Beneath the scuttle: Patterns of bacterial abundance, diversity and distribution in an insect dark taxon

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

Abstract

Symbiotic microorganisms can have diverse multi-level effects on wild insects: they can dramatically influence host life history traits, affect population structure and dynamics, shape host evolutionary trajectories, and influence multi-species communities. Despite their importance, our understanding of microbiota in natural insect communities remains limited, especially for ‘dark taxa’ - the hyperdiverse, abundant, broadly distributed, yet severely understudied clades that dominate our planet. Here, we resolve patterns in the abundance, diversity, and distribution of microbiota in natural communities of one such dark taxon, the scuttle flies (Diptera: Phoridae). Combining high-throughput amplicon sequencing of host and bacterial marker genes, including bacterial quantification, with joint species distribution modelling, we characterise the microbiome of 1,842 scuttle flies representing ca. 186 species from six sites in Northern Sweden. In particular, we ask how the microbiome is structured by the host species, genotype, and sex of the host individual, as well as by site and season.

We found fly individuals to vary in bacterial abundance by up to five orders of magnitude, with major variation in bacterial community composition. Strains of facultative endosymbionts *Wolbachia* and *Rickettsia* showed high host specificity: both fly COI genotypes and *Wolbachia* 16S rRNA and COI genotypes differed considerably in their distribution across genotypes of partners. Joint species distribution modelling revealed that symbiont distribution and abundance were largely explained by host species and genotypes as well as sex. In contrast, 16S rRNA genotypes of most other abundant bacterial genera in the dataset, including *Serratia*, *Providencia*, *Carnobacterium* and *Pseudomonas*,

were broadly distributed across host species and populations. Models indicated that much of the variation occurred at the level of individuals. Also, facultative endosymbiont infections had no consistent effect on the presence and abundance of other microbes.

This broad survey of microbiota in a non-model insect clade highlights the importance of facultative endosymbionts in shaping microbial community profiles and reveals striking variation among microbial clades in their distributional patterns. It also showcases the potential of integrative sequencing-based approaches in revealing the complexity and ecological significance of symbiotic microorganisms in diverse multi-species natural communities.

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

1. Introduction

Since their symbiotic origins (Vosseberg et al., 2024), eukaryotes have formed diverse associations with taxonomically and functionally diverse bacteria and other microbes that have often played dramatically important roles in their evolution and biology (McFall-Ngai et al., 2013). Some of these associations (described as 'closed' by Perreau & Moran, 2022) have been stable over tens or hundreds of millions of years and are strictly heritable. By contributing essential nutrients, such associations have enabled the evolutionary success and diversification of entire host clades on limited nutritional resources (Baumann, 2005; Cornwallis et al., 2023). Other host-microbe symbioses ('mixed') are mostly transmitted maternally but can sometimes switch hosts (horizontal transfer). They have diverse and often major effects on host reproduction, defence, and sometimes nutrition as well (Kaur et al., 2021; Oliver et al., 2010). Many eukaryotes form less specialized ('open'), taxonomically and functionally diverse associations which nevertheless can have major effects (Engel & Moran, 2013). Notably, many organisms do not seem to form stable associations with microbes at all (Hammer et al., 2019). On the other hand, all organisms are frequently attacked by diverse parasites and pathogens (Danilova, 2006).

These different functional categories have shaped hosts' biology at many levels that are relatively well-understood in insects (Douglas, 2015). The range of microbial symbionts' effects on insect life-history traits - nutrition, reproduction, and defence - is relatively best-known (Kaur et al., 2021; Oliver et al., 2010). Evidence is mounting that these individual-level effects translate to population processes and patterns (Jaenike et al., 2010; Jaenike, 2012; Kriesner et al., 2013), host evolutionary trajectories (Cornwallis et al., 2023; Moran et al., 2008), and effects at the level of multi-species communities (Frago et al., 2012). There is little doubt that through such combined effects, microbiota have shaped the global biodiversity patterns, and will continue playing key roles in ecosystems rapidly changing under anthropogenic pressures (Łukasik & Kolasa, 2024; Sudakaran et al., 2017).

The alphaproteobacterial genus *Wolbachia* showcases the diversity of symbionts' effects at different levels of biological organisation (Kaur et al., 2021). These endosymbionts infect about half of all insect species and many other arthropods and nematodes (Weinert et al., 2015; Zug & Hammerstein, 2012). Although historically viewed mainly as parasitic reproductive manipulators inducing male killing, parthenogenesis, feminisation, and cytoplasmic incompatibility, their effects cover a broad range between mutualism and parasitism. Various strains can enhance host fitness by

providing essential nutrients such as B vitamins (Hosokawa et al., 2010; Moriyama et al., 2015) or granting resistance to natural enemies, especially viruses (Teixeira et al., 2008). Through such a combination of effects, strains can rapidly sweep through populations and alter their genetic structure (Kriesner et al., 2013). They have been found to move among species at least occasionally (Turelli et al., 2018; Vancaester & Blaxter, 2023), potentially driving adaptation to environmental challenges and opportunities (Hague et al., 2020; Kaur et al., 2021; Łukasik & Kolasa, 2024).

Despite their massive importance, insect microbiota are poorly known outside of model species. Some biological systems such as honey bees (Zheng et al., 2018) or pea aphids (Shigenobu & Yorimoto, 2022) 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 (Li & Wiens, 2023; Stork, 2018), our understanding of microbiota is nonexistent, or at best, very limited. Among the primary reasons is the sheer diversity of insects and microorganisms, which poses a significant challenge in generating, handling and analysing vast amounts of multidimensional ecological data. Another limitation is the lack of reliable and cost-effective workflows for processing large numbers of wild-caught insects, encompassing both host identity validation and microbiome characterisation. Existing protocols suffer from the lack of standardisation and other issues such as the lack of controls for different types of contamination that can dramatically alter results (Knight et al., 2018; Salter et al., 2014). Indeed, a recent meta-analysis reported that over two-thirds of insect microbiota studies lack any contamination control and thus cannot be considered reliable (Williamson et al., 2024). Also, existing approaches rarely consider and address critical questions such as microbial abundance estimation (Hammer et al., 2019). 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 of low abundance, with contamination not controlled. As a result, we know very little about the patterns and primary drivers of microbial composition and abundance across the arthropod diversity. Host biology, especially the diet, has been identified as an important driver (Cornwallis et al., 2023). Host taxonomy is another driver (Baumann, 2005); however, even within 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 (Lange et al., 2023). Similarly, the differences between host sexes in microbiota diversity and abundance have rarely been considered outside of the context of reproductive manipulation (Hurst & Frost, 2015; Perlmutter & Bordenstein, 2020). 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.

Scuttle flies (Diptera: Phoridae) are a quintessential example of a “dark taxon”: they are hyperdiverse, abundant, and globally distributed, yet remain poorly understood (Hartop et al., 2022; Srivathsan et al., 2023). They exhibit diverse life strategies, including parasitism, predation, herbivory, and saprophagy (reviewed by Disney, 1994). Such varied life histories allow them to occupy numerous ecological niches, from bustling urban centres (Hartop et al., 2015) and subterranean caves (Disney et al., 2011) to intertidal zones (Cheng & Hogue, 1974) and the cavities of pitcher plants (Disney, 1991).

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 (Srivathsan et al., 2019). 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 (Pennington et al., 2017; Sakamoto et al., 2024).

The study of dark taxa has greatly benefitted from recent advances in molecular workflows (Hartop et al., 2022; Srivathsan et al., 2018, 2019, 2021; Wang et al., 2018), particularly in Megabarcoding - high-throughput single-specimen barcoding (Chua et al., 2023). Using this approach, the study of scuttle flies has rapidly expanded to large-scale analyses illuminating the distribution and abundance of phorid communities on a broader scale than previously possible (Hartop et al., 2024). Existing barcoding data combined with the diversity, accessibility, and straightforward sampling and processing of scuttle flies provide a promising opportunity to illuminate the hidden microbial diversity within this enigmatic group. This research could also offer valuable insights applicable to broader ecological and biological systems.

In this study, we comprehensively characterise the abundance, composition, and distribution patterns of microbiota across a large sample of scuttle fly communities from Northern Sweden. To understand the level and drivers of variation in microbial communities of individual flies within this dark taxon, we ask how the individual microbiota are structured by the site and the season, the identity of the host species, and the sex and the genotype of the host individual. This we do at two levels: across the full microbial community and for the dominant microbe in the dataset, *Wolbachia*. We addressed all three parts by applying a combination of innovative laboratory and analytical approaches - the simultaneous characterisation of host and bacterial marker genes, bacterial quantification, and joint species distribution modelling - to 1,842 taxonomically diverse phorid flies sorted from Malaise traps.

2. Materials and methods

2.1. Field collections

The experimental insects were collected using Malaise traps as a part of the Swedish Insect Inventory Project (SIIP) (Karlsson et al., 2020). 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. 2). Within each site, a Malaise trap operated during the 2018 vegetation season in two periods: late spring (June to July) and summer (July to August) (Supplementary Table S1). During the collection, bottles containing 95% ethanol were exchanged between every 7 to 31 days due to access challenges. After collection, insects were transported to Station Linné, Sweden, the samples were stored at - 20 °C until further processing. Staff taxonomists then sorted out specimens from the fly family Phoridae, with no attempt to identify or sort them to lower taxonomic levels. We then randomly selected for processing 88 males and the same number of females representing each of the six sites and two seasons (except for three combinations where 87, 86

and 84 were available), totalling 2,105 insects. As explained further, data for 1,842 of these samples (87.5%) passed our stringent quality criteria and were used for the analyses (Supplementary Table S1).

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 (Buczek et al., 2023). 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 (Tourlousse et al., 2017), placed in a 96-well PCR plate, and purified using SPRI beads (Rohland & Reich, 2012). DNA was eluted in 20 µl of TE buffer.

The amplicon libraries were prepared using a two-step PCR procedure (method 4 in Glenn et al., 2019). 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 (Parada et al., 2016), and for a 418 bp portion of the mitochondrial cytochrome oxidase I (COI) gene, primers BF3: CCHGAYATRGCHTTYCCHCG and BR2: TCDGGRTGNCCRAARAAYCA (Elbrecht et al., 2019). The PCR reaction mix also contained 500 copies of another plasmid carrying an artificial 16S rRNA target, Ec5001 (Tourlousse et al., 2017), 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 (Zhang et al., 2014), VSEARCH (Rognes et al., 2016), and USEARCH

(Edgar, 2010) 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 ≤ 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 (Antich et al., 2021). 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 (Edgar, 2016), and taxonomic reference databases: MIDORI version GB 239 (Leray et al., 2018), with alphaproteobacterial sequences added, for the COI data and SILVA (Quast et al., 2012) 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 (Hartop et al., 2022) 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 (Hartop et al., 2022), 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 bacterial COI data primarily comprised sequences assigned to the order Rickettsiales, which we used in 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. 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 resolve the overall structure of the insect microbiome, we characterised individual microbiomes in terms of the overall bacterial abundance, the identity and distribution of diverse bacteria at 16S rRNA OTU and ZOTU levels, and the identity and distribution of *Wolbachia* at COI ZOTU level.

To examine the drivers of variation in insect microbiomes, we built Joint Species Distribution Models (JSDMs; (Warton et al., 2015)) of the data using HMSC (Hierarchical Modelling of Species Communities; (Ovaskainen et al., 2017; Ovaskainen & Abrego, 2020). For these analyses, we selected a subset of the decontaminated data corresponding to 75 species represented by at least 5 individuals - 1634 samples in total. As the abundance response in the models, we considered the estimated copy numbers of either the 50 most abundant 16S-V4 OTUs or the 100 most dominant 16S-V4 ZOTUs. For the most dominant 24 *Wolbachia* COI ZOTUs, we used information about the number of reads instead. To test for effects of the season (summer versus spring), the site (n = 6), the sex of the host individual (male versus female), the host insect species (n = 75), the host insect genotype (n = 328), and the sampling unit (n=1634), we fitted a separate HMSC model to each response.

Due to the zero-inflated nature of the data (i.e. the many absences of individual taxa, corresponding to counts of zero), we modelled abundances through a hurdle approach: for each of the datasets (16S-V4 OTUs, 16S-V4 ZOTUs, *Wolbachia* COI ZOTUs) we first modelled presence-absences, then abundance conditional on presence. In modelling presence-absences, we assumed a Bernoulli distribution and a probit link function; in modelling abundances conditional on presence, we log-transformed and scaled the response, assuming a normal distribution. Season and sex were modelled as fixed effects, site, host insect species, host insect genotype, and host insect individual as community-level random effects modelled through latent variables.

We assumed the default prior distribution of HMSC (Ovaskainen & Abrego, 2020). We fitted the model using the Markov Chain Monte Carlo (MCMC) approach implemented in the R-package Hmsc (Tikhonov et al., 2020). 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 (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 most 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, ca. 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 (Hartop et al., 2022). Obtained species presented significant variation in genotype numbers and abundance (Fig. 1). 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. 2A, Supplementary Table S2).

