# What lurks in the dark? - An innovative framework for studying diverse wild insect microbiota

Karol H. Nowak[1,2*], Emily Hartop[3], Monika Prus-Frankowska[1], Mateusz Buczek[1], Michał Kolasa[1,4], Tomas Roslin[5,6], Otso T. Ovaskainen[7,8], Piotr Łukasik[1*]

1) Institute of Environmental Sciences, Faculty of Biology, Jagiellonian University, Kraków, Poland
2) Doctoral School. PhD Programme in Biology, Faculty of Biology, Jagiellonian University, Kraków, Poland
3) Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, NO-7491, Norway
4) Department of Farm Animal Biodiversity Conservation and Horse Breeding, National Research Institute of Animal Production, Balice, Poland
5) Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland
6) Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden
7) Department of Biological and Environmental Science, Faculty of Mathematics and Science, University of Jyväskylä, Jyväskylä, Finland
8) Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland

* To whom correspondence should be addressed.

E-mails: karol.h.nowak@doctoral.uj.edu.pl, p.lukasik@uj.edu.pl

All supporting data available at: https://github.com/KarolHub/Phorid-Microbiota.

# Abstract

Symbiotic microorganisms can profoundly impact insects, their life history traits, population dynamics, and evolutionary trajectories. However, microbiota remain poorly understood in natural insect communities, especially in ‘dark taxa’ - hyperdiverse, yet understudied clades.

Here, we implemented a novel multi-target amplicon sequencing approach to study microbiota in complex, species-rich communities. It combines four methodological innovations: (1) To establish a host taxonomic framework, we sequenced amplicons of host marker gene (COI) and reconstructed barcodes alongside microbiota characterisation. (2) To assess microbiota abundance, we incorporated synthetic spike-in-based quantification. (3) To improve the phylogenetic resolution for the dominant endosymbiont, *Wolbachia*, we analysed bycatch data from the COI amplicon sequencing. (4) To investigate the primary drivers of host-microbe associations in massive multi-dimensional dataset, we performed HMSC modelling.

Applying this approach to 1,842 wild-caught scuttle flies (Diptera: Phoridae) from northern Sweden, we organised them into 480 genotypes and 186 species, and gained unprecedented insights into their microbiota. We found orders-of-magnitude differences in bacterial abundance and massive intrapopulational variation in microbiota composition. Differences emerged in patterns and drivers among microbial functional categories: the distribution and abundance of facultative endosymbionts (*Wolbachia*, *Rickettsia*, *Spiroplasma*) were shaped by host species, genotype and sex. In contrast, many other bacterial taxa were broadly distributed across species and sites.

This study highlights facultative endosymbionts as key players in insect microbiota and reveals striking variation in distributional patterns of microbial clades. It also demonstrates the power of integrative sequencing approaches in uncovering the ecological complexity and significance of symbiotic microorganisms in multi-species natural communities.

**Keywords:** Arthropods, microbiome, endosymbiosis, next-generation sequencing, holobiont, Hierarchical Modelling of Species Communities, HMSC, joint species distribution models

# 1. Introduction

Microbial symbioses have played tremendously important roles in the biology and evolution of higher organisms [1]. The many levels at which different functional categories of symbionts have shaped hosts' biology are relatively well-understood in insects [2]. The range of microbial symbionts' effects on insect life-history traits - nutrition, reproduction, and defence [3,4] translate to population processes and patterns [5–7], host evolutionary trajectories [8,9], and effects at the level of multi-species communities [10]. There is little doubt that through such combined effects, microbiota have shaped the global biodiversity, and will continue to play key roles in ecosystems rapidly changing under anthropogenic pressures [11,12].

Despite their massive importance, insect microbiota are poorly known outside of model species. Some biological systems, including honey bees [13] or pea aphids [14] have been the focus of extensive research on symbiosis diversity, function, mechanisms, and various other aspects. However, for some 6 million of the remaining insect species on our planet [15,16], our understanding of microbiota is non-existent or, at best, very limited. Among the primary reasons is the sheer diversity of insects and microorganisms, which poses a significant challenge in generating and handling vast amounts of multidimensional ecological data. We lack reliable and cost-effective workflows for processing large numbers of wild-caught insects, encompassing host identity validation and microbiome characterisation, while including controls for different types of contamination [17–19]. Additionally, existing approaches rarely consider and address questions about microbial abundance, which is critical, as insects differ dramatically in microbial load [20]. Data analysis is another neglected area, as the value of conventional approaches, including diversity estimation and comparison, is questionable when microbial communities are dominated by one or few strains or when

the abundance is low and the contamination is not controlled. Finally, analytical solutions for the reconstruction of microbiome patterns and drivers in highly multidimensional datasets are lacking.

As a result, we know very little about the microbial composition and abundance across the arthropod diversity. Host biology, especially the diet, has been identified as an important driver [8], alongside host taxonomy [21]. However, even within a species, there can be major differences among populations, within a population over time, and among individuals sampled at the same time from a single population [22]. Similarly, the differences in microbiota between host sexes have rarely been considered outside of the context of reproductive manipulation [23,24]. Interactions among these and other possible microbiota drivers can make analyses especially challenging. At the same time, we would anticipate that microbial taxa and functional categories would differ in how the infections are determined. A systematic assessment of these and other host-microbe interaction drivers would require a combination of comprehensive laboratory methods applied at a large scale and robust data analysis approaches, but also suitable study organisms.

Much of the global insect diversity is made up of so-called ‘dark taxa’: insect clades that are hyperdiverse, abundant, and broadly distributed yet remain poorly studied [25]. The study of dark taxa has greatly benefitted from recent advances in molecular workflows [26,27], particularly in Megabarcoding - high-throughput barcoding of single specimens [28], which has enabled large-scale analyses of hyperdiverse insects on a broader scale than previously possible [29]. Growing barcoding data provide a promising opportunity to illuminate the hidden microbial diversity across these enigmatic clades.

Scuttle flies (Diptera: Phoridae) are a quintessential example of a “dark taxon”: they are hyperdiverse, abundant, and globally distributed but remain poorly understood [25,30]. They exhibit diverse life strategies, including parasitism, predation, herbivory, and saprophagy [31]. Such varied life histories allow them to occupy numerous ecological niches, from bustling urban centres [32] and subterranean caves [33] to intertidal zones [34] and the cavities of pitcher plants [35]. Despite their ubiquity and diverse ecological roles, our understanding of their diversity, distribution, and biology

remains grossly incomplete - highlighted by recent reporting of more species captured over eight weeks at a single Ugandan site than had been previously described from the entire Afrotropical bioregion [26]. Given our poor state of knowledge about scuttle flies generally, it is not surprising that little is known about their interactions with microbial symbionts, with the few studies that addressed questions about these associations limited to single-species investigations with shallow sampling depth [36,37].

In this study, we leverage a novel combination of molecular and analytical approaches to comprehensively characterise microbiota abundance, composition, and distribution in diverse scuttle flies as a model dark taxon across six sites in Northern Sweden. We employ high-resolution COI amplicon sequencing for host identification, bacterial 16S-V4 rRNA sequencing to resolve microbiota composition, using synthetic spike-ins to quantify microbial abundance, and state-of-the-art Hierarchical Modelling of Species Communities (HMSC) to disentangle the ecological drivers of microbial variation. By integrating these innovative techniques, we assess how microbial communities are shaped by site, season, host species identity, and individual host traits such as sex and genotype. This innovative methodological pipeline provides an unprecedented resolution of host-microbiota interactions in a hyperdiverse insect group, advancing our understanding of the ecological and evolutionary forces shaping microbial associations in wild communities.

# 2. Materials and methods

## *2.1. Field collections*

Samples were collected using Malaise traps as a part of the Swedish Insect Inventory Project (SIIP) [38]. The collection was carried out in northern Sweden at six sites, separated by distances varying from less than 50 m to about 390 km (for exact locations, see Supplementary Table S1 and Fig. 1). Selected samples come from late spring (June to July) and summer (July to August) of 2018

(Supplementary Table S1). After collection, samples in 95% ethanol were stored at -20 °C in Station Linné, Sweden, where taxonomists manually sorted the scuttle flies from the samples. We randomly selected 88 males and 88 females for processing from each of the six sites and two seasons (except for three combinations where only 87, 86 and 84 individuals were available; see Supplementary Table S1 for details), totalling 2,105 insects.

## *2.2 DNA extraction, amplicon library preparation and sequencing*

The insects were processed and prepared for sequencing in the Molecular Ecology Laboratory at the Institute of Environmental Sciences of Jagiellonian University, following a protocol comprehensively described recently [39]. First, each insect was placed in a 2 ml tube with 190 µl of the lysis buffer (0.4 M NaCl, 10 mM Tris-HCl, 2 mM EDTA, 2% SDS), 10 µl proteinase K and different-sized ceramic beads. After two rounds of 30-second homogenisation on a homogenizer (Bead Ruptor Elite, Omni International), the samples were incubated for 2 hours at 55 °C. Subsequently, 40 µl aliquot of the homogenate was combined with 2 µl of the solution containing 10,000 or 20,000 copies (depending on the batch) of plasmids carrying artificial 16S rRNA amplification targets, Ec5502 [40], placed in a 96-well PCR plate, and purified using SPRI beads [41]. DNA was eluted in 20 µl of TE buffer.

The amplicon libraries were prepared using a two-step PCR procedure (see method 4 [42]). In the first step, two marker regions of interest were simultaneously amplified using QIAGEN Multiplex Master Mix and template-specific primers with variable-length inserts and Illumina adapter stubs. Specifically, for the V4 region of bacterial 16S rRNA, we used primers 515F: GTGYCAGCMGCCGCGGTAA and 806R: GGACTACNVGGGTWTCTAAT [43], and for a 418 bp portion of the mitochondrial cytochrome oxidase I (COI) gene, primers BF3: CCHGAYATRGCHTTYCCHCG and BR2: TCDGGRTGNCCRAARAAYCA [44]. The PCR reaction mix also contained 500 copies of another plasmid carrying an artificial 16S rRNA target,

Ec5001 [40], but it was not used in this study and subsequently filtered out from the data. SPRI bead-purified first-step PCR products were used as a template for the second indexing PCR. The libraries were verified on an agarose gel and based on band intensities pooled approximately equimolarly. The purified pool was quality checked using BioAnalyzer 2100 (Agilent), combined with pools representing other projects, and submitted to Swedish National Genomics Infrastructure for sequencing on Illumina NovaSeq SPrime flow cells (2x250 bp read mode). To control for contamination from different sources, for each of the 88 biological samples processed, we used eight different controls: negative controls for the DNA extraction (3), first PCR (2), and indexing PCR stages (1), as well as positive controls (2). Raw sequence data for all samples that passed filtering criteria and controls are available under GenBank BioProject accession number PRJNA1099311.

## *2.3. Bioinformatic analysis of amplicon data*

The amplicon data were processed using scripts and files available at the GitHub repository, together with supplementary data: https://github.com/KarolHub/Phorid-Microbiota.

### *2.3.1. Basic read sorting and processing*

Initially, we split the demultiplexed reads into batches corresponding to COI and 16S-V4 targets using a custom Python script (Multisplit.py), which divided the files based on primer sequences and trimmed the primers. Further steps relied on the custom LSD.py script, a wrapper tool executing several steps using PEAR [45], VSEARCH [46], and USEARCH [47] programs. Specifically, we merged forward and reverse reads into contigs, discording those outside of the expected length range (400-430 for COI and 250-260 bp for 16S-V4), or with lower quality regions (Phred Score $\leq$ 30). Datasets representing both regions were then dereplicated to retain only unique sequences per sample and their occurrence counts, with singleton genotypes discarded. We denoised each library separately using the UNOISE algorithm [48]. Then, we clustered the zero-radius operational taxonomic units (ZOTUs) into operational taxonomic units (OTUs) using a 97% identity cutoff. Finally, we assigned

taxonomy using the SINTAX algorithm [49], and taxonomic reference databases: MIDORI version GB 239 [50], with alphaproteobacterial sequences added, for the COI data and SILVA [51] version 138 SSU for 16S rRNA data. We retained identification to the lowest level that had at least an 80% classification confidence score.

### *2.3.2. COI data analysis*

The COI data contained a mix of sequences of different origins, including those representing insect mitochondrial genes, nuclear pseudogenes, alphaproteobacterial symbionts, and others. To obtain insect barcodes, we discarded sequences shorter or longer than the anticipated 418 nucleotides, as well as bacterial, non-arthropod, unassigned, and chimeric sequences. Next, we assessed the number of reads in the most abundant OTU classified as an arthropod and calculated its ratio relative to all Arthropoda OTUs combined. We excluded libraries from further analyses if they had fewer than 300 reads of the dominant OTU or if the aforementioned read ratio was less than 95%. Then, we designated the most abundant sequence within the dominant OTU of each sample as the individual's barcode, dereplicated these barcodes, and clustered them with sequences from earlier barcoding studies [25] to generate a dendrogram. Given the complex taxonomy of phorid flies, we manually reviewed the proximity of barcodes and sequences of established phorid species within the dendrogram to determine which groups of barcodes form distinct species-level clades. Once we had the sequences divided into species-level clusters of barcodes, we compared them with sequences of named species from the same study [25], to obtain the taxonomy labels. The species names were assigned if a sequence matched the reference by at least 97%. Variants with no close matches were compared against the NCBI nt database using the BLAST algorithm, retaining labels with a match of 97% or higher.

The COI data contained sequences assigned to the order Rickettsiales, which we separated for further analyses. We filtered out non-*Wolbachia* sequences, chimeric sequences, and those of length other than 418 bp. This resulted in tables on *Wolbachia* COI ZOTU and OTU read number distribution among samples, analogous to standard bacterial 16S rRNA-based tables.

### *2.3.3. 16S rRNA amplicon data filtering and abundance estimation*

We decontaminated the 16S-V4 data from 1,845 samples with reliable insect COI barcodes using a custom QUACK.py script. This script classified as contaminants and removed ZOTUs unless the percentage of reads representing that ZOTU in at least one experimental sample was at least five times greater than the maximum percentage in any of the blank samples [52,53]. It also removed sequences classified as Archaea, Eukaryota, chloroplasts, mitochondria and chimaeras. After calculating the ratio between the spike-in and remaining symbiont reads in each sample, the spike-in sequences were also removed. Afterwards, we calculated the absolute abundance of non-contaminant bacterial 16S rRNA copies in each sample by multiplying the symbiont to extraction spike-in read number ratio, the number of inserted extraction spike-in copies (10,000 or 20,000), and the inverse proportion of insect homogenate used during the DNA extraction (5). The abundance was estimated separately for each bacterial ZOTU, with the total bacterial abundance in the sample estimated as the summed abundance of individual ZOTUs. Three samples without extraction spike-in reads were discarded, along with control samples, resulting in the final total number of phorid samples of 1842. For subsequent comparisons and modelling, we used tables on bacterial OTU and ZOTU estimated 16S rRNA copy number distributions across samples.

