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

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

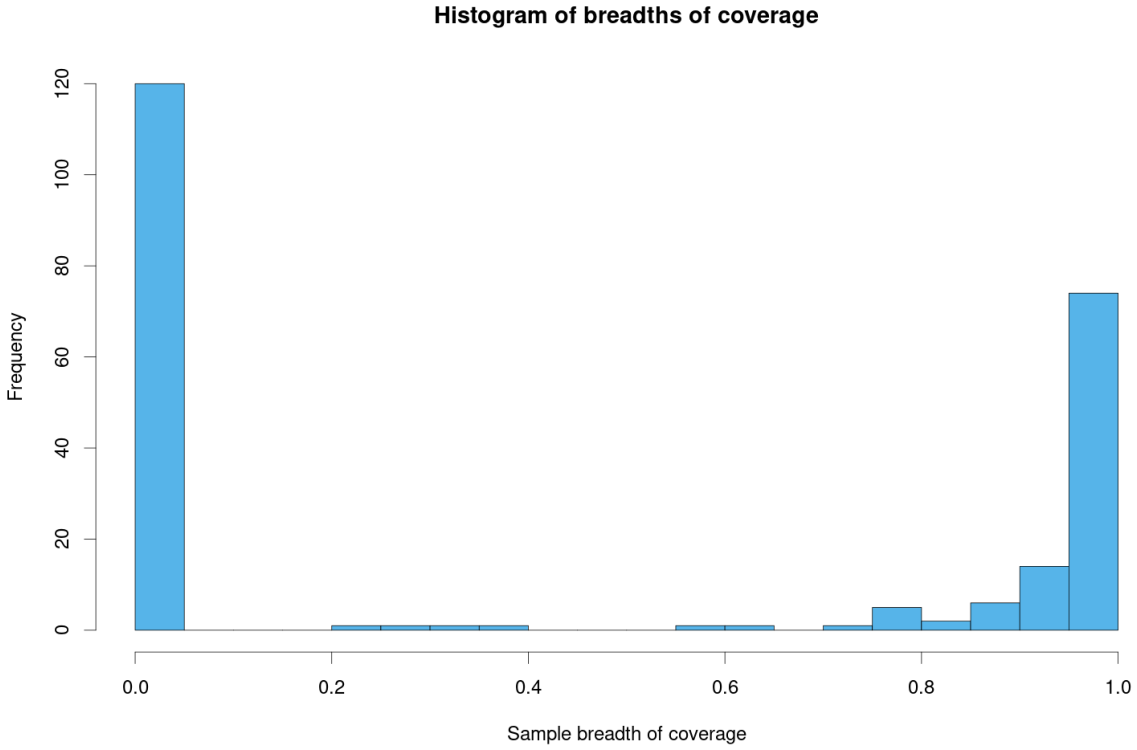

Figure S1. Histogram of the highest breadths of coverage obtained when mapping the Illumina paired-end libraries against the concatenation of *Wolbachia* reference genomes for the competitive mapping step. To obtain the histogram, the breadth value for the reference genome that obtained the highest breadth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