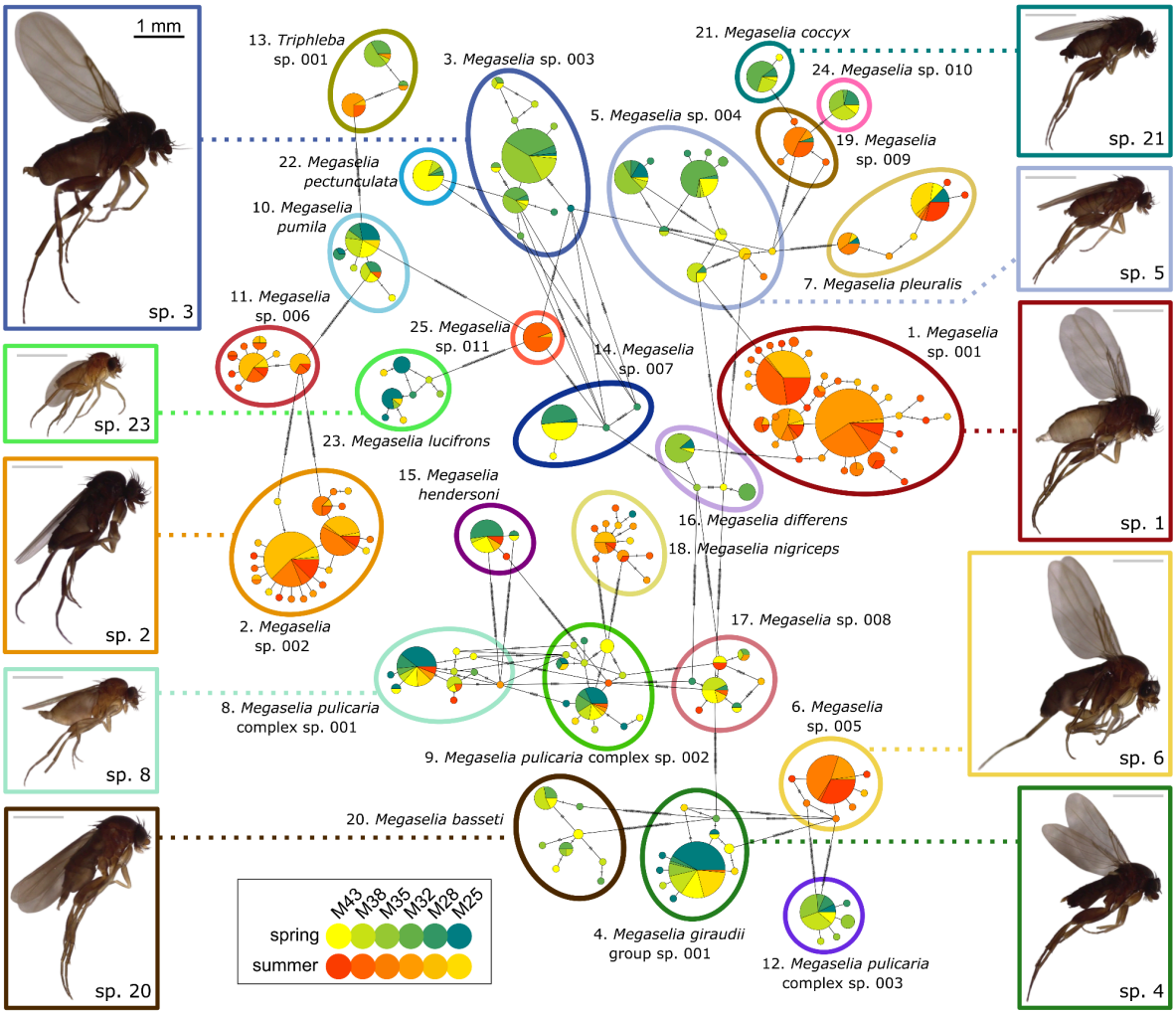
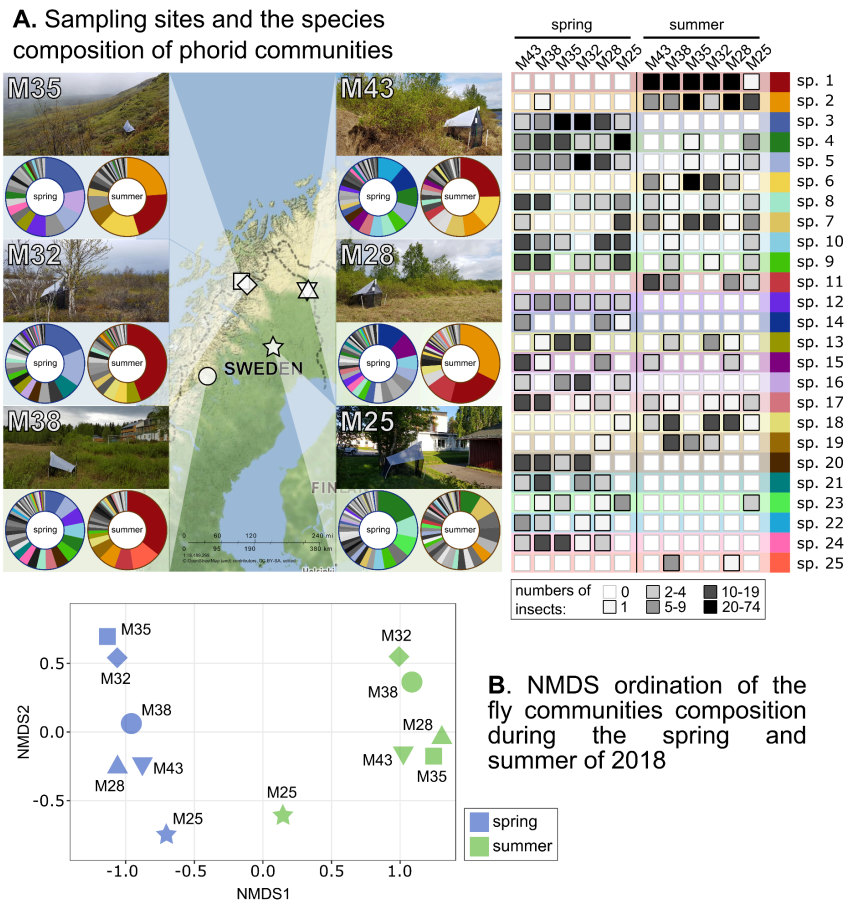


Fig. 1. Haplotype network of the 25 most abundant scuttle fly species, coloured by site and season. Selected species are represented by photos of individually barcoded individuals from various sites.

Genotypes of the remaining 161 species are not shown. The network was drawn using PopArt (Leigh & Bryant, 2015). Representative individuals for 10 selected species are pictured, with all images scaled uniformly for consistency.

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. 2A). 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. 2A). NMDS based on Bray-Curtis distances clearly separated spring and summer community samples (Fig. 2B).



B. NMDS ordination of the fly communities composition during the spring and summer of 2018

Fig. 2. The distribution and diversity of sampled phorid flies across two seasons and six sites in Northern Sweden. **A.** Map of the sampling sites and the species composition of fly communities across sites and dates (for both sexes combined), and the distribution of the 25 most common species across sites and seasons. **B.** A graphical representation of differences in phorid community composition among sites and seasons (NMDS plot based on Bray-Curtis dissimilarity).

3.2. Abundance and composition of phorid microbiota

Across 1,842 phorid samples, a total of 28M (M = million) reads scored as non-contaminants. They represented 58.1K ZOTUs and 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 (Buczek et al., 2024; Tourlousse et al., 2017), 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. 3A).

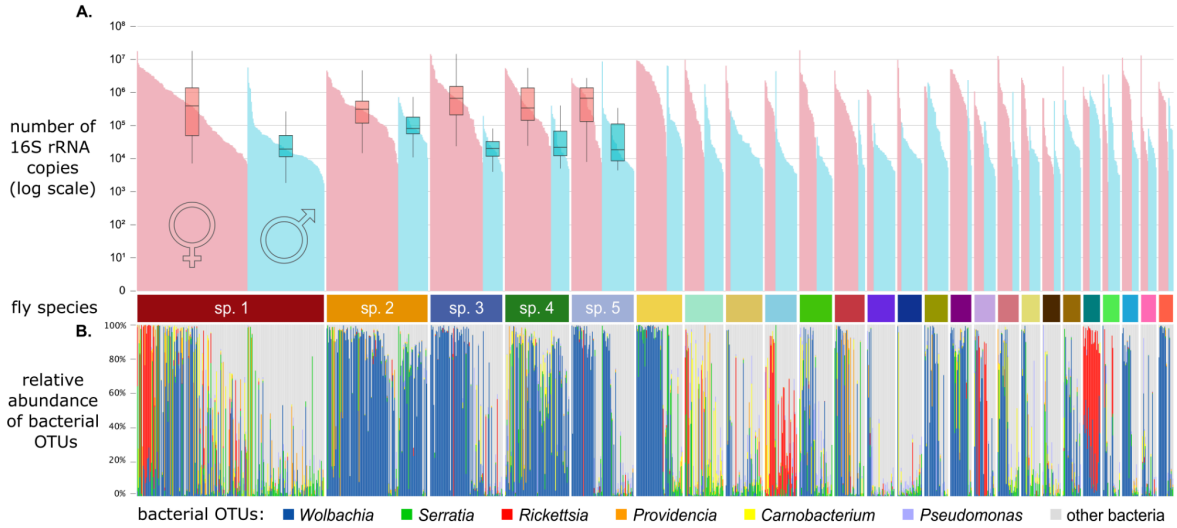


Fig. 3: Distribution of the absolute abundance and composition of bacteria across 1,193 specimens of 25 dominant phorid species. **A.** The distribution of bacterial 16S-V4 rRNA copy numbers across individual insects, estimated based on quantification spike-ins added during the DNA extraction step. Here, each individual 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). On average, the facultative endosymbiont *Wolbachia* OTU1 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 (Table 1; 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. 3B). For example, *Rickettsia* comprised a particularly large share of the microbiota of females of species 1 with high overall bacterial abundance.

Table 1: The prevalence and abundance of the 50 most common 16S-V4 OTUs in the 75 fly species used in the modelling. We considered individuals as infected when they had at least 100 estimated copies of 16S rRNA. The mean number of 16S copies was calculated only for the infected individuals.

16S OTU label	Class	Genus	Infected species [%]	Infected females [%]	Log_{10} 16S copy no. in infected females (mean ± SD)	Infected males [%]	Log_{10} 16S copy no. in infected males (mean ± SD)
OTU1	Alphaproteobacteria	*Wolbachia*	90.7	85.6	5.05 ± 1.21	40.4	3.87 ± 1.13
OTU23	Alphaproteobacteria	*Wolbachia*	26.7	2.9	3.69 ± 1.23	0.9	3.70 ± 1.23
OTU5	Alphaproteobacteria	*Rickettsia*	80.0	29.3	3.87 ± 1.41	13.9	3.63 ± 1.18
OTU43	Alphaproteobacteria	*Sphingomonas*	97.3	45.8	2.81 ± 0.51	40.4	2.70 ± 0.47
OTU36	Alphaproteobacteria	*Hepatincola*	13.3	2.1	3.78 ± 1.52	0.6	2.50 ± 0.36
OTU78	Alphaproteobacteria	*Rhizobium*	82.7	23.0	2.70 ± 0.52	14.8	2.61 ± 0.52
OTU100	Alphaproteobacteria	*Allorhizobium*	84.0	26.4	2.75 ± 0.52	14.8	2.62 ± 0.43
OTU28	Betaproteobacteria	*Janthinobacterium*	76.0	21.9	3.00 ± 0.83	12.9	2.55 ± 0.40
OTU115	Betaproteobacteria	*Burkholderia*	69.3	16.0	2.79 ± 0.58	11.5	2.56 ± 0.43
OTU133	Betaproteobacteria	*Collimonas*	68.0	17.0	2.79 ± 0.45	12.0	2.63 ± 0.43
OTU6	Gammaproteobacteria	*Serratia*	98.7	79.4	3.45 ± 1.04	68.3	3.01 ± 0.84
OTU7	Gammaproteobacteria	*Providencia*	94.7	48.7	3.47 ± 1.18	30.8	2.75 ± 0.69
OTU8	Gammaproteobacteria	*Pseudomonas*	97.3	67.4	3.24 ± 0.90	59.1	2.80 ± 0.57
OTU12	Gammaproteobacteria	*Pseudomonas*	86.7	35.2	2.93 ± 0.68	27.2	2.66 ± 0.51
OTU13	Gammaproteobacteria	*Erwinia*	97.3	59.6	3.10 ± 0.83	43.6	2.84 ± 0.61
OTU24	Gammaproteobacteria	*Rickettsiella*	65.3	18.1	2.94 ± 0.85	8.5	2.67 ± 0.47
OTU44	Gammaproteobacteria	*Legionella*	49.3	8.2	2.88 ± 0.77	4.9	2.68 ± 0.56
OTU51	Gammaproteobacteria	*Stenotrophomonas*	81.3	17.5	2.95 ± 0.78	10.8	2.69 ± 0.59
OTU57	Sphingobacteriia	*Pedobacter*	85.3	24.1	2.84 ± 0.58	23.3	2.63 ± 0.45
OTU47	Sphingobacteriia	*Pedobacter*	89.3	23.8	2.84 ± 0.63	22.2	2.63 ± 0.41
OTU18	Flavobacteriia	*Chryseobacterium*	81.3	30.4	2.94 ± 0.66	22.9	2.66 ± 0.52
OTU33	Flavobacteriia	*Chryseobacterium*	73.3	13.7	2.85 ± 0.68	12.0	2.58 ± 0.55
OTU9	Mollicutes	*Spiroplasma*	62.7	7.9	3.67 ± 1.44	9.1	3.89 ± 1.42
OTU10	Bacilli	*Carnobacterium*	97.3	61.9	3.52 ± 1.05	33.1	2.79 ± 0.59
OTU59	Bacilli	*Lactobacillus*	56.0	17.9	2.77 ± 0.57	16.3	2.73 ± 0.43
OTU49	Bacilli	*Enterococcus*	76.0	19.9	2.88 ± 0.73	13.0	2.65 ± 0.44
OTU32	Bacilli	*Brochothrix*	69.3	14.3	2.93 ± 0.84	11.0	2.62 ± 0.43
OTU128	Bacilli	*Bacillus*	69.3	8.5	2.71 ± 0.45	8.5	2.76 ± 0.54
OTU20	Actinomycetes	*Amycolatopsis*	100.0	65.2	3.04 ± 0.55	67.1	2.92 ± 0.51
OTU19	Actinomycetes	*Streptomyces*	97.3	64.7	3.05 ± 0.56	62.5	2.89 ± 0.50
OTU31	Actinomycetes	*Mycobacterium*	100.0	65.7	2.91 ± 0.47	65.9	2.76 ± 0.45
OTU38	Actinomycetes	*Umezawaea*	98.7	50.5	2.84 ± 0.49	55.0	2.78 ± 0.49
OTU35	Actinomycetes	*Streptomyces*	97.3	52.5	2.86 ± 0.48	53.6	2.85 ± 0.50
OTU46	Actinomycetes	*Kribbella*	96.0	38.9	2.79 ± 0.47	42.9	2.73 ± 0.44
OTU53	Actinomycetes	*Pseudonocardia*	97.3	43.2	2.72 ± 0.43	48.0	2.64 ± 0.43
OTU63	Actinomycetes	*Pseudonocardia*	93.3	42.0	2.70 ± 0.42	45.2	2.70 ± 0.40
OTU37	Actinomycetes	*Cryocola*	96.0	46.2	2.90 ± 0.58	35.3	2.64 ± 0.42
OTU65	Actinomycetes	*Nocardia*	86.7	31.4	2.84 ± 0.56	31.2	2.66 ± 0.42
OTU81	Actinomycetes	*Goodfellowiella*	92.0	29.9	2.66 ± 0.43	29.2	2.68 ± 0.45
OTU52	Actinomycetes	*Aeromicrobium*	94.7	38.0	2.78 ± 0.49	34.7	2.63 ± 0.38
OTU26	Actinomycetes	*Pseudarthrobacter*	73.3	24.5	3.08 ± 0.81	13.7	2.64 ± 0.52
OTU48	Actinomycetes	*Rhodococcus*	92.0	32.5	2.86 ± 0.60	25.8	2.66 ± 0.48
OTU92	Actinomycetes	*Actinospica*	62.7	18.6	2.81 ± 0.51	16.3	2.63 ± 0.47
OTU77	Actinomycetes	*Corynebacterium*	73.3	20.6	2.72 ± 0.49	10.8	2.67 ± 0.52
OTU62	Actinomycetes	*Turicella*	50.7	5.0	2.96 ± 0.86	6.4	2.64 ± 0.54
OTU60	Actinomycetes	*Rhodococcus*	77.3	20.3	2.97 ± 0.70	14.4	2.52 ± 0.40
OTU76	Actinomycetes	*Kocuria*	69.3	10.4	2.45 ± 0.31	9.4	2.63 ± 0.50
OTU93	Actinomycetes	*Cellulomonas*	70.7	16.8	2.65 ± 0.52	10.4	2.64 ± 0.40
OTU17	unassigned	unassigned	98.7	46.4	3.15 ± 0.67	42.8	2.94 ± 0.67
OTU30	unassigned	unassigned	89.3	49.5	3.08 ± 0.64	47.5	2.90 ± 0.51

