Article

MBE section: Discoveries

Life and death of selfish genes: comparative genomics reveals the dynamic evolution of cytoplasmic incompatibility.

Julien Martinez[1,2], Lisa Klasson[3], John J Welch[1], Francis M Jiggins[1].

[1]Department of Genetics, University of Cambridge, UK

[2] MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow G61 1QH, UK.

[3]Molecular Evolution, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden.

Corresponding authors:

Julien Martinez, julien.martinez@glasgow.ac.uk ;

Francis M Jiggins, fmj1001@cam.ac.uk.

Abstract

Cytoplasmic incompatibility is a selfish reproductive manipulation induced by the endosymbiont *Wolbachia* in arthropods. In males *Wolbachia* modifies sperm, leading to embryonic mortality in crosses with *Wolbachia*-free females. In females*, Wolbachia* rescues the cross and allows development to proceed normally. This provides a reproductive advantage to infected females, allowing the maternally-transmitted symbiont to spread rapidly through host populations. We identified homologs of the genes underlying this phenotype, *cifA* and *cifB,* in 52 of 71 new and published *Wolbachia* genomes sequences. They are strongly associated with cytoplasmic incompatibility. There are up to seven copies of the genes in each genome, and phylogenetic analysis shows that *Wolbachia* frequently acquires new copies due to pervasive horizontal transfer between strains. In many cases the genes have subsequently acquired loss-of-function mutations to become pseudogenes. As predicted by theory, this tends to occur first in *cifB,* whose sole function is to modify sperm, and then in *cifA,* which is required to rescue the cross in females. Although *cif* genes recombine, recombination is largely restricted to closely related homologs. This is predicted under a model of coevolution between sperm modification and embryonic rescue, where recombination between distantly related pairs of genes would create a self-incompatible strain. Together, these patterns of gene gain, loss and recombination support evolutionary models of cytoplasmic incompatibility.

Introduction

Maternally-transmitted bacteria in the genus *Wolbachia* commonly manipulate the reproduction of their arthropod hosts by inducing cytoplasmic incompatibility (CI). In the simplest case, CI causes embryonic mortality in crosses between symbiont-infected males and uninfected females (unidirectional CI). Because *Wolbachia*-infected females can still reproduce successfully with infected males, this provides them with a fitness advantage. Above a certain threshold in *Wolbachia* frequency, CI allows the infection to rapidly spread in the host population, even if it induces moderate fitness costs (Turelli et al. 1992). This is thought to have contributed to the remarkable evolutionary success of *Wolbachia*, which is estimated to infect around half of terrestrial arthropod species (Weinert et al. 2015).

Cytological studies have revealed that CI results from cytogenetic defects affecting the paternal chromosomes in early embryogenesis (Callaini et al. 1997; Tram and Sullivan 2002). As *Wolbachia* is not present in mature sperm, this suggests that *Wolbachia* modifies the sperm of infected males during spermatogenesis (the modification function). This leads

to development failing unless the zygote inherits a *Wolbachia* from the mother that is able to rescue embryonic development (the rescue function) (Callaini et al. 1997; Tram and Sullivan 2002). When the male and female parents are infected with different *Wolbachia* strains, it is also common to find that the cross is incompatible (bidirectional CI) (O'Neill and Karr 1990; Bordenstein et al. 2001; Atyame et al. 2014). This suggests the modification and rescue factors must match each other for development to proceed normally (Poinsot et al. 2003).

Recent work on the *Wolbachia* strains *w*Mel and *w*Pip from *Drosophila melanogaster* and *Culex pipiens* has found that the bacterial genes *cifA* and *cifB* are sufficient to induce CI (Beckmann et al. 2017; Lepage et al. 2017; Chen et al. 2019; Shropshire and Bordenstein 2019). In both strains *cifA* is located directly upstream of *cifB,* and it is thought that *cifA* and *cifB* are transcribed as a single operon (Beckmann et al. 2017) (although this has been questioned by Shropshire and Bordenstein 2019). In both *Wolbachia* genomes the genes are found within a prophage called WO, and the proteins encoded by the two genes bind each other (Beckmann et al. 2017; Chen et al. 2019).

