THE UNIVERSITY OF SOUTH BOHEMIA IN ČESKÉ BUDĚJOVICE
FACULTY OF SCIENCE

Bioinformatic identification and characterisation of *Wolbachia* prevalence in a tropical skipper butterfly taxon (Hesperiidae: Eudaminae) using whole-genome sequencing data

Bachelor Thesis

Andrea Kochova

Supervisor: Pedro de Gusmão Ribeiro, PhD.

Consultant: RNDr. Pável Matos-Maraví, PhD.

České Budějovice 2026

BIBILIOGRAPHICAL DETAIL

Kochova A., 2026: Bioinformatic identification and characterisation of *Wolbachia* prevalence in a tropical skipper butterfly taxon (Hesperiidae: Eudaminae) using whole-genome sequencing data. Bc. Thesis, in English. – 44 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

ANNOTATION

The bachelor's thesis aims to investigate *Wolbachia* infections in neotropical skipper butterflies (Hesperiidae: Eudaminae) with the use of whole-genome sequencing (WGS) data. The research focuses on detecting infection, evaluating supergroup signal, and describing infection character based on similarity.

DECLARATION

I declare that I am the author of this qualification thesis and that in writing it I have used the sources and literature displayed in the list of used sources only. I further declare that I have used the Grammarly, Writefull, and ChatGPT generative artificial intelligence tool(s) or service(s) in accordance with academic ethics for the purpose of refining text structure, correcting grammatical and lexical mistakes.

České Budějovice, 06.08.2026.

Student's signature

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Pedro de Gusmão Ribeiro, PhD, for his guidance, constructive feedback, and support throughout the development of this thesis. I am also grateful to my consultant, RNDr. Pável Matos-Maraví, PhD, for his scientific advice, valuable insights, and contribution to this work.

Finally, I would like to thank my family - my parents Hirsto and Elitsa Kochovi, and sister Elena, and my friends in Linz, Corinna Strasser and Vadym Hurkin, for their continuous encouragement, patience, and support throughout my studies and during the preparation of this thesis.

Abstract

Wolbachia is one of the most widespread intracellular endosymbionts infecting a large proportion of arthropods. However, its infection dynamics remain incompletely understood, even in well-studied groups such as Lepidoptera, and particularly in studies spanning many species across multiple genera. Within Lepidoptera, the Eudaminae subfamily offers a compelling system for studying *Wolbachia*–insect relationship, in a broader multi-genera context, since the increasing availability of genomic resources for the group makes its species well suited to detecting *Wolbachia* from host whole-genome data. Here, we aimed to detect and quantify *Wolbachia* infection across Eudaminae using whole-genome sequencing data from 77 specimens spanning 18 genera, analysed by two approaches: screening assembled contigs for similarity to *Wolbachia* proteins, and mapping reads against *Wolbachia* genomes. The assembly-based screen flagged 26 samples, while the read-mapping classification identified 20 as infected, with a prevalence of 25.97% infected samples. Supergroup B predominated over supergroup A, comprising 13 B-only infections, three A-only, and four A+B co-infections, and detectable infection occurred in seven of the eighteen genera. First, our results compare whole-genome *Wolbachia* detection methods showing that while the read-mapping approach is more stringent and appropriate in cases where there is high data heterogeneity, it is also more computationally demanding and requires more steps. Thus, both methods are complementary, and their use should be considered in face of study needs. Additionally, these results provide the first multi-genera assessment of *Wolbachia* in Eudaminae, a clade where the symbiont had not previously been surveyed at this scale.

1. Introduction

1.1. *Wolbachia* – a widespread endosymbiont, its diversity, and evolution

Wolbachia is an intracellular, maternally inherited bacterium that belongs to the order Rickettsiales. *Wolbachia* was first described in mosquitoes, and since then, it has become known as one of the most widespread endosymbionts, infecting approximately 40-50% of all terrestrial arthropod species and 20-30% of insect species (Weinert et al., 2015; Duplouy & Hornett, 2018; Vancaester & Blaxter, 2023).

Its large prevalence is partly attributed to its ability to manipulate host reproduction in various ways such that it enhances its own transmission through host populations. Several reproductive manipulation strategies have been documented: cytoplasmic incompatibility (Kageyama & Traut, 2004; Werren et al., 2008; Shropshire & Bordenstein, 2016), male killing (Hurst & Majerus, 1993; Dyson et al., 2002), feminisation of genetic males (Hiroki et al., 2002; Hiroki et a., 2004; Narita et al., 2011), and parthenogenetic induction (Stouthamer & Werren, 1993). Cytoplasmic incompatibility (CI) is the most widespread reproductive manipulation strategy, resulting in reduced rates of hatched eggs when infected males mate with uninfected females or females infected with other *Wolbachia* strains. Moreover, this phenomenon is reversed when an infected male mates with a similarly infected female (*i.e.*, one having the same CI strain). The genetic mechanism underlying this phenomenon involves CI factor genes called *cifA/cifB* and *cidA/cidB*, which were described respectively in the *w*Mel and *w*Pip *Wolbachia* strains (Beckmann et al., 2017; Le Page et al., 2017) . It is the simultaneous expression of *cifA* and *cifB* that induces CI by modifying male sperm, while the presence of *cifA* in the egg of a female, infected with the same strain as the male counterpart, is responsible for rescuing embryonic viability (Shropshire et al., 2018), and, in CI-infected organisms, the sex ratio of the host populations is not altered (Engelstädter & Hurst, 2009; Duplouy & Hornett, 2018).

On the other hand, three other types of reproductive manipulation can alter the sex ratio of populations: male killing (MK), feminisation, and parthenogenesis induction (PI). In male killing, which is well known in Lepidoptera, male offspring are killed early in development, producing a female-biased offspring. Male-killing *Wolbachia* have been documented in butterflies, most notably in *Acraea encedon*, where more than 80% of females in Ugandan populations are infected (Jiggins et al., 2001), and in *Hypolimnas bolina*, where Samoan populations presented a sex ratio of 100 females to every male (Dyson et al., 2002; Dyson & Hurst, 2004). At the molecular level, male killing in Lepidoptera has recently been linked to the

Wolbachia gene *oscar*, which interferes with the male sex-determination pathway of the host (Katsuma et al., 2022). In feminisation, genetic males develop as phenotypic females, also biasing the sex ratio towards the transmitting sex. In butterflies, this is best characterised in *Eurema mandarina*, which carries a feminising *Wolbachia* strain, *wFem* (Hiroki et al., 2002; Narita et al., 2011). Moreover, treating infected larvae with antibiotics produces adults with intermediate sexual traits, and the effect depends on the timing of treatment, indicating that the symbiont interacts with its host continuously through development rather than at a single point (Narita et al., 2007). Lastly, parthenogenesis induction can also produce female-biased populations, but confirmed cases are restricted to haplodiploid hosts (*i.e.,* those where unfertilised eggs become males and fertilised eggs females), such as parasitoid wasps, with no confirmed examples in diplodiploid groups like butterflies (Ma & Schwander, 2017).

Wolbachia can also influence host evolution through lateral gene transfer (LGT), by which fragments of the bacterial genome become integrated into the host genome. Such transfers can range from short segments to nearly complete bacterial genomes, and, in some cases, they contribute to chromosomal rearrangements or novel gene expression with beneficial effects on host fitness (Nikoh et al., 2008; Werren et al., 2008; Dunning Hotopp et al., 2011; Li et al., 2022). Lepidoptera in particular show among the highest rates of *Wolbachia*-derived LGT of any insect group (Li et al., 2022), so that bacterial sequence residing in the host genome can be common feature of their genomic data. Importantly, laterally transferred genes from *Wolbachia* can undergo gene erosion or higher substitution rates over time, thus accumulating more mutations, since their function tends to be lost (Nikoh et al., 2008; Tvedte et al., 2022).

Wolbachia diversity is typically classified into major phylogenetic lineages, referred to as supergroups. So far, 17 supergroups have been described (A-F, H-Q, and S) (Kaur et al., 2021). In insects, the dominating supergroups are A and B, while the other supergroups are mostly associated with various arthropod or nematode groups (Duplouy & Hornett, 2018; Kaur et al., 2021; Mahmood et al., 2023; Vancaester & Blaxter, 2023), and while most insect clades are infected by supergroup A strain, butterflies are mostly infected by supergroup B strains (Vancaester & Blaxter, 2023). Furthermore, within the supergroups, *Wolbachia* infections can also be distinguished at the strain level, with a strain representing a specific *Wolbachia* variant that may differ from others in the same supergroup, in its genome, and in the phenotypes it induces in the host. Standardised strain typing has commonly relied on multilocus sequence typing (MLST), which uses sequence variation across conserved genes to distinguish strains and compare their relationships across host species (Baldo et al., 2006).

1.2. *Wolbachia* vertical and horizontal transfer

It has been suggested that *Wolbachia* infection in closely related host species is the result of vertical transmission from a shared ancestor, a process commonly referred to as cladogenetic transmission or cospeciation (Bandi et al., 1998; Raychoudhury et al., 2009). This view is a macroevolutionary extension of the actual mechanism by which *Wolbachia* is transmitted from one generation through the next, vertically, via maternal inheritance. Thus, under strict vertical transmission, over broad evolutionary timescales, host and symbiont phylogenies would be expected to show substantial congruence, reflecting the codivergence between the two lineages (Werren et al., 2008). Such signatures of codivergence have been documented in filarial nematodes (Bandi et al., 1998, Lefoulon et al., 2016), in insects in *Nasonia* wasps (Raychoudhury et al., 2009) and in *Nomada* bees (Gerth & Bleidorn, 2016).

Alternatively, *Wolbachia* can spread across species through horizontal transfer. This process occurs when unrelated hosts acquire the bacterium via sharing the same food sources (*e.g.,* hostplants) (Duron et al., 2010; Chrostek & Gerth, 2019), predation and cannibalism (Le Clec'h et al., 2013), introgression (Turelli et al., 2018; Shastry et al., 2022) or parasitism (Heath et al., 1999; Huigens et al., 2004). In this scenario, the phylogenies of *Wolbachia* and their hosts are uncorrelated, resulting in the loss of signals of codivergence across evolutionary time (Werren et al., 2008). Interestingly, *Wolbachia* often transfer preferentially among closely related species due to host-specialisation (Jiggins et al., 2001; Werren et al., 2008; Russell et al., 2009) as related hosts are more likely to share similar ecological niches and physical environments. Therefore, closely related species tend to harbour closely related strains, and the resulting pattern can resemble codivergence, particularly when taxonomic sampling is limited. In such cases, horizontal transfer can still be the underlying process of transmission, and, in fact, this process seems to be the dominant way by which *Wolbachia* is able to transfer across species in arthropods. For example, screening genomic data from a total of 368 insect species, Vancaester and Blaxter (2023) assembled 110 complete *Wolbachia* genomes from 93 infected species and found little to no congruence between host and symbiont phylogenies, concluding that horizontal transmission has been the dominant pattern shaping *Wolbachia* distribution across insects. In Lepidoptera, a recent global survey also showed that horizontal transfer is the main mechanism by which the bacterium can spread within the clade (Yang et al., 2026), a pattern that has been shown on both macroevolutionary scale (Ahmed et al., 2016), and in microevolutionary scale among closely related species (Ribeiro et al., 2025). Moreover, in filiarial nematodes, a group in which cospeciation long persisted as the main paradigm of

Wolbachia transmission (Bandi et al., 1998), frequent losses, symbiont replacement and double-infections have shifted the field towards a more complex scenario including horizontal transfer across species as a relevant process (Lefoulon et al., 2016; Lefoulon et al., 2020; Vancaester et al., 2025).

1.3. Whole-genome sequencing and the detection of *Wolbachia*

Most commonly, *Wolbachia* infections are identified through PCR amplification of conserved bacterial loci, including 16S rRNA, *ftsZ*, *wsp*, and the multilocus sequence typing (MLST) gene set (Baldo et al., 2006; Simões et al., 2011). However, the growing availability of whole-genome sequencing (WGS) data has shifted detection towards bioinformatic approaches, by which *Wolbachia* are detected directly from the whole genome sequencing data of the host. One approach is to assemble host contigs from this data and to "mine" *Wolbachia* contigs that are assembled as a byproduct, for instance, by using BLAST (Altschul et al., 1990) to detect *Wolbachia* proteins or nucleotide sequences. Such a process requires conservative filters so that alignments do not represent just potential insertions of the bacterium into the host genome, but true *Wolbachia* contigs. For instance, Ribeiro and colleagues (2025) used strict filters (*e.g.,* e-value < $1\times e^{-10}$, bitscore >300, and percent-identity >95%), while percent identity had to be lowered to >90% to accommodate potential insertions that can erode with time (Nikoh et al., 2008; Tvedte et al., 2022). Additionally, BUSCO (Simão et al., 2015) can be used to assess the completeness of these contigs according to a curated database of bacteria orthologues (*e.g.,* bacteria_odb10), adding another layer of confirmation that the assembled contigs aligned with BLAST are *Wolbachia* (Ribeiro et al., 2025). Alternatively, infections can be screened using taxonomic classifiers such as Kraken 2 and MetaPhlAn 4, which assign sequencing reads to reference taxa (Wood et al., 2019; Lu et al., 2022; Blanco-Míguez et al., 2023). Potential infections can then be confirmed by mapping reads against a *Wolbachia* reference genome using programs such as Bowtie2 v.2.4.2.(Langmead & Salzberg, 2012) and then assessing mapping statistics with SAMtools (Danecek et al., 2021). With this, infection assignments can be evaluated quantitatively through breadth of coverage (*i.e.,* the proportion of the reference covered), depth of mapped reads, or the number of *Wolbachia*-mapped reads normalised by library size to allow comparison across samples (reads per million, RPM).