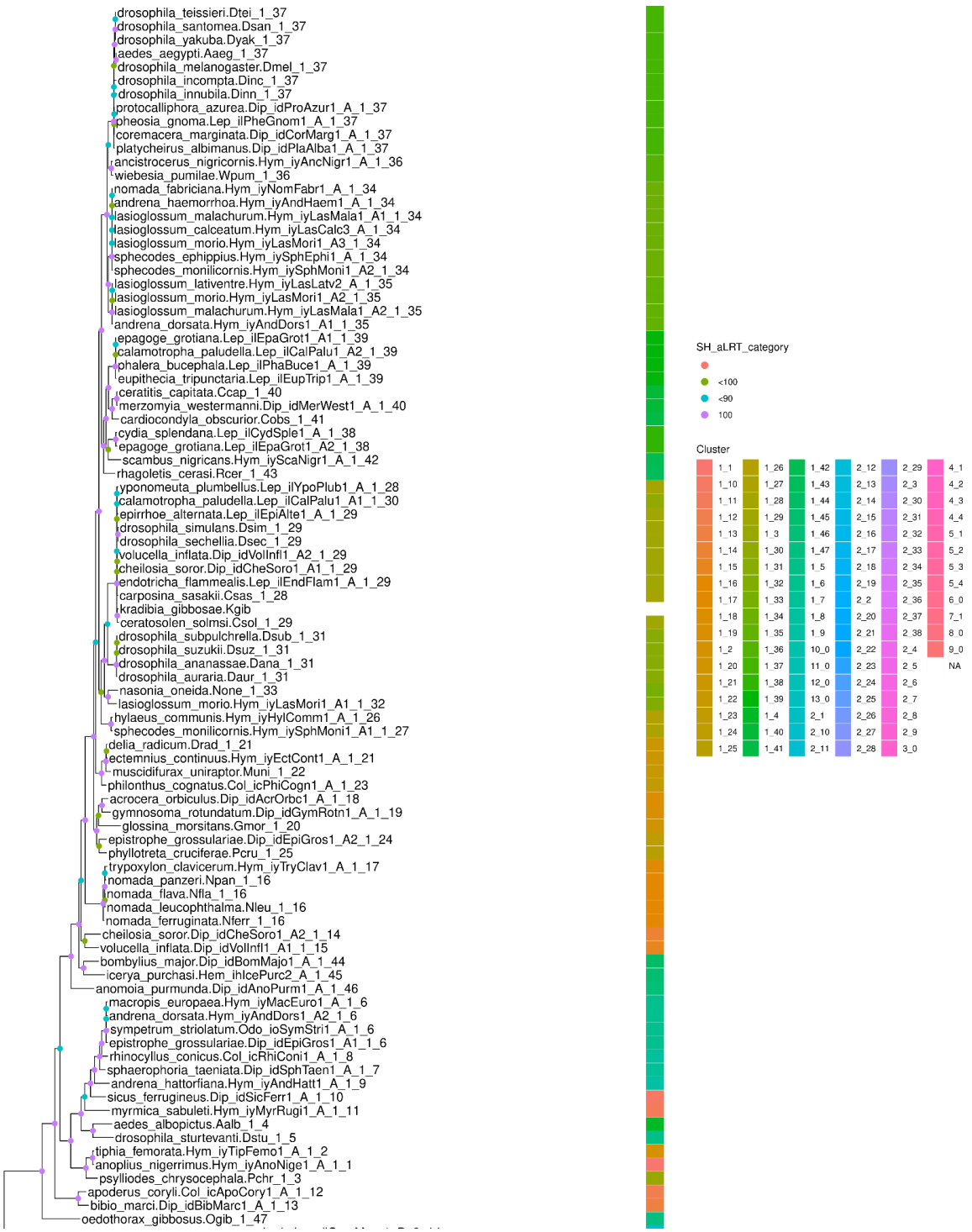

Figure S2. Phylogenetic tree of all the *Wolbachia* genomes, reconstructed using Maximum Likelihood. Supergroup A is shown in this section of the tree. Node colors indicate bootstrap support, while the heatmap indicates the clusters formed by dRep based on an average nucleotide identity (ANI) threshold 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.

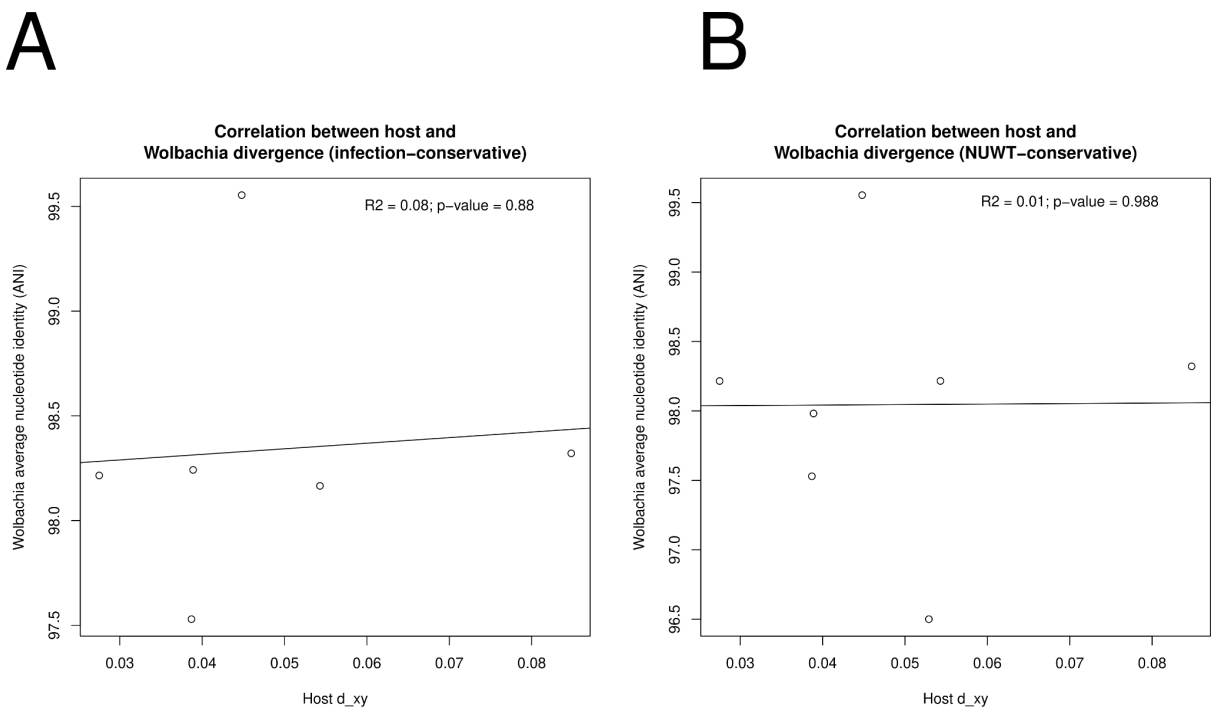

Figure S3. Correlation between host d_{xy} and *Wolbachia* average nucleotide identity (ANI)