In *Drosophila,* expressing $cifA^{wMel}$ in the female germline rescues CI in crosses with *w*Mel-infected males (Shropshire et al. 2018; Shropshire and Bordenstein 2019). Unexpectedly, in transgenic flies the modification of sperm requires both $cifA^{wMel}$ and $cifB^{wMel}$ to be expressed in the male germline (Lepage et al. 2017; Shropshire and Bordenstein 2019; Shropshire et al. 2020). This suggests that both genes are required for the modification function but only one gene for rescue and has been referred to as the 'two-by-one' model of CI by some authors (Shropshire & Bordenstein, 2019). One hypothesis to explain this is that *cifA* and *cifB* together modify sperm in a way that is lethal unless reversed by *cifA* in the early embryo. Other authors have proposed that *cifB* may be the only toxin in sperm and that *cifA* could act as an antitoxin both in embryos and in sperm (Beckmann, Bonneau, et al. 2019). In this model, the requirement of *cifA* in sperm would be due to the fact that it protects maturing sperm cells from the toxic effect of *cifB* (Beckmann, Bonneau, et al. 2019). This model requires further validation as experiments have so far failed to detect *cifB* being transferred to females on mating (Beckmann and Fallon 2013).

There appear to be at least two distinct molecular mechanisms by which these genes cause CI, which are illustrated by two paralogous pairs of *cifA*-*cifB* genes in *w*Pip. In one case *cifB* has two PD-(D/E)XK domains, which encode DNase activity and are required for the modification of sperm (Chen et al. 2019). The other *cifB* paralog in *w*Pip has two PD-(D/E)XK domains that lack the residues required for nuclease activity (Chen et al. 2019). Instead, a deubiquitylating domain, which cleaves ubiquitin from proteins, is required for the sperm modification function (Beckmann et al. 2017; Beckmann, Sharma, et al.

2019). Based on these distinct molecular functions, *cif* genes with DNase activity are also known as *cinA* and *cinB,* and *cif* genes with deubiquitinase activity are known as *cidA* and *cidB* (Beckmann, Bonneau, et al. 2019). Following Beckmann, Bonneau, et al. (2019), we use the *cif* terminology to refer to cytoplasmic incompatibility factors regardless of their mode of action.

Homologs of *cifA* and *cifB* have been discovered in other *Wolbachia* genomes, allowing us to address questions around their evolution (Lepage et al. 2017; Lindsey et al. 2018). The *cifA* and *cifB* phylogenies are strongly congruent, which is compatible with a model where modification and rescue factors must be matched (Lindsey et al. 2018). In contrast, the *cif* gene and *Wolbachia* phylogenies are incongruent indicating that these genes are often transferred between *Wolbachia* genomes (Lindsey et al. 2018). CI genes are also commonly associated with prophage sequences on the *Wolbachia* chromosome, suggesting that phage-mediated transfer may underlie their mobility (Lepage et al. 2017; Lindsey et al. 2018). Additionally, *Wolbachia* genomes often contain multiple pairs of the *cifA* and *cifB* genes, which may explain complex patterns of bidirectional incompatibility (Bonneau et al. 2018). Finally, homologs of *cifA* and *cifB* frequently display signs of pseudogenization, carrying mutations that disrupt their open reading frame or introduce premature stop codons (Asselin et al. 2018; Meany et al. 2018; Turelli et al. 2018). Therefore, the evolution of CI genes appears to be highly dynamic, being punctuated with acquisition events through horizontal transfer and losses by pseudogenization. However, the rate at which these events occur and to what extent they are governed by neutral processes or selection acting on the CI phenotype remains to be explored.