The use of host genome data to detect *Wolbachia* has been extended to the mining of public sequence repositories, where *Wolbachia* genomes are reconstructed from data originally generated for other purposes. Scholz and colleagues (2020) screened 31,825 publicly available sequencing samples from more than 500 likely host species. They identified 1,793 as

Wolbachia-positive and assembled 1,166 metagenome-assembled genomes (MAGs), raising the number of host species with an assembled *Wolbachia* genome from 33 to 55. However, such short-read MAGs are frequently incomplete or fragmented and tend to collapse repetitive regions, which are largely present in *Wolbachia* genomes (Wu et al., 2004; Scholz et al., 2020). Because of this, they used a conservative quality control in their dataset, retaining 1,005 of the 1,166 genomes after filtering on coverage, core-gene content, genome length, and evidence of chimeric assembly arising from coinfection (Scholz et al., 2020).

One way of overcoming some of the constraints present when using short-read MAG assemblies is using long-read sequencing. For instance, long reads can usually span repetitive regions entirely, resulting in less fragmented assemblies that generate more complete and circular genomes in comparison to short reads (Vancaester and Blaxter, 2023). This sequencing strategy has been largely adopted by large sequencing initiatives such as the Earth BioGenome Project (https://www.earthbiogenome.org/) and the Darwin Tree of Life (https://www.darwintreeoflife.org), and it is becoming a standard for comparative genomics projects and to be used as reference-level genomes, which, for *Wolbachia* screening, is the best option when read-mapping approaches are used. For example, one large-scale application of long-reads assembled genomes was the screening of 368 insect species sequenced by the Darwin Tree of Life project (https://www.darwintreeoflife.org/data/). Vancaester and Blaxter (2023) recovered 110 *Wolbachia* genomes from 93 host species, 77 of them assembled as complete circular chromosomes and including the first complete genomes from Odonata and Orthoptera with high accuracy, allowing for identification of co-infecting strains and *Wolbachia* regions inserted into the genomes of the hosts (Vancaester & Blaxter, 2023).

Genome data additionally enable analyses that marker genes such as the MLST typing genes cannot support, including phylogenomic inference from hundreds of single-copy orthologues, assignment of strains to supergroups, characterisation of prophage WO content, and direct screening for host-manipulation loci such as the *cif* gene pairs. This context and the accumulation of publicly available assemblies can further move the study of *Wolbachia* from detection and pattern assessment towards a more mechanistic view of how the bacterium can infect natural populations, from the molecular perspective to large-scale community and macroevolutionary level investigation.

1.4. *Wolbachia* in Lepidoptera

Lepidoptera are an animal clade including butterflies and moths and is posed as one of the most important insect models in population genomics, ecology, chemical ecology, and biogeography. They marked the formal beginning of the study of coevolution in a paper that described how the coevolutionary relationship between, in this case, butterflies and plants, was central to macroevolutionary species diversification (Ehrlich and Raven, 1964). Nowadays, butterflies and moths have become one of the most studied insects in symbiont-insect relationships, and, especially, with *Wolbachia* (Duplouy & Hornett, 2018). In Lepidoptera, most major host reproductive manipulation strategies caused by *Wolbachia* have been reported, including male killing, feminisation, and cytoplasmic incompatibility, but not parthenogenesis (Duplouy & Hornett, 2018). Additionally, while Lepidoptera are mostly infected by supergroup B, contrary to other insects, double supergroup A+B infections in a host individual have also been noted (Twort et al., 2022; Vancaester & Blaxter, 2023; Ribeiro et al., 2025).

Lepidoptera possess several genomic features that distinguish them among insects. Their chromosomes are holocentric (*i.e.,* without centralised centromeres) and highly conserved in numbers, most species having haploid chromosome numbers of n = 29-31, but with wide variation ranging from 5 to 223 in known species (Wright et al., 2024). Genome sizes are modest, ranging from ~200 Mb to ~1,897 Mb, a variation mostly attributed to presence of repetitive DNA and transposable element content (Liu et al., 2020). Lepidoptera are also female-heterogametic, with females carrying a Z and a W chromosome and males two Z chromosomes.

Lepidoptera also represent a suitable system for genomic investigation. Five years ago, they were already one of the insect orders with the most available genomic resources, including reference-level genomes (250 at the time) and more scattered assemblies (870 at the time). Now, with large-scale transnational efforts focusing on the generation of Lepidoptera genomic data (*e.g.,* Project Psyche https://www.projectpsyche.org/), they have become the eukaryotic order with the most sequenced reference-level genomes (Wright et al., 2025). In NCBI, a total of 4,800 Lepidoptera genomes are available, 2,429 of which are reference-level, which represents ~46% of all Insecta genomes in NCBI and ~44% of all Insecta reference genomes (https://www.ncbi.nlm.nih.gov/datasets/genome/; search: Lepidoptera, Insecta; filters: Reference Genome). Consequently, considering many of these genomes come from natural population samples, these numbers make Lepidoptera the largest natural resource to investigate the genomics of the relationship between insect and *Wolbachia* bacteria.

Studies have shown that *Wolbachia* infection dynamics in Lepidoptera are diverse, often characterised by high turnover, strain specificity, and pervasive horizontal transfer (Ahmed et al., 2016; Twort et al., 2022; Ribeiro et al., 2025; Yang et al., 2026). Most of what is known, however, derives primarily from either single-species studies or broad screens with limited taxonomic depth, so the scale at which infections are gained, lost, and exchanged remains poorly resolved. The genomic resources available for the order make it possible to address these questions across multiple phylogenetic scales, from strain diversity within populations to patterns of transmission among closely related species.

1.5. Hesperiidae and the Eudaminae subfamily

The Hesperiidae, also known as the skipper butterflies, are a diverse family of Lepidoptera. Although understudied, it has been shown that they present a genome size variation ranging 293 Mb – 645 Mb (Liu et al., 2020), with haploid chromosome numbers ranging from n = 16 to n = 50, most species presenting n = 31 across 46 analysed species (source GoaT database https://goat.genomehubs.org/; serach: Hesperiidae, chromosome number).

Within Hesperiidae, Eudaminae is a diverse subfamily comprising approximately 600 recognised species, with most species harboured throughout the Neotropical region, which ranges from central Mexico to the southern part of South America, but with high species diversity in other tropical areas such as Asia and Africa (Warren et al., 2009; Li et al., 2019; Zhang et al., 2023). It is one of the subfamilies with the highest diversification rates among Lepidoptera (Kawahara et al., 2023), which poses interesting evolutionary questions such as understanding the mechanisms behind their convergent wing colour patterns. Indeed, the convergent phenotypes across its phylogeny have posed difficulties for their taxonomy assessment (Warren et al., 2008; Warren et al., 2009), which have been further disentangled recently in phylogenomic studies (Li et al., 2019). Many of Eudaminae's populations occur in either the same or overlapping geographical areas and have similar diel activities, which means that their occurrence is sympatric and often syntopic (Devries et al., 2008; Zhang et al., 2023), increasing the likelihood of interactions among closely related taxa and the opportunities for horizontal transfer of endosymbionts like *Wolbachia*. Additionally, recent genomic analyses of closely related Eudaminae species show that frequent *Wolbachia* infection is maintained through a high turnover dynamic that should be the result of pervasive horizontal transmission (Ribeiro et al., 2025). This demonstrates that Eudaminae is a valuable system for investigating patterns of infection turnover and modes of acquisition.

1.6. Overview of the thesis

In this thesis, I have screened for *Wolbachia* in whole-genome sequencing data from 77 skipper butterfly specimens, representing 18 genera collected across five Neotropical countries: Peru, Cuba, Panama, Brazil and Jamaica. The data were generated for the sequencing of the butterfly genomes and are therefore dominated by host-derived reads, so *Wolbachia* was detected bioinformatically from the same datasets.

Detection was based on complementary approaches: (1) one derived from the assembly of butterfly contigs, the "mining" of *Wolbachia* contigs (present in the assembly as a byproduct of infection), identified with a combination of protein detection with BLASTX (Altschul et al., 1990) and BUSCO (Simão et al., 2015); (2) the second, using a read-mapping approach that mainly consists of mapping total sequenced reads (*i.e*., reads, containing both host and *Wolbachia)* to a *Wolbachia* reference genome, filtering *Wolbachia* reads and re-mapping them to obtain mapping statistics such as mapping coverage (*i.e.,* distribution of reads across the reference) and mean depth. Most of my interpretation of results, however, is based on the read-mapping approach since in the assembly-based screen, a single matching contig is enough to flag a sample, and hit counts are too variable to set a meaningful cutoff, so it can only distinguish presence from absence. The read-mapping approach, by contrast, measures breadth of coverage across complete chromosome-level references, which supports a defensible threshold and reduces false positives. Lastly, strain exploration is read-mapping based, used as a primary description in Eudaminae, and depends on the reference set used. Thus, they are not intended as a formal strain characterisation.

The aims are to: (i) curate raw short-read data and assemble *de novo* contigs; (ii) classify contigs of *Wolbachia* origin by sequence similarity; and (iii) confirm infections through orthologue completeness and read mapping against published references using BUSCO.

2. Materials and methods

2.1. Sample collection, DNA extraction and Illumina sequencing

Members of the research group collected most specimens with the assistance of external collaborators across the Neotropical region. Samples were collected in five countries: Peru (n = 35), Cuba (n = 22), Panama (n = 16), Brazil (n = 3) and Jamaica (n = 1), resulting in taxonomically varied dataset, comprising representatives of 18 distinct genera. A small subset of specimens could not be taxonomically assigned with strong confidence at the time of

sampling, but these were retained for downstream analyses and are further discussed in Section 2.2. All species and their corresponding sample code are in Table S1.

Genomic DNA extraction was also performed by other team members in the laboratory prior to the present study. To extract the DNA, one to three legs were used per sample, following the manufacturer's protocols of QIAGEN DNEasy Blood and Tissue DNA isolation kit. Subsequently, the extracted DNA was sent to Novogene, UK, for library preparation and Illumina sequencing, which was performed on a llumina NovaSeq 6000 system on an S4 v1.5, PE 2x150 flowcell, aiming for approximately 5Gb of sequencing data per butterfly sample on average.

2.2. Quality Control (QC), *de novo* genome assembly, and taxonomic investigation

All following scripts and the raw reports that are too large for the thesis text are provided in a GitHub repository, hereafter referred to as Supplementary Material M1.

We performed an initial quality assessment of the raw Illumina reads using FastQC v.0.11.9 (Andrews, 2010) and aggregated its results using MultiQC v.1.20 (Ewels et al., 2016). Based on these QC reports, we identified sequencing artefacts, including commonly seen adapter contamination, duplication and poly-G tails. We used FastP v.0.22.0 (Chen et al., 2018) for read trimming and filtering, the default parameters being sufficient in most cases. Still, persistent adapter content contamination required custom adapter removal via the *--adapter_fasta* option, in combination with an adapter FASTA file. To reduce the PCR duplicates and overrepresented sequences, we performed deduplication using the *clumpify.sh* utility in the BBMap suite v.35.85 (Bushnell, 2014). The duplicate removal was carried out with the optical duplicate detection and spatial adjacency enabled. Finally, to confirm the cleaning was successful, we reassessed the reads with FastQC and MultiQC (Supplementary Material M1).

We assembled the clean reads into contiguous sequences (contigs) with SPAdes v.3.15.4 (Bankevich et al., 2012), without a reference genome (*i.e., de novo* assembly), optimised through multiple *k*-mer sizes (21, 33 and 77). The resulting contigs were then used as input for other downstream analyses.

To validate species identity for specimens of uncertain taxonomy, mitochondrial genome assembly was performed and barcode sequences obtained (*i.e.,* the COI or *cox1*). Firstly, we retrieved reference mitochondrial genomes in GenBank format from NCBI (NCBI Accession NC_024649 – *Lobocla bifasciatus;* and NCBI Accession NC_030602 – *Achalarus lyciades*) and used them for reference mitochondrial assemblies with MitoFinder v.1.4.2 (Allio

et al., 2020). The barcode sequence for these specimens was then retrieved manually and we inferred the species identity with the Barcode of Life Data Systems (BOLD systems) using the default 'Animal Library – Public' databse (Ratnasingham et al., 2024). BOLD reports that allowed us to identify these specimens and the output from MitoFinder v.1.4.2 from which barcodes were manually extracted are deposited in the Supplementary Material (Supplementary Material M1).

2.3. *Wolbachia* identification: assembly-based and read-mapping approaches

As the whole-genome sequencing data from the butterfly specimens contained mainly host-derived reads, the detection of *Wolbachia* required a workflow able to identify bacterial signal within the host genomic dataset and subsequently evaluate whether the signal was consistent. Therefore, the analysis combined two main complementary approaches: (1) one meant to provide a first screening of potentially infected samples, based on the presence of *Wolbachia* contigs generated as a byproduct during the *de novo* assembly of total read data from the host samples (as described in section 2.2); (2) the second one meant to assess the same samples based on mapping total reads against *Wolbachia* reference genomes, filtering *Wolbachia*-only reads, remapping the filtered *Wolbachia* reads against the same reference genomes and finally obtaining mapping statistics. The associated and specific tools accompanying these approaches are described below.

In the first approach, to identify potential *Wolbachia* contigs and initially screen infected samples, the assembled *de novo* contigs were searched with BLASTX v.2.10.0 (Altschul et al., 1990) against a *Wolbachia* reference protein database that contained representatives of supergroup A (NCBI Accession: GCF_000008025.1, *w*Mel from *Drosophila melanogaster*) and supergroup B (NCBI Accession: GCF_000073005.1, *w*Pip from *Culex quinquefasciatus*). Two references were sufficient at this stage, as this approach aimed to detect the presence of *Wolbachia*-associated sequence rather than to assign it to a supergroup. Since *Wolbachia* proteins are conserved across supergroups, protein-level hits reliably establish presence but cannot distinguish supergroups, and, additionally, the translated search was specifically used, as protein-level similarity can detect bacterial coding sequences even when nucleotide-level similarity is reduced. To filter contigs that could reliably be associated with *Wolbachia*, we filtered the BLASTX tables based on three strict criteria: e-value ($1\times e^{-10}$), bitscore (>300), and percent identity (>95%). Then, to evaluate whether conserved bacterial orthologues were recovered from these putative *Wolbachia* contigs, we assessed them with BUSCO v.5.2.2 (Simão et al., 2015), using the built-in *bacteria_odb10* reference set. In our study, BUSCO

assessed whether these contigs contained conserved bacterial genes and how completely a bacterial contig was recovered, complementing the *Wolbachia*-level identification obtained with BLASTX. BUSCO was also discussed in the context of LGT, and it helped characterise some intermediate samples according to the second, read-mapping approach (see Discussion section 4.3). Lastly, since our aim at this stage was to detect *Wolbachia* presence rather than to recover insertions, we retained a strict percent-identity threshold (>95%) rather than relaxing it, accepting that this may exclude more divergent or degraded sequences.