### *2.3.4. Analyses*

To examine the drivers of variation in insect microbiomes we used Hierarchical Modelling of Species Communities (HMSC; [54,55], a method of joint species distribution modelling (JSDM; [56]). We modelled 75 species that were represented by at least 5 individuals for a total of 1634 samples. We fitted separate HMSC models for each of three response variables: (1) estimated copy numbers for the 50 most abundant 16S-V4 OTUs, (2) estimated copy numbers for the 100 most dominant 16S-V4 ZOTUs, and (3) read counts of the 24 most dominant Wolbachia COI ZOTUs. For covariates, we used effects of season (spring vs. summer), site (n = 6), sex of the host (male vs. female), host species (n =

75), host genotype (n = 328), and sampling unit (n = 1,634). This approach allowed us to quantify how these ecological and host-related factors shape microbial community composition and abundance across a hyperdiverse insect assemblage.

Because our data were zero-inflated—with many taxa being absent from individual samples, resulting in numerous zero counts—we used a hurdle modelling approach to separately analyse presence-absence and abundance. First, we modelled the presence-absence of each taxon using a Bernoulli distribution with a probit link function. Then, for taxa that were present, we modelled abundance conditional on presence by log-transforming and scaling the response, assuming a normal distribution. In our models, season and sex were included as fixed effects, while site, host species, host genotype, and individual host insect were treated as community-level random effects, structured using latent variables to account for shared variation across taxa. This approach allowed us to better capture the ecological and host-related factors influencing microbiota distributions.

We fitted the model using the Markov Chain Monte Carlo (MCMC) approach implemented in the R-package HMSC [57] with default prior distributions [55]. The posterior distribution was sampled with four chains, with 375 * thin MCMC iterations per chain. We ignored the first 125 * thin iterations as burn-in and thinned the remaining iterations by thin to yield 250 samples per chain and hence 1000 samples in total. The thinning parameter was increased as thin = 1, 10, 100, … until satisfactory MCMC convergence was reached. We considered MCMC convergence satisfactory if the third quartile of the potential scale reduction factors over the parameters measuring the responses of the taxa to season and sex was at most 1.05; this was reached with thin = 1000 for 16S-V4 ZOTU and *Wolbachia* COI ZOTU presence-absence models, and with thin = 100 for the other four model components.

To evaluate model fit, we focused on explanatory and predictive powers – the former based on fitting the models to all data and the latter through two-fold cross-validation. As metrics for evaluation, we used AUC for presence-absence models and $R^2$ for models on abundance conditional on presence.

To examine the importance of the fixed and random effects in structuring microbiome variation, we applied a variance partitioning approach to the explained variation. In other words, we examined what proportion of the overall variation explained by the model could be attributed to the individual factors and covariates. To examine the effects of sex and season, we counted the proportion of taxa that showed either a positive or negative response with at least 95% posterior probability. To examine patterns of co-variation not explained by sex or season, we examined which pairs of bacterial taxa showed a positive or negative association for each random effect (site, species, genotype and individual) with at least 90% posterior probability. The latter approach will reveal what microbial taxa appear to covary in their presence and abundance beyond what can be expected based on the fixed effects of sex and season.

# 3. Results

After stringent quality filtering of 2,105 amplicon sequencing libraries for COI and 16S-V4 marker regions of individual phorid flies, we retained on average 77 individuals for each of the 24 site-season-sex combinations: 936 males and 906 females, or 1,842 individuals in total (Supplementary Table S1). For individuals passing the filtering criteria, we obtained, on average, 27.1K (K = thousand) COI reads (min 373, max 191.9K, median 20.4K) (Supplementary File S1). We then used these data to reconstruct insect barcodes and classify specimens to species (Supplementary Table S2; see section 3.1), but also to reconstruct and compare COI sequences of *Wolbachia* (see section 3.3). We obtained, on average, 29.7K 16S-V4 reads (range 682 - 295.7K, median 19.5K) per experimental library, but after filtering and removal of putative laboratory and reagent-derived contaminants and removing spike-in control reads, we retained, on average, 15.2K 16S-V4 non-contaminant reads (range 29 - 227K, median 5.4K) (Supplementary Table S3). We used these data to reconstruct microbial abundance and diversity across individuals (section 3.2). For statistical modelling of the data (section

3.4) we used all data types listed above, focusing on the 75 abundant species (for specific cut-offs, see Methods).

### *3.1. Abundance, diversity and distribution of phorid flies*

Of the COI amplicon data for individual phorid flies, around 99.5% of reads were classified as matching insects. On average, the most abundant insect ZOTU in each library, designated as the individual's barcode, accounted for 98.8% of the filtered insect reads (min 52.2%, max 100%, median 99.5%). The COI barcodes of 1,842 phorid individuals represented 480 genotypes, which we grouped into 186 species (Supplementary Table S2) through a comparison with previously sequenced and morphologically characterised material [25]. These species varied in genotype numbers and abundance (Fig. 1a). For example, the five most abundant species were represented by between 10 and 33 genotypes each (85 genotypes in total), but 59 of these genotypes were represented by a single individual only. Many of these low-abundance genotypes differed from the more abundant genotypes by a single nucleotide. 142 of the COI genotypes, representing 438 individuals, were confidently assigned to 63 recognised, named species, mostly within the genus *Megaselia*. A further 42 genotypes were assigned to two species complexes: the *Megaselia pulicaria* complex (132 individuals) and the *Megaselia giraudii* complex (82 individuals). Most of the samples (295 genotypes, 1176 individuals) were assigned to the genus level, and one genotype consisting of 14 individuals was assigned to the family level only. Given the low prevalence of individual genotypes, there was little indication of geographic structuring within species (Fig. 1, Supplementary Table S2). To allow unequivocal reference to each species (whether recognised or not), we assigned unique numbers to all species - generally, numbering them consecutively after sorting them by decreasing abundance, although subsequent changes to the sample set caused some discrepancies. We are referring to species by these numbers throughout the text (e.g., "sp. 1" and "sp. 15" - rather than "*Megaselia* sp. 001" and "*Megaselia hendersoni*") (Fig. 1a, Supplementary Table S2).

Species' representation in the dataset was highly skewed, with a few abundant species and a long tail of rare ones (Supplementary Table S2). Within the same season, we found an overlapping set of abundant species across all sites (Fig. 1b). For example, the five most abundant species were found at all six sites. At the same time, there was little overlap in community composition among seasons, with most species well-represented either in spring or in the summer (Fig. 1b). NMDS based on Bray-Curtis distances clearly separated spring and summer community samples (Supplementary Fig. S1).

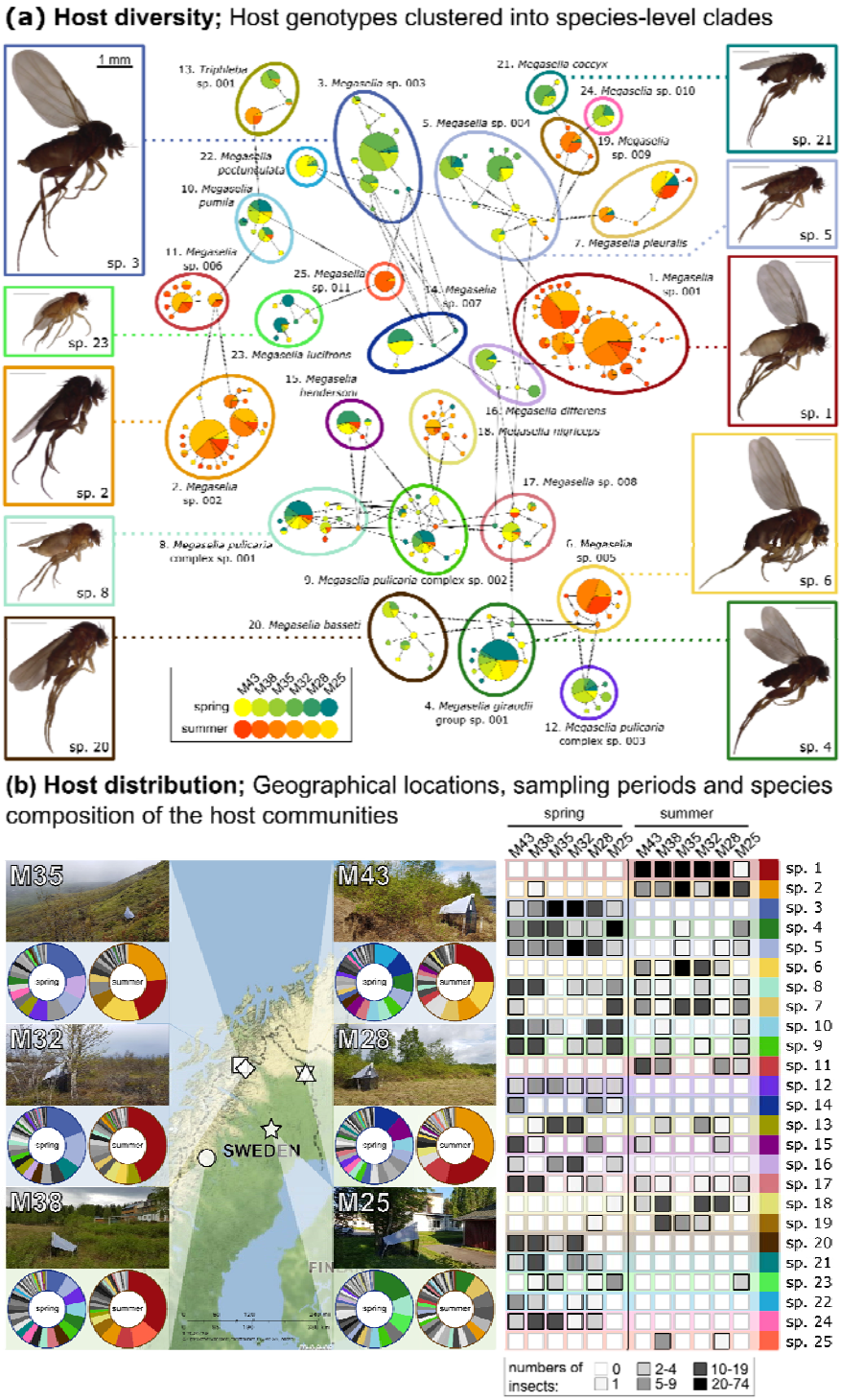



**Fig. 1:** (**a**) Haplotype network of the 25 most abundant scuttle fly species highlights the diversity among the sampled insects and showcases challenges with their delimitation. Individual pie charts represent fly COI genotypes, coloured by collection site and season. Selected species are represented

by photos of individually barcoded individuals from various sites, with all images scaled uniformly for consistency. Genotypes of the remaining 161 species are not shown. The network was drawn using PopArt [58]. (**b**) The distribution and diversity of the 25 most commonly sampled scuttle fly species from Northern Sweden highlights community composition variation among the two seasons and overlap among the six sampled sites.

### *3.2. Abundance and composition of phorid microbiota*

Across 1,842 phorid samples, a total of 28M (M = million) reads scored as non-contaminants and represented 58.1K ZOTUs that clustered into 9.6K 97% OTUs (Supplementary File S3). Using the ratio of quantification spike-in to non-contaminant reads as an estimate of the abundance of specific bacterial ZOTUs, OTUs, and total bacteria [39,40], we were able to assess and interpret microbial abundance and diversity at the same time.

In terms of microbe abundance, we observed high variation in the symbiont to spike-in ratios across the samples. The ratio ranged between 0.015 and 354.8 (median 1.38), and with 10K or 20K artificial template spike-in copies added to one-fifth of insect homogenate before purification, these ratios translate to bacterial 16S rRNA copy numbers per insect ranging between 766 and 24M (median 815.4K) (Supplementary File S3). Even among individuals representing the same sex and species, the range in 16S rRNA copy numbers typically spanned between two to four orders of magnitude. At the same time, microbial abundance was often higher in females than males of the same species, by about an order of magnitude on average (Fig. 2a).

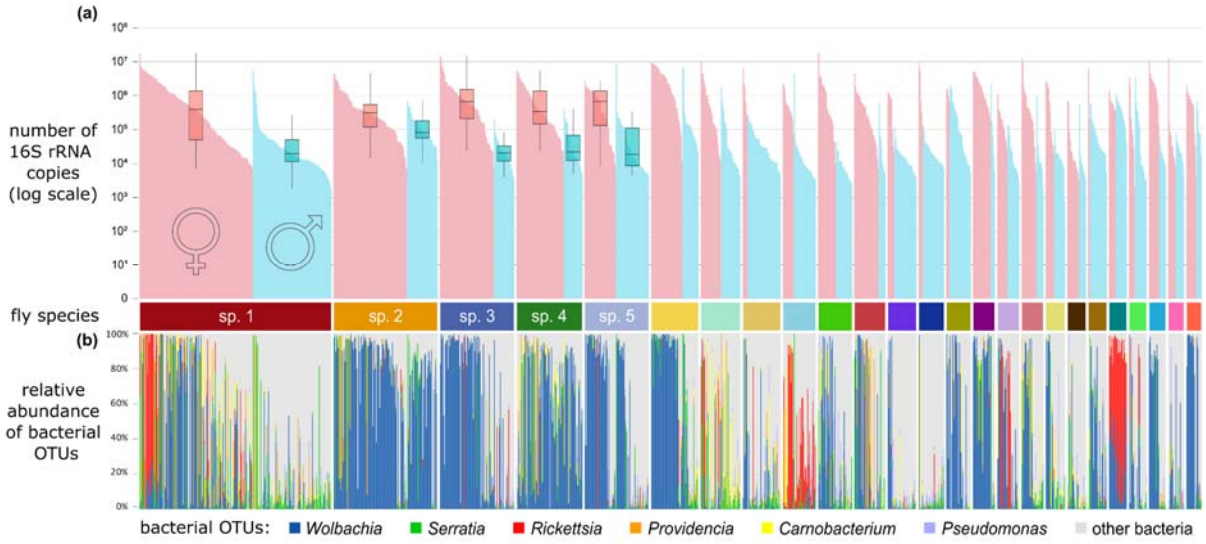



**Fig. 2:** Simultaneous inspection of bacterial absolute abundance and microbial diversity data indicates strong variation across 1,193 specimens of the 25 dominant scuttle fly species. (**a**) The distribution of bacterial 16S-V4 rRNA copy numbers across individual insects, estimated based on quantification spike-ins added during the DNA extraction step. Here, each individual fly is represented by a narrow vertical bar, with bars sorted first by species assignment, then by sex, and then by abundance. (**b**) The relative abundances of the six most abundant bacterial 97% OTUs across individual insects sorted as in panel **a**. The category 'other bacteria' shows all the remaining bacterial OTUs combined.