The identification of the *cifA* and *cifB* genes allows us to revisit theoretical predictions about the evolution of CI. A curious feature of CI is that it involves sperm being modified in males, and yet *Wolbachia* is not transmitted from males to future generations. This leads to the prediction that in randomly mating populations the ability to modify sperm is selectively neutral, so the genes involved can be lost by mutation and genetic drift (Turelli 1994; Hurst and McVean 1996). Once sperm modification has been lost, the rescue function will be free to suffer a similar fate (Turelli 1994; Hurst and McVean 1996). It has been suggested that the long-term maintenance of CI can be explained by kin selection in structured populations (*Wolbachia* in females benefits from related males inducing CI (Frank 1997)), but theoretical analyses suggest this may be a weak force that acts only under specific circumstances (Haygood and Turelli 2009). This has led to the suggestion that CI may be maintained by a process of clade selection where CI is necessary for the horizontal transfer of *Wolbachia* to new species (Hurst and McVean 1996). There is an analogy between these models, as here *Wolbachia* populations are structured depending on which species they infect (clade

selection model) as opposed to structure in space (kin selection model). Theory also predicts that multiple *Wolbachia* strains with different rescue and modification factors can invade populations (Frank 1998; Vautrin et al. 2007). A similar logic predicts that *Wolbachia* variants that acquire novel CI crossing types will invade infected populations provided they retain compatibility with the resident strain (Charlat et al. 2001). At the molecular level, this could be achieved by *Wolbachia* genomes accumulating *cifA*-*cifB* paralogs.

Here we explore the evolution of *cifA* and *cifB* using publicly available and newly sequenced *Wolbachia* genomes. We found the genes in most *Wolbachia* genomes, often in multiple copies. While the protein domains required for CI are widely conserved, the domain structure of divergent homologs found in some *Wolbachia* strains and related bacterial genera suggests they may play a diverse range of molecular functions. The *cif* genes are lost on relatively short time scales, with the modification function usually being lost before the rescue function. This is compensated by pervasive horizontal transfer of functional *cif* genes between symbiont genomes, in many instances across long phylogenetic distances. Finally, recombination between *cif* genes is largely restricted to closely related homologs, supporting the hypothesis that genetic divergence leads to the diversification of CI compatibility types.

Results

New genome sequences of *Wolbachia* from *Drosophila*

We sequenced eleven *Wolbachia* genomes from eleven different *Drosophila* hosts, with mean sequencing depths ranging from 27 to 40x (Table 1). Assembly sizes were within the range of typical *Wolbachia* genomes, with the genomes of the strains *w*Stv and *w*Nik forming a single scaffold. The number of near-universal, single-copy genes from the BUSCO proteobacteria database in the assemblies was similar to published reference genomes of *Wolbachia*, indicating that genomes are near complete (Table 1; repeating the analysis on published genomes of *w*Mel, *w*Au, *w*Ha, *w*No and *w*Ri yielded 181, 184, 183, 182, 182 complete BUSCO genes respectively). One of these strains, *w*Tri, has been previously sequenced by Turelli et al. (2018). Our sequence differed by 114 SNPs, was more intact and contained an additional pair of *cif* genes. We named our strain *w*Tri-2. The newly sequenced strains all cluster within *Wolbachia* supergroup A, like most *Wolbachia* isolated from *Drosophila* hosts in our dataset (Figure 1).

The *cif* genes are widespread in the genomes of *Wolbachia* and related *Rickettsiales*

To examine the evolution of *cif* genes we combined our newly sequenced genomes with published sequences (Table S1). This gave 71 *Wolbachia* genome sequences, in which we identified 129 and 128 homologs of *cifA* and *cifB* respectively (Figure 1; Table S2). Synteny was highly conserved—in 115 cases (~89%) *cifA* was located immediately upstream of *cifB* (Figure 1). A single *cifA* homolog and three *cifB* homologs broke this pattern and were present without their partner. Interestingly, all four of these genes carry mutations that disrupt their open reading frame (ORF) suggesting that they are pseudogenes. The remaining genes not found in pairs were located at the end of contigs and/or were partially sequenced, preventing us from drawing conclusions about synteny. As the operon status of these genes is disputed (Beckmann, Bonneau, et al. 2019; Shropshire et al. 2019), we will refer to the 115 syntenic pairs of genes as *cifA*–c*ifB* genes.