In the second approach, we used read-mapping to assess samples independently of the assembly-based screening. Since the first approach identifies *Wolbachia* sequences from the assembled contigs, mapping complements it by evaluating the total read data against complete, chromosome-level, reference genomes of *Wolbachia*, quantifying the breadth of coverage and the relative number of reads supporting it. For this, we first selected a total of 12 reference *Wolbachia* genomes based on the work by Twort and colleagues (2022), which mostly represent supergroups A and B, the only ones known to infect butterflies, and supergroups C, F, and D, increasing the exploratory power of the analysis (used reference genomes are in Table S2). Additionally, each reference genome comes from a different and identified *Wolbachia* strain and we acknowledge that more *Wolbachia* genomes are available in Twort and colleagues (2022), but these were excluded either because they were suppressed from NCBI or because they do not contain a fully assembled chromosome, but fragmented contigs or scaffolds. The references were then transformed into an appropriate database with Bowtie2 v.2.4.2. (Langmead & Salzberg, 2012) and the mapping was performed with the same program, resulting in SAM files. After converting SAM files to BAM and sorting the BAM files, we used BEDTools v.2.26.0 (Quinlan & Hall, 2010) to convert them to FASTQ format, obtaining mostly *Wolbachia*-filtered reads. We then remapped these reads against the same set of 12 reference *Wolbachia* genomes, consequently removing alignments not supported as proper pairs and producing a BAM containing only *Wolbachia*-derived reads.

We then used the mapped BAM files to obtain the breadth of coverage and mean depth of every sample against each reference with SAMtools v.1.14 (Danecek et al., 2021). However, mean depth was not used as an infection metric, as it is directly comparable across samples only when sequencing effort is equal, which was not the case in the present study. We thus established two criteria regarding the obtained metrics to identify infection in our samples. The primary criterion was the actual breadth of coverage (*i.e.,* the proportion of the reference genome covered by reads), for which a conservative threshold was applied at 70% breadth. This

choice reflected the distribution of values across samples, the limitations of the source material (collection preservation, locality, depth of sequencing, and suboptimal tissue – leg), and the fact that breadth directly reflects how complete a *Wolbachia* genome is represented across the reads. As a second criterion, we normalised the number of Wolbachia-mapping reads by the total number of cleaned reads in the corresponding library, expressed as reads per million (RPM), to make read abundance comparable across samples in place of mean depth. The total number of reads per specimen was determined from the cleaned paired-end FASTQ files before any *Wolbachia*-specific filtering. For each sample, read counts were obtained separately from the R1 and R2 files using *seqkit stats*, and the two values were summed to calculate the total number of reads. These totals were subsequently used as the denominator for reads-per-million (RPM) normalisation. The formula used was:

$$(N\ Wolbachia\ reads \div N\ total\ cleaned\ reads) \times 10^6$$

Importantly, since RPM is an abundance measure spanning several orders of magnitude, proportional (fold) differences are more informative than absolute differences and, thus, a $\log_{10}$ transformation was applied (Table S3; Figure S1). Finally, in our study, samples falling below the 70% breadth threshold were designated as uninfected, which denotes the absence of a sufficiently strong and genome-wide signal rather than concluding absence of *Wolbachia.* Every script used, including program parameters, is provided in Supplementary Material M1. Additionally, a table containing the mapping metrics for all samples, including number of mapped reads, across the 12 references, is included in Supplementary Material M1.

As part of the rationale of the read-mapping approach, we also intended to provide an initial description of the supergroups and strains infecting our samples. However, this thesis did not extensively characterise strains since short Illumina reads, as the ones used to obtain our genomic datasets, make it challenging to phase haplotypes of closely related strains in read-mapping analyses (Vancaester & Blaxter, 2023), which can underestimate read-mapping identifications. Moreover, we did not perform phylogenetic placement based on conserved single-copy orthologues. This could be limited by inconsistent orthologue recovery across our samples, since incomplete assemblies and duplicated signals would limit the set of orthologues shared by all samples. In addition, phylogenetic placement remains dependent on the reference genomes included, so samples whose closest relatives are not represented in the available references could be placed without informative resolution. Thus, our initial description can narrow the identification to the best-supported candidate supergroups and strains based on

similarity, but it cannot provide a true strain-level characterisation of the infected samples. We are using the results derived from this mapping as a description that can guide future analyses in Eudaminae. Lastly, insertions of bacterial sequence into the host genome were not systematically characterised with any of the methods, although mapping patterns compatible with such insertions are considered in the context of LGT in Discussion, section 4.3.

3. Results

3.1. Dataset overview and data Quality Control

The dataset comprised 77 skipper butterfly samples collected across five Neotropical countries: Peru (n = 35), Cuba (n = 22), Panama (n = 16), Brazil (n = 3), and Jamaica (n = 1). After taxonomic curation, the material represented 18 genera. Following taxonomic investigation of specimens with uncertain placement, the final genus-level distribution was as follows: *Telegonus* (n = 6), *Cecropterus* (n = 13), *Spicauda* (n = 5), *Aguna* (n = 3), *Chioides* (n = 14), *Urbanus* (n = 18), *Spathilepia* (n = 1), *Autochton* (n = 3), *Bungalotis* (n = 2), *Cogia* (n = 1), *Dyscophellus* (n = 1), *Ectomis* (n = 2), *Entheus* (n = 1), *Epargyreus* (n = 1), *Euriphellus* (n = 1), *Narcosius* (n = 1), *Nascus* (n = 3), and *Porphyrogenes* (n = 1).

Quality assessment of the raw Illumina reads identified the main sequencing artefacts as adapter contamination, duplicated reads, and poly-G tails. These were removed by adapter trimming, quality filtering, and deduplication. Post-filtering quality assessment confirmed consistently high-quality read sets across the dataset. The MultiQC summary statistics showed a median read length of 150 bp, a median sequencing output of approximately 33 million reads per sample, and an average GC content of 36.6%. FastQC and MultiQC reports are provided in Supplementary Material M1.

3.2. Assembly-based screening for *Wolbachia* presence

The initial assembly-based screening identified candidate *Wolbachia*-associated sequences in 26 out of the 77 analysed samples (33,76% of the analysed dataset). Within this method, a specimen was screened as potentially infected when at least one assembled contig was retained after the applied BLASTX table-filtering criteria (see section 2.3), against the *w*Mel (supergroup A representative) and *w*Pip (supergroup B representative) protein references. The number of retained contigs, as counted by the extracted number of contig nodes from a filtered table, and including the elimination of duplicated hits, varied substantially among the candidate-positive samples, ranging from one in PMLBSc023 to 1,762 in PMLBSc040 (for node count extraction script, see Supplementary Material M1, *table_filters.sh*; the full list of

extracted node counts is in Supplementary Material M1). This stage was deliberately permissive, as a single qualifying contig was sufficient for a specimen to be classified as potentially infected. Therefore, the 26 butterfly specimens in which *Wolbachia* sequences were identified by BLASTX represent an initial screen rather than a final estimate of detectable infection prevalence. Additionally, BUSCO assessment of the candidate *Wolbachia* contig set showed uneven recovery of conserved bacterial orthologues among specimens with complete BUSCO recovery ranging from 0% to approximately 86.3%, with a mean of 58.1% across the evaluated assemblies (Figure 1). Four samples, namely PMLBSc023, PMLBSc033, PMLBSc065, and PMLBSc072, contained no complete BUSCOs, whereas 13 assemblies recovered at least 80% of the expected orthologues. The remaining ones showed intermediate levels of completeness. BUSCOs' full completeness assessment per sample is provided in Supplementary Material M1.

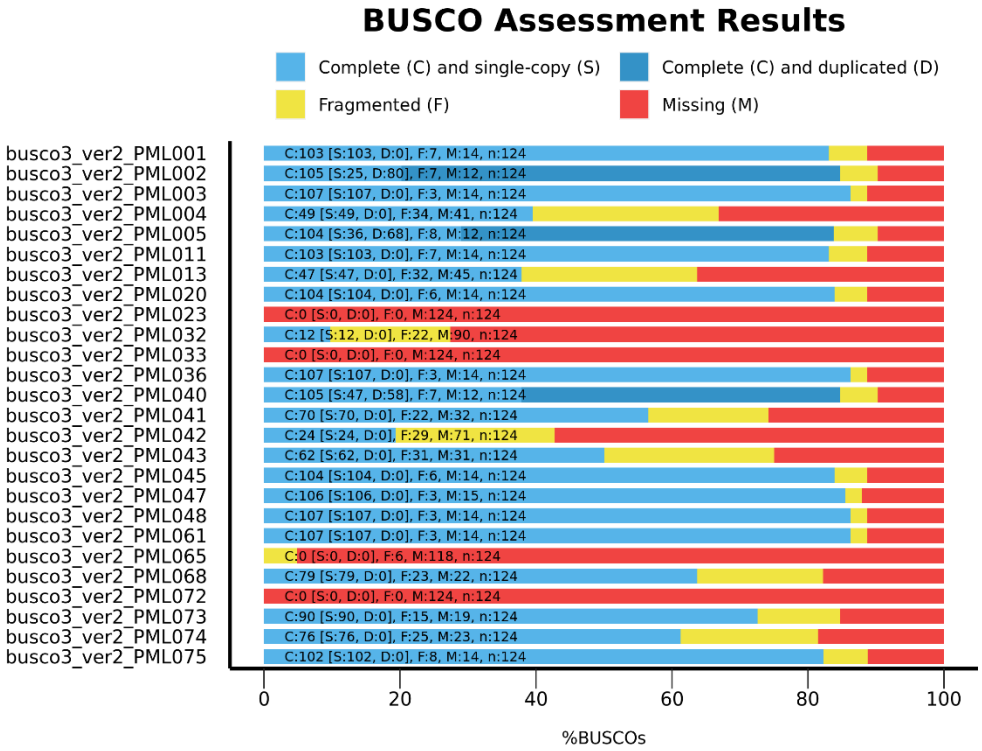

Figure 1: BUSCO completeness of candidate Wolbachia-associated assemblies identified by the initial BLASTX screen. Each horizontal bar represents the BUSCO completeness profile for an individual sample. Blue segments indicate complete BUSCOs (single-copy or duplicated), yellow segments represent fragmented BUSCOs, and red segments indicate missing BUSCOs. Completeness ranged from 0% to 86.3%.

3.3. Read-mapping infection detection and classification

Read mapping against the 12 selected *Wolbachia* reference genomes provided the main evidence for the final infection classification. Based on the combined evaluation of breadth of coverage and normalised mapped-read abundance (RPM, and $\log_{10}$ RPM), 20 out of the 77

analysed samples were retained as infected, corresponding to an infection prevalence of 25.97%. Among the infected samples, 3 carried supergroup A only, 13 carried supergroup B only, and, notably, 4 showed evidence of A+B co-infection (Figure 2; Table 1).

Table 1: *Wolbachia* infection status of the 20 positive skipper samples identified by the read-mapping approach. For each sample, the best-supported reference in supergroups A and B is shown with its breadth of coverage and normalised read abundance (RPM). Reference labels denote closest similarity, not formal strain identity.

Sample	Supergroup	Strain signal (based on reference)	Breadth of coverage (%)	RPM	Infection status	$\log_{10}$ (RPM)
PMLBSc001	A	*w*Mel	92.2	1843	A	3.265
	B	*w*Pip	7.2	135		2.13
PMLBSc002	A	*w*Mel	96.2	4722	A + B	3.674
	B	*w*No	88	1550		3.19
PMLBSc003	A	*w*Irr	5.4	64	B	1.808
	B	*w*Pip	80.6	1333		3.125
PMLBSc004	A	*w*Irr	3.2	22	B	1.346
	B	*w*Pip	70.6	498		2.698
PMLBSc005	A	*w*Mel	96.1	4940	A + B	3.694
	B	*w*No	88	1710		3.233
PMLBSc011	A	*w*Mel	96	5878	A + B	3.769
	B	*w*No	66.5	544		2.736
PMLBSc013	A	*w*Irr	3.3	19	B	1.29
	B	*w*Pip	70.4	466		2.669
PMLBSc020	A	*w*Mel	95	8354	A	3.922
	B	*w*Pip	10.3	304		2.482
PMLBSc036	A	*w*Irr	3.6	35	B	1.54
	B	*w*Pip	79.6	1654		3.219
PMLBSc040	A	*w*Mel	96.7	3154	A + B	3.499
	B	*w*No	89.9	1833		3.263
PMLBSc041	A	*w*Ri	4.8	12	B	1.072
	B	*w*No	77.7	253		2.403
PMLBSc043	A	*w*Ri	6.3	14	B	1.15
	B	*w*Pip	73.9	264		2.421
PMLBSc045	A	*w*Mel	93.3	1230	A	3.09
	B	*w*Pip	6.1	44		1.642
PMLBSc047	A	*w*Irr	3.3	24	B	1.375
	B	*w*Pip	77	360		2.556
PMLBSc048	A	*w*Irr	5.6	48	B	1.677
	B	*w*Pip	82.9	2306		3.363
PMLBSc061	A	*w*Irr	4.9	35	B	1.545

	B	*w*Pip	82.3	1550		3.19
PMLBSc068	A	*w*Ri	6.8	41	B	1.613
	B	*w*Pip	76	790		2.898
PMLBSc073	A	*w*Ri	7	17	B	1.236
	B	*w*Pip	77	316		2.5
PMLBSc074	A	*w*Ri	6.5	26	B	1.42
	B	*w*Pip	75.2	475		2.676
PMLBSc075	A	*w*Mel	17.2	49	B	1.688
	B	*w*Pip	78.6	735		2.867

The primary metric used for the infection classification was the breadth of coverage at the threshold of 70%. When the breadth value per sample was compared across the dataset, 52 samples fell below 6% and the 20 considered infected fell above 70%, while only five occupied the intermediate range: PMLBSc032 (58.5%), PMLBSc033 (16.0%), PMLBSc042 (66.3%), PMLBSc065 (31.4%), and PMLBSc072 (13.2%) (Table S3). Notably, despite their intermediate breadth, all five samples reached $\log_{10}$ RPM values within the range of the samples above the threshold (Table S3), and, additionally, they presented low number or no recovered BUSCO orthologues at all based on the assembly-based approach. Two extreme cases, samples PMLBSc031 and PMLBSc046, presented very low breadth of coverage (0.14% and 0.30%, respectively) but $\log_{10}$ RPM also above the gap which separated all samples with breadth higher than 70%. All of these samples were considered uninfected based on the breadth threshold, and they are further discussed in Discussion section 4.3.