Although the number of detected 16S-V4 97% OTUs was high, scuttle fly communities were dominated by relatively few bacterial clades (Supplementary Table S4). The most common OTU1, taxonomically assigned to the genus *Wolbachia*, was represented by 30.3% of decontaminated bacterial reads in a library. The next-most prevalent symbiont OTUs, *Serratia*, *Rickettsia*, *Providencia*, *Carnobacterium*, and *Pseudomonas*, comprised an average of 5.5%, 5.1%, 3.0%, 2.9%, and 2.7% of reads in a library, respectively. We identified multiple additional, less abundant bacterial OTUs representing the same six genera plus a long list of other genera (Supplementary Table S4). These OTUs were broadly distributed across species, but their abundances showed non-random distribution

patterns among individuals within species (Table 1; Fig. 2b). For example, *Rickettsia* comprised a particularly large share of the microbiota of females of species 1 with high overall bacterial abundance.

The analysis of abundance and distribution of dominant bacterial clades revealed large variation among individuals, even of the same species, population, and sex - as well as the complex interactions among likely drivers of microbial abundance and composition (Fig. 3). For example, when present, *Rickettsia* can be a quantitatively dominant microbe in females of phorid sp. 1, but its prevalence varies widely among populations. At a specific site (M38), *Providencia* is common and abundant in both sexes of phorid sp. 1 and sp. 2, but it is much less common or abundant at other sites or in co-occurring phorid sp. 3.

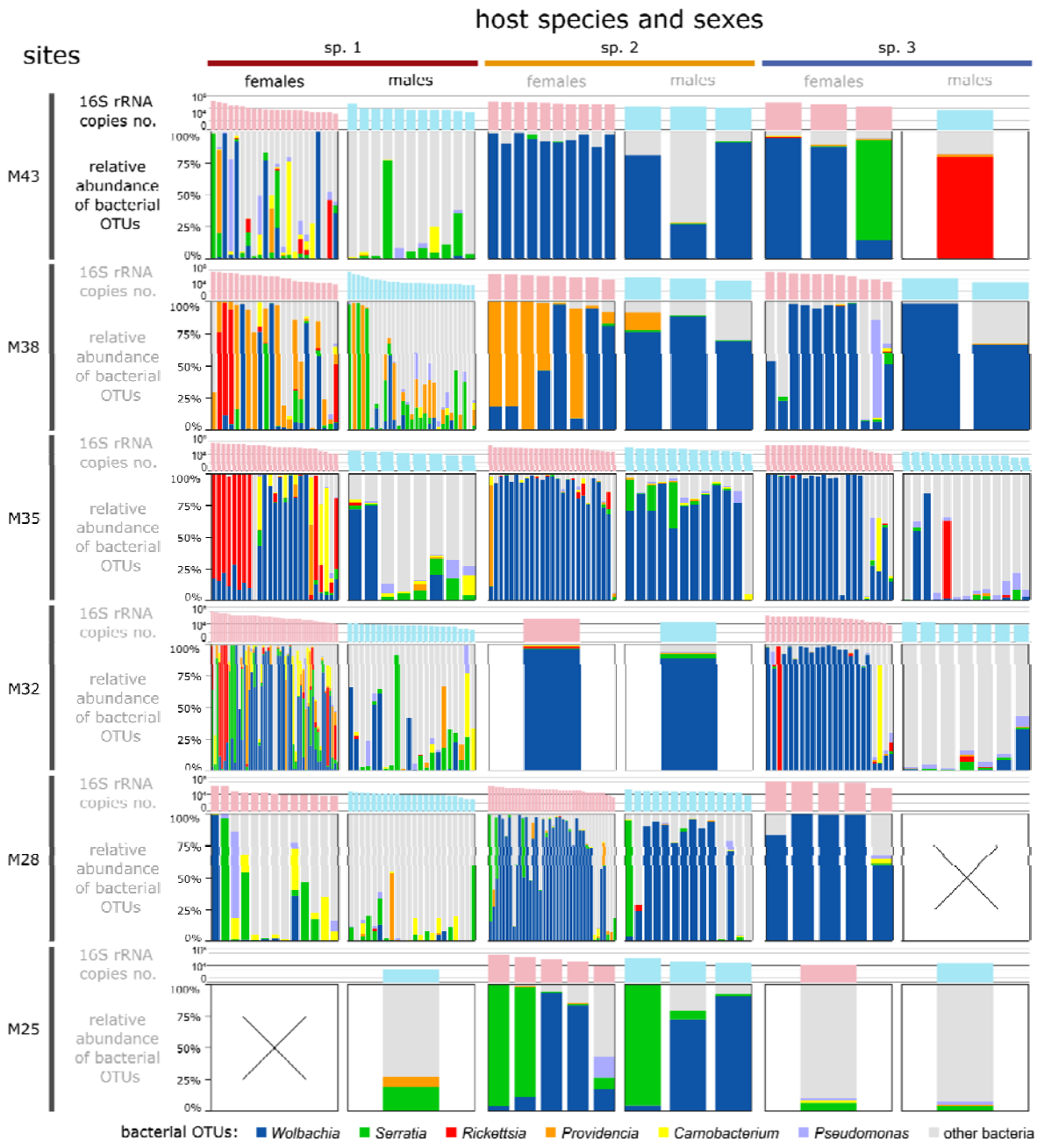



**Fig. 3:** The comparison of total bacterial abundance and relative abundance of dominant bacterial OTUs showcased the extent and patterns of variation among three most abundant phorid species, two sexes, and six sites. Each panel represents a particular species-sex-site combination, with the top bars indicating bacterial abundance and the lower stacked bars indicating the relative abundance of the most abundant bacterial 16S OTUs (as in Fig. 2). Individual columns of varying width represent individual flies.

Inspection of the data at higher phylogenetic resolution - 16S rRNA ZOTUs instead of 97% OTUs - provided additional information about the host-microbe association patterns. Bacterial OTUs typically comprise several relatively abundant genotypes (ZOTUs), and these OTUs differ in how their component ZOTUs are distributed across host species (Fig. 4). Specifically, insect species were found to preferentially associate with only one or two specific 16S rRNA genotypes of facultative endosymbiont *Wolbachia* and *Rickettsia*. Different species associate with different facultative endosymbiont genotypes, and conversely, different facultative endosymbiont genotypes colonise different sets of species. In contrast, for most other abundant bacterial OTUs including *Serratia*, *Providencia*, *Carnobacterium*, and *Pseudomonas*, associations appear much less specific. Most insect species harbour several genotypes from these OTUs, and genotypes are broadly distributed across host species (Supplementary File S3).

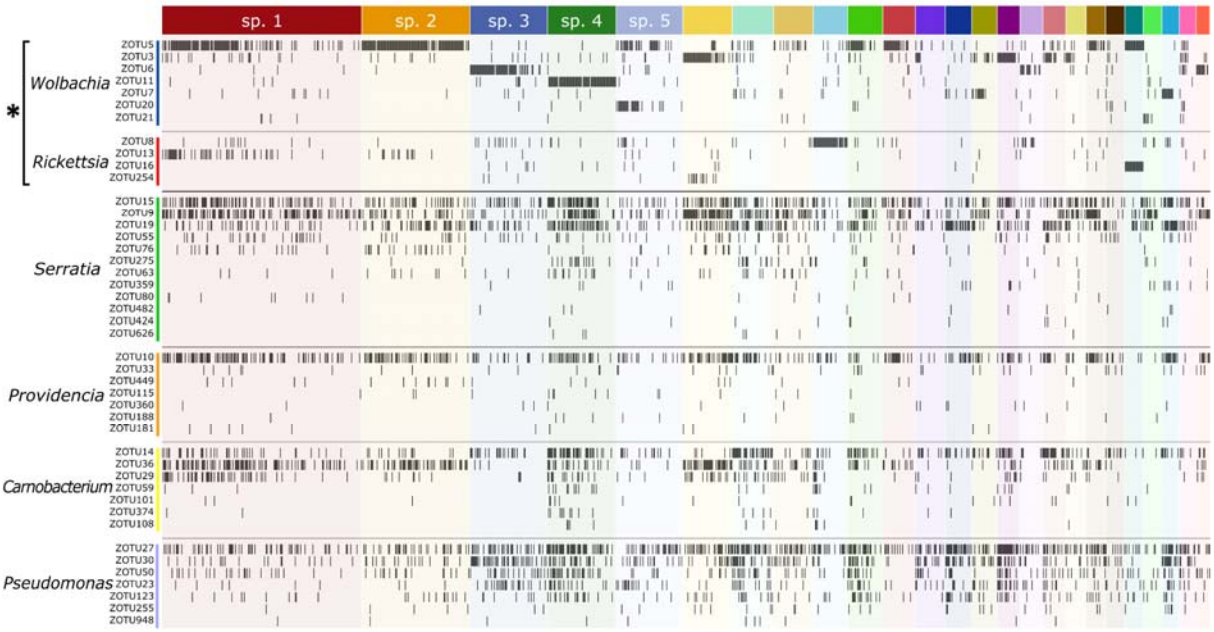



**Fig. 4**: The distributions of dominant bacterial genotypes (ZOTUs) within most abundant OTUs across host species reveal differences in specificity among bacterial functional categories: facultative endosymbionts (denoted with an asterisk) form relatively specific associations with hosts, unlike other

microbes. Insect individuals, represented by columns, are sorted by species, sex, and decreasing bacterial abundance (as in Fig. 2). Rows represent bacterial genotypes (ZOTUs) sorted into 97% OTUs, and black lines indicate symbiont genotypes represented by at least 100 copies of 16S rRNA in specific insect individuals.

We noted that cytochrome oxidase (I) amplicon data for Alphaproteobacteria provide even greater phylogenetic resolution. Specifically, *Wolbachia* comprised most of the 0.5% of non-insect COI sequences in the dataset. Based on these COI data, *Wolbachia* was represented by 2 or more reads and thus scored as present in 862 (47%) of samples, with an average of 269 reads per sample (max 4,875, median 100). These *Wolbachia* COI reads represented 258 unique ZOTUs grouped into 23 97% OTUs. A comparison of the 16S-V4 and COI data revealed > 90% agreement among the two datasets regarding the *Wolbachia* presence in the samples, after applying cutoffs likely to exclude much of likely cross-contamination (3 COI reads vs 400 reads of 16S rRNA) (Supplementary Table S5).

The comparison of *Wolbachia* COI OTU distributions across phorid species (Fig. 5a) resembles *Wolbachia* 16S rRNA genotype distribution patterns (Fig. 4). These data show that phorid species typically associate with a single dominant *Wolbachia* COI OTU and differ considerably in infection prevalence. COI genotype-level data for both hosts and symbionts (Fig. 5b) provide much more detailed information: they show that genotypes of the same fly species frequently associate with different genotypes of *Wolbachia*, with substantial variation among genotypes in infection prevalence. Interestingly, we observe some of the same *Wolbachia* COI genotypes across genotypes of different phorid species. For example, *Wolbachia* ZOTU206 infects most genotypes of phorid species 2, at high prevalence, but also colonizes several genotypes of sp. 7, sp. 9, and some less abundant species.

Within species, we found even greater specificity in host-symbiont interactions with respect to sex and site (Fig. 5c). For example, in sp. 5., *Wolbachia* COI ZOTU495 colonised most host genotypes, and tends to achieve higher prevalence in females, only sometimes infecting males. In contrast, *Wolbachia* ZOTU333 infected all individuals - males and females - of one of the abundant phorid COI genotypes,

and a single processed male of another genotype, while being absent from others. Interestingly, two other *Wolbachia* genotypes (ZOTU206 & ZOTU486) infected sp. 5 individuals representing different genotypes, all collected in summer rather than spring.