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. 4). 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.

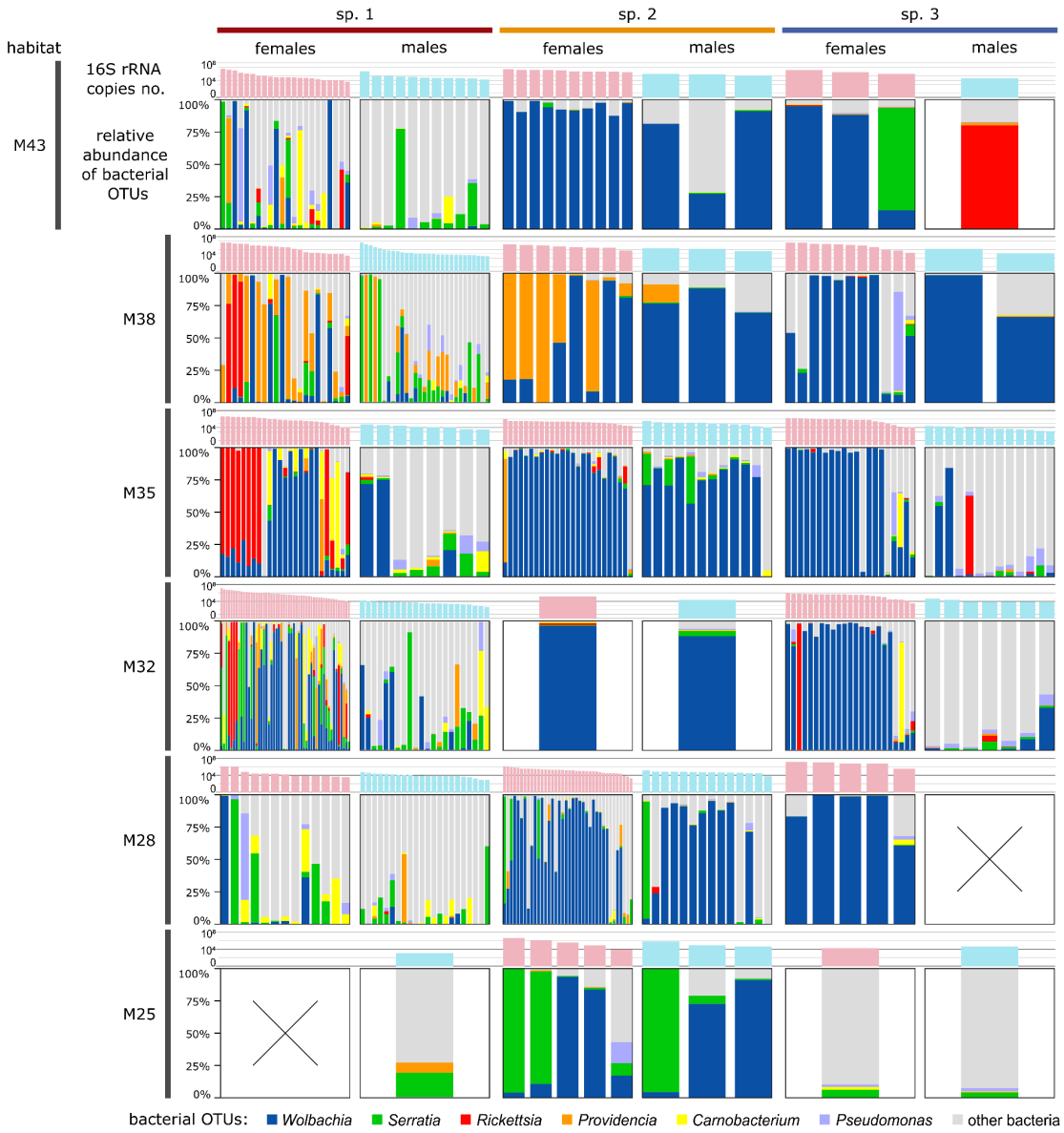


Fig. 4: Variation in bacterial absolute abundance and relative abundance of dominant bacterial OTUs within three abundant phorid species across six sites. Individual columns of varying width show individuals, with colours identifying the relative proportion of 16S-V4 OTU reads.

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 from each other in how their component ZOTUs are distributed across host species (Fig. 5). Specifically, insect species were found to preferentially associate with only one or two specific 16S rRNA genotypes of facultative endosymbiont *Wolbachia* and *Rickettsia*. Furthermore, different species associate with different facultative endosymbiont genotypes. Conversely, different facultative endosymbiont genotypes colonize 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 (Fig. 5; Supplementary File S3).

Fig. 5. The distribution of the most abundant genotypes of six dominant bacterial OTUs across 1,193 specimens of the 25 most abundant scuttle fly species. Insect individuals, represented by columns, are sorted by species, sex, and decreasing bacterial abundance (as in Fig. 3). 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.

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 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. 6A) resembles *Wolbachia* 16S rRNA genotype distribution patterns (Fig. 5). These data show that these fly 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. 6B) provide much more detailed information: they show that genotypes of the same fly species frequently associate with different genotypes of *Wolbachia*, with substantial variation 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 many less abundant species.

Within species, we found even greater specificity in host-symbiont interactions with respect to sex and site (Fig. 6C). For example, in sp. 5., *Wolbachia* COI ZOTU495 colonised most host genotypes, and tend 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.

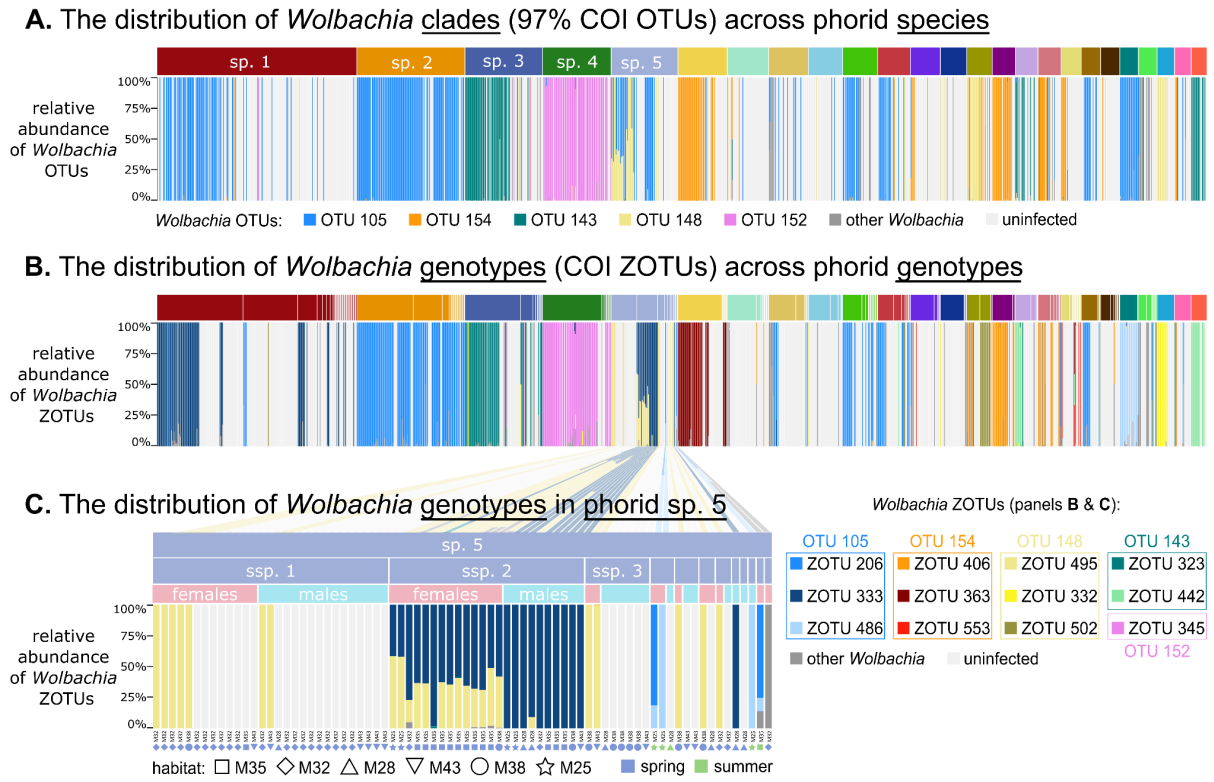


Fig. 6: The distribution of *Wolbachia* infections across the 25 most abundant phorid species, based on bacterial COI data. **A.** The relative abundance of *Wolbachia* COI OTUs among all *Wolbachia* COI reads across individuals represented by columns, and sorted by species, sex and decreasing abundance (as in Fig. 2); **B.** The relative abundance of *Wolbachia* COI genotypes (ZOTUs) across individual flies sorted into species and 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

With patterns dissected from different perspectives above, we observed substantial variation in the prevalence and abundance of microbiota at all levels: among host species, sexes, and genotypes, as well as among sites and seasons. Variation was also evident at different levels of resolution: microbial species equivalents (16S rRNA 97% OTUs), 16S rRNA genotypes, and for *Wolbachia*, *also* COI genotypes. To summarise the relative contribution of different factors to this overall variation, we turn to the HMSC analyses.

3.3.1 Results of the HMSC models

Overall, our predictors of presence vs. absence of bacterial 16S rRNA OTUs and ZOTUs (i.e. the site, the season, the host species, and the sex and the genotype of the host individual) accounted for a substantial proportion of the variation observed (Table 2; columns Presence/absence - 16S-V4 OTUs and 16S-V4 ZOTUs). Both the explanatory and predictive powers of the HMSC models were high (AUC ≥ 0.69). The effect of the host individual, followed by the site, explained the largest share of variance in the occurrence of 16S rRNA OTUs and ZOTUs. Of the microbial taxa showing significant responses to season, most were more prevalent in the spring than in the summer. Among the taxa responding significantly to the sex of the host insect, most microbial taxa were more prevalent in females than in male hosts.

