota transplantation and phage therapy against multi-drug resistance bacteria.

In this study, we utilize a bee model to investigate the host specificity and cophylogeny of the "animal-gut bac-phage" tripartite system. Eusocial corbiculate bees include honeybees (*Apis* spp.) and bumble bees (*Bombus* spp.), and the *Apis* genus was further split into the Eastern honeybee *A. cerana* and the Western honeybee *A. mellifera* due to geographic isolation[13]. These bee species share nearly the same set of gut microbiota that comprise approximately ten phylotypes only[14], and most of these phylotypes have further diverged into multiple "sequence-discrete populations" (SDPs). Both phylotype and SDP are phylogenetic units that separate an organism from the rest of the community by genetic and genealogical discontinuity. While phylotype is often defined a threshold identity (e.g., ≥97%) of 16 s rRNA genes and therefore approximates genus- or species-level taxon, SDP represents a finer subdivision of phylotype and is characterized by high (e.g., ≥95%) genome-wide average nucleotide identity (gANI)[15]. The simple and conserved composition of the gut microbiota in bees makes it easy to obtain the high-quality contigs for the gut bacteria though metagenomic sequencing. Furthermore, due to the limited number of the bacterial phylotypes, it is simpler to make host prediction for the gut phages and, subsequently, to examine the interaction between phages and bacteria. Last but not least, the divergence of the bacterial phylotypes, such as *Gilliamella* and *Snodgrassella*, into SDPs have been known to mirror the evolutionary history of their respective bee hosts[16–18]. During the coevolution, the bacteria of different SDPs have developed elaborate mechanisms to adapt to their specific bee hosts, including the recognition of neuro signals and the formation of biofilm[19,20]. These features make bee a particularly convenient animal model for further testing the downstream cophylogeny between gut bacteriome and phageome.

Currently, a few studies have characterized the gut phageome in *A. mellifera*, which identified a large repertoire of novel phages with remarkable genetic diversity[21–23]. However, knowledge regarding the phageome of *A. cerana* and *Bombus* remains unknown. Through comprehensive sequencing of the paired gut bacteriome and phageome of *A. cerana*, *A. mellifera* and *B. terrestris*, we have not only gauged the diversity of the phageome and bacteriome but also assessed the quantitative relationship between them. By conducting comparative genomic and phylogenetic analyses, we assessed the chained cophylogeny in the "animal-gut bacteria-phage" system, and compared the evolutionary rates between the tripartite members. Collectively, this study provides an unprecedented groundwork for future investigations into the ecological, evolutionary and functional host-microbe interactions.

Results

Conserved and codiversified bacteria in the bee gut

We collected *A. cerana* and *A. mellifera* from Zhejiang, Yunnan, and Hainan provinces in China, respectively (Fig. 1a). At each region, *A. cerana* and *A. mellifera* colonies were kept in the same apiary, with two colonies being collected for each species. Meanwhile, commercialized *B. terrestris* was reared in the same apiaries for a duration of two months. By sequencing the genomes of these bee samples, a total of 5053 single-copy orthologous genes were conserved among all samples across the three bee species. A phylogenetic tree based on concatenated core coding sequences was built, in which the bee samples were firstly clustered by species and further by regions (Fig. 1b). The branch length of the tree was roughly proportional with the divergence time inferred from the fossils that the genus *Apis* and *Bombus* split approximately 82 million years ago and the species *A. cerana* and *A. mellifera* split approximately 11.8 million years ago (Fig. 1b)[24,25].

We conducted shotgun sequencing of the gut bacteriome, and confirmed that all individual gut communities were dominated by ten core bacterial phylotypes (Fig. 1c). The principal co-ordinates analysis (PCoA) showed that the composition was distinguishable between bee species (Anosim test, $r = 0.319$, $p = 0.001$; Fig. 1d) but not between geographical regions (Anosim test, $r = 0.095$, $p = 0.166$). The functional profile inferred from the entire gut bacteriome showed significantly different abundance in a few categories between bee species, including carbon degradation, oxygen and sulfur metabolism (Kruskal–Wallis test, $p < 0.05$; Fig. 1c).

Since *Gilliamella* and *Snodgrassella* were the only two phylotypes that accounted for >1% of abundance among all samples, they were chosen for further profiling at SDP level. As revealed in the core gene-based phylogenetic trees for both *Gilliamella* and *Snodgrassella*, the metagenome-assembled genomes (MAGs) derived from this study and the isolate genomes downloaded from Genbank database were organized into three distinct and robust major SDPs, which aligned with the respective bee species rather than the geographical origin (Fig. 2a; Supplementary Fig. 1). The strong cophylogenetic signal demonstrated that coevolution existed between the bee host and the two gut bacteria (Parafit test, $p = 0.006$ for *Gilliamella* & $p = 0.002$ for *Snodgrassella*).

The genetic divergence between the SDPs in both *Gilliamella* and *Snodgrassella* was remarkable. The average amino acid distance for the conserved genes between SDPs was 12–18%; and the gANI was 78–81% (Supplementary Fig. 2). Even within the 16S gene sequences, there were substantial differences between the SDPs, with the maximal distance of up to 3.3% for *Gilliamella* and 1.8% for *Snodgrassella* (Supplementary Fig. 2). Gene content of the bacterial SDPs were clustered by bee species instead of by region (Fig. 2b). The core-pan curves showed that both *Gilliamella* and *Snodgrassella* possessed a closed pangenome (Supplementary Fig. 2). The ratio of pan- and core-genome size for *Gilliamella* (5.95) was much higher than that for *Snodgrassella* (3.75), indicating a much more expanded auxiliary gene repertoire for *Gilliamella*. To evaluate the extent of recombination within bacterial genomes, we not only examined the individual gene-based trees for their topology discordance with the concatenated gene tree, but also detected recombination events within conserved chromosomal regions (Supplementary Fig. 3). Both approaches showed a higher frequency of recombination in bacteria compared to the bees, and yet the majority of recombination events in bacteria occurred within SDPs rather than between SDPs (Supplementary Fig. 3).

The number of mapped reads showed the dominance of host-specific SDPs in each sample, which indicated that *Gilliamella* and *Snodgrassella* were transmitted in bees mainly in a vertical way (Fig. 2c). However, each host was simultaneously colonized by non-native SDPs, albeit in smaller proportions. Notably, nearly one quarter of both *Gilliamella* and *Snodgrassella* in a *B. terrestris* sample raised in Hainan were assigned to the *A. cerana*-specific SDP (Fig. 2c). Furthermore, the non-native SDPs from Hainan were not genetically similar to the *A. cerana*-specific SDPs from