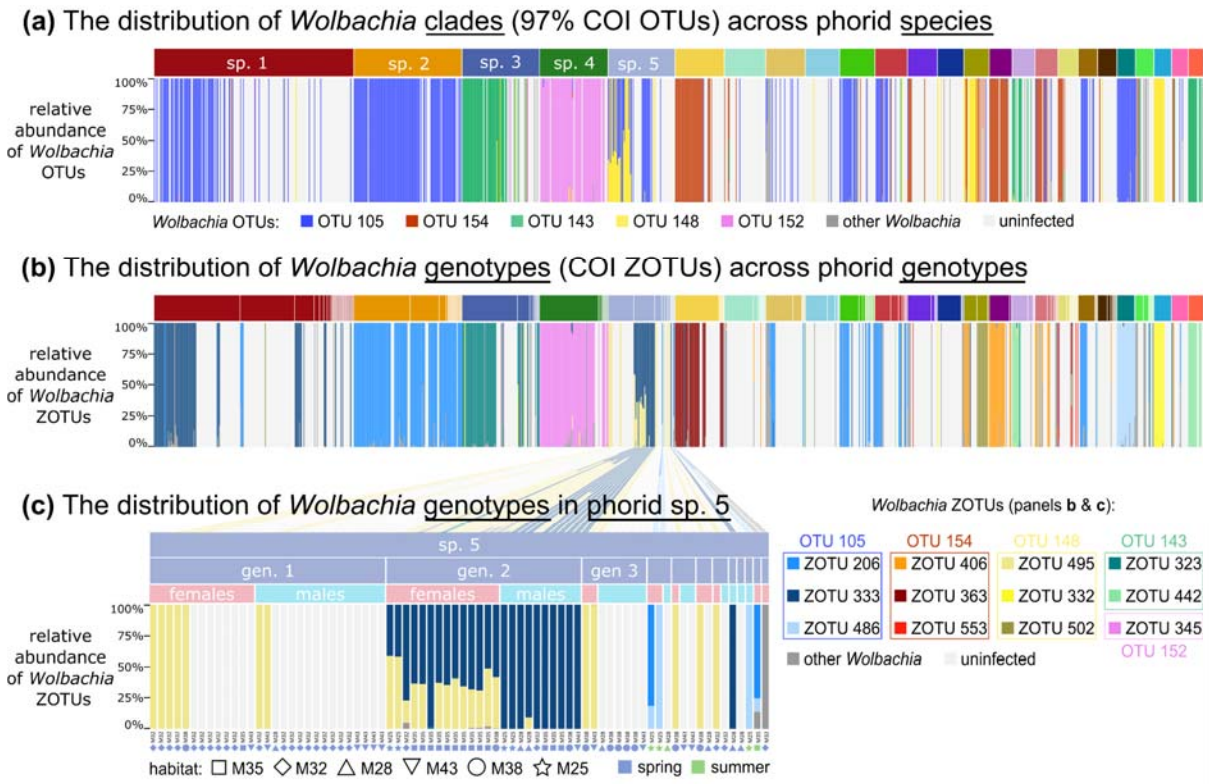



**Fig. 5:** *Wolbachia* COI amplicon data, analysed at different phylogenetic resolution scales, provides much more precise information about the patterns of host-symbiont interactions than 16S rRNA amplicon data. (**a**) The relative abundance of *Wolbachia* COI OTUs among all *Wolbachia* COI reads, across 25 most abundant scuttle fly species. Individual flies are represented by columns, and sorted by species, sex and decreasing abundance (as in Fig. 2 and 4); (**b**) The relative abundance of *Wolbachia* COI genotypes (ZOTUs) across individual flies sorted into species and additionally COI genotypes; (**c**) The relative abundance of *Wolbachia* genotypes (ZOTUs) among genotypes, sexes, and populations of sp. 5 (*Megaselia* sp. 004), highlights strong specificity in interactions. In all panels, individuals with fewer than 2 *Wolbachia* COI reads are considered uninfected and are represented by white columns.

## *3.3. Drivers of microbiota variation*

Within our dataset, we observed substantial variation in the prevalence and abundance of microbiota. To quantify the relative contributions of host traits, environmental factors, and microbial resolution levels to this variation, we turned to HMSC analyses. We reconstructed the drivers of presence/absence and abundance conditional on presence at three levels of resolution: 16S rRNA OTUs, 16S rRNA ZOTUs, and for *Wolbachia* only, COI ZOTUs.

### *3.3.1 Results of the HMSC models*

Our predictors (site, season, host species, host sex and host genotype) explained a substantial proportion of the variation in bacterial 16S rRNA OTU and ZOTU presence-absence (Supplementary Table S6), with high explanatory and predictive power (AUC ≥ 0.69). The effects of host individual and site contributed the most to variance, with microbial taxa generally more prevalent in the spring than summer and more abundant in female than male hosts. For abundance conditional on presence, the explanatory power remained high, but the predictive power was low due to the strong host individual effect (Table S6). Some individuals had a high abundance of most bacterial taxa, while in others, microbial abundance was consistently low for most bacterial taxa, amplifying trends in presence-absence. Unlike presence-absence, abundance was generally higher in summer than in spring, indicating that spring samples contained many taxa in low abundance, whereas summer samples had fewer taxa in high abundance.

For *Wolbachia* COI genotypes, presence-absence was strongly shaped by host species and genotype, with additional effects of season and site. Model performance was exceptionally high (Supplementary Table S6; AUC ≥ 0.90), and most strains were more prevalent in spring and in female hosts. *Wolbachia* abundance conditional on presence showed high explanatory power (AUC = 0.94),

but low predictive power, with individual variation driving 36% of the observed differences. Strain-specific abundances of *Wolbachia* were consistently higher in females than males.

At the level of individual microbial taxa, explained variance (Tjur's $R^2$) varied significantly, with more prevalent taxa generally showing higher values (Supplementary Fig. S2). The relative contributions of specific drivers also differed among taxa (Fig. 6).

### *3.3.2 Drivers of variation in bacterial clade occurrence and abundance*

The HMSC model explained up to 40% of the variance in the presence-absence of individual 16S-V4 OTUs, with much variation occurring among host individuals (Fig. 6a and Supplementary Fig. S3a). However, for facultative endosymbionts (*Wolbachia*, *Rickettsia*, *Spiroplasma*), variance was primarily driven by host species, sex, and genotypes, while *Lactobacillus* showed the strongest seasonal variation (Fig. 6a and Supplementary Fig. S3a). For abundance conditional on presence, the model explained up to 70% of variance, with lowest contribution of the host individual factor in facultative endosymbionts, exhibiting distinct patterns: *Wolbachia* varied most with host species, sex and genotype, *Rickettsia* showed a strong host genotype effect, and *Spiroplasma* exhibited minimal sex-based differences.

Similar patterns emerged for the 16S-V4 ZOTUs, where *Wolbachia* ZOTUs showed the highest Tjur $R^2$ values for both presence/absence and abundance conditional on presence, with variation attributed to host species and genotype. Facultative endosymbionts differed from other taxa, in which host individual effects explained most variation (Supplementary Fig. S3b). This was particularly clear in Actinomycetes, which showed the strongest host individual effect (Supplementary Fig. S3b).

For *Wolbachia* strains identified via COI (Supplementary Fig. S3c), variance in presence/absence was primarily shaped by host genotype, species, sex, or site, dependent on the

ZOTU. However, on the overall, a relatively low share of raw variance was explained. Abundance conditional on presence showed a greater influence of the host individual but remained significantly affected by host species, genotype, and sex.

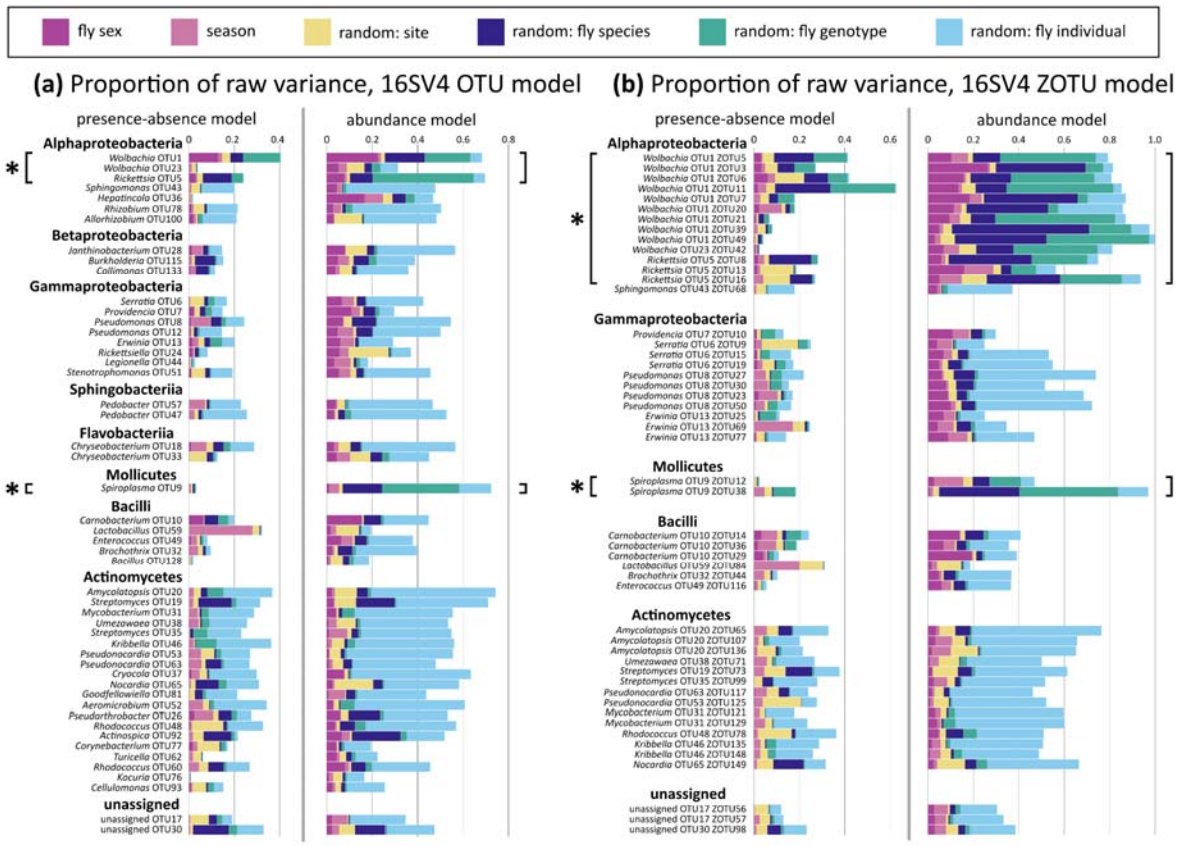



**Fig. 6**: Proportion of variance explained per microbial (a) 16S-V4 OTU and (b) 16S-V4 ZOTU by both the presence-absence and the abundance conditional on presence models, showing that facultative endosymbionts (denoted with asterisks) are highly influenced by the host sex, species and genotype factors when compared with other bacteria. For 16S-V4 ZOTUs, 50 of the most abundant ZOTUs are shown out of 100 included in the models.

### *3.3.3 Residual associations among microbial taxa*

After accounting for systematic effects, we examined residual associations among microbial taxa to identify influences beyond shared external drivers. For 16S rRNA OTU model, at the host species level, numerous associations were evident, with taxa showing positive co-occurrence within groups and negative co-occurrence between groups, particularly for Actinomycetes and various Proteobacteria (Fig. 7). At the host genotype level, Actinomycetes taxa formed two distinct clusters, with contrasting associations with other clades, though abundance associations were less consistent. At the individual host level, most taxa, except Alphaproteobacteria, Mollicutes, and some Actinomycetes, showed positive associations, indicating that host individuals either had high or low microbial diversity. Facultative endosymbionts showed limited interactions with other taxa. Site-level patterns of associations were mostly positive, with some exceptions, and mixed trends were observed for abundance conditional on presence, where more taxa showed negative associations.

Patterns observed for the 16S OTU models were similar to the ones shown for the 16S ZOTU models (Supplementary Fig. S4). In contrast, for the *Wolbachia* models, few significant associations were detected, likely due to low statistical power from the limited occurrence of specific ZOTUs across host species and individuals (Supplementary Fig. S4).

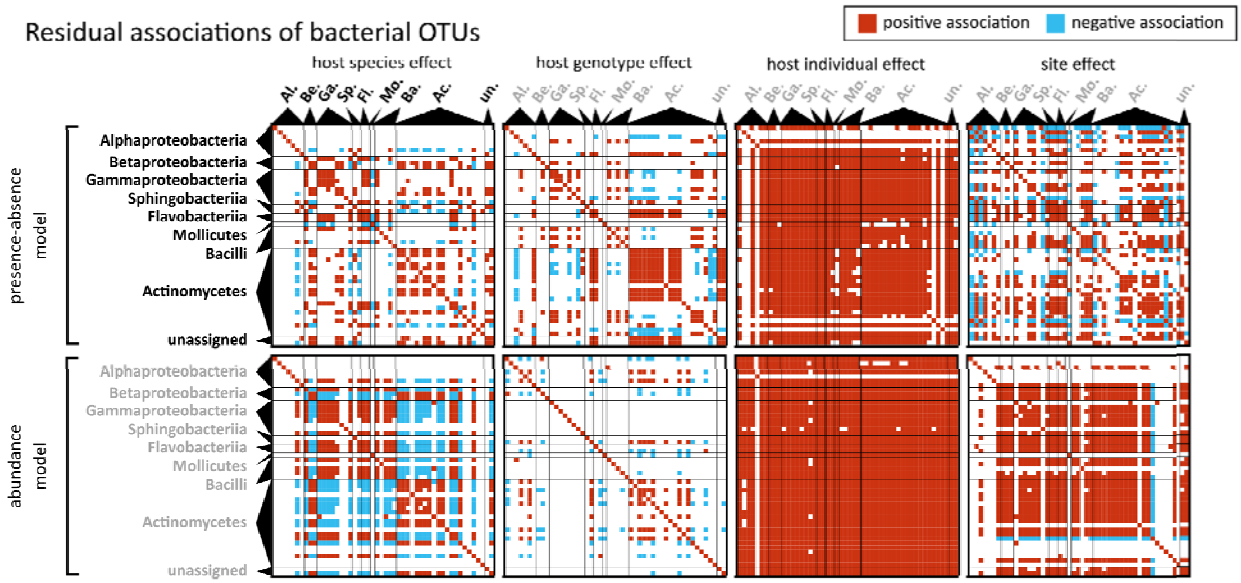



**Fig. 7:** Residual associations among bacterial taxa explain how OTUs may influence each others' presence and abundance (conditional on presence) once shared external drivers are accounted for. The columns of panels correspond to the four random levels included in the HMSC models, and the rows of panels represent two model components. Within each panel, the bacterial taxa were ordered in the same way as in the presence-absence part of the model at the level of the insect host species. Red colour indicates pairs of taxa with a positive residual association and blue colour pairs of taxa with a negative residual association, with a posterior probability of 0.90 used as a threshold.

## 4. Discussion

Insect biodiversity largely consists of poorly known 'dark taxa', such as scuttle flies [15,30,59]. In these groups, addressing even basic questions, such as about their interactions, pose considerable difficulties [29,60]. The situation is changing with the advent of new conceptual approaches combined with progress in sequencing, bioinformatics, and statistical modelling [11,17]. Our study demonstrates how these innovations, especially when applied in combination, can greatly enhance our understanding of interactions within the diverse wild insect communities.

## *4.1. What have we learnt about scuttle flies?*

This study shows how diverse communities of a dark taxon can be accurately resolved *at the same time* as their associations with microbes or other organisms. This approach opens up entirely new prospects for community ecology. To date, most studies of ecological communities have focused on characterising species' Grinnellian niches – that is, how communities are shaped by the match between abiotic conditions and species requirements [61–63]. This has largely neglected how species occurrence is determined by their interactions with other species – the Eltonian niche [64,65]. This study characterises both of these niches in one go. By simultaneously identifying the host species (here: scuttle flies) and their associated taxa (here: bacteria) across sites, we may effectively characterise entire wild communities of multi-partite symbiotic entities, or holobionts [66]. The current study focuses on a relatively small set of sites, thus constraining analyses of the Grinnellian niche – but another study shows how this can be achieved for broader sample collection [29]. By adding the microbial dimension, we show the way forward for the integration of the two [67].

