Supplementary Materials for: *Wolbachia* infections in recently speciated butterflies

Eric Toro-Delgado, Roger Vila, Konrad Lohse, Dominik R. Laetsch, Gerard Talavera

Supplementary methods

Sampling and data collection

Butterflies were collected between 2006-2019 in multiple European countries. All samples were either dried and stored in absolute ethanol at -20°C or directly flash frozen at -80°C in a dry shipper and stored at -80°C. Collection dates and localities for all samples are available in Tables S1A-S1B. Butterfly species were identified based on morphology and, in cases of cryptic taxa, using the barcoding region of the cytochrome-c oxidase subunit one gene (see Ebdon *et al*., 2021 for details). The species were selected by their phylogenetic proximity based on a complete phylogeny of European butterflies (Wiemers *et al*., 2020) and also on phylogenies of specific genera (Peña *et al*., 2015; Klečková *et al*., 2023); a total of 18 species pairs were selected, representing all butterfly families (expect Riodinidae, which is monotypic in Europe). These species are very heterogeneous, with a wide range of sizes (12-46mm of forewing length), hostplant families (e.g. Poaceae, Fabaceae, Rosaceae) and habitats (from lowland to montane specialists), and range in split times from 0.92 Mya (*Colias*) to 8.5 Mya (*Satyrus*). Therefore, they constitute a representative sample of recent butterfly speciation events, and a sizeable fraction of all European butterfly sister species pairs (total of ca. 70). In eight cases, specimens of a third species in the same genus were also obtained and sequenced, resulting in a resequence dataset of 44 species. These additional species were included in analyses that are not based on comparing the pairs.

To tackle the presented questions, two types of genomic data were generated. First of all, for a small set of reference specimens covering all studied butterfly genera, DNA was extracted from thoracic tissue and both PacBio Sequel libraries and Illumina paired-end libraries were generated at an expected coverage of 50X to generate genome assemblies of *Wolbachia*. We refer to these samples as the “genomic reference individuals”. Secondly, for a larger set of specimens (between four and thirteen) for all species of interest, DNA was extracted from thoracic tissue and Illumina paired-end libraries were generated at an expected coverage of 25X. We refer to these samples as the “WGS resequence individuals”. In total, 273 specimens were sequenced.

In addition, we also obtained *Wolbachia* reference genomes from NCBI RefSeq database (Sayers *et al*. (2022); downloaded on October 7th 2022) and from the Darwin Tree of Life (DToL) project (Vancaester and Blaxter (2023); downloaded on December 11th, 2022).

Analyses

Genome assembly

Illumina libraries were trimmed on both their 5’ and 3’ ends using FASTP (Chen *et al*., 2018), with a mean quality threshold of 20 and a window size of 4 bases. Afterwards they were mapped against the closest available reference genome to check coverage, and those with mean coverage lower than 5X were discarded.

For each host butterfly, the nuclear genome had been previously assembled from the corresponding read libraries. If the butterfly is infected, the genome assembly can already contain the *Wolbachia* genome as one of the contigs, but this is not always the case. Therefore, we used BlobTools (Laetsch and Blaxter, 2017a) to assess the taxonomic identities of the contigs obtained during assembly of the butterfly genomes. If *Wolbachia* contigs were found and BUSCO (Manni *et al*., 2021) completeness (based on the rickettsiales_odb10 dataset) was at least 90%, they were kept as reference *Wolbachia* genomes. If *Wolbachia* contigs were found but with a lower BUSCO completeness, PacBio and Illumina libraries were mapped back to the assembled butterfly genome using minimap2 (Li, 2018), and we extracted all reads from both libraries that did not map to any contig or that mapped to either *Wolbachia* contigs or to contigs of uncertain taxonomic assignment. Using these reads, a new long-read assembly with short-read polishing was carried out, using NextDenovo (Hu *et al*., 2023) and HAPO-G (Aury and Istace, 2021). If an assembly could not be obtained with this strategy, a long-read assembly with Flye and short-read polishing with HAPO-G was conducted, the minimap2 and BlobTools steps were repeated to partition the reads mapping to *Wolbachia* contigs only, and assembly with these reads was attempted. A summary of the generated assemblies can be found in Table S2.

Phylogenetic inference

The newly assembled *Wolbachia* genomes, together with those obtained from DToL, were annotated using Prokka (Seemann, 2014), using the proteomes of all RefSeq *Wolbachia* genomes to guide the annotation. The resulting proteomes, together with the RefSeq ones, were used to infer orthogroups using Orthofinder (Emms and Kelly, 2015) and KinFin (Laetsch and Blaxter, 2017b), with the exceptions of GCF_918315375.1 and GCF_918308635.1, which were discarded due to them being represented in less than 20% of the orthogroups. These two genomes also had very low BUSCO completeness (<40%). A total of 250 single-copy orthologs (SCOs) were recovered that were present at 95% of the genomes. These SCOs were then aligned using MAFFT (Katoh, 2002), trimmed using trimAl (Capella-Gutiérrez *et al*., 2009) and concatenated with FASconCAT-G (Kück and Longo, 2014) or SuperCRUNCH (Portik and Wiens, 2020) to build a supermatrix that was used as input to make a maximum-likelihood phylogeny of all *Wolbachia* genomes with IQTree (Minh *et al*., 2020). Substitution model parameters were estimated separately for each orthogroup, using the GTR+Γ substitution model. 1,000 replicates were run for both the ultra fast bootstrap with hill-climbing NNI and the SH-aLRT.

Since this phylogeny was consistent with previous results with regards to supergroup classification (Vancaester and Blaxter, 2023) and all our assembled genomes belonged to supergroup B, the same approach was repeated to build a phylogeny of only supergroup B *Wolbachia*, keeping three supergroup A taxa as outgroup. This phylogeny was based on 493 SCOs present in at least 95% of the strains.

In addition, we also built a butterfly phylogeny with the reference genome of each butterfly species. For species without a reference genome, we mapped the highest coverage Illumina library among the resequence individuals to the closest reference genome available using minimap2. Then, with the output from BUSCO for the reference genome, we used BEDTools (Quinlan and Hall, 2010) to select those regions corresponding to the BUSCO genes, extended 1000bp upstream and downstream to allow for some variation in the exact positioning of the BUSCOs in the resequencing individual. The BAM files were subsetted with SAMtools (Danecek *et al*., 2021) to keep only the reads mapped to the selected regions and with mapping quality > 20, and variants were called from these subsetted BAM files and filtered based on quality ≥ 20 and number of reads ≥ 8 using BCFtools (Danecek *et al*., 2021). Afterwards, a phylogeny was built with the same approach as for *Wolbachia*, but using the set of single-copy BUSCOs from the BUSCO analysis of each genome or consensus sequence. A total of 4,186 BUSCOs present in at least 95% of the samples were used.

Detection of *Wolbachia* presence in Illumina libraries

A 2-step competitive mapping approach was used to assess the occurrence of *Wolbachia* in each butterfly genus. First, the Illumina libraries were mapped with minimap2 against a reference consisting of the concatenation of all *Wolbachia* reference genomes from RefSeq and DToL, plus the ones assembled in this study. As before, the RefSeq genomes GCF_918315375.1 and GCF_918308635.1 were excluded from the reference. The resulting BAM files were processed with inStrain (Olm *et al*., 2021) to obtain the breadth of coverage (defined as the proportion of bases in the genome that are covered by reads) for each of the genomes present in the reference FASTA. Following inStrain's documentation, we did not set any mapping quality threshold, since the competitive mapping strategy implies that many reads will map well to multiple genomes and will be assigned artificially low mapping qualities. In addition, the competitive mapping implies the genomes "compete" to attract the reads and thus do not get as good mapping as if the reads were mapped to each of them alone. For this reason, to get results closer to that of mapping to each genome individually, the ten reference genomes with the highest breadth in each sample of a given genus were selected, and the Illumina reads were mapped against the concatenation of this reduced number of reference genomes and analysed again with inStrain.

inStrain provides an expected breadth value, which is the breadth that would be expected at the obtained coverage if the reads truly came from exactly the same genome they are mapping to. If a library has an observed breadth much lower than the expected one, it is unlikely to be infected with that genome. Based on the observed and expected breadths for the Illumina libraries of the genomic reference individuals, we established a threshold of having an observed breadth lower than the expected one by no more than 5% against at least one reference genome to consider a sample as infected (i.e. if expected breadth is 95%, then observed breadth should be >90%). However, this would cause some samples with an observed breadth slightly below the 5% threshold, and mapping to a strain already found in other specimens of the same species, to be considered uninfected. This occurred in samples with low coverages, for which the difference in expected and observed breadth due to the competitive mapping approach is likely to be higher given the nonlinear relationship between coverage and breadth (Port *et al*., 1995). Therefore, if a sample was mapping to a strain that is already considered to occur in that species based on other samples, we allowed observed breadth to be up to 10% below the expected breadth. We did not focus on coverage depth because high values can be achieved due to repetitive elements or other components of a genome receiving many reads while the rest receives no or few reads, and thus depth can be misleading.

