Symbiont loss and gain, rather than co-diversification shapes honeybee gut microbiota diversity and function

Aiswarya Prasad[1], Asha D Pallujam[2], Rajath Siddaganga[3], Ashwin Suryanarayanan[3], Florent Mazel[1], Axel Brockmann[3], Sze H Yek[2,4] and Philipp Engel[1,*]

[1]Department of Fundamental Microbiology, University of Lausanne, Switzerland

[2]School of Science, Monash University Malaysia, Malaysia

[3]National Centre for Biological Sciences-Tata Institute of Fundamental Research, Bengaluru, Karnataka, India

[4]Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Malaysia

[*]To whom correspondence should be addressed

e-mail: philipp.engel@unil.ch

Keywords

Apis, metagenomics, host specificity, co-diversification, microorganisms, symbiosis

Abstract

Studying gut microbiota evolution across animals is crucial for understanding symbiotic interactions but is hampered by the lack of high-resolution genomic data. Honeybees, with their specialized gut microbiota and well-known ecology, offer an ideal system to study this evolution. Using shotgun metagenomics on 200 honeybee workers from five species, we recovered thousands of metagenome-assembled genomes, identifying several novel bacterial species. While microbial communities were mostly host-specific, we found both specialist and generalist bacteria, even among closely related species, with notable variation between host species. Some generalists emerged host-specific only at the strain level, suggesting recent host switches. Unexpectedly, we found little evidence of co-diversification between hosts and symbionts. Instead, symbiont gains, losses, and replacements led to functional differences, such as the ability to degrade pollen-derived pectin. Our results provide new insights into gut microbiota evolution and uncover the functional potential of the previously underexplored gut microbiota of these important pollinators.

Introduction

Animals harbor specialized microbial communities in their gut, which can influence health and disease susceptibility through diverse functional capabilities[1,2]. These communities are often host-specific, where the gut microbiota of individuals of the same host species are more similar in composition to each other than to those of different host species[3]. In some animal groups, it has also been observed that more closely related host species harbor more similar gut microbiota in composition, a pattern referred to as phylosymbiosis[4]. This host specificity at the community level emerges from the host-restricted distribution of individual community members. Dispersal limitation and host-filtering processes have been suggested to contribute to the host-specific distribution of community members in the animal gut[3,5].

Some gut bacterial clades (*e.g.*, in termites[6] and hominids[7–11]) are host specific and their phylogenies are congruent with that of their hosts. However, only a small fraction of the analyzed bacteria showed signs of co-diversification (a pattern different from phylosymbiosis[12]) in these studies, especially when considering the evolutionary timescales of the host and microbes. On the other hand, several examples of generalist gut symbionts exist[13,14]. This variation in the distribution patterns of different gut bacteria suggests they do not follow the same evolutionary trajectory across various animal species. Indeed, a recent study[15] showed that host specificity could vary across bacterial lineages in the animal gut (*i.e.*, there are generalists and specialists), and this may not be explained by microbial phenotype alone[16], with host ecology playing an important role[3]. This raises the question of why generalist and specialist gut bacteria exist and whether specialist bacteria primarily evolved *via* repeated switches between hosts or through co-diversification with hosts.

Our current understanding of the evolution and diversification of gut microbial communities across hosts is limited by at least four technical and conceptual factors. (1) Previous studies have focused on highly diverse and variable systems across host individuals[17], making it challenging to understand whether gut communities change by acquisition of new members or by *in situ* evolution. (2) Significant differences in the relatedness and ecology of the compared hosts hinder the identification of the drivers of community change over evolutionary time scales and the prediction of their functional consequences. (3) Samples from different animals often come from different geographic regions, making geography an important confounding factor that could explain the host-restricted distribution observed for some gut symbionts[9,18]. (4) Most comparative studies are based on 16S rRNA analysis, which lacks the resolution to determine the distribution of closely related bacterial species, strains and functional gene content across hosts[19].

Comparative studies of the microbiota of closely related animals with similar phenotypes and geographic distribution at high taxonomic resolution are needed to overcome these limitations. The gut microbiota of honeybee workers is a promising model offering the combined advantage of a simple yet specialized microbial community shared among related host species with similar ecology, social behavior, and overlapping geographic ranges[20–22]. Prominent honeybee gut microbiota members include *Lactobacillus*, *Bombilactobacillus*, *Bifidobacterium*, and several proteobacteria. These bacterial lineages from honeybees are usually monophyletic but belong to host-specific species, of which several can coexist simultaneously in the same host individual[23–25]. Thus, the predominant members of the honeybee gut microbiota have likely been acquired in a common ancestor of social bees and subsequently diversified in honeybees by separation into

different host species and distinct ecological niches in the gut[26–28]. Co-diversification has been suggested as an important mode of evolution[22,26,29–31], but formal tests of across all microbial community members have not yet been conducted[20]. So far, most comparative studies are based on either 16S rRNA gene amplicon sequencing with low genetic resolution or relatively few isolate genomes biased towards isolates from the managed Western honeybee *Apis mellifera*[26,29,32–37]. Both approaches provide limited quantitative insights into the distribution of divergent strains and species across hosts. Deep shotgun metagenomics allows the sequencing of all of thegenomes in a given sample, characterization of the distribution of community members at high resolution (down to the strain level) across hosts, and assessment of the functional consequences of differences among hosts. However, this approach has only been used to study the Western honeybee (*Apis mellifera*)[22,29,38,39] and the Asian honeybee (*Apis cerana*)[40,41].

In this study, we applied shotgun metagenomics to individual worker bees of the five most prevalent species of honeybees described so far, and spanning all three major clades of honeybees, namely Micrapis (dwarf honeybee): *A. florea* and *A. andreniformis*, Megapis (giant honeybee): *A. dorsata*, Apis: *A. mellifera* and *A. cerana*. While *A. mellifera* is globally distributed due to human introduction, the other honeybee species are mainly found across Asia[21]. All five honeybee species share several traits, such as a complex social lifestyle and a diet mainly consisting of processed pollen and nectar. However, they differ in certain key aspects, including body and colony size, nesting habit (open versus cavity nesting) and migrating ability[42].

From a total of 200 shotgun metagenomes from the gut of individual worker bees, sampled across Peninsular Malaysia and South India, we generated a comprehensive dataset of

metagenome-assembled genomes (MAGs), expanding the catalog of the bacterial genomes available from these important pollinators and model for microbiome research. We reveal striking differences in host specificity across host and bacterial species and identify previously unrecognized but prevalent members of the bee gut microbiota that expand the functional repertoire of these communities in specific host species. Overall, our results show that ancient gut microbiota-host interactions can be maintained in the absence of strict co-diversification.

Results

An expanded catalog of bacterial genomes from the gut of five honeybee species

We shotgun sequenced (35M ±19M Illumina paired-end reads per sample) the hindguts of 200 individual worker bees from 40 different colonies of five honeybee species sampled in South India and various locations across peninsular Malaysia **(Fig. 1A, Supplementary Table S1, Supplementary Fig. S1**).

We recovered a total of 1,959 high- and medium-quality metagenome-assembled genomes (MAGs, *i.e.* completeness score > 50% and contamination <5%) (**Supplementary Fig. S2**), including 156 from *A. florea*, 163 from *A. andreniformis*, 869 from *A. dorsata*, 324 from *A. mellifera* and 413 from *A. cerana*) (**Fig. 1B, Supplementary Table S2**). The MAGs were discriminated into 150 species spanning 23 genera (prevalence > 10%) using 95% ANI (average nucleotide identity) clustering (**Supplementary Fig. S3, Supplementary Fig. S4**). A majority of these species belonged to genera previously reported to be predominant in the gut microbiota of honeybees, including *Lactobacillus* (formerly Firm5), *Bombilactobacillus* (formerly Firm4), *Bifidobacterium*, *Gilliamella*, *Snodgrassella*, *Frischella*, *Bartonella* and *Apibacter* (constituted by 91 / 150 species clusters), but also lesser known genera such as *Dysgonomonas* (family Dysgonomonadaceae), *Saezia* (family Burkholderiaceae), *Pectinatus* (family Selenomonadaceae), *Entomomonas* (family Pseudomonadacea), *Pluralibacter* (family Enterobacteriaceae), WRHT01 (Desulfovibrionaceae family), CALYQQ01 (family Enterobacteriaceae) and CALYQJ01 (family Lactobacillaceae) (constituted by 21 / 150 species). Around 47% of the identified bacterial species (71 / 150 species) have not been cultured to date, with many belonging to the predominant genera (41 / 71 species) and with genomic information

currently lacking in public databases (**Fig. 1B**). The established collection of MAGs massively expands the catalog of bacterial genomes from the gut environment of these underexplored honeybee species.

Most honeybee species harbor host-specific microbial communities in their gut

We assessed the taxonomic diversity of the gut microbiota across the five honeybee species using our collection of MAGs. Between 7 - 56M reads from each sample mapped to these MAGs (**Supplementary Table S3**), providing sufficient depth of bacterial reads for MAG recovery and bacterial species detection from individual guts (**Supplementary Fig. S5**). Rarefaction curves show that our sampling effort from different colonies of each honeybee species (**Fig. 1A**) was sufficient to cover most of the species-level diversity in each honeybee species (**Supplementary Fig. S6**).

Prevalence of the identified genera varied greatly across host species (**Fig. 1C**). *Lactobacillus*, *Bombilactobacillus*, and *Bifidobacterium* were consistently present across all five honeybee species (**Fig. 1B**). All other genera were only found in a subset of the five host species or occurred sporadically throughout. Notably, samples of *A. dorsata* contained several additional genera in high prevalence, such as *Dysgonomas* (90.9%), *Pectinatus* (70.5%), and *Saezia* (90.9%), which were rarely found in the other host species.

Comparison of the species-level bacterial diversity revealed that *A. dorsata* contained the most diverse microbiota, followed by *A. mellifera*, *A. cerana* and then the two dwarf honeybees *A. florea* and *A. andreniformis* (**Fig. 1D**). This pattern holds when considering only species of the

shared genera (*i.e.*, genera with a prevalence of at least 10% in each host), except that *A. dorsata* and *A. mellifera* were not significantly different from each other (**Supplementary Fig S7**). Interestingly, the bacterial biomass per individual gut was the highest in *A. mellifera*, followed by *A. dorsata*, *A. cerana* and the dwarf bees, and hence, body size cannot necessarily explain the differences in alpha-diversity (**Fig. 1E**). Most bacterial species displayed significant strain level diversity as measured by the percentage of single nucleotide variants (SNVs) per bacterial species within each bee gut with the possible co-occurrence of several closely related strains (**Supplementary Fig. S8**). The Western honeybee *A. mellifera* hosted significantly more strains than other hosts for most bacterial species, while the dwarf honeybees harbored the least strain level diversity (**Fig. 1F**).

Taxonomic composition (**Supplementary Fig. S9-10**) was host-specific at the species level as suggested by β-diversity analyses (PERMANOVA Sorenen distance Ω^2=0.6777, p=0.001, N=195), but with no distinction between the two dwarf bee species (**Fig. 1G**) and no influence of geography (Ω^2=0.009, p=0.086, N=131 excluding two host species - *A. andreniformis* and *A. mellifera* - sampled unequally across locations) (**Supplementary Fig. S11**). We found no evidence for phylosymbiosis, as there was no correlation between bacterial community distance and host divergence (Spearman's p-value = 0.49, mean q-value = 0.619±0.204 across 1000 mantel tests, **Fig. 1H**). For example, *A. andreniformis* and *A. mellifera* are the most distant host species but have much smaller distances in bacterial community composition than other pairs of host species.

Host specificity of individual community members varies across bacteria and host species

Specificity at the community level emerges because of the host-specific distribution of individual community members[15]. We calculated the specificity per bacterial species based on Rohde's Index[43], which considers the prevalence of each community member across samples from different hosts (**Fig. 2A, Supplementary Table S4**).