## *4.2. What have we learnt about scuttle fly microbiota?*

Our study resolves massive variation in the microbiota of scuttle flies at several levels of organization. We show a large variation in bacterial abundance – a rarely assessed, although critical aspect of microbiota [19,20,40]. We present up to four orders of magnitude differences in the number of 16S rRNA copies, even among individuals representing the same population and sex. This is likely to correlate with large differences in bacterial cell density, although variations in rRNA operon copy numbers within bacterial genomes [68,69] and copies of the genome per cell [70] can also play a role. In turn, bacterial density should correlate with the functional significance of bacteria in the host biology, even if the density-dependence of bacterial effects on the host [71,72] makes the direction of these effects hard to predict.

This variation in bacterial abundance is linked to the community composition. In particular, facultative endosymbionts *Wolbachia* and *Rickettsia* tend to be highly abundant when present, driving the overall abundance among species and sexes. These widespread microbes show highly specific patterns in their distribution, with host species, genotype, and sex explaining a large share of the overall variance across statistical models. Pronounced variation in the association of facultative endosymbiont vs host genotypes suggests that these symbioses are dynamic at relatively short timescales - with symbiont strains moving among host lines, spreading, and likely affecting the host population structure, but rarely persisting for a long time [3,73,74]. At the same time, the broad distribution of some *Wolbachia* COI genotypes across different phorid species indicates that they are relatively frequently transmitted horizontally among distantly related host individuals [75].

*Wolbachia* is thought to spread in host populations through the combination of reproductive manipulation and beneficial effects, including protection against natural enemies and nutrient provisioning. Our community-level data strongly suggest that reproductive manipulation is common in the phorid system. Specifically, major differences in *Wolbachia* prevalence and abundance between males and females (Supplementary Table S6) argue for a negative link between infection and male functions. These patterns are consistent with expectations for some types of reproductive manipulation: male killing, feminisation of genetic males, or parthenogenesis induction [3,76]. Unfortunately, we lack information about sex ratios in the surveyed populations since our sampling and specimen selection strategy were not designed to assess sex ratios accurately. Nonetheless, in eight out of 25 most abundant species, the sex ratio among processed specimens significantly deviated from an expected 1:1 sex ratio towards females, and in six others towards males (chi-squared test, $p < 0.05$). The current tentative evidence for reproductive manipulation needs to be systematically validated through the combination of genomics and experiments.

Conversely, many of the remaining abundant bacteria represent generalist clades. For example, the genus *Pseudomonas* is known from a wide range of environments and host organisms that it influences in a multitude of ways. While hundreds of *Pseudomonas* species have been formally named,

their environmental variation is much greater [77]. The same applies to most other abundant bacterial genera in our dataset, including *Serratia* [78], *Carnobacterium* [79], *Providencia* [80], and others. To date, research on these bacterial genera has focused on a limited set of species of direct relevance to human health and economy, and our understanding of their taxonomic and functional diversity in natural ecosystems has remained limited. At the same time, our 16S rRNA genotype data highlight the surprisingly broad distribution of bacterial genotypes across phorid species, suggesting their recent and likely ongoing transmission among distantly related hosts. Here, the relatively slow rate of evolution in the rRNA gene [81] constrains further resolution of phorid-bacteria relationships. Higher-resolution tools, especially genomics, will be necessary to conclude how these microbes are distributed and move among hosts and environments and what their effects may be.

Overall, our data suggest that adult scuttle flies may not typically form specific associations with microbes other than heritable facultative endosymbionts. We suspect that at the time of flies' emergence from pupas, their body surfaces and digestive tracts may typically be largely sterile - and that only later do they become colonised by some of the versatile opportunistic bacteria present in the environment.

### *4.3. What do scuttle flies teach us about broader insect–microbe interaction patterns?*

Perhaps the most universal finding of this study, and the key to interpreting most patterns, is the striking contrast between microbial functional categories - facultative endosymbionts and other bacteria - in their associations with hosts [82]. In future studies, it will be critical to distinguish among functional categories when assessing insect-microbe associations. At the same time, when characterising the microbiota of a poorly known host group, distribution patterns can provide insights into microbial biology and its role in the host's biology.

Facultative endosymbioses, represented here by *Wolbachia*, *Rickettsia*, and *Spiroplasma*, are often thought to form relatively dynamic associations with hosts. We have convincing examples of relatively recent symbiont host switches among related species [74,83] and across distantly related species that form natural communities [84], as well as selective sweeps that they may drive [5,7,85,86]. These processes can dramatically influence insect population structure, performance in changing environmental conditions, or interactions with other organisms [11]. Conversely, this can affect species' responses to ecological and environmental challenges and opportunities, their longer-term evolutionary trajectories, as well as broader community processes [4,87]. However, we do not know how often these processes take place or what their primary drivers and likely consequences are [73,75]. We believe that the patterns observed in phorids - including a relatively short age of most host-symbiont associations, frequent transmission of some *Wolbachia* lineages across host species, and common reproductive manipulation - will eventually prove to be general features of facultative endosymbiont infections across diverse insects. Pioneering methodological approaches such as those used in our study will enable the systematic testing of these hypotheses about facultative endosymbiosis stability and infection dynamics.

In contrast, microbial clades other than facultative endosymbionts are likely to have much broader distributions. Most genera listed in Supplementary Table 4 have been reported repeatedly from many diverse insects from around the world - for example, from butterflies or colonies of army ants from Neotropics [52,53,88]. It is tempting to think of these microbes as highly mobile across global environments and host clades. However, comparative data from vertebrates show how host biology may shape the specificity of bacteria, including strains from broadly distributed genera [89]. The systematic assessment of microbial diversity, transmission patterns, and roles across taxa and environments would require not only massive sampling and stringent contamination control [18] but also markers that provide much greater phylogenetic resolution than a short rRNA gene region - ideally, entire genomes.

At the same time, scuttle flies do not represent the full range of known insect-microbe associations. Many other insects form highly stable associations with specific bacterial clades, often maintained through nutritional dependencies, whether with obligate endosymbionts [9] or with host-adapted gut microbiota [82,90]. We would normally expect bacteria representing these categories to be always present and abundant within the host species - and in surveys spanning multiple species, to show evidence of co-diversification with hosts [91,92]. However, it is possible that the near-universal associations with *Wolbachia* or *Rickettsia* that some scuttle fly species form may be due to the indispensable effects of these microbes, as reported from some other insects [93,94].

### *4.4. Advancing the study of wild-insect microbiota: a comprehensive approach*

This study introduces a novel methodological framework for understanding wild insect communities and their microbial associations, offering insights far beyond the capabilities of standard marker gene amplicon sequencing or other methodologies. Our approach incorporates four key innovations: 1: host DNA-based identification done alongside bacterial community characterisation, 2: quantification of bacterial abundance, 3: the use of COI amplicons for high-resolution characterisation of *Wolbachia*, and 4: the application of advanced modelling to decipher complex community patterns of hosts and microbes. By combining these techniques at scale, we provide a uniquely nuanced and biologically relevant understanding of microbiota diversity and its ecological role in wild insect communities.

DNA barcoding has already revolutionized the understanding of arthropod biodiversity [95]. In this study, while characterising microbiota, we simultaneously barcoded their hosts, providing a robust framework for reconstructing insect microbiota patterns with virtually unprecedented resolution. As shown in other insect systems, COI data may also enable correcting misidentifications and reconstructing patterns such as parasitoid infections - potentially important confounding variables in microbiota studies [96].

We incorporated synthetic spike-ins to estimate microbial abundance, a step that adds a crucial dimension to our understanding of microbiota. Such quantification has been repeatedly recommended due to the known variation in within-host bacterial abundance [17,19,20], but our approach shows how it can be implemented in large-scale studies. While qPCR has been the most popular quantification approach, spike-ins [40] are an extremely cost-effective alternative of obtaining consistent abundance estimates spanning some six orders of magnitude. The patterns reconstructed are biologically realistic, as revealed by the non-random bacterial distributions observed within scuttle fly species. Quantification helps to assess samples where bacterial abundances are high and community composition reliable, but also identify samples with low bacterial abundances, where we should treat the community composition information with caution. This is preferable to discarding such samples for being below the detection limit. While this approach awaits more systematic validation, we highly recommend incorporating spike-in-based quantification into microbiota studies across small organisms.

COI amplicon data for *Wolbachia* provide much higher phylogenetic resolution than standard 16S rRNA amplicon data and thus yield more accurate insights into relationships among insects and these key facultative endosymbionts. This was realised over two decades ago when COI was incorporated into the multi-locus sequence typing (MLST) scheme [97]. Interestingly, the realisation that *Wolbachia* co-amplifies with insect mitochondrial COI [98] has motivated dedicated efforts to block its amplification [99]. Here, we have considered this co-amplification as an opportunity rather than a problem. The > 90% overlap between 16S rRNA-based and COI-based *Wolbachia* detection (depending on the cutoffs used - K. H. Nowak, unpublished) suggests that this symbiont can be reliably detected based on COI data alone. Increased phylogenetic resolution greatly enhances our ability to reconstruct key distribution and transmission patterns. This suggests exciting prospects for extracting infection information from rapidly growing barcoding datasets.

Finally, our implementation of the Hierarchical Modelling of Species Communities proved indispensable for systematically reconstructing host and microbiota patterns and drivers. HMSC is

ideally suited to deal with the complexities of these multi-dimensional systems, but its applications in the reconstruction of insect-bacteria interaction networks have so far been limited [100]. On the other hand, in other host-microbe systems, the method has proven its power to reconstruct complex processes and patterns and their drivers [101]. In particular, it offers a way for disentangling species associations caused by a joint preference for the same environmental conditions from interactions among species. As we address broader questions about biological interactions in complex multi-species communities from environmentally diverse habitats undergoing rapid anthropogenic changes, HMSC will serve as an essential statistical framework for understanding such complexity.

The solutions outlined above - combined with other recommended approaches such as contamination filtering [17,19,102] - help address the key limitations of amplicon sequencing and put this powerful, if somewhat out-of-fashion technique back into the mainstream of microbiota research. The amplicon sequencing approach can be applied in tandem with complementary techniques that vary in breadth and depth - high-throughput (meta-) barcoding-based insect surveys on one end and high-resolution metagenomics and metatranscriptomics on the other [11,103]. Such a combination of tools and approaches can revolutionise our understanding of the significance of microbiota in natural ecosystems during the era of major anthropogenic changes.

## *4.5. Conclusions*

In this paper, we have shown how a combination of quantification-enabled marker gene amplicon sequencing and statistical modelling, when applied to large collections of diverse specimens, can reveal ecological patterns of fundamental importance in a model dark taxon. By sequencing amplicons of the host marker gene (COI) and reconstructing barcodes alongside microbiota characterisation, we establish the critical host taxonomic framework even when species are impossible to tell apart or undescribed. By incorporating synthetic spike-in-based quantification, we cost-

effectively assess microbiota abundance, a critical, if largely neglected, aspect of host-microbe interactions. By analysing the bacterial bycatch data from the COI amplicon sequencing, we improve the phylogenetic resolution for the dominant insect endosymbiont, *Wolbachia*. By implementing HMSC modelling, we reconstruct the primary drivers of host-microbe associations in a massive multi-dimensional dataset. More generally, we show how methodological and statistical innovations allow us to address questions about microbe associations across diverse natural communities at a combination of breadth and resolution that has, to our knowledge, not been achieved in a non-vertebrate system. Further refinement and broad implementation of these approaches will aid the research community's endeavours to understand the significance of microbial associations in the function and resilience of natural communities to rapidly changing conditions.

## Acknowledgements

The project was supported by the Polish National Agency for Academic Exchange grant PPN/PPO/2018/1/00015 and Polish National Science Centre grants 2018/31/B/NZ8/01158 and 2021/41/B/NZ8/04526 (to PŁ). EH's work on Swedish Phoridae was funded by the Swedish Taxonomy Initiative grant 2016-203 4.3. OO and TR were funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506: ERC-Synergy project LIFEPLAN). OO was additionally funded by the Research Council of Finland (grant no. 336212 and 345110). The authors also acknowledge support from the National Genomics Infrastructure in Stockholm. We thank Dave Karlsson, Esben Bøggild, trap owners, and sample sorters for the Artdatabanken-funded Swedish Insect Inventory Project that provided the samples for this project.

## Supplementary item legends

Supplementary figures, tables and files are available at:

https://github.com/KarolHub/Phorid-Microbiota/tree/main/Supplementary_materials

Supplementary Fig. S1: Ordination of host communities' composition among sites and seasons (NMDS plot based on Bray-Curtis dissimilarity).

Supplementary Fig. S2: Relationship between the explained variance ($R^2$) of bacterial taxa presence and their prevalence.

Supplementary Fig. S3: Proportion of variance explained per microbial taxon by the presence-absence model and the abundance conditional of presence model at the levels of microbial 16S-V4 OTUs (a), 16S-V4 ZOTUs (b) and *Wolbachia* COI ZOTUs (c).

Supplementary Fig. S4: Residual associations in the presence-absence model and the abundance conditional of presence model among bacterial taxa at the levels of microbial 16S-V4 OTUs (a), 16S-V4 ZOTUs (b) and *Wolbachia* COI ZOTUs (c).

Supplementary Table S1: Details of the sampling sites and the numbers of processed insect specimens passing the quality criteria.

Supplementary Table S2: List of sample barcodes, species and genotype assignments, along with their sexes, sampling locations and accession numbers.

Supplementary Table S3: Numbers and proportions of 16S rRNA reads classified as contaminants, spike-ins, and symbionts across the samples, identified by the QUACK script.

Supplementary Table S4: List of 16S rRNA OTUs, their prevalence and abundance across samples.

Supplementary Table S5: Comparison of *Wolbachia* presence between COI and 16S-V4 datasets.

Supplementary Table S6: Results of the HMSC models

Supplementary File 1: Compressed table of COI ZOTUs read numbers and taxonomic assignments across the samples.

Supplementary File 2: Compressed, decontaminated table of 16S-V4 ZOTU read numbers and taxonomic assignments across the samples.

Supplementary File 3: Compressed, decontaminated table of estimated 16S-V4 ZOTU copies and taxonomic assignments across the samples.