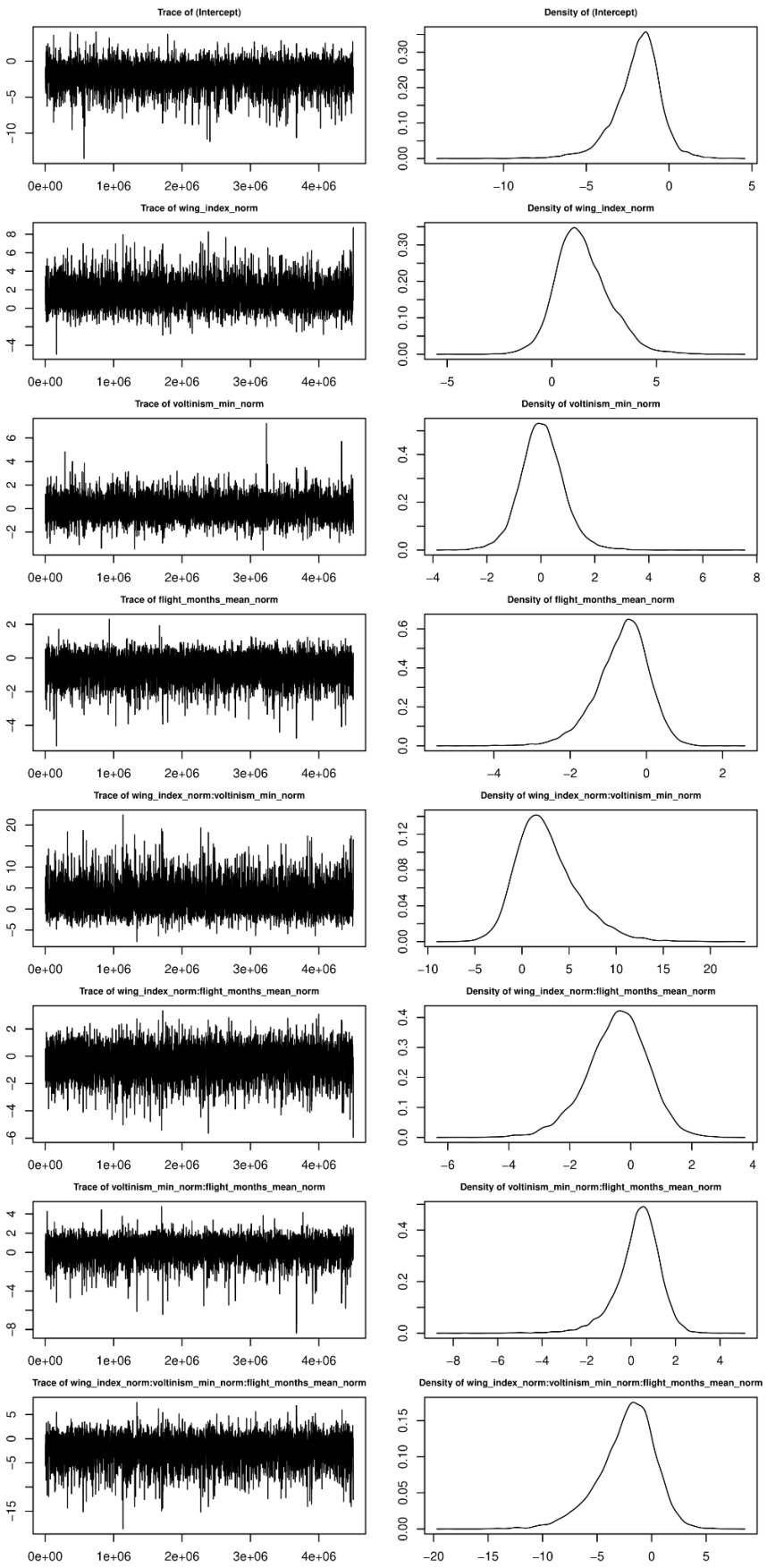

Figure S4. Trace plots and posterior distributions for the coefficients of the MCMCglmm model of number of strains per species, with the infection-conservative data.