In addition, if a sample had an observed breadth in the best-mapping genome of less than 10%, it was considered as uninfected in all cases. This left some samples with intermediate breadth values but that were below the expected breadth by more than 10%. Due to these samples, there was no clear gap in the breadth distribution, preventing establishing a clear cutoff to separate infected and uninfected samples (Figure S1). Also, large sized nuclear *Wolbachia* insertions (NUWTs), sometimes consisting of entire genomes, have been detected in insects (Hotopp *et al*., 2007; Kondo *et al*., 2002), but how often they occur and whether they are very conserved is unclear. Therefore, these specimens may either have NUWTs or be infected with strains that are not well represented by the available genomes. For this reason, we considered two versions of the infection status data: an "infection-conservative" dataset, in which samples with uncertainty were considered as uninfected (which would imply they have NUWTs), and a "NUWT-conservative" dataset, in which samples with uncertainty have been considered as infected. Subsequent analyses involving the WGS resequencing individuals were performed over both versions of the data. Since results were consistent, we only report those of the infection-conservative version.

To determine the number and identity of *Wolbachia* strains in an infected specimen, we examined the pattern of competitive mapping across samples in each genus. Given multiple genomes of sufficiently high similarity, if one of them obtains a high breadth, the rest will do so too. This allowed the detection

of mapping to sets of genomes, for example, a sample may map well to genomes A and B (but not to C and D), another to C and D (but not to A and B), and a third one, to all of them; in such case the first sample would be infected by a strain, the second one by a different strain, and the last one by both strains. A clustering analysis with dRep (Olm *et al*., 2017) with an average nucleotide identity (ANI) threshold of 99% recovered clusters of genomes consistent with the mapping behavior and also with the clades in the *Wolbachia* phylogeny (Figure S2). Species mapping to genomes belonging to the same clusters were considered as being infected by the same strains, which were named after the genome that dRep assigned as representative of that cluster based on several quality metrics, such as N50.

The software versions and parameters used for each program are available in Table S3. All analyses were run using GNU Parallel (Tange, 2022).

Statistical analysis of *Wolbachia* presence and diversity in sister pairs of European butterflies

To investigate the extent to which *Wolbachia* is maintained via vertical transmission and thus the potential for co-cladogenesis, we built a correlation between mean host divergences and mean *Wolbachia* divergences. For *Wolbachia*, we used average nucleotide identity (ANI) as implemented in FastANI (Jain *et al*., 2018) as a measure of strain similarity. To obtain mean ANI between two species in a pair, we added the ANI values of all pairwise comparisons of all *Wolbachia* occurrences in one species against all occurrences in the other species, and divided by the number of comparisons. Pairwise comparisons were computed both ways, since ANI is asymmetric. For the hosts, we used mean gene divergences among the sister species (d_{xy}), obtained from Ebdon *et al*. (2021).

In addition, we also calculated if the mean ANI between sister pairs is significantly higher than expected by chance. To do this, we drew an empirical p-value from a null distribution of mean ANI values generated as follows: for each replicate, randomly assign the species in pairs without repeating any, obtain the mean ANI of each pair in the same way as for the true pairs, and compute their mean. With all these replicates (10,000, including the observed value), generate a distribution of mean ANI values in random pairs, and compute how many values are greater than or equal to the observed value.

To assess potential evidence of a role of *Wolbachia* in host speciation, we calculated if the number of strains shared by sister taxa is higher or lower than expected by chance by generating a null distribution of mean numbers of shared strains in randomized species pairs as explained above, but using a matrix of strain-sharing instead of ANI, and counting the number of strains occurring in both species in each randomized pair. A total of 10,000 replicates was computed.

To assess potential relationships between ecological traits of butterflies and *Wolbachia* infections, we fitted a Poisson generalised linear mixed model with log link function in a Bayesian framework using the MCMCglmm R package (Hadfield, 2010). We modelled the number of strains found in a given species as a function of wing index (a standardised measure of fore wing length, which is a proxy for dispersal ability), voltinism (the number of annual generations), mean number of flight months, and all possible interactions. As voltinism can vary across a species range, we selected the minimum value, which could act as a bottleneck to the spread of *Wolbachia*. Species traits were obtained from Middleton-Welling *et al*. (2020). Since the observations of the dependent variable are phylogenetically correlated, we accounted for this by including the inverse branch length matrix obtained from the previously inferred phylogeny. In addition, since this model does not depend on species being in pairs, we also included additional data for the other species, resulting in 44 species in total (i.e. N=44).

To assess how the relationship between species pairs affects their similarity in *Wolbachia* status, we built Bayesian logistic regression with MCMCglmm, using strain sharing (i.e. whether a pair has or not

some strains in common) as the response variable. Only the nine (in the infection-conservative data) or twelve (in the NUWT-conservative data) pairs in which at least one of the species is infected were considered (i.e. N=9 and N=12). As predictor variables we included the time since the split of the species, the degree of sympatry (i.e. the extent to which species ranges overlap) and their interaction. We fitted two models per data version, one with split time in generations and another with split time in million years ago (Mya). Values for these variables were obtained from Ebdon *et al*. (2021). Since the model considers pairs, no phylogenetic covariance matrix was included.

For all statistical analyses, infection statuses were derived based on our data, without considering infection reports from the literature. We took this decision because many studies are based on MLST and are thus not directly comparable to our analyses of infection status and strain identity. Analyses were performed under R version 4.3.0. The scripts for the statistical analyses are available at: https://github.com/etd530/MSc_thesis_SoM

Supplementary references

Aury, J. M. and Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genomics and Bioinformatics.

Capella-Gutiérrez, S. *et al*. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25, 1972–1973.

Chen, S. *et al*. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34, i884–i890. Danecek, P. *et al*. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10, 1–4.

Ebdon, S. *et al*. (2021). The Pleistocene species pump past its prime: Evidence from European butterfly sister species. Molecular Ecology, 30, 3575–3589.

Emms, D. M. and Kelly, S. (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology, 16, 157.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. Journal of Statistical Software, 33(2), 1–22.

Hotopp, J. C. D. *et al*. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science, 317, 1753–1756.

Hu, J. *et al*. (2023). An efficient error correction and accurate assembly tool for noisy long reads. bioRxiv , page 2023.03.09.531669.

Jain, C. *et al*. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications 2018 9:1 , 9, 1–8.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30, 3059–3066.

Klečková, I. *et al*. (2023). Climatic niche conservatism and ecological diversification in the Holarctic cold-dwelling butterfly genus *Erebia*. Insect Systematics and Diversity, 7.

Kondo, N. *et al*. (2002). Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proceedings of the National Academy of Sciences, 99, 14280–14285.

Kück, P. and Longo, G. C. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Frontiers in Zoology, 11, 81.

Laetsch, D. R. and Blaxter, M. L. (2017a). BlobTools: Interrogation of genome assemblies. F1000Research 2017 6:1287 , 6, 1287.

Laetsch, D. R. and Blaxter, M. L. (2017b). KinFin: Software for taxon-aware analysis of clustered protein sequences. G3 Genes—Genomes—Genetics, 7, 3349–3357.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094–3100.

Manni, M. *et al*. (2021). BUSCO: Assessing genomic data quality and beyond. Current Protocols, 1, e323.

Middleton-Welling, J. *et al*. (2020). A new comprehensive trait database of European and Maghreb butterflies, Papilionoidea. Scientific Data, 7, 351.

Minh, B. Q. *et al*. (2020). Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution, 37, 1530–1534.

Olm, M. R. *et al*. (2017). drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. The ISME Journal 2017 11:12, 11, 2864–2868.

Olm, M. R. *et al*. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nature Biotechnology 2021 39:6 , 39, 727–736.

Peña, C. *et al*. (2015). Adaptive radiations in butterflies: evolutionary history of the genus *Erebia* (nymphalidae: Satyrinae). Biological Journal of the Linnean Society, 116, 449–467.

Port, E. *et al*. (1995). Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics, 26, 84–100.

Portik, D. M. and Wiens, J. J. (2020). SuperCRUNCH: A bioinformatics toolkit for creating and manipulating supermatrices and other large phylogenetic datasets. Methods in Ecology and Evolution, 11, 763–772.

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

Sayers, E. W. *et al*. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Research, 50, D20.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30, 2068–2069.

Tange, O. (2022). GNU Parallel 20220722 (’Roe vs Wade’).

Vancaester, E. and Blaxter, M. (2023). Phylogenomic analysis of *Wolbachia* genomes from the Darwin Tree of Life biodiversity genomics project. PLOS Biology, 21, e3001972.

Wiemers, M. *et al*. (2020). A complete time-calibrated multi-gene phylogeny of the European butterflies. ZooKeys, 938, 97–124.