Across bacterial 16S rRNA OTUs and ZOTUs, the results on abundance conditional on presence contrasted with the results of presence-absence analyses. While these models reached relatively high explanatory power, their predictive power was low (Table 2; columns Abundance conditional on presence, 16S-V4 OTUs and 16S-V4 ZOTUs). This is because, in the abundance conditional on presence models, the random effect of the host individual (which cannot be utilised for prediction) explained a very high amount of variation (Table 2; Fig. 7), with associations among taxa being mostly positive (Fig. 8). Thus, some host individuals had a high abundance of most bacterial taxa, whereas other host individuals had a low abundance of most bacterial taxa, amplifying the corresponding result of the presence-absence models. Consistently with the presence-absence models, abundance conditional on presence was higher in females than in male host individuals. In contrast to the presence-absence patterns, abundance conditional on presence was generally higher in summer than in spring. Thus, spring samples were characterised by the presence of many taxa in low abundance, whereas summer samples were characterised by the presence of fewer taxa in high abundance.

In terms of the presence vs. absence of specific COI genotypes of *Wolbachia* (Table 2, column presence/absence - *Wolbachia* COI ZOTUs), the explanatory and predictive powers of the model were exceptionally high (AUC ≥ 0.90). Presences vs. absences were strongly influenced by host species and host genotype, with additional imprints of season and site. Most *Wolbachia* strains were more

prevalent in the spring than in the summer and more prevalent in female hosts than male hosts. In terms of the abundance of specific COI strains of *Wolbachia* conditional on presence (Table 2, column Abundance conditional on presence - *Wolbachia* COI ZOTUs), the results mimicked those across other microbial clades: explanatory power was exceptionally high (AUC = 0.94), but predictive power was very low – as the highest proportion of variation (36%) occurred among individuals. Abundances differed between sexes, and where a difference emerged, the strain-specific abundance of *Wolbachia* was always higher in females than males.

Table 2. Results of the HMSC models. Model fit is averaged over the taxa included in each model and given in units of AUC for the presence-absence models (first three columns) and in units of R^2 for the abundance conditional on presence models (latter three columns). Variance partitioning is also averaged over the taxa included in each model and given in units of percentage (%) of explained variation. Responses to sex and season show the percentages of bacterial taxa that showed a preference for either season or either sex with at least 95% posterior probability.

Response	Presence/absence			Abundance conditional on presence		
	16S-V4 OTUs	16S-V4 ZOTUs	*Wolbachia* COI ZOTUs	16S-V4 OTUs	16S-V4 ZOTUs	*Wolbachia* COI ZOTUs
Model fit [%]						
Explanatory power	80	83	99	46	53	94
Predictive Power	69	71	90	11	14	1
Variance partitioning [%]						
FE: Season	14	16	23	7	9	8
FE: Sex	5	5	19	11	13	20
RE: Host species	16	13	27	10	11	14
RE: Host genotype	11	13	24	7	10	15
RE: Host individual	36	28	2	55	48	36
RE: Site	18	26	4	10	9	7
Responses to season and sex [%]						
spring > summer	70	66	46	2	0	0
spring < summer	8	15	21	58	74	0
female > male	56	48	75	92	94	38
female < male	2	6	0	0	0	0

At the level of individual microbial taxa, the total variance explained (here measured by Tjur's R^2) differed significantly among taxa. In general, more prevalent microbial taxa exhibited higher values of Tjur R^2 (Supplementary Figure S1). The contributions of specific drivers of variance also differed among taxa (Fig. 7).

3.3.2 Drivers of variation in bacterial clade occurrence and abundance

Of the occurrence (presence/absence) of individual 16S-V4 OTUs, the HMSC model explained up to 40% of the variance (Fig. 7A). Much of this variation was found among host individuals. For facultative endosymbionts: *Wolbachia* and *Rickettsia*, however, a different pattern emerged: here much variation occurred among host species, sex, and genotypes. For *Lactobacillus*, most variation occurred among seasons (Fig. 7A).

Of abundance conditioned on presence, the HMSC model explained up to 70% of abundance variation in individual OTUs (Fig. 7A). Notably, facultative endosymbionts (i.e., *Wolbachia, Rickettsia* and *Spiroplasma*) showed little variation among host individuals and different contributions of specific drivers. For *Wolbachia*, we found much variation associated with host species, sex and genotype. For *Rickettsia,* the effect of host genotype was particularly pronounced; for *Spiroplasma,* the effect of sex was minimal.

Similarly to microbial 16S-V4 OTUs (above), presence-absence patterns of individual 16S-V4 ZOTUs showed high variability across taxa – including ZOTUs belonging to the same OTUs (Fig. 7B). Values of Tjur R^2 were highest for *Wolbachia*, where most variation was attributed to host species and host genotype. Variation at the level of the insect individual was most pronounced among Actinomycetes (Fig. 7B).

For drivers of variation in abundance, facultative endosymbionts in Alphaproteobacteria and Mollicutes clearly stood out from the rest of the taxa (Fig. 7B). Here, most of the variation was associated with fly species and genotype, whereas, for other taxa, the host individual contributed the most variation (Fig. 7B).

At the level of presence of individual *Wolbachia* strains resolved by the COI marker (Fig. 7C), very little variation occurred among host individuals. Rather, the variance was dominated by other contributions, which differed among individual *Wolbachia* ZOTUs. Even within the same OTU, some *Wolbachia* genotypes showed a strong imprint of host genotype or host species, while others were more affected by the host sex or site (Fig. 7C). In terms of abundance (conditional on presence), we note a high contribution of the host individual in comparison with presence-absence. Nonetheless, host species, host genotype and host sex all contributed substantially to overall variation, with some variation among individual ZOTUs.

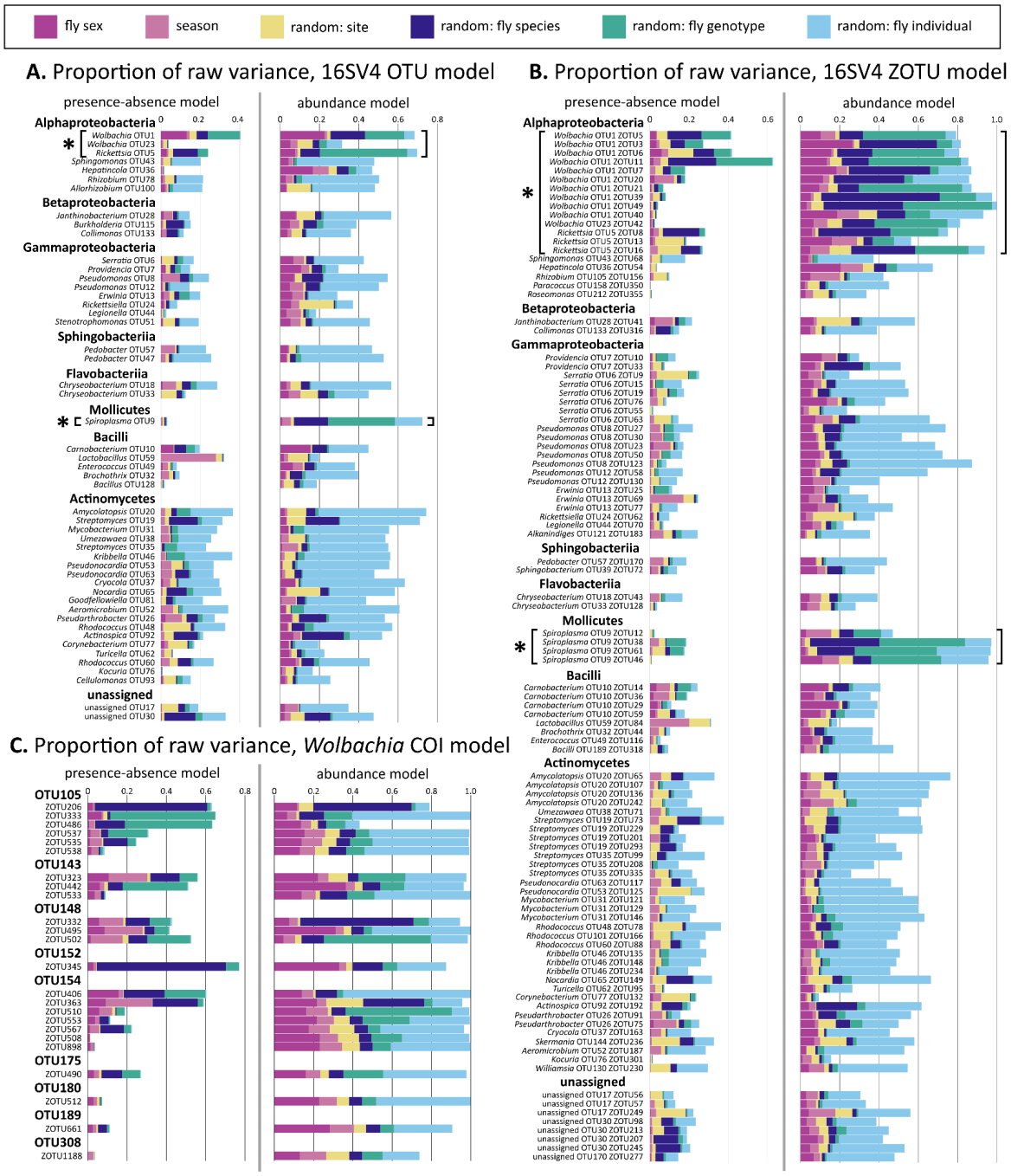


Fig. 7. Proportion of variance explained per microbial taxon by the presence-absence model and the abundance conditional of presence models at different levels of variation: A) microbial 16S-V4 OTUs, B) 16S-V4 ZOTUs, C) *Wolbachia* COI ZOTUs. In (A) and (B), asterisks (*) identify bacterial taxa known as facultative endosymbionts.

3.3.3 Residual associations among microbial taxa

After accounting for the joint impact of systematic effects (3.3.2), we turned to patterns of residual associations. This approach revealed what microbial taxa may influence each other's presence and abundance *beyond* responding to the same external drivers. Such associations can be addressed at the level of each random factor, i.e. within host species, genotypes, and sites.

At the level of host species, numerous statistically supported associations in the presence or absence of microbial taxa were evident (Fig. 8A). The microbial taxa showed positive residual co-occurrence within the groups and negative residual co-occurrence among the groups. These blocks were associated with microbial taxonomy: for example, in terms of presence, Actinomycetes OTUs were often positively associated with each other, and with some Betaproteobacteria, Gammaproteobacteria, and unclassified. The patterns were even more striking in terms of OTU abundance conditional on presence, where Actinomycetes split into two clusters. One cluster of Actinobacteria OTUs was positively associated with each other and most Betaproteobacteria and unassigned. The other cluster was positively associated with several other taxa such as Gammaproteobacteria and Bacilli. However, the two clusters were negatively associated with each other. These general presence-absence and abundance patterns at the species level were largely replicated for ZOTUs (Fig. 8B).

At the level of host genotypes, a block of positive associations in presence-absence emerged among several OTUs of Actinomycetes and several Gammaproteobacteria, Bacilli, and unassigned. For ZOTUs, Actinomycetes again split into two clusters. One of these clusters was negatively associated with many genotypes from several other clades (including the other Actinomycetales cluster); in contrast, the other Actinomycetales cluster was associated positively in terms of presence with Bacilli, Gammaproteobacteria and Alphaproteobacteria. Interestingly, in terms of abundance conditional on presence, associations were less consistent and relatively few statistically supported associations in abundance were discovered, especially for ZOTUs.

At the level of host individuals, practically all microbial taxa except Alphaproteobacteria, Mollicutes and some Actinomycetes showed consistent positive associations among each other, in both presence-absence and abundance conditional on presence, at OTU and ZOTU levels. This suggests that host individuals tended to have either very high or very low bacterial diversity, with little structuring among individual microbial taxa. Interestingly, most Alphaproteobacterial facultative endosymbionts showed no statistically supported associations with other microbes in terms of presence-absence, at OTU and ZOTU levels. Facultative endosymbionts in the class Mollicutes showed positive associations with other microbes at the level of OTUs but not ZOTUs. This suggests a limited effect of facultative endosymbionts on the occurrence of most other microbial taxa. At the same time, in terms of abundance conditional on presence, most microbial clades including facultative endosymbionts associated positively with each other.

At the level of sites, we found strong but mixed patterns of associations in the presence/absence of OTUs or ZOTUs. Most statistically supported associations were positive, with the notable exception of two *Wolbachia* OTUs, *Spiroplasma, Legionella* and a few Actinomycetes OTUs, which formed negative associations with almost all other taxa. In terms of abundance conditional on

presence, associations were also largely positive, except for some Alphaproteobacteria and Actinomycetes taxa.

Among individual strains of *Wolbachia* (Fig. 8C), we detected relatively fewer statistically supported associations at any level. Some *Wolbachia* COI ZOTUs were positively or negatively associated with others in terms of presence-absence mostly at the levels of host species and sites. There were very few significant associations in terms of abundance conditional on presence. Nonetheless, this may be due to low statistical power, as most *Wolbachia* COI ZOTUs occurred in only a minority of host species and individuals.