Seventy-three percent (52/71) of *Wolbachia* strains carry at least one homolog of *cifA* or *cifB* (Figure 1), and almost half the *Wolbachia* genomes carry multiple *cifA*–c*ifB* homologs (35/71). The largest number was in the strains infecting *Diploeciton nevermanni* and *Gerris buenoi,* both of which had seven syntenic *cifA*–c*ifB* genes (Figure 1). The counts of gene pairs across strains did not differ from a Poisson distribution (Cameron-Trivedi test for Poisson equidispersion on per strain counts of intact syntenic *cifA*–c*ifB* genes in supergroups A and B: $z = 1.38$, $p = 0.16$ (Cameron and Trivedi 1990)).

To examine the distribution of *cifA*–c*ifB* genes across bacterial strains we reconstructed the *Wolbachia* phylogeny using a set of 28 single-copy genes that were present in over 95% of the genomes. The *cifA* and *cifB* genes are widespread across the *Wolbachia* supergroups A and B, which contain most of the genomes analysed (Figure 1). Six symbiont strains belong to other *Wolbachia* supergroups, and none of their genomes contained *cifA* or *cifB.* However, syntenic *cifA*–c*ifB* genes were identified in the genomes of *Rickettsia gravesii, Rickettsia amblyommatis* and *Occidentia massiliensis* which infect ticks, and in a plasmid found in *Rickettsia felis* strain *LSU-Lb,* which infects the booklouse *Liposcelis bostrychophila* (Figure 2 and S1; Table S2).

The *cif* genes are associated with cytoplasmic incompatibility

To examine the association between *cif* genes and phenotype, we compiled a list of the effects that each *Wolbachia* strain has on host reproduction (Figure 1C, Table S1). We found that there was a significant association between the presence of *cifA*–*cifB* genes (without

ORF-disrupting mutations) and published reports of a strain inducing CI (Table 2; Fisher Exact Test: $p < 0.0001$). The only case of CI-inducing strain without these genes was *w*VitB, however, this may be an omission from the genome sequence. Indeed, we found what appeared to be fragments of *cifA* and *cifB* on very small contigs in the *w*VitB genome assembly. Another CI-inducing strain, *w*Rec, carries a syntenic *cifA*–c*ifB* pair where *cifB* has a frameshift mutation towards its 3' end, downstream of predicted protein domains known to have a role in sperm modification (Figure 2, Table S2). Although we counted this as a pseudogene it is possible that it encodes a functional protein. Finally, *w*Yak induces weak CI (Cooper et al. 2017) and both copies of cifB in the genome contain stop codons (Cooper et al. 2019 ; Figure 2).

There were 14 strains that are reported not to induce CI (Table 2). Among them, the strain *w*Bor was the only one to carry a pair of *cifA*–c*ifB* genes without ORF-disrupting mutations, although this strain is only known to induce male-killing. As described below, both of these genes lack domains that are conserved across all CI-inducing strains. Among the other non-CI strains, the genes were absent from two *Wolbachia* strains that induce parthenogenesis (*w*Tpre and *w*Uni), one strain suspected to induce parthenogenesis (*w*Fol), two strains that do not induce CI or other clear phenotypic effects (*w*Au and *w*Tro) and five strains thought to be mutualists (*w*Cle in bed bugs and the four strains from nematodes) (Figure 1). Finally, the male-killing strain *w*Bif and two strains with no clear reproductive phenotype (*w*Spc and *w*Neo) were only found to carry *cif* gene pairs with ORF-disrupting mutations. Overall, the strong association between syntenic *cif* genes and the CI phenotype suggests a single evolutionary origin of this reproductive manipulation in *Wolbachia* (Table 2).

A new and diverse group of *cif* genes is found in *Wolbachia and other Rickettsiales*

Reconstruction of the *cifA-cifB* gene tree revealed new diversity among these genes. While most sequences fell into four clades that have been named Types I-IV, the genes from two *Rickettsia* species, a *Rickettsia* plasmid, *Occidentia massiliensis* and several *Wolbachia* strains were basal to these four types on the midpoint rooted tree (Figure 2; Figure S1). The divergence among these sequences is frequently greater than between Types I-IV, and we are unable to root the phylogeny with high confidence without a reliable outgroup. However, midpoint rooting suggests that this may be a paraphyletic group with the other *cif* genes nested within it. We propose to call this diverse assemblage of *cifA-cifB* homologs "Type V".