Supplementary figures

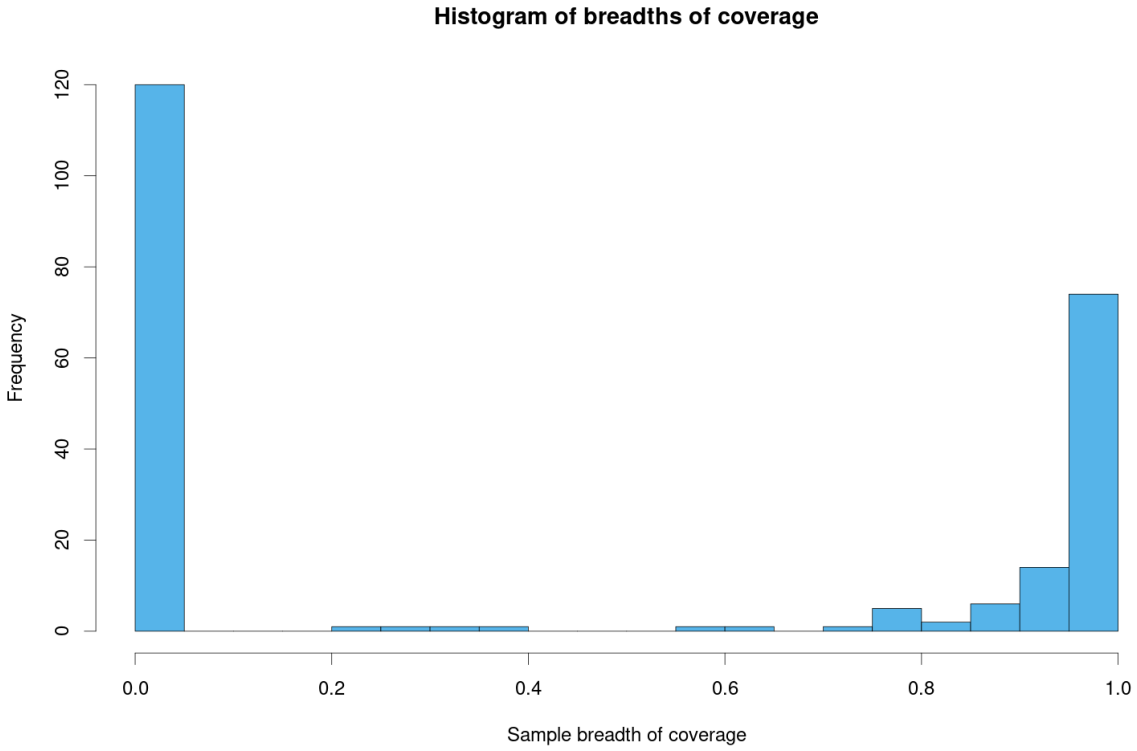


Figure S1. Histogram of the highest breadths of coverage obtained when mapping the Illumina paired-end libraries against the concatenation of *Wolbachia* reference genomes for the competitive mapping step. To obtain the histogram, the breadth value for the reference genome that obtained the highest breadth of coverage was selected for each Illumina library.

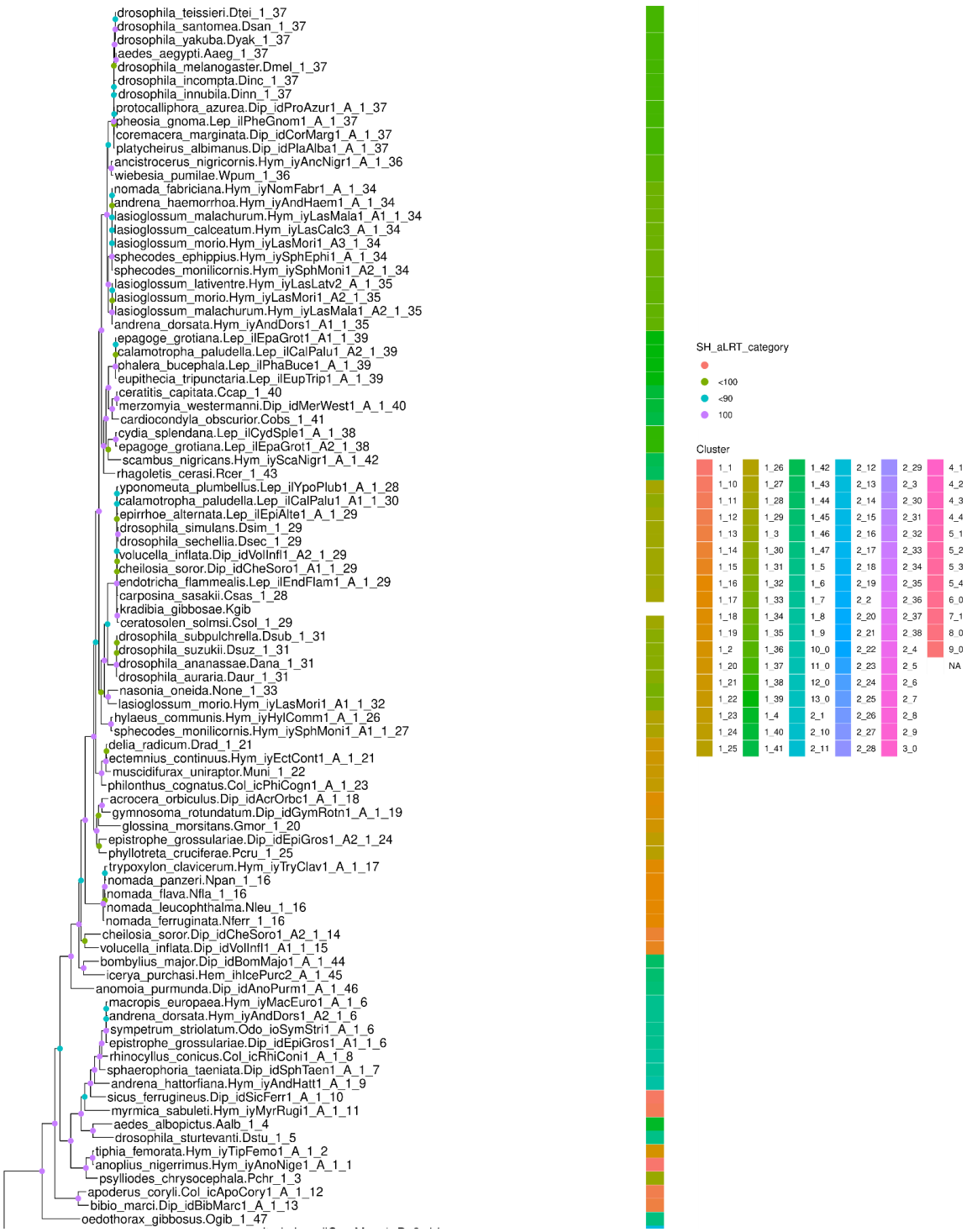
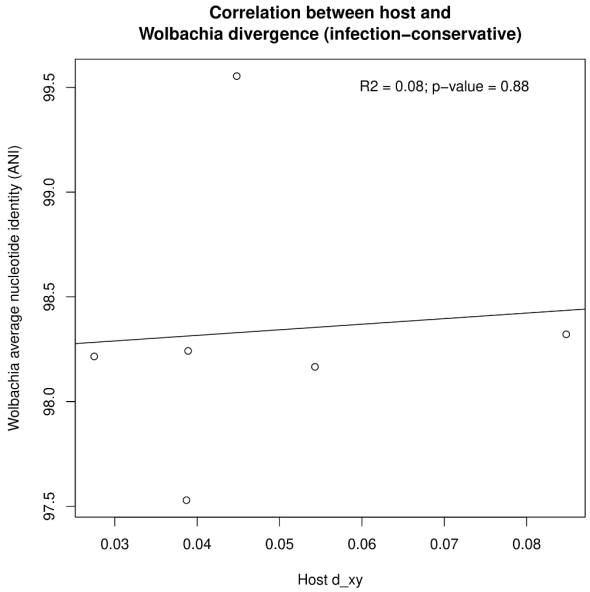


Figure S2. Phylogenetic tree of all the *Wolbachia* genomes, reconstructed using Maximum Likelihood. Supergroup A is shown in this section of the tree. Node colors indicate bootstrap support, while the heatmap indicates the clusters formed by dRep based on an average nucleotide identity (ANI) threshold of 99%. Cluster names are also indicated at the end of each tip label.

Figure S2 (continued). Phylogenetic tree of all the *Wolbachia* genomes, reconstructed using Maximum Likelihood. Supergroup B (sister to supergroup A) and the remaining supergroups (in the basal clade) are shown. Node colors indicate bootstrap support, while the heatmap indicates the clusters formed by dRep based on an average nucleotide identity (ANI) threshold of 99%. Cluster names are also indicated at the end of each tip label.

