bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

An endosymbiont harvest: Phylogenomic analysis of *Wolbachia* genomes from the Darwin Tree of Life biodiversity genomics project

Emmelien Vancaester[1*], Mark Blaxter[1]

[1] Tree of Life, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK

* Corresponding author:

Emmelien Vancaester ORCID https://orcid.org/0000-0002-9177-8808 ev3@sanger.ac.uk

Mark Blaxter ORCID https://orcid.org/0000-0003-2861-949X mb35@sanger.ac.uk

Keywords: *Wolbachia* alphaproteobacteria, endosymbiosis, host-symbiont coevolution, *Wolbachia* phage, cytoplasmic incompatibility

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

The Darwin Tree of Life project aims to sequence all described terrestrial and aquatic eukaryotic species found in Britain and Ireland. Reference genome sequences are generated from single individuals for each target species. In addition to the target genome, sequenced samples often contain genetic material from microbiomes, endosymbionts, parasites and other cobionts. *Wolbachia* endosymbiotic bacteria are found in a diversity of terrestrial arthropods and nematodes, with supergroups A and B the most common in insects. We identified and assembled 110 complete *Wolbachia* genomes from 93 host species spanning 92 families by filtering data from 368 insect species generated by the Darwin Tree of Life project. From 15 infected species we assembled more than one *Wolbachia* genome, including cases where individuals carried simultaneous supergroup A and B infections. Different insect orders had distinct patterns of infection, with Lepidopteran hosts mostly infected with supergroup B, while infections in Diptera and Hymenoptera were dominated by A-type *Wolbachia*. Other than these large-scale order-level associations, host and *Wolbachia* phylogenies revealed no (or very limited) cophylogeny. This points to the occurrence of frequent host switching events, including between insect orders, in the evolutionary history of the *Wolbachia* pandemic. While supergroup A and B genomes had distinct GC% and GC skew, and B genomes had a larger core gene set and tended to be longer, it was the abundance of active and pseudogenised copies of bacteriophage WO who was a strong determinant of *Wolbachia* genome size. Mining raw genome data generated for reference genome assemblies is a robust way of identifying and analysing cobiont genomes and giving greater ecological context for their hosts.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

The natural world is a complex web of interactions between living species. These interactions can be mutualistic, commensal, pathogenic, parasitic, predatory or inconsequential, but each individual lives alongside a rich diversity of cobionts. Most eukaryotes associate intimately with a specific microbiota, and are commonly infected by a range of microbial and other pathogens. For some microbial associates the distinction between mutualism and pathogenicity or parasitism is fuzzy. For example, *Wolbachia* (Proteobacteria; Alphaproteobacteria; Rickettsiales; Anaplasmataceae; Wolbachieae) are found living intracellularly in a range of terrestrial arthropods and nematodes. No free-living *Wolbachia* are known: the association is essential for their survival. In contrast, infection with *Wolbachia* can be beneficial to hosts, but is not usually essential.

Wolbachia were first identified as mosquito endobacteria that were maternally transmitted, through the oocyte, and that induced a range of reproductive manipulations on their hosts[1,2]. The most common manipulation by *Wolbachia* is to induce cytoplasmic incompatibility (CI). Under CI, infected females are able to mate productively with all males, but uninfected females are only able to mate with uninfected males (as mating with CI-inducing *Wolbacha*-infected males results in zygotic death). This asymmetry in fitness can drive spread of the CI-inducing *Wolbachia*. Other reproductive manipulations include feminisation of genetic males[3], male killing[4] and induction of parthenogenesis in females[5]. All these manipulations promote the transmission of infected oocytes to the next host generation, and thus boost the spread of *Wolbachia*. In most species that can be infected, populations are a mix of infected and infection-free individuals, and hosts can evolve to resist infection[6,7]. While *Wolbachia* are often described as reproductive parasites, association with *Wolbachia* can sometimes have beneficial effects, providing nutritional supplementation to phloem-feeding hemiptera[8], and enhancing host immunity to viruses and *Plasmodium* parasites[9]. Indeed the host immunity-boosting phenotype may explain the initial spread of *Wolbachia* in previously uninfected populations. In nematodes, elimination of *Wolbachia* induces host sterility, and antibiotic treatment is an effective addition to pharmacological treatment of human-infecting, *Wolbachia*-positive filarial nematodes[10].

Wolbachia infection of terrestrial arthropods is very common, with nearly half of all insect species predicted to be infected[11]. *Wolbachia* can be classified using molecular phylogenetic analyses into a series of supergroups[12,13]. Supergroups C, D and J are found only in filarial nematodes, supergroups E and F are found in both nematodes and insects, and supergroups A, B and S (and others for which full genome data are not available) are found only in arthropods. Supergroups A and B are the most common *Wolbachia* found in terrestrial insects.

Analysis of *Wolbachia* biology has been expanded by the determination of genome sequences for many isolates. The genome sequences for *Wolbachia* from over 90 host species are publicly available, and mining of host genomic raw sequence data identified a large number of additional partial genomes[14]. This understanding, that cobiont genomes can be assembled from the “contamination” present in the data generated for a target host, has been especially useful for the unculturable *Wolbachia*. We now have the opportunity to survey for the presence of *Wolbachia* genomes at an unprecedented scale, as the Darwin Tree of Life (DToL) project aims to sequence all described terrestrial and aquatic eukaryotic species found in Britain and Ireland[15]. This project is using high-accuracy long read and chromatin conformation long range

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sequencing to generate and release publicly available chromosomal genome assemblies, meeting exact standards of contiguity and completeness, for thousands of protists, fungi, plants and animals. Several hundred terrestrial arthropod assemblies are already available (https://portal.darwintreeoflife.org).

The DToL project sequences genomes from individual, wild-caught specimens of target species, and thus will also generate data for the cobiome present in each specimen at the time of sampling. Where possible, DToL processing usually avoids body parts or tissues that are expected to have a high relative mass of cobionts. In smaller-bodied species, where the whole organism is extracted, and in cases where *Wolbachia* disseminates widely within an organism it is inevitable that cobiont genomes will be sequenced alongside the host genome.

Using k-mer classification tools, it is possible to efficiently and correctly separate out cobiont data from that of the host, and to deliver clean host assemblies[16–18]. The cobiont data are then available for independent assembly and analysis. Here we present a survey of the first 368 terrestrial arthropod genome datasets produced in DToL for the presence of *Wolbachia*, and assemble over 100 new *Wolbachia* genomes. We use these to explore patterns and processes in bacterial genome evolution and coevolution of *Wolbachia* with its hosts and with its own bacteriophage parasites. Lepidopteran hosts were mostly infected with supergroup B, while infections in Diptera and Hymenoptera were mainly caused by A-type *Wolbachia*. However, host and *Wolbachia* phylogenies revealed no (or very limited) cophylogeny. We show that while B genomes tended to be longer compared to supergroup A, genome size in *Wolbachia* is correlated with the level of integration of its double-stranded bacteriophage WO.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Screening a diverse set of insect genome data for *Wolbachia* infections

We screened raw genomic sequence data and primary assemblies for 368 insect species (204 Lepidoptera, 61 Diptera, 52 Hymenoptera, 24 Coleoptera, 9 Hemiptera, 5 Trichoptera, 4 Orthoptera, 3 Ephemeroptera, 3 Plecoptera, 2 Odonata and 1 Neuroptera) generated by DToL for the presence of *Wolbachia* (Table S1) using the small subunit ribosomal RNA (SSU rRNA) as a marker gene. *Wolbachia* SSU sequences were detected in 111 (30%) of the species. This degree of infection is similar to previous estimates (ranging from 22%[19,20] to 40%[11] of all arthropods). While the DToL project aims to sequence eukaryotes from across Britain and Ireland, 82% of the samples screened were sampled from the Wytham Woods Ecological Observatory, Oxfordshire (https://www.wythamwoods.ox.ac.uk/)[21]. No correlation between sampling location and incidence level was detected, with 29% of all samples collected in Wytham Woods being *Wolbachia* positive, reflective of the overall incidence level (Figure S1).

Wolbachia prevalence and infection intensity varies between species and between populations within a species[22,23]. As only one individual was analysed for each taxon screened, the true level of infection within the insect biota surveyed by DToL is likely much higher. Incidence was lower in Coleoptera (4/24, 17%) compared to Lepidoptera (55/204, 27%), Diptera (21/61, 34%) and Hymenoptera (23/52, 44%) (Figure 1A). We observed an equal prevalence of infection in samples identified as female (39/138, 28%) and male (45/153, 29%) (Figure 1B).