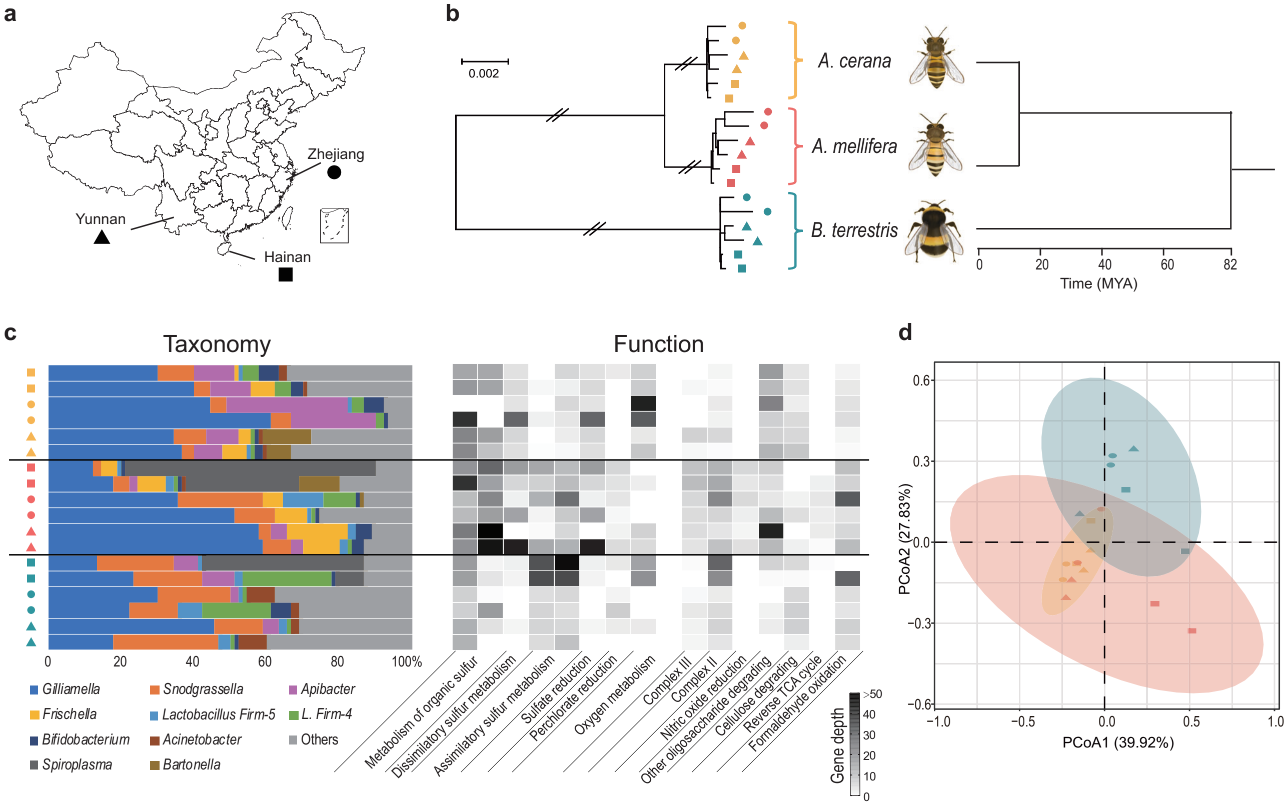

Fig. 1 | Bee model and their gut bacteriome at the phylotype level. **a** Locations where bees are sampled. **b** Maximum likelihood phylogenetic tree comprising the three bee species based on concatenated coding nucleotide sequence (left tree) and the evolutionary timescale inferred by the TimeTree database (right tree). For the left tree, the branches marked by slashes are shortened 10-fold for an enlarged view of the terminal branches. **c** Taxonomic and functional profile of bee gut bacteriome. The taxonomic profile (left plot) shows the relative abundances of phylotypes; the functional profile (right plot) shows the sequencing depth of genes that are assigned to each category. Only the functions that show significant difference between the bee species are listed. **d** PCoA plot based on the pairwise Bray–Curtis dissimilarity between samples. The shape and color of the dots that represents individual samples in (**c**, **d**) are the same as (**a**, **b**).

Hainan samples but to that from other regions (Supplementary Fig. 4). This indicated dispersal of the same SDP across broad regions, which explained why the phylogenetic tree of bacterial SDPs was not clustered by region.

Host specificity and cophylogeny of bacteriophages in the bee gut

We enriched the virus-like particles (VLPs) for each bee gut sample and subsequently sequenced the virome. According to the quality control analysis with the ViromeQC tool, 16 out of 18 samples showed a VLP enrichment score higher than 5, 13 of which had the score even higher than 10 (Supplementary Table 1), suggesting that most of these samples did not suffer severe bacterial or eukaryotic contamination. By clustering with vConTACT software, the derived phage sequences together with the prophage sequences predicted from the bacterial isolate genome and the MAGs were assigned to 794 VCs. Only 32 VCs (4.41%) were taxonomically annotated, with the others not overlapping with the RefSeq phages, suggesting the gut phages may have a restricted ecological niche primarily within bee gut.

Few VCs were conserved among all samples or even among all samples from a particular bee species, indicating the absence of core phageome in the bee gut. PCoA based on VC abundance can moderately cluster samples by bee species (Anosim test, $r = 0.549$, $p = 0.001$; Fig. 3a). Compared to the bacteriome, the phageome showed a much more labile composition as revealed by the Shannon index (Fig. 3b). The Shannon index of bacteriome and phageome were not correlated with each other (Pearson correlation, $p = 0.247$; Fig. 3c). According to the Bray-Curtis dissimilarity, nevertheless, the phageome composition was significantly correlated to the bacteriome composition at the phylotype level (Mantel test, $r = 0.331$, $p = 0.001$; Supplementary Fig. 5), indicating a reciprocal interaction between the two kinds of communities.

In our efforts to predict the host of these phages, we identified 95 VCs and 27 VCs originating from *Gilliamella* and *Snodgrassella*, respectively. Notably, most of VCs from *Snodgrassella* were more or less related to each other in gene content (Fig. 3d), suggesting they belonged to a few viral families only. Visualization of vConTACT analysis did not distinguish VCs from different bee species or from different geographical regions (Fig. 3d). Compared to *Gilliamella* phages, *Snodgrassella* phages had an extremely low abundance (<1%) in half of the samples; the four samples with abundance > 10% all happened to be from Zhejiang region (Fig. 3e). Similar to that at the phylotype level, for both *Gilliamella* and *Snodgrassella*, the composition of their phages was also moderately correlated with the bacterial composition at the SDP level based on the Bray-Curtis dissimilarity (Mantel test, $p < 0.05$; Supplementary Fig. 5). This suggested that the interaction between the phages and their bacterial host was also present at SDP level. For *Gilliamella*, in particular, a linear relationship of relative abundance was observed between the phages and the bacterium, with the slope of the linear regression being significantly <1 (Fig. 3e). Thus, *Gilliamella* phages did not proliferate as rapidly as their bacterial host did. The same trend was not observed in *Snodgrassella*.

Moving forward, we checked host specificity of the gut phages. Although 470 (64.8%) VCs showed their members originated from a single bee species, there was a clear trend that VCs with a greater number of members were more likely to be found in multiple hosts (Fig. 4a). Among these, 63 VCs appeared in all of the three bee species, which belong to host generalist. According to the Parafit test, majority of these VCs codiversified with their bacterial host; but these significant cophylogenetic signals