A



B

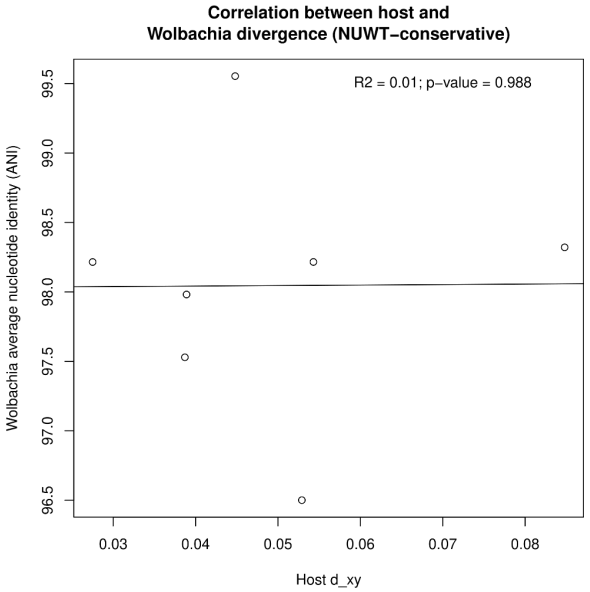


Figure S3. Correlation between host d_{xy} (obtained from Ebdon *et al.*, 2021) and *Wolbachia* average nucleotide identity (ANI) computed by FastANI.

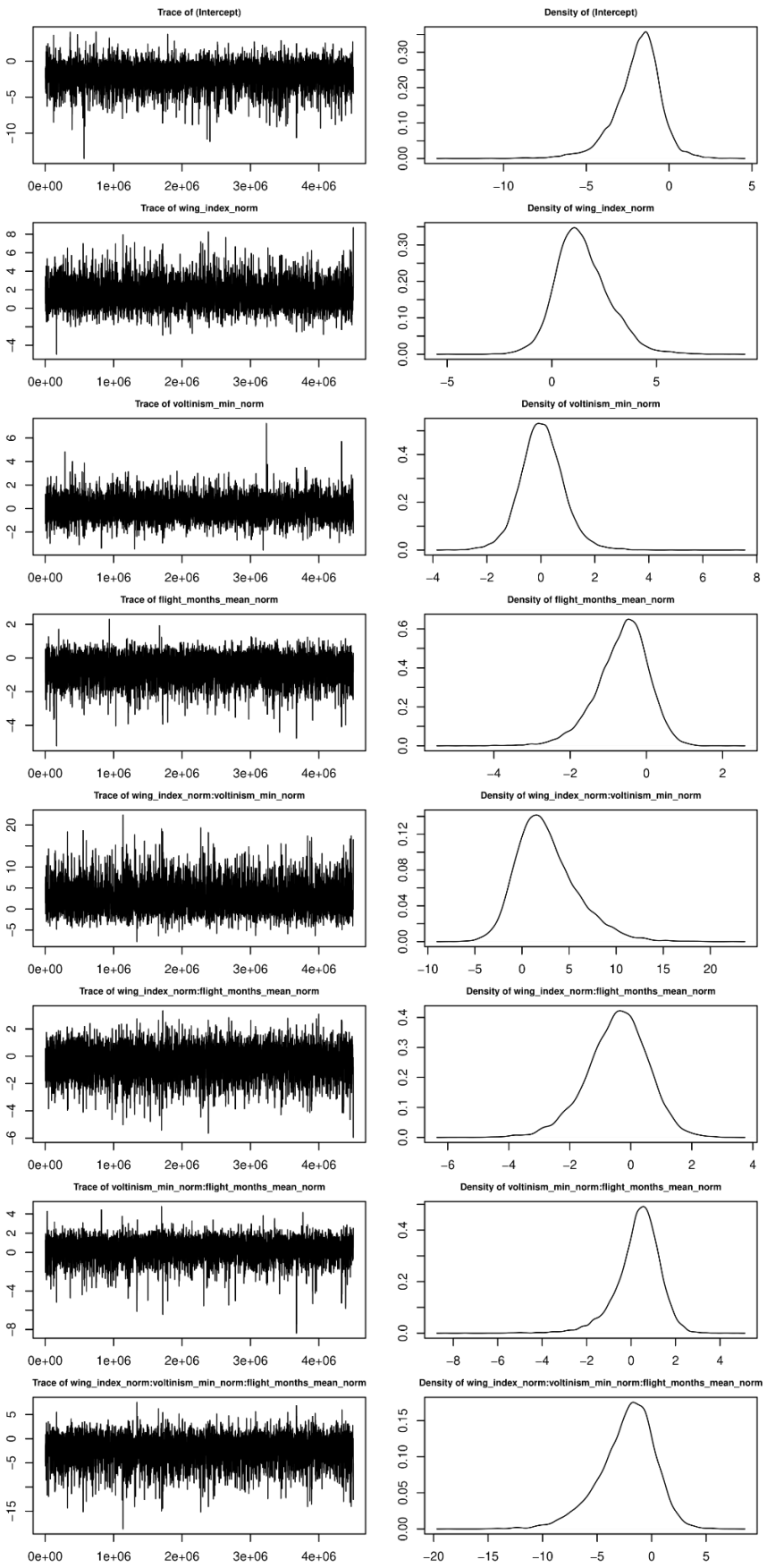


Figure S4. Trace plots and posterior distributions for the coefficients of the MCMCglmm model of number of strains per species, with the infection-conservative data.

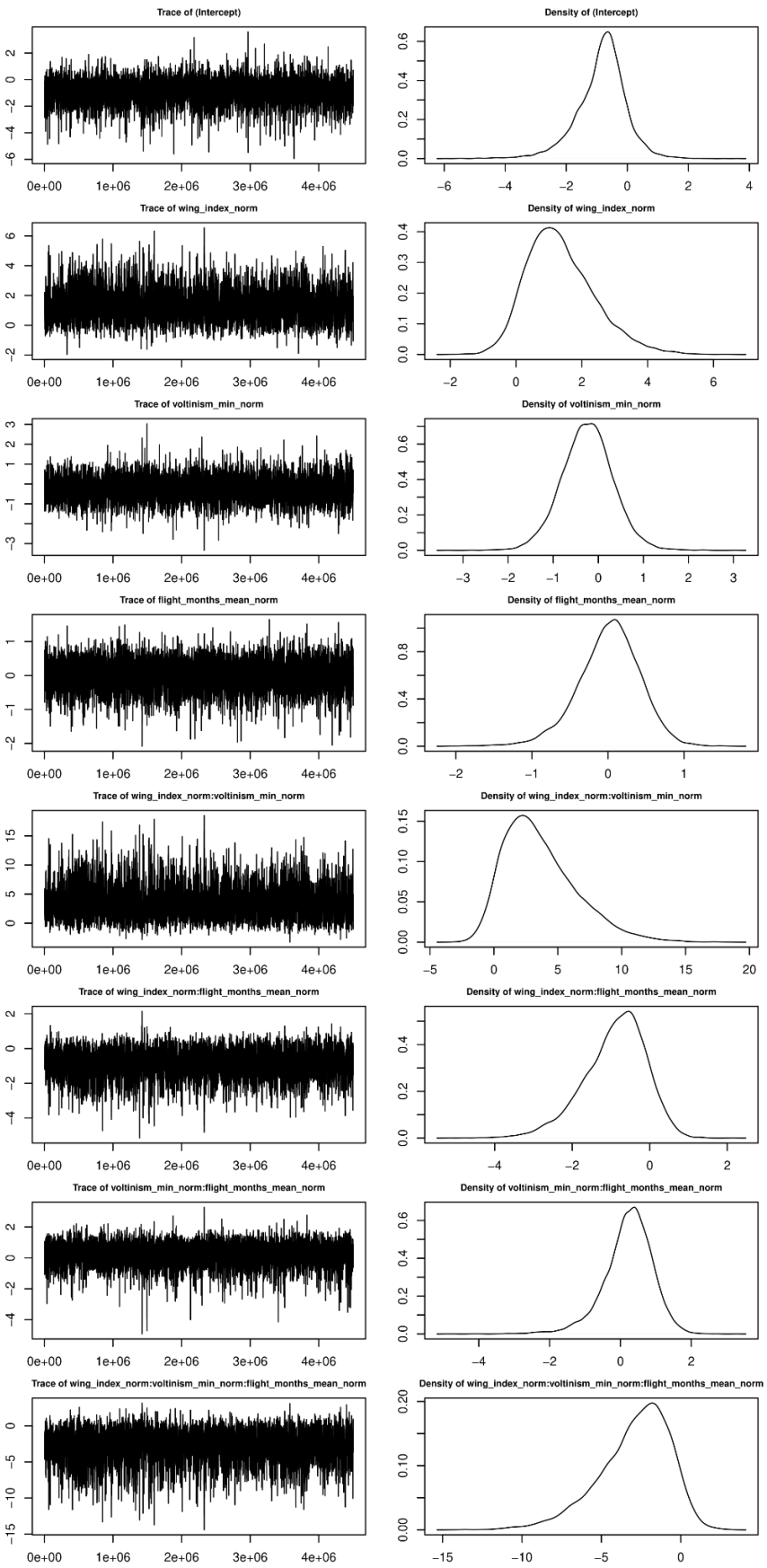


Figure S5: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of number of strains per species, with the NUWT-conservative data.