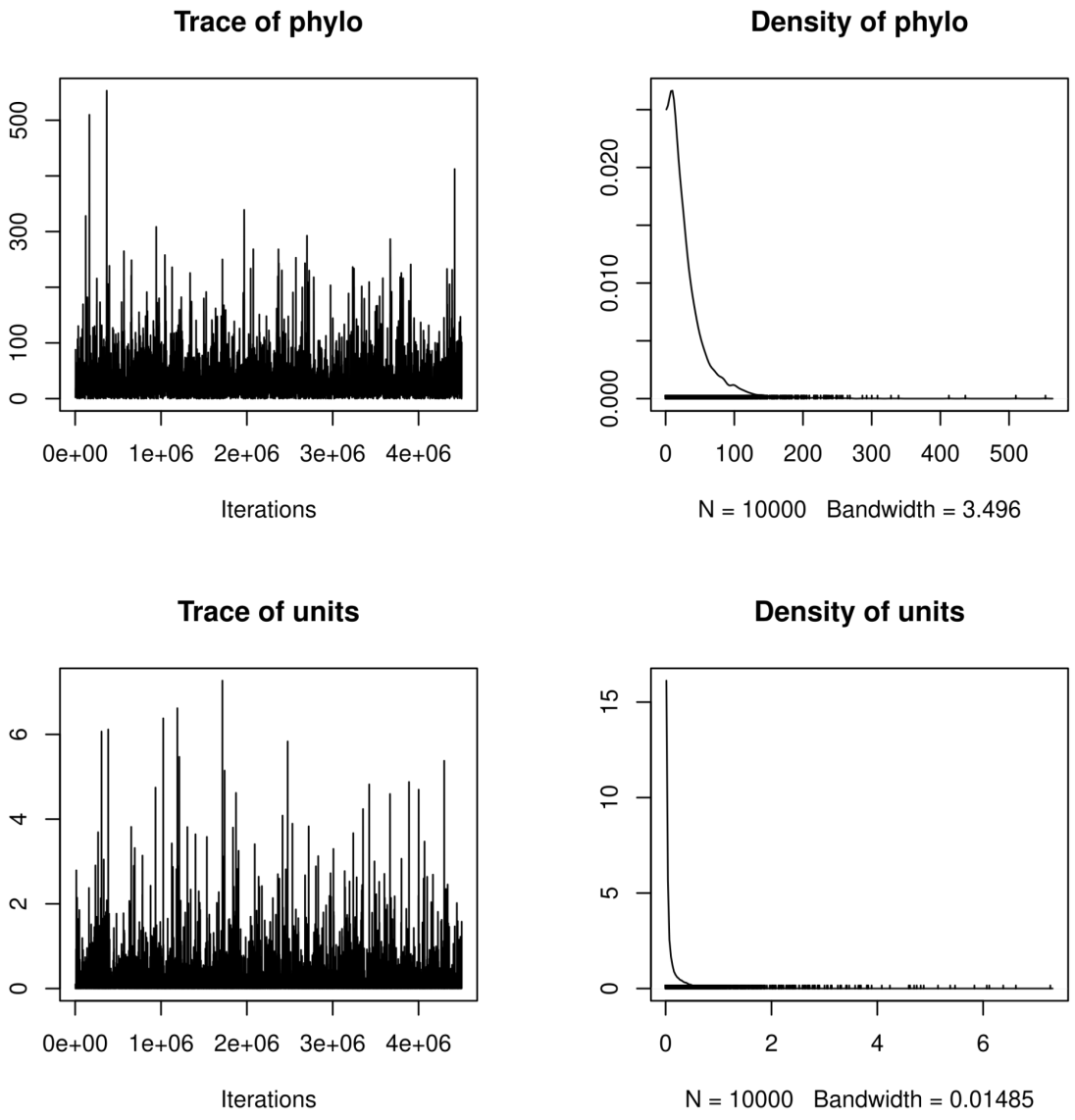

Figure S5: Trace plots and posterior distributions for the phylogenetic and residual variance of the MCMCglmm model of number of strains per species with the infection-conservative data.