Overall, we found large differences in the specificity of individual bacterial species both within and between genera. For example, the genus *Bombilactobacillus* mainly consisted of generalist species found across all five honeybee species, while the genera *Lactobacillus* or *Bifidobacterium* contained specialists and generalists (**Fig. 2A**). Some bacterial species showed an intermediate level of specificity as they were shared only among a subset of the five host species and were more prevalent in one host than another (*e.g.*, one of the *Snodgrassella* species). Few shared genera consisted of only specialist species (*Apibacter* and *Commensalibacter*). Notably, there was a clear difference in the specificity of bacterial species across hosts. The microbiota of *A. dorsata* and *A. mellifera* were mainly composed of specialist species, while the microbiota of the other smaller honeybee species (*A. cerana*, *A. florea*, and *A. andreniformis*) was primarily composed of shared species (**Fig. 2A**)

Bacterial species may be shared between host species but exhibit host specificity at the strain level. To look for such strain-level host specificity, we measured the extent of shared SNVs between samples for each bacterial species shared between two or more host species using popANI[44]. We found both bacterial species that segregate and species that do not segregate by host species in terms of strain-level composition (**Fig. 2B, Fig. 2C**). This pattern was observed

across all analyzed genera (**Fig. 2B, Fig. 2C**). Moreover, in some cases, we also found that samples segregated by sampling location (**Fig. 2B, Fig. 2C**). Overall, 75% of the species were host-specific at either strain or species level.

Host switches and species gains and losses interrupted the host-specific evolution of gut microbiota members of honeybees

Co-diversification has been described in other host-associated microbiomes as an important driver of host specificity patterns[7,10,11]. To test for evidence of co-diversification in the honeybee gut microbiota, we constructed core genome phylogenies for all prevalent genera using the collection of all high- and medium- quality MAGs assembled from the five honeybee species (**Fig. 3A**), including several MAGs from each bacterial species within the genus. We used genus-level phylogenies to test for co-diversification because divergence time within genera was roughly equivalent to divergence times in the *Apis* genus (see methods). While most honeybee-derived MAGs from a given bacterial genus were monophyletic relatives to bacteria found in bumble bees and stingless bees, they belonged to distinct clades corresponding to the species clusters determined by ANI (**Supplementary Fig. S3**). In agreement with the observed differences in host specificity, a subset of the identified species-level clades were exclusive to one host species (24% - 47% of all nodes), while others were mixed. To test for co-diversification, we used the Hommola cospeciation test, a generalization of the Mantel test for arbitrary host-parasite interaction graphs[45] on each node, including those representing several MAGs/strains within bacterial species. We then carried out second-order permutation tests (N=100) with shuffled tip labels to account for the rate of false discoveries as described in previously published work[12] (see Methods for more details and an illustration of the approach). Overall, we detected no more significant co-diversifying nodes than expected by chance in any

of the nine tested genera (i.e. those for which trees contained host-specific clades from three or more host species) (**Supplementary Table S5 and Supplementary Table S6**). In other words, nodes that were marked significant in the individual comparison were likely false positives based on the FDRs determined by the second-order permutation tests. In each genus there were at least as many significant nodes from the comparisons between the symbiont trees and host trees with shuffled labels (between none and three) as there were from the actual comparison (between none and four). Closer inspection of the subtrees from these nodes showing the highest signals of potential co-diversification for each genus showed little congruence with host phylogeny (**Supplementary Fig. S12, Supplementary Fig. S13**). In summary, we show that the strength of the signal of co-diversification between honeybees and their gut bacteria is not higher than expected by chance despite high host specificity. Instead, as seen in the phylogenies (**Fig. 3A**), deep-branching host-specific clades are interrupted by generalist clades, indicating recent host switch, and deep-branching generalist clades (sometimes show host-specificity at the strain level) neighbor host-specific clades suggesting recent host specialization, exemplifying the dynamic history of these associations.

Functional differences in the gut microbiota across honeybees

To test whether the host specificity observed at the taxonomic level translates into functional host specificity of the honeybee gut microbiota, we performed a functional analysis of the MAGs. Of 4,759,622 prokaryotic ORFs in our MAG database, more than half (2,685,895 ORFs) were assigned a KO (KEGG Ortholog) identifier. The rarefaction curve analysis of these KOs indicated that most of the functional diversity across the analyzed samples was recovered with our MAG database (**Supplementary Fig. S14**). Overall, the number of KOs per sample followed

a similar trend as the taxonomic diversity, with *A. dorsata* having the most and *A. florea* having the fewest KOs (**Fig. 4A**).

Multivariate analysis of normalized KO counts (and the Robust Aitchison distance) showed that the functional composition of the gut microbiota of five honeybee species was distinct, even between *A. florea* and *A. andreniformis*, which overlapped in composition at the taxonomic level (Ω^2=0.148, p=0.001) (**Fig. 4B**). To understand which bacteria drive these differences, we performed a similar analysis using counts of all KO families in each high- and medium-quality MAG across all samples. As expected, MAGs of bacterial species from the same genus clustered across all host species, suggesting that they occupy similar functional niches and that the overall functional capacity of the bee microbiota is conserved across hosts (**Fig. 4C**). Yet, there were some important differences. Some of the genera only found in *A. dorsata* formed distinct clusters (*e.g.*, *Dysgonomonas*), suggesting that these bacteria occupy unique functional niches in the microbiota of this host species. Other genera (*e.g.*, *Saezia* and *Entomomonas*) clustered with community members prevalent in other honeybee species (*e.g.*, *Snodgrassella*) but absent from *A. dorsata* (**Fig. 4C**). Closer inspection of their metabolic profiles indicated that both *Saezia* and *Entomomonas* like *Snodgrassella* lack major carbohydrate metabolism pathways (**Supplementary Fig. S15**), suggesting potential functional replacements. Finally, although MAGs from the same genus mostly overlapped in terms of their functional profile, we found some cases where MAGs of the same bacterial genus clustered by honeybee species from which they were recovered, suggesting host-specific differences in the functional repertoire at the species-level (and potentially, adaptation) within these genera (**Supplementary Fig. S16)**.

Since pollen-derived glycan (*i.e.*, pectin and hemicellulose) breakdown has been described as one of the characteristic catabolic functions of the gut microbiota *of A. mellifera*[22,46], we specifically looked at the Carbohydrate-active enzyme (CAZyme) repertoire of the microbiota across the five host species. We detected 19,562 CAZyme families among the prokaryotic ORFs in our MAG database. Overall, the composition of these gene families was host-specific (**Supplementary Fig. S17**). While all five honeybee species harbored a similar set of CAZymes for pectin, hemicellulose, and cellulose degradation, *A. dorsata* contained a more extensive and diverse set of such gene families, especially many of which are involved in pectin degradation (**Fig. 4D**). To understand which bacteria drove these differences, we considered the number of genes belonging to CAZyme families from each bacterial species. Previously recognized members contributed most families to degrade pollen-derived glycans, specifically *Gilliamella* and *Bifidobacterium*[22,46] (**Fig. 4E**), except in the giant honeybee *A. dorsata,* where the genus *Dysgonomonas* also contributed key enzyme families involved in pectin degradation such as PL1_2, PL11_1, GH28, GH78, GH138, or CE12 (**Fig. 4F**, **Supplementary Fig. S18**). Interestingly, the CAZyme genes of *Dysgonomonas* were often organized in large genomic islands reminiscent of specialized PULs (polysaccharide utilization loci) found in gut symbionts of mammals[47], suggesting that this community member contributes to the capability of this particular honeybee species to degrade plant-derived fibers (**Fig. 4F**). Moreover, the fact that the MAGs of *Dysgonomonas* found in *Apis dorsata* form deep-branching clades, only distantly related to species isolated from cockroaches[48] and human stool[49], suggests host specialization.

Discussion

Many animals harbor host-specific bacterial communities in their gut[3–5,15]. However, it is still unclear whether and why host specificity varies among closely related host species or bacterial lineages, and which evolutionary processes underlie the observed patterns of specificity and their functional consequences. Our comparative shotgun metagenomic analysis of the gut microbiota of five honeybee species provides novel insights into these questions. It advances our understanding of the evolutionary processes that govern gut microbiota composition across closely related animal species.

We find that the community composition of the gut microbiota of three (*A. mellifera*, *A. cerana*, *A. dorsata*) of the five analyzed honeybee species is host-specific In contrast to *A. mellifera*, *A. cerana* and *A. dorsata*, which harbor host-specific communities, the gut microbiota of the two dwarf honeybee species, *A. florea* and A. *andreniformis*, were indistinguishable from each other, suggesting that host specificity can vary between closely related host species.

Our genome-level data, coupled with our broad sampling design, allowed us to move beyond community-level patterns and to identify host specificity of individual community members. We found that the vast majority of the analyzed community members were host specific, but also that host specificity emerges at different taxonomic levels, for example, at the genus (such as *Saezia* and *Dysgonomonas* in *A. dorsata*), species (such as many *Gilliamella*, *Snodgrassella* and *Lactobacillus* species), and strain level (such as *Bombilactobacillus mellis-36_2*). Considering that all of these bacteria belong to clades specific to social bees, this suggests that the host-specific association between different gut symbionts and their honeybee hosts has been

established across different time points in their evolutionary history. These findings also highlight that patterns of host-specific associations can be overlooked without strain-level/genomic resolution.

Another key finding of our analysis is that the degree of host specificity varies across bacteria (i.e., there are generalist and specialist species). This has also been observed in other systems[15,16,50] and linked to bacterial phenotypes, such as aerotolerance and spore-formation, that affect dispersal ability[15,51–54]. However, in addition to previous work, we observed unexpected high variability in host specificity between closely related bacteria, i.e., species belonging to the same genus. This suggests that microbial traits, such as aerotolerance, usually conserved within genera, cannot explain fully variation in specificity patterns. Instead, we find host species in which they are found to be an important determinant of the specificity of individual community members. Most of the highly host-specific bacterial species were found in *A. dorsata* and *A. mellifera,* while the communities of *A. cerana* and the two dwarf honeybees were dominated by shared bacterial species. *A. mellifera* and *A. dorsata* guts may harbor more specific ecological niches[55] (a narrow range of conditions or resources that an organism can exploit) than the other bee species, facilitating the colonization and maintenance of additional 'specialist' bacteria. This is consistent with the observation that *A. cerana* and the two dwarf bees harbor significantly fewer bacterial species per individual bee than *A. dorsata* and *A. mellifera*. Several different factors could contribute to these patterns. Ecological niches in the host gut can vary even among closely related host species due to differences in host physiology, diet, or gut structure. For example, the host-specific colonization of *S. alvi* of *A. mellifera* may be supported by specific organic acids secreted by the host into the gut lumen and utilized by this bacterium[56]. Slight

differences in dietary preferences, e.g the amount or type of pollen, may allow different combinations of species to co-exist in one host species compared to another[57]. Variation in body and colony size between *Apis* species may determine the carrying capacity and hence bacterial diversity in the gut, as proposed previously[26,40]. Another possibility is that host species may exert different extents of control over their gut microbiota, for examples, e.g. via the immune system, as suggested in other animals[58].

In contrast to studies of other social bees[26,27,30,59], we find little evidence for co-diversification in the honeybee gut microbiota, despite high host specificity. Indeed, the overall cophylogenetic signal was not higher than expected by chance. This might be because previous analyses did not apply extensive tests and corrections[12], use genome-scale phylogenies, or sample the full breadth of microbiota diversity as they relied on isolate genomes. Our results markedly contrast with co-diversification findings in mammals and, in particular, hominids[6–11]. Those studies, using a similar approach, found that the ratio of the number of significantly co-diversifying clades observed and expected only by chance was one order of magnitude higher than our dataset (~12 vs. ~1.2). Additionally, clear examples of highly congruent host and symbiont phylogenies can be found in these hominid studies, which is not the case here. Even in hominids, only 16%-36%[7,9,11] of all tested bacterial clades showed statistically significant signals of co-diversification, suggesting that the evolution of only a minority of the microbiota is driven by co-diversification. Co-diversification might be the exception rather than the rule in both mammals and social bees.

Overall, our results show that ancient and specific gut microbiota-host interactions can be maintained in the absence of prevalent co-diversification[60,61]. Instead, host switches, gains and losses between related species could have been frequent, suggesting that they govern compositional changes in gut microbiota over evolutionary time. This is surprising given the social lifestyle of bees that facilitate within-host species transmission mode. In contrast, ecological opportunities for spillover and invasion of the gut microbiota are not uncommon, maybe due to the sympatric distribution of *Apis* species in Asia. Furthermore, since all five honeybee species have similar dietary preferences, gut structure, and physiology and lack an adaptive immune system, the barrier to colonization between native and non-native bacteria may be low. This is supported by experiments with microbiota-free bees, which showed that bacterial strains isolated from honeybees can colonize a non-native honeybee species and sometimes even outcompete native bacterial strains[26,28]. The eco-evolutionary mechanisms underlying host specificity remain unclear, and the host-specific clades and those exhibiting evidence of recent host switches identified in this study present promising candidates for investigating these processes further. To better understand the evolutionary dynamics of the bee gut microbiota, we suggest that it will be key for future research to identify the rate and determinants of host switches, gain and losses.