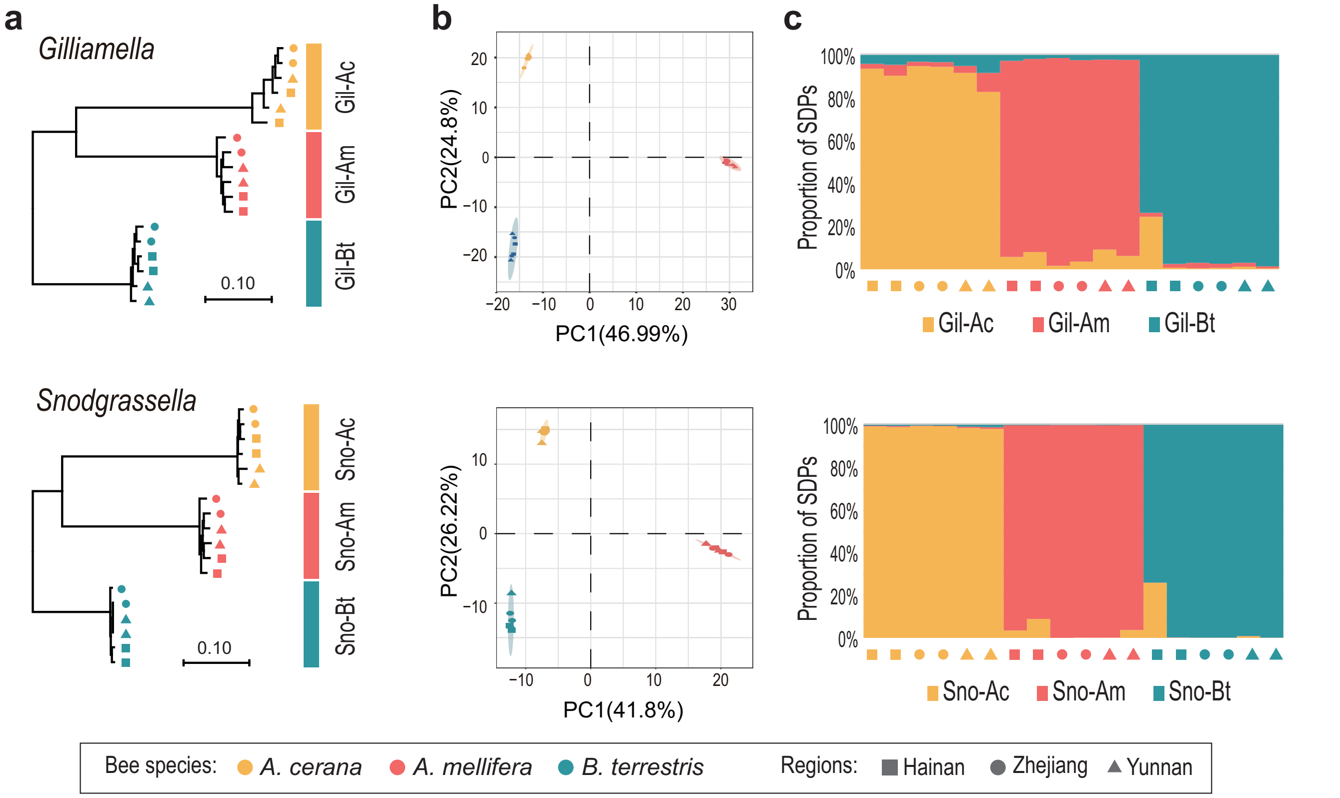

Fig. 2 | Genetic comparison between SDPs and compositional abundance of gut bacteriome at the SDP level. a Core genome-based maximum-likelihood phylogeny for *Gilliamella* and *Snodgrassella*. Gil-Ac means *A. cerana*-specific *Gilliamella* SDP (i.e., the *Gilliamella* SDP that is native to *A. cerana*); Am and Bt mean *A. mellifera* and *B. terrestris*, respectively; Sno means *Snodgrassella*. **b** Principal component analysis (PCA) of gene content of *Gilliamella* and *Snodgrassella* SDPs. The gene content is inferred from the MAGs of the native SDP for each sample. **c** Gut community composition across samples at the SDP level. The shape and color of individual samples in (**a**, **b**, **c**) follow the scheme in the bottom box.

obviously stemmed from the clustering of large number of (pro)phages from the same bee species, even though the phages from other bee species also appeared in the same clade (Supplementary Fig. 6). Only two VCs from *Gilliamella* and one VC from *Snodgrassella* had their phylogenetic tree perfectly separated into three clades according to their bee and bacterial hosts (Fig. 4b; Supplementary Fig. 6). When we assigned the phages into a finer taxonomic level known as vOTU, which took 90% gANI as threshold, >80% of vOTUs had a single bee superhost, regardless of the number of members the vOTUs possessed (Fig. 4c). While this finding indicated host specificity at the vOTU level, it also explained the aforementioned significant cophylogenetic signals for the gut phages.

For phage members of the same vOTU, those originating from the same bee species showed slightly lower genetic distance than that from different bee species (Fig. 4d). This observation indicated that segregation of superhost blocked the transmission of phage across bee species to a minor extent. The impact of superhost segregation on the gene content variation of phages was also small. Considering the trend that less gene content was shared between phage members with decreasing genetic similarity, the slope of the linear fit was slightly steeper for phages within the same bee species in comparison to those between different bee species (Fig. 4e).

Genetic disparity and functional collaboration in the tripartite system

The cophylogeny observed among bee, the gut bacteria *Gilliamella* and *Snodgrassella*, as well as their respective (pro)phages VC59, VC413 and VC44 strongly suggested a codiversification scenario. The ancestral sequences and gene content state of the three parties were reconstructed at three stages during the bee's evolution, i.e., genus, species and isolate. At the isolate stage, it became apparent that the synonymous substitution rate (Ks) of *Gilliamella*, *Snodgrassella* and the three VCs was higher than that of bee by one or two orders of magnitude (Fig. 5). In other word, the bacteria and phages had diverged to the species or even genus level at this timepoint. Given that Ks is almost free from selection pressure, the different Ks values between the three parties may reflect their different generation time. At the genus stage, however, the Ks values for the three parties were somewhat similar, indicating that Ks had approached saturation.

The non-synonymous substitution rate (Ka) increased in similar pace with Ks (Fig. 5). Ka/Ks is a common indicator of selective pressures on coding genes. At the isolate stage, Ka/Ks for bee floated around 0.35, but fell down to nearly 0.1 at the species and genus stage. This suggested that bee underwent the purification of missense mutations during isolate divergence and finished this process at the species stage. In contrast, Ka/Ks for gut bacteria differed relatively little at the three stages, indicating their purification process has been rapidly completed at the bee's isolate stage. In terms of gene content, bee varied with less than 15% between the compared genera, whereas bacteria began to exceed 15% at the bee's isolate stage. Phages exhibited even higher Ka, Ka/Ks and gene content difference than their bacterial host, which might be explained by the weakness of their DNA repair system, a lack of functional constraint and/or the influence of positive selection pressure. It is noteworthy, however, that the displayed evolutionary pattern of phages came from only three VCs and may not well represent the entire gut phageomes of bees or other animals.

The functional variation at the three evolutionary stages was evaluated by taking two categories of genes for enrichment analysis: genes under positive selection and genes involving gain or loss during evolution. The former genes generated more enriched functions than the latter, indicating that positive selection may represent a more common approach for elevating fitness for the whole holobiont. Many of the enriched functions were