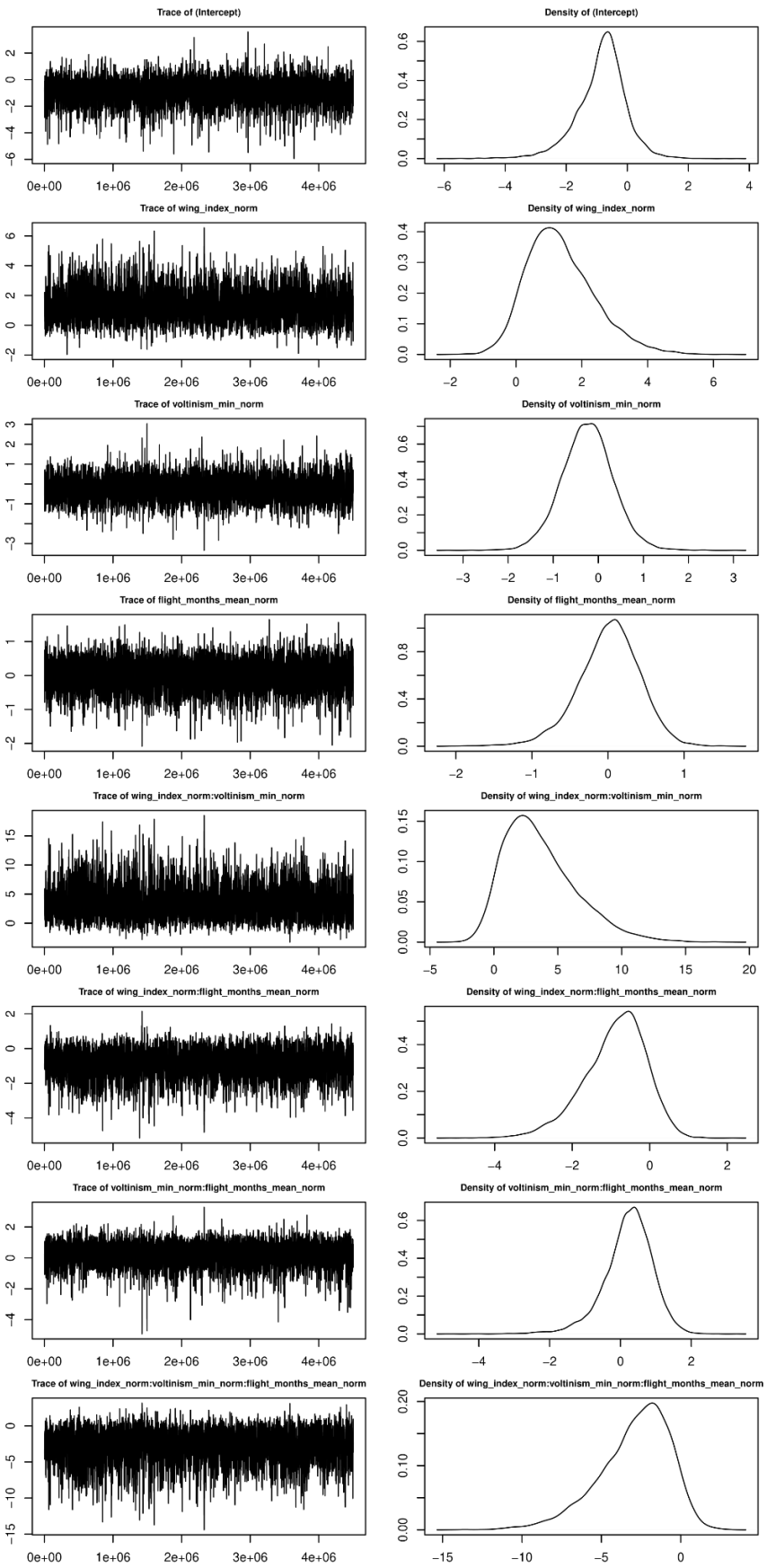

Figure S6: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of number of strains per species, with the NUWT-conservative data.