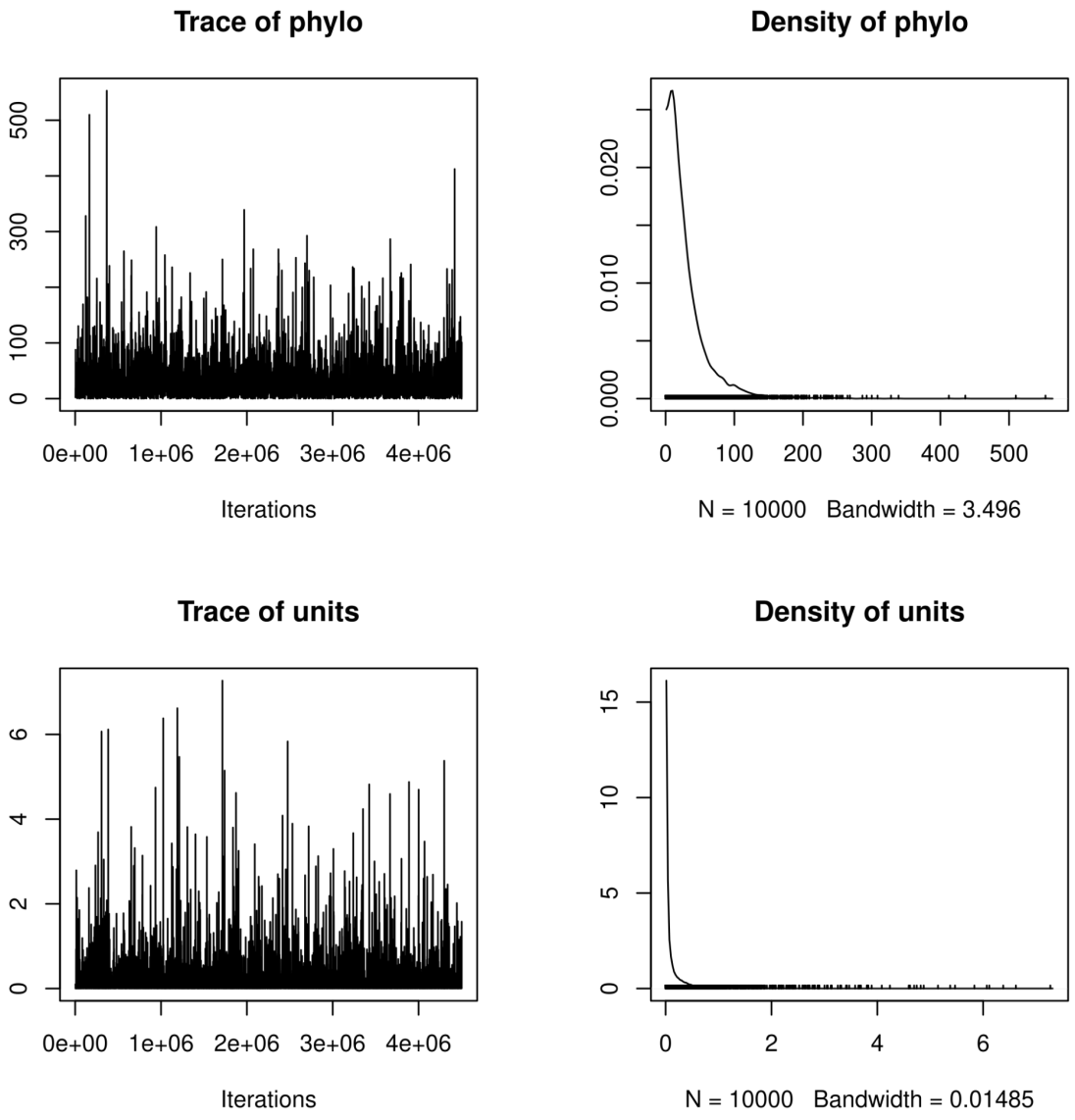


Figure S6: Trace plots and posterior distributions for the phylogenetic and residual variance of the MCMCglmm model of number of strains per species with the infection-conservative data.

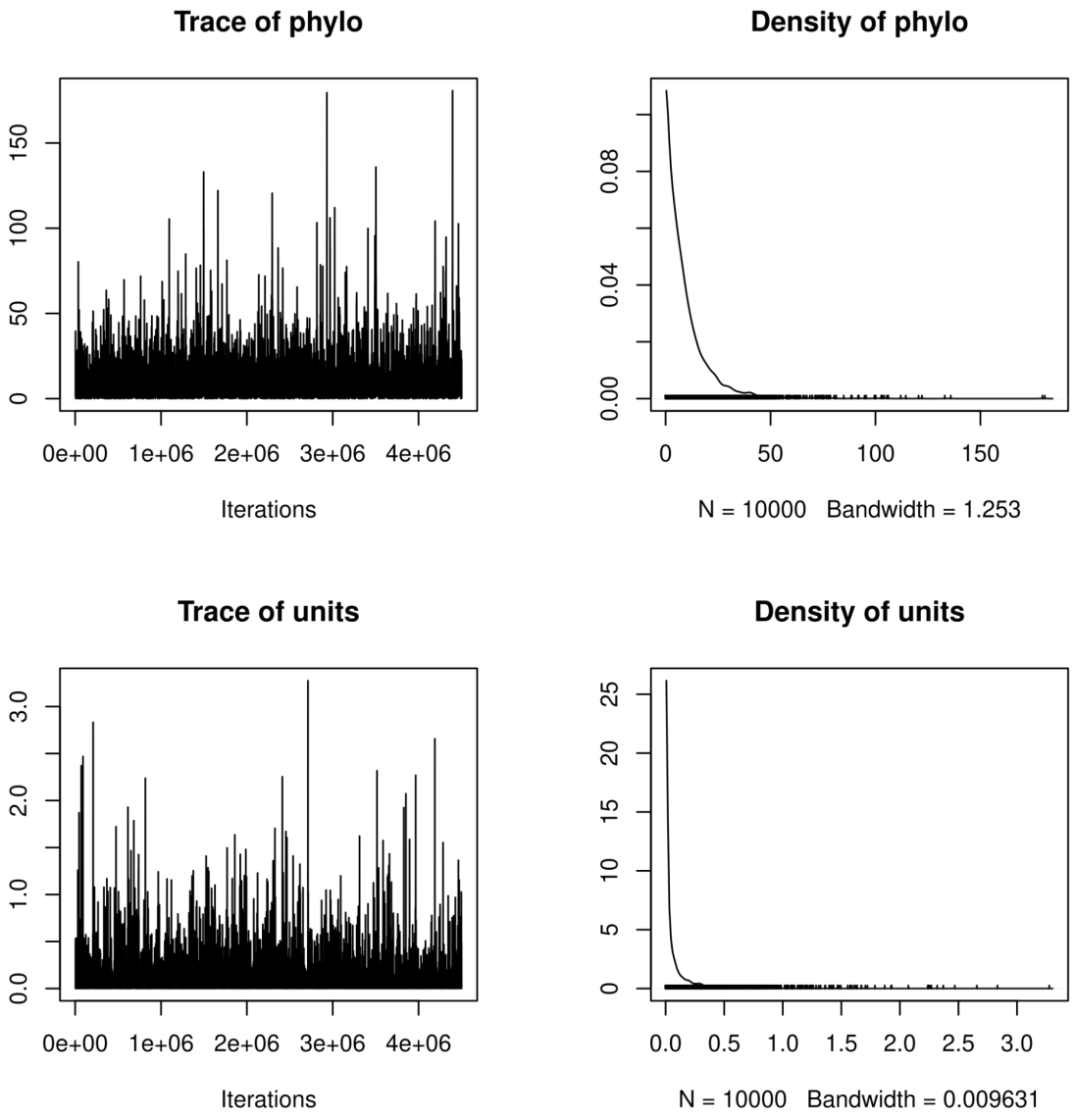


Figure S7: Trace plots and posterior distributions for the phylogenetic and residual variances of the MCMCglmm model of number of strains per species, with the NUWT-conservative data.

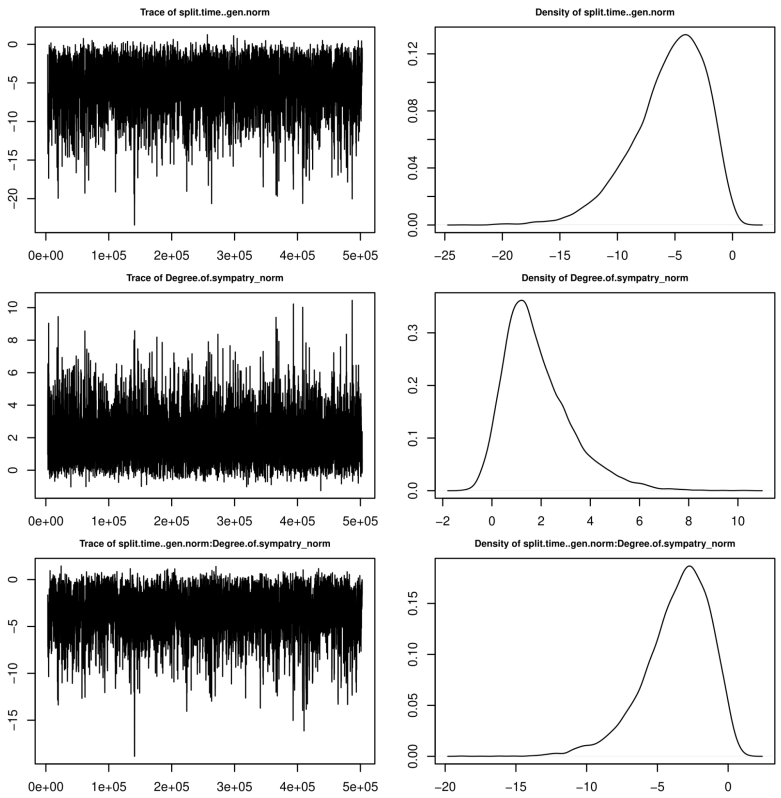


Figure S8: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the infection-conservative data and split times in number of generations.