Additionally, one near-threshold exception, the supergroup B signal in PMLBSc011, was evaluated separately because it occurred together with an unambiguous supergroup A infection and was supported by 66.5% breadth and comparatively high normalised abundance. The justification for retaining this signal is provided in Discussion section 4.3.

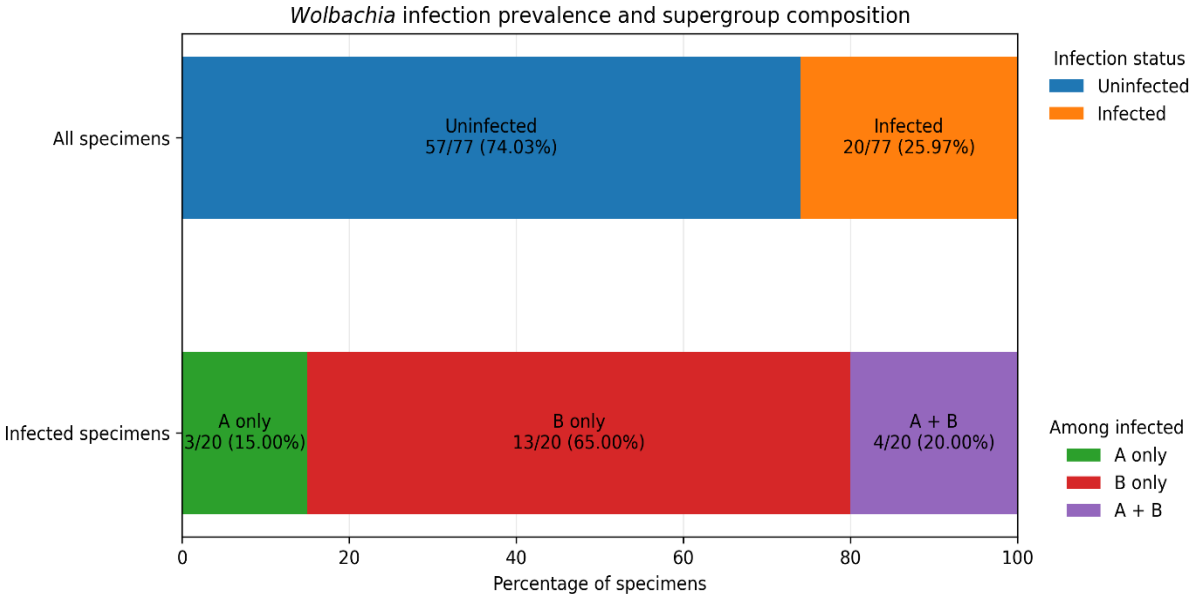

Figure 2: Mapping-based Wolbachia infection prevalence and supergroup composition across the analysed Eudaminae dataset. Of the 77 analysed samples, 20 were classified as infected and 57 as uninfected under the conservative mapping-based framework. Among the 20 infected samples, 3 carried supergroup A only, 13 carried supergroup B only, and 4 showed evidence of A+B co-infection. Percentages in the upper bar are calculated relative to the complete dataset, whereas percentages in the lower bar are calculated relative to the infected subset.

To complement our findings using breadth of coverage, mapped-read abundance was normalised by library size and expressed on a logarithmic scale, as it spanned several orders of magnitude. This also contrasted infected and uninfected samples: samples exceeding the 70% breadth threshold occurred in the higher-RPM part of the distribution, whereas the low-breadth samples were mostly concentrated at lower RPM values (Figure S1), with a few samples further discussed in Discussion section 4.3. Across the $\log_{10}$ RPM distribution, the largest gap occurred between PMLBSc025 (RPM 67.7; $\log_{10} = 1.831$) and PMLBSc046 (RPM 183.8; $\log_{10} = 2.264$), with no samples falling within this interval of approximately 0.43 log units, and all samples with breadth higher than 70% falling above $\log_{10} = 2.264$ (Table S3, filtered by ascending $\log_{10}$ RPM values for visualization). Because some samples below the 70% threshold also had a $\log_{10}$ RPM higher than $\log_{10} = 2.264$, the gap is treated in this study as a boundary capable of distinguishing all samples considered as infected by breadth, but not as a main threshold or as evidence of two discrete statistical populations.

For the infected samples, the mapping profiles were next interpreted in terms of closest-reference similarity. Considering that 4 supergroup A strains, 4 supergroup B strains, and additional supergroup C, D, and F were present in the mapping procedure, supergroup A-associated signals were consistently most similar to the *w*Mel reference, while supergroup B-associated signals were most similar to either *w*Pip or *w*No. The A+B co-infections combined a *w*Mel supergroup A signal with a *w*No supergroup B signal (Table 1). These assignments should be interpreted cautiously as the closest supported similarity among the tested references, rather than as a definitive strain characterisation.

3.4. Distribution across genera

Mapping-supported *Wolbachia* infections were detected in seven of the eighteen represented genera: *Cecropterus*, *Chioides*, *Autochton*, *Spicauda*, *Nascus*, *Telegonus*, and *Urbanus* (Figure 3). The largest numbers of infected specimens were observed in *Cecropterus*, with 6 of 13 specimens, and *Chioides*, with 5 of 14. All three sampled *Autochton* specimens were infected. Infections were also detected in 2 of 3 *Nascus* specimens and 2 of 5 *Spicauda* specimens. *Telegonus* and *Urbanus* each contained one infected specimen, corresponding to 1 of 6 and 1 of 18 specimens, respectively. No infections were detected in the remaining eleven genera. Because the number of specimens differed substantially among genera, these values are presented descriptively rather than as directly comparable genus-level prevalence estimates.

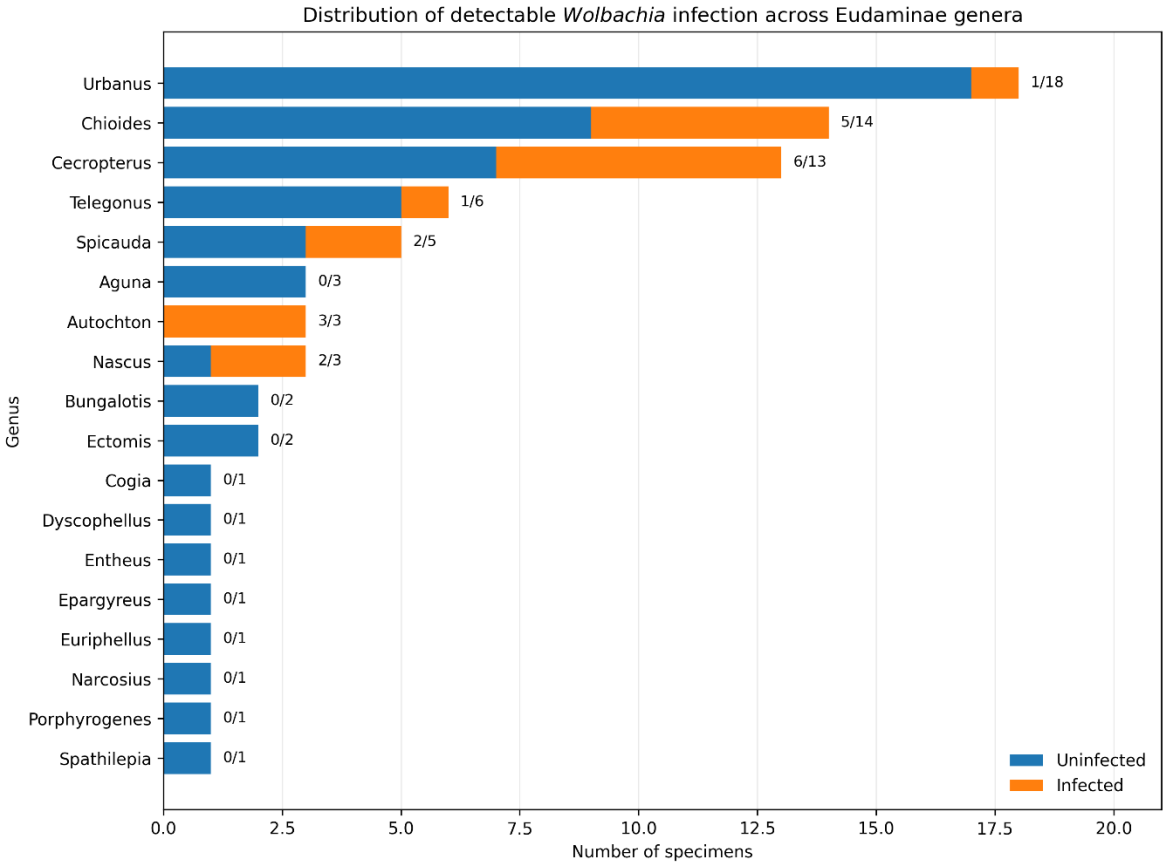

Figure 3: Distribution of detectable Wolbachia signal across Eudaminae genera based on the read mapping approach. Each horizontal bar represents the total number of specimens analysed per genus, partitioned into uninfected and infected individuals.

4. Discussion

4.1. Conceptual framework and key findings

The study aimed to detect and quantify *Wolbachia* infections across a dataset of neotropical Eudaminae butterflies (skippers). As opposed to traditional marker-based screening approaches, such as the analysis of amplified fragments of the MLST loci, the analysis in this thesis used host genomic datasets and searched for evidence of *Wolbachia*-associated sequence signal within them. The framework of *Wolbachia* detection used two main approaches: (1) protein-level similarity of assembled contigs to screen for presence of *Wolbachia* proteins, and (2) read mapping against available complete *Wolbachia* reference genomes for classification. By using two approaches, this thesis aimed to account for the heterogeneity of the dataset, which varied in the depth of sequencing of host DNA, and for the use of leg tissue as the source material, in which *Wolbachia* densities may be low and infection signal correspondingly weak.

The read-mapping analyses detected *Wolbachia* signal in 20 of the 77 samples, corresponding to an overall infection rate of 25.97%, while the assembly based one screened

26 out of the 77 samples to be potentially infected. Importantly, samples without a detectable signal cannot be interpreted as definitively uninfected, particularly if low titre infection was present, since the study's approach depends on recoverable bacterial signal in the host-derived WGS data. All 20 read-mapping infected samples were also flagged by the assembly-based screen, while six additional samples were only flagged by the latter (PMLBSc023, PMLBSc032, PMLBSc033, PMLBSc042, PMLBSc065, PMLBSc072). Differences between both approaches, and particular characteristics of these samples is further discussed in sections 4.2 and 4.3.

The detected infection signal was predominantly associated with supergroup B, while supergroup A was less frequent, in accordance with *Wolbachia*-Lepidoptera literature and other *Wolbachia*-insect broader studies (Duplouy & Hornett, 2018; Vancaester & Blaxter, 2023). Furthermore, the read-mapping reference comparisons indicated that the supergroup A-associated signal was consistently *w*Mel, while the supergroup B-associated signal was more variable, mainly being a *w*Pip signal with a smaller number of *w*No cases. The four A+B samples followed the same supported reference pattern, *w*Mel + *w*No. This suggests a pattern of strong affinity between the identified *Wolbachia* and specific used references within the dataset, but not a definitive strain-level characterisation.

Detectable *Wolbachia* was distributed across seven of the eighteen represented genera, with the highest numbers observed in *Cecropterus* and *Chioides*. This heterogeneous distribution suggests that infection is not evenly structured across the sampled host taxa. Nevertheless, since the sampling depth and number of representative samples differed strongly among genera, these patterns should be interpreted descriptively rather than as robust genus-level prevalence estimates.

The butterfly family Hesperiidae, and especially the Eudaminae subfamily are understudied outside the scope of taxonomic phylogenetics (Li et al., 2019; Toussaint et al., 2018; Warren et al., 2008; Warren et al., 2009). Recent efforts have, however, substantially expanded the knowledge about this clade by investigating the evolutionary ecology of defences against predation (Linke et al., 2022), diversification of wing morphology (Matos-Maraví et al., 2025) and *Wolbachia*-Lepidoptera relationship (Ribeiro et al., 2025). Here, for the first time, I work on a multi-genera scale within this clade and show both an infection assessment and an initial description of the strains that might be present in these butterflies, providing relevant information that can be used for future studies in the field.

4.2. Methodological evaluation of the infection detection approaches

The two complementary detection approaches differed in their intended purpose and strictness. The main difference between both approaches (*i.e.*, 26 detected with the assembly based and 20 with the read-mapping) was expected since the assembly-based approach was deliberately permissive, with a sample considered potentially positive when at least one assembled contig was retained after the BLASTX filtering criteria were applied. Consequently, even a small amount of conserved *Wolbachia*-associated sequence could result in a candidate-positive classification. In contrast, the mapping-based approach required read support to be distributed across a substantial proportion of a complete reference genome, therefore producing a more conservative result.