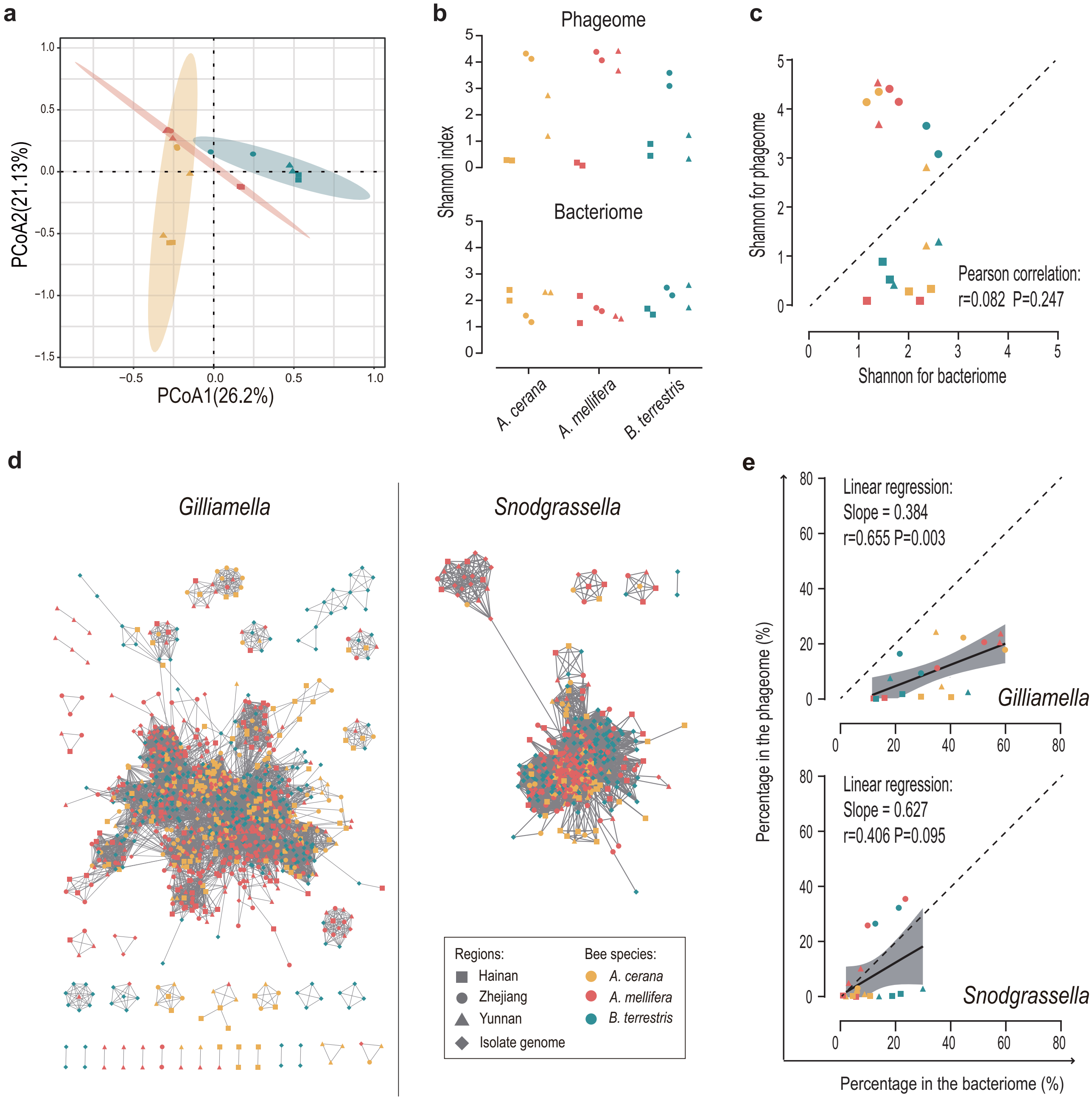

Fig. 3 | Gut phageome and its quantitative relationship with gut bacteriome. **a** PCoA plot based on the Bray-Curtis dissimilarity of phageome between samples. **b** Compositional evenness of phageome and bacteriome as measured by Shannon index for each sample. **c** Pearson correlation of Shannon index between phageome and bacteriome. **d** The vConTACT-based network diagram, in which each node represents a (pro)phage sequence from either VLP metagenome or bacterial genomes. Contigs that have significantly similar sequences are connected to each other. **e** Relative abundance of the bacteria (the percentage of the bacteria in the entire bacteriome) and their phages (the percentage of the phages that take the bacteria as the host in the entire phageome). The solid line and the gray area represent the line fit and the 95% confidence interval. The shape and color of individual samples in all panels follow the scheme in the bottom middle box.

related with degradation and biosynthesis of fatty acid, amino acid, and vitamin, several of which were shared by bee and *Gilliamella* (Fig. 6).

Since only three VCs had their members perfectly clustered according to the bee source, their small number of genes did not permit enrichment analysis. Instead, we opted to analyze the enriched functions of all phage genes while considering the genome of bacterial host as background. In the case of *Gilliamella*, its (pro)phages enriched the coding potential to influence the bacteria in metabolism of amino acid, the main ingredient of pollen (Supplementary Fig. 7). This implies that the (pro)phages may also aid in food digestion and nutrient utilization. On the other hand, much fewer biological processes were enriched from the VCs of *Snodgrassella*, possibly due to their much smaller gene repertoire.

Discussion

In this study, we employed bee model to investigate the cophylogeny of "animal-gut bacteria-phage" tripartite system. As consistent with previous studies[20], our study revealed that the gut bacterial community had a highly stable phylotype composition across bee species and even genus whereas the phylotypes further split into SDPs that aligned with the bee hosts. Taking *Gilliamella* for example, the 16S rRNA sequence similarity was less than

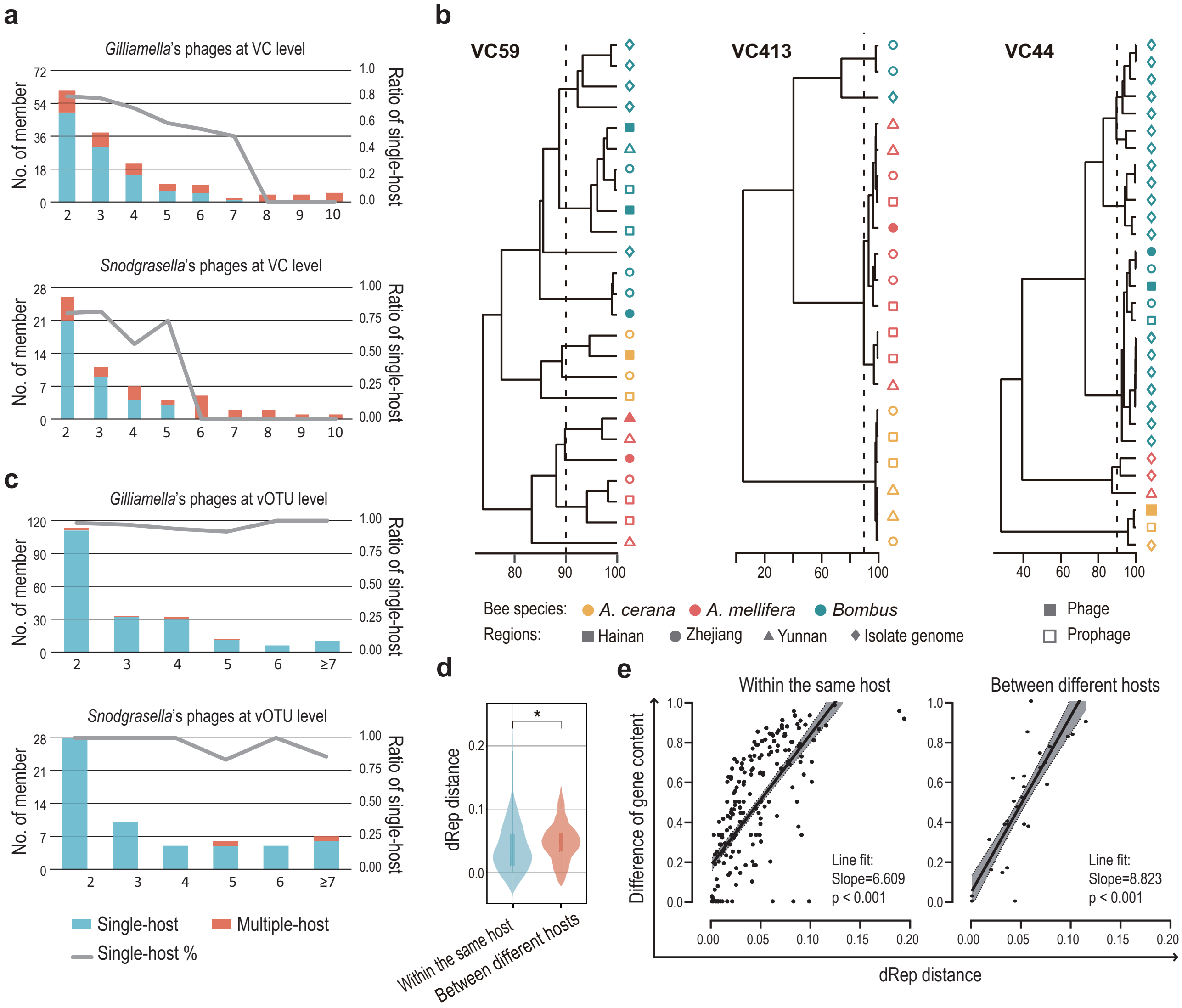

Fig. 4 | Host specificity and cophylogeny of phages in the bee gut. a Histogram showing the number of VCs (left y-axis) derived from a single bee species or from multiple bee species at different VC size (x-axis). The VC size refer to the total number of phage and prophage members contained in the VC. The gray line using the right y-axis shows the percentage of VCs derived from a single bee species. **b** Phylogeny of the three VCs that shows perfect cophylogeny with the bee and bacterial hosts. VC59 and VC413 are from *Gilliamella*, and VC413 is from *Snodgrassella*. The neighbor-joining tree is based on dRep distance, with the *x*-axis representing gANI. **c** Histogram showing the number of vOTUs (left *y*-axis) derived from a single bee species or from multiple bee species. **d** The genetic distance (by dRep software) for phages within the same bee species and for phages between different bee species. The box-and-whisker plots within the violin plots indicate the median and the quartiles. **p* < 0.05 by two-sided Student's *t*-test. **e** The relationship between the genetic distance and the variation of gene content for the gut phages. The solid line and the gray area represent the line fit and the 95% confidence interval.