# References


1. McFall-Ngai M *et al.* 2013 Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci.* **110**, 3229–3236. (doi:10.1073/pnas.1218525110)

2. Douglas AE. 2015 Multiorganismal Insects: Diversity and Function of Resident Microorganisms. *Annu. Rev. Entomol.* **60**, 17–34. (doi:10.1146/annurev-ento-010814-020822)

3. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, Bordenstein SR, Bordenstein SR. 2021 Living in the endosymbiotic world of Wolbachia: A centennial review. *Cell Host Microbe* **29**, 879–893. (doi:10.1016/j.chom.2021.03.006)

4. Oliver KM, Degnan PH, Burke GR, Moran NA. 2010 Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. *Annu. Rev. Entomol.* **55**, 247–266. (doi:10.1146/annurev-ento-112408-085305)

5. Jaenike J. 2012 Population genetics of beneficial heritable symbionts. *Trends Ecol. Evol.* **27**, 226–232. (doi:10.1016/j.tree.2011.10.005)

6. Jaenike J, Unckless R, Cockburn SN, Boelio LM, Perlman SJ. 2010 Adaptation via Symbiosis: Recent Spread of a *Drosophila* Defensive Symbiont. *Science* **329**, 212–215. (doi:10.1126/science.1188235)

7. Kriesner P, Hoffmann AA, Lee SF, Turelli M, Weeks AR. 2013 Rapid Sequential Spread of Two

Wolbachia Variants in Drosophila simulans. *PLoS Pathog.* **9**, e1003607. (doi:10.1371/journal.ppat.1003607)

8. Cornwallis CK, Van ’T Padje A, Ellers J, Klein M, Jackson R, Kiers ET, West SA, Henry LM. 2023 Symbioses shape feeding niches and diversification across insects. *Nat. Ecol. Evol.* **7**, 1022–1044. (doi:10.1038/s41559-023-02058-0)

9. Moran NA, McCutcheon JP, Nakabachi A. 2008 Genomics and Evolution of Heritable Bacterial Symbionts. *Annu. Rev. Genet.* **42**, 165–190. (doi:10.1146/annurev.genet.41.110306.130119)

10. Frago E, Dicke M, Godfray HCJ. 2012 Insect symbionts as hidden players in insect–plant interactions. *Trends Ecol. Evol.* **27**, 705–711. (doi:10.1016/j.tree.2012.08.013)

11. Łukasik P, Kolasa MR. 2024 With a little help from my friends: the roles of microbial symbionts in insect populations and communities. *Philos. Trans. R. Soc. B Biol. Sci.* **379**, 20230122. (doi:10.1098/rstb.2023.0122)

12. Sudakaran S, Kost C, Kaltenpoth M. 2017 Symbiont Acquisition and Replacement as a Source of Ecological Innovation. *Trends Microbiol.* **25**, 375–390. (doi:10.1016/j.tim.2017.02.014)

13. Zheng H, Steele MI, Leonard SP, Motta EVS, Moran NA. 2018 Honey bees as models for gut microbiota research. *Lab Anim.* **47**, 317–325. (doi:10.1038/s41684-018-0173-x)

14. Shigenobu S, Yorimoto S. 2022 Aphid hologenomics: current status and future challenges. *Curr. Opin. Insect Sci.* **50**, 100882. (doi:10.1016/j.cois.2022.100882)

15. Li X, Wiens JJ. 2023 Estimating Global Biodiversity: The Role of Cryptic Insect Species. *Syst. Biol.* **72**, 391–403. (doi:10.1093/sysbio/syac069)

16. Stork NE. 2018 How Many Species of Insects and Other Terrestrial Arthropods Are There on

Earth? *Annu. Rev. Entomol.* **63**, 31–45. (doi:10.1146/annurev-ento-020117-043348)

17. Knight R *et al.* 2018 Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422. (doi:10.1038/s41579-018-0029-9)

18. Salter SJ *et al.* 2014 Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87. (doi:10.1186/s12915-014-0087-z)

19. Williamson E, Hammer T, Hogendoorn K, Eisenhofer R. 2024 Blanking on Blanks: Few Insect Microbiota Studies Control for Contaminants - A Systematic Review. (doi:10.1101/2024.08.30.610589)

20. Hammer TJ, Sanders JG, Fierer N. 2019 Not all animals need a microbiome. *FEMS Microbiol. Lett.* **366**, fnz117. (doi:10.1093/femsle/fnz117)

21. Baumann P. 2005 BIOLOGY OF BACTERIOCYTE-ASSOCIATED ENDOSYMBIONTS OF PLANT SAP-SUCKING INSECTS. *Annu. Rev. Microbiol.* **59**, 155–189. (doi:10.1146/annurev.micro.59.030804.121041)

22. Lange C *et al.* 2023 Impact of intraspecific variation in insect microbiomes on host phenotype and evolution. *ISME J.* **17**, 1798–1807. (doi:10.1038/s41396-023-01500-2)

23. Hurst GDD, Frost CL. 2015 Reproductive Parasitism: Maternally Inherited Symbionts in a Biparental World. *Cold Spring Harb. Perspect. Biol.* **7**, a017699. (doi:10.1101/cshperspect.a017699)

24. Perlmutter JI, Bordenstein SR. 2020 Microorganisms in the reproductive tissues of arthropods. *Nat. Rev. Microbiol.* **18**, 97–111. (doi:10.1038/s41579-019-0309-z)

25. Hartop E, Srivathsan A, Ronquist F, Meier R. 2022 Towards Large-Scale Integrative Taxonomy

(LIT): Resolving the Data Conundrum for Dark Taxa. *Syst. Biol.* **71**, 1404–1422. (doi:10.1093/sysbio/syac033)

26. Srivathsan A, Hartop E, Puniamoorthy J, Lee WT, Kutty SN, Kurina O, Meier R. 2019 Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. *BMC Biol.* **17**, 96. (doi:10.1186/s12915-019-0706-9)

27. Srivathsan A, Lee L, Katoh K, Hartop E, Kutty SN, Wong J, Yeo D, Meier R. 2021 ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biol.* **19**, 217. (doi:10.1186/s12915-021-01141-x)

28. Chua PYS, Bourlat SJ, Ferguson C, Korlevic P, Zhao L, Ekrem T, Meier R, Lawniczak MKN. 2023 Future of DNA-based insect monitoring. *Trends Genet.* **39**, 531–544. (doi:10.1016/j.tig.2023.02.012)

29. Hartop E, Lee L, Srivathsan A, Jones M, Peña-Aguilera P, Ovaskainen O, Roslin T, Meier R. 2024 Resolving biology's dark matter: species richness, spatiotemporal distribution, and community composition of a dark taxon. *BMC Biol.* **22**, 215. (doi:10.1186/s12915-024-02010-z)

30. Srivathsan A *et al.* 2023 Convergence of dominance and neglect in flying insect diversity. *Nat. Ecol. Evol.* **7**, 1012–1021. (doi:10.1038/s41559-023-02066-0)

31. Disney H. 1994 *Scuttle Flies*. Dordrecht: Springer Netherlands.

32. Hartop EA, Brown BV, Disney RHL. 2015 Opportunity in our Ignorance: Urban Biodiversity Study Reveals 30 New Species and One New Nearctic Record for Megaselia (Diptera: Phoridae) in Los Angeles (California, USA). *Zootaxa* **3941**. (doi:10.11646/zootaxa.3941.4.1)

33. Disney RH, Taylor S, Slay ME, Krejca JK. 2011 New species of scuttle flies (Diptera: Phoridae) recorded from caves in Nevada, USA. *Subterr. Biol.* **9**, 73–84. (doi:10.3897/subtbiol.9.2511)

34. Cheng L, Hogue CL. 1974 New distribution and habitat records of biting midges and mangrove flies from the coasts of southern Baja California, Mexico (Diptera: Ceratopogonidae, Culicidae, Chironomidae, and Phoridae. *Entomol. News* **85**, 211–218.

35. Disney RHL. 1991 The aquatic Phoridae (Diptera). *Insect Syst. Evol.* **22**, 171–191. (doi:10.1163/187631291X00048)

36. Pennington MJ, Rothman JA, Jones MB, McFrederick QS, Gan J, Trumble JT. 2017 Effects of contaminants of emerging concern on Megaselia scalaris (Lowe, Diptera: Phoridae) and its microbial community. *Sci. Rep.* **7**, 8165. (doi:10.1038/s41598-017-08683-7)

37. Sakamoto JM, Shikano I, Rasgon JL. 2024 Microbiomes of Two Pest Fly Species of Pennsylvania Mushroom Houses. *Insects* **15**, 525. (doi:10.3390/insects15070525)

38. Karlsson D, Hartop E, Forshage M, Jaschhof M, Ronquist F. 2020 The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. *Biodivers. Data J.* **8**, e47255. (doi:10.3897/BDJ.8.e47255)

39. Buczek M *et al.* 2024 A new tool in a toolbox: Addressing challenges in high-throughput microbiota surveys across diverse wild insects. (doi:10.1101/2024.08.26.609764)

40. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. 2016 Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res.* , gkw984. (doi:10.1093/nar/gkw984)

41. Rohland N, Reich D. 2012 Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946. (doi:10.1101/gr.128124.111)

42. Glenn TC *et al.* 2019 Adapterama II: universal amplicon sequencing on Illumina platforms (TaggiMatrix). *PeerJ* **7**, e7786. (doi:10.7717/peerj.7786)

43. Parada AE, Needham DM, Fuhrman JA. 2016 Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**, 1403–1414. (doi:10.1111/1462-2920.13023)

44. Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D. 2019 Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* **7**, e7745. (doi:10.7717/peerj.7745)

45. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014 PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620. (doi:10.1093/bioinformatics/btt593)

46. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016 VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584. (doi:10.7717/peerj.2584)

47. Edgar RC. 2010 Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461. (doi:10.1093/bioinformatics/btq461)

48. Antich A, Palacin C, Wangensteen OS, Turon X. 2021 To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics* **22**, 177. (doi:10.1186/s12859-021-04115-6)

49. Edgar RC. 2016 SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. (doi:10.1101/074161)

50. Leray M, Ho S-L, Lin I-J, Machida RJ. 2018 MIDORI server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics* **34**, 3753–3754. (doi:10.1093/bioinformatics/bty454)

51. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012 The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.

*Nucleic Acids Res.* **41**, D590–D596. (doi:10.1093/nar/gks1219)

52. Łukasik P, Newton JA, Sanders JG, Hu Y, Moreau CS, Kronauer DJC, O'Donnell S, Koga R, Russell JA. 2017 The structured diversity of specialized gut symbionts of the New World army ants. *Mol. Ecol.* **26**, 3808–3825. (doi:10.1111/mec.14140)

53. Valdivia C, Newton JA, Von Beeren C, O'Donnell S, Kronauer DJC, Russell JA, Łukasik P. 2023 Microbial symbionts are shared between ants and their associated beetles. *Environ. Microbiol.* **25**, 3466–3483. (doi:10.1111/1462-2920.16544)

54. Ovaskainen O, Tikhonov G, Norberg A, Guillaume Blanchet F, Duan L, Dunson D, Roslin T, Abrego N. 2017 How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20**, 561–576. (doi:10.1111/ele.12757)

55. Ovaskainen O, Abrego N. 2020 *Joint Species Distribution Modelling: With Applications in R.* 1st edn. Cambridge University Press. (doi:10.1017/9781108591720)

56. Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, Hui FKC. 2015 So Many Variables: Joint Modeling in Community Ecology. *Trends Ecol. Evol.* **30**, 766–779. (doi:10.1016/j.tree.2015.09.007)

57. Tikhonov G, Opedal ØH, Abrego N, Lehikoinen A, De Jonge MMJ, Oksanen J, Ovaskainen O. 2020 Joint species distribution modelling with the r □package H msc. *Methods Ecol. Evol.* **11**, 442–447. (doi:10.1111/2041-210X.13345)

58. Leigh JW, Bryant D. 2015 popart□: full□feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116. (doi:10.1111/2041-210X.12410)

59. Caruso V, Hartop E, Chimeno C, Noori S, Srivathsan A, Haas M, Lee L, Meier R, Whitmore D. 2024 An integrative framework for dark taxa biodiversity assessment at scale: A case study using

*Megaselia* (Diptera, Phoridae). *Insect Conserv. Divers.* **17**, 968–987. (doi:10.1111/icad.12762)

60. Roslin T. 2024 Understanding insect movements through space and time is vital for safeguarding global ecosystem services. *One Earth* **7**, 537–540. (doi:10.1016/j.oneear.2024.03.007)

61. Grinnell J. 1917 Field Tests of Theories Concerning Distributional Control. *Am. Nat.* **51**, 115–128. (doi:10.1086/279591)

62. Holt RD. 2009 Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proc. Natl. Acad. Sci.* **106**, 19659–19665. (doi:10.1073/pnas.0905137106)

63. Soberón J. 2019 A Grinnellian Niche Perspective on Species-Area Relationships. *Am. Nat.* **194**, 760–775. (doi:10.1086/705898)

64. Colwell RK, Rangel TF. 2009 Hutchinson's duality: The once and future niche. *Proc. Natl. Acad. Sci.* **106**, 19651–19658. (doi:10.1073/pnas.0901650106)

65. Hirzel AH, Le Lay G. 2008 Habitat suitability modelling and niche theory. *J. Appl. Ecol.* **45**, 1372–1381. (doi:10.1111/j.1365-2664.2008.01524.x)

66. Bordenstein SR, Theis KR. 2015 Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. *PLOS Biol.* **13**, e1002226. (doi:10.1371/journal.pbio.1002226)

67. Gravel D *et al.* 2019 Bringing Elton and Grinnell together: a quantitative framework to represent the biogeography of ecological interaction networks. *Ecography* **42**, 401–415. (doi:10.1111/ecog.04006)

68. Klappenbach JA. 2001 rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res.* **29**, 181–184. (doi:10.1093/nar/29.1.181)

69. Louca S, Doebeli M, Parfrey LW. 2018 Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* **6**, 41. (doi:10.1186/s40168-018-0420-9)

70. Pecoraro V, Zerulla K, Lange C, Soppa J. 2011 Quantification of Ploidy in Proteobacteria Revealed the Existence of Monoploid, (Mero-)Oligoploid and Polyploid Species. *PLoS ONE* **6**, e16392. (doi:10.1371/journal.pone.0016392)

71. Koga R *et al.* 2022 Single mutation makes Escherichia coli an insect mutualist. *Nat. Microbiol.* **7**, 1141–1150. (doi:10.1038/s41564-022-01179-9)