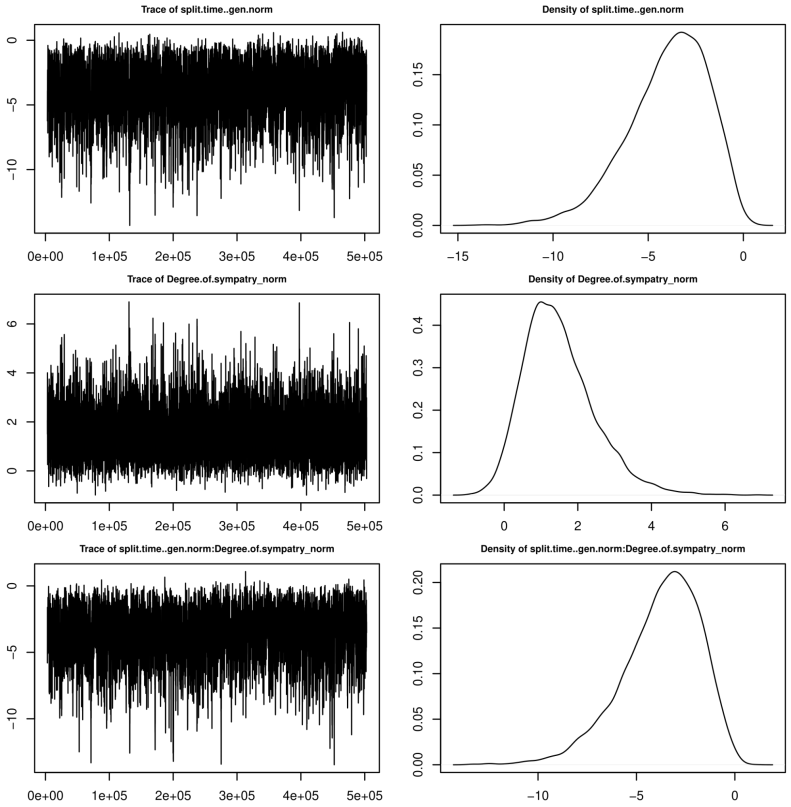


Figure S9: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the NUWT-conservative data and split times in number of generations.

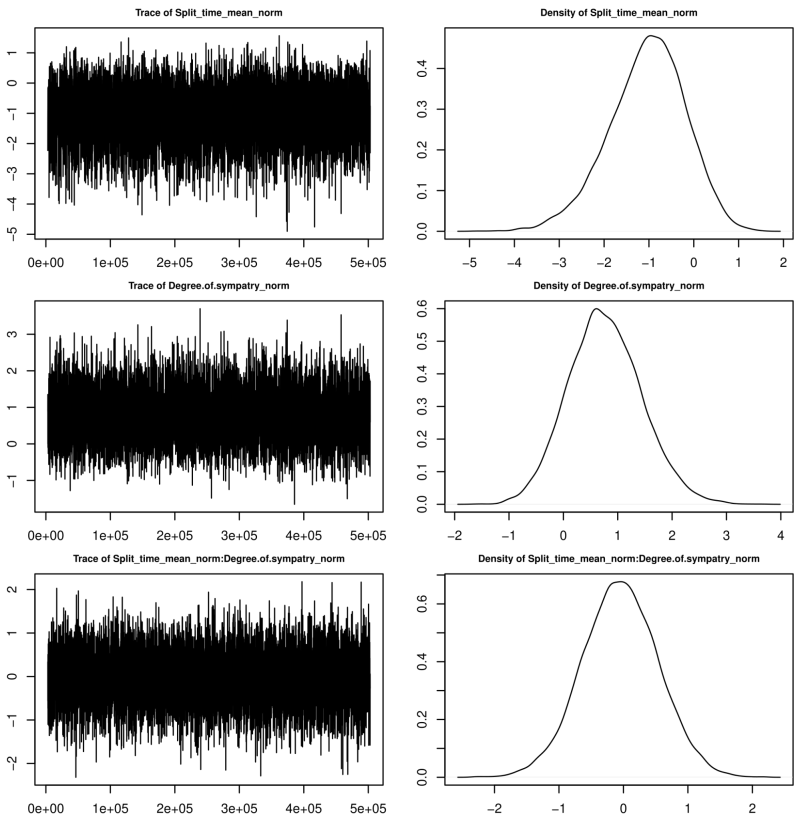


Figure S10: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the infection-conservative data and split times in million years.

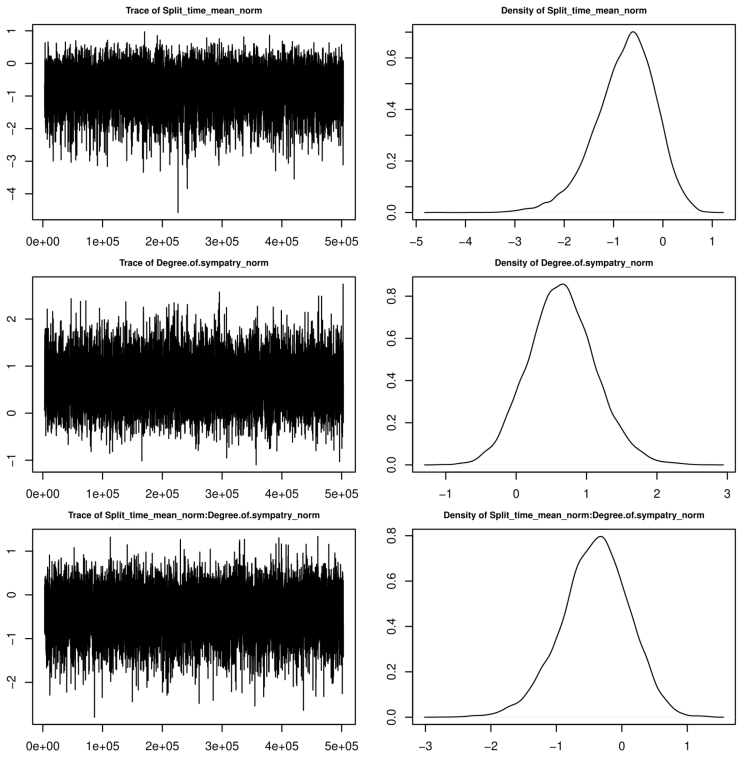


Figure S11: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the NUWT-conservative data and split times in million years.

Supplementary tables

Table S1A. List of all butterfly specimens analysed, with their *Wolbachia* infection status (infection-conservative version) and metadata. The meaning of the columns is as follows: Sample ID, the code used to identify a particular sample in the dataset; Genus, the taxonomic genus of the sample; Species; the specific epithet of the species to which that sample belongs; Reference_individual, whether that sample corresponds to a reference sample, for which PacBio Sequel libraries were also generated; Clear_absence; whether that sample presented very low depth and breadth of coverage when mapped against *Wolbachia* genomes and thus was confidently considered as not infected; large NUWTs present; whether this sample presented large NUWTs (nuclear *Wolbachia* transfers); *Wolbachia* presence; whether that sample was considered as infected with *Wolbachia*; Closest reference *Wolbachia*; the name of the *Wolbachia* genome representing the cluster that obtained the highest breadth of coverage in the competitive mapping stage, coinfections are separated with commas; Host nuclear mean coverage (Illumina), the mean coverage obtained when mapping the Illumina library against the butterfly genome of that species or, when not available, the closest genome available; Closest *Wolbachia* coverage (Illumina); the mean coverage obtained when mapping the Illumina library against the *Wolbachia* genome that obtained the highest breadth in the competitive mapping stage; Relative *Wolbachia*/host nuclear coverage, the ratio of *Wolbachia* and host mean coverages; Sex, the sex of the butterfly specimen; Date, the date in which the specimen was collected; Collector, the person who collected the specimen; Locality; the name of the locality in which the specimen was collected; Island, the name of the island (when applicable) in which the specimen was collected; State, the name of the state in which the specimen was collected; Country, the name of the country in which the specimen was collected; Continent, the name of the continent in which the specimen was collected; Latitude, the latitude of the coordinates of the point of collection of the specimen; Longitude, the longitude of the coordinates of the point of collection of the specimen; Tissue extracted, the tissue that was used for the DNA extraction