The host-specific nature of the microbiota of these honeybees translates into host-specific microbial functional potential. In particular, the specific repertoire of conserved members and the gain of new members with unique potential significantly impacted microbiota functions. For example, honeybee-specific lineages of *Dysgonomonas*, a genus of polysaccharolytic Bacteroidetes also present in the gut of cockroaches, termites, and humans[48,49,62], greatly

expanded the repertoire of pectin and hemicellulose degradation genes in the gut microbiota of *A. dorsata* without necessarily replacing another gut symbiont. *A. dorsata* may be distinct in its broad exposure to more diverse microbes compared to other honeybees, given its highly migratory nature[63]. Interestingly, differences in diet may not explain differences in microbiome diversity, considering that *A. dorsata* tends to forage on fewer mass-flowering plants[64] , and plant-pollinator network analysis suggests that *A.florea*, *A. cerana*, and *A. dorsata* show substantial overlap in the flowers they visit[65]. In summary, the honeybee gut microbial community shows a highly dynamic evolution despite the presence of a few conserved lineages.

Our results form the basis for more detailed functional characterization of diverse bacterial gut symbionts previously not recognized as important members of the honeybee gut microbiome. It vastly expands our knowledge of bee microbiota obtained chiefly from the managed Western honeybee. More generally, it advances our understanding of host specificity in gut microbiota-host associations by showing that even ancient gut microbiota-host interactions can be maintained in the absence of prevalent co-diversification via dynamic gain, losses and exchanges of microbes. This history of frequent turnover can have a critical impact on microbiome functions and suggest that apparently conserved ancient gut microbiota–host associations are susceptible to important change over evolutionary time.

Methods

Sampling and storage

All honeybee workers (10-50 individuals per colony) were collected from colonies across locations described in the metadata sheet (**Supplementary Table S1)**. Most colonies were wild except those *A. mellifera* samples labeled "Langstroth hive" under the hive description column. *A. cerana* samples from Langstroth hives were local bees that were temporarily housed in these hives with minimal intervention but not managed as *A. mellifera* normally are. Bees were stored in small ventilated boxes upon collection and transferred to absolute ethanol within a day. Bees were then stored at -20°C until dissection and DNA isolation, sufficiently preserving the sample for sequencing the microbiome[66]. Our sampling strategy prioritized covering many colonies across locations to capture the diversity of each host species rather than many individuals from the same colony, which was done before[38,40].

Dissection and DNA extraction

To decrease the ratio of DNA from the host cells to bacterial cells, we only used the hindgut which contains most of the bacterial biomass. To do so, we carefully opened the cuticle, removed the stinger and venom glands, and, finally, cut the tissue near the pylorus, below the malpighian tubules removing the midgut and crop. Then, the hindgut tissue was transferred to a bead-beating tube containing 478 µL of sterile PBS and 2.0 mm sterile glass beads and subjected to bead-beating for 30s at 6 m/s in FastPrep-24 5G (MP Biomedicals). Next, 0.1 mm zirconia beads along with 250 µL of 2x CTAB, 2 µL of β-mercaptoethanol and 20 µL of proteinase K (20 mg/mL) was added and bead-beating was carried out for 30s at 6 m/s in FastPrep-24 5G (MP Biomedicals). The homogenate was incubated for 1h at 56°C. Next, 750 µL of Phenol-

chloroform isoamyl alcohol (25:24:1, v/v) at pH 8.0 was added and mixed thoroughly, followed by centrifugation for 10 min at 16,000 xg. The upper aqueous layer was transferred to a new tube, and 500 μL of chloroform was added and mixed thoroughly. The mixture was centrifuged for 10 min at 16,000 xg, and the 300 μL of the upper phase was mixed with 900 μL of chilled absolute ethanol and incubated at -20°C overnight. Next, the precipitated DNA was pelleted by centrifugation for 30 mins at 16,000 xg at 4°C. After two washes with 70% ethanol, it was resuspended in 50 μL of nuclease-free water. Next, it was treated with RnaseA and purified using magnetic beads (Clean NGS). DNA concentration was measured using Qubit™ (Thermo Fischer Scientific). Bacterial absolute abundance was measured using qPCR with primers targeting the 16S rRNA region as described earlier[67] with 0.1x dilution of gut homogenate.

Metagenomic sequencing

Library preparation was carried out using the Nextera XT library kit (Illumina), according to manufacturer instructions, with the number of PCR cycles chosen to account for the amount of DNA in samples. Illumina sequencing was carried out using Novaseq (2x 150 bp) with a target of 10-50 million microbial reads. Since there was a significant difference between host species in the total gut microbial content as determined by qPCR using universal primers, first, a run of sequencing was carried out without normalization of the amount of samples added to the library to estimate the percentage of host DNA and microbial DNA in the library made from each sample. This was useful in ensuring that there was enough DNA from the bacterial community in all the samples, including those where the library might have been dominated by host DNA due to lower bacterial biomass. Further sequencing was then carried out after pool correction, where the library was adjusted in such a way when possible that at least 15 million microbial reads

were obtained (based on expectation from the first shallow sequencing), even for the samples with the lowest percentage of microbial DNA. All library preparation and sequencing was done at the Lausanne Genomic Technologies facility (GTF), UNIL. Details about all further bioinformatic analysis, including the code used, can be found on GitHub (https://github.com/Aiswarya-prasad/honeybee-cross-species-metagenomics) and Zenodo (https://zenodo.org/doi/10.5281/zenodo.13732977) in the repositories.

Mapping to databases

To assess the amount of bacterial, host, and other reads in each sample, reads were mapped to host and MAG databases. The host database comprised draft or complete genomes of all the four host species (GCA_000184785.2, GCA_000469605.1, GCA_001442555.1, GCF_003254395.2) except *Apis andreniformis* since no genome was available in public databases for this host species. A database was compiled using representative MAGs identified for each 95% ANI cluster or bacterial species with the highest quality score (calculated as described by dRep[68] using various MAG characteristics, including checkM contamination and completeness). Trimmed and quality-filtered reads were mapped to the host database using BWA-mem[69] followed by the MAG database (and *vice versa* to account for non-specific mapping). Three samples were removed from the downstream analysis as their metagenomic reads did not map the host that they were labeled as (**Supplementary Table S3**) and might have faced cross-contamination during sample processing. Samples for which relatively few reads mapped to the MAG database often were also low in bacterial biomass as measured by qPCR (**Supplementary Fig. S16**). Indicating that the sequencing and workflow to detect microbial diversity in each sample was otherwise robust. Initial taxonomic profiling of the reads using mOTUs[70] detected the expected core marker genes from various honeybee gut microbiota members.

Assembly and binning of MAGs

Raw reads were trimmed and quality-filtered before assembly using Trimmomatic[71] and bbtools. We did not carry out host filtering at this stage as we have no host genome for *A. andreniformis* and did not want to bias our samples due to non-specific mapping. For some samples that had a large number of reads (19/200 samples) due to which assemblies were not successful, bbnorm was used to down-sample regions of high depth. For samples with too many host reads due to which assemblies failed (3/200 samples), host-filtered reads were used. The detailed workflow used for assembly and binning into MAGs is described in GitHub (https://github.com/Aiswarya-prasad/honeybee-cross-species-metagenomics). Briefly, contigs were assembled using metaSpades[72]. Quality-filtered reads were mapped against assembled scaffolds (all against all samples), and then scaffolds were binned into MAGs using coverage information as implemented in MetaBAT2[73]. The quality and completeness of the resulting MAGs were assessed using CheckM[74] and low (< 50% completeness or >5% contamination), medium (50-90% completeness and <5% contamination), high (> 90% completeness and <5% contamination) MAGs were then clustered into groups averaging 95% average nucleotide identity (ANI) using dRep[68]. To further ensure that our MAGs represent true species and are not assembly and binning artifacts, we inferred phylogenies, including isolate genomes along with our MAGs. The rationale is that if MAGs are artifacts, they would rarely cluster together in the phylogeny and would instead be placed apart from isolate genomes. This new analysis shows that our MAGs form discrete phylogenetic clades, which include MAGs assembled from several different samples. Further, these clades often include closely related isolate genomes in cases where several strains were included from the same species. MAGs belonging to clusters without isolate genomes represent novel species not isolated before (**Supplementary Fig S4**). We are, therefore,

confident that the MAGs represent true species and can be used to infer meaningful phylogenies. Other studies in the field, which have used similar approaches to ours, are similarly able to establish meaningful phylogenies from MAGs[7].

Taxonomic annotation was carried out using GTDB-tk[75]. No manual curation of MAGs was performed; however, Kaiju[76] (nr database) and Kraken2[77] (custom database including bacteria, viruses and honeybee) classification of scaffolds screened by random sampling further confirmed that scaffolds from the same MAG matched each other and the taxonomic assignment of the MAG. Each cluster of MAGs was named to reflect the GTDB species name or, in cases where none was assigned, a unique identifier, including the genus name was used. Combined with the inspection of ANI heatmaps (**Supplementary Fig. S3**), a final set of species names was assigned to each of the 95% ANI clusters.

Taxonomic profiling

InStrain[44] was used according to its recommended workflow for taxonomic profiling at the bacterial species and strain level. Trimmed and quality-filtered reads were competitively mapped to the database of representative MAGs using the tool bowtie2[78], which presents mapping scores in a format compatible with InStrain. Among many approaches to detecting microbial species in a dataset, this stands out in its appropriate filters applied to mapped reads, including read ANI, which helps differentiate genomes of closely related species (<95% ANI) and breadth of the genome mapped to avoid falsely identifying species as present due to high coverage of a region shared with others or horizontally transferred. Moreover, genome coverage of a given species can be misleading when comparing communities of varying diversity, as an equally abundant species would have higher coverage in a community with low diversity. A species was

considered present if at least 50% of its genome (comprising all scaffolds) was covered by the reads. The justification for the choice of this threshold is exemplified for several Snodgrassella species in **Supplementary Fig. S5**: briefly: there is a clear gap between samples where a given species is detected (in large green circles) or not (in small red circles). We inspected such plots for all the species represented in the database and visually verified that the breath coverage we used (50%) was valid across species. These plots are included with the dataset found in the Zenodo repository (https://zenodo.org/doi/10.5281/zenodo.13732977).

Relative coverage, when used, was obtained by dividing the coverage of each species by that of the sum of all and qPCR copy numbers of the bacterial 16S rRNA region were used to estimate absolute abundance for comparison between samples where mentioned. For samples from India (n=50), qPCR values were interpolated based on the median of the host species. The underlying data (presence-absence, relative abundance etc.) and normalization approach are explicitly indicated for each downstream analysis section. After filtering samples as mentioned earlier and removing samples with fewer than 1 million mapping quality filtered reads mapping to the representative database of MAGs, a total of 195 samples were included in the analysis (*A. mellifera* 34, *A. cerana* 45, *A. dorsata* 44, *A. florea* 42 and *A. andreniformis* 30).

Host specificity and strain-level profiles

Host specificity was calculated for each bacterial species as the Rohde's index of specificity[43] using the prevalence of the species across hosts. Strain-level composition profiles for each bacterial species represented in the MAG database were generated per sample using InStrain in

database mode. The percentage of variant sites was calculated as the percentage ratio of the number of SNVs to the total length of the MAG covered. A site is counted as an SNV if the site is covered at least 5x and the allele is present at a frequency of at least 0.05. Regions mapping with a read ANI of <0.92 (might be cross-mapping of reads from a different species) or coverage of <1x were not considered. To compare the strain profile of samples, the popANI measure, as comprehensively described by InStrain[44], was used. Briefly, a value from 0 to 1 was calculated for each species for each pair of samples in which the genome was covered at least by 5x and at least half of the genome was covered in both samples. A position of the genome was considered if there was more than one allele with a frequency of at least 0.05. If both samples share no alleles in that position, it is counted as a popANI substitution. The value for that pair of samples was then calculated as the ratio of positions containing popANI substitutions and the number of positions compared. This value was used to create a matrix of Jaccard distances between all pairs of samples. Finally, each of the bacterial species was either considered host-specific, if they were detected in only one host species, or shared if not. Shared species were considered host-specific at the strain level if samples showed significant clustering by host species ($p < 0.01$) according to a PERMANOVA test.