72. Medina Munoz M, Spencer N, Enomoto S, Dale C, Rio RVM. 2020 Quorum sensing sets the stage for the establishment and vertical transmission of Sodalis praecaptivus in tsetse flies. *PLOS Genet.* **16**, e1008992. (doi:10.1371/journal.pgen.1008992)

73. Bailly-Bechet M, Martins-Simões P, Szöllősi GJ, Mialdea G, Sagot M-F, Charlat S. 2017 How Long Does Wolbachia Remain on Board? *Mol. Biol. Evol.* **34**, 1183–1193. (doi:10.1093/molbev/msx073)

74. Turelli M *et al.* 2018 Rapid Global Spread of wRi-like Wolbachia across Multiple Drosophila. *Curr. Biol.* **28**, 963-971.e8. (doi:10.1016/j.cub.2018.02.015)

75. Sanaei E, Charlat S, Engelstädter J. 2021 *Wolbachia* host shifts: routes, mechanisms, constraints and evolutionary consequences. *Biol. Rev.* **96**, 433–453. (doi:10.1111/brv.12663)

76. Engelstädter J, Hurst GDD. 2009 The Ecology and Evolution of Microbes that Manipulate Host Reproduction. *Annu. Rev. Ecol. Evol. Syst.* **40**, 127–149. (doi:10.1146/annurev.ecolsys.110308.120206)

77. Lalucat J, Gomila M, Mulet M, Zaruma A, García-Valdés E. 2022 Past, present and future of the

boundaries of the Pseudomonas genus: Proposal of Stutzerimonas gen. Nov. *Syst. Appl. Microbiol.* **45**, 126289. (doi:10.1016/j.syapm.2021.126289)

78. Williams DJ *et al.* 2022 The genus Serratia revisited by genomics. *Nat. Commun.* **13**, 5195. (doi:10.1038/s41467-022-32929-2)

79. Leisner JJ, Laursen BG, Prévost H, Drider D, Dalgaard P. 2007 *Carnobacterium:* positive and negative effects in the environment and in foods. *FEMS Microbiol. Rev.* **31**, 592–613. (doi:10.1111/j.1574-6976.2007.00080.x)

80. Dong X, Jia H, Yu Y, Xiang Y, Zhang Y. 2024 Genomic revisitation and reclassification of the genus *Providencia*. *mSphere* **9**, e00731-23. (doi:10.1128/msphere.00731-23)

81. Ochman H, Elwyn S, Moran NA. 1999 Calibrating bacterial evolution. *Proc. Natl. Acad. Sci.* **96**, 12638–12643. (doi:10.1073/pnas.96.22.12638)

82. Perreau J, Patel DJ, Anderson H, Maeda GP, Elston KM, Barrick JE, Moran NA. 2021 Vertical Transmission at the Pathogen-Symbiont Interface: Serratia symbiotica and Aphids. *mBio* **12**, e00359-21. (doi:10.1128/mBio.00359-21)

83. Shropshire D, Conner WR, Vanderpool D, Hoffmann AA, Turelli M, Cooper BS. 2023 Rapid host switching of Wolbachia and even more rapid turnover of their phages and incompatibility-causing loci. (doi:10.1101/2023.12.04.569981)

84. Vancaester E, Blaxter M. 2023 Phylogenomic analysis of Wolbachia genomes from the Darwin Tree of Life biodiversity genomics project. *PLOS Biol.* **21**, e3001972. (doi:10.1371/journal.pbio.3001972)

85. Himler AG *et al.* 2011 Rapid Spread of a Bacterial Symbiont in an Invasive Whitefly Is Driven by Fitness Benefits and Female Bias. *Science* **332**, 254–256. (doi:10.1126/science.1199410)

86. Zhao D, Zhang Z, Niu H, Guo H. 2023 Pathogens are an important driving force for the rapid spread of symbionts in an insect host. *Nat. Ecol. Evol.* **7**, 1667–1681. (doi:10.1038/s41559-023-02160-3)

87. Ferrari J, Vavre F. 2011 Bacterial symbionts in insects or the story of communities affecting communities. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 1389–1400. (doi:10.1098/rstb.2010.0226)

88. Ravenscraft A, Berry M, Hammer T, Peay K, Boggs C. 2019 Structure and function of the bacterial and fungal gut microbiota of Neotropical butterflies. *Ecol. Monogr.* **89**, e01346. (doi:10.1002/ecm.1346)

89. Song SJ *et al.* 2020 Comparative Analyses of Vertebrate Gut Microbiomes Reveal Convergence between Birds and Bats. *mBio* **11**, e02901-19. (doi:10.1128/mBio.02901-19)

90. Engel P, Moran NA. 2013 The gut microbiota of insects – diversity in structure and function. *FEMS Microbiol. Rev.* **37**, 699–735. (doi:10.1111/1574-6976.12025)

91. Fisher RM, Henry LM, Cornwallis CK, Kiers ET, West SA. 2017 The evolution of host-symbiont dependence. *Nat. Commun.* **8**, 15973. (doi:10.1038/ncomms15973)

92. Perreau J, Moran NA. 2022 Genetic innovations in animal–microbe symbioses. *Nat. Rev. Genet.* **23**, 23–39. (doi:10.1038/s41576-021-00395-z)

93. Dedeine F, Boulétreau M, Vavre F. 2005 Wolbachia requirement for oogenesis: occurrence within the genus Asobara (Hymenoptera, Braconidae) and evidence for intraspecific variation in A. tabida. *Heredity* **95**, 394–400. (doi:10.1038/sj.hdy.6800739)

94. Hosokawa T, Koga R, Kikuchi Y, Meng X-Y, Fukatsu T. 2010 *Wolbachia* as a bacteriocyte-associated nutritional mutualist. *Proc. Natl. Acad. Sci.* **107**, 769–774. (doi:10.1073/pnas.0911476107)

95. Meier R, Lawniczak MKN, Srivathsan A. 2024 Illuminating Entomological Dark Matter with DNA Barcodes in an Era of Insect Decline, Deep Learning, and Genomics. *Annu. Rev. Entomol.* (doi:10.1146/annurev-ento-040124-014001)

96. Kolasa M, Kajtoch Ł, Michalik A, Maryańska□Nadachowska A, Łukasik P. 2023 Till evolution do us part: The diversity of symbiotic associations across populations of *Philaenus* spittlebugs. *Environ. Microbiol.* **25**, 2431–2446. (doi:10.1111/1462-2920.16473)

97. Baldo L *et al.* 2006 Multilocus Sequence Typing System for the Endosymbiont *Wolbachia pipientis*. *Appl. Environ. Microbiol.* **72**, 7098–7110. (doi:10.1128/AEM.00731-06)

98. Smith MA *et al.* 2012 Wolbachia and DNA Barcoding Insects: Patterns, Potential, and Problems. *PLoS ONE* **7**, e36514. (doi:10.1371/journal.pone.0036514)

99. Bleidorn C, Henze K. 2021 A new primer pair for barcoding of bees (Hymenoptera: Anthophila) without amplifying the orthologous coxA gene of Wolbachia bacteria. *BMC Res. Notes* **14**, 427. (doi:10.1186/s13104-021-05845-9)

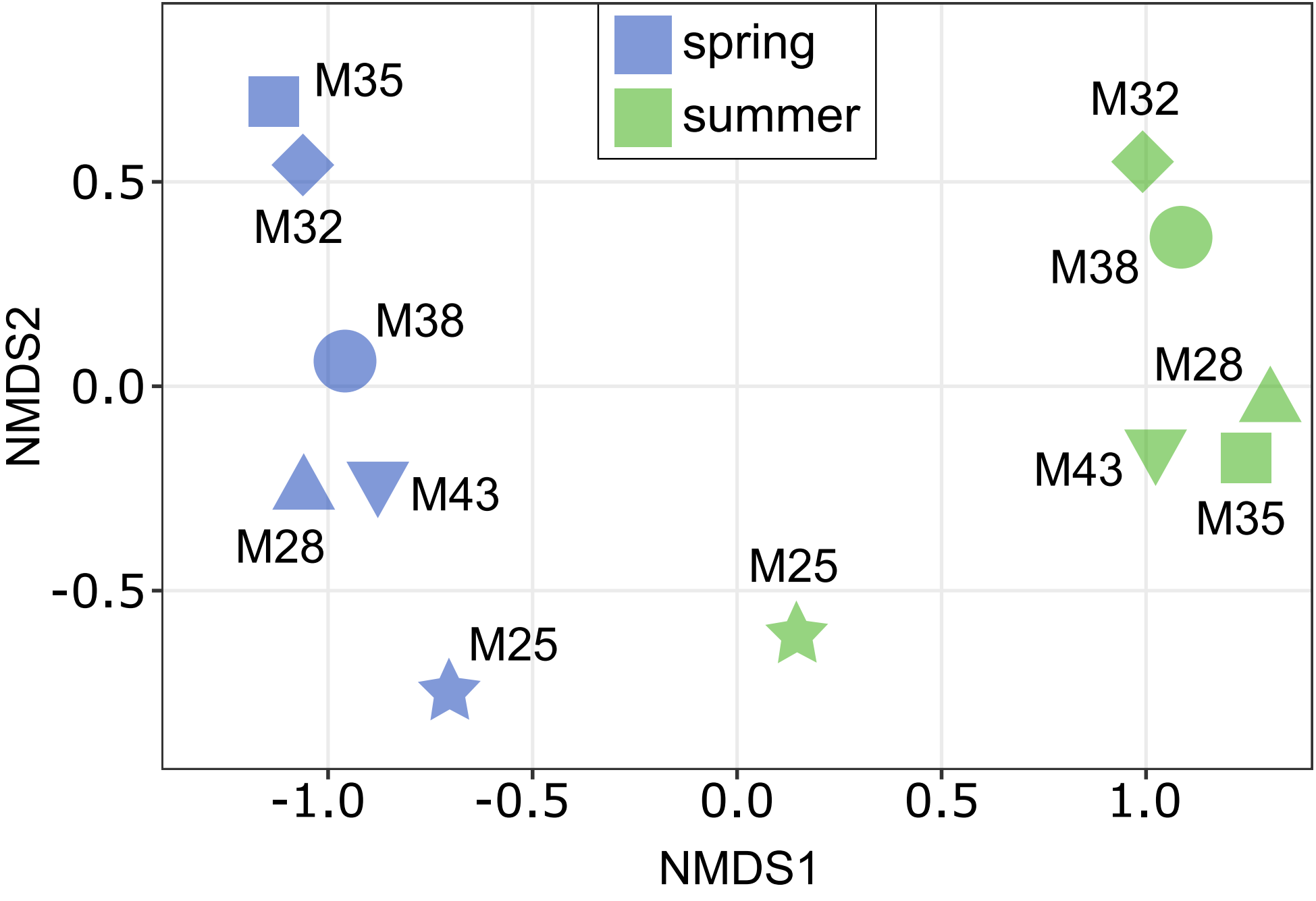
100. Minard G, Tikhonov G, Ovaskainen O, Saastamoinen M. 2019 The microbiome of the *Melitaea cinxia* butterfly shows marked variation but is only little explained by the traits of the butterfly or its host plant. *Environ. Microbiol.* **21**, 4253–4269. (doi:10.1111/1462-2920.14786)

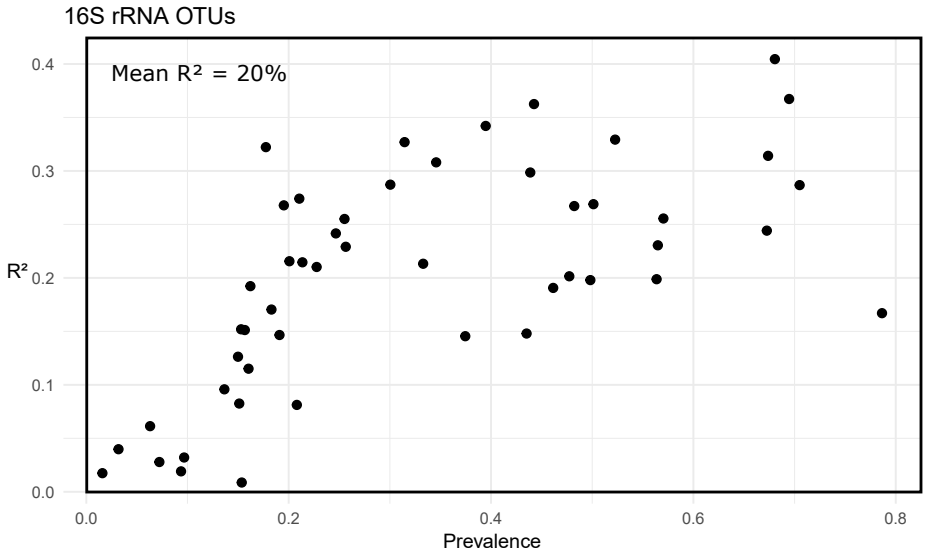
101. Parisy B *et al.* 2024 Arctic plant-fungus interaction networks show major rewiring with environmental variation. *Commun. Earth Environ.* **5**, 735. (doi:10.1038/s43247-024-01902-w)

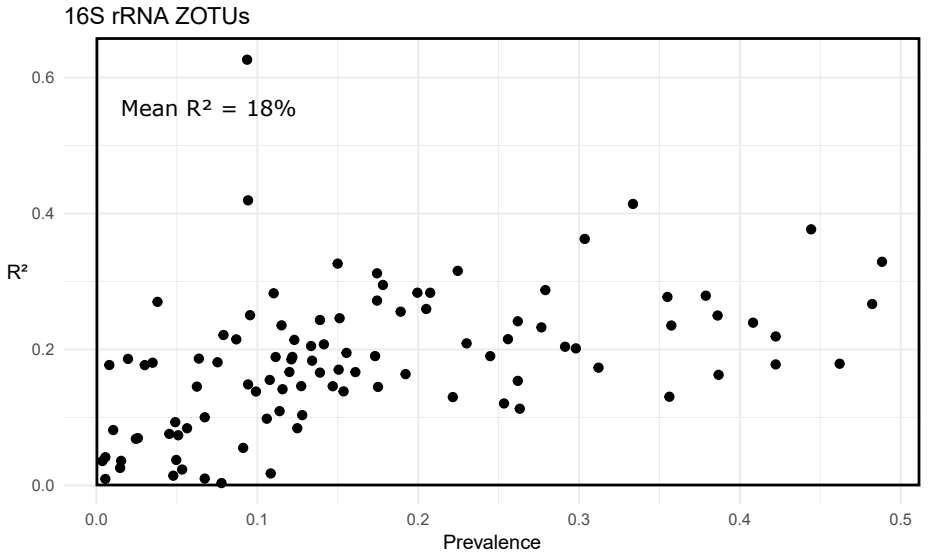
102. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018 Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226. (doi:10.1186/s40168-018-0605-2)