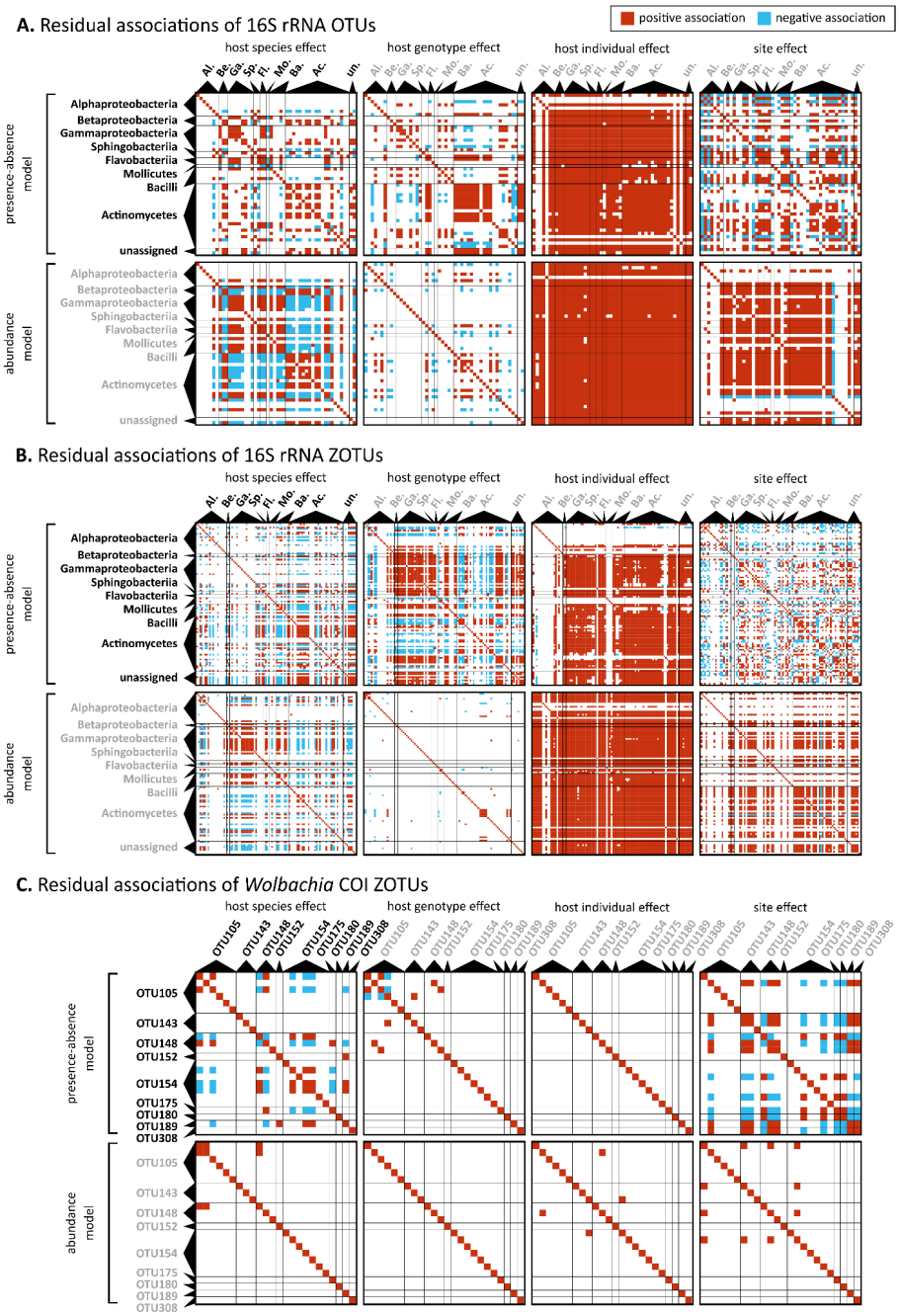


Fig. 8. Residual associations among bacterial taxa. The columns of panels correspond to the four random levels included in the HMSC models, and the rows of panels represent two model components. Within each panel, the bacterial taxa were ordered in the same way as in the presence-absence part of the 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

Biodiversity largely consists of 'dark taxa' such as scuttle flies, which dominate in terms of abundance and diversity (Caruso et al., 2024; Li & Wiens, 2023; Srivathsan et al., 2023). However, the large share of undescribed species and the difficulty of identifying the few known species have made it challenging to address even basic ecological questions (Hartop et al., 2024). Understanding how these species interact with one another is arguably one of the most fundamental ones (Roslin, 2024). New conceptual approaches combined with progress in sequencing, bioinformatics, and statistical modelling are now changing the playing field (Knight et al., 2018; Łukasik & Kolasa, 2024). This study showcases how several such innovations, especially when applied in combination, can greatly enhance our understanding of interactions among wild non-model organisms. Below, we explain what this study has taught us about scuttle flies and their microbiota, and what scuttle flies have taught us about the patterns and processes of host-microbe interactions relevant beyond this 'dark taxon'. Further, we discuss the utility of the key methodological innovations used in this study and explore new avenues for the characterisation of insect-microbe associations across diverse natural communities that they open up.

4.1. What have we learnt about scuttle flies?

While scuttle flies are essentially a pitch-dark taxon (Hartop et al., 2024), the current study shows how scuttle fly communities 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 the Grinnellian niche of species – i.e., how communities are formed from the match between abiotic conditions and species requirements (Grinnell, 1917; Holt, 2009; Soberón, 2019). This has largely neglected the other niche dimension, i.e. how species occurrence is determined by their interactions with other species (the Eltonian niche (Colwell & Rangel, 2009; Hirzel & Le Lay, 2008)). The current approach allows us to characterise the Grinnellian and the Eltonian niche of species in one go. By simultaneously identifying the host species (here: scuttle flies) and their associated taxa (here: microbes) across sets of sites, we may effectively characterise the entire wild communities of multi-partite symbiotic entities, or holobionts (Bordenstein & Theis, 2015). 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 (Hartop et al., 2024). By adding the microbial dimension, we show the way forward for the integration of the two (Gravel et al., 2019).

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 large variation in bacterial abundance – a critical aspect of microbiota rarely assessed in microbiome studies (Hammer et al., 2019; Tourlousse et al., 2017; Williamson et al., 2024), with 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, despite variations in rRNA operon copy numbers within bacterial genomes (Klappenbach, 2001; Louca et al., 2018), and copies of the genome per cell (Pecoraro et al., 2011). In turn, this should correlate with the functional significance of bacteria in the host biology, even if the density-dependence of bacterial effects on the host (Koga et al., 2022; Medina Munoz et al., 2020) 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 variation among species and sexes. These two widespread microbes show highly specific patterns in their distribution, with host species, genotype, and sex explaining a large share of the overall variance in their presence/absence and abundance at the levels of 16S rRNA OTU and genotype and, for *Wolbachia*, COI genotypes. 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 (Bailly-Bechet et al., 2017; Kaur et al., 2021; Turelli et al., 2018). 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 (Sanaei et al., 2021).

Wolbachia is thought to invade 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 as well as abundance among males and females (Table 2, Figure 9) argue for a negative link between infection and male functions. These patterns are consistent with expectations for some types of reproductive manipulation: male killing, feminization of genetic males, or parthenogenesis induction (Engelstädter & Hurst, 2009; Kaur et al., 2021). 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 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.

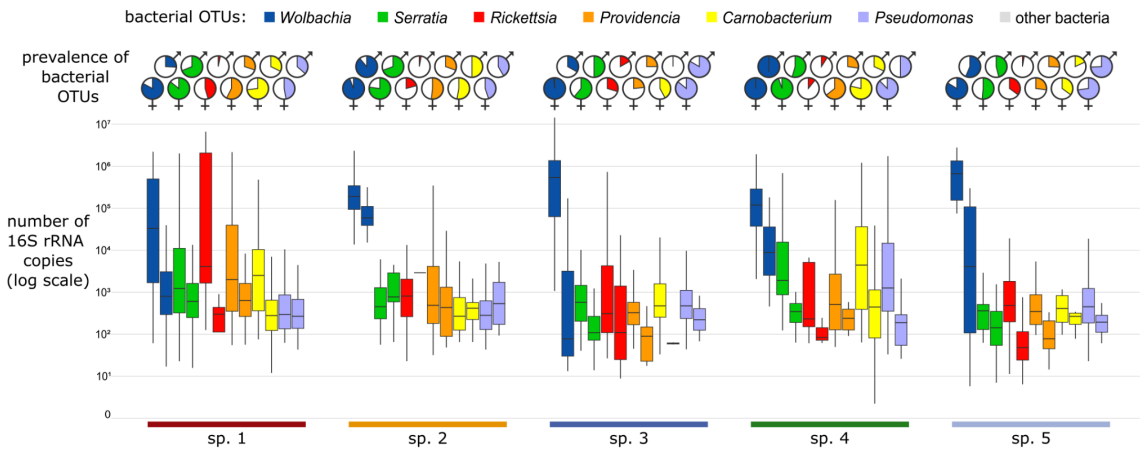


Fig. 9: The prevalence of infection with the six most abundant bacterial OTUs across males and females of the five most abundant scuttle fly species, and the estimated absolute abundance of 16S rRNA gene copies in individuals scored as infected. The differences among sexes in the prevalence and abundance of *Wolbachia,* evident in most species, are consistent with expectations for some types of reproductive manipulations. Insects were scored as infected if the estimated 16S rRNA copy number of a particular bacterial OTU was at least 100.

On the other hand, many of the abundant phorid-associated bacteria other than facultative endosymbionts represent clades broadly distributed across environments. 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 (Lalucat et al., 2022). The same applies to most other abundant bacterial genera in our dataset, including *Serratia* (Williams et al., 2022), *Carnobacterium* (Leisner et al., 2007), *Providencia* (Dong et al., 2024), and others. As research on these bacterial genera to date has focused on a limited set of species of direct relevance to human health and economy, the 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 (Ochman et al., 1999) 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 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?

Several aspects of the microbial associations reported here are novel for insects. This includes the massive within-species variation in the abundance and prevalence of particular bacteria, and high host species- and genotype-specificity among facultative endosymbionts contrasted with broad distributions of other microbes. Given the lack of comparative data, it is plausible that these features are not unique to phorids. In fact, we propose that these observations are representative of general patterns, relevant to the broad insect diversity – resolved in our study by the rare combination of sampling breadth and comprehensive methods.

Perhaps the most universal finding of this study, and the key to the interpretation of most patterns, is the striking contrast among microbial functional categories in their host associations (Perreau et al., 2021). Two functional categories appear to have dominated the microbiota of phorids: facultative endosymbionts and environmentally-acquired bacteria, of which the latter form non-specialized associations with hosts. For future studies, it will be critical to distinguish among these and other functional categories when assessing insect-microbe association patterns, drivers, and likely consequences. At the same time, when characterising microbiota in a poorly known host clade, the distribution patterns can teach us about microbial biology and its role in the biology of the host.

Facultative endosymbioses, apparently represented in phorids by *Wolbachia*, *Rickettsia*, and *Spiroplasma* are often thought to form relatively dynamic associations with hosts. The whitefly *Bemisia tabaci,* the dipteran genus *Drosophila*, and growing genomic datasets for diverse insects provide examples of relatively recent symbiont host switches among related species (Shropshire et al., 2023; Turelli et al., 2018) and across distantly related species that form natural communities (Vancaester & Blaxter, 2023), as well as selective sweeps that they may drive (Himler et al., 2011; Jaenike, 2012; Kriesner et al., 2013; Zhao et al., 2023). These processes can dramatically influence insect population structure, performance in changing environmental conditions, or interactions with other organisms (Łukasik & Kolasa, 2024). Conversely, this can affect species' responses to ecological and environmental challenges and opportunities, their longer-term evolutionary trajectories, as well as broader community processes (Ferrari & Vavre, 2011; Oliver et al., 2010). However, we do not know how often these processes take place, or what are their primary drivers and likely consequences (Bailly-Bechet et al., 2017; Sanaei et al., 2021). 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 systematic testing of these hypotheses about facultative endosymbiosis stability and infection dynamics.

Likewise, microbial clades commonly encountered in phorids but not appearing to form specific associations are likely to have much broader distributions. Most genera listed in Table 2 have been repeatedly reported from many diverse insects from around the world - for example, from neotropical butterflies, or colonies of army ants (Łukasik et al., 2017; Ravenscraft et al., 2019; Valdivia et al., 2023). It is tempting to think of these microbes as highly mobile across global environments and host

clades. On the other hand, comparative data from vertebrates show how host biology may shape bacterial specificity, including strains from broadly distributed genera (Song et al., 2020). The systematic assessment of microbial diversity, transmission patterns, and roles across taxa and environments would require not only massive sampling and stringent contamination control (Salter et al., 2014) but also markers that provide much greater phylogenetic resolution than a short rRNA region - ideally, entire genomes.

At the same time, it is important to remember that phorids do not encompass the full range of insect-microbe associations known. Many insects form highly stable associations with specific bacterial clades, often maintained through nutritional dependencies: whether with obligate endosymbionts (Moran et al., 2008) or with host-adapted gut microbiota (Engel & Moran, 2013; Perreau et al., 2021). We would normally expect bacteria representing this category to be always present and abundant within the host species - and in surveys spanning multiple species, generally to show evidence of co-diversification with hosts (Fisher et al., 2017; Perreau & Moran, 2022). It is possible that 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 (Dedeine et al., 2005; Hosokawa et al., 2010).

4.4. Towards a comprehensive understanding of wild-insect microbiota

The novel sample processing, data analysis and interpretation strategies that we implemented in this study provide an accurate and comprehensive picture of microbial associations of diverse wild insects, well beyond what standard methods allow. Our workflow incorporates four major innovations that have rarely been used in insect microbiome studies - and definitely not together on a large scale.