The DToL species were sequenced using PacBio Sequel II HiFi highly accurate long read platform, generating consensus raw reads of 10-20 kb with base level accuracy of >99% (~Q30-40). These long, accurate reads are ideal for assembly, particularly for bacterial genomes where the information content per base is higher than in repeat-rich eukaryotes. The average sequence length of HiFi reads identified as being derived from *Wolbachia* was 12 kb, indistinguishable from host HiFi reads. We separated and assembled all *Wolbachia* reads in each positive sample and screened these assemblies to identify complete genomes. We generated 110 complete genomes, from 93 species, of which 77 were circular (Table S2). The average completeness of these genomes, assessed using BUSCO, was 99.3%, with a mean duplication level of 0.37%. The mean genome size of the new genomes was 1.47 Mb, which is significantly larger than the average genome size of public *Wolbachia* genomes (1.32 Mb; Wilcoxon rank sum test, p-value = 4.576×10^{-9}) (Figure S2). This is likely because it is possible to assemble across repeated loci (such as integrated *Wolbachia* phage) with the long, accurate HiFi reads. The mean number of contigs generated for the 33 genomes that could not be circularised was 2.12 (ranging from 1 to 6).

The dataset includes the first *Wolbachia* genomes assembled from two insect orders, Odonata (dragonflies and damselflies) and Orthoptera (grasshoppers and crickets). Both species of dragonfly surveyed (Odonata) harboured *Wolbachia* (Figure 1A). The largest circular *Wolbachia* genome generated, 2.19 Mb, was isolated from the blue-tailed damselfly. This is the longest complete *Wolbachia* genome yet reported (Figure 5A). Although in most samples infection by only a single *Wolbachia* strain was detected, 15 of 93 specimens (16%) were

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

infected with at least two *Wolbachia* genomes. Within *Phalera bucephala* (Lepidoptera) and *Lasioglossum morio* (Hymenoptera) three genomes were assembled, while all other co-infections involved two strains.

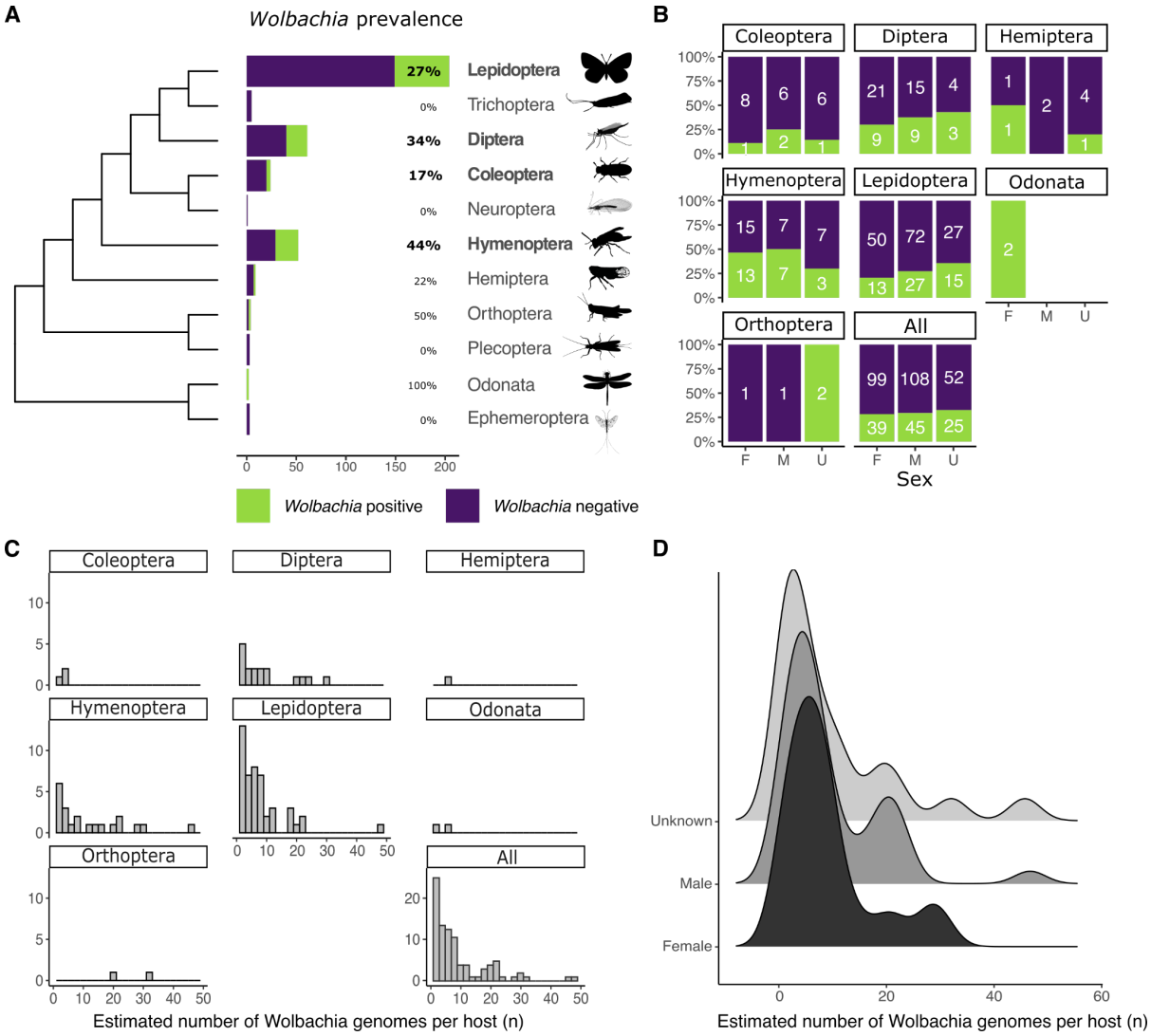
Having chromosomally-complete insect host genomes, as well as complete *Wolbachia* allows for the estimation of the relative numbers of *Wolbachia* genomes per host genome. Most infected hosts tightly control *Wolbachia* proliferation and have a relative abundance below ten *Wolbachia* genomes per host nuclear genome. Particularly high abundances were observed in *Thymelicus sylvestris* (48 *Wolbachia per* host) and *Athalia cordata* (47) (Table S2) (Figure 1C). The mean relative abundance in different taxonomic orders lay between 3 and 12, except for the two crickets (Orthoptera), *Chorthippus brunneus* and *Chorthippus parallelus*, which have a 33 and 20 *Wolbachia* genome copies per host genome, respectively (Figure 1C). No significant difference was observed between relative *Wolbachia* abundance and sex of the host (Figure 1D), with both male and female having a mean between nine and ten copies.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1: Prevalence and relative abundance of *Wolbachia* in DToL insect genomes.

A, B Prevalence of *Wolbachia* in insect hosts, split by taxonomic order (A) and by sex (B). The cladogram of insect ordinal relationships is based on Misof et al[24]. Orders with more than 10 analysed species are shown in bold. Silhouettes are from PhyloPic (http://phylopic.org/). Sex of insects was classified as F (female), M (male) or U (unknown, where not recorded on collection).

C, D The estimated number of *Wolbachia* genomes per copy of the host nuclear genome split by taxonomic order (C) and by sex (D).



bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Wolbachia phylogeny suggests frequent host switching events

We selected 93 high-contiguity and high-completeness *Wolbachia* genomes from the public INSDC databases, including genomes from *Wolbachia* infecting Nematoda (13 genomes), Arachnida (4), Isopoda (1) and several orders of Hexapoda (75) (Table S3). Adding the 110 newly assembled genomes yielded a dataset of over 200 high-quality assemblies. We annotated all protein-coding genes in those genomes using Prodigal[25], and clustered the predicted protein sets into orthologous groups using OrthoFinder2[26]. The resulting 634 near-single copy genes were used to infer a phylogeny of *Wolbachia* (Figure 2A, Figure S3). From this phylogeny we assigned each genome to the previously defined *Wolbachia* supergroups[12,13]. All newly assembled *Wolbachia* genomes belonged to either supergroup A or B. While Lepidoptera were predominantly infected with supergroup B *Wolbachia* (42/53, 80%), *Wolbachia* supergroup A was most frequent in all other insect classes (46/57, 81%). It has been previously observed that supergroup B is the most common *Wolbachia* type in Lepidoptera[19,27–29]. Of the 15 species where co-infections occurred, *Endotricha flammealis*, *Phalera bucephala*, *Philonthus cognatus*, *Protocalliphora azurea* and *Sphaerophoria taeniata* were co-infected with strains from both A and B supergroups, and the other ten co-infections were of distinct strains within the same supergroup (Table S2).

Wolbachia generally do not show strict cophylogeny with their hosts[7,23]. This pattern was also observed when comparing host and *Wolbachia* phylogenies for the supergroup A and B genomes (Figure 2B). Closely related insect species may be infected by dissimilar *Wolbachia* strains and conversely, closely related *Wolbachia* can infect a diverse set of insects. For example, the *Wolbachia* strains infecting the hoverfly *Eupeodes latifasciatus* and four Lepidoptera (*Pararge aegeria*, *Celastrina argiolus*, *Hylaea fasciara*, *Watsonella binaria*) (Figure 2C) share over 99% nucleotide identity. Because most of our new samples came from a single site (Wytham Woods Genomic Observatory) we were also able to explore the horizontal transfer of *Wolbachia* between hosts in a local context. Wytham Woods-derived *Wolbachia* were no more likely to be related than any other *Wolbachia* subset (Figure S4).