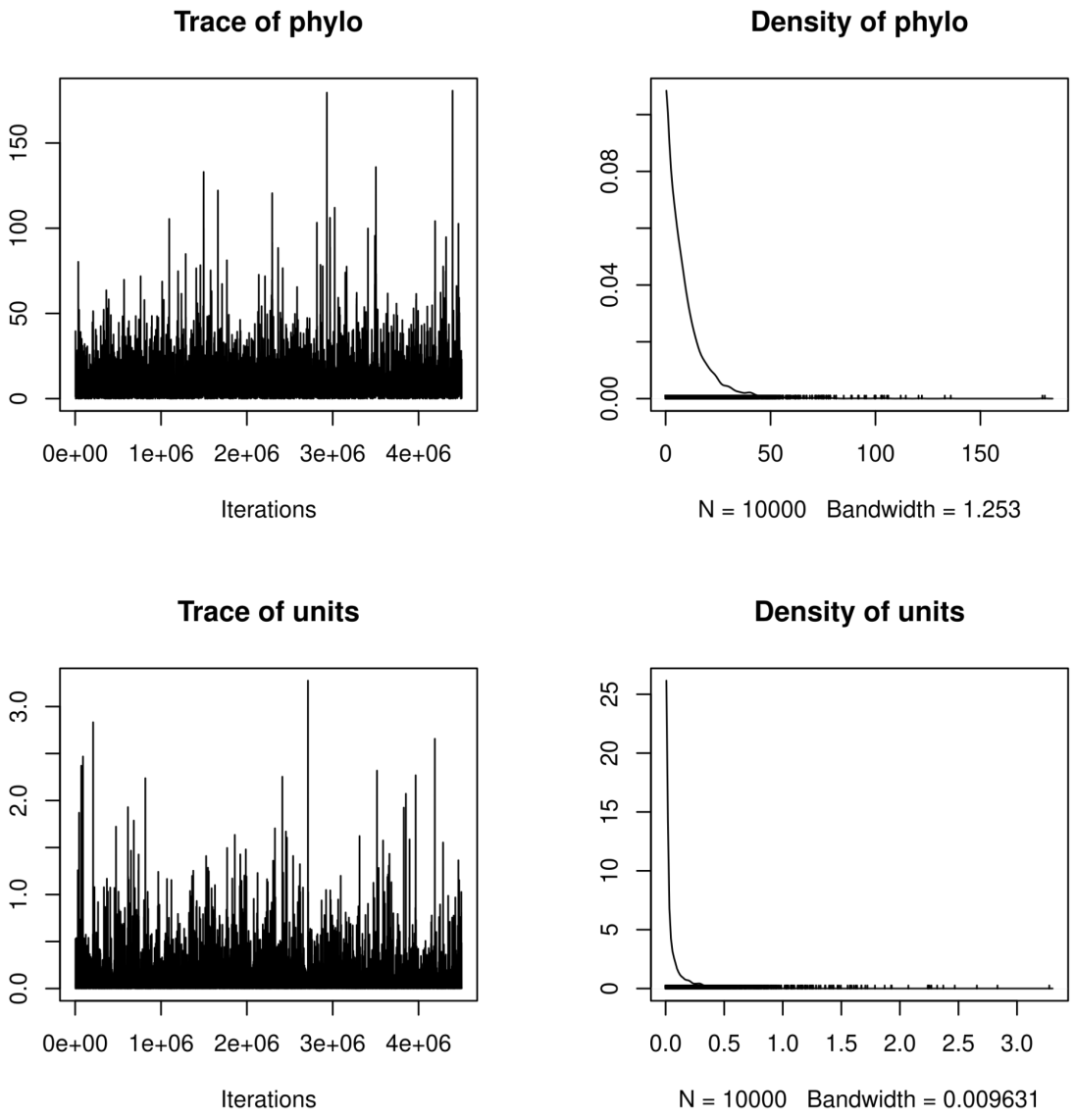

Figure S7: Trace plots and posterior distributions for the phylogenetic and residual variances of the MCMCglmm model of number of strains per species, with the NUWT-conservative data.

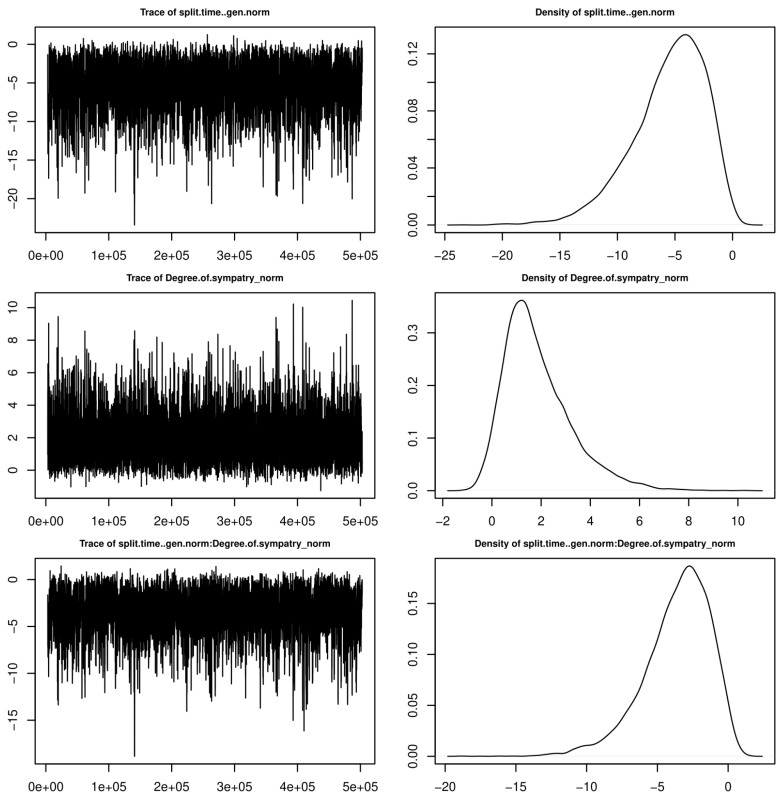

Figure S8: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the infection-conservative data and split times in number of generations.