We found multiple protein domains, including the *cifB* nuclease and deubiquitinase domains known to be involved in sperm modification (see below for a detailed description of protein

domain conservation). Type V *cifB* genes tend to be longer and possess a diverse array of domains, suggesting that they may perform a variety of molecular functions (Figure 2; Figure S1B). Among these genes we identified toxin domains (Latrotoxin, RTX, pore-forming and salivary-gland toxins), a protease domain (OTU-like cysteine protease) and domains involved in protein-protein interactions (tetratricopeptide and ankyrin repeats). The ankyrin repeat domain also shows strong similarities with DNA-binding proteins (probabilities ~100% in the HHpred search).

All five *cifA-cifB* types are associated with CI. Type I genes from *w*Mel and *w*Pip, and Type IV genes from *w*Pip have been experimentally linked to CI (Beckmann et al. 2017; Lepage et al. 2017; Chen et al. 2019). Additionally, we found CI-inducing strains such as *w*No and *w*Stri that carry only Type III or Type V genes. Finally, in the CI-inducing strain *w*Ri the only *cifA-cifB* genes without signs of pseudogenization belong to Type II.

The *cifA* and *cifB* genes codiverge with recombination restricted to closely related genes

The *cifA* and *cifB* proteins bind each other, and in a comparison of two *Wolbachia* strains the proteins encoded by syntenic pairs of genes bound more strongly than heterologous proteins (Beckmann et al. 2017; Chen et al. 2019). This led to the suggestion that coevolution of binding affinities between the proteins could underlie the divergence of CI crossing types (Beckmann, Bonneau, et al. 2019). Consistent with this and in agreement with earlier studies (Lepage et al. 2017; Lindsey et al. 2018), syntenic *cifA* and *cifB* genes show strong phylogenetic congruence (Mantel test p-value < 0.0001; Figure S2A-B; Figure 3A). Strikingly, there is no case where recombination has brought together *cifA* and *cifB* genes from different Types (Figure 3A). Nonetheless, the two trees are not identical. Using multiple approaches to recombination detection on the concatenated alignment of *cifA* and *cifB* we identified 83 well-supported recombination events (Table S4; note that some events may have been counted multiple times). Manual inspection of the sequences frequently revealed clear recombination breakpoints (Figure S3A-C). However, all but one of these events involved sequences of the same Type (82/83 events; Table S4). This pattern of recombination tending to occur between closely related sequences was strongly supported, as the mean genetic distance between inferred parental sequences was significantly lower than expected by chance (Figure S4). Manual inspection of the only recombination event involving parental sequences belonging to different Types revealed no clear breakpoint in the recombinant sequence, suggesting that this event could be a false-positive created from low-quality alignment between highly divergent homologs (Figure S3D). Together, these

results indicate that *cifA* and *cifB* recombination is largely restricted to closely related sequences, perhaps because their binding affinities have coevolved.

A conserved protein domain architecture is associated with cytoplasmic incompatibility

To gain insights into the molecular basis of cytoplasmic incompatibility and its evolution, we used a comparative approach to identify the protein domains associated with the trait. A complication is that closely related domains may be annotated in some sequences but not others depending on whether they meet an arbitrary significance threshold. To avoid this, we first searched for domains using conventional tools, and then used these sequences to create a *Wolbachia*-specific HMM profile for each domain that was used to repeat the search.