Phylogeny construction and co-diversification test

To gain a rough estimate of the taxonomic level at which to expect co-diversification, we computed a core genome phylogeny using the nucleotide sequence of 120 bacterial marker genes for all medium and high-quality MAGs identifying core genes using GTDB-tk, aligning in a codon-aware manner using MACSE[79] and constructed using IQ-TREE[80]. From this tree, the median substitutions per site were 0.014 for MAGs of the same species, 0.39 for different species of the same genus, and 2.65 for different genera (**Supplementary Table S7**). Bacteria evolve at

a rate in the order of magnitude of 10^{-3} substitutions per site per million years[81]. Since the *Apis* clade is about 30 million years old[82], gut microbes that diverged during this timeline could be as divergent as different species within a genus or closer, confirming that genus-level trees are appropriate for these tests.

OrthoFinder[83] was used to infer orthologous gene families of all medium and high-quality MAGs for each genus along with previously published genomes of the same genus from other environments as outgroup (details of genomes used in metadata files included in the Github (https://github.com/Aiswarya-prasad/honeybee-cross-species-metagenomics) and Zenodo (https://zenodo.org/doi/10.5281/zenodo.13732977) repository. Phylogenies were constructed using IQ-TREE using the amino acid sequences of genus-level single-copy core genes aligned by MAFFT[84]. All medium and high-quality MAGs (including several per bacterial species) were included in the phylogenies in order to preserve maximum information and power of resolution. Since medium-quality MAGs were included in phylogenies, core genes can be randomly missing in some MAGs. Naively defining core genes as those present in all MAGs would result in the exclusion of most orthologs. Hence, as implemented in OrthoFinder, when <1000 single-copy orthologs are present in all MAGs, it iteratively identifies the orthologs that are found once in N-1 MAGs until either considering one less MAG does not double the number of orthologs or more than half the MAGs are dropped. Core single-copy orthologs identified were in agreement with those orthologs identified as core by mOTUpan[85], which uses a Bayesian approach informed by the completeness score of MAGs. The number of orthologous gene families (OG) used for core genome phylogenies and the minimum percentage of MAGs of the genus in which each OG was detected (*e.g.*, 100% means all OGs were detected in all MAGs and 50% means every OG was found in 50% or more of the MAGs): Apibacter - 1096 (94.5%), Apilactobacillus - 704 (81.8%),

Bartonella - 705 (81.5%), Bifidobacterium - 118 (84.0%), Bombilactobacillus - 194 (89.4%), CALYQJ01 - 883 (89.5%), CALYQQ01 - 1221 (84.6%), Commensalibacter - 1120 (96.6%), Dysgonomonas - 223 (86.5%), Entomomonas - 1144 (91.7%), Frischella - 984 (85.9%), Gilliamella - 189 (88.0%), Pectinatus - 1143 (97.3%), Pluralibacter - 1419 (100.0%), Saezia - 1000 (92.3%), Snodgrassella - 799 (91.5%), WRHT01 - 1897 (100.0%). Further information about the parameters and best-fit model chosen to compute the phylogenies can be found in the log files included in the corresponding directory containing IQ-TREE results in the Zenodo (https://zenodo.org/doi/10.5281/zenodo.13732977) repository.

A host phylogeny inferred based on complete mtDNA sequences of honeybee species rooted by a bumble species published before[86] was used for co-diversification tests. Tests were carried out using the Hommola cospeciation test[87] (implemented in Python) on all nodes containing at least 7 MAGs from at least three host species and with a maximum tip-tip distance not exceeding 1 for 100 permutations. This node-by-node comparison was then carried out on trees with tip labels artificially permuted randomly across (N=100) second-order permutation test to estimate the number of nodes displaying significant results by chance, as described in an earlier study[12]. See (**Supplementary Fig. S20**) for a detailed schematic of this approach. Thresholds were set at various levels for comparison of the number of significant nodes in the original tree and the median number of significant nodes in tip label permuted trees: relaxed ($p<0.05$), medi um ($p<0.05$ and $r>0.75$) and strict ($p<0.01$ and $r>0.75$).

Gene functional annotation and profiling

Open reading frames (ORFs) were identified on all assembled scaffolds of each sample using Prodigal[88]. ORFs marked partial or shorter than 300 bp were removed. Further, those ORFs

detected on scaffolds identified as being of eukaryotic origin by Whokaryote[89] or unclassified as bacteria by Kaiju (nr database) were also excluded. The remaining ORFs were then annotated using DRAM[90] to obtain KEGG ortholog (KO) identities. Finally, the gene catalog comprising all these ORFs was assigned to CAZyme families as implemented in Cayman[91].

Trimmed and quality-filtered reads from each sample were mapped back to their respective assemblies using bowtie2. Low-quality (Q<20) alignments were removed, and coverage per gene/ORF was calculated using BEDtools[92]. Genes were considered detected if >5 reads were mapped to it and > 90% of the bases in the gene were covered by at least a read. The abundance of each gene was then calculated in terms of reads per kilobase per million (RPKM) calculated as follows:

$$RPKM = \frac{\#\ Mapped\ reads}{\left(\frac{Gene\ length}{1000}\right) * \left(\frac{\#Total\ reads}{10^6}\right)}$$

For KOs and CAZymes, summed RPKMs of constituent genes were used. To estimate the average count of CAZymes of interest (all families other than GT and AA families) in each genus per individual, the RPKM of all CAZymes in each genus was summed per sample, and then the average of this sum across samples was converted to a relative proportion to represent the relative contribution of each genus on average in each host species and plotted. Only genera with a prevalence >0.05 in the respective host were considered.

To infer biological pathways represented within each MAG, we collected the KOs detected within each MAG and used MinPath[93], a parsimony approach which aims to explain the set of protein (KO) families using the minimum number of pathways. Further, we used MaAsLin 2[94],

which was developed to assess multivariable associations of microbial community features with complex metadata. With the presence-absence matrix of pathways as the dataset and 'genus' as the explanatory variable we identified the most likely functional similarities driving certain genera to cluster together in the PCoA plot of all KOs per MAG (**Fig. 4C**).

Statistical tests and bioinformatic analyses

All details about the specific parameters used for each tool and their versions, along with conda environment specifications, are included in the Github (https://github.com/Aiswarya-prasad/honeybee-cross-species-metagenomics) and Zenodo (https://zenodo.org/doi/10.5281/zenodo.13732977) repositories. In general, bioinformatic workflows were implemented using Snake make (v7.28.1) to parallelize jobs on a computing cluster at University of Lausanne, preprocessing and filtering using Python scripts, visualization using ggplot with several related packages in R and iTOL for phylogenetic trees and arrangement and decoration of plots using Adobe Illustrator. Unless otherwise specified, statistical tests of the results of each section were carried out as follows. Significance in boxplots was tested using the Wilcoxon Rank Sum test (significant if $p<0.05$). Distance matrices were calculated using the method indicated in their respective plots. The package Betapart[95] was used to estimate the turnover component of the Jaccard and Sorensen distances. Multivariate analyses were visualized using PcoA or NMDS and tested using the adonis2 function implemented in the R package Vegan. The effect size was adjusted for sample size to get Ω^2 using the adonis_OmegaSq function from the package Russel88/MicEco (https://zenodo.org/badge/latestdoi/83547545). To test for phylosymbiosis, we minimized bias due to pseudo-replication by performing 1000 permutations. Each permutation included five randomly selected samples, one of each host species, and tested for correlation between bacterial

and host distances using a Mantel test (as implemented in the Vegan package). The p-values were adjusted using the FDR method.

Acknowledgments

We thank the following people and organizations for supporting honeybee sampling and access. The Nature Inspired team (John Chan, Steven Wong, & JC Tan), Mr. HoneyBee Farm, The Orang Asli community (Temengor, Perak), Sherwani Honey Bee, Mr. Wee, Mr. Steve, and MY Bee Saviour. The Cayman team (Quinten Ducarmon, Nicolai Karcher, and Christian Schudoma) for their support with early access to their software and custom annotation of our dataset. The Lausanne Genomics Technology Facility (GTF) team at the University of Lausanne. This work was supported by a Spirit grant (grant number 189496), an SNSF Consolidator grant ('GLOBEE,' grant number 213860), and the NCCR Microbiomes, a National Centre of Competence in Research (grant number 180575), funded by the Swiss National Science Foundation.

Author contributions

Conceptualization – PE, SHY, AP, FM; Methodology – AP, FM, PE; Software – AP; Validation – AP; Formal analysis – AP; Investigation – ADP, AP, RS, AS; Resources – PE, SHY, AB; Data curation – AP; Writing - Original Draft – AP; Review and Editing – PE, FM, AP, AB, SHY; Visualization – AP; Supervision – PE, SHY, FM, AB; Project administration – PE, SHY, AB; Funding acquisition – PE, SHY

Competing interests: All authors declare they have no competing interests.

Data availability: Raw metagenomic data has been deposited to the NCBI Sequence Read Archive (SRA) under the Project ID PRJNA1157353. The collection of MAGs, some important

intermediate files, and tables can be found in the Zenodo repository (https://zenodo.org/doi/10.5281/zenodo.13732977).

Code availability: Code and details of parameters and software used are available at https://github.com/Aiswarya-prasad/honeybee-cross-species-metagenomics archived at https://zenodo.org/doi/10.5281/zenodo.13732977

Figures

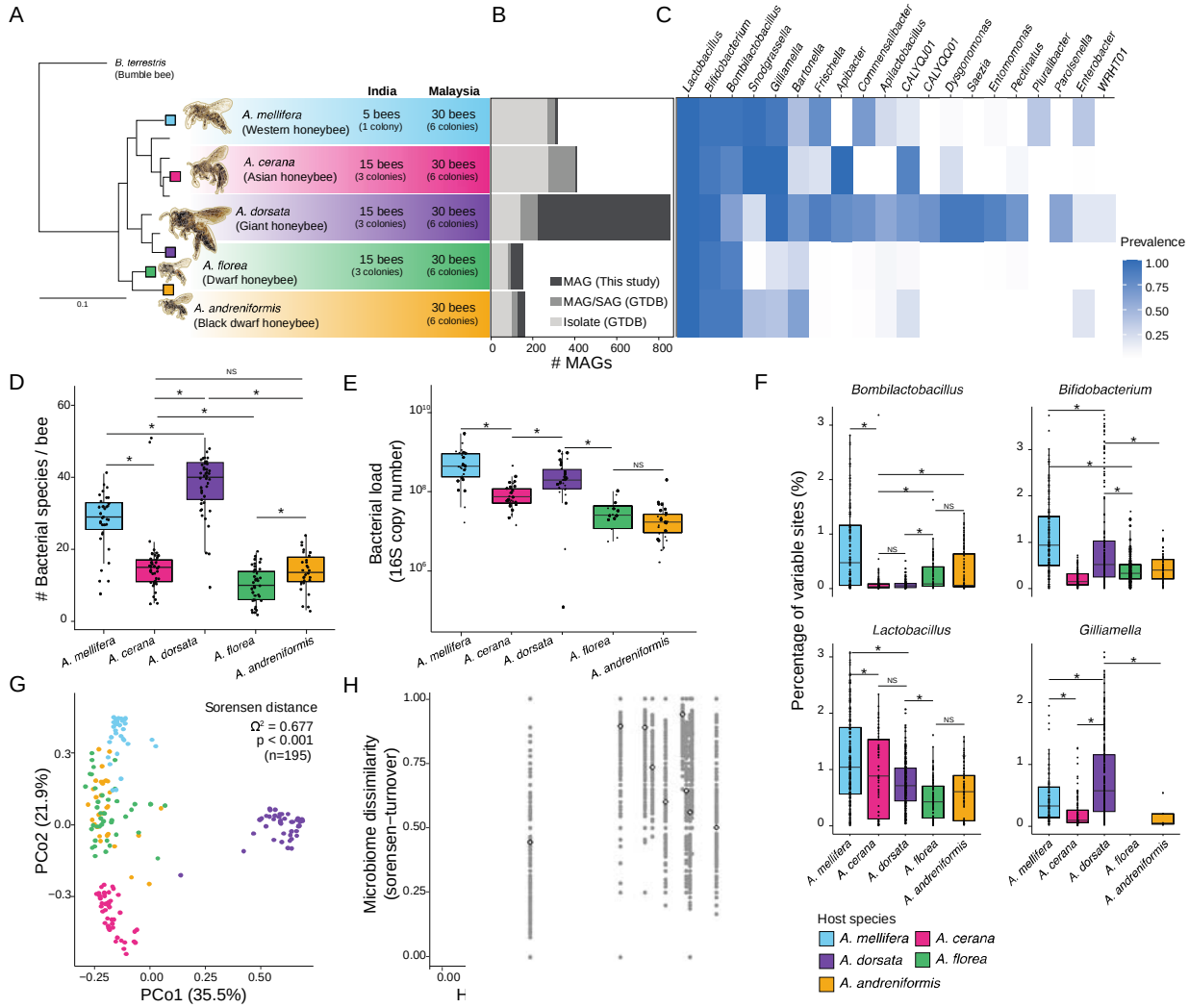


Figure 1. Community-level patterns in the honeybee gut microbiome across taxonomic scales. (A) Apis bee phylogeny with tips indicating extant species and the colored tips species included in this study. Pictures of bees [96] are to-scale relative to each other. Numbers and text next to species names indicate the sampling design. (B) Barplot showing the number of MAGs recovered from this study for which no isolate genomes of the same species (>95% ANI) were found in GTDB, those that were represented by MAGs or SAGs in GTDB and those that are similar to published genomes from isolated bacteria represented in GTDB. (C) Heatmap with the prevalence of each genus within each host species showing genera with prevalence > 10%

excluding the following three: *Klebsiella* (*K. pneumonia*) – 10% in *A. mellifera*; Pantoea (P. dispersa) – 17% in *A. mellifera*; *Melissococcus* (*M. plutonius*) – 13% in *A. andreniformis*). (D) Boxplot of the number of bacterial species in each honeybee gut across host species. (E) Boxplot of the number of 16S rRNA gene copies per 0.1μL of individual bee gut homogenate across host species. (F) Boxplot of the percentage of variable sites per individual of each host species shown for four most prevalent genera.Each point represents the percentage of variable sites found in each bacteria species of the genus within each individual of that host species. (G) PCoA plot based on the Sorensen distance across host species. (H) Scatter plot of microbiota dissimilarity (as the turnover component of Sorenson distance) as a function of host species divergence based on 16S+12S rRNA[97] with white points with thick black outlines marking the median of values for the comparisons of all pairs of samples from a given pair of host species.