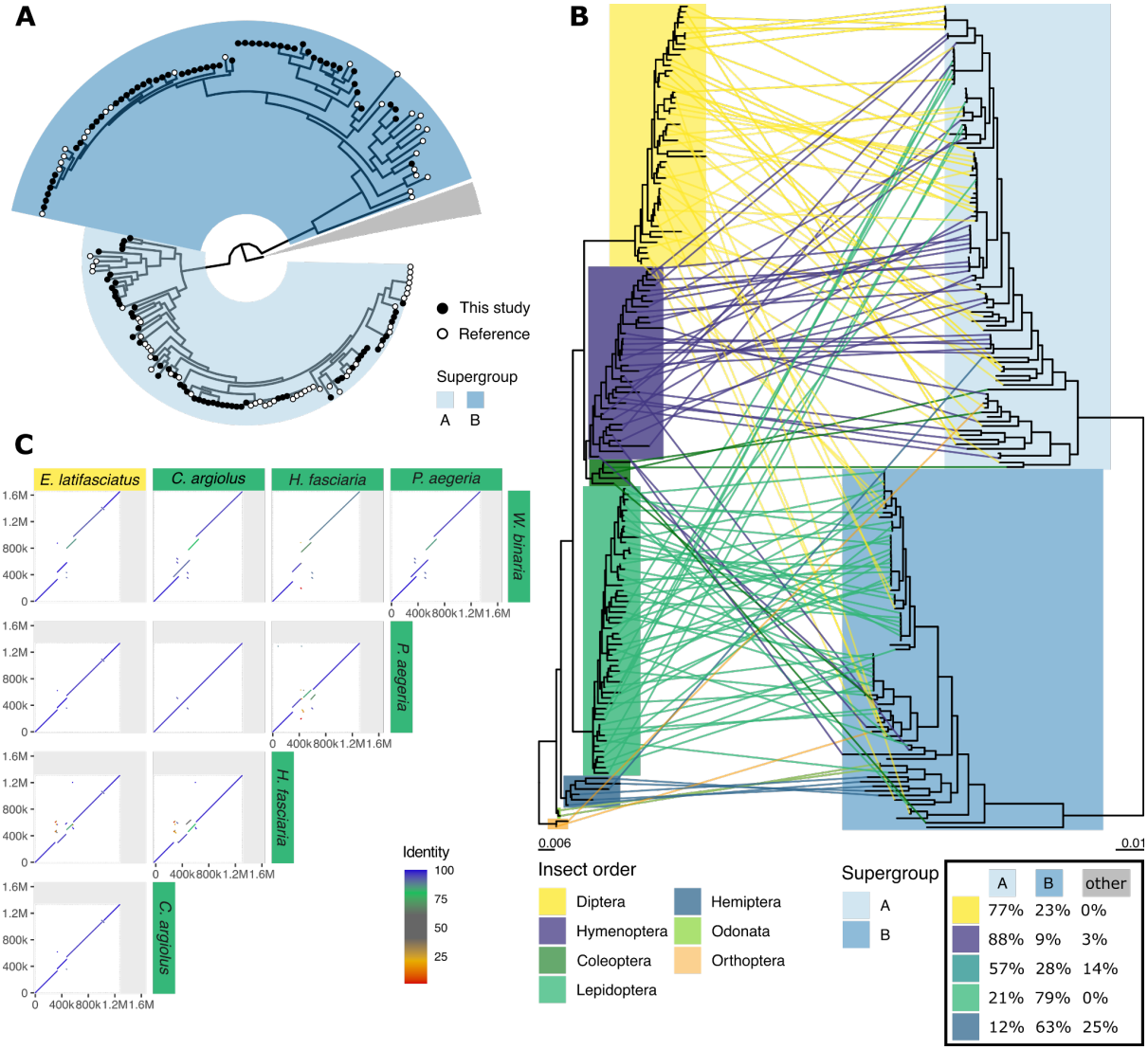
bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2: *Wolbachia* DToL genomes expand known phylogeny.

A Circular phylogeny of supergroup A and B *Wolbachia*, visualised with the root placed between the A and B supergroups and the remaining supergroups (C,D,E,F, J, S; nodes collapsed as grey wedge), highlighting newly sequenced genomes (black tip labels) and genomes from public databases (white).

B Incongruence between host topology (left) and supergroup A and B *Wolbachia* topology (right) is shown as a tanglegram. Overview of the supergroups infecting diverse insect orders is given in a table (inset, bottom right).

C Example of a host switching event, where the *Wolbachia* of the hoverfly *Eupeodes latifasciatus* has high nuclear sequence identity and genome colinearity to four *Wolbachia* genomes assembled from Lepidoptera.



bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Intrinsic properties of *Wolbachia* distinguish supergroups

The completeness of the new genomes, and in particular the circular assemblies achieved for 77 of them, permits analyses of genome properties that are not possible with fragmented and partial genomes. All circularised genomes, including those from public databases, were rotated to start at the presumed origin of replication. The average pairwise whole genome nucleotide identity between all *Wolbachia* genomes ranged between 77.3% and 100.0%, with at least 92.8% and 93.5% identity within supergroup A and B, respectively (Figure 3A). The number of breakpoints interrupting pairwise whole-genome alignments was counted, normalised for the total alignable length, and compared to average nucleotide identity (ANI) of the compared genomes (Figure 3A). A significant correlation was observed between nucleotide divergence and the number of breakpoints in supergroups A (0.90, $p < 2.2e^{-16}$, Spearman correlation) and B (0.69, $p < 2.2e^{-16}$, Spearman correlation) (Figure 3A).

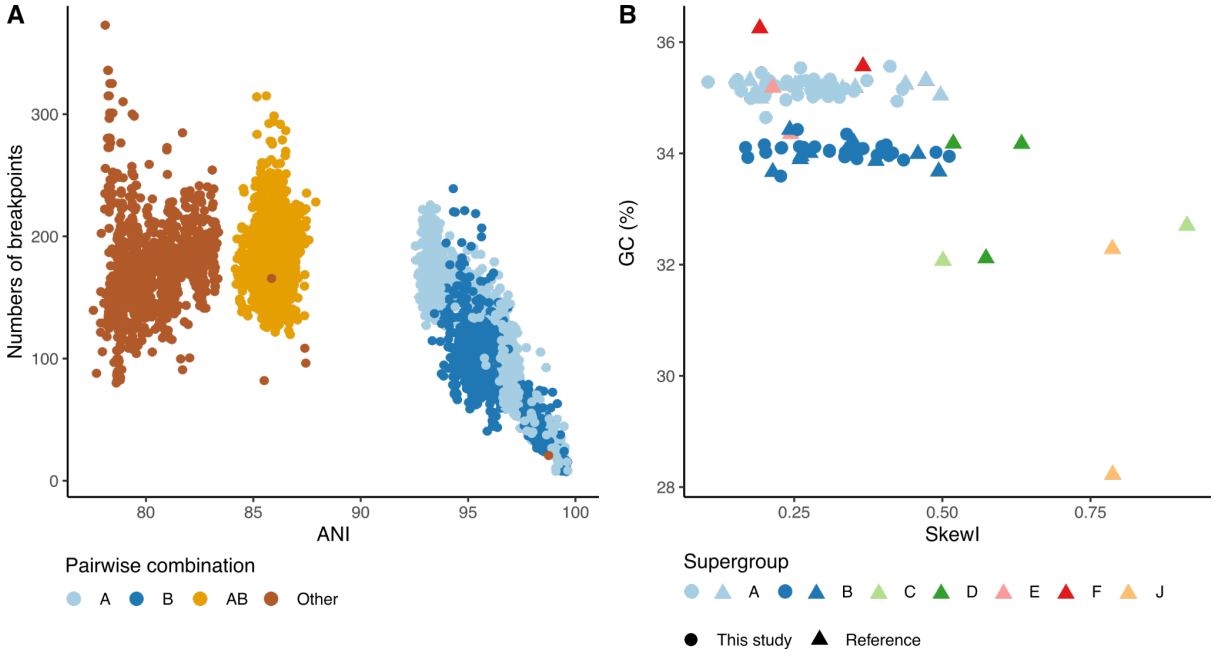
GC skew accumulates in stable bacterial genomes through differential mutation pressures on leading versus lagging strands. Genomes that have undergone frequent rearrangement are expected to have lower overall GC skew, which can be summarised across the genome as a single metric, SkewI[30]. Genomes from supergroups A and B had distinct GC contents (Figure 3B), with supergroup A having a higher mean GC (35.2%, standard deviation 0.15%), compared to B (34.0%, standard deviation 0.16%) (two-sample t-test p-value $< 2.2e^{-16}$). Genomes from other supergroups had distinct GC content, often very different from A and B genomes, but as so few examples have been sequenced, general patterns are not discernible. In both A and B supergroups SkewI values were relatively low, but genomes from *Wolbachia* from nematode hosts (C, D, J) had higher SkewI values (Figure 3B). A high degree of GC skew had already been observed in the *Wolbachia* strain infecting *Dirofilaria immitis* (supergroup C)[31]. This suggests that nematode-associated *Wolbachia* have retained chromosome stability across a long timeframe, not observed in supergroups A and B, also evident by their large number of re-arrangements (Figure 3A).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3: Comparative genomics of *Wolbachia*.