The induction of CI by the Type IV *cifB* paralog in wPip (*cinB*) requires DNAse activity due to its PD-(D/E)XK nuclease domains (Chen et al. 2019). However, the Type I *cifB* paralog in *w*Pip (*cidB*) has lost the catalytic residues associated with DNase activity and instead CI requires a deubiquitinase domain that functions to deconjugate ubiquitin from proteins (Beckmann et al. 2017; Beckmann, Sharma, et al. 2019). We found the two PD-(D/E)XK nuclease domains were highly conserved across the *cifB* tree (Figure 2; Figure S1B; Table S5). As is normally the case, these domains were associated with a 5' AAA-ATPase domain (Knizewski et al. 2007). The catalytic D-E-K residues were conserved throughout *cifB* evolution until they were lost in the common ancestor of the Type I genes (Figure S6**,** Lindsey *et al.* 2018; Chen *et al.* 2019). This coincided with the Type I genes acquiring the deubiquitinase domain, and this domain is conserved across Type I. This suggests that the molecular basis of CI has been conserved across the Type II, III and IV genes, and then changed (Chen et al., 2019) in the ancestor of the Type I sequences. Interestingly, the deubiquitinase domain is also present in some but not all Type V *cifB* homologs, suggesting that these might induce CI through a mechanism similar to that of the Type I genes (Figure 2; Figure S1B; Table S5). The sparse distribution of the deubiquitinase domain raises questions about its origin. A phylogenetic reconstruction of the deubiquitinase domain's amino acid sequence grouped Type I and V deubiquitinase domains as two monophyletic clades (Figure S7), so the domain does not appear to have been exchanged by recombination. Instead, the deubiquitinase domain must either have been acquired independently by the Type I and V genes, or have been present in the common ancestor of the *cif* genes and subsequently lost multiple times.

The *cifA* gene has a more conserved domain structure than *cifB* (Figure 2; Figure S1A; Table S5). The two domains that we identified were annotated as functioning in apoptosis regulation and RNA-binding, but we caution that the support for these annotations being true positives was less than 85%. There are six *cifA* genes that have lost the apoptosis regulator-like domain (including two putative pseudogenes) and none of these are known to induce CI (Figure S1A). If some of these strains were found to induce CI, it would be of interest to test if they can also rescue this effect, to assess if this domain is required for the rescue activity of *cifA* proteins.

The *cif* genes frequently transfer between distantly related *Wolbachia* genomes and insect hosts

The *cif* genes are often found in the vicinity of prophage genes, a cluster of genes known to undergo intense genomic rearrangements and horizontal transfer between *Wolbachia* genomes (Bordenstein and Wernegreen 2004). The link between *cif* genes and mobile genetic elements is reinforced by the presence of the genes on a plasmid found in *Rickettsia.* There are numerous cases of the genes being exchanged between distantly related bacterial hosts, including multiple instances of transfer between the *Wolbachia* supergroups A and B (Figure 3B). This suggests that horizontal transfer of *cif* genes between distantly-related symbionts is common, sometimes even crossing the bacterial genus boundaries as exemplified by the presence of *cifA*-*cifB* homologs in *Rickettsia* genomes and plasmids (Figure 2; Figure S1). Nonetheless, there is partial congruence between the phylogenies of *Wolbachia* and their *cif* genes (Mantel test p-value < 0.0001; Figure 3B; Figure S2C-D). This could result from *cif* genes being maintained for long enough to be co-inherited with the *Wolbachia* genome during speciation events or the genes frequently transferring between closely related bacterial strains.

Since the *cifA* and *cifB* proteins interact with targets in the arthropod host, their distribution may be constrained by the arthropod phylogeny rather than the *Wolbachia* phylogeny. Despite the *cifA* and *cifB* genes frequently transferring between distantly related arthropods, there is a highly significant tendency for closely related *cif* genes to be found in the same host order (Mantel test comparing *cifA*-*cifB* pairwise distances and insect orders: p-value < 0.0001; Figure S2E-F). However, we would caution that this association could be the result of the *Wolbachia* phylogeny being correlated with the arthropod phylogeny (Mantel test comparing *Wolbachia* pairwise distances and insect orders: p-value < 0.0001; Figure S2G-H) and these effects are difficult to disentangle.

Pseudogenes are common and the loss of sperm modification usually pre-dates the loss of the rescue

In randomly mating populations there is no selection for CI-inducing *Wolbachia* to modify sperm (Turelli 1994), so a gene whose only function is sperm-modification (*cifB*) is predicted to eventually lose its function. Once this occurs there is no selection to maintain the rescue function, so *cifA* can also be lost. Excluding partially sequenced genes, putative loss-of-function mutations were found in 12.1% of *cifA* sequences (15/124) and 27.8% of *cifB* sequences (27/97), suggesting that the modification of sperm function is lost more frequently than the rescue function (Figure S1A-B**;** Fisher Exact Test: $p = 0.02$). The fixation rate of loss-of-function mutations per site was not significantly different between the two genes (*cifA*: 2.4×10^{-4} mutations/amino acid; *cifB*: 3×10^{-4} mutations/amino acid; Fisher's exact test: $p = 0.64$).