First, we base our identification of both hosts and microbes on DNA barcoding, the approach that has already revolutionized the understanding of arthropod biodiversity (Meier et al., 2024). In insect microbiota surveys that address questions across large multi-species collections and/or dark taxa, DNA-based methods remain the only plausible technique for establishing the host identity and the essential phylogenetic framework. Host barcoding done simultaneously with microbiota characterisation is certainly the most cost-effective approach. Our approach, which enabled the delimitation of 1842 individuals into 480 genotypes and 186 species, permitted the organisation of unknown diversity into biologically meaningful units, generating valuable data on host diversity in its own right. However, more critically, these barcoding data provided a robust framework for reconstructing insect microbiota patterns that have provided virtually unprecedented resolution. In other insect systems, COI data may enable correcting misidentifications and reconstructing patterns such as parasitoid infections - potentially important confounding variables in microbiota studies (Kolasa et al., 2023).

Second, we quantify microbial abundance. Abundance information as an added dimension revolutionises the picture of the microbiota. While it has been known that insects differ dramatically in their microbial abundance, and while microbiota quantification has been repeatedly recommended (Hammer et al., 2019; Knight et al., 2018; Williamson et al., 2024), our approach shows how this can be achieved. Incorporating quantification spike-ins into sample processing and analysis workflows (Tourlousse et al., 2017) is an extremely cost-effective way of obtaining consistent abundance estimates spanning some six orders of magnitude. The patterns reconstructed are biologically realistic,

as revealed by non-random bacterial distributions within phorid species. Quantification helps not only to identify samples where bacterial abundances are high and community composition reliable but also to score samples with low bacterial abundances as such and treat community composition information with caution. This is clearly 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 or samples.

Third, we show that 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. Indeed, the phylogenetic value of the *Wolbachia* COI (or cox1) gene was realized over two decades ago when it was incorporated into the multi-locus sequence typing (MLST) scheme (Baldo et al., 2006). At the same time, the realization that the *Wolbachia* gene co-amplifies with insect mitochondrial COI in barcoding efforts (Smith et al., 2012) has motivated dedicated efforts to block its amplification (Bleidorn & Henze, 2021). Here we have considered *Wolbachia* COI 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, offering exciting prospects to extract infection information from rapidly growing barcoding datasets. At the same time, the improved phylogenetic resolution provided by COI amplicons was instrumental in reconstructing key biological patterns. They included indicating the rapid turnaround of this symbiont across phorid genotypes and its transmission across species - but also differences among symbiont genotypes in associations that they form.

Last but not least, joint species distribution modelling proved indispensable for systematically reconstructing the microbiota patterns and drivers. Where the massive dimensionality of microbial data offers a challenge to most analysis frameworks, Hierarchical Modelling of Species Communities has been developed specifically to deal with this complexity. However, its applications in the reconstruction of insect-bacteria interaction networks have so far been limited (Minard et al., 2019). On the other hand, in other host-microbe systems, the method has proved its power to reconstruct complex processes and patterns and their drivers (Parisy et al., 2024). In particular, it offers a way for disentangling species associations caused by joint preference for the same environmental conditions from actual interactions among the species. As we try to 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.

Despite rapid technological progress that revolutionized biodiversity monitoring, with further opportunities emerging (Van Klink et al., 2024), amplicon-sequencing methods will likely remain the tool of choice for addressing questions about microbiota, or other diverse biological interactions, across large numbers of diverse specimens. The solutions outlined above - combined with other recommended approaches such as contamination filtering (Davis et al., 2018; Knight et al., 2018; Williamson et al., 2024) - help address the key limitations of amplicon sequencing and put this powerful 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 (Łukasik & Kolasa, 2024). Such a combination of

tools and approaches can revolutionize 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 dark insect taxon. By simultaneously assessing the abundance and composition of bacteria across the host taxonomic framework, we find that individual flies vary enormously in their microbial associations, with likely consequences for any service/disservice delivered by the microbiota. We find striking differences among microbial functional categories: whereas facultative endosymbionts showcase specificity combined with occasional transmission among species, other abundant microbes in the dataset form largely unspecialized associations. 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 so far 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 P.Ł.). 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 figure, tables and files are available at:
https://github.com/KarolHub/Phorid-Microbiota/tree/main/Supplementary_materials

Supplementary Figure S1: Relationship between the explained variance (R^2) of bacterial taxa presence and their prevalence.

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, spikeins, 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 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

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, *22*(1), 177. https://doi.org/10.1186/s12859-021-04115-6

Bailly-Bechet, M., Martins-Simões, P., Szöllősi, G. J., Mialdea, G., Sagot, M.-F., & Charlat, S. (2017). How Long Does Wolbachia Remain on Board? *Molecular Biology and Evolution*, *34*(5), 1183–1193. https://doi.org/10.1093/molbev/msx073

Baldo, L., Dunning Hotopp, J. C., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R., Hayashi, C., Maiden, M. C. J., Tettelin, H., & Werren, J. H. (2006). Multilocus Sequence Typing System for the Endosymbiont *Wolbachia pipientis*. *Applied and Environmental Microbiology*, *72*(11), 7098–7110. https://doi.org/10.1128/AEM.00731-06

Baumann, P. (2005). BIOLOGY OF BACTERIOCYTE-ASSOCIATED ENDOSYMBIONTS OF PLANT SAP-SUCKING INSECTS. *Annual Review of Microbiology*, *59*(1), 155–189. https://doi.org/10.1146/annurev.micro.59.030804.121041

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 Research Notes*, *14*(1), 427. https://doi.org/10.1186/s13104-021-05845-9

Bordenstein, S. R., & Theis, K. R. (2015). Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. *PLOS Biology*, *13*(8), e1002226. https://doi.org/10.1371/journal.pbio.1002226

Buczek, M., Kolasa, M. R., Prus-Frankowska, M., Lipowska, M., Nowak, K. H., Azarbad, H., Franco, D. C., Marszałek, M., Roslin, T., Michalik, A., & Łukasik, P. (2024). *A new tool in a toolbox: Addressing challenges in high-throughput microbiota surveys across diverse wild insects*. https://doi.org/10.1101/2024.08.26.609764

Buczek, M., Prus-Frankowska, M., & Łukasik, P. (2023). *Quantitative multi-target amplicon sequencing workflow v1*. https://doi.org/10.17504/protocols.io.36wgq351ylk5/v1

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 Conservation and Diversity*, *17*(6), 968–987. https://doi.org/10.1111/icad.12762

Cheng, L., & Hogue, C. L. (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. *Entomological News*, *85*(7–8), 211–218.

Chua, P. Y. S., Bourlat, S. J., Ferguson, C., Korlevic, P., Zhao, L., Ekrem, T., Meier, R., & Lawniczak, M. K. N. (2023). Future of DNA-based insect monitoring. *Trends in Genetics*, *39*(7), 531–544. https://doi.org/10.1016/j.tig.2023.02.012

Colwell, R. K., & Rangel, T. F. (2009). Hutchinson's duality: The once and future niche. *Proceedings of the National Academy of Sciences*, *106*(supplement_2), 19651–19658. https://doi.org/10.1073/pnas.0901650106

Cornwallis, C. K., Van 'T Padje, A., Ellers, J., Klein, M., Jackson, R., Kiers, E. T., West, S. A., & Henry, L. M. (2023). Symbioses shape feeding niches and diversification across insects. *Nature Ecology & Evolution*, *7*(7), 1022–1044. https://doi.org/10.1038/s41559-023-02058-0

Danilova, N. (2006). The evolution of immune mechanisms. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *306B*(6), 496–520. https://doi.org/10.1002/jez.b.21102

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. https://doi.org/10.1186/s40168-018-0605-2

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*(5), 394–400. https://doi.org/10.1038/sj.hdy.6800739

Disney, R. H. L. (1991). The aquatic Phoridae (Diptera). *Insect Systematics & Evolution*, *22*(2), 171–191. https://doi.org/10.1163/187631291X00048

Disney, R. H. L. (1994). *Scuttle Flies: The Phoridae*. Springer Netherlands. https://doi.org/10.1007/978-94-011-1288-8

Disney, R. H., Taylor, S., Slay, M. E., & Krejca, J. K. (2011). New species of scuttle flies (Diptera: Phoridae) recorded from caves in Nevada, USA. *Subterranean Biology*, *9*, 73–84. https://doi.org/10.3897/subtbiol.9.2511

Dong, X., Jia, H., Yu, Y., Xiang, Y., & Zhang, Y. (2024). Genomic revisitation and reclassification of the genus *Providencia*. *mSphere*, *9*(3), e00731-23. https://doi.org/10.1128/msphere.00731-23

Douglas, A. E. (2015). Multiorganismal Insects: Diversity and Function of Resident Microorganisms. *Annual Review of Entomology*, *60*(1), 17–34. https://doi.org/10.1146/annurev-ento-010814-020822

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2016). *SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences*. https://doi.org/10.1101/074161

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M.,

Zakharov, E. V., Hebert, P. D. N., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, *7*, e7745. https://doi.org/10.7717/peerj.7745

Engel, P., & Moran, N. A. (2013). The gut microbiota of insects – diversity in structure and function. *FEMS Microbiology Reviews*, *37*(5), 699–735. https://doi.org/10.1111/1574-6976.12025

Engelstädter, J., & Hurst, G. D. D. (2009). The Ecology and Evolution of Microbes that Manipulate Host Reproduction. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 127–149. https://doi.org/10.1146/annurev.ecolsys.110308.120206

Ferrari, J., & Vavre, F. (2011). Bacterial symbionts in insects or the story of communities affecting communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1569), 1389–1400. https://doi.org/10.1098/rstb.2010.0226

Fisher, R. M., Henry, L. M., Cornwallis, C. K., Kiers, E. T., & West, S. A. (2017). The evolution of host-symbiont dependence. *Nature Communications*, *8*(1), 15973. https://doi.org/10.1038/ncomms15973

Frago, E., Dicke, M., & Godfray, H. C. J. (2012). Insect symbionts as hidden players in insect–plant interactions. *Trends in Ecology & Evolution*, *27*(12), 705–711. https://doi.org/10.1016/j.tree.2012.08.013

Glenn, T. C., Pierson, T. W., Bayona-Vásquez, N. J., Kieran, T. J., Hoffberg, S. L., Thomas Iv, J. C., Lefever, D. E., Finger, J. W., Gao, B., Bian, X., Louha, S., Kolli, R. T., Bentley, K. E., Rushmore, J., Wong, K., Shaw, T. I., Rothrock Jr, M. J., McKee, A. M., Guo, T. L., … Faircloth, B. C. (2019). Adapterama II: Universal amplicon sequencing on Illumina platforms (TaggiMatrix). *PeerJ*, *7*, e7786. https://doi.org/10.7717/peerj.7786

Gravel, D., Baiser, B., Dunne, J. A., Kopelke, J., Martinez, N. D., Nyman, T., Poisot, T., Stouffer, D. B., Tylianakis, J. M., Wood, S. A., & Roslin, T. (2019). Bringing Elton and Grinnell together: A quantitative framework to represent the biogeography of ecological interaction networks. *Ecography*, *42*(3), 401–415. https://doi.org/10.1111/ecog.04006

Grinnell, J. (1917). Field Tests of Theories Concerning Distributional Control. *The American*

Naturalist, *51*(602), 115–128. https://doi.org/10.1086/279591

Hague, M. T. J., Caldwell, C. N., & Cooper, B. S. (2020). Pervasive Effects of *Wolbachia* on Host Temperature Preference. *mBio*, *11*(5), e01768-20. https://doi.org/10.1128/mBio.01768-20

Hammer, T. J., Sanders, J. G., & Fierer, N. (2019). Not all animals need a microbiome. *FEMS Microbiology Letters*, *366*(10), fnz117. https://doi.org/10.1093/femsle/fnz117

Hartop, E. A., Brown, B. V., & Disney, R. H. L. (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*(4). https://doi.org/10.11646/zootaxa.3941.4.1

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 Biology*, *22*(1), 215. https://doi.org/10.1186/s12915-024-02010-z

Hartop, E., Srivathsan, A., Ronquist, F., & Meier, R. (2022). Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa. *Systematic Biology*, *71*(6), 1404–1422. https://doi.org/10.1093/sysbio/syac033

Himler, A. G., Adachi-Hagimori, T., Bergen, J. E., Kozuch, A., Kelly, S. E., Tabashnik, B. E., Chiel, E., Duckworth, V. E., Dennehy, T. J., Zchori-Fein, E., & Hunter, M. S. (2011). Rapid Spread of a Bacterial Symbiont in an Invasive Whitefly Is Driven by Fitness Benefits and Female Bias. *Science*, *332*(6026), 254–256. https://doi.org/10.1126/science.1199410

Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, *45*(5), 1372–1381. https://doi.org/10.1111/j.1365-2664.2008.01524.x

Holt, R. D. (2009). Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences*, *106*(supplement_2), 19659–19665. https://doi.org/10.1073/pnas.0905137106

Hosokawa, T., Koga, R., Kikuchi, Y., Meng, X.-Y., & Fukatsu, T. (2010). *Wolbachia* as a

bacteriocyte-associated nutritional mutualist. *Proceedings of the National Academy of Sciences*, *107*(2), 769–774. https://doi.org/10.1073/pnas.0911476107

Hurst, G. D. D., & Frost, C. L. (2015). Reproductive Parasitism: Maternally Inherited Symbionts in a Biparental World. *Cold Spring Harbor Perspectives in Biology*, *7*(5), a017699. https://doi.org/10.1101/cshperspect.a017699

Jaenike, J. (2012). Population genetics of beneficial heritable symbionts. *Trends in Ecology & Evolution*, *27*(4), 226–232. https://doi.org/10.1016/j.tree.2011.10.005