The complete BUSCO recovery in the 26 candidate-positive samples ranged from 0% to 86.3% (Figure 1; Supplementary Material M1), further demonstrating the uneven quality of the bacterial contigs obtained through assembly. The wide range of completeness values is consistent with substantial variation in effective bacterial coverage, bacterial abundance, and assembly fragmentation among samples. Additionally, it could be a signal of lateral gene transfer (LGT), which would yield a similarly fragmented set of recovered orthologues, and this is further discussed in a broader context in section 4.3. Similar genome-mining studies have also shown that low effective coverage may produce detectable *Wolbachia* sequence without yielding a credible or sufficiently complete bacterial assembly (Twort et al., 2022) and, thus, the assembly results should therefore be interpreted as a presence screening instead of genome-level infection calls.

Read mapping was considered more suitable for the final classification because breadth of coverage evaluates whether the observed signal is randomly distributed across the reference genome rather than restricted to one or a few conserved regions. Indeed, in the most recent methodological guideline for *Wolbachia* detection from host-genome data, different instances of read-mapping approaches are recommended (Valerio et al., 2023), including other methods that directly use reads as input such as screening *Wolbachia* with Kraken2 (Wood et al., 2019; Lu et al., 2022). We did not, however, use Kraken2 in our approaches because our understanding is that it serves as an initial screening, which was covered by our assembly-based approach. Read-mapping was additionally important because protein sequences can be conserved across *Wolbachia* supergroups and, consequently, the occurrence of BLASTX hits against both the supergroup A and B protein references could not by itself distinguish a genuine

A+B co-infection from cross-hits involving homologous proteins. Thus, in our study, the assembly-based analysis was used to screen for presence, and in turn potential infection, whereas the assessment of actual infection, infection per genus, closest-reference similarity, and co-infection was based on the read-mapping results.

Taken together, these differences show that the two approaches are best treated as complementary rather than a single combined classifier. The overlap between both approaches in terms of *Wolbachia*-positive samples showed that the assembly screen recovered every confidently infected sample while additionally flagging weaker or partial signals that the stricter breadth of coverage criterion later excluded. Thus, the assembly-based approach is a reliable and computationally lighter screening method, and it can be sufficient when coverage is high and uniform across samples, and that the source tissue is more appropriate (*e.g.*, abdomen). The read-mapping approach is stricter and better suited for heterogeneous data in which sequencing coverage and bacterial abundance vary, although it requires additional steps and greater computational effort. Thus, the choice between the two approaches largely depends on the study design, available tissue, and sequencing depth potential. Finally, we recommend always using read mapping when recovery of *Wolbachia* contigs is highly uneven, and BUSCOs show many samples with high number of fragmented or missing BUSCOs.

4.3. A deeper look into intermediate samples, interpretation of metrics, and LGT in the context of the approaches used.

Because the classification depended on recoverable bacterial signal from limited source material, in our study "uninfected" denotes the absence of a sufficiently strong and genome-wide, as stated in methods. Breadth of coverage was treated as the primary criterion because a high value requires *Wolbachia* reads distributed across much of the reference, working as a conservative metric in the face of sample and source material limitations. RPM was treated as complementary, since it quantifies the abundance of *Wolbachia*-mapping reads but cannot assert how much of the genome they cover. In our study, the two metrics largely agreed, since the samples at or above 70% breadth (*i.e.,* considered infected) were set apart, in terms of abundance, from most of those with breadth lower than 70% by a log-scale RPM gap wider than any separation within the infected group itself (Figure S1; Table S3). Five samples (PMLBSc032, PMLBSc033, PMLBSc042, PMLBSc065, and PMLBSc072, Table S3), however, presented intermediate breadth (between 6% and 70%) and are further discussed.

First, PMLBSc033 was weak on both metrics (breadth = 16.0%; RPM = 38.6; $\log_{10}$ = 1.587) and recovered no complete BUSCOs, being directly considered as uninfected. In contrast, PMLBSc072 had low breadth of coverage (13.2%) but high RPM value (536.5; $\log_{10}$ RPM = 2.730), comparable to several samples exceeding the 70% breadth threshold. Similarly, PMLBSc065 had an RPM of 186.8 ($\log_{10}$ RPM = 2.271) but only moderate breadth of coverage (31.4%), which remained well below the threshold. The pattern of both samples means that their reads were concentrated on one or a few fractions of the genome, which can arise from their accumulation on conserved or repetitive regions. Such an accumulation can, for instance, result from cross-mapping from another supergroup at their shared conserved regions. Alternatively, shallow sequencing (such as for PMLBSc072; Table S3) could produce the same low breadth if *Wolbachia* is genome-wide but present at low titre or poorly represented in the leg tissue, so that only its best-represented regions are recovered. The pattern could also represent fragmented or partial matches belonging to insertions resulting from LGT, which can deteriorate over time (Nikoh et al., 2008; Tvedte et al., 2022). This would also agree with the fact that PMLBSc072 did not recover any type of BUSCO orthologues, while PMLBSc065 recovered only a few fragmented.

Additionally, PMLBSc032 and PMLBSc042 covered more than half of the reference (58.5% and 66.3%), but still below the threshold, with abundances (RPM 230.1 and 187.2 respectively; $\log_{10}$ = 2.362 and 2.272 respectively; Table S3) falling at the low end of the samples with breadth above 70%. However, both had large libraries (71.1 and 111.5 million reads) and reached the highest breadth and most complete BUSCOs of the five (12 and 24, Figure 1; Supplementary Material M1). The below threshold breadth is, therefore, unlikely reflecting insufficient sequencing, cross-mapping or LGT, and more likely representing a true infection at low titre excluded by the conservative breadth cutoff. Finally, outside the intermediate range previously discussed, two samples (PMLBSc031 and PMLBSc046), reached comparable $\log_{10}$ RPM values (2.411 and 2.264) on negligible breadth (0.14% and 0.30%). They were classified as uninfected without further consideration, but they still highlight, as extreme cases, how RPM can reach the range of infected samples with almost no breadth.

Sample PMLBSc011 represents a similar case, but here our analysis helped to retain rather than to exclude this intermediate sample. Its supergroup A signal was unambiguous (96.0% breadth of coverage against *w*Mel; RPM = 5,877.6; $\log_{10}$ = 3.769), while its additional

supergroup B signal fell just below the threshold at 66.5% breadth against *w*No (Table 1). Notably, however, the breadth of the B signal (66.5%) is essentially identical to that of the excluded PMLBSc042 (66.3%), so breadth alone does not separate the two cases. We then decided to retain the supergroup B infection of PMLBSc011 based on a combination of assessments. First, its abundance (RPM = 544; $\log_{10} = 2.736$) is roughly two to three times that of PMLBSc042, placing it within the range of the confidently infected samples rather than at its lower end (Table S3). Second, as mentioned previously, a dominant supergroup A infection is a plausible source of a false B signal when mapping, since the supergroup A can cross-map onto the conserved regions shared by both supergroups. Consequently, such cross-mapping would be confined to those regions and be of lower breadth, but the B signal of PMLBSc011 spans two-thirds of the reference, which is inconsistent with the cross-mapping artefact and more likely indicates a distinct B genome. A near-threshold signal from a single supergroup, such as that of PMLBSc042, lacks this second signal for comparison, so its breadth stays ambiguous, and is treated conservatively as below threshold.

Another consideration is related to LGT, since *Wolbachia* sequences are known to be integrated into lepidopteran genomes with notable frequency (Li et al., 2022). One consequence is that samples recovering few or no conserved bacterial orthologues under BUSCO, as in the intermediate-breadth cases above, could reflect a *Wolbachia*-derived insertion in the host genome rather than infection, in cases in which the inserted gene or genes would undergo accumulation of mutation, fragmentation, and degradation over time (Nikoh et al., 2008; Tvedte et al., 2022). However, the read-mapping procedure cannot establish an insertion on its own, because mapping locates reads along a reference genome, but it cannot show whether the corresponding sequence is within the host genome. Therefore, confirming a genuine transfer event would require placing the *Wolbachia*-matching regions in their host genomic context (Dunning Hotopp et al., 2007), for instance by identifying flanking host genes, which was beyond the scope of this study.

Lastly, all reported strain-associated signals must be understood as the closest-reference similarities rather than formal strain identifications. The mapping analysis compared the samples with a restricted set of 12 available reference genomes (Table S2), meaning that the highest-supported reference represents the closest genome within that particular reference set and not necessarily the actual strain infecting the specimen. Furthermore, short-read data limits the ability to separate or phase closely related genomes when multiple strains are present, while

fragmented or uneven sequence recovery reduces the number of comparable loci available for phylogenetic inference (Vancaester & Blaxter, 2023). Thus, a formal phylogenetic placement was not performed, since, under the present data limitations, constructing a phylogeny would not resolve incomplete reference representation or short-read phasing and could instead provide an unjustified impression of precise strain-level assignment. Finally, the mapping results provide a conservative initial description of supergroup signal and reference affinity that can guide subsequent targeted genomic analyses.

4.4 Infection prevalence: supergroup assessment

Under the mapping-based classification, detectable *Wolbachia* infection was identified in 20 of the 77 analysed specimens, corresponding to an overall prevalence of 25.97% (Figure 2). Published estimates indicate that *Wolbachia* prevalence can vary substantially among Lepidoptera taxa, for instance, in *Acraea* butterflies in which infection frequencies of approximately 30% have been reported, whereas prevalence of approximately 80% has been observed in the species of the African *Bicyclus* (Jiggins et al., 2001; Duplouy & Brattström, 2018). On this basis, the prevalence observed in the present dataset falls in the lower end of this reported range. However, these values should not be treated as directly equivalent, as the studies differ in host taxonomy, sampling effort per species, tissue source, and detection methodology.

Furthermore, the infected subset showed a clear predominance of the supergroup B-associated signal. Among the infected specimens, 13 carried only supergroup B, three carried only supergroup A, and four had detectable signals from both supergroups (co-infection) (Figure 2). Thus, including the four co-infections, supergroup B was present in 17 out of the 20 infected samples, whereas supergroup A was present in seven, which is consistent with previous observations about the prevalence of supergroup B over supergroup A in Lepidoptera. Additionally, the presence of four A+B co-infections also agrees with previous genomic and marker-based studies showing that multiple *Wolbachia* lineages can coexist within individual lepidopteran hosts (Duplouy & Hornett, 2018; Twort et al., 2022; Vancaester & Blaxter, 2023).

4.5 *Wolbachia* across Eudaminae: distribution and dynamics

Detectable *Wolbachia* infection was distributed unevenly across the represented Eudaminae genera. Mapping-supported infections occurred in seven of the eighteen genera: *Cecropterus*, *Chioides*, *Autochton*, *Spicauda*, *Nascus*, *Telegonus*, and *Urbanus* (Figure 3). All three sampled *Autochton* specimens were infected, as were six of the thirteen *Cecropterus*, five of the fourteen *Chioides*, two of the five *Spicauda*, and two of the three *Nascus*. In contrast,

only one infection was detected among six *Telegonus* specimens and one among eighteen *Urbanus* specimens. No infection was detected in the remaining eleven genera. Importantly, the genus-level values must be interpreted in relation to the highly uneven sampling structure, as thirteen of the eighteen genera were represented by three or fewer specimens, and all eleven genera without a detected infection belonged to this low-sampled group. Consequently, the absence of detectable signal in these genera cannot be treated as absolute evidence that *Wolbachia* is absent from the genus itself. Similarly, the detection of infection in all three *Autochton* specimens does not demonstrate that the infection is fixed or consistently prevalent throughout that genus.

Considerable heterogeneity of *Wolbachia* presence has also been reported in other Lepidoptera systems. A broad genomic screen by Twort et al. (2022) detected *Wolbachia* in only 20 out of 106 surveyed lepidopteran species and found that many identified host-symbiont associations had not been previously recorded. At a geographical scale, infection frequencies in North American *Lycaeides* butterflies among sampled localities ranged from complete absence to whole-population infection, with several *Wolbachia* lineages showing different host species and geographical distributions (Shastry et al., 2022). Similarly, Ahmed and colleagues. (2016) identified identical or closely related *Wolbachia* lineages in different butterfly and moth hosts, together with cases of multiple infection. Overall, these studies show that *Wolbachia* distribution in Lepidoptera is rarely uniform across a clade. Infection status can vary between closely related host species, differ between populations of the same species across localities (Shastry et al., 2022; Ribeiro et al., 2025), and involve distinct bacterial lineages that each associate with particular hosts. Ribeiro and colleagues (2025), for instance, examined four closely related species of the Eudaminae genus *Spicauda*, showing that the four differed notably in infection status. In this study, infections were frequent in *S. simplicius* and *S. procne*, whereas no infection was detected in *S. tanna* or *S. teleus*, and the infecting *Wolbachia* lineages were host-specific with *S. simplicius* including individuals carrying infections from both supergroups A and B. Additionally, their phylogenetic and demographic analyses supported frequent turnover of *Wolbachia* in *Spicauda* populations rather than stable inheritance of a single ancestral infection across the genus. The present multi-genera dataset cannot resolve infection dynamics at the same population or phylogenetic level, but its heterogeneous distribution extends the observation of uneven *Wolbachia* occurrence across a substantially broader taxonomic representation of Eudaminae.

Additionally, the ecology of Eudaminae makes opportunities for transfer possible: many skipper species occur sympatrically or syntopically and overlap in their daily activity and ecological environments (DeVries et al., 2008; Ribeiro et al., 2025). Such overlap may facilitate indirect contact through shared host plants or parasites, while inter-specific hybridisation may provide an additional route among closely related species (Shastry et al., 2022; Twort et al., 2022). However, eco-evolutionary studies in Eudaminae are still scarce, and all of these processes would require appropriate assessment to help clarify how *Wolbachia* is transferred in the clade. Thus, the present results are only compatible with this hypothesis and do not demonstrate horizontal transmission, as neither host-symbiont phylogenetic congruence nor direct ecological transmission routes were tested.

Taken together, the scattered genus-level distribution, the occurrence of supergroup B-associated signal in several distinct hosts, and the A+B co-infections described in Section 4.3 are compatible with a dynamic infection system shaped by repeated acquisition, persistence, and loss, with horizontal transmission being a plausible driver of such dynamics. Previous studies have inferred frequent horizontal transfer in Lepidoptera from the sharing of closely related bacterial lineages among different host species and from incongruence between host and *Wolbachia* relationships (Ahmed et al., 2016; Shastry et al., 2022; Ribeiro et al., 2025).