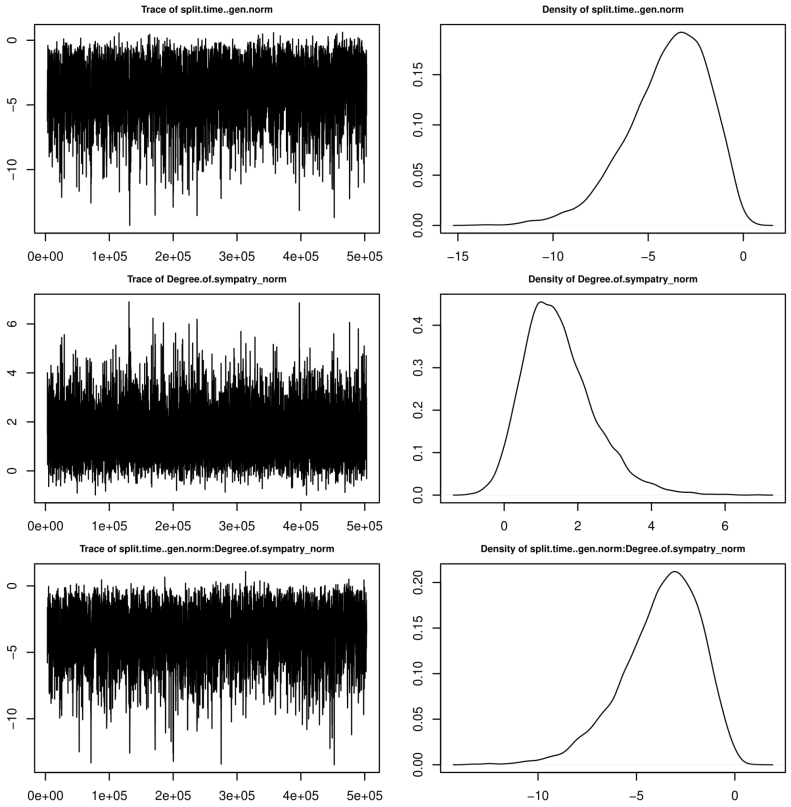

Figure S9: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the NUWT-conservative data and split times in number of generations.

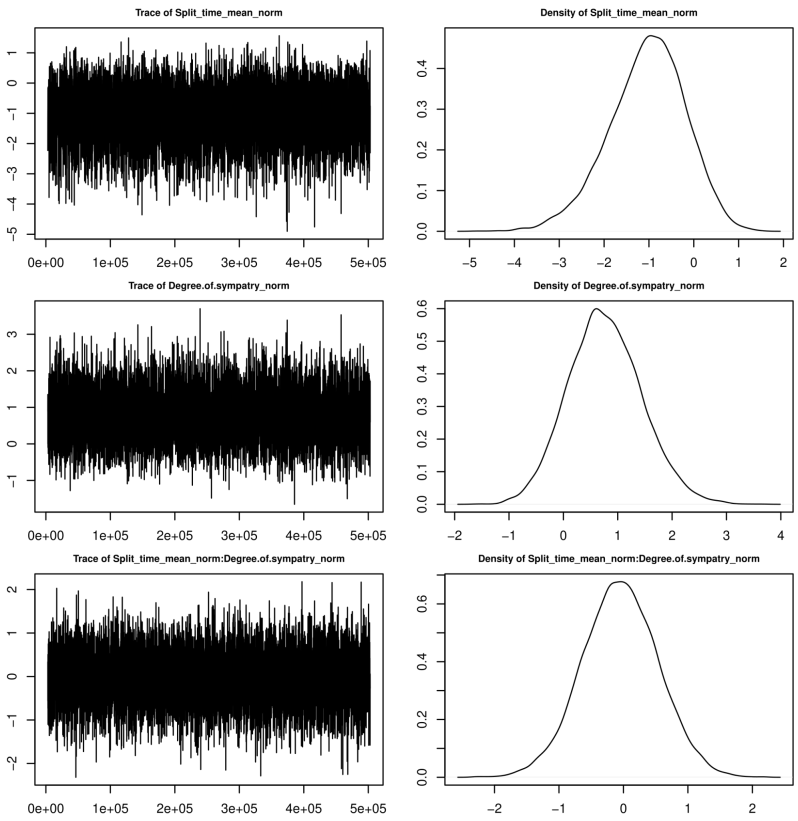

Figure S10: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the infection-conservative data and split times in million years.