A Average nucleotide identity (ANI) plotted against the number of breakpoints in comparisons within A supergroup genomes, within B, between A and B and between other supergroup *Wolbachia*.

B Index of skewness compared to GC content for all circularised *Wolbachia* genomes.



bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Conservation and diversity in gene content of *Wolbachia*

Wolbachia, because they are sheltered within the cells of their hosts, may be relatively isolated from other bacteria, and thus have somewhat closed pan-genomes. One route to acquisition and sharing of new genes is through the *Wolbachia* phage (WO phage), which alongside the essential phage particle structural genes carry a cargo of genes that have been implicated in host manipulation. We re-annotated all 203 *Wolbachia* with the same, standard gene finding toolkit, Prodigal, to normalise annotations. While this may have lost careful manual revision in previously determined gene sets, it avoids issues of data incompatibility. Gene number correlated with genome size, and the average gene number in the newly assembled set of supergroup A and B *Wolbachia* was larger than in A and B genomes from the public databases (Figure S5). Comparing all genomes, the mean number of predicted genes was larger in supergroup B (1467) compared to A (1385).

We used OrthoFinder with default settings to define clusters of orthologous proteins across all *Wolbachia* genomes. Each genome contained between 0 and 184 novel, strain-specific genes (average 19). These novel genes were shorter than all genes (average gene length overall was 875 nucleotides or ~290 amino acids, while novel genes averaged 434 nucleotides or ~145 amino acids). As expected, supergroups which were not well represented often contained more strain-specific genes. For example, wCfeT from supergroup E (which infects cat fleas, *Ctenocephalides felis*) uniquely encoded genes for pantothenate (panC-panG-panD-panB)[32] and thiamine (thiG-thiC) biosynthesis. Nonetheless, out of the ten genomes with most strain-specific genes, seven belonged to either supergroup A or B. These novel genes were not preferentially associated with WO phage regions (Figure S6) but the majority (78%) had annotations that associated them with transposon and mobile element function. This suggests that much of the novelty arose through mobile elements other than WO phage. Other than clusters with one or two members, the most frequently observed cluster sizes were 203±2. These clusters contained the single-copy (and near-single-copy) orthologs deployed in phylogenetic analyses (Figure 4A). Overall, the majority of the proteins encoded in the *Wolbachia* genomes were members of orthology clusters that were present in at least 95% of all strains.

The abundant sampling of supergroup A and B genomes allowed us to address and compare the sizes of the core- and pan-proteomes of these groups. The larger genome and proteome size found in supergroup B was reflected in a larger core proteome (Figure 4B), but supergroup A had a larger pan-proteome (Figure 4B). While the core proteomes differed, very few of the protein families that were part of each supergroup's core proteome were unique to that supergroup. One supergroup-restricted set of protein families was found to comprise the operon for arginine transport (ArtM, ArtQ and ArtP and the repressor of arginine degradation ArgR)[33], which was uniquely detected and conserved in supergroup A (present in 83/103 or 80% of all *Wolbachia* A genomes). Although the periplasmic arginine-specific binding protein (ArtI or ArtJ) was not detected, the presence of this ATP-binding cassette-type (ABC) transporter suggests that these *Wolbachia* are acquiring arginine from their hosts.

The operon producing biotin (vitamin B7)[34] was detected in seven of the 110 new genomes, all belonging to supergroup A (Figure 4C). One derived from *Icerya purchasi* (Hemiptera) and six were from Hymenoptera (two from *Lasioglossus malacharum*, which carried two strains, and single strains from three *Andrena* and a *Nomada* species). The biotin synthesis cluster has been described previously from a restricted but diverse set of supergroups, including two

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A genomes from additional *Nomada* bee hosts. This distribution suggests possible ecological linkage[35], as *Andrena* bees are kleptoparasitized by *Nomada* cuckoo bees and phylogenetic analyses of both the biotin gene clusters and the *Wolbachia* core proteomes show close relationships between this cluster of genomes (Figure S7). The gene cluster is strongly conserved in physical organisation of all six necessary genes (bioA-D,F,H). In the genomic region immediately surrounding the operon we identified recombinase and transposase genes, as well as ankyrin repeat containing genes and toxin-antitoxin cytoplasmic incompatibility Cin gene pairs. In three genomes (from *Andrena dorsata, Nomada fabricium* and one of the *L. malacharum* strains) the operon was independently disrupted by transposases. The region containing the biotin operon thus has the hallmarks of a “virulence island” that may be mobile between genomes, and may have accrued additional genes (ankyrin, Cin) that hitchhike with the biotin operon.

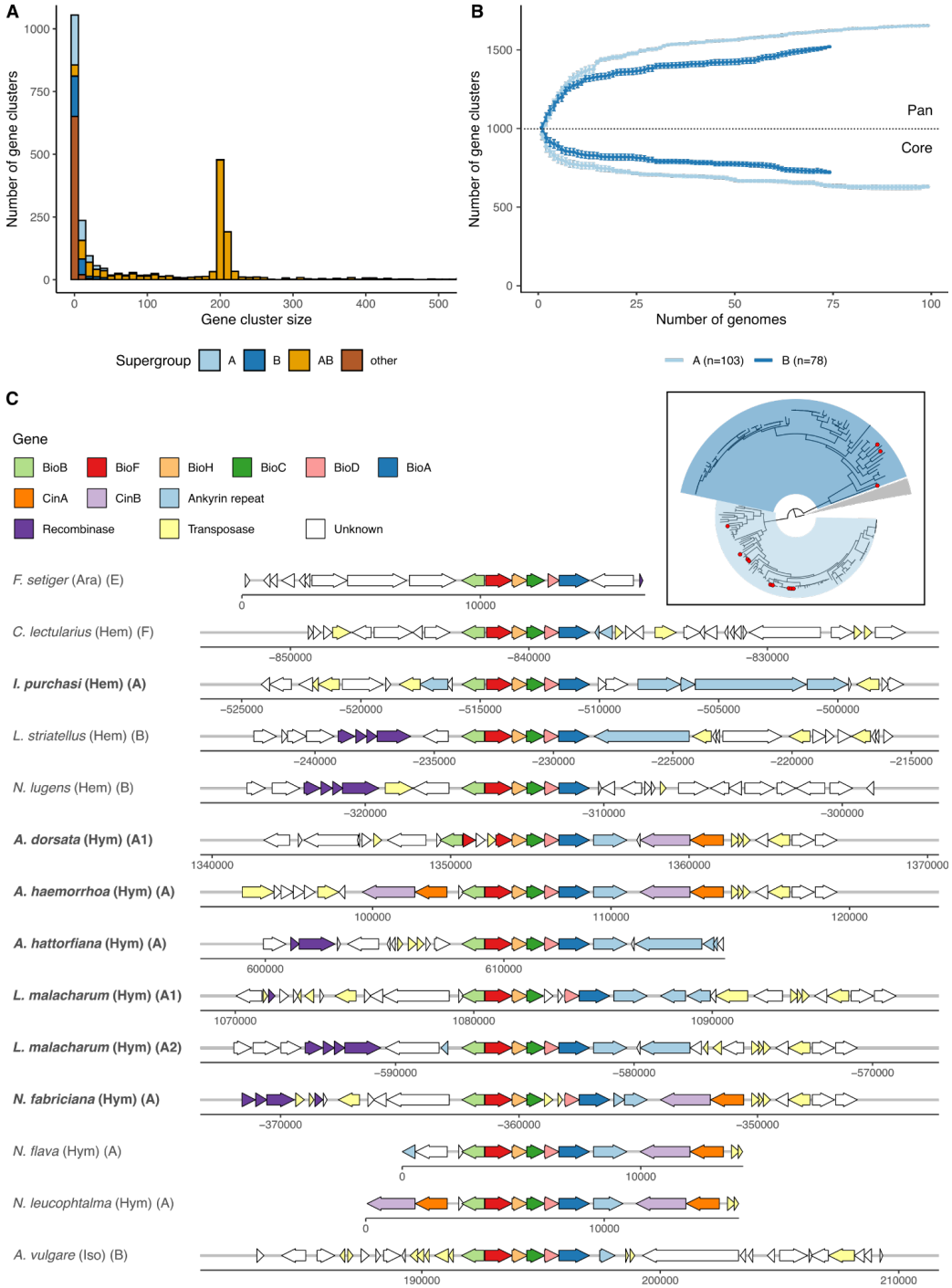
bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4: Exploration of *Wolbachia* protein-coding gene diversity.