Despite these constraints, this study provides the first broad multi-genera assessment of *Wolbachia* prevalence in a butterfly subfamily, Eudaminae, extending previous work focused on four *Spicauda* species to a dataset comprising 18 genera and 33 distinct species, establishing an initial map of detectable infections.

4.6 Limitations and future directions

The main limitation of this study is that *Wolbachia* detection depended on the recovery of sufficient bacterial signal from host-derived WGS data. As discussed throughout this thesis, samples classified as uninfected should be interpreted as lacking detectable signal under the applied framework rather than as being definitively free of *Wolbachia*. Low-abundance infections may have remained undetected, particularly because DNA was extracted from leg tissue and because the samples differed in age, origin, storage conditions, and sequencing characteristics. Even though the conservative 70% breadth of coverage threshold reduced the likelihood of fragmented signals being classified as infections, it may also have excluded genuine low titre or poorly recovered infections. Furthermore, the read-mapping descriptions

are strongly influenced by the reference dataset used and, consequently, cannot reflect the real strain identity of the sample.

Uneven sampling across host taxa also limited biological interpretation. Thirteen of the eighteen genera were represented by three or fewer specimens, preventing robust comparisons of prevalence among genera. Similarly, the dataset contained specimens from different geographical localities and collection periods, making it difficult to separate taxonomic, geographical, and temporal effects on infection distribution.

Future work should therefore prioritise more balanced sampling, with multiple individuals from each species and genus and repeated sampling across populations and geographical regions. Targeted validation with PCR amplification of *Wolbachia* typing genes on uncertain or low-signal samples could help distinguish genuinely uninfected specimens from low titre infections. Furthermore, increased sequencing depth, targeted enrichment of *Wolbachia* DNA, long-read sequencing, or the usage of more appropriate tissue, such as abdomens, could also improve genome recovery and enable more reliable separation of co-infecting strains. Finally, reconstructing both host and *Wolbachia* phylogenies would allow infection gains, losses, vertical persistence, and potential horizontal transfers to be tested directly.

5 Conclusion

The study aimed to detect, quantify, and initially characterise *Wolbachia* infections across 77 Neotropical Eudaminae specimens. A whole-genome sequencing-based framework combined assembly-based screening with conservative read mapping against selected reference genomes. The assembly screen identified 26 candidate-positive samples, while 20 samples were retained as infected under the mapping-based classification, corresponding to a detectable prevalence of 25.97%. Supergroup B predominated over supergroup A, and four samples showed evidence of A+B co-infection. The detected signals were most similar to *w*Mel, *w*Pip, or *w*No, with all co-infections showing a *w*Mel-like and *w*No-like combination. Detectable infection occurred in seven of the eighteen represented genera, although uneven sampling limited genus-level comparisons. These observed patterns offer a first reference point for *Wolbachia* in the Eudaminae, and the framework applied here provides a reproducible basis on which future strain-level and phylogenetic analyses can build to resolve how these infections are acquired and maintained.

6 Acknowledgements

We acknowledge the

computational resources provided by the CESNET LM2015042 and the CERIT Scientific Cloud

LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures. This work was supported by the Czech Science Foundation (GAČR grant 22-35084J to

P.M.M.) (THANK OTHERS IF NECESSARY)

7 Data Availability

The data presented in this project belong to the laboratory of Dr. Pável Matos and are currently part of two ongoing projects. Because those projects are not yet published, the data are not publicly available at this time, but they can be obtained from the laboratory upon request.

8 References

Ahmed, M. Z., Breinholt, J. W., & Kawahara, A. Y. (2016). Evidence for common horizontal transmission of *Wolbachia* among butterflies and moths. *BMC Evolutionary Biology 2016 16:1*, *16*(1), 118-. https://doi.org/10.1186/s12862-016-0660-x

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892. https://doi.org/10.1111/1755-0998.13160

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2

Andrews, S. (2010). FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc

Baldo, L., Hotopp, J. C. D., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R., … & 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

Bandi, C., Anderson, T. J. C., Genchi, C., & Blaxter, M. L. (1998). Phylogeny of *Wolbachia* in filarial nematodes. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1413), 2407. https://doi.org/10.1098/rspb.1998.0591

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455. https://doi.org/10.1089/cmb.2012.0021

Beckmann, J. F., Ronau, J. A., & Hochstrasser, M. (2017). A *Wolbachia* deubiquitylating enzyme induces cytoplasmic incompatibility. *Nature Microbiology 2017 2:5*, *2*(5), 17007-. https://doi.org/10.1038/nmicrobiol.2017.7

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., … & Segata, N. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, *41*(11), 1633–1644. https://doi.org/10.1038/s41587-023-01688-w

Bushnell, B. (2014). *BBMap: A fast, accurate, splice-aware aligner.* 9th Annual Genomics of Energy & Environment Meeting, Walnut Creek, CA, March 17-20. https://sourceforge.net/projects/bbmap/..

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chrostek, E., & Gerth, M. (2019). Is anopheles gambiae a natural host of *Wolbachia*? *MBio*, *10*(3). https://doi.org/10.1128/mbio.00784

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

DeVries, P. J., Austin, G. T., & Martin, N. H. (2008). Diel activity and reproductive isolation in a diverse assemblage of Neotropical skippers (Lepidoptera: Hesperiidae). *Biological Journal of the Linnean Society*, *94*(4), 723–736. https://doi.org/10.1111/J.1095-8312.2008.01037.X

Duplouy, A., & Brattström, O. (2018). *Wolbachia* in the genus *Bicyclus*: a forgotten player. *Microbial Ecology*, *75*(1), 255–263. https://doi.org/10.1007/S00248-017-1024-9

Duplouy, A., & Hornett, E. A. (2018). Uncovering the hidden players in Lepidoptera biology: The heritable microbial endosymbionts. *PeerJ*, *2018*(5). https://doi.org/10.7717/peerj.4629

Duron, O., Wilkes, T. E., & Hurst, G. D. D. (2010). Interspecific transmission of a male-killing bacterium on an ecological timescale. *Ecology Letters*, *13*(9), 1139–1148. https://doi.org/10.1111/J.1461-0248.2010.01502.X

Shropshire, D.J., On, J., Layton, E. M., Zhou, H., & Bordenstein, S. R. (2018). One prophage WO gene rescues cytoplasmic incompatibility in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(19), 4987–4991. https://doi.org/10.1073/pnas.1800650115

Dunning Hotopp, J. C., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M., … & Werren, J. H. (2007). Widespread lateral gene tansfer from intracellular cacteria to multicellular eukaryotes. *Science*, *317* (5845), 1753–1756. https://doi.org/10.1126/science.1142490

Dunning Hotopp, J. C. (2011). Horizontal gene transfer between bacteria and animals. *Trends in Genetics*, 27 (4), 157–163). https://doi.org/10.1016/j.tig.2011.01.005

Dyson, E. A., & Hurst, G. D. D. (2004). Persistence of an extreme sex-ratio bias in a natural population. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(17), 6520–6523. https://doi.org/10.1073/pnas.0304068101

Dyson, E. A., Kamath, M. K., & Hurst, G. D. D. (2002). *Wolbachia* infection associated with all-female broods in *Hypolimnas bolina* (Lepidoptera: Nymphalidae): evidence for horizontal transmission of a butterfly male killer. *Heredity*, *88*(3), 166–171. https://doi.org/10.1038/sj.hdy.6800021

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*, 127–149. https://doi.org/10.1146/annurev.ecolsys.110308.120206

Ehrlich, P. R., & Raven, P. H. (1964). Butterflies and plants: A study in coevolution. *Evolution*, *18*(4), 586. https://doi.org/10.2307/2406212

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Gerth, M., & Bleidorn, C. (2016). Comparative genomics provides a timeframe for *Wolbachia* evolution and exposes a recent biotin synthesis operon transfer. *Nature Microbiology 2016 2:3*, *2*(3), 16241-. https://doi.org/10.1038/nmicrobiol.2016.241

Heath, B. D., Butcher, R. D. J., Whitfield, W. G. F., & Hubbard, S. F. (1999). Horizontal transfer of *Wolbachia* between phylogenetically distant insect species by a naturally occurring mechanism. *Current Biology*, *9*(6), 313–316. https://doi.org/10.1016/S0960-9822(99)80139-0

Hiroki, M., Kato, Y., Kamito, T., & Miura, K. (2002). Feminization of genetic males by a symbiotic bacterium in a butterfly, *Eurema hecabe* (Lepidoptera: Pieridae). *Naturwissenschaften*, *89*(4), 167–170. https://doi.org/10.1007/S00114-002-0303-5

Hiroki, M., Tagami, Y., Miura, K., & Kato, Y. (2004). Multiple infection with *Wolbachia* inducing different reproductive manipulations in the butterfly *Eurema hecabe*. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1549), 1751–1755. https://doi.org/10.1098/RSPB.2004.2769

Huigens, M. E., De Almeida, R. P., Boons, P. A. H., Luck, R. F., & Stouthamer, R. (2004). Natural interspecific and intraspecific horizontal transfer of parthenogenesis–inducing *Wolbachia* in *Trichogramma* wasps. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1538), 509–515. https://doi.org/10.1098/rspb.2003.2640

Hurst, G. D. D., & Majerus, M. E. N. (1993). Why do maternally inherited microorganisms kill males? *Heredity*, *71*(1), 81–95. https://doi.org/10.1038/hdy.1993.110

Jiggins, C. D., Naisbit, R. E., Coe, R. L., & Mallet, J. (2001). Reproductive isolation caused by colour pattern mimicry. *Nature 2001 411:6835*, *411*(6835), 302–305. https://doi.org/10.1038/35077075

Kageyama, D., & Traut, W. (2004). Opposite sex-specific effects of *Wolbachia* and interference with the sex determination of its host *Ostrinia scapulalis*. *Proceedings of the Royal Society B: Biological Sciences*, *271*(1536), 251–258. https://doi.org/10.1098/rspb.2003.2604

Katsuma, S., Hirota, K., Matsuda-Imai, N., Fukui, T., Muro, T., Nishino, K., … & Kiuchi, T. (2022). A *Wolbachia* factor for male killing in lepidopteran insects. *Nature Communications 2022 13:1*, *13*(1), 6764-. https://doi.org/10.1038/s41467-022-34488-y

Kaur, R., Shropshire, J. D., Cross, K. L., Leigh, B., Mansueto, A. J., Stewart, V., … & 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

Kawahara, A. Y., Storer, C., Carvalho, A. P. S., Plotkin, D. M., Condamine, F. L., Braga, M. P., … & Lohman, D. J. (2023). A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nature Ecology & Evolution 2023 7:6*, *7*(6), 903–913. https://doi.org/10.1038/s41559-023-02041-9

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Le Clec'h, W., Chevalier, F. D., Genty, L., Bertaux, J., Bouchon, D., & Sicard, M. (2013). Cannibalism and predation as paths for horizontal passage of *Wolbachia* between terrestrial isopods. *PLOS ONE*, *8*(4), e60232. https://doi.org/10.1371/journal.pone.0060232

Lefoulon, E., Bain, O., Makepeace, B. L., D'Haese, C., Uni, S., Martin, C., & Gavotte, L. (2016). Breakdown of coevolution between symbiotic bacteria *Wolbachia* and their filarial hosts. *PeerJ*, *2016*(3). https://doi.org/10.7717/peerj.1840

Lefoulon, E., Clark, T., Guerrero, R., Cañizales, I., Cardenas-Callirgos, J. M., Junker, K., … & Slatko, B. E. (2020). Diminutive, degraded but dissimilar: *Wolbachia* genomes from filarial nematodes do not conform to a single paradigm. *Microbial Genomics*, *6*(12), 1–21. https://doi.org/10.1099/mgen.0.000487

Le Page, D. P., Metcalf, J. A., Bordenstein, S. R., On, J., Perlmutter, J. I., Shropshire, J. D., Layton, E. M., … & Bordenstein, S. R. (2017). Prophage WO genes recapitulate and enhance *Wolbachia*-induced cytoplasmic incompatibility. *Nature 2017 543:7644*, *543*(7644), 243–247. https://doi.org/10.1038/nature21391

Li, W., Cong, Q., Shen, J., Zhang, J., Hallwachs, W., Janzen, D. H., & Grishin, N. V. (2019). Genomes of skipper butterflies reveal extensive convergence of wing patterns. *Proceedings of the National Academy of Sciences*, *116*(13), 6232–6237. https://doi.org/10.1073/pnas.1821304116

Li, Y., Liu, Z., Liu, C., Shi, Z., Pang, L., Chen, C., … & Shen, X. X. (2022). HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell*, *185*(16), 2975-2987. https://doi.org/10.1016/j.cell.2022.06.014

Linke, D., Elias, M., Klečková, I., Mappes, J., & Matos-Maraví, P. (2022). Shape of evasive prey can be an important cue that triggers learning in avian predators. *Frontiers in Ecology and Evolution*, *10*, 910695. https://doi.org/10.3389/fevo.2022.910695

Liu, G., Chang, Z., Chen, L., He, J., Dong, Z., Yang, J., … & Li, X. (2020). Genome size variation in butterflies (Insecta, Lepidotera, Papilionoidea): a thorough phylogenetic comparison. *Systematic Entomology*, *45*(3), 571–582. https://doi.org/10.1111/syen.12417

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., … & Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*, *17*(12), 2815. https://doi.org/10.1038/s41596-022-00738-y

Ma, W. J., & Schwander, T. (2017). Patterns and mechanisms in instances of endosymbiont-induced parthenogenesis. *Journal of Evolutionary Biology*, *30*(5), 868–888. https://doi.org/10.1111/jeb.13069

Mahmood, S., Nováková, E., Martinů, J., Sychra, O., & Hypša, V. (2023). Supergroup F *Wolbachia* with extremely reduced genome: transition to obligate insect symbionts. *Microbiome*, *11*(22), 1–16. https://doi.org/10.1186/s40168-023-01462-9