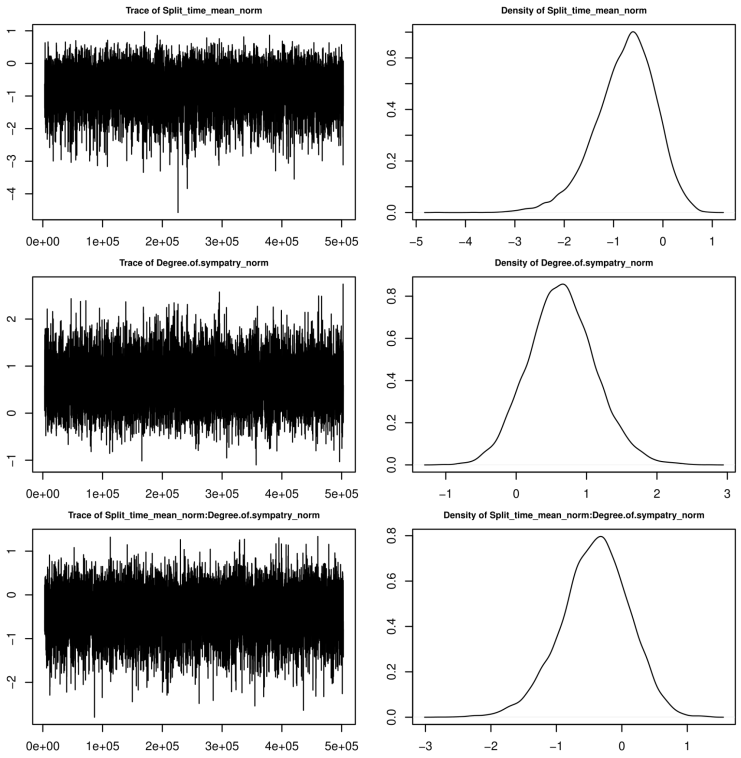

Figure S11: Trace plots and posterior distributions for the coefficients of the MCMCglmm model of strain sharing, with the NUWT-conservative data and split times in million years.

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

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

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

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 (config file)	2.4.0	Long-read genome assembly	[General]
			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/Concatenati

			on.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"). In 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

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

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

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, nmber 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

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, nmber 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

Table S7: Summary statistics for the MCMCglmm models of strain sharing and number of strains.

Model	Term	Posterior mean	lower 95% CI	upper 95% CI	Effective sample size	pMCMC
Strain sharing, infection-conservative, split time in generations	split.time..gen.norm	-5.522	-11.854	-0.159	1509	0.008
	Degree.of.sympatry_norm	1.854	-0.438	4.671	4077	0.075
	split.time..gen.norm:Degree.of.sympatry_norm	-3.605	-8.263	0.395	1553	0.043
Strain sharing, infection-conservative, split time in Mya	Split_time_mean_norm	-1.038	-2.734	0.585	9652	0.207
	Degree.of.sympatry_norm	0.767	-0.519	2.124	10000	0.251
	Split_time_mean_norm:Degree.of.sympatry_norm	-0.049	-1.156	1.165	10592	0.933
Strain sharing, NUWT-conservative, split time in generations	split.time..gen.norm	-3.968	-8.097	-0.283	1824	0.008
	Degree.of.sympatry_norm	1.487	-0.209	3.429	6053	0.061
	split.time..gen.norm:Degree.of.sympatry_norm	-3.744	-7.844	-0.365	1686	0.007
Strain sharing, NUWT-conservative, split time in Mya	Split_time_mean_norm	-0.765	-2.010	0.371	10000	0.175
	Degree.of.sympatry_norm	0.648	-0.268	1.620	10000	0.170
	Split_time_mean_norm:Degree.of.sympatry_norm	-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_norm	1.473	-0.796	4.126	3272	0.198
	voltinism_min_norm	0.044	-1.585	1.616	4470	0.974
	flight_months_mean_norm	-0.630	-2.043	0.626	4241	0.326
	wing_index_norm:voltinism_min_norm	2.638	-2.793	9.633	2854	0.401
	wing_index_norm:flight_months_mean_norm	-0.449	-2.449	1.474	3383	0.671
	voltinism_min_norm:flight_months_mean_norm	0.316	-1.826	2.024	4338	0.622
	wing_index_norm:voltinism_min_norm:flight_months_mean_norm	-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_norm	1.359	-0.500	3.475	1836	0.134
	voltinism_min_norm	-0.243	-1.416	0.845	3218	0.659
	flight_months_mean_norm	0.013	-0.836	0.778	6212	0.926
	wing_index_norm:voltinism_min_norm	3.629	-1.125	9.315	1668	0.116
	wing_index_norm:flight_months_mean_norm	-0.902	-2.617	0.506	1900	0.223
	voltinism_min_norm:flight_months_mean_norm	0.222	-1.151	1.566	5019	0.645
	wing_index_norm:voltinism_min_norm:flight_months_mean_norm	-2.915	-7.546	0.949	1710	0.121