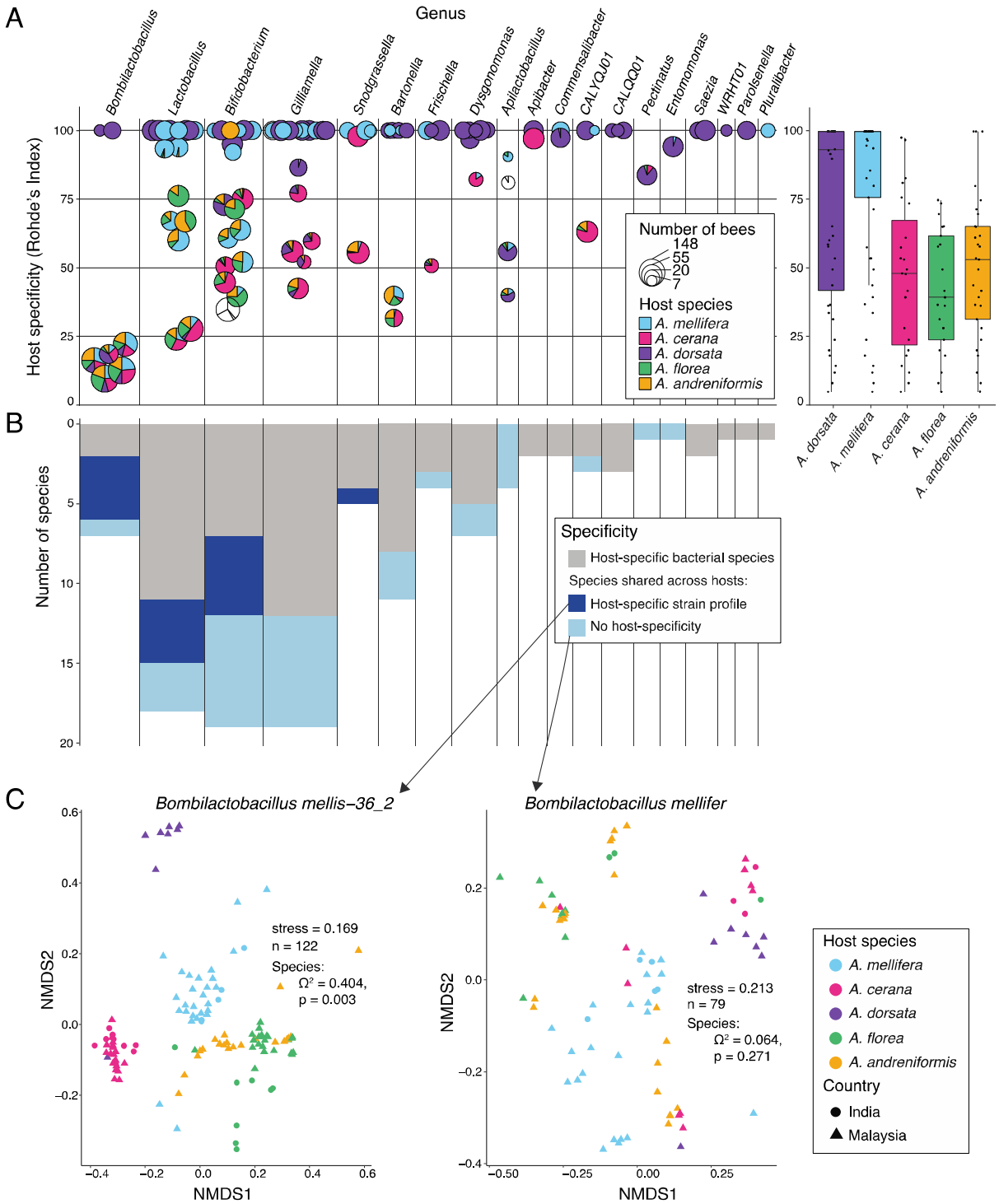


Figure 2. Host-specificity of individual gut microbial lineages across honeybee hosts. (A) Scatter-pie plot of Rohde's index of specificity with each circle representing a bacterial species

with the size representing the number of bees in which the species was found, and the pie chart showing the proportion of samples belonging to each honeybee host. The adjacent boxplot shows Rohde's specificity index for each species found in each honeybee host. (B) Distribution of species richness per genus across three categories of specificity levels: host-specific at the species level (grey), host-specific at the strain level, but the species is detected across hosts (dark blue) or not specific even at the strain level (light blue). (C) NMDS plots from a strain level generalist and specialist bacterial species are included as an example. The Jaccard dissimilarity was calculated in terms of strain level composition (popANI) tested using PERMANOVA (significant if the p-value is < 0.05). Only bacterial species with enough coverage to recover SNPs in at least five samples spanning at least two host species were included.

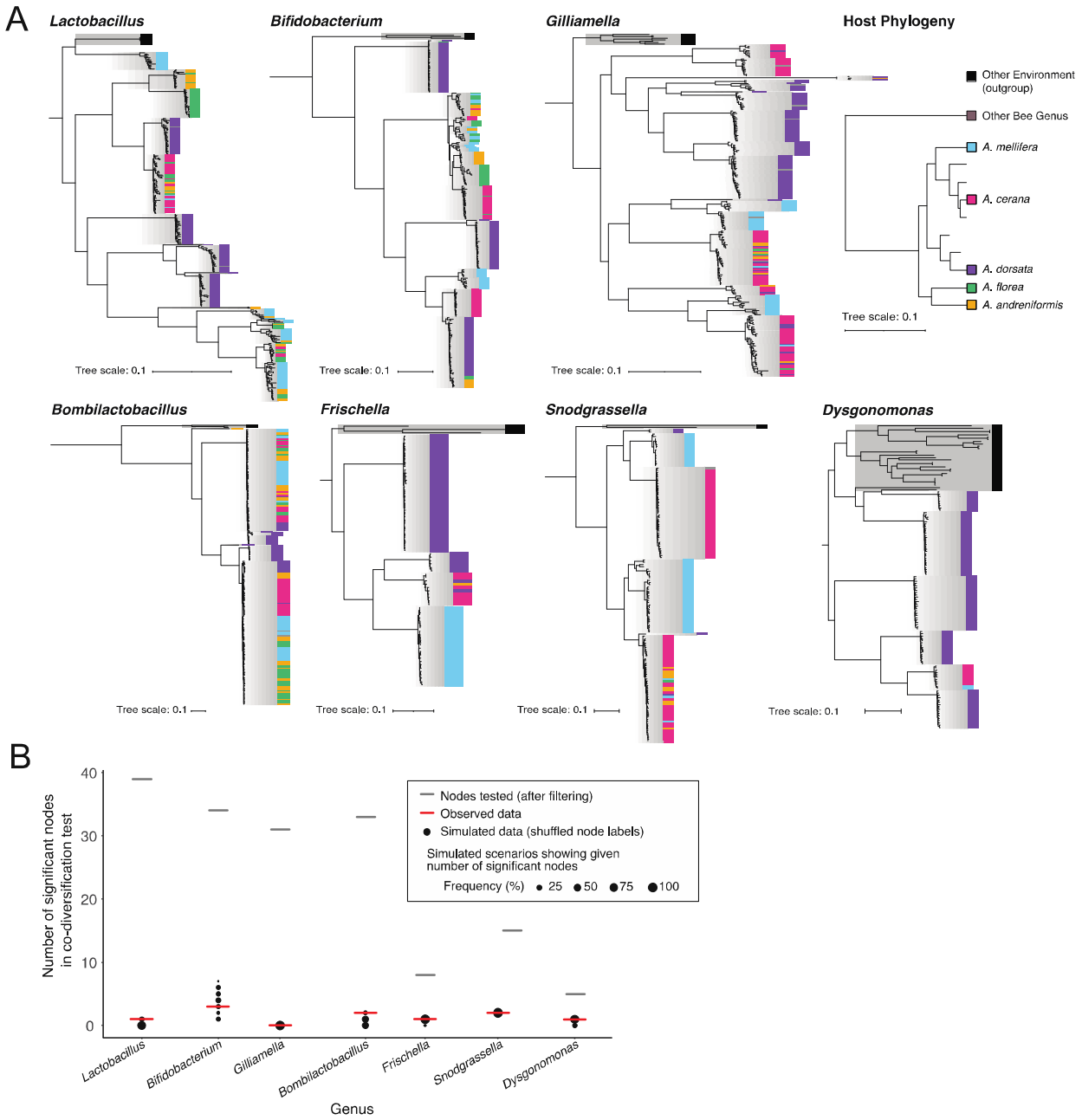


Figure 3. Bacterial genus-level core genome phylogenies of MAGs. (A) Core genome phylogenies of MAGs from various genera found in more than one host species with the color at the leaf tip indicating the host species from which the MAG was recovered and solid grey boxes the set of genomes included to represent outgroup species of the genus from environments other than honeybee guts. The scale bar represents 0.1 amino-acid substitutions per site. The phylogenies are maximum likelihood trees inferred by IQ-TREE (see Methods). (B) Scatter plot

representing the number of nodes significant (under strict threshold) for co-diversification in the actual node-by-node comparison (red line) compared with that in the second-order permutation test comparisons (N=100) with host tree node labels shuffled. The other short horizontal line (grey) denotes the number of nodes compared after excluding those that were filtered (did not have at least 7 MAGs from at least 3 different host species).