Matos-Maraví, P., Linke, D., Ribeiro, P., Lamas, G., Freitas, A. V. L., Núñez, R., ... & Antonelli, A. (2025). Environmental variation promotes convergent evolution and rapid diversification of wing shape and color in skipper butterflies. *BioRxiv*, 2025.12.27.696669. https://doi.org/10.64898/2025.12.27.696669

Narita, S., Kageyama, D., Nomura, M., & Fukatsu, T. (2007). Unexpected mechanism of symbiont-induced reversal of insect sex: Feminizing *Wolbachia* continuously acts on the butterfly *Eurema hecabe* during larval development. *Applied and Environmental Microbiology*, *73*(13), 4332–4341. https://doi.org/10.1128

Narita, S., Kageyama, D., Hiroki, M., Sanpei, T., Hashimoto, S., Kamitoh, T., & Kato, Y. (2011). *Wolbachia*-induced feminisation newly found in *Eurema hecabe*, a sibling species of *Eurema mandarina* (Lepidoptera: Pieridae). *Ecological Entomology*, *36*(3), 309–317. https://doi.org/10.1111/j.1365-2311.2011.01274.x

Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M., & Fukatsu, T. (2008). *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome Research*, *18*(2), 272–280. https://doi.org/10.1101/gr.7144908

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ratnasingham, S., Wei, C., Chan, D., Agda, J., Agda, J., Ballesteros-Mejia, L., … & Hebert, P. D. N. (2024). *Chapter 26 BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data*. https://doi.org/10.1007/978-1-0716-3581-0_26

Raychoudhury, R., Baldo, L., Oliveira, D. C. S. G., & Werren, J. H. (2009). Modes of acquisition of *Wolbachia*: horizontal transfer, hybrid introgression, and codivergence in the *Nasonia* species complex. *Evolution*, *63*(1), 165–183. https://doi.org/10.1111/j.1558-5646.2008.00533.x

Ribeiro, P., Butenko, A., Linke, D., Ghanavi, H. R., Meier, J. I., Wahlberg, N., & Matos-Maraví, P. (2025). Pervasive horizontal transmission of *Wolbachia* in natural populations of closely related and widespread tropical skipper butterflies. *BMC Microbiology*, *25*(1), 5. https://doi.org/10.1186/S12866-024-03719-1

Russell, J. A., Goldman-Huertas, B., Moreau, C. S., Baldo, L., Stahlhut, J. K., Werren, J. H., & Pierce, N. E. (2009). Specialization and geographic isolation among *Wolbachia* symbionts from ants and lycaenid butterflies. *Evolution; International Journal of Organic Evolution*, *63*(3), 624–640. https://doi.org/10.1111/j.1558-5646.2008.00579.x

Scholz, M., Albanese, D., Tuohy, K., Donati, C., Segata, N., & Rota-Stabelli, O. (2020). Large scale genome reconstructions illuminate *Wolbachia* evolution. *Nature Communications*, *11*(1), 5235. https://doi.org/10.1038/s41467-020-19016-0

Shastry, V., Bell, K., Buerkle, C. A., Fordyce, J., Forister, M., Gompert, Z., … & Nice, C. (2022). A continental-scale survey of *Wolbachia* infections in blue butterflies reveals evidence of interspecific transfer and invasion dynamics. *G3 Genes|Genomes|Genetics, 12*(10), 1-17. https://doi.org/10.1093/g3journal/jkac213

Shropshire, J. D., & Bordenstein, S. R. (2016). Speciation by symbiosis: The microbiome and behavior. *mBio 7*(2). https://doi.org/10.1128/mBio.01785-15

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simões, P. M., Mialdea, G., Reiss, D., Sagot, M. F., & Charlat, S. (2011). Wolbachia detection: an assessment of standard PCR protocols. *Molecular Ecology Resources*, *11*(3), 567–572. https://doi.org/10.1111/J.1755-0998.2010.02955.X

Stouthamer, R., & Werren, J. H. (1993). Microbes associated with parthenogenesis in wasps of the genus *Trichogramm*a. *Journal of Invertebrate Pathology*, *61*(1), 6–9. https://doi.org/10.1006/jipa.1993.1002

Toussaint, E. F. A., Breinholt, J. W., Earl, C., Warren, A. D., Brower, A. V. Z., Yago, M., … & Kawahara, A. Y. (2018). Anchored phylogenomics illuminates the skipper butterfly tree of life. *BMC Evolutionary Biology 2018 18:1*, *18*(1), 101-. https://doi.org/10.1186/S12862-018-1216-Z

Turelli, M., Cooper, B. S., Richardson, K. M., Ginsberg, P. S., Peckenpaugh, B., Antelope, C. X., … & Hoffmann, A. A. (2018). Rapid Global Spread of *w*Ri-like *Wolbachia* across Multiple Drosophila. *Current Biology*, *28*(6), 963-971.e8. https://doi.org/10.1016/j.cub.2018.02.015

Tvedte, E. S., Gasser, M., Zhao, X., Tallon, L. J., Sadzewicz, L., Bromley, R. E., … & Dunning Hotopp, J. C. (2022). Accumulation of endosymbiont genomes in an insect autosome followed by endosymbiont replacement. *Current Biology*, *32*(12), 2786–2795. https://doi.org/10.1016/j.cub.2022.05.024

Twort, V. G., Blande, D., & Duplouy, A. (2022). One's trash is someone else's treasure: sequence read archives from Lepidoptera genomes provide material for genome reconstruction of their endosymbionts. *BMC Microbiology 2022 22:1*, *22*(1), 209-. https://doi.org/10.1186/S12866-022-02602-1

Valerio, F., Twort, V. G., & Duplouy, A. (2023). A Worked Example of Screening Genomic Material for the Presence of *Wolbachia* Infection. In A. M. Fallon (Ed.), *Wolbachia. Methods in Molecular Biology* (pp. 275–299). Humana. https://doi.org/10.1007/978-1-0716-3553-7_17

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

Vancaester, E., Oldrieve, G. R., Reid, A., Koutsovoulos, G., Laetsch, D. R., Makepeace, B. L., … & Blaxter, M. (2025). Ghosts of symbionts past: the hidden history of the dynamic association between filarial nematodes and their *Wolbachia* endosymbionts. *G3: Genes, Genomes, Genetics*, *15*(12), 1–14. https://doi.org/10.1093/g3journal/jkaf226

Warren, A. D., Ogawa, J. R., & Brower, A. V. Z. (2008). Phylogenetic relationships of subfamilies and circumscription of tribes in the family Hesperiidae (Lepidoptera: Hesperioidea). *Cladistics*, *24*(5), 642–676. https://doi.org/10.1111/j.1096-0031.2008.00218.x

Warren, A. D., Ogawa, J. R., & Brower, A. V. Z. (2009). Revised classification of the family Hesperiidae (Lepidoptera: Hesperioidea) based on combined molecular and morphological data. *Systematic Entomology*, *34*(3), 467–523. https://doi.org/10.1111/j.1365-3113.2008.00463.x

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). https://doi.org/10.1098/rspb.2015.0249

Werren, J. H., Baldo, L., & Clark, M. E. (2008). *Wolbachia*: master manipulators of invertebrate biology. *Nature Reviews Microbiology 2008 6:10*, *6*(10), 741–751. https://doi.org/10.1038/nrmicro1969

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1891-0

Wright, C. J., Stevens, L., Mackintosh, A., Lawniczak, M., & Blaxter, M. (2024). Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nature Ecology and Evolution*, *8*(4), 777–790. https://doi.org/10.1038/s41559-024-02329-4

Wright, C. J., Wahlberg, N., Vila, R., Mutanen, M., Matos-Maraví, P., Lucek, K., … & Zucco, G. (2025). Project Psyche: reference genomes for all Lepidoptera in Europe. *Trends in Ecology & Evolution*, *40*(12), 1234–1250. https://doi.org/10.1016/j.tree.2025.10.007

Wu, M., Sun, L. V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J. C., … & Eisen, J. A. (2004). Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements. *PLOS Biology*, *2*(3), e69. https://doi.org/10.1371/JOURNAL.PBIO.0020069

Yang, F., Ilinsky, Y. Y., Wei, X., Duan, F., Zhu, J., & Jiang, W. (2026). Evolutionary ecology of *Wolbachia*-lepidopteran insects relationships: Multifaceted effects, transmission patterns and practical applications. *Ecological Entomology*. https://doi.org/10.1111/een.70082

Zhang, J., Cong, Q., & Grishin, N. V. (2023). Thirteen new species of butterflies (Lepidoptera: Hesperiidae) from Texas. *Insecta Mundi*, 0969, 1-58. https://pmc.ncbi.nlm.nih.gov/articles/PMC9880598/

9 Acknowledgements

We acknowledge the computational resources provided by CESNET LM2015042 and the CERIT Scientific Cloud LM2015085 under the programme “Projects of Large Research, Development, and Innovation Infrastructures.” This work was supported by the Czech Science Foundation (GAČR grant No. 22-35084J awarded to P. Matos-Maraví).

10 Supplementary Material

M1: GitHub Repository: https://github.com/AndreaKochova/BachelorThesis_USB/tree/main. The repository contains an explanatory README.md file explaining all its contents.

Table S1: Species and collection locality metadata for all 77 Eudaminae samples.

Sample	Order	Family	Subfamily	Tribe	Subtribe	Genus	Species	Subspecies	Country	Place
PMLBSc001	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Spicauda	simplicius		Panama	Argos
PMLBSc002	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Autochton	itylus		Peru	Pozuzo
PMLBSc003	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	dorantes	Peru	Pozuzo
PMLBSc004	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	dorantes	Peru	Pozuzo
PMLBSc005	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Autochton	itylus		Peru	Pozuzo
PMLBSc006	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	longipennis		Peru	Tarapoto
PMLBSc007	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Spicauda	teleus		Peru	Tarapoto
PMLBSc008	Lepidoptera	Hesperiidae	Eudaminae	Entheini	NA	Entheus	priassus nr		Peru	Tarapoto
PMLBSc009	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	esmeraldus		Peru	Tarapoto
PMLBSc010	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	longipennis		Peru	Tarapoto
PMLBSc011	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Spicauda	simplicius		Peru	Tarapoto
PMLBSc012	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	velinus		Peru	Tarapoto
PMLBSc013	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	dorantes	Peru	Tarapoto
PMLBSc014	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	longipennis		Peru	Tarapoto
PMLBSc015	Lepidoptera	Hesperiidae	Eudaminae	Oileidini	Typhedanina	Cogia	undulatus		Peru	Tarapoto
PMLBSc016	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Telemiadina	Ectomis	pervivax nr		Peru	Tarapoto
PMLBSc017	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Telegonus	fulgerator		Peru	Tarapoto
PMLBSc018	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	longipennis		Peru	Tarapoto
PMLBSc019	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	longipennis		Peru	Tarapoto
PMLBSc020	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Nascus	solon		Peru	Tarapoto
PMLBSc021	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Epargyreus	socus nr		Peru	M. de Dios
PMLBSc022	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Loboclina	Aguna	spicata		Peru	M. de Dios
PMLBSc023	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Telemiadina	Ectomis	auginus		Peru	M. de Dios
PMLBSc024	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Narcosius	nazareus		Peru	M. de Dios
PMLBSc025	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Nascus	phocus		Peru	M. de Dios
PMLBSc026	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Dyscophellus	erythras nr		Peru	M. de Dios
PMLBSc027	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Bungalotis	corentinus		Peru	M. de Dios
PMLBSc028	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Bungalotis	midas nr		Peru	M. de Dios
PMLBSc029	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Porphyrogenes	omphale		Peru	M. de Dios
PMLBSc030	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	marmorosa		Cuba	Mayabeque
PMLBSc031	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	
PMLBSc032	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	santiago	Cuba	Habana
PMLBSc033	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	dorantes	Brazil	Sao Paulo
PMLBSc034	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Spicauda	simplicius		Brazil	Bahia
PMLBSc035	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	churchi		Jamaica	
PMLBSc036	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Telegonus	talus		Cuba	
PMLBSc037	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	
PMLBSc038	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Telegonus	habana	habana	Cuba	Santiago de Cuba
PMLBSc039	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	marmorosa		Cuba	Holguín
PMLBSc040	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Autochton	itylus		Peru	Pozuzo
PMLBSc041	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	pronta nr		Peru	Tarapoto
PMLBSc042	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Peru	Tarapoto
PMLBSc043	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Peru	Tarapoto
PMLBSc044	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Loboclina	Aguna	aguna sp.		Peru	Tarapoto
PMLBSc045	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Nascus	solon		Peru	M. de Dios
PMLBSc046	Lepidoptera	Hesperiidae	Eudaminae	Phocidini	NA	Euriphellus	euribates		Peru	M. de Dios
PMLBSc047	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	santiago	Cuba	
PMLBSc048	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	santiago	Cuba	
PMLBSc049	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	
PMLBSc050	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	Cienfuegos
PMLBSc051	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Spicauda	simplicius		Brazil	Paraíba
PMLBSc052	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Loboclina	Aguna	asander	haitiensis	Cuba	Habana
PMLBSc053	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	Habana
PMLBSc054	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	marmorosa		Cuba	Artemisa
PMLBSc055	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Telegonus	habana	habana	Cuba	Holguín
PMLBSc056	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Telegonus	xagua	xagua	Cuba	Guantánamo
PMLBSc057	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Telegonus	xagua	xagua	Cuba	Guantánamo
PMLBSc058	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	
PMLBSc059	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus	domingo	Cuba	
PMLBSc060	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Cuba	Holguín
PMLBSc061	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Cecropterus	dorantes	santiago	Cuba	Habana
PMLBSc062	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	marmorosa		Cuba	
PMLBSc063	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa
PMLBSc064	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa
PMLBSc065	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc066	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc067	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa
PMLBSc068	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc069	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa

Table S1: Species and collection locality metadata for all 77 Eudaminae samples.