Jaenike, J., Unckless, R., Cockburn, S. N., Boelio, L. M., & Perlman, S. J. (2010). Adaptation via Symbiosis: Recent Spread of a *Drosophila* Defensive Symbiont. *Science*, *329*(5988), 212–215. https://doi.org/10.1126/science.1188235

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. *Biodiversity Data Journal*, *8*, e47255. https://doi.org/10.3897/BDJ.8.e47255

Kaur, R., Shropshire, J. D., Cross, K. L., Leigh, B., Mansueto, A. J., Stewart, V., Bordenstein, S. R., & Bordenstein, S. R. (2021). Living in the endosymbiotic world of Wolbachia: A centennial review. *Cell Host & Microbe*, *29*(6), 879–893. https://doi.org/10.1016/j.chom.2021.03.006

Klappenbach, J. A. (2001). rrndb: The Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research*, *29*(1), 181–184. https://doi.org/10.1093/nar/29.1.181

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Koga, R., Moriyama, M., Onodera-Tanifuji, N., Ishii, Y., Takai, H., Mizutani, M., Oguchi, K., Okura, R., Suzuki, S., Gotoh, Y., Hayashi, T., Seki, M., Suzuki, Y., Nishide, Y., Hosokawa, T., Wakamoto, Y., Furusawa, C., & Fukatsu, T. (2022). Single mutation makes Escherichia coli an

insect mutualist. *Nature Microbiology*, *7*(8), 1141–1150. https://doi.org/10.1038/s41564-022-01179-9

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. *Environmental Microbiology*, *25*(11), 2431–2446. https://doi.org/10.1111/1462-2920.16473

Kolasa, M., & Łukasik, P. (2023). *With a little help from my friends: The roles of microbial symbionts in insect populations and communities.* https://doi.org/10.32942/X2BS45

Kriesner, P., Hoffmann, A. A., Lee, S. F., Turelli, M., & Weeks, A. R. (2013). Rapid Sequential Spread of Two Wolbachia Variants in Drosophila simulans. *PLoS Pathogens*, *9*(9), e1003607. https://doi.org/10.1371/journal.ppat.1003607

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. *Systematic and Applied Microbiology*, *45*(1), 126289. https://doi.org/10.1016/j.syapm.2021.126289

Lange, C., Boyer, S., Bezemer, T. M., Lefort, M.-C., Dhami, M. K., Biggs, E., Groenteman, R., Fowler, S. V., Paynter, Q., Verdecia Mogena, A. M., & Kaltenpoth, M. (2023). Impact of intraspecific variation in insect microbiomes on host phenotype and evolution. *The ISME Journal*, *17*(11), 1798–1807. https://doi.org/10.1038/s41396-023-01500-2

Leigh, J. W., & Bryant, D. (2015). popart: Full‐feature software for haplotype network construction. *Methods in Ecology and Evolution*, *6*(9), 1110–1116. https://doi.org/10.1111/2041-210X.12410

Leisner, J. J., Laursen, B. G., Prévost, H., Drider, D., & Dalgaard, P. (2007). *Carnobacterium:* positive and negative effects in the environment and in foods. *FEMS Microbiology Reviews*, *31*(5), 592–613. https://doi.org/10.1111/j.1574-6976.2007.00080.x

Leray, M., Ho, S.-L., Lin, I.-J., & Machida, R. J. (2018). MIDORI server: A webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics*, *34*(21), 3753–3754. https://doi.org/10.1093/bioinformatics/bty454

Li, X., & Wiens, J. J. (2023). Estimating Global Biodiversity: The Role of Cryptic Insect Species. *Systematic Biology*, *72*(2), 391–403. https://doi.org/10.1093/sysbio/syac069

Louca, S., Doebeli, M., & Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, *6*(1), 41. https://doi.org/10.1186/s40168-018-0420-9

Łukasik, P., & Kolasa, M. R. (2024). With a little help from my friends: The roles of microbial symbionts in insect populations and communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1904), 20230122. https://doi.org/10.1098/rstb.2023.0122

Łukasik, P., Newton, J. A., Sanders, J. G., Hu, Y., Moreau, C. S., Kronauer, D. J. C., O'Donnell, S., Koga, R., & Russell, J. A. (2017). The structured diversity of specialized gut symbionts of the New World army ants. *Molecular Ecology*, *26*(14), 3808–3825. https://doi.org/10.1111/mec.14140

McFall-Ngai, M., Hadfield, M. G., Bosch, T. C. G., Carey, H. V., Domazet-Lošo, T., Douglas, A. E., Dubilier, N., Eberl, G., Fukami, T., Gilbert, S. F., Hentschel, U., King, N., Kjelleberg, S., Knoll, A. H., Kremer, N., Mazmanian, S. K., Metcalf, J. L., Nealson, K., Pierce, N. E., … Wernegreen, J. J. (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences*, *110*(9), 3229–3236. https://doi.org/10.1073/pnas.1218525110

Medina Munoz, M., Spencer, N., Enomoto, S., Dale, C., & Rio, R. V. M. (2020). Quorum sensing sets the stage for the establishment and vertical transmission of Sodalis praecaptivus in tsetse flies. *PLOS Genetics*, *16*(8), e1008992. https://doi.org/10.1371/journal.pgen.1008992

Meier, R., Lawniczak, M. K. N., & Srivathsan, A. (2024). Illuminating Entomological Dark Matter with DNA Barcodes in an Era of Insect Decline, Deep Learning, and Genomics. *Annual Review of Entomology*. https://doi.org/10.1146/annurev-ento-040124-014001

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. *Environmental Microbiology*, *21*(11), 4253–4269. https://doi.org/10.1111/1462-2920.14786

Moran, N. A., McCutcheon, J. P., & Nakabachi, A. (2008). Genomics and Evolution of Heritable Bacterial Symbionts. *Annual Review of Genetics*, *42*(1), 165–190. https://doi.org/10.1146/annurev.genet.41.110306.130119

Moriyama, M., Nikoh, N., Hosokawa, T., & Fukatsu, T. (2015). Riboflavin Provisioning Underlies *Wolbachia* 's Fitness Contribution to Its Insect Host. *mBio*, *6*(6), e01732-15. https://doi.org/10.1128/mBio.01732-15

Ochman, H., Elwyn, S., & Moran, N. A. (1999). Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences*, *96*(22), 12638–12643. https://doi.org/10.1073/pnas.96.22.12638

Oliver, K. M., Degnan, P. H., Burke, G. R., & Moran, N. A. (2010). Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. *Annual Review of Entomology*, *55*(1), 247–266. https://doi.org/10.1146/annurev-ento-112408-085305

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

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. *Ecology Letters*, *20*(5), 561–576. https://doi.org/10.1111/ele.12757

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403–1414. https://doi.org/10.1111/1462-2920.13023

Parisy, B., Schmidt, N. M., Cirtwill, A. R., Villa-Galaviz, E., Tiusanen, M., Klütsch, C. F. C., Aspholm, P. E., Raundrup, K., Vesterinen, E. J., Wirta, H., & Roslin, T. (2024). Arctic

plant-fungus interaction networks show major rewiring with environmental variation. *Communications Earth & Environment*, *5*(1), 735. https://doi.org/10.1038/s43247-024-01902-w

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*(1), e16392. https://doi.org/10.1371/journal.pone.0016392

Pennington, M. J., Rothman, J. A., Jones, M. B., McFrederick, Q. S., Gan, J., & Trumble, J. T. (2017). Effects of contaminants of emerging concern on Megaselia scalaris (Lowe, Diptera: Phoridae) and its microbial community. *Scientific Reports*, *7*(1), 8165. https://doi.org/10.1038/s41598-017-08683-7

Perlmutter, J. I., & Bordenstein, S. R. (2020). Microorganisms in the reproductive tissues of arthropods. *Nature Reviews Microbiology*, *18*(2), 97–111. https://doi.org/10.1038/s41579-019-0309-z

Perreau, J., & Moran, N. A. (2022). Genetic innovations in animal–microbe symbioses. *Nature Reviews Genetics*, *23*(1), 23–39. https://doi.org/10.1038/s41576-021-00395-z

Perreau, J., Patel, D. J., Anderson, H., Maeda, G. P., Elston, K. M., Barrick, J. E., & Moran, N. A. (2021). Vertical Transmission at the Pathogen-Symbiont Interface: Serratia symbiotica and Aphids. *mBio*, *12*(2), e00359-21. https://doi.org/10.1128/mBio.00359-21

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Ravenscraft, A., Berry, M., Hammer, T., Peay, K., & Boggs, C. (2019). Structure and function of the bacterial and fungal gut microbiota of Neotropical butterflies. *Ecological Monographs*, *89*(2), e01346. https://doi.org/10.1002/ecm.1346

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open

source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946. https://doi.org/10.1101/gr.128124.111

Roslin, T. (2024). Understanding insect movements through space and time is vital for safeguarding global ecosystem services. *One Earth*, *7*(4), 537–540. https://doi.org/10.1016/j.oneear.2024.03.007

Sakamoto, J. M., Shikano, I., & Rasgon, J. L. (2024). Microbiomes of Two Pest Fly Species of Pennsylvania Mushroom Houses. *Insects*, *15*(7), 525. https://doi.org/10.3390/insects15070525

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Sanaei, E., Charlat, S., & Engelstädter, J. (2021). *Wolbachia* host shifts: Routes, mechanisms, constraints and evolutionary consequences. *Biological Reviews*, *96*(2), 433–453. https://doi.org/10.1111/brv.12663

Shigenobu, S., & Yorimoto, S. (2022). Aphid hologenomics: Current status and future challenges. *Current Opinion in Insect Science*, *50*, 100882. https://doi.org/10.1016/j.cois.2022.100882

Shropshire, D., Conner, W. R., Vanderpool, D., Hoffmann, A. A., Turelli, M., & Cooper, B. S. (2023). *Rapid host switching of Wolbachia and even more rapid turnover of their phages and incompatibility-causing loci*. https://doi.org/10.1101/2023.12.04.569981

Smith, M. A., Bertrand, C., Crosby, K., Eveleigh, E. S., Fernandez-Triana, J., Fisher, B. L., Gibbs, J., Hajibabaei, M., Hallwachs, W., Hind, K., Hrcek, J., Huang, D.-W., Janda, M., Janzen, D. H., Li, Y., Miller, S. E., Packer, L., Quicke, D., Ratnasingham, S., … Zhou, X. (2012). Wolbachia and DNA Barcoding Insects: Patterns, Potential, and Problems. *PLoS ONE*, *7*(5), e36514. https://doi.org/10.1371/journal.pone.0036514

Soberón, J. (2019). A Grinnellian Niche Perspective on Species-Area Relationships. *The American Naturalist*, *194*(6), 760–775. https://doi.org/10.1086/705898

Song, S. J., Sanders, J. G., Delsuc, F., Metcalf, J., Amato, K., Taylor, M. W., Mazel, F., Lutz, H. L., Winker, K., Graves, G. R., Humphrey, G., Gilbert, J. A., Hackett, S. J., White, K. P., Skeen, H. R., Kurtis, S. M., Withrow, J., Braile, T., Miller, M., … Knight, R. (2020). Comparative Analyses of Vertebrate Gut Microbiomes Reveal Convergence between Birds and Bats. *mBio*, *11*(1), e02901-19. https://doi.org/10.1128/mBio.02901-19

Srivathsan, A., Ang, Y., Heraty, J. M., Hwang, W. S., Jusoh, W. F. A., Kutty, S. N., Puniamoorthy, J., Yeo, D., Roslin, T., & Meier, R. (2023). Convergence of dominance and neglect in flying insect diversity. *Nature Ecology & Evolution*, *7*(7), 1012–1021. https://doi.org/10.1038/s41559-023-02066-0

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W. X., Bertrand, D., Ng, A. H. Q., Boey, E. J. H., Koh, J. J. Y., Nagarajan, N., & Meier, R. (2018). A Min ION ™-based pipeline for fast and cost-effective DNA barcoding. *Molecular Ecology Resources*, *18*(5), 1035–1049. https://doi.org/10.1111/1755-0998.12890

Srivathsan, A., Hartop, E., Puniamoorthy, J., Lee, W. T., Kutty, S. N., Kurina, O., & Meier, R. (2019). Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. *BMC Biology*, *17*(1), 96. https://doi.org/10.1186/s12915-019-0706-9

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S. N., Wong, J., Yeo, D., & Meier, R. (2021). ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. *BMC Biology*, *19*(1), 217. https://doi.org/10.1186/s12915-021-01141-x

Stork, N. E. (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, *63*(1), 31–45. https://doi.org/10.1146/annurev-ento-020117-043348

Sudakaran, S., Kost, C., & Kaltenpoth, M. (2017). Symbiont Acquisition and Replacement as a Source of Ecological Innovation. *Trends in Microbiology*, *25*(5), 375–390.

https://doi.org/10.1016/j.tim.2017.02.014

Teixeira, L., Ferreira, Á., & Ashburner, M. (2008). The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. *PLoS Biology*, *6*(12), e1000002. https://doi.org/10.1371/journal.pbio.1000002

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., De Jonge, M. M. J., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the r‐package H msc. *Methods in Ecology and Evolution*, *11*(3), 442–447. https://doi.org/10.1111/2041-210X.13345

Tourlousse, D. M., Yoshiike, S., Ohashi, A., Matsukura, S., Noda, N., & Sekiguchi, Y. (2017). Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Research*, gkw984. https://doi.org/10.1093/nar/gkw984

Turelli, M., Cooper, B. S., Richardson, K. M., Ginsberg, P. S., Peckenpaugh, B., Antelope, C. X., Kim, K. J., May, M. R., Abrieux, A., Wilson, D. A., Bronski, M. J., Moore, B. R., Gao, J.-J., Eisen, M. B., Chiu, J. C., Conner, W. R., & Hoffmann, A. A. (2018). Rapid Global Spread of wRi-like Wolbachia across Multiple Drosophila. *Current Biology*, *28*(6), 963-971.e8. https://doi.org/10.1016/j.cub.2018.02.015

Valdivia, C., Newton, J. A., Von Beeren, C., O'Donnell, S., Kronauer, D. J. C., Russell, J. A., & Łukasik, P. (2023). Microbial symbionts are shared between ants and their associated beetles. *Environmental Microbiology*, *25*(12), 3466–3483. https://doi.org/10.1111/1462-2920.16544

Van Klink, R., Sheard, J. K., Høye, T. T., Roslin, T., Do Nascimento, L. A., & Bauer, S. (2024). Towards a toolkit for global insect biodiversity monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1904), 20230101. https://doi.org/10.1098/rstb.2023.0101

Vancaester, E., & Blaxter, M. (2023). Phylogenomic analysis of Wolbachia genomes from the Darwin Tree of Life biodiversity genomics project. *PLOS Biology*, *21*(1), e3001972. https://doi.org/10.1371/journal.pbio.3001972

Vosseberg, J., Van Hooff, J. J. E., Köstlbacher, S., Panagiotou, K., Tamarit, D., & Ettema, T. J. G.