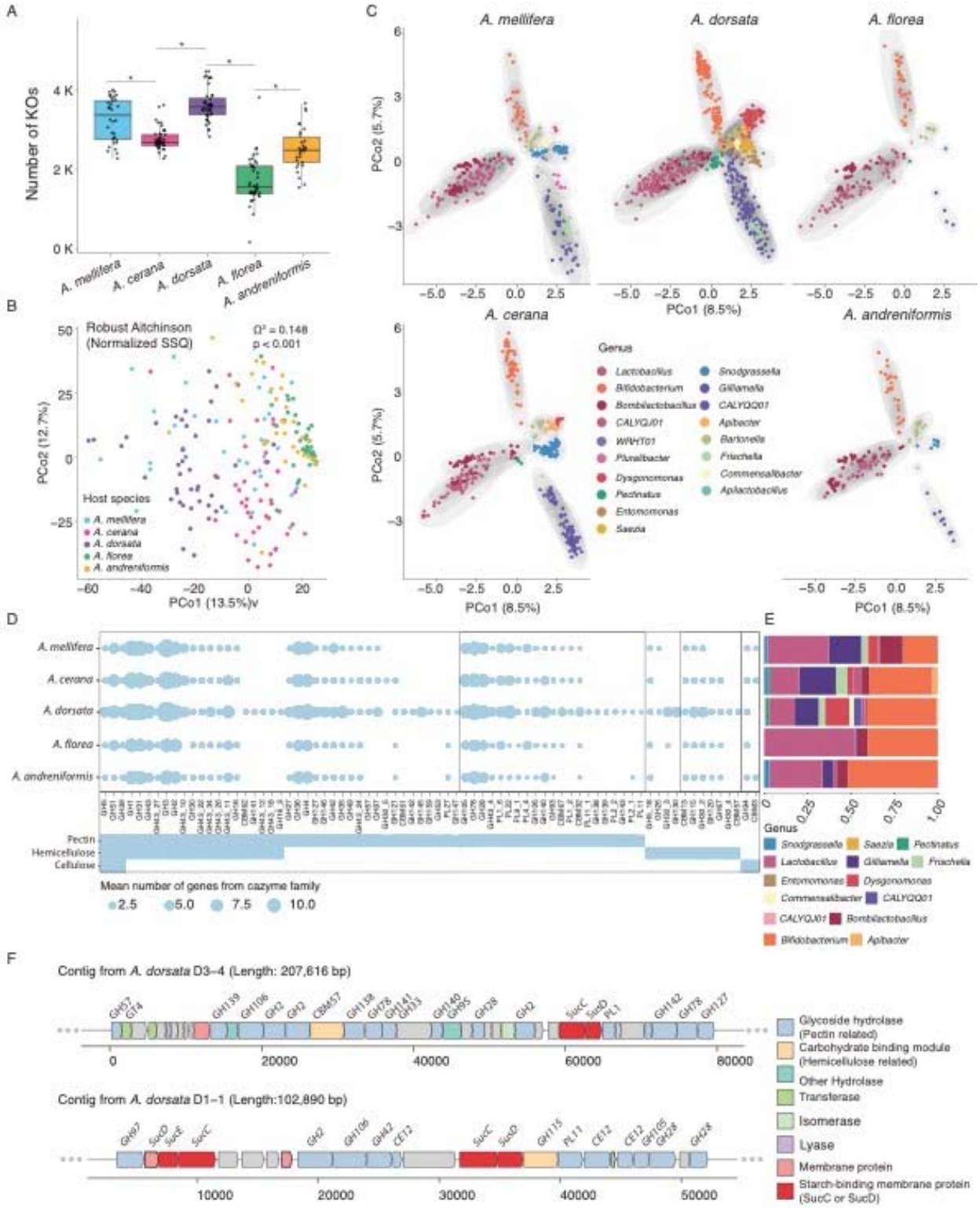


Figure 4. Functional potential of honeybee gut microbiota across host species. (A) Boxplot of the number of KEGG ortholog (KO) families detected per individual of each host species Wilcox test ($p < 0.05$) indicated on the plot. (B) PCoA of KO family composition across all

samples, visualized using the Robust Aitchison distance on a matrix of reads per kilobase per million (RPKM) summed across genes belonging to the respective KO family. Each point on the plot represents an individual honeybee sample and the color the host species identity of the sample. PERMANOVA test indicated on the plot tests the effect of host species identity. (C) PCoA of all KO families from each MAG using Robust-Aitchison distance on the RPKM matrix of KOs. Each point represents a bacterial MAG colored by the bacterial genus it belongs to. MAGs obtained from different host species were separated across panels for clearer visualization. The axes across the five panels have been maintained to the same limits. So, the panels can be superposed.. (D) Dotplot of the mean number of genes from each CAZyme family with a potential role in pectin, hemicellulose, or cellulose degradation per individual of each host species. (E) Barplot of the proportion of genes from CAZyme families contributed by each bacterial genus. (F) Loci within Dysgonomonas MAGs encoding putative polysaccharide utilization loci (PULs).

References

1. Levin, D. *et al.* Diversity and functional landscapes in the microbiota of animals in the wild. *Science* **372**, eabb5352 (2021).
2. McFall-Ngai, M. *et al.* Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci.* **110**, 3229–3236 (2013).
3. Mallott, E. K. & Amato, K. R. Host specificity of the gut microbiome. *Nat. Rev. Microbiol.* **19**, 639–653 (2021).
4. Lim, S. J. & Bordenstein, S. R. An introduction to phylosymbiosis. *Proc. R. Soc. B* (2020) doi:10.1098/rspb.2019.2900.
5. Mazel, F. *et al.* Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? *mSystems* **3**, e00097-18 (2018).
6. Arora, J. *et al.* Evidence of cospeciation between termites and their gut bacteria on a geological time scale. *Proc. R. Soc. B Biol. Sci.* **290**, 20230619 (2023).
7. Sanders, J. G. *et al.* Widespread extinctions of co-diversified primate gut bacterial symbionts from humans. *Nat. Microbiol.* **8**, 1039–1050 (2023).
8. Moeller, A. H. *et al.* Cospeciation of gut microbiota with hominids. *Science* **353**, 380–382 (2016).
9. Rühlemann, M. C. *et al.* Functional host-specific adaptation of the intestinal microbiome in hominids. *Nat. Commun.* **15**, 326 (2024).
10. Good, B. H. Limited codiversification of the gut microbiota with humans. 2022.10.27.514143 Preprint at https://doi.org/10.1101/2022.10.27.514143 (2023).
11. Suzuki, T. A. *et al.* Codiversification of gut microbiota with humans. *Science* **377**, 1328–1332 (2022).

12. Moeller, A. H., Sanders, J. G., Sprockett, D. D. & Landers, A. Assessing co-diversification in host-associated microbiomes. *J. Evol. Biol.* **n/a**, (2023).
13. Song, S. J. *et al.* Comparative Analyses of Vertebrate Gut Microbiomes Reveal Convergence between Birds and Bats. *mBio* **11**, 10.1128/mbio.02901-19 (2020).
14. Groussin, M. *et al.* Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **8**, 14319 (2017).
15. Mazel, F., Guisan, A. & Parfrey, L. W. Transmission mode and dispersal traits correlate with host specificity in mammalian gut microbes. *Mol. Ecol.* **33**, e16862 (2024).
16. Raulo, A. *et al.* Social and environmental transmission spread different sets of gut microbes in wild mice. *Nat. Ecol. Evol.* **8**, 972–985 (2024).
17. Douglas, A. E. Simple animal models for microbiome research. *Nat. Rev. Microbiol.* **17**, 764–775 (2019).
18. Moeller, A. H. Partner fidelity, not geography, drives co-diversification of gut microbiota with hominids. *Biol. Lett.* **21**, 20240454 (2025).
19. Nayfach, S. & Pollard, K. S. Toward Accurate and Quantitative Comparative Metagenomics. *Cell* **166**, 1103–1116 (2016).
20. Mazel, F., Prasad, A. & Engel, P. Host specificity of gut microbiota associated with social bees: patterns and processes. *Microbiol. Mol. Biol. Rev.* (2025).
21. *Honeybees of Asia*. (Springer Berlin Heidelberg, Berlin, Heidelberg). doi:10.1007/978-3-642-16422-4.
22. Engel, P., Martinson, V. G. & Moran, N. A. Functional diversity within the simple gut microbiota of the honey bee. *Proc. Natl. Acad. Sci.* **109**, 11002–11007 (2012).

23. Ellegaard, K. M. *et al.* Genomic changes underlying host specialization in the bee gut symbiont *Lactobacillus* Firm5. *Mol. Ecol.* **28**, 2224–2237 (2019).
24. Sarton-Lohéac, G. *et al.* Deep Divergence and Genomic Diversification of Gut Symbionts of Neotropical Stingless Bees. *mBio* **0**, e03538-22 (2023).
25. Ellegaard, K. M. *et al.* Extensive intra-phylotype diversity in lactobacilli and bifidobacteria from the honeybee gut. *BMC Genomics* **16**, 284 (2015).
26. Kwong, W. K. *et al.* Dynamic microbiome evolution in social bees. *Sci. Adv.* **3**, e1600513 (2017).
27. Li, Y., Leonard, S. P., Powell, J. E. & Moran, N. A. Species divergence in gut-restricted bacteria of social bees. *Proc. Natl. Acad. Sci.* **119**, e2115013119 (2022).
28. Zhang, W. *et al.* Genomic features underlying the evolutionary transitions of *Apibacter* to honey bee gut symbionts. *Insect Sci.* **29**, 259–275 (2022).
29. Wu, J. *et al.* Honey bee genetics shape the strain-level structure of gut microbiota in social transmission. *Microbiome* **9**, 225 (2021).
30. Koch, H., Abrol, D. P., Li, J. & Schmid-Hempel, P. Diversity and evolutionary patterns of bacterial gut associates of corbiculate bees. *Mol. Ecol.* **22**, 2028–2044 (2013).
31. Luo, S., Zhang, X. & Zhou, X. Temporospatial dynamics and host specificity of honeybee gut bacteria. *Cell Rep.* **43**, 114408 (2024).
32. Niode, N. J., Adji, A., Rimbing, J., Tulung, M. & Tallei, T. E. Composition and diversity of bacteria from giant Asian honeybee Apis dorsata gut. *Biodiversitas J. Biol. Divers.* **22**, (2021).

33. Ganeshprasad, D. N., Jani, K., Shouche, Y. S. & Sneharani, A. H. *Gut Bacterial Inhabitants of Open Nested Honey Bee, Apis Florea*. https://www.researchsquare.com/article/rs-225332/v1 (2021) doi:10.21203/rs.3.rs-225332/v1.
34. Gruneck, L., Gentekaki, E., Khongphinitbunjong, K. & Popluechai, S. Distinct gut microbiota profiles of Asian honey bee (Apis cerana) foragers. *Arch. Microbiol.* **204**, 187 (2022).
35. Gruneck, L., Khongphinitbunjong, K. & Popluechai, S. Gut microbiota associated with two species of domesticated honey bees from Thailand. *Symbiosis* **83**, 335–345 (2021).
36. Khan, K. A. *et al.* Gut microbial diversity in Apis cerana indica and Apis florea colonies: a comparative study. *Front. Vet. Sci.* **10**, (2023).
37. Ge, Y., Jing, Z., Diao, Q., He, J.-Z. & Liu, Y.-J. Host Species and Geography Differentiate Honeybee Gut Bacterial Communities by Changing the Relative Contribution of Community Assembly Processes. *mBio* **12**, e00751-21 (2021).
38. Ellegaard, K. M. & Engel, P. Genomic diversity landscape of the honey bee gut microbiota. *Nat. Commun.* **10**, 446 (2019).
39. Regan, T. *et al.* Characterisation of the British honey bee metagenome. *Nat. Commun.* **9**, 4995 (2018).
40. 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.e7 (2020).
41. 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).

42. Fouks, B. *et al.* The genomic basis of evolutionary differentiation among honey bees. *Genome Res.* **31**, 1203–1215 (2021).

43. Rohde, K. & Rohde, P. P. How to measure ecological host specificity. *Vie Milieu* (2008).

44. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).

45. Hommola, K., Smith, J. E., Qiu, Y. & Gilks, W. R. A Permutation Test of Host–Parasite Cospeciation. *Mol. Biol. Evol.* **26**, 1457–1468 (2009).

46. Zheng, H. *et al.* Division of labor in honey bee gut microbiota for plant polysaccharide digestion. *Proc. Natl. Acad. Sci.* **116**, 25909–25916 (2019).

47. Wardman, J. F., Bains, R. K., Rahfeld, P. & Withers, S. G. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nat. Rev. Microbiol.* **20**, 542–556 (2022).

48. Vera-Ponce de León, A., Jahnes, B. C., Duan, J., Camuy-Vélez, L. A. & Sabree, Z. L. Cultivable, Host-Specific Bacteroidetes Symbionts Exhibit Diverse Polysaccharolytic Strategies. *Appl. Environ. Microbiol.* **86**, e00091-20 (2020).

49. Bilen, M. *et al.* Dysgonomonas massiliensis sp. nov., a new species isolated from the human gut and its taxonogenomic description. *Antonie Van Leeuwenhoek* **112**, 935–945 (2019).

50. Gile, G. H. Protist symbionts of termites: diversity, distribution, and coevolution. *Biol. Rev.* **99**, 622–652 (2024).

51. Tung, J. *et al.* Social networks predict gut microbiome composition in wild baboons. *eLife* **4**, e05224 (2015).

52. Kearney, S. M. *et al.* Endospores and other lysis-resistant bacteria comprise a widely shared core community within the human microbiota. *ISME J.* **12**, 2403–2416 (2018).

53. Valles-Colomer, M. *et al*. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).

54. Moeller, A. H., Suzuki, T. A., Phifer-Rixey, M. & Nachman, M. W. Transmission modes of the mammalian gut microbiota. *Science* **362**, 453–457 (2018).