A Histogram of protein family size per supergroup.

B Rarefaction analysis of pan- and core proteomes of supergroups A and B, based on 500,000 random addition-order permutations of co-occurring orthogroups excluding novel genes.

C Synteny of the biotin cluster shows conserved gene order and punctuated pattern of species presence (inset, species with biotin cluster present are highlighted with red circles).



bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

WO prophage insertions expand genome size

Wolbachia can itself be infected by double-stranded DNA temperate bacteriophages, WO phage, which can integrate in the genome of its host as a prophage. Four modules are necessary for construction and function of phage particles during the lytic stage: head, baseplate, tail and fibre, and inserted and pseudogenised *WO* phage can be identified and discriminated based on the presence and completeness of these components. Regions of a *Wolbachia* genome flanked by WO phage modules are likely to form components that are transduced by the phage during infection of new cells, “cargo” loci that form the Eukaryotic Association Module (EAM)[36,37]. All the *Wolbachia* genomes were screened for prophage regions using essential module genes from previously annotated WO insertions (Table S4). Prophage regions were deemed complete when all four modules were observed with at least 80% of genes of each module present. An abundance of putative intact and pseudogenised WO phage were identified. For example the supergroup B *Wolbachia* from *Ischneura elegans* (the bluetail damselfly; the largest *Wolbachia* genome assembled) contained three putative intact prophage and nine WO phage fragments (Figure 5A) summing to 0.8 Mb of the genome.

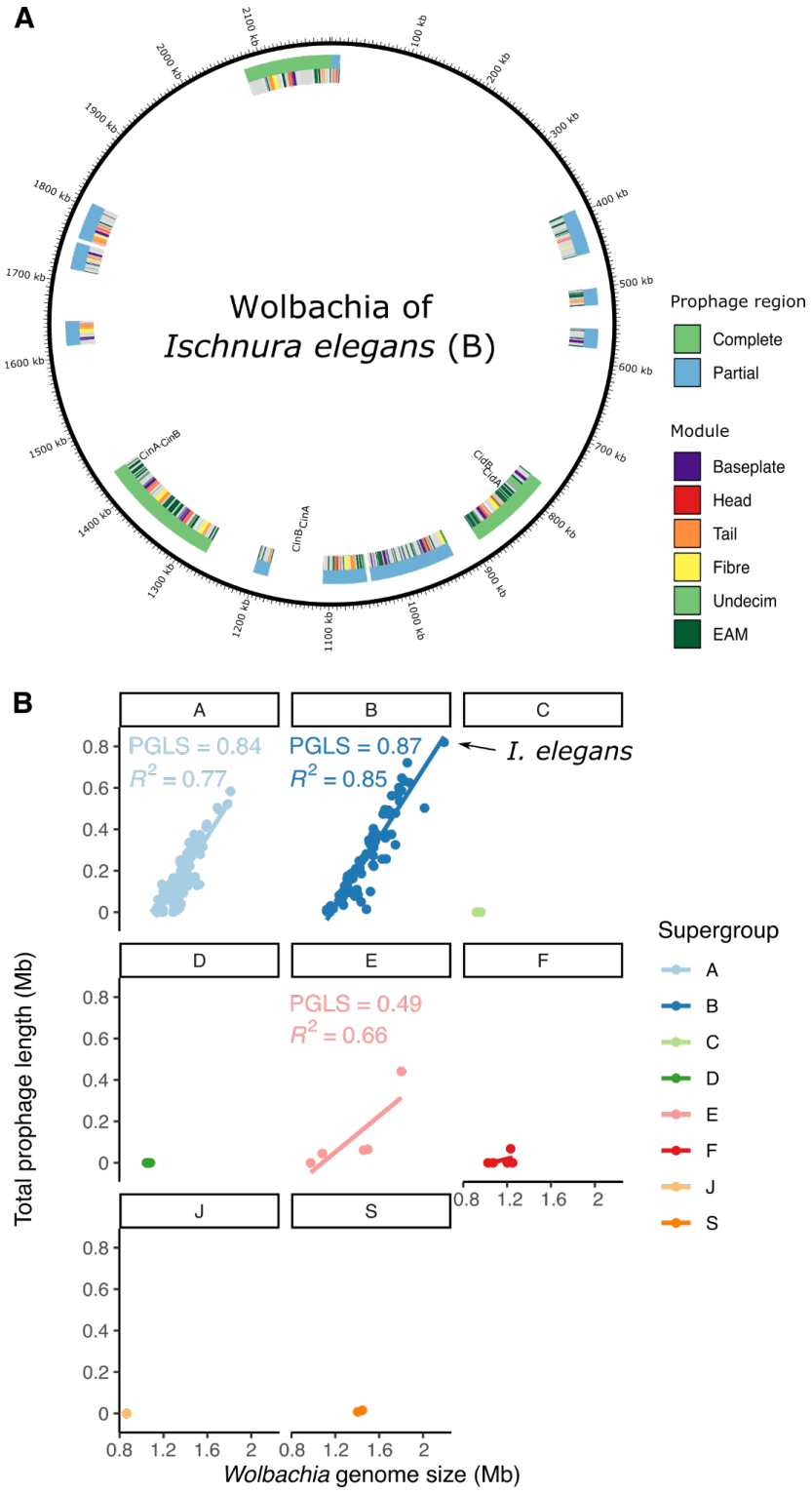
The fraction of total prophage region in each genome ranged from 0-38%. Nematode-associated *Wolbachia* typically are not infected by WO phage[38] and no prophage regions were detected in genomes of supergroup C, D, J and nematode-infecting F (Figure 5B). A significant correlation was found between genome size and WO prophage span in supergroups A and B (Figure 5B). This association was robust to correction for phylogenetic relatedness of the genomes (model fit increased to 0.84 and 0.87 respectively with p-values $<10^{-16}$).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 5: WO prophage in *Wolbachia*

A Annotation of the WO prophage integrated in the genome of the *Wolbachia* strain infecting *Ischnura elegans.*

B *Wolbachia* genome size is strongly correlated with integrated prophage span in supergroups with WO phage association. Phylogenetic Generalised Least Squares (PGLS) analyses were performed to assess the correlation between prophage length and genome size in a phylogenetically-aware manner.



bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Toxins are often associated with mobile elements

We identified several potential cargo genes within intact and fragmented prophage. These included transposases and integrases associated with mobile elements, and other loci previously associated with eukaryotic manipulation, such as cytoplasmic incompatibility loci and ankyrin repeat containing genes, as expected from the EAM model[36,37].

Wolbachia produces a suite of toxins[39] which can have dramatic effects on their hosts, such as cytoplasmic incompatibility (CI). The CI phenotype is caused by two adjacent genes, CifA and CifB, which function as a toxin-antitoxin pair[40,41]. Phylogenetic analysis classified most *Wolbachia* Cif gene pairs into four types (I-IV)[42]. A fifth type (V) is much more variable in structure. The toxin component can have nuclease activity (in which case the gene pair is frequently referred to as CinA-CinB), deubiquitinase (CidA-CidB) or both (CndA,CndB)[43]. All type II, III and IV pairs have nuclease domains, while all type I have deubiquitinase and most have nuclease[42]. Two hundred and sixty one full-length and likely functional Cif pairs were detected in 133 of the 181 (73%) supergroup A and B genomes. One Cif pair was detected in most genomes, but many had several, with seven copies in the *Wolbachia* strain infecting the holly tortrix moth (*Rhopobota naevana*). Most of the gene pairs (93) contained a deubiquitinase domain (type I, Cid), while the other four types occurred in roughly equal proportions (II: 40, III: 43, IV:35 and V:50). Many pairs (177/261; 69%) were located in the predicted EAM of the prophage.

Loci encoding additional toxins such as RelE/RelB, Spaid-like and latrotoxin were identified in multiple *Wolbachia* genomes, frequently in prophage regions (71/586 [12%], 136/597 [23%] and 227/382 [59%] genes, respectively). The Tc pore-forming toxin complex, which consists of two genes TcA and TcB-C, was detected in a limited number of A and B supergroup genomes and also showed a predisposition to occur within prophage (13/69 [19%] and 9/35 [26%], respectively). Additional toxin-encoding loci had limited presence in different subgroups and were not associated with prophage regions. ParD/ParE only occurred in supergroups A, B and E, and FIC only occured in supergroups A, E, F and S. The type IV toxin-antitoxin gene pair AbiEii/AbiGii-AbiEi, which protects against the spread of phage infection[44], was only detected in two genomes in supergroup E. It is noteworthy that these two genomes had very low levels of prophage-derived DNA (4.3% of their genome span).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Isolation of cobiont genomes, and specifically *Wolbachia* genomes, from shotgun high-throughput sequencing data has been established for many years[45]. In the field of prokaryotic and eukaryotic microbial metagenomics, metagenome-assembled genomes (MAGs) are likely to be the only way to access many unculturable microbial genomes, even if the species they derive from are hyperabundant[46,47]. The abundance of raw sequencing data in the International Nucleotide Sequence Database Collaboration (INSDC) databases has been an attractive prospecting ground for microbial associates of eukaryotic target species. To date, most raw data available for such searches have been short reads from Illumina and other platforms. These reads are too short to partition efficiently into bins corresponding to putative distinct genomes. Preliminary assembly of such datasets is more likely to be able to separate cobionts from target genomes. These approaches have been applied to hunt for *Wolbachia* with a recent *tour de force* generating nearly 1,200 *Wolbachia* MAGs from publicly available data[14]. However, these MAGs suffer from the expected issues of low completeness (due to low effective coverage), fragmentation (due to coverage and sequence repeat issues), undetected contamination and inability to distinguish co-infecting strains. Moreover, the biased nature of public data meant that these derived from only 37 different host species.