97% between SDPs, and the gANI values were even below 83%. These values exceeded the common cut-off for species in modern bacterial taxonomy and are even comparable to inter-genus distance, e.g., between *Escherichia coli* and *Salmonella enterica*[26]. The substantial genetic divergence between the SDPs explains why recombination occur exclusively within SDPs instead of between SDPs. Consequently, the coexistence of SDPs within an individual bee or colony, which results possibly from the secondary host sympatry of *A. cerana* and *A. mellifera*, would not reverse the process of host-bacteria codiversification.

Cophylogeny can be driven by a variety of ecological and/or evolutionary processes. From the view of functional adaptation, two prerequisites are required for formation or subsequent maintenance of cophylogeny: (1) strong functional dependencies between symbiotic parties and (2) faithful transmission of the functional dependencies across generations (e.g., vertical transmission)"[4]. At both the phylotype and SDP level, the gut bacterial community may encompass vast genes involving metabolism of sugar, amino acid and vitamin, which may compensate for digestion and nutrient availability that bees may not manage on their own. Nevertheless, it is worth noting that reciprocal functional dependencies between bee and the gut bacteria and the resulting competitiveness of native SDP over non-native SDP is not so strong to produce absolute host fidelity. This is supported by that fact that a minor proportion of non-native SDP is occasionally present in the bee gut as found by both field collection and co-inoculation experiments in the lab[17,27]. The existence of non-native SDPs suggests their dispersal mediated by social contact among bee colonies, e.g., through shared floral resources or hive robbing by sympatric bee species. While most of the laterally transmitted SDPs would be eliminated from the non-native bee host soon due to their poorer fitness, a few of them are likely fixed and inherited in a long term, leading to imperfect phylogenetic clustering by bee hosts[17].

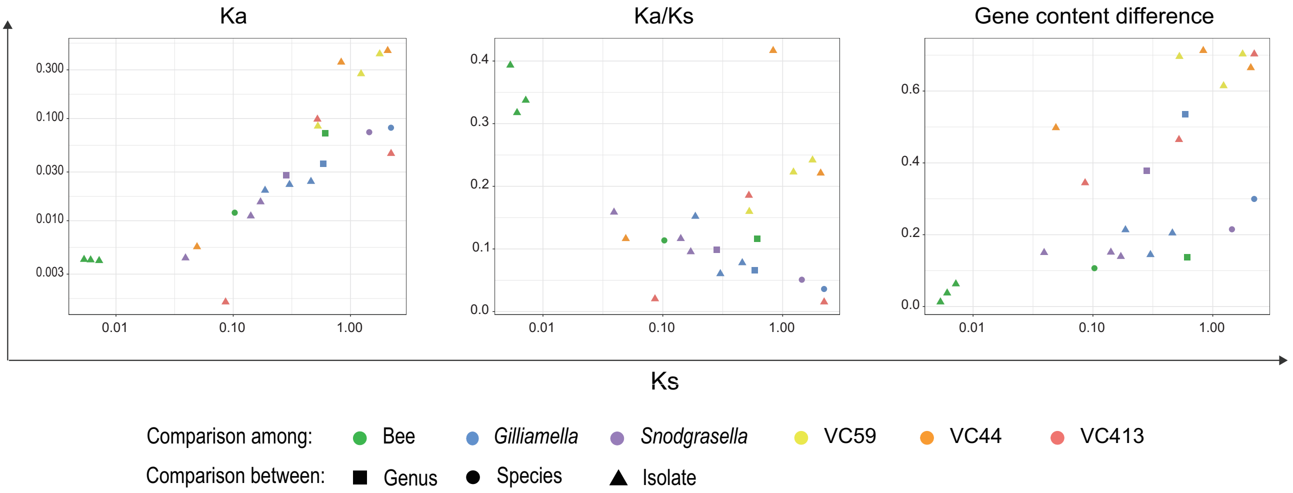

Fig. 5 | Evolutionary rates of bee, gut bacteria and phages. For bee, comparison between 'Genus' refers to comparison between the ancestor of *Apis* and the ancestor of *B. terrestris*; comparison between 'Species' refers to comparison between the ancestor of *A. cerana* and the ancestor of *A. mellifera*. The three dots for the 'Isolate' stages represent comparison between *A. cerana* samples, between *A. mellifera* samples, and between *B. terrestris* samples, respectively. The three stages for the gut bacteria and phages in this figure also refer to the bee's stages instead of their own. For the three phage VCs, the genetic differences at the 'Genus' and 'Species' stage are too huge to be precisely anticipated and plotted. The values of synonymous substitution rate (Ks), nonsynonymous substitution rate (Ka), and Ka/Ks in this figure are not for individual genes but the average for all conserved orthologous genes. Gene content difference refers to the number of unshared genes divided to the number of total orthologous genes involved in the comparison.

This study performs a comprehensive investigation into the host specificity of bee's gut phages. Most of the bee phages share no sequence similarity with phages found in other animals. Currently, there are only a few literatures working on the gut phageomes of insects, including honey bee, *Melipona quadrifasciata* (a stingless bee), and mosquito[21–23,28,29]. These literatures all reported that the gut phages from insects had little overlap with the reference phage sequences. The distinct gene repertoire likely arises from niche specialization of the insect's gut bacteria. Notably, the level of typing resolution shows critical importance on host specificity of both gut bacteria and phages. For bacteria, phylotypes are stable across bee species while SDPs are host specific. For phages, most of them are host generalist at the VC level but host specialist at the vOTU level. Actually, both phages and bacteria in the gut may repeatedly generate mutations, disperse and infect new host, adapt to and propagate in the new host, gradually evolve into new species, and mutate for host switch again. With a higher mutation rate, phages are better at adapting to new environment, namely they have a much higher frequency of successful host switch into new bacterial hosts (at the SDP level) compared to the bacterial SDPs switching into their non-native bee host, which eventually leads to the phages' poorer host specificity. Only three VCs were identified to show perfect cophylogeny with host, indicating their potentially important contribution to the fitness of bees or gut bacteria. Meanwhile, their perfect cophylogeny also point to another possibility that these VCs are defective phages that can only replicate in the presence of helper virus. Since long-term stay within the bacterial genome attempts to synchronize the evolutionary pace of the phages and bacteria, the real evolutionary rate of phages at the lytic state is probably faster than that inferred in this study.

We measured the quantitative relationship between bacteriome and phageome in bee. By the Mantel test, we observed that the bacteriome's composition at both phylotype and SDP level was associated with the phageome's composition. There are mainly two models describing the virus-host dynamics[30]. The "Kill-the-Winner" represents the classical predator-prey model, in which phages predominantly prey on fast-growing bacteria. In contrast, the "piggyback-the-winner" model suggests that, as bacterial host become more abundant, phages forego rapid replication and opt instead to integrate themselves into their host, thereby reducing their numbers. Our findings indicate that *Gilliamella* phages replicate more slowly than *Gilliamella* itself, aligning with the "piggyback-the-winner" model. The enriched pathways for *Gilliamella* phages involve some metabolic functions, implying that they can modulate the metabolic state rather than hijack the bacterial cell. Unlike *Gilliamella* that is located on top of the mucus layer, *Snodgrassella* grows in contact with the ileum epithelium and does not directly interact with food[31]. This may explain why *Snodgrassella* and its phages possess fewer auxiliary genes with metabolic functions. However, due to the phageome's greater compositional variability and the absence of core phage members, it is important to note that the impact of phage on bee's health is probably more indirect and less significant than that of bacteria.

Within the "animal-gut bacteria-phage" tripartite system, the evolutionary rates displayed by the bee, the bacteria and the phages followed basically the same pattern, e.g., Ka, Ks and gene content variation accumulate during evolution while Ka/Ks declines as a result of purification of harmful mutations. Nevertheless, the exact values from the three biological entities differed greatly. Probably, this disparity stems from their inherent biological characteristics such as generation time and error-rate for DNA replication, which depend more on the taxonomic attribute than on the interactions between the three biological entities. In other words, the different evolutionary rates would also be observed between totally unrelated animals, bacteria and phages. To reconcile the different evolutionary rates within a frame of coevolution, selection on the whole holobiont, instead of on the individual parties separately, is required[32]. For the bee model in this study, one of the selections can be food digestion and nutrient utilization since the functional variation for bee, bacteria and phage all center around metabolism of sugar, amino acid and vitamin. However, such functional codependency tends to be not strong enough, and the absolutely faithful transmission is also difficult to achieve in the gut ecosystem. We therefore postulate that rigorous cophylogeny of "animal-gut bacteria-phage" tripartite system cannot widely apply to most of gut microbes. In this context, the very few bacteria and phages that show the cophylogenetic signal are of particular research value due to their potential benefits to host's fitness.