Sample	Order	Family	Subfamily	Tribe	Subtribe	Genus	Species	Subspecies	Country	Place
PMLBSc070	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa
PMLBSc071	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Spathilepia	clonius		Panama	Gamboa
PMLBSc072	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc073	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc074	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc075	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Chioides	catillus		Panama	Gamboa
PMLBSc076	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa
PMLBSc077	Lepidoptera	Hesperiidae	Eudaminae	Eudamini	Eudamina	Urbanus	proteus		Panama	Gamboa

Table S2: Reference genomes used for the mapping procedure based on referenced used in Twort et al., (2022).

Strain	Accession	Host	Host classification	Supergroup	Assembly level	Contig L50	Chromosome RefSeq
*w*Pip	GCF_000073005.1	*Culex quinquefasciatus*	Diptera	B	Chromosome	1	NC_010981.1
*w*Mel	GCF_000008025.1	*Drosophila melanogaster*	Diptera	A	Complete Genome	1	NC_002978.6
*w*No	GCF_000376585.1	*Drosophila simulans*	Diptera	B	Complete Genome	1	NC_021084.1
*w*Ri	GCF_000022285.1	*Drosophila simulans Riverside*	Diptera	A	Complete Genome	1	NC_012416.1
*w*Vol	GCF_000530755.1	*Onchocerca volvulus*	Nematoda	C	Complete Genome	1	NZ_HG810405.1
*w*Cle	GCF_000829315.1	*Cimex lectularius*	Hemiptera	F	Complete Genome	1	NZ_AP013028.1
*w*Tpre	GCF_001439985.1	*Trichogramma pretiosum*	Hymenoptera	B	Chromosome	2	NZ_CM003641.1
*w*Oo	GCF_000306885.1	*Onchocera ochengi*	Nematoda	C	Complete Genome	1	NC_018267.1
*w*Ha	GCF_000376605.1	*Drosophila simulans*	Diptera	A	Complete Genome	1	NC_021089.1
*w*Bm	GCF_000008385.1	*Brugia malayi*	Nematoda	D	Complete Genome	1	NC_006833.1
*w*Alb-B	GCF_004171285.1	*Aedes albopictus*	Diptera	B	Complete Genome	1	NZ_CP031221.1
*w*Irr	GCF_009732755.1	*Haematobia irritans*	Diptera	A	Complete Genome	1	NZ_CP037426.1

Table S3*:* Read-mapping metrics for all 77 specimens, ranked by ascending $\log_{10}$ RPM. Columns give maximum breadth of coverage across the 12 *Wolbachia* references, whether it met the 70% threshold, *Wolbachia*-mapping and total reads, RPM, $\log_{10}$ RPM, and the $\log_{10}$ RPM difference from the sample below. *Largest interval in the ranking, marking the operational abundance boundary discussed in the text.

Sample	Max. breadth of coverage (%)	Breadth of coverage ≥ 70%	Wolbachia-mapping reads	Total raw reads	Max. RPM	$\log_{10}$ RPM	Δ $\log_{10}$
PMLBSc059	0.0774	No	123	133,014,386	0.92	-0.036	
PMLBSc009	0.0815	No	70	65,854,034	1.06	0.025	0.062
PMLBSc017	0.1290	No	68	58,811,898	1.16	0.064	0.039
PMLBSc037	0.0074	No	112	80,192,102	1.40	0.146	0.082
PMLBSc049	0.1863	No	248	171,372,740	1.45	0.161	0.015
PMLBSc058	0.0466	No	206	130,020,734	1.58	0.199	0.037
PMLBSc052	0.1446	No	274	165,898,470	1.65	0.217	0.019
PMLBSc012	0.2409	No	94	52,368,134	1.79	0.253	0.035
PMLBSc053	0.0272	No	250	137,227,690	1.82	0.260	0.007
PMLBSc060	0.1446	No	238	120,092,554	1.98	0.297	0.037
PMLBSc030	0.0602	No	155	78,082,280	1.99	0.299	0.002
PMLBSc050	0.1176	No	257	123,024,260	2.09	0.320	0.021
PMLBSc062	0.0295	No	292	131,478,782	2.22	0.346	0.026
PMLBSc044	0.2931	No	313	136,427,994	2.29	0.360	0.013
PMLBSc007	0.1146	No	152	62,457,486	2.43	0.386	0.026
PMLBSc054	0.0560	No	367	138,140,708	2.66	0.425	0.039
PMLBSc051	0.1313	No	350	130,229,386	2.69	0.430	0.005
PMLBSc034	0.0823	No	198	72,154,496	2.74	0.438	0.008
PMLBSc014	0.0809	No	145	49,261,644	2.94	0.468	0.031
PMLBSc021	0.0363	No	211	69,997,494	3.01	0.479	0.010
PMLBSc067	0.5811	No	132	41,024,934	3.22	0.508	0.029
PMLBSc022	0.0718	No	192	56,827,908	3.38	0.529	0.021
PMLBSc039	0.0294	No	170	49,886,214	3.41	0.533	0.004
PMLBSc063	0.1192	No	136	33,901,540	4.01	0.603	0.070
PMLBSc035	0.1010	No	217	52,738,364	4.11	0.614	0.011
PMLBSc066	0.9766	No	144	34,768,724	4.14	0.617	0.003

Sample	Max. breadth of coverage (%)	Breadth of coverage ≥ 70%	Wolbachia-mapping reads	Total raw reads	Max. RPM	$\log_{10}$ RPM	$\Delta \log_{10}$
PMLBSc064	0.4005	No	158	37,151,884	4.25	0.628	0.011
PMLBSc006	0.0489	No	268	62,527,514	4.29	0.632	0.004
PMLBSc018	0.0456	No	311	65,968,058	4.71	0.673	0.041
PMLBSc057	0.0235	No	266	53,616,450	4.96	0.695	0.022
PMLBSc015	0.0709	No	280	53,503,218	5.23	0.719	0.023
PMLBSc055	0.0626	No	253	47,894,330	5.28	0.723	0.004
PMLBSc010	0.1686	No	268	49,685,684	5.39	0.732	0.009
PMLBSc038	0.0523	No	381	58,012,528	6.57	0.818	0.086
PMLBSc019	0.2548	No	446	63,005,162	7.08	0.850	0.032
PMLBSc070	0.1938	No	719	92,047,590	7.81	0.893	0.043
PMLBSc016	0.0820	No	459	57,520,352	7.98	0.902	0.009
PMLBSc027	0.0590	No	705	74,128,556	9.51	0.978	0.076
PMLBSc069	0.5918	No	1,408	145,122,928	9.70	0.987	0.009
PMLBSc023	5.8355	No	716	66,229,834	10.81	1.034	0.047
PMLBSc008	0.0183	No	732	67,129,602	10.90	1.037	0.004
PMLBSc026	0.1021	No	618	46,520,488	13.28	1.123	0.086
PMLBSc028	0.1814	No	781	49,085,722	15.91	1.202	0.078
PMLBSc056	0.2748	No	879	53,769,742	16.35	1.214	0.012
PMLBSc076	0.5294	No	3,032	178,566,806	16.98	1.230	0.016
PMLBSc029	0.0748	No	909	49,185,216	18.48	1.267	0.037
PMLBSc071	0.2084	No	2,756	116,382,448	23.68	1.374	0.108
PMLBSc077	1.5890	No	4,553	184,412,694	24.69	1.393	0.018
PMLBSc033	16.0236	No	2,471	63,972,186	38.63	1.587	0.194
PMLBSc024	0.3061	No	3,002	49,747,446	60.34	1.781	0.194
PMLBSc025	0.0475	No	3,226	47,653,264	67.70	1.831	0.050
PMLBSc046	0.3015	No	9,415	51,220,216	183.81	2.264	0.434*
PMLBSc065	31.3502	No	6,084	32,568,196	186.81	2.271	0.007
PMLBSc042	66.2510	No	20,863	111,451,990	187.19	2.272	0.001
PMLBSc032	58.5415	No	16,357	71,076,658	230.13	2.362	0.090

Sample	Max. breadth of coverage (%)	Breadth of coverage ≥ 70%	Wolbachia-mapping reads	Total raw reads	Max. RPM	$\log_{10}$ RPM	$\Delta \log_{10}$
PMLBSc041	77.7011	Yes	30,620	121,201,382	252.64	2.403	0.041
PMLBSc031	0.1431	No	15,332	59,458,212	257.86	2.411	0.009
PMLBSc043	73.8764	Yes	32,272	122,413,994	263.63	2.421	0.010
PMLBSc073	76.9949	Yes	74,076	234,447,822	315.96	2.500	0.079
PMLBSc047	77.0494	Yes	64,682	179,750,308	359.84	2.556	0.056
PMLBSc013	70.4134	Yes	27,201	58,345,776	466.20	2.669	0.112
PMLBSc074	75.2141	Yes	61,351	129,271,956	474.59	2.676	0.008
PMLBSc004	70.5796	Yes	26,737	53,645,348	498.40	2.698	0.021
PMLBSc072	13.2021	No	3,121	5,817,224	536.51	2.730	0.032
PMLBSc075	78.6387	Yes	93,474	127,090,904	735.49	2.867	0.137
PMLBSc068	75.9773	Yes	41,315	52,285,374	790.18	2.898	0.031
PMLBSc045	93.2664	Yes	148,961	121,059,546	1230.48	3.090	0.192
PMLBSc003	80.5972	Yes	98,553	73,907,502	1333.46	3.125	0.035
PMLBSc061	82.2965	Yes	199,299	128,614,534	1549.58	3.190	0.065
PMLBSc036	79.5608	Yes	114,052	68,952,498	1654.07	3.219	0.028
PMLBSc001	92.1785	Yes	90,797	49,277,328	1842.57	3.265	0.047
PMLBSc048	82.9002	Yes	356,479	154,605,870	2305.73	3.363	0.097
PMLBSc040	96.6946	Yes	381,252	120,880,352	3153.96	3.499	0.136
PMLBSc002	96.2165	Yes	298,889	63,301,992	4721.64	3.674	0.175
PMLBSc005	96.0565	Yes	316,117	63,986,652	4940.36	3.694	0.020
PMLBSc011	96.0228	Yes	439,075	74,703,074	5877.60	3.769	0.075
PMLBSc020	94.9610	Yes	506,605	60,641,386	8354.11	3.922	0.153

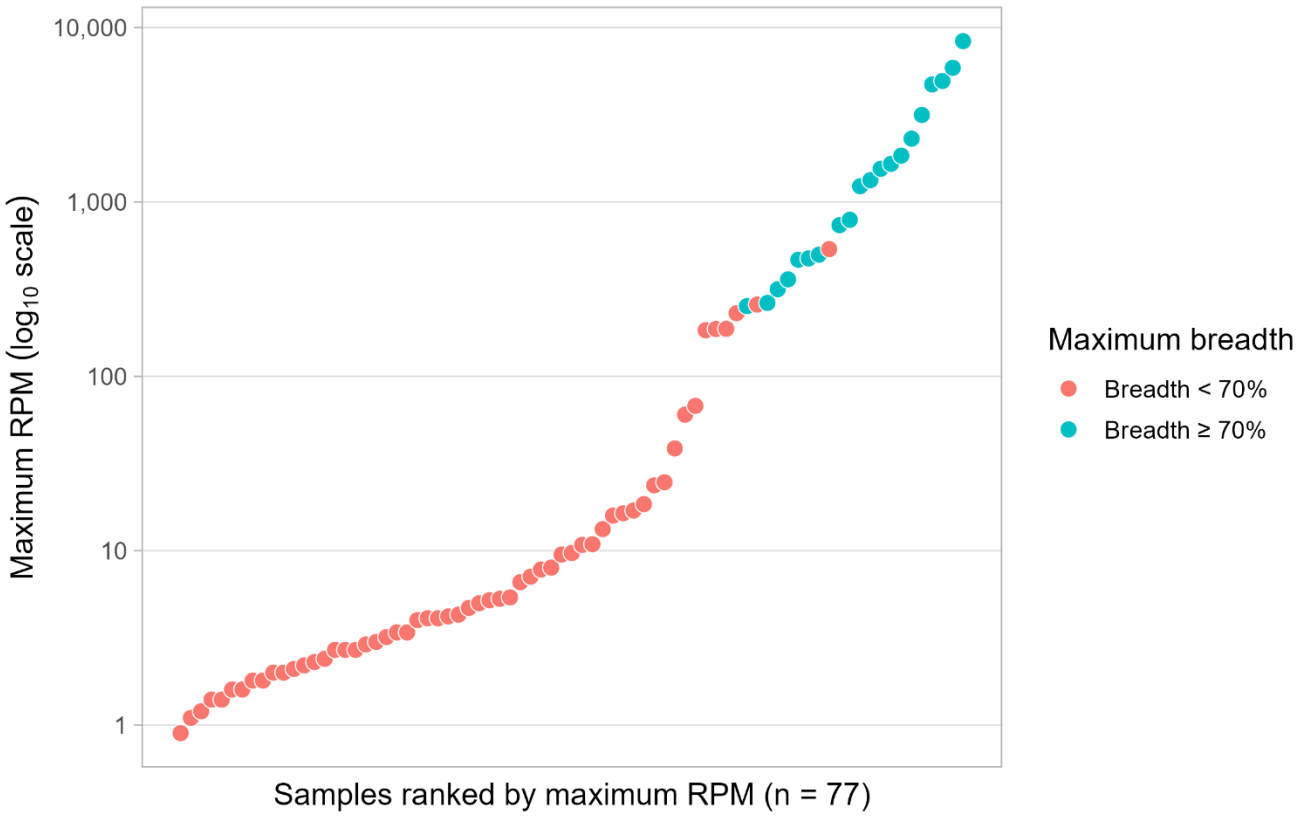

Figure S1: Distribution of maximum normalised Wolbachia-mapping read abundance across the 77 analysed samples. Samples are ranked by maximum reads per million total reads (RPM), with RPM displayed on a logarithmic scale. Coloured points indicate whether the maximum breadth of coverage was above or below the 70% operational threshold. The colours represent the primary breadth criterion rather than final infection status.