We generated 110 *Wolbachia* assemblies from 368 terrestrial arthropod HiFi datasets, and 77 of these were fully circular genome assemblies. The genomes were uniformly of high completeness (Figure S2). Due to the high intrinsic base quality of HiFi reads (Q30 to Q40; from one error in 1000 to one error in 10,000) we were able to distinguish insertions of *Wolbachia* DNA into the host genome from true components of the *Wolbachia* genome, and to independently assemble even closely related strains with confidence. As we were screening raw data from a biodiversity genomics programme that aims to sample a wide phylogenetic diversity of hosts, the new *Wolbachia* genomes presented here more than double the number of different host species from which *Wolbachia* genomes have been assembled. The assembled genomes include the first representatives isolated from Odonata (damselflies) and Orthoptera (crickets). In 16 additional datasets we identified likely *Wolbachia* content but were not able to produce credible genome assemblies (see Supplemental Data, Table S2). This was usually because the *Wolbachia* sequence was present in very low effective coverage (~ threefold) but in some samples no credible assembly was generated despite high coverage. These datasets may contain multiple recombining strains, or contain large insertions in the host genome and deserve further exploration.

The distribution of *Wolbachia* in insect hosts is a function of the balance between co-speciation (vertical transmission of *Wolbachia* among daughters of the host species) and horizontal transmission where strains move between species. Transmission among insect hosts was the dominant pattern underpinning *Wolbachia* distribution, but we identified two features of the distribution, one local and one general, that are of note. Lepidoptera were more likely to be infected with supergroup B *Wolbachia* than A, and Hymenoptera, Diptera and Coleoptera were more likely to be infected with supergroup A strains. Multi locus sequence typing (MLST) has previously shown that supergroup B is the most common *Wolbachia* type in Lepidoptera[19,27–29]. This suggests some non-exclusive specialisation of *Wolbachia* on their hosts, which may be driven by *Wolbachia* genetics, host genetics or (less likely) a distinct set of ecological transmission routes in each insect group. Many of our genomes derived from insects were collected at one site, the Wytham Woods Genomic Observatory (Figure S1) but this subset

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was no more closely related than other genomes from widely separated sites (Figure S5). It is likely that the mobility of hosts, including through seasonal migration, means that sampling from one geographical site is a valid approximation of more global sampling.

Close ecological association between host species may promote sharing of *Wolbachia* isolates and localised genetic exchange, for example within predator-prey systems. The close similarity of *Wolbachia* genomes from *Andrena* solitary bees and their *Nomada* cuckoo bee kleptoparasites, and the shared occurrence of the biotin synthesis operon (Figure 4C) may be a case of transmission within an ecological network. The presence of the biotin operon in *Wolbachia* of insects that largely or solely feed on low-protein plant fluids (nectar or phloem) suggests that the *Wolbachia* may be offering nutritional support to their hosts[48], and thus that this cluster of genomes may have been positively selected for their mutualist tendencies. Other genes whose distribution among isolates is driven by horizontal gene transfer, including by mobile elements such as phage, might be expected to have a distribution that is not explained by overall genome relatedness, and might reflect ecological association. We note that previous work has suggested that horizontal transmission rather than cospeciation may also explain closely related *Wolbachia* in closely related taxa. For example, genomic divergence between closely related *Wolbachia* in sister *Drosophila* species was too low to be the product of independent evolution since the last common ancestor of the flies[49,50].

Wolbachia can promote reproductive success of their hosts[1,2], and thus their own Darwinian fitness, through reproductive manipulations such as CI. The loci underpinning CI are a diverse set of toxin-antitoxin gene pairs. Our survey of *Wolbachia* identified many additional CI gene pairs, mainly of the I Cid type and mostly associated with WO phage. Many genomes had more than one toxin-antitoxin pair, and some individual hosts were infected with multiple *Wolbachia* strains carrying different CI gene pairs. These CI genes likely mediate conflict between *Wolbachia* strains and the ecosystem of toxin deposition and rescue in individual zygotes must be complex. Interestingly we identified CI gene pairs next to 5 of the 14 biotin synthesis operons, suggesting that the mobile elements that transduce this presumably mutualist physiology are also engaged in CI conflict.

One striking feature of the genomes assembled from the HiFi reads was that their average span was ~10% greater than the average size of previously-assembled *Wolbachia* genomes. As we also observed a correlation between content of WO phage in the genome and genome size (Figure 5B), we speculate that the lower average size of previous assemblies may be because the presence of near-identical segments of phage and other mobile elements led to collapse of repeats and artificial underestimation of true genome size. This underestimation of genome size may also have biased understanding of WO phage diversity and of the diversity of genes that can be transduced by the phage. WO phage carry genes necessary for production of phage particles and cargo genes that have been hypothesised to form an Eukaryotic Association Module (EAM)[36,37]. The increased genome size and increased resolution of WO phage copies might also mean increased gene content and diversity, and an increased set of common EAM loci. We estimated the pan-proteome of A and B supergroup strains, and found that the A supergroup had a higher pan-proteome but a smaller core proteome than supergroup B. Coupled with the observation of host-association bias between these supergroups, and other major genomic features such as GC proportion, this suggests that these divergent groups have followed very distinct evolutionary trajectories, despite evidence for transduction of loci between supergroups, and perhaps have evolved distinct physiologies and host-manipulation or -cooperation strategies. We note that the average

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

nucleotide identity (ANI) between A and B supergroup strains, and between strains from all supergroups, is relatively low (within-supergroup identity >93%, between supergroup identity <88%). This pattern of significant phylogenetic separation between supergroups suggests, as others have noted, that these supergroups have the features expected of bacterial species[33].

The Darwin Tree of Life project[15] is one of a growing constellation of biodiversity genomics initiatives worldwide that, under the banner of the Earth BioGenome Project[51], intend to "sequence life for the future of life" (https://www.earthbiogenome.org). These projects, based around ecological, regional or taxonomic lists of target species, will lay the foundations for biological research, bioindustry and conservation for the next decades. While their focus is to generate reference genomes for eukaryotic species, these projects will also yield critical resources for the study of the microbial cobionts – mutualists, pathogens, parasites and commensals – that live on and in eukaryotic organisms. Our understanding of *Wolbachia* and other common endosymbionts will thrive on a rich harvest of cobiont genomes from the tens to hundreds of thousands of host genomes that will be generated in the next decade. The assembly of 110 high-quality *Wolbachia* genomes shows the power of the long read data now being generated and the analytic approach that allowed these low complexity metagenomes to be effectively separated into their constituent parts. Analysis of these genomes revealed a propensity to infect different insect orders among supergroups, while simultaneously pinpointing to several host switching events during the course of the *Wolbachia* pandemic. Moreover, we observed that genome size in *Wolbachia* is correlated with the abundance of active and pseudogenised copies of bacteriophage WO.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Detection and assembly of *Wolbachia* genomes from DToL species data

DToL raw data are generated from whole or partial single specimens, and thus contain sequence from any cobionts in or on the specimen at the time of sampling. We screened data for 368 insect genomes generated by the Darwin Tree of Life project[15] for the presence of the intracellular endosymbiont *Wolbachia* (Table S1) using a marker gene scan approach by searching for the small subunit rRNA locus. The prokaryotic 16S rRNA alignment from RFAM (RF00177)[52] was transformed into a HMMER profile and the profile was used to screen contigs with nhmmscan[53]. We defined a positive match as having an e-value $<10^{-150}$ or an aligned length of >1000 nucleotides. Putative positive regions were extracted from the sequences, and compared to the SILVA SSU database (version 138.1)[54] using sina[55]. Matches were filtered to retain only those with >90% identity. Taxonomic classification of each positive was determined via a consensus rule of 80 percent of the top 20 best hits, using both the NCBI[56] and SILVA[57] taxonomies.