Overall, this study integrates phage into the analysis of cophylogeny in gut microbiome and therefore represents a significant advancement toward understanding the coevolutionary history of the "animal-gut bacteria-phage" system. While only a few bacteria and phages have been identified as forming a chained cophylogeny with bees, there exist notable differences in their evolutionary rates between the three parities. Nevertheless, their

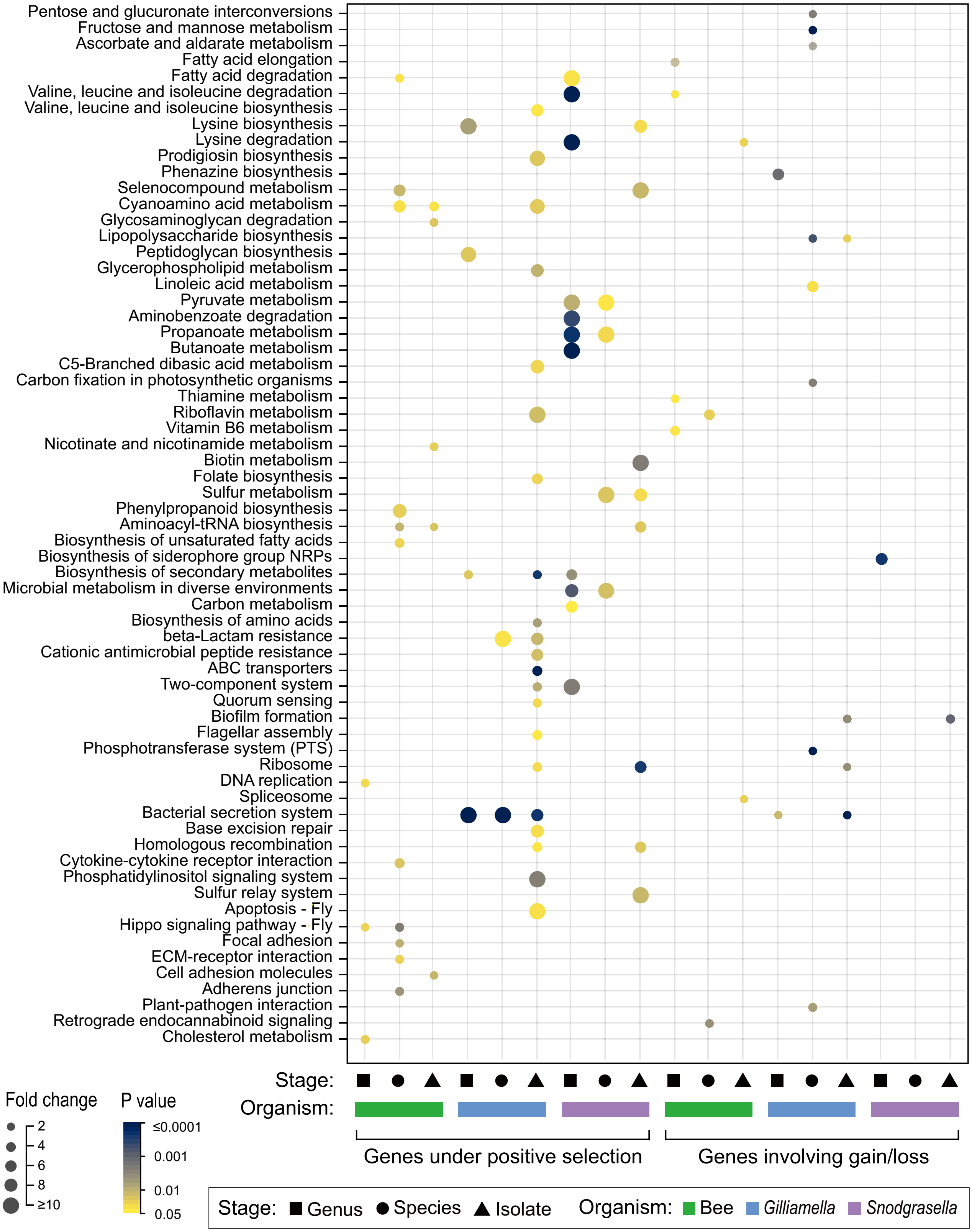

Fig. 6 | Functional codependency in the "bee-gut bacteria-phage" system. The functional enrichment analysis is performed on genes generated on three dimensions: gene categories, including the genes under positive selection and the genes involving gain or loss between clades; organisms, including bee, *Gilliamella* and *Snodgrassella*; stages, including comparison between genus, between species, and between isolates. The functional annotation refers to KEGG. The fold change refers GeneRatio divided by BgRatio, which is calculated by the ClusterProfiler package.

functional variations are simultaneously associated with food digestion and nutrient utilization, highlighting the tripartite coevolution that results in an enhanced ecological fitness for the whole holobiont. To gain a deeper understanding of how gut microbes displaying cophylogenetic signals benefit their bee hosts, further in-depth and mechanistic studies in controlled laboratory conditions are necessary. While our bee model-based study provides a framework for future coevolution studies, it is imperative to expedite the exploration of the "animal-gut bacteria-phage" tripartite system in humans and mice. Despite technical obstacles caused by the complicated gut microbiota in mammals, this will provide insights into the significance of cophylogeny in humans and offer a novel perspective on the relevance of the gut ecosystem to human health.

Methods

Sample collection

A. cerana and *A. mellifera* workers were collected in Zhejiang (30.308 N, 120.092E), Yunnan (25.135 N, 102.756E) and Hainan (19.825 N, 109.695E) provinces in China, respectively. At each region, *A. cerana* and *A. mellifera* colonies have been reared in the same apiaries for years. Meanwhile, commercialized *B. terrestris* colonies (Biobest, China) were reared in the same apiaries for a duration of two months to allow potential transmission of gut bacteria and phages.

Bee genome sequencing and bioinformatic analyses

The genomic DNA was extracted from bee thoraces using Magnetic universal genomic DNA kit (TianGen Biotech, China). Genomic sequencing libraries were generated using NEB Next Ultra DNA Library Prep Kit for Illumina (New England Biolabs, USA), and were then sequenced on the Illumina Nova-seq 6000 platform with 150-bp paired-end reads.

The reads were de novo assembled into contigs by using MegaHit v1.2.9 with default settings[33]. Genes in the bee genomes were predicted by using GeneID v1.4.4[34]. For construction of orthologous gene relationship, a reference gene pool was built. All complete and draft genomes of *A. cerana*, *A. mellifera* and *B. terrestris* were downloaded from NCBI Genbank database. Orthologous gene families of their genes were built using OrthoFinder v2.5.2[35]. The 0/1 matrix indicating gene presence/absence was subject to principal components analysis (PCA) by R software v4.0.3.

Single-copy conserved genes were considered as core gene families and were used for inferring core genome-based phylogeny. The coding sequences were aligned at the protein level with Mafft v7.471[36], and back-translated to nucleotides. The resulting concatenated alignments was processed by IQtree v2.2.0 for construction of maximum likelihood (ML) tree with GTR + F + I model[37]. The evolutionary time trees comprising the three bee species was meanwhile inferred by Timetree database v5[24].

Gut bacteriome sequencing and bioinformatic analysis

For *A. cerana* and *A. mellifera*, approximately 100 foragers were collected at entrances for each colony. For *B. terrestris*, 30 workers were collected in the nest. After anesthetization of the bees by CO_2, guts were pulled out, pooled, and homogenized with a bead-beater. The homogenates were firstly centrifuged at $500 \times g$ for 5 min to remove debris and then centrifuged at $5000 \times g$ for 5 min to pellet the bacterial cells. Bacterial DNA was extracted from the bacterial pellets using a CTAB-based DNA extraction protocol[18].