55. Engel, P. & Moran, N. A. The gut microbiota of insects – diversity in structure and function. *FEMS Microbiol. Rev.* **37**, 699–735 (2013).

56. Quinn, A. *et al*. Host-derived organic acids enable gut colonization of the honey bee symbiont Snodgrassella alvi. *Nat. Microbiol.* **9**, 477–489 (2024).

57. Brochet, S. *et al.* Niche partitioning facilitates coexistence of closely related honey bee gut bacteria. *eLife* **10**, e68583 (2021).

58. Wilde, J., Slack, E. & Foster, K. R. Host control of the microbiome: Mechanisms, evolution, and disease. *Science* **385**, eadi3338 (2024).

59. Powell, E., Ratnayeke, N. & Moran, N. A. Strain diversity and host specificity in a specialized gut symbiont of honeybees and bumblebees. *Mol. Ecol.* **25**, 4461–4471 (2016).

60. Janz, N. Ehrlich and Raven Revisited: Mechanisms Underlying Codiversification of Plants and Enemies. *Annu. Rev. Ecol. Evol. Syst.* **42**, 71–89 (2011).

61. Agosta, S. J., Janz, N. & Brooks, D. R. How specialists can be generalists: resolving the ‘parasite paradox’ and implications for emerging infectious disease. *Zool. Curitiba* **27**, 151–162 (2010).

62. Yang, Y. *et al.* Dysgonomonas macrotermitis sp. nov., isolated from the hindgut of a fungus-growing termite. *Int. J. Syst. Evol. Microbiol.* **64**, 2956–2961 (2014).

63. Dyer, F. C. & Seeley, Th. D. Colony migration in the tropical honey beeApis dorsata F. (Hymenoptera: Apidae). *Insectes Sociaux* **41**, 129–140 (1994).

64. Raffiudin, R. *et al.* The effect of land cover on the foraging behavior and pollen in the honey of the giant bee Apis dorsata in Sumatra. *Front. Bee Sci.* **2**, (2024).

65. Stewart, A. B., Sritongchuay, T., Teartisup, P., Kaewsomboon, S. & Bumrungsri, S. Habitat and landscape factors influence pollinators in a tropical megacity, Bangkok, Thailand. *PeerJ* **6**, e5335 (2018).

66. Hammer, T. J., Dickerson, J. C. & Fierer, N. Evidence-based recommendations on storing and handling specimens for analyses of insect microbiota. *PeerJ* **3**, e1190 (2015).

67. Kešnerová, L. *et al*. Disentangling metabolic functions of bacteria in the honey bee gut. *PLOS Biol*. **15**, e2003467 (2017).

68. 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).

69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

70. Ruscheweyh, H.-J. *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* **10**, 212 (2022).

71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

72. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

73. Kang, D. D. *et al*. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

74. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

75. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

76. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).

77. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

78. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

79. Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E. J. P. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE* **6**, e22594 (2011).

80. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

81. Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc. Natl. Acad. Sci.* **96**, 12638–12643 (1999).

82. Su, Y.-C., Chiu, Y.-F., Warrit, N., Otis, G. W. & Smith, D. R. Phylogeography and species delimitation of the Asian cavity-nesting honeybees. *Insect Syst. Divers.* **7**, 5 (2023).

83. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

84. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

85. Buck, M., Mehrshad, M. & Bertilsson, S. mOTUpan: a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. *NAR Genomics Bioinforma.* **4**, lqac060 (2022).

86. Carr, S. M. Multiple mitogenomes indicate Things Fall Apart with Out of Africa or Asia hypotheses for the phylogeographic evolution of Honey Bees (Apis mellifera). *Sci. Rep.* **13**, 9386 (2023).

87. The scikit-bio development team. scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. (2020).

88. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

89. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb. Genomics* **8**, 000823 (2022).

90. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

91. Ducarmon, Q. R. *et al.* Large-scale computational analyses of gut microbial CAZyme repertoires enabled by Cayman. 2024.01.08.574624 Preprint at https://doi.org/10.1101/2024.01.08.574624 (2024).

92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

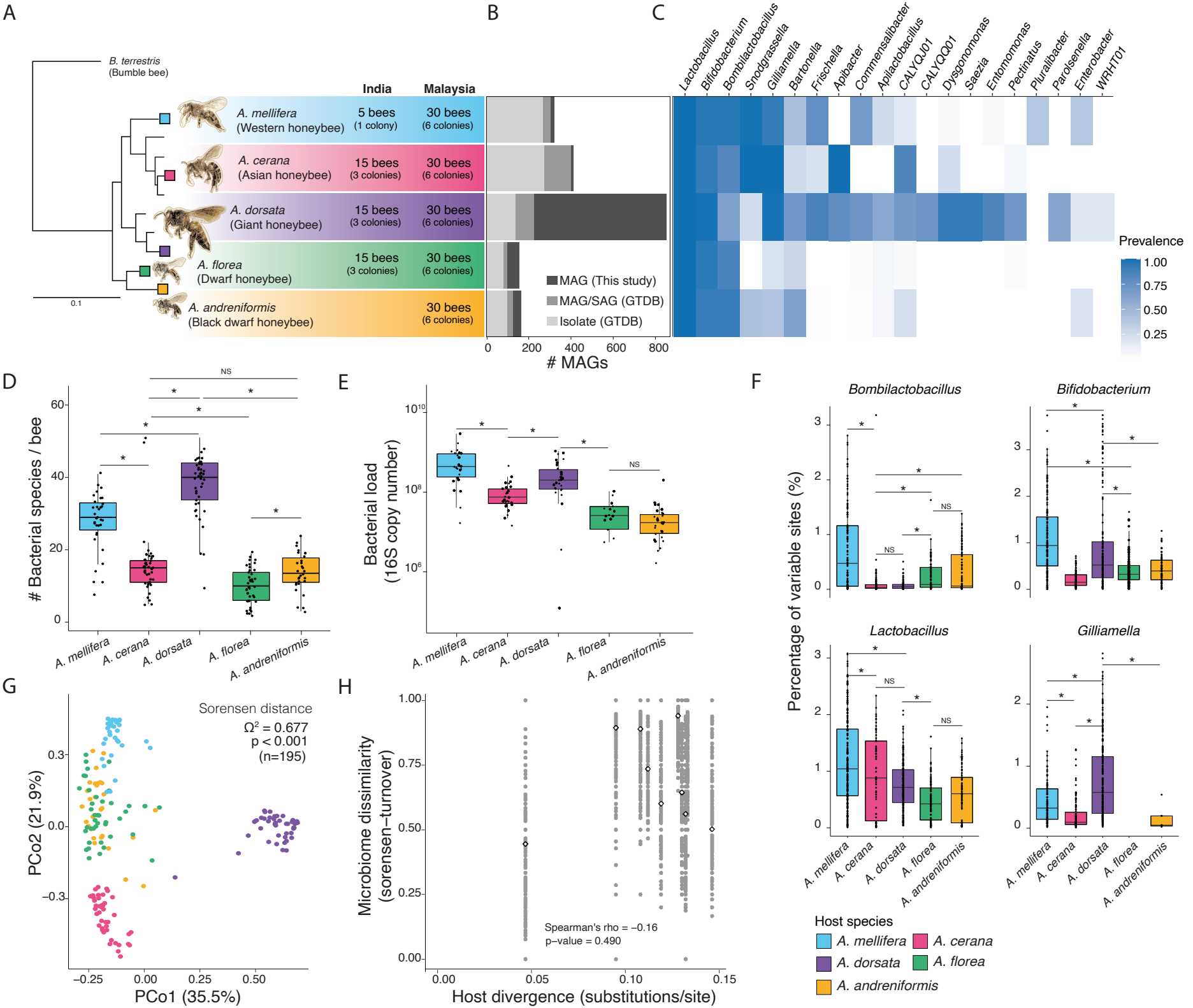
93. Ye, Y. & Doak, T. G. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLOS Comput. Biol.* **5**, e1000465 (2009).

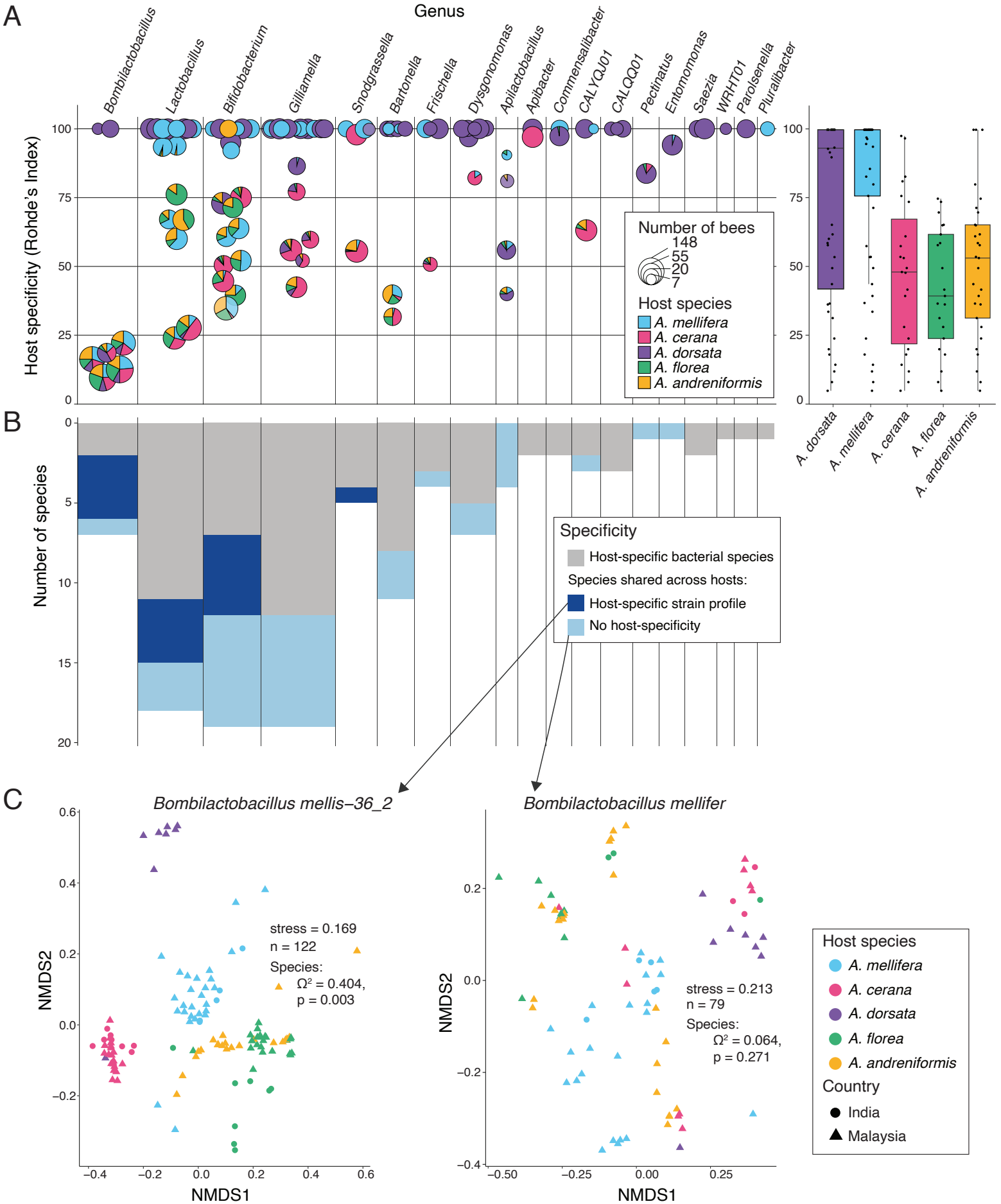
94. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLOS Comput. Biol.* **17**, e1009442 (2021).