For *Wolbachia*-positive samples, all PacBio HiFi reads were analysed using kraken2[58] with a custom database consisting of a genome from a species closely related to the host, all RefSeq genomes of Anaplasmataceae and reference genomes of additionally detected cobionts downloaded using NCBI datasets and masked using dustmasker[59]. Horizontal transfer of fragments of endosymbiont and organellar DNA to the nuclear genome is a common phenomenon. To avoid inadvertently classifying nuclear *Wolbachia* insertions (NUWTs) as deriving from an independent bacterial replicon, *Wolbachia* reads identified by kraken2 were mapped to the insect genome assembly and only contigs fully covered by these reads were retained. The *Wolbachia* reads were also independently re-assembled using several assembly tools: flye (version 2.9) (flye --pacbio-hifi {reads} -o {dir} -t {threads} --asm-coverage 50 --genome-size 1.6m --scaffold)[60], hifiasm (version 0.14) (hifiasm -o {prefix} -t {threads} {reads} -D 10 -l 1 -s 0.999)[61] and hifiasm-meta (version 0.1-r022) (hifiasm_meta -o {prefix} -t {threads} {reads} -l 1)[62]. The several assemblies generated for each sample were ranked based on their completeness using BUSCO version 5.2.2[63] and the Rickettsiales_odb10 dataset, alignment to reference genomes using nucmer (version 4.0.0)[64], evenness of coverage and circularity. The best (most complete, single-contig circular preferred) assembly per sample was chosen. For samples where 10X Genomics Chromium data were available, polishing was performed using FreeBayes-called variants[65] from 10X short reads aligned with LongRanger. The host origin, span and completeness of all *Wolbachia* detected is presented in Table S2.

Collation of *Wolbachia* genome dataset, gene prediction and orthology inference

All available *Wolbachia* genomes were downloaded from NCBI GenBank on 01/02/2022, and supplemented with assemblies generated from short-read insect datasets by Scholz et al.[14]. This dataset contained replicate genomes for very closely related *Wolbachia* from the same host, and many fragmented and partial assemblies. Only the most contiguous assembly per host species was retained. These genomes were renamed using the schema

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

"R_Xyz_GenSpec_§", where Xyz is the first three letters of the insect order of the host, GenSpec is an abbreviation derived from the generic and specific epithets of the host, and § indicates the supergroup. Retained assemblies were assessed for the presence of contamination by performing a contig analysis by kraken2 using a database of only circular *Wolbachia* genomes. A list of all removed contigs can be found in Table S3. Furthermore, we only included database-sourced *Wolbachia* genomes with at least 90% BUSCO completeness[63] and at most 3% duplication with the Rickettsiales_odb10 dataset (Table S3). The exception to this filtering was the inclusion of genomes belonging to the most divergent supergroup S.

All of the publicly available and newly assembled genomes were annotated using Prodigal (version 2.6.3)[25]. Protein families were inferred using OrthoFinder (version 2.4.0)[26]. We identified 624 protein families which were single-copy in more than 95% of all *Wolbachia* genomes. These were individually aligned using mafft in automatic mode (version 7.490)[66]. Individual maximum likelihood gene trees were calculated using iqtree (version 2.1.4) (iqtree -s {alignment} -nt {threads})[67], and coalescence of these gene trees was determined using ASTRAL (version 5.7.4)[68]. The individual alignments were trimmed using trimAl (version 1.4)[69], and concatenated to form a supermatrix. This was used to infer a maximum likelihood phylogeny with iqtree using 1000 ultrafast bootstrap approximation iterations (version 2.1.4) (iqtree -s {supermatrix} -m LG+G4 -bb 1000 -nt {threads})[67]. The insect topology was subsampled from Chesters and al[70]. Incongruence in topology between the insect host and *Wolbachia*, host phylogeny was determined with ggtree in R[71].

Intrinsic genomic properties

All circular genomes were rotated to start with HemE (OG0000716) on the positive strand, as this gene is located next to the origin of replication[72]. All pairwise alignments were calculated using nucmer (version 4.0.0)[64], and breakpoints were inferred and adjusted for the aligned coverage. Average nucleotide diversity was calculated using FastANI (version 1.33)[73]. GC and GC skew index values were calculated for all genomes using SkewIT[30].

Gene content analysis

To functionally annotate predicted genes, both Prokka (version 1.14.6)[74] and InterProScan (version 5.54-87.0)[75] were run. The synteny plot of the biotin locus was created using gggenes[76]. All six genes that make up the biotin locus (BioA-D, BioF, BioH) were individually aligned with mafft in automatic mode (version 7.490)[66] and transformed into a concatenated nucleotide alignment. A phylogenetic tree was built using the model GTR+F+G4 in iqtree (version 2.1.4)[67]. Genes responsible for cytoplasmic incompatibility (CI) were identified by a BLAST search[77] using the following genes as queries: CidA: WP_010962721.1, WP_182158704.1; CidB: WP_010962722.1, WP_182158703.1 and CinA: CAQ54402.1; CinB: CAQ54403.1. Only pairs of identified neighbouring genes (e-value $1e^{-30}$, coverage 80-120%) were retained.

WO prophage analysis

A list of known prophage sequences was generated based on annotated regions described in the literature [37,40,78] (Table S4) for a set of genomes (R_Dip_DroSim_A, R_Hym_NasVit_A,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

554 R_Dip_DroAna_A, R_Dip_HaeIrr_A, R_Hym_CerSol_A and R_Hym_WiePum_A) and linked
555 to their respective gene families. Each *Wolbachia* genome was screened for continuous
556 stretches of linked prophage genes with at most five other genes in-between and these were
557 annotated as prophage regions if they contained at least one gene from one of the four core
558 phage modules (head, baseplate, tail, fibre). This permitted detection of novel prophage-
559 associated genes. Regions which contained at least 5 of 6 head, 7 of 8 baseplate, 5 of 6 fibre
560 and 5 of 6 tail module genes were deemed complete. Genomic maps of prophage integration
561 were created with circos[79]. Phylogenetic Generalised Least Squares analyses were performed
562 to assess the correlation between prophage length and genome size using the ape R
563 package[80], using a Brownian model of evolution and the phylogenetic tree in Figure 2A. R
564 squared values were calculated using the package rr2[81].

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We thank our many colleagues in the Darwin Tree of Life project – from field collectors to data curators – for the production of the raw data we analysed. We also thank Tree of Life colleagues, especially Claudia Weber, Charlotte Wright and Ellen Cameron for fruitful discussions and Andrew Varley, James Torrance and Shane McCarthy for help with sequence deposition. This research was funded by the Wellcome Trust Grants 206194 and 218328. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data availability

The raw data for each species analysed is available under BioProject PRJEB40665 (https://www.ebi.ac.uk/ena/browser/view/PRJEB40665). Darwin Tree of Life species and data are collated in the project portal at https://portal.darwintreeoflife.org.The *Wolbachia* genome assemblies are currently in progress for sequence deposition in INSDC, but can already be accessed on Zenodo (https://doi.org/10.5281/zenodo.7092419).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Yen, J. H. & Barr, A. R. New Hypothesis of the Cause of Cytoplasmic Incompatibility in Culex pipiens L. *Nature* **232**, 657–658 (1971).
2. Yen, J. H. & Barr, A. R. The etiological agent of cytoplasmic incompatibility in Culex pipiens. *Journal of Invertebrate Pathology* **22**, 242–250 (1973).
3. Bordenstein, S. R., O'Hara, F. P. & Werren, J. H. Wolbachia-induced incompatibility precedes other hybrid incompatibilities in Nasonia. *Nature* **409**, 707–710 (2001).
4. Hurst, G. D. D. *et al.* Male–killing Wolbachia in two species of insect. *Proc. R. Soc. Lond. B* **266**, 735–740 (1999).
5. Stouthamer, R., Breeuwer, J. A. J., Luck, R. F. & Werren, J. H. Molecular identification of microorganisms associated with parthenogenesis. *Nature* **361**, 66–68 (1993).
6. Hornett, E. A. *et al.* Evolution of Male-Killer Suppression in a Natural Population. *PLoS Biol* **4**, e283 (2006).
7. Werren, J. H., Baldo, L. & Clark, M. E. Wolbachia: master manipulators of invertebrate biology. *Nat Rev Microbiol* **6**, 741–751 (2008).
8. Nikoh, N. *et al.* Evolutionary origin of insect– *Wolbachia* nutritional mutualism. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10257–10262 (2014).
9. Pan, X. *et al.* The bacterium Wolbachia exploits host innate immunity to establish a symbiotic relationship with the dengue vector mosquito Aedes aegypti. *ISME J* **12**, 277–288 (2018).
10. Hoerauf, A., Mand, S., Adjei, O., Fleischer, B. & Büttner, D. W. Depletion of wolbachia endobacteria in Onchocerca volvulus by doxycycline and microfilaridermia after ivermectin treatment. *The Lancet* **357**, 1415–1416 (2001).
11. Zug, R. & Hammerstein, P. Still a Host of Hosts for Wolbachia: Analysis of Recent Data Suggests That 40% of Terrestrial Arthropod Species Are Infected. *PLoS ONE* **7**, e38544 (2012).
12. Zhou, W., Rousset, F. & O'Neill, S. Phylogeny and PCR–based classification of Wolbachia strains using wsp gene sequences. *Proc. R. Soc. Lond. B* **265**, 509–515 (1998).
13. Glowska, E., Dragun-Damian, A., Dabert, M. & Gerth, M. New Wolbachia supergroups detected in quill mites (Acari: Syringophilidae). *Infection, Genetics and Evolution* **30**, 140–146 (2015).
14. Scholz, M. *et al.* Large scale genome reconstructions illuminate Wolbachia evolution. *Nat Commun* **11**, 5235 (2020).
15. The Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences* **119**, e2115642118 (2022).
16. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics* **10**, 1361–1374 (2020).
17. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
18. Regan, T. *et al.* Characterisation of the British honey bee metagenome. *Nat Commun* **9**, 4995 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