(2024). The emerging view on the origin and early evolution of eukaryotic cells. *Nature*, *633*(8029), 295–305. https://doi.org/10.1038/s41586-024-07677-6

Wang, W. Y., Srivathsan, A., Foo, M., Yamane, S. K., & Meier, R. (2018). Sorting specimen‐rich invertebrate samples with cost‐effective NGS barcodes: Validating a reverse workflow for specimen processing. *Molecular Ecology Resources*, *18*(3), 490–501. https://doi.org/10.1111/1755-0998.12751

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Weinert, L. A., Araujo-Jnr, E. V., Ahmed, M. Z., & Welch, J. J. (2015). The incidence of bacterial endosymbionts in terrestrial arthropods. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1807), 20150249. https://doi.org/10.1098/rspb.2015.0249

Williams, D. J., Grimont, P. A. D., Cazares, A., Grimont, F., Ageron, E., Pettigrew, K. A., Cazares, D., Njamkepo, E., Weill, F.-X., Heinz, E., Holden, M. T. G., Thomson, N. R., & Coulthurst, S. J. (2022). The genus Serratia revisited by genomics. *Nature Communications*, *13*(1), 5195. https://doi.org/10.1038/s41467-022-32929-2

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

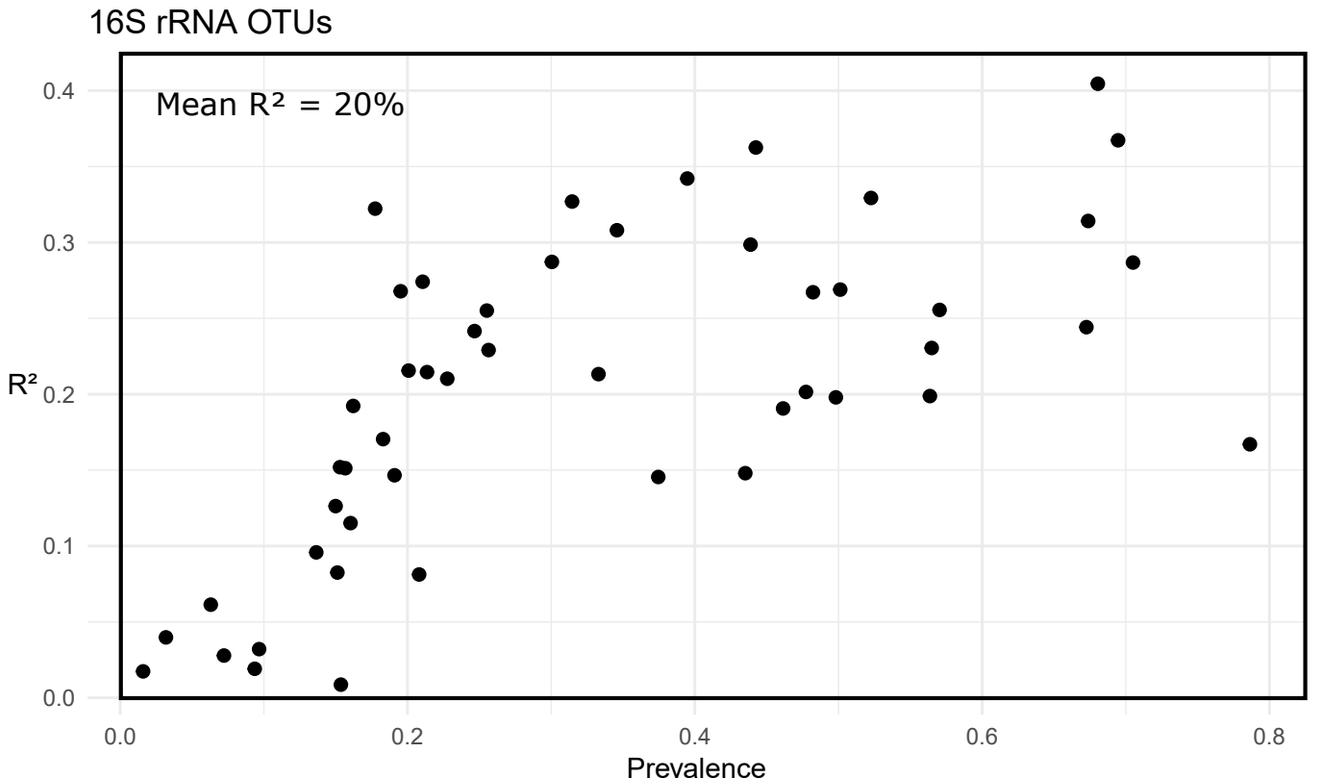
Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

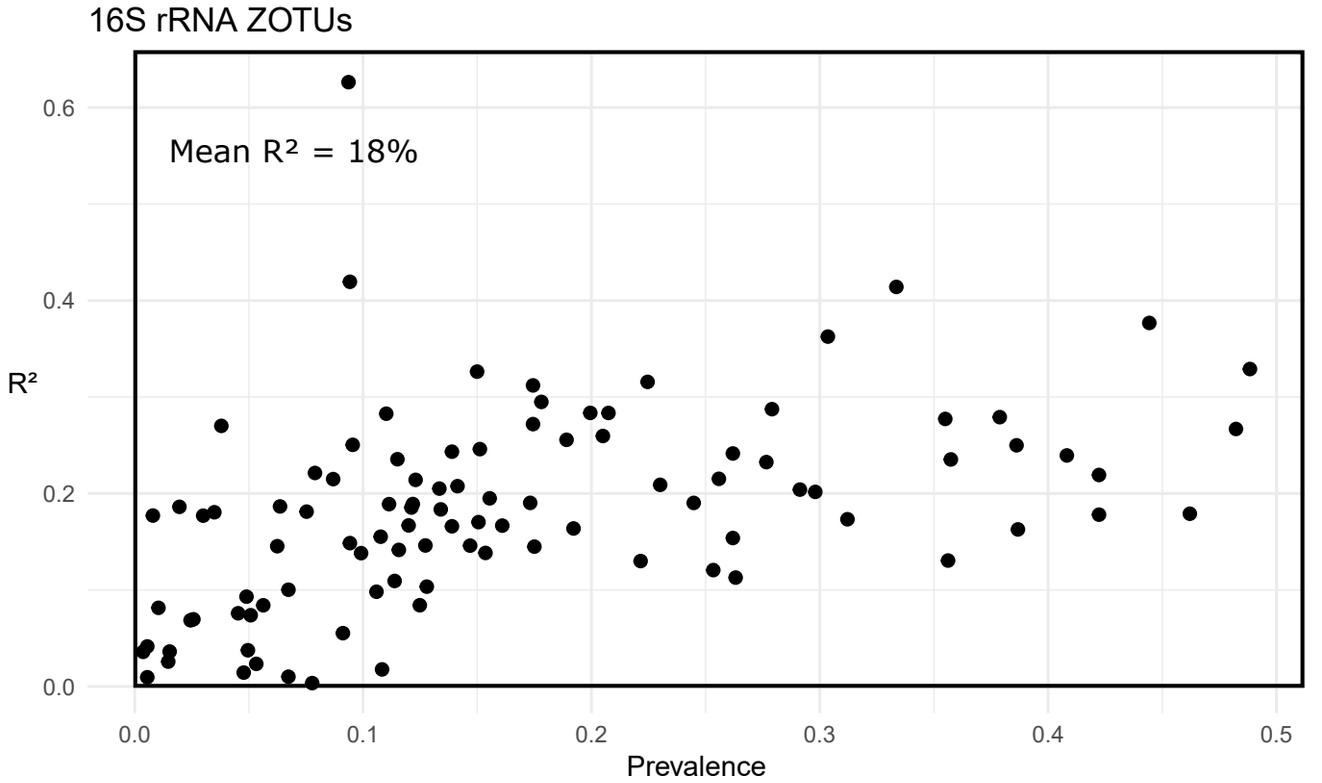
Zhao, D., Zhang, Z., Niu, H., & Guo, H. (2023). Pathogens are an important driving force for the rapid spread of symbionts in an insect host. *Nature Ecology & Evolution*, *7*(10), 1667–1681. https://doi.org/10.1038/s41559-023-02160-3

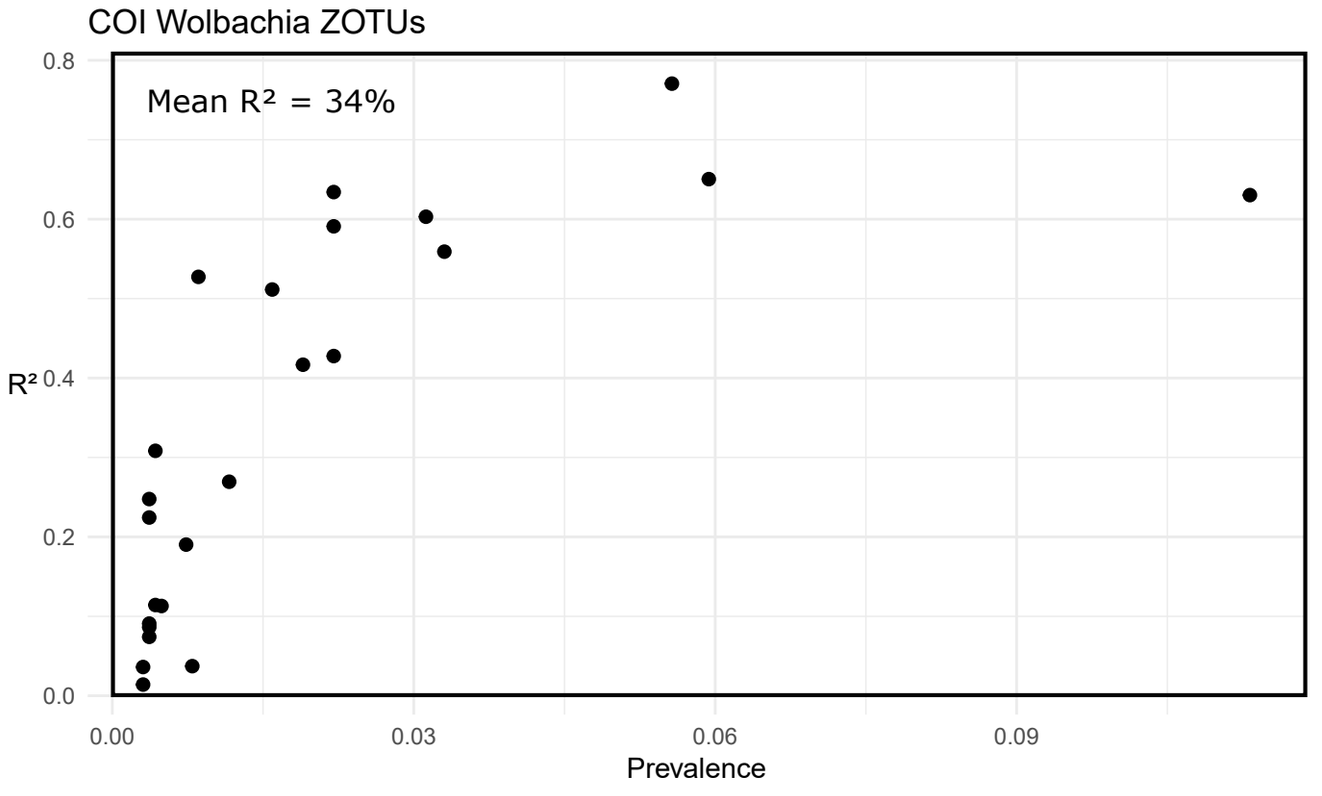
Zheng, H., Steele, M. I., Leonard, S. P., Motta, E. V. S., & Moran, N. A. (2018). Honey bees as models

for gut microbiota research. *Lab Animal*, *47*(11), 317–325. https://doi.org/10.1038/s41684-018-0173-x

Zug, R., & Hammerstein, P. (2012). Still a Host of Hosts for Wolbachia: Analysis of Recent Data Suggests That 40% of Terrestrial Arthropod Species Are Infected. *PLoS ONE*, *7*(6), e38544. https://doi.org/10.1371/journal.pone.0038544


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