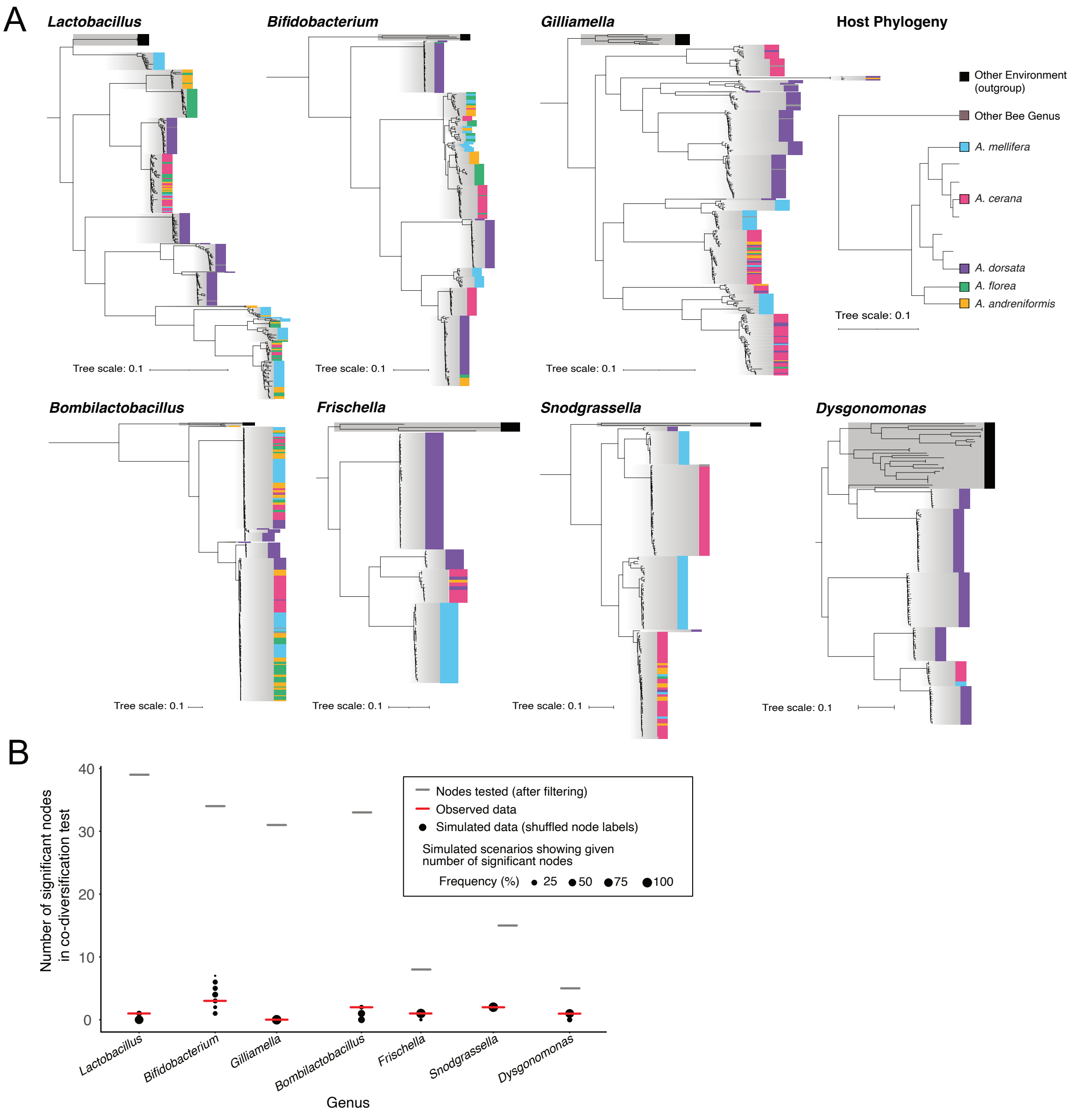
95. Baselga, A. & Orme, C. D. L. betapart: an R package for the study of beta diversity. *Methods Ecol. Evol.* **3**, 808–812 (2012).

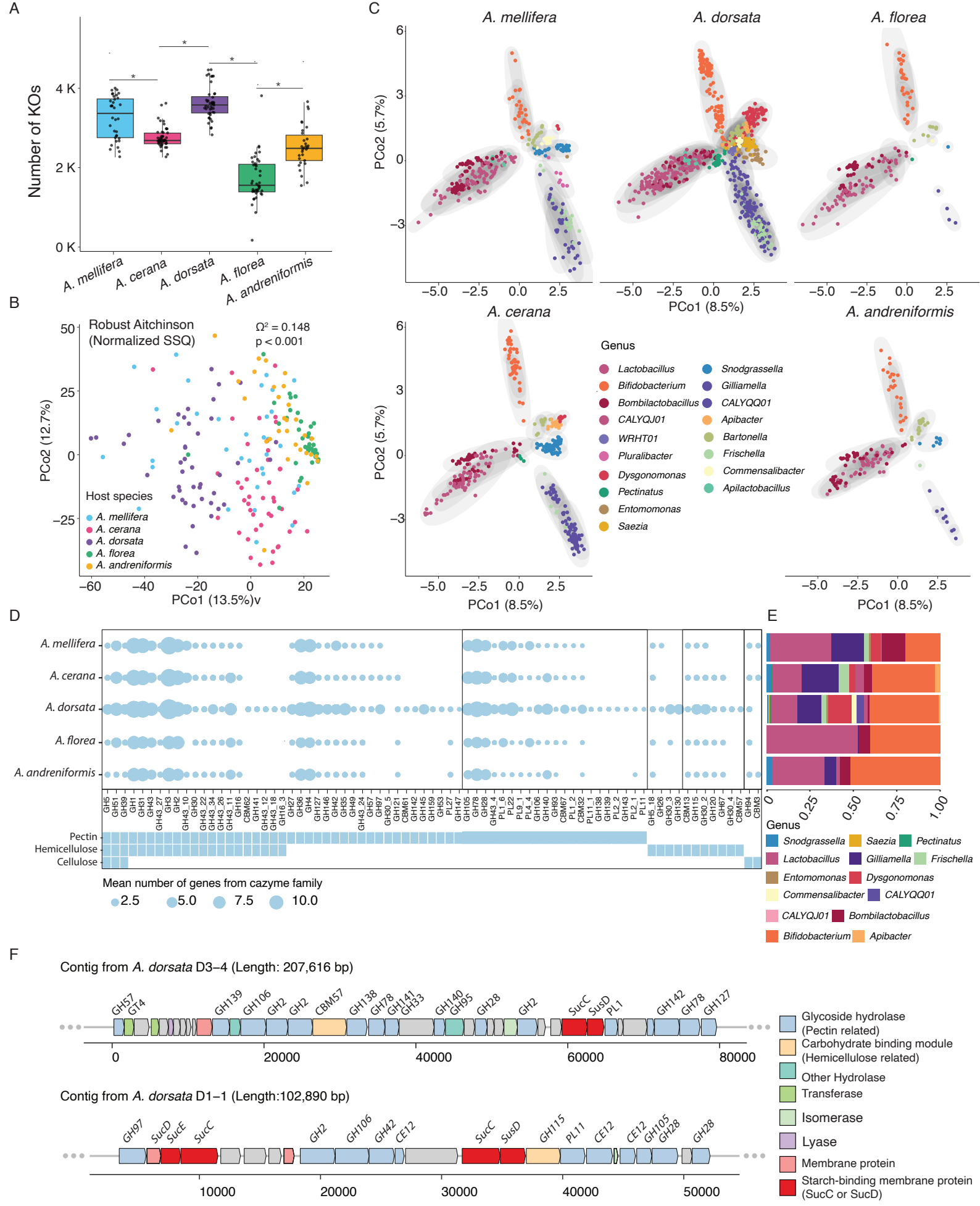
96. Kotthoff, U., Wappler, T. & Engel, M. Miocene honey bees from the Randeck Maar of southwestern Germany (Hymenoptera, Apidae). *ZooKeys* **96**, 11–37 (2011).

97. Ghonche-Golan, S., Nazemi-Rafie, J. & Rezapanah, M. The relationship study among Apis spp. using mitochondrial markers, Procrustes coordinates and residuals of geometric morphometric method. *Biologia (Bratisl.)* **77**, 2519–2530 (2022).


A
B. terrestris
(Bumble bee)
India
Malaysia
A. mellifera
(Western honeybee)
5 bees
(1 colony)
30 bees
(6 colonies)
A. cerana
(Asian honeybee)
15 bees
(3 colonies)
30 bees
(6 colonies)
A. dorsata
(Giant honeybee)
15 bees
(3 colonies)
30 bees
(6 colonies)
A. florea
(Dwarf honeybee)
15 bees
(3 colonies)
30 bees
(6 colonies)
A. andreniformis
(Black dwarf honeybee)
30 bees
(6 colonies)
0.1
B
MAG (This study)
MAG/SAG (GTDB)
Isolate (GTDB)
0
200
400
600
800
MAGs
C
Lactobacillus
Bifidobacterium
Bombilactobacillus
Snodgrassella
Gilliamella
Bartonella
Frischella
Apibacter
Commensalibacter
Apilactobacillus
CALYQJ01
CALYQQ01
Dysgonomonas
Saezia
Entomomonas
Pectinatus
Pluralibacter
Parolsenella
Enterobacter
WRHT01
Prevalence
1.00
0.75
0.50
0.25
D
Bacterial species / bee
60
40
20
0
NS
*
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
E
Bacterial load
(16S copy number)
10^{10}
10^{8}
10^{6}
*
NS
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
F
Bombilactobacillus
Bifidobacterium
Lactobacillus
Gilliamella
Percentage of variable sites (%)
3
2
1
0
*
NS
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
Host species
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
G
Sorensen distance
$\Omega^2 = 0.677$
$p < 0.001$
(n=195)
PCo2 (21.9%)
0.3
0.0
−0.3
−0.25
0.00
0.25
0.50
PCo1 (35.5%)
H
Microbiome dissimilarity
(sorensen−turnover)
1.00
0.75
0.50
0.25
0.00
Spearman's rho = −0.16
p−value = 0.490
0.00
0.05
0.10
0.15
Host divergence (substitutions/site)


A
Genus
Bombilactobacillus
Lactobacillus
Bifidobacterium
Gilliamella
Snodgrassella
Bartonella
Frischella
Dysgonomonas
Apilactobacillus
Apibacter
Commensalibacter
CALYQJ01
CALQQ01
Pectinatus
Entomomonas
Saezia
WRHT01
Parolsenella
Pluralibacter
Host specificity (Rohde's Index)
Number of bees
148
55
20
7
Host species
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
B
Number of species
Specificity
Host-specific bacterial species
Species shared across hosts:
Host-specific strain profile
No host-specificity
C
Bombilactobacillus mellis–36_2
Bombilactobacillus mellifer
stress = 0.169
n = 122
Species:
Ω² = 0.404,
p = 0.003
stress = 0.213
n = 79
Species:
Ω² = 0.064,
p = 0.271
NMDS1
NMDS2
Host species
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
Country
India
Malaysia


A
Lactobacillus
Bifidobacterium
Gilliamella
Host Phylogeny
Other Environment (outgroup)
Other Bee Genus
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
Tree scale: 0.1
Bombilactobacillus
Frischella
Snodgrassella
Dysgonomonas
B
Number of significant nodes in co-diversification test
Nodes tested (after filtering)
Observed data
Simulated data (shuffled node labels)
Simulated scenarios showing given number of significant nodes
Frequency (%) 25 50 75 100
Lactobacillus
Bifidobacterium
Gilliamella
Bombilactobacillus
Frischella
Snodgrassella
Dysgonomonas
Genus


A
Number of KOs
4 K
2 K
0 K
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
B
Robust Aitchinson
(Normalized SSQ)
Ω² = 0.148
p < 0.001
PCo2 (12.7%)
PCo1 (13.5%)v
Host species
A. mellifera
A. cerana
A. dorsata
A. florea
A. andreniformis
C
A. mellifera
A. dorsata
A. florea
A. cerana
A. andreniformis
PCo2 (5.7%)
PCo1 (8.5%)
Genus
Lactobacillus
Bifidobacterium
Bombilactobacillus
CALYQJ01
WRHT01
Pluralibacter
Dysgonomonas
Pectinatus
Entomomonas
Saezia
Snodgrassella
Gilliamella
CALYQQ01
Apibacter
Bartonella
Frischella
Commensalibacter
Apilactobacillus
D
Pectin
Hemicellulose
Cellulose
Mean number of genes from cazyme family
2.5
5.0
7.5
10.0
E
Genus
Snodgrassella
Saezia
Pectinatus
Lactobacillus
Gilliamella
Frischella
Entomomonas
Dysgonomonas
Commensalibacter
CALYQQ01
CALYQJ01
Bombilactobacillus
Bifidobacterium
Apibacter
F
Contig from A. dorsata D3–4 (Length: 207,616 bp)
Contig from A. dorsata D1–1 (Length:102,890 bp)
Glycoside hydrolase (Pectin related)
Carbohydrate binding module (Hemicellulose related)
Other Hydrolase
Transferase
Isomerase
Lyase
Membrane protein
Starch-binding membrane protein (SucC or SucD)