19. West, S. A., Cook, J. M., Werren, J. H. & Godfray, H. C. J. *Wolbachia* in two insect host–parasitoid communities. *Molecular Ecology* **7**, 1457–1465 (1998).
20. Duron, O. *et al.* The diversity of reproductive parasites among arthropods: Wolbachia do not walk alone. *BMC Biol* **6**, 27 (2008).
21. Savill, P., Perrins, C., Kirby, K. & Fisher, N. *Wytham woods: Oxford's ecological laboratory*. (Oxford Univ. Press, 2011).
22. Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A. & Werren, J. H. How many species are infected with Wolbachia? – a statistical analysis of current data: Wolbachia infection rates. *FEMS Microbiology Letters* **281**, 215–220 (2008).
23. Ahmed, M. Z., Breinholt, J. W. & Kawahara, A. Y. Evidence for common horizontal transmission of Wolbachia among butterflies and moths. *BMC Evol Biol* **16**, 118 (2016).
24. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
25. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
26. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
27. Russell, J. A. *et al.* Specialization and geographic isolation among *Wolbachia* symbionts from ants and lycaenid butterflies. *Evolution* **63**, 624–640 (2009).
28. Werren, J. H. & Windsor, D. M. *Wolbachia* infection frequencies in insects: evidence of a global equilibrium? *Proc. R. Soc. Lond. B* **267**, 1277–1285 (2000).
29. Tagami, Y. & Miura, K. Distribution and prevalence of Wolbachia in Japanese populations of Lepidoptera: Wolbachia in Japanese Lepidoptera. *Insect Molecular Biology* **13**, 359–364 (2004).
30. Lu, J. & Salzberg, S. L. SkewIT: The Skew Index Test for large-scale GC Skew analysis of bacterial genomes. *PLoS Comput Biol* **16**, e1008439 (2020).
31. Comandatore, F. *et al.* Supergroup C Wolbachia, mutualist symbionts of filarial nematodes, have a distinct genome structure. *Open Biol.* **5**, 150099 (2015).
32. Mahmood, S., Nováková, E., Martinů, J., Sychra, O. & Hypša, V. *Extremely reduced supergroup F* Wolbachia *: transition to obligate insect symbionts*. http://biorxiv.org/lookup/doi/10.1101/2021.10.15.464041 (2021) doi:10.1101/2021.10.15.464041.
33. Ellegaard, K. M., Klasson, L., Näslund, K., Bourtzis, K. & Andersson, S. G. E. Comparative Genomics of Wolbachia and the Bacterial Species Concept. *PLoS Genet* **9**, e1003381 (2013).
34. Gerth, M. & Bleidorn, C. Comparative genomics provides a timeframe for Wolbachia evolution and exposes a recent biotin synthesis operon transfer. *Nat Microbiol* **2**, 16241 (2017).
35. Gerth, M., Röthe, J. & Bleidorn, C. Tracing horizontal Wolbachia movements among bees (Anthophila): a combined approach using multilocus sequence typing data and host phylogeny. *Mol Ecol* **22**, 6149–6162 (2013).
36. Bordenstein, S. R. & Bordenstein, S. R. Eukaryotic association module in phage WO genomes from Wolbachia. *Nat Commun* **7**, 13155 (2016).
37. Bordenstein, S. R. & Bordenstein, S. R. Widespread phages of endosymbionts: Phage WO

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genomics and the proposed taxonomic classification of Symbioviridae. *PLoS Genet* **18**, e1010227 (2022).

38. Gavotte, L. *et al.* A Survey of the Bacteriophage WO in the Endosymbiotic Bacteria Wolbachia. *Molecular Biology and Evolution* **24**, 427–435 (2006).
39. Massey, J. H. & Newton, I. L. G. Diversity and function of arthropod endosymbiont toxins. *Trends in Microbiology* **30**, 185–198 (2022).
40. LePage, D. P. *et al.* Prophage WO genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. *Nature* **543**, 243–247 (2017).
41. Beckmann, J. F., Ronau, J. A. & Hochstrasser, M. A Wolbachia deubiquitylating enzyme induces cytoplasmic incompatibility. *Nat Microbiol* **2**, 17007 (2017).
42. Martinez, J., Klasson, L., Welch, J. J. & Jiggins, F. M. Life and Death of Selfish Genes: Comparative Genomics Reveals the Dynamic Evolution of Cytoplasmic Incompatibility. *Molecular Biology and Evolution* **38**, 2–15 (2021).
43. Beckmann, J. F. *et al.* The Toxin–Antidote Model of Cytoplasmic Incompatibility: Genetics and Evolutionary Implications. *Trends in Genetics* **35**, 175–185 (2019).
44. Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. C. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Research* **42**, 4590–4605 (2014).
45. Kumar, S. & Blaxter, M. L. Simultaneous genome sequencing of symbionts and their hosts. *Symbiosis* **55**, 119–126 (2011).
46. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**, 105–114 (2021).
47. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
48. Ju, J.-F. *et al.* Wolbachia supplement biotin and riboflavin to enhance reproduction in planthoppers. *ISME J* **14**, 676–687 (2020).
49. Conner, W. R. *et al.* Genome comparisons indicate recent transfer of *w* Ri-like *Wolbachia* between sister species *Drosophila suzukii* and *D. subpulchrella*. *Ecol Evol* **7**, 9391–9404 (2017).
50. Turelli, M. *et al.* Rapid Global Spread of wRi-like Wolbachia across Multiple Drosophila. *Current Biology* **28**, 963-971.e8 (2018).
51. Lewin, H. A. *et al.* The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115635118 (2022).
52. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **49**, D192–D200 (2021).
53. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011).
54. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590–D596 (2013).
55. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

698 56. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and
699 tools. *Database* **2020**, baaa062 (2020).
700 57. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic
701 frameworks. *Nucl. Acids Res.* **42**, D643–D648 (2014).
702 58. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome*
703 *Biology* **20**, 257 (2019).
704 59. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST
705 implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**, 1028–1040 (2006).
706 60. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using
707 repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
708 61. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
709 assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
710 62. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads with
711 hifiasm-meta. *Nat Methods* **19**, 671–674 (2022).
712 63. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update:
713 Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for
714 Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–
715 4654 (2021).
716 64. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS*
717 *Computational Biology* **14**, e1005944 (2018).
718 65. Garrison, E. & Marth, G. *Haplotype-based variant detection from short-read sequencing*.
719 http://arxiv.org/abs/1207.3907 (2012) doi:10.48550/arXiv.1207.3907.
720 66. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:
721 Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
722 67. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
723 the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
724 68. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation.
725 *Bioinformatics* **30**, i541–i548 (2014).
726 69. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
727 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
728 70. Chesters, D. Construction of a Species-Level Tree of Life for the Insects and Utility in
729 Taxonomic Profiling. *Syst Biol* syw099 (2016) doi:10.1093/sysbio/syw099.
730 71. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing
731 Associated Data on Phylogeny Using *Ggtree*. *Molecular Biology and Evolution* **35**, 3041–3043
732 (2018).
733 72. Ioannidis, P. *et al.* New criteria for selecting the origin of DNA replication in Wolbachia and
734 closely related bacteria. *BMC Genomics* **8**, 182 (2007).
735 73. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput
736 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114
737 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.20.508647; this version posted September 21, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

738 74. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069
739 (2014).
740 75. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Research* **33**,
741 W116–W120 (2005).
742 76. Wilkins, D. gggenes. https://github.com/wilkox/gggenes.
743 77. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
744 search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
745 78. Miao, Y., Xiao, J. & Huang, D. Distribution and Evolution of the Bacteriophage WO and Its
746 Antagonism With Wolbachia. *Front. Microbiol.* **11**, 595629 (2020).
747 79. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res*
748 **19**, 1639–1645 (2009).
749 80. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary
750 analyses in R. *Bioinformatics* **35**, 526–528 (2019).
751 81. Ives, A. R. R2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs. *Systematic*
752 *Biology* **68**, 234–251 (2019).