Sequencing of the gut bacteriome followed the same protocol as the bee genome sequencing. To infer de novo metagenome assemblies, paired-end reads were firstly mapped against the reference bee genomes using Bowtie v2.4.1[38]. The GenBank accession number of reference genomes for *A. cerana*, *A. mellifera* and *B. terrestris* are GCF_001442555.1 (ACSNU-2.0), GCA_003254395.2 (Amel_HAv3.1) and GCF_000214255.1 (Bter_1.0), respectively. After filtering off the host derived reads, the unmapped reads for each sample were assembled independently using MegaHit. Putative open reading frames (ORFs) were predicted with Prokka v1.14.6 using the metagenome flag[39]. The functional profile of the entire gut bacteriome was inferring by using METABOLIC v4.0[40], with the assembled contigs and raw reads being the input.

To obtain the taxonomic information at the phylotype level, the ORF sequences from all samples were pooled together and a non-redundant gene catalog was obtained by CD-HIT v4.7 with parameters "-c 0.95, -aS 0.9"[41]. DIAMOND v0.9 was used to search the non-redundant protein sequences against the NCBI microNR database v13[42]. The taxonomic information of each gene was determined using the lowest common ancestor-based algorithm (LCA) implemented in MEGAN v6.20[43]. Then each contig was assigned to a phylotype to which the largest number of ORFs in this contig was assigned. The abundance of each phylotype was measured by mapping reads against contig sequences by Bowtie and summing the read numbers of contigs assigned to the phylotype. Both the calculation of Shannon index and the PCoA analysis, which were based on the phylotype abundance, was calculated using R vegan package v2.6.2[44].

The taxonomic assignment at the SDP level following the methods from Ellegaard and Engel[16] with small modifications. All available published genomes of *Gilliamella* and *Snodgrassella* strains were downloaded from NCBI Genbank database (Supplementary Table 2). The orthologous relationship was constructed by using OrthoFinder, and the ML phylogenetic tree based on concatenated conserved genes was made following the same methods for bee as above. Based on the tree phylogeny, the isolates were clearly assigned into *A. cerana*, *A. mellifera* and *Bombus* clades. When all genes from the isolate genomes made a reference gene database, query ORFs from samples were searched against the database with NCBI Blastn v2.11.0 with a minimum percentage identity of 80% and a minimum alignment length of 150 bp. According to the best hit, each ORF was assigned to one of *A. cerana*, *A. mellifera* and *Bombus* clade. Then each contig was assigned to a host-specific SDP to which the largest number of ORFs in this contig was assigned; and all ORFs in the contig were re-assigned to the same SDP as their contig was. For each sample, therefore, all of the contigs corresponding to the sample's bee source constituted the MAGs of its native SDP, whereas the contigs not corresponding to the sample's bee source constituted the MAGs of non-native SDP. Last, reads for each sample were mapped to the MAGs for abundance estimation of native and non-native SDPs. For each sample's native SDP, its 0/1 gene matrix indicating gene presence/absence in the MAG was subject to PCA by R. This 0/1 matrix was also used for core-pan genome analysis by using PanGP v1.0.1[45]. The genes conserved among all samples' native SDP were extracted for core genome phylogeny following the same method for bee as above. Pairwise ANI values between the MAGs of all samples' native SDP were calculated using FastANI v1.33[26].

Three approaches were adopted for assessing the recombination in bacterial genomes. First, for each of the single-copy conserved genes, the gene's ML tree was constructed by IQtree and then was compared with the aforementioned core genome-based ML tree by using the treedist program in Philip v3.69[46], with the branch score distance method. The larger the distance, the more likely the gene was involved in recombination. Second, still for each of the conserved genes, the PhiPack software was run to produce a *p*-value of recombination[47], with p-value less than 0.05 being considered to involve recombination. Third, recombinant segments within the concatenated alignment of core genes were detected by using FastGear[48]. The extent of recombination in bee genomes were also assessed using the first and the second approach.

Gut phageome sequencing and bioinformatic analyses

Extraction of VLPs followed the method by Bonilla-Rosso et al.[23] and He et al.[49]. Briefly, the supernatant obtained from centrifuging the homogenates of gut tissues (as aforementioned in the section "Gut bacteriome sequencing and bioinformatic analysis") was filtered through 0.22-μm-pore-sized filters (BIOFIL, China), followed by PEG6000 (GENERAYBIO, China) flocculation and centrifugation at $25,000 \times g$ for 2 h to concentrate VLP. The VLP solution was further incubated with DNase I (Takara, China) to remove free DNA. Next, the VLPs were extracted with TE buffer, lysed by SDS and proteinase K (Vazyme, China), and incubated with CTAB/NaCl. Finally, the viral DNA was extracted with phenol/chloroform/isoamyl alcohol protocol.

To assess the quality of enrichment of the viromes, the raw fastq data were processed with the ViromeQC v1.0[50]. Genome sequencing and de novo assembly followed the same method as the bacteriome sequencing. The phage sequences were identified using VIRSorter v2.2.2 from the derived contigs[51]. Meanwhile, putative prophage sequences were identified from the isolate genomes of *Gilliamella* and *Snodgrassella* downloaded from the public database as well as the MAGs of bacteriome derived from this study. The phage and prophage sequences were pooled and processed with vContACT v2, which generated probabilistic VCs with high similarity[52]. The reference viral sequences organized by the ICTV bacterial viruses subcommittee were also added for the vContACT clustering analysis[53]. Taxonomic annotation of VC was determined based on the vContACT results: if a viral sequence was within the same VC as a reference genome, then the viral sequence was considered to belong to the same genus as the reference genome does. The vContACT clustering results were visualized by Cytoscape v3.7.1[54].

The abundance of phage members in each sample was measured by the number of reads mapped to the phage sequences by Bowtie. The abundance of the phage members of the same VC were summed to obtain the VC abundance in each sample. To assess the quantitative relationship between baceriome and phageome, the Mantel test was used to compare the Bray-Curtis dissimilarity for the bacteriome and for the phageome by R ape package v5.6[55], for which the bacteriome referred to the phylotype level and the phageome referred to the VC level. Linear fit of Bray-Curtis dissimilarity between bacteriome and phageome was run by using the lm function in R. The hypothesis of cophylogeny between bee and bacteria and between bacteria and phages was tested using the Parafit function in R ape package[56].

Host prediction for phages was based on two strategies. First, CRISPR spacers were predicted from the isolate genomes and MAGs using CRISPR Recognition Tool v1.2[57]. This CRISPR-spacer collection was subsequently searched against the phage sequences using the blastn software with the additional settings -ungapped and -perc_identity 100. Blast hits indicated the host relationship. Second, phage members within the same VC by vCONTACT2 analysis were considered to have the same host phylotype. The quantitative relationship between bacteriome at the SDP level and phageome was also determined: the Bray–Curtis dissimilarity for the bacteriome was calculated based on the SDP abundance, and the Bray-Curtis dissimilarity for the phageome was calculated for the phylotype-specific VCs.

At a finer level of taxonomy, the phages and prophages were also assigned to vOTU by using dRep v3.4.1[58], with the default 90% of ANI as the threshold. The genetic difference between phage members was defined as 100% minus ANI calculated by dRep. The gene content difference between phage members was defined as the number of unshared protein clusters divided by total non-redundant protein clusters of the two compared phages. The information of protein clusters was generated from the vContACT analysis as described above.

Tripartite system analysis

The sequences of nodes that represented ancestral *A. cerana*, *A. mellifera* and *Apis* were inferred by IQtree, with the aligned conserved genes and the corresponding ML tree as the input file and -asr as the input parameter. The ancestral states of gene presence/absence that reflected gene gain and loss during divergence were determined by Dollo parsimony algorithm in Count v9.1106[59].