See supplementary file Table S1A.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S1B. List of all butterfly specimens analysed, with their *Wolbachia* infection status (NUWT-conservative version) and metadata. The meaning of the columns is as follows: Sample ID, the code used to identify a particular sample in the dataset; Genus, the taxonomic genus of the sample; Species; the specific epithet of the species to which that sample belongs; Reference_individual, whether that sample corresponds to a reference sample, for which PacBio Sequel libraries were also generated; Clear_absence; whether that sample presented very low depth and breadth of coverage when mapped against *Wolbachia* genomes and thus was confidently considered as not infected; large NUWTs present; whether this sample presented large NUWTs (nuclear *Wolbachia* transfers); *Wolbachia* presence; whether that sample was considered as infected with *Wolbachia*; Closest reference *Wolbachia*; the name of the *Wolbachia* genome that obtained the highest breadth of coverage in the competitive mapping stage, coinfections are separated with commas; Host nuclear mean coverage (Illumina), the mean coverage obtained when mapping the Illumina library against the butterfly genome of that species or, when not available, the closest genome available; Closest *Wolbachia* coverage (Illumina); the mean coverage obtained when mapping the Illumina library against the *Wolbachia* genome that obtained the highest breadth in the competitive mapping stage; Relative *Wolbachia*/host nuclear coverage, the ratio of *Wolbachia* and host mean coverages; Sex, the sex of the butterfly specimen; Date, the date in which the specimen was collected; Collector, the person who collected the specimen; Locality; the name of the locality in which the specimen was collected; Island, the name of the island (when applicable) in which the specimen was collected; State, the name of the state in which the specimen was collected; Country, the name of the country in which the specimen was collected; Continent, the name of the continent in which the specimen was collected; Latitude, the latitude of the coordinates of the point of collection of

the specimen; Longitude, the longitude of the coordinates of the point of collection of the specimen; Tissue extracted, the tissue that was used for the DNA extraction.

See supplementary file Table S1B.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S2. Quality metrics of the obtained genome assemblies. The column 'Final' indicates if an assembly was kept for further analyses or not.

See supplementary file Table_S2.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S3. Arguments used to run the different programs involved in the study.

Program	Version	Function	Example
FASTP	0.23.2	Read trimming	fastp -i {reads forward} -I {reads reverse} -o {trimmed reads forwards} -O {trimmed reads reverse} --cut by quality5 --cut by quality3 --cut window size 4 --cut mean quality 20 --html {sample name}.html --thread 16
BLASTn	2.13.0+	Similarity search of sequences	blastn -query {ref_genome} -out {outdir} -num_threads 20 -max_target_seqs 10 -max_hsps 1 -db {nt_database} -evalue 1e-25 -outfmt '6 qseqid staxids bitscore std'
minimap2	2.22-r1101	Align PacBio reads	minimap2 -t 20 -ax map-pb {ref_genome} {pacbio_reads}
		Align Illumina paired-end reads (interleaved)	minimap2 -t 20 -ax sr {ref_genome} {sr_reads1}
BlobTools	1.1.1	Create database	blobtools create -i {ref_genome} -t {blast_results} -b {pacbio_bamfile} -b {illumina_bamfile} -o {out_folder}
		Create taxonomic assignment files	blobtools view -i {blobDB} -o {out_prefix} -r all --hits
		Make BlobPlots	blobtools plot -i {blobDB} -o {out_prefix} -r order
		Partition Illumina paired-end reads	blobtools bamfilter -b {illumina_bamfile} -i {contig_IDs} -o {outfile}
Flye	2.9.1-b1780	Long-read metagenome assembly	flye --pacbio-raw {pacbio_reads} --out-dir {out_dir} -t 40 -i 1 --meta
NextDenovo	2.4.0	Long-read	[General]

(config file)		genome assembly	job_type = local
			job_prefix = {out_prefix}
			task = all # 'all', 'correct', 'assemble'
			rewrite = yes # yes/no
			deltmp = yes
			rerun = 0
			parallel_jobs = 5
			input_type = raw
			read_type = clr
			input_fofn = {input_files_list.fofn}
			workdir = {work_dir}
			
			[correct_option]
			read_cutoff = 1k
			genome_size = 1500000
			pa_correction = 5
			sort_options = -m 40g -t 20 -k 80
			minimap2_options_raw = -t 12
			correction_options = -p 13
			seed_cutoff = 10000
			
			[assemble_option]
			minimap2_options_cns = -t 12
			nextgraph_options = -a 1

HAPO-G	1.3.4	Assembly polishing	hapog --genome {assembly} --pe1 {illumina_reads1} --pe2 {illumina_reads2} -u --output {out_file} --threads 20
Pilon	1.24	Assembly polishing	java -Xmx20G -jar ~/pilon-1.24.jar --genome {genome_assembly} --changes --vcf --tracks --fix all,circles --iupac --frags {bam_file_illumina_against_assembly} --output {out_prefix} --outdir {out_dir}
BUSCO	5.4.2	Evaluate genome quality	busco -m genome -i {genome_assembly} -o {out_dir} -l {busco_database} -c 30
Prokka	1.14.6	Genome annotation	prokka -force --outdir {out_dir} --prefix {out_prefix} --cdsrnaolap --addgenes --addmrna --gcode 11 --kingdom bacteria --genus Wolbachia --cpus 20 --mincontiglen 1000 --proteins {reference_proteomes} {genome_assembly}
OrthoFinder	2.5.4	Find single-copy orthologs (SCOs)	orthofinder -f {proteomes_folder} -n wolbachia -p {out_folder} -a 60 -t 60
KinFin	1.1	Recover SCOs with a given missing data percent	kinfin -g OrthoGroups.txt -c config.txt -s SequenceIDs.txt --target_count 1 --target_fraction 0.95 --min 0 --max 1
MAFFT	7.508	Align SCOs	mafft --thread 10 --genafpair --maxiterate 1000 {fasta} > {out_file}

FASconCAT-G	1.05.1	Build supermatrix	~/FASconCAT-G/FASconCAT-G_v1.05.1.pl -s -l -p -p -j
SuperCRUNCH	1.3.2	Build supermatrix	python ~/SuperCRUNCH/supercrunch-scripts/Concatenation.py -i {folder_with_alignments} -o {out_folder} --informat fasta --outformat phylip -s dash
BEDtools	2.30.0	Generate BED file of extended BUSCO regions	grep -v "^#" {busco_all_tsv}| awk '$2=="Complete"' | cut -f3,4,5 | bedtools slop -i - -g {genome_file} -b 1000 | bedtools sort -i - -g {genome_file} | bedtools merge -i - > {out_bedfile}
SAMtools	1.6	Subset BAM file based on BED file and quality	samtools view -q 20 -bL {bed_file} {bam_file} > {output_bam} && samtools index {output_bam}
BCFtools	1.17	Call variants	bcftools mpileup --threads 12 -f {genome} {subsetted_bam}| bcftools call --threads 12 -mv -Oz -o {gzipped_vcf}
		Normalise and filter variants	bcftools norm -f {genome} {gzipped_vcf} -Oz | bcftools view - -e 'QUAL<20 | DP<8' -Oz -o {gzipped_filtered_vcf}
		Call consensus sequence with IUPAC codes	cat {genome} | bcftools consensus --haplotype I {gzipped_filtered_vcf} > {consensus_seq}
IQTree	2.2.0.3	Build phylogeny	iqtree -s {supermatrix} -p {partitions_file} -m PROTGAMMAGTR -bb 1000 -bnni -alrt 1000 -nt 30 -safe -pre {out_prefix}
inStrain	1.6.4	Evaluate competitive mapping	inStrain profile {bamfile} {concatenated_genomes} --stb {contigs_to_genomes_map} -o {out_folder} -p 30 --database_mode
dRep	3.4.3	Dereplicate *Wolbachia* genome set	dRep dereplicate {out_folder} -g {genomes} --S_ani 0.99 --S_algorithm fastANI
FastANI	1.33	Compute average nucleotide identity (ANI)	fastANI --rl reference_genomes.txt --ql query_genomes.txt -o {out_file} --matrix --visualize -t 30

Table S4. inStrain analysis of the competitive mapping of the Illumina libraries against the *Wolbachia* genomes. The first sheet contains the results for the genomic reference individuals, while the remaining sheets contain the results for all samples in each genus. For each genus, there is one

sheet containing the results when mapping against the full set of *Wolbachia* reference genomes (with suffix ".vs.all", and another sheet with the results of the competitive mapping against a reduced set comprised of the best mapping genomes form the former mapping (with suffix ".vs.top"). Each sheet contains results for all the samples of that genus separated by an empty row; the columns indicate the reference genome in question, the coverage it obtained, the observed breadth (number of bases of that genome covered by at least one read), the expected breadth (theoretical expectation of breadth based on the coverage if that were the genome from which the reads were generated), and the difference between expected and observed breadth (E-O).