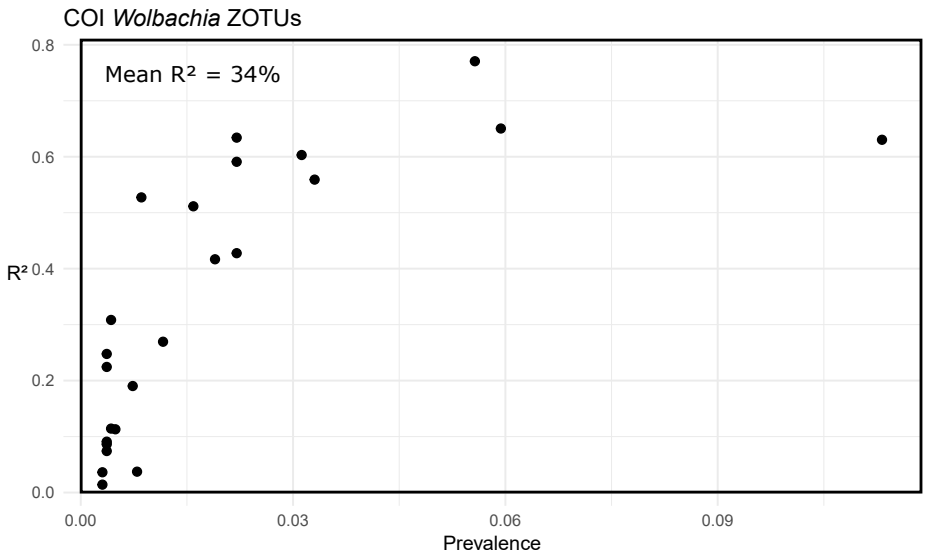
103. Andriienko V, Buczek M, Meier R, Srivathsan A, Łukasik P, Kolasa MR. 2024 Implementing

high-throughput insect barcoding in microbiome studies: impact of non-destructive DNA extraction on microbiome reconstruction. *PeerJ* **12**, e18025. (doi:10.7717/peerj.18025)

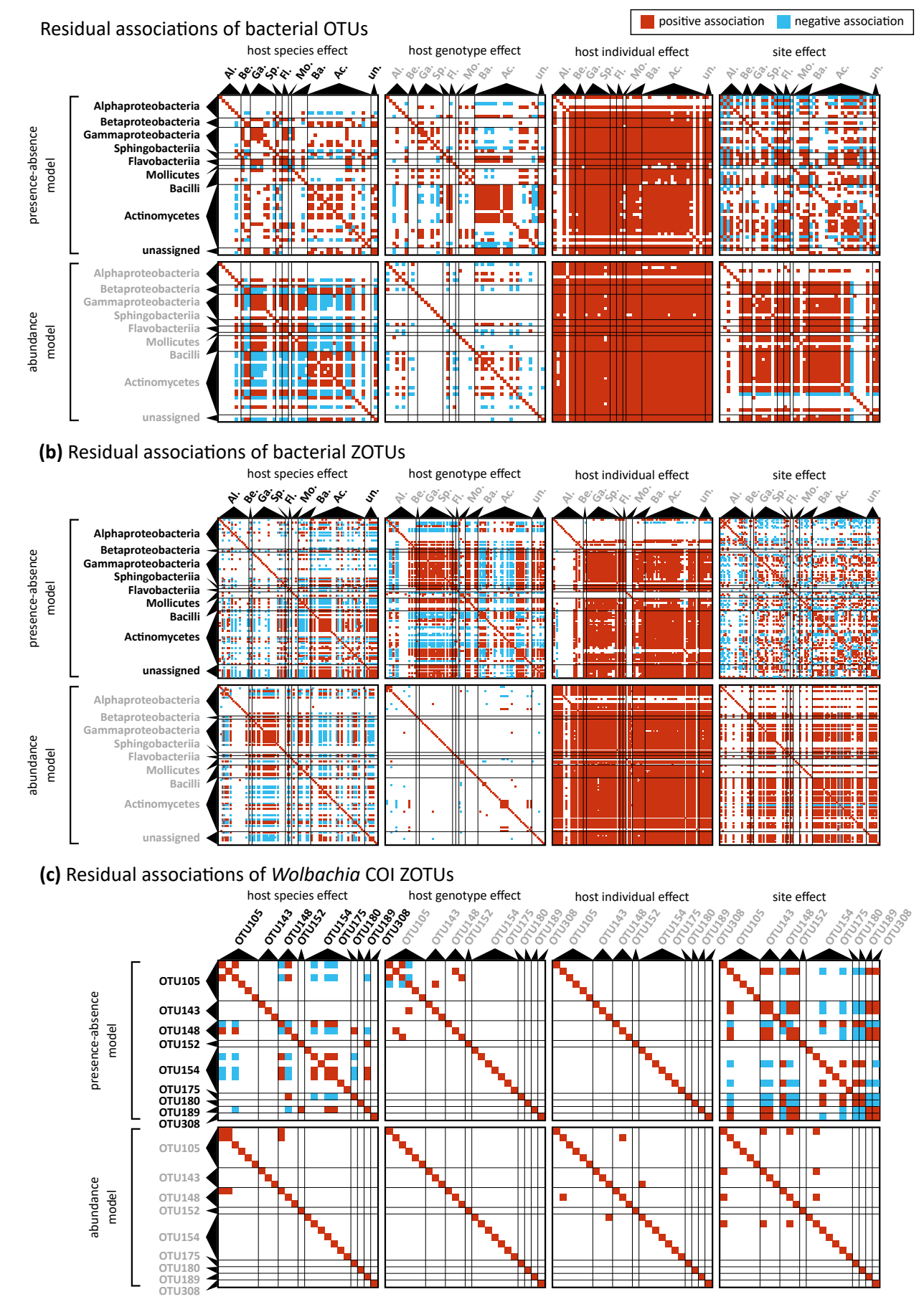

spring
summer
M35
M32
M38
M28
M43
M25
M32
M38
M28
M43
M35
M25
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
NMDS2
NMDS1


16S rRNA OTUs
Mean R² = 20%
R²
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
Prevalence


16S rRNA ZOTUs
Mean R² = 18%
R²
0.0
0.2
0.4
0.6
0.0
0.1
0.2
0.3
0.4
0.5
Prevalence


COI *Wolbachia* ZOTUs
Mean R² = 34%
R²
0.0
0.2
0.4
0.6
0.8
0.00
0.03
0.06
0.09
Prevalence

fly sex
season
random: site
random: fly species
random: fly genotype
random: fly individual
(a) Proportion of raw variance, 16SV4 OTU model
presence-absence model
abundance model
Alphaproteobacteria
Wolbachia OTU1
Wolbachia OTU23
Rickettsia OTU5
Sphingomonas OTU43
Hepatincola OTU36
Rhizobium OTU78
Allorhizobium OTU100
Betaproteobacteria
Janthinobacterium OTU28
Burkholderia OTU115
Collimonas OTU133
Gammaproteobacteria
Serratia OTU6
Providencia OTU7
Pseudomonas OTU8
Pseudomonas OTU12
Erwinia OTU13
Rickettsiella OTU24
Legionella OTU44
Stenotrophomonas OTU51
Sphingobacteriia
Pedobacter OTU57
Pedobacter OTU47
Flavobacteriia
Chryseobacterium OTU18
Chryseobacterium OTU33
Mollicutes
Spiroplasma OTU9
Bacilli
Carnobacterium OTU10
Lactobacillus OTU59
Enterococcus OTU49
Brochothrix OTU32
Bacillus OTU128
Actinomycetes
Amycolatopsis OTU20
Streptomyces OTU19
Mycobacterium OTU31
Umezawaea OTU38
Streptomyces OTU35
Kribbella OTU46
Pseudonocardia OTU53
Pseudonocardia OTU63
Cryocola OTU37
Nocardia OTU65
Goodfellowiella OTU81
Aeromicrobium OTU52
Pseudarthrobacter OTU26
Rhodococcus OTU48
Actinospica OTU92
Corynebacterium OTU77
Turicella OTU62
Rhodococcus OTU60
Kocuria OTU76
Cellulomonas OTU93
unassigned
unassigned OTU17
unassigned OTU30
(b) Proportion of raw variance, 16SV4 ZOTU model
presence-absence model
abundance model
Alphaproteobacteria
Wolbachia OTU1 ZOTU5
Wolbachia OTU1 ZOTU3
Wolbachia OTU1 ZOTU6
Wolbachia OTU1 ZOTU11
Wolbachia OTU1 ZOTU7
Wolbachia OTU1 ZOTU20
Wolbachia OTU1 ZOTU21
Wolbachia OTU1 ZOTU39
Wolbachia OTU1 ZOTU49
Wolbachia OTU1 ZOTU40
Wolbachia OTU23 ZOTU42
Rickettsia OTU5 ZOTU18
Rickettsia OTU5 ZOTU13
Rickettsia OTU5 ZOTU16
Sphingomonas OTU43 ZOTU68
Hepatincola OTU36 ZOTU54
Rhizobium OTU105 ZOTU156
Paracoccus OTU158 ZOTU350
Roseomonas OTU212 ZOTU355
Betaproteobacteria
Janthinobacterium OTU28 ZOTU41
Collimonas OTU133 ZOTU316
Gammaproteobacteria
Providencia OTU7 ZOTU10
Providencia OTU7 ZOTU33
Serratia OTU6 ZOTU9
Serratia OTU6 ZOTU15
Serratia OTU6 ZOTU19
Serratia OTU6 ZOTU76
Serratia OTU6 ZOTU55
Serratia OTU6 ZOTU63
Pseudomonas OTU8 ZOTU27
Pseudomonas OTU8 ZOTU30
Pseudomonas OTU8 ZOTU23
Pseudomonas OTU8 ZOTU50
Pseudomonas OTU8 ZOTU123
Pseudomonas OTU12 ZOTU58
Pseudomonas OTU12 ZOTU130
Erwinia OTU13 ZOTU25
Erwinia OTU13 ZOTU69
Erwinia OTU13 ZOTU77
Rickettsiella OTU24 ZOTU62
Legionella OTU44 ZOTU70
Alkanindiges OTU121 ZOTU183
Sphingobacteriia
Pedobacter OTU57 ZOTU170
Sphingobacterium OTU39 ZOTU72
Flavobacteriia
Chryseobacterium OTU18 ZOTU43
Chryseobacterium OTU33 ZOTU128
Mollicutes
Spiroplasma OTU9 ZOTU12
Spiroplasma OTU9 ZOTU38
Spiroplasma OTU9 ZOTU61
Spiroplasma OTU9 ZOTU46
Bacilli
Carnobacterium OTU10 ZOTU14
Carnobacterium OTU10 ZOTU36
Carnobacterium OTU10 ZOTU29
Carnobacterium OTU10 ZOTU59
Lactobacillus OTU59 ZOTU84
Brochothrix OTU32 ZOTU44
Enterococcus OTU49 ZOTU116
Bacilli OTU189 ZOTU318
Actinomycetes
Amycolatopsis OTU20 ZOTU65
Amycolatopsis OTU20 ZOTU107
Amycolatopsis OTU20 ZOTU136
Amycolatopsis OTU20 ZOTU242
Umezawaea OTU38 ZOTU71
Streptomyces OTU19 ZOTU73
Streptomyces OTU19 ZOTU229
Streptomyces OTU19 ZOTU201
Streptomyces OTU19 ZOTU293
Streptomyces OTU35 ZOTU99
Streptomyces OTU35 ZOTU208
Streptomyces OTU35 ZOTU335
Pseudonocardia OTU63 ZOTU117
Pseudonocardia OTU53 ZOTU125
Mycobacterium OTU31 ZOTU121
Mycobacterium OTU31 ZOTU129
Mycobacterium OTU31 ZOTU146
Rhodococcus OTU48 ZOTU78
Rhodococcus OTU101 ZOTU166
Rhodococcus OTU60 ZOTU88
Kribbella OTU46 ZOTU135
Kribbella OTU46 ZOTU148
Kribbella OTU46 ZOTU234
Nocardia OTU65 ZOTU149
Turicella OTU62 ZOTU95
Corynebacterium OTU77 ZOTU132
Actinospica OTU92 ZOTU192
Pseudarthrobacter OTU26 ZOTU91
Pseudarthrobacter OTU26 ZOTU75
Cryocola OTU37 ZOTU163
Skermania OTU144 ZOTU236
Aeromicrobium OTU52 ZOTU187
Kocuria OTU76 ZOTU301
Williamsia OTU130 ZOTU230
unassigned
unassigned OTU17 ZOTU56
unassigned OTU17 ZOTU57
unassigned OTU17 ZOTU249
unassigned OTU30 ZOTU98
unassigned OTU30 ZOTU213
unassigned OTU30 ZOTU207
unassigned OTU30 ZOTU245
unassigned OTU170 ZOTU277
(c) Proportion of raw variance, Wolbachia COI model
presence-absence model
abundance model
OTU105
ZOTU206
ZOTU333
ZOTU486
ZOTU537
ZOTU535
ZOTU538
OTU143
ZOTU323
ZOTU442
ZOTU533
OTU148
ZOTU332
ZOTU495
ZOTU502
OTU152
ZOTU345
OTU154
ZOTU406
ZOTU363
ZOTU510
ZOTU553
ZOTU567
ZOTU508
ZOTU898
OTU175
ZOTU490
OTU180
ZOTU512
OTU189
ZOTU661
OTU308
ZOTU1188


Residual associations of bacterial OTUs
positive association
negative association
host species effect
host genotype effect
host individual effect
site effect
Al. Be. Ga. Sp. Fl. Mo. Ba. Ac. un.
presence-absence model
abundance model
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Sphingobacteriia
Flavobacteriia
Mollicutes
Bacilli
Actinomycetes
unassigned
(b) Residual associations of bacterial ZOTUs
(c) Residual associations of Wolbachia COI ZOTUs
OTU105
OTU143
OTU148
OTU152
OTU154
OTU175
OTU180
OTU189
OTU308