Ka, Ks and Ka/Ks were calculated by the codeml program in PAML v4.9[60]. In detail, NSsites was specified to be '0 1 2 7 8'. In these models, Model 0 was used to calculate one Ka/Ks ratio for all branches; comparison between M1a and M2a and between M7 and M8 was performed to identify positively selected sites through Bayes Empirical Bayes analysis. Genes with Ka/Ks >1 or containing positively selected sites were regarded as genes under positive selection.

Function annotation was performed by using the eggNOG-mapper server[61], which included annotation by Gene Ontology (GO) and KEGG database. Functional enrichment analysis was performed with the clusterProfiler package v4.0 in R. The enriched GO terms were visualized by REVIGO online service in the Treemap form[62].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw reads were deposited in NCBI's Sequence Read Archive (SRA) database under project accession no. PRJNA1026499, PRJNA1027301, PRJNA1027315, PRJNA1027983, PRJNA1027985, PRJNA1028003, PRJNA1029451, PRJNA1029457, PRJNA1029459. The detailed accession numbers for each dataset are listed in Supplementary Table 3. The source data in this study are available at https://doi.org/10.5281/zenodo.12673332.

Code availability

The scripts used in this study are available at https://doi.org/10.5281/zenodo.12673332.

Received: 23 February 2024; Accepted: 20 August 2024;
Published online: 27 August 2024

References

1. Belkaid, Y. & Harrison, O. J. Homeostatic immunity and the microbiota. *Immunity* **46**, 562–576 (2017).
2. McFall-Ngai, M. et al. Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. USA* **110**, 3229–3236 (2013).
3. Blasco-Costa, I., Hayward, A., Poulin, R. & Balbuena, J. A. Next-generation cophylogeny: unravelling eco-evolutionary processes. *Trends Ecol. Evol.* **36**, 907–918 (2021).
4. Groussin, M., Mazel, F. & Alm, E. J. Co-evolution and co-speciation of host-gut bacteria systems. *Cell Host Microbe* **28**, 12–22 (2020).
5. Groussin, M. et al. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **8**, 14319 (2017).
6. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109.e1099 (2021).
7. Simmonds, P. et al. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).
8. Yang, K. et al. Alterations in the gut virome in obesity and type 2 diabetes mellitus. *Gastroenterology* **161**, 1257–1269.e1213 (2021).
9. Norman, J. M. et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).
10. Gogokhia, L. & Round, J. L. Immune-bacteriophage interactions in inflammatory bowel diseases. *Curr. Opin. Virol.* **49**, 30–35 (2021).
11. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
12. Gogarten, J. F. et al. Primate phageomes are structured by superhost phylogeny and environment. *Proc. Natl. Acad. Sci. USA* **118**, e2013535118 (2021).
13. Engel, M. S. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; Apis). *J. Hymenopt. Res.* **8**, 165–196 (1999).
14. Kwong, W. K. & Moran, N. A. Gut microbial communities of social bees. *Nat. Rev. Microbiol.* **14**, 374–384 (2016).
15. Meziti, A. et al. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. *ISME J.* **13**, 767–779 (2019).
16. Ellegaard, K. M. & Engel, P. Genomic diversity landscape of the honey bee gut microbiota. *Nat. Commun.* **10**, 446 (2019).
17. Ellegaard, K. M., Suenami, S., Miyazaki, R. & Engel, P. Vast differences in strain-level diversity in the gut microbiota of two closely related honey bee species. *Curr. Biol.* **30**, 2520–2531.e2527 (2020).
18. Wu, Y. et al. Genetic divergence and functional convergence of gut bacteria between the Eastern honey bee Apis cerana and the Western honey bee Apis mellifera. *J. Adv. Res.* **37**, 19–31 (2022).
19. Powell, J. E., Leonard, S. P., Kwong, W. K., Engel, P. & Moran, N. A. Genome-wide screen identifies host colonization determinants in a

bacterial gut symbiont. *Proc. Natl. Acad. Sci. USA* **113**, 13887–13892 (2016).

20. Wu, J. et al. Honey bee genetics shape the strain-level structure of gut microbiota in social transmission. *Microbiome* **9**, 225 (2021).
21. Deboutte, W. et al. Honey-bee-associated prokaryotic viral communities reveal wide viral diversity and a profound metabolic coding potential. *Proc. Natl. Acad. Sci. USA* **117**, 10511–10519 (2020).
22. Busby, T. J., Miller, C. R., Moran, N. A. & Van Leuven, J. T. Global composition of the bacteriophage community in honey bees. *mSystems* **7**, e0119521 (2022).
23. Bonilla-Rosso, G., Steiner, T., Wichmann, F., Bexkens, E. & Engel, P. Honey bees harbor a diverse gut virome engaging in nested strain-level interactions with the microbiota. *Proc. Natl. Acad. Sci. USA* **117**, 7355–7362 (2020).
24. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
25. Almeida, E. A. B. et al. The evolutionary history of bees in time and space. *Curr. Biol.* **33**, 3409–3422.e3406 (2023).
26. Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
27. Kwong, W. K. et al. Dynamic microbiome evolution in social bees. *Sci. Adv.* **3**, e1600513 (2017).
28. Shi, C. et al. Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. *Microbiome* **7**, 121 (2019).
29. Caesar, L. & Haag, K. L. Tailed bacteriophages (Caudoviricetes) dominate the microbiome of a diseased stingless bee. *Genet Mol. Biol.* **46**, e20230120 (2024).
30. Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
31. Kwong, W. K. & Moran, N. A. Cultivation and characterization of the gut symbionts of honey bees and bumble bees: description of Snodgrassella alvi gen. nov., sp. nov., a member of the family Neisseriaceae of the Betaproteobacteria, and Gilliamella apicola gen. nov., sp. nov., a member of Orbaceae fam. nov., Orbales ord. nov., a sister taxon to the order 'Enterobacteriales' of the Gammaproteobacteria. *Int. J. Syst. Evol. Microbiol* **63**, 2008–2018 (2013).
32. Rosenberg, E. & Zilber-Rosenberg, I. The hologenome concept of evolution after 10 years. *Microbiome* **6**, 78 (2018).
33. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
34. Parra, G., Blanco, E. & Guigó, R. Geneid in drosophila. *Genome Res.* **10**, 511–515 (2000).
35. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
36. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47**, W5–W10 (2019).
37. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
39. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
40. Zhou, Z. et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**, 33 (2022).
41. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
42. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
43. Gautam, A., Felderhoff, H., Bağci, C. & Huson, D. H. Using AnnoTree to get more assignments, faster, in DIAMOND+MEGAN microbiome analysis. *mSystems* **7**, e0140821 (2022).
44. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
45. Zhao, Y. et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* **30**, 1297–1299 (2014).
46. Felsenstein, J. PHYLIP—phylogeny inference package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
47. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
48. Mostowy, R. et al. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol. Biol. Evol.* **34**, 1167–1182 (2017).
49. He, T. et al. Environmental viromes reveal the global distribution signatures of deep-sea DNA viruses. *J. Adv. Res.* **57**, 107–117 (2023).
50. Zolfo, M. et al. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).
51. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
52. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
53. Turner, D. et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* **168**, 74 (2023).
54. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
55. Paradis, E. & Schliep, K. P. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018). **3**.
56. Legendre, P., Desdevises, Y. & Bazin, E. A statistical test for host-parasite coevolution. *Syst. Biol.* **51**, 217–234 (2002).
57. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinforma.* **8**, 209 (2007).
58. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
59. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
60. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
61. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
62. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant number 32072798), Science and Technology Department of Zhejiang Province, China (grant number 2021C02068-8); China Agriculture Research System of MOF and MARA (grant number CARS-44).

Author contributions

Y.F. and H. Zheng conceived the study. N.Z., G.C., G.Z., K.D., Y.Z., and H. Zhao collected samples and generated metagenome data. Y.F., R.W., Q.C., T.S., and Z.W. performed data analysis. Y.F. drafted the manuscript. Y.F., Y.C., F.H., and H. Zheng revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-024-00557-x.

Correspondence and requests for materials should be addressed to Ye Feng or Huoqing Zheng.

Reprints and permissions information is available at http://www.nature.com/reprints