See supplementary file Table_S4.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S5A. Infection status for all the butterfly species analysed in this study, considering the infection-conservative approach. Genus, species genus; Species; species specific epithet; Total samples, total number of samples screened for that species; *Wolbachia* presence, binary variable indicating if *Wolbachia* was detected on a given species; Infected samples, number of samples with *Wolbachia*; *Wolbachia* prevalence, proportion of samples infected with *Wolbachia*; NUWT presence, whether NUWTs were detected on a given species; NUWT prevalence, proportion of samples in which NUWTs were detected; Number of strains, number of strains detected in a given species; *Wolbachia* in literature, whether a given species is described as infected, uninfected, or is not found in the literature; Reference individual, code of the genomic reference individual of that species; Supergroup (ref. ind.), *Wolbachia* supergroup of the infection in the genomic reference individual of that species; NCBI TaxID, the taxonomic ID number of that species; Supergroup (literature), the supergroup to which the *Wolbachia* infection in the literature belongs; Strain, strain classification from the literature based on MLST and wsp markers; Prevalence in literature, the prevalence of *Wolbachia* in that species found in the literature; Reference, bibliographic reference of the study that reports the infection status of that species; Geographic region, region of procedence of the screened specimens in the literature; Closest RefSeq hit, *Wolbachia* RefSeq genome that is the closest match to the one in the genomic reference individuals; DToL, whether a *Wolbachia* genome used in this study was obtained from the Darwin Tree of Life project; Comments, additional comments on the infection status.

See supplementary file Table_S5A.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S5B. Infection status for all the butterfly species analysed in this study, considering the NUWT-conservative approach. Genus, species genus; Species; species specific epithet; Total samples, total number of samples screened for that species; *Wolbachia* presence, binary variable indicating if *Wolbachia* was detected on a given species; Infected samples, number of samples with *Wolbachia*; *Wolbachia* prevalence, proportion of samples infected with *Wolbachia*; NUWT presence, whether NUWTs were detected on a given species; NUWT prevalence, proportion of samples in which NUWTs were detected; Number of strains, number of strains detected in a given species; *Wolbachia* in literature, whether a given species is described as infected, uninfected, or is not found in the literature; Reference individual, code of the genomic reference individual of that species; Supergroup (ref. ind.), *Wolbachia* supergroup of the infection in the genomic reference individual of that species; NCBI TaxID, the taxonomic ID number of that species; Supergroup (literature), the supergroup to which the *Wolbachia* infection in the literature belongs; Strain, strain classification from the literature based on MLST and wsp markers; Prevalence in literature, the prevalence of *Wolbachia* in that species found in the literature; Reference, bibliographic reference of the study that reports the infection status of that species; Geographic region, region of procedence of the screened specimens in the literature; Closest RefSeq hit, *Wolbachia* RefSeq genome that is the closest match to the one in the genomic reference individuals; DToL, whether a *Wolbachia* genome used in this study was

obtained from the Darwin Tree of Life project; Comments, additional comments on the infection status.

See supplementary file Table_S5B.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S6A. Comparison of the infection status across sister pairs of butterfly species, being infection-conservative. Genus, species genus; Species 1; first species of the pair; π sp.1, nucleotidic diversity of the first species; Gen y−1 sp.1; generations per year of the first species; Species 2, second species in the pair; π sp.2, nucleotidic diversity of the second species; Gen y−1 sp.2; generations per year of the second species; d_xy, mean genetic divergence between the species in the pair; d_a, net genetic divergence between the species in the pair; split time (gen), split time in number of generations; Split time (MYA), split time in million years ago; F_st, fixation index; Degree of sympatry, proportion of range overlap between the two species; Contact zone, whether the species in the pair have a contact zone; Known to hybridize, whether the two species in the pair are known to produce hybrids; None infected; whether none of the species in the pair are infected with *Wolbachia*; One infected, whether one species in the pair is infected and the other is not; Both infected, whether both species in the pair are infected; Presence of shared strains in the pair; whether the species pair has at least one strain in common; Presence of specific strains within the pair; whether there are strains in at least one species of the pair that are absent in the other species of the pair; Presence of species-specific strains; whether at least one species in the pair has at least one strain that was not detected in any other species (either in the pair or outside); Presence of strains shared outside the pair (e.g. with other genera), whether at least one species in the pair has some strain that is also found in another species not form the pair; strain_num_sp1, number of strains in the first species of the pair; strain_num_sp2, number of species in the second species of the pair; num_shared_strains, number of strains that are found in both species of the pair.

See supplementary file Table_S6A.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S6B. Comparison of the infection status across sister pairs of butterfly species, being NUWT-conservative. Genus, species genus; Species 1; first species of the pair; π sp.1, nucleotidic diversity of the first species; Gen y−1 sp.1; generations per year of the first species; Species 2, second species in the pair; π sp.2, nucleotidic diversity of the second species; Gen y−1 sp.2; generations per year of the second species; d_xy, mean genetic divergence between the species in the pair; d_a, net genetic divergence between the species in the pair; split time (gen), split time in number of generations; Split time (MYA), split time in million years ago; F_st, fixation index; Degree of sympatry, proportion of range overlap between the two species; Contact zone, whether the species in the pair have a contact zone; Known to hybridize, whether the two species in the pair are known to produce hybrids; None infected; whether none of the species in the pair are infected with *Wolbachia*; One infected, whether one species in the pair is infected and the other is not; Both infected, whether both species in the pair are infected; Presence of shared strains in the pair; whether the species pair has at least one strain in common; Presence of specific strains within the pair; whether there are strains in at least one species of the pair that are absent in the other species of the pair; Presence of species-specific strains; whether at least one species in the pair has at least one strain that was not detected in any other species (either in the pair or outside); Presence of strains shared outside the pair (e.g. with other genera), whether at least one species in the pair has some strain that is also found in another species not form the pair; strain_num_sp1, number of strains in the first species of the pair; strain_num_sp2, number of species in the second species of the pair; num_shared_strains, number of strains that are found in both species of the pair.

See supplementary file Table_S6B.xlsx on https://github.com/etd530/MSc_thesis_SoM

Table S7: Summary statistics for the MCMCglmm models of strain sharing and number of strains. Posterior mean indicates the mean of the posterior distribution obtained from the Markov Chain Monte Carlo (MCMC) process; lower and upper 95% CI indicate the 95% credibility interval of the distribution; effective sample size indicates the effective sample size of the markov chain after thinning and accounting for non-independence of the iterations; pMCMC indicates the p-value based on the values of the posterior distribution.

Model	Term	Posterior mean	lower 95% CI	upper 95% CI	Effective sample size	pMCMC
Strain sharing, infection-conservative, split time in generations	Split time	-5.522	-11.854	-0.159	1509	0.008**
	Degree of sympatry	1.854	-0.438	4.671	4077	0.075
	Split time : degree of sympatry	-3.605	-8.263	0.395	1553	0.043*
Strain sharing, infection-conservative, split time in Mya	Split time	-1.038	-2.734	0.585	9652	0.207
	Degree of sympatry	0.767	-0.519	2.124	10000	0.251
	Split time : degree of sympatry	-0.049	-1.156	1.165	10592	0.933
Strain sharing, NUWT-conservative, split time in generations	Split time	-3.968	-8.097	-0.283	1824	0.008**
	Degree of sympatry	1.487	-0.209	3.429	6053	0.061
	Split time : degree of sympatry	-3.744	-7.844	-0.365	1686	0.007**
Strain sharing, NUWT-conservative, split time in Mya	Split time	-0.765	-2.010	0.371	10000	0.175
	Degree of sympatry	0.648	-0.268	1.620	10000	0.170
	Split time : degree of sympatry	-0.411	-1.489	0.536	10000	0.429
Number of strains, infection-conservative	Intercept	-1.899	-4.992	0.653	6258	0.112
	Wing index	1.473	-0.796	4.126	3272	0.198
	Minimum voltinism	0.044	-1.585	1.616	4470	0.974
	Mean number of flight months	-0.630	-2.043	0.626	4241	0.326
	Wing index : voltinism	2.638	-2.793	9.633	2854	0.401
	Wing index : flight months	-0.449	-2.449	1.474	3383	0.671
	Voltinism : flight months	0.316	-1.826	2.024	4338	0.622
	Wing index : voltinism : flight months	-2.282	-7.623	2.503	3019	0.342
Number of strains, NUWT-conservative	Intercept	-0.842	-2.548	0.695	7450	0.220
	Wing index	1.359	-0.500	3.475	1836	0.134
	Minimum voltinism	-0.243	-1.416	0.845	3218	0.659
	Mean number of flight months	0.013	-0.836	0.778	6212	0.926
	Wing index : voltinism	3.629	-1.125	9.315	1668	0.116
	Wing index : flight months	-0.902	-2.617	0.506	1900	0.223
	Voltinism : flight months	0.222	-1.151	1.566	5019	0.645
	Wing index : voltinism : flight months	-2.915	-7.546	0.949	1710	0.121