To examine the order in which *cifA* and *cifB* become pseudogenes, we visually inspected the 87 fully-sequenced *cifA*-*cifB* pairs. We identified 38 independent mutational events that cause the ORF to be disrupted (i.e. not double-counting mutations inherited through speciation or duplication events; Figure 2). These putative loss-of-function mutations include single base pair substitutions introducing premature stop codons, indels producing frameshifts, insertion of transposable elements (*w*Stv pair 1) and short inversions (*w*Yak pair 1) (Table S2). The majority of *cifA*-*cifB* gene pairs show no signs of pseudogenization (63/87), while six carry ORF-disrupting mutations in both genes. In cases where just one gene carried a mutation, it was more often *cifB* than *cifA* (14 versus 4 instances; Figure 2; binomial test: $p = 0.03$). This suggests sperm modification function is usually being lost before the rescue function. Interestingly, in two cases where only *cifA* appears to be pseudogenized (*Wolbachia* in *Nomada* bees), the *cifB* genes are ~50% shorter than their close relatives and lack the nuclease domains conserved in all other *cifB* genes (Figure 2). Therefore, these *cifB* genes may be non-functional despite the absence of ORF-disrupting mutations. In a third case, *cifA* is pseudogenized in *w*Spc, a strain that does not modify sperm in its host *Drosophila subpulchrella* despite carrying an intact *cifB* gene, meaning that there should be no selection acting to maintain the rescue function in this strain (Figure 1).

Once the CI genes lose their function, it has been predicted that the *Wolbachia* infection will also be lost from the population. This should happen unless other phenotypes such as the provision of fitness benefits to the host are maintaining the symbiont, or the *Wolbachia* genomes harbour additional *cif* genes that are still functional. If the infection is maintained, non-functional CI genes may accumulate further loss-of-function mutations or be eliminated

from the *Wolbachia* genomes, for instance by excision of the prophage harbouring the CI genes. By looking at the phylogenetic distribution of putative loss-of-function mutations, it is possible to infer whether non-functional *cif* genes slowly degenerate or are quickly eliminated. The majority of loss-of function mutations are located on the terminal branches of the CI gene phylogenies (Figure 2; Figure S1A-B). The only exceptions are a few mutations co-inherited between closely-related *Wolbachia* strains (*w*Neo/*w*Orie; *Wolbachia* from *Nomada* bees) or through a putative duplication in the *w*Ri genome (pair 1 and 3 are identical). This suggests that pseudogenized *cif* genes are rarely horizontally-transmitted between *Wolbachia* genomes or maintained for long enough to be inherited through speciation events.

Closely related *cif* genes commonly coexist in the same genome

CI favours female hosts that have the greatest reproductive compatibility with males in the population. Therefore, *Wolbachia* strains that induce a new CI crossing type can invade *Wolbachia*-infected populations provided they are compatible with the resident strain (Charlat et al. 2001). Such a mutant could arise if *Wolbachia* acquires an additional pair of *cif* genes, and this may explain why *Wolbachia* genomes commonly harbour multiple *cif* paralogs. However, additional *cif* genes should only spread if they induce a different crossing type from the genes already present in the genome. If closely related genes confer the same crossing type, then this would mean that closely related *cif* homologs are unlikely to be found within the same genome. However, we found no support for this as in our dataset - putatively functional *cif* homologs in the same genome have very similar levels of divergence to *cif* homologs found in different genomes (Figure S5). A similar argument applies to the loss of *cif* genes—if a genome contains paralogs that induce the same crossing type then one of the paralogs could be lost by mutation. However, again there is no evidence for this process. Looking within genomes, the genetic distance between intact *cif* genes and putative pseudogenes was similar to the genetic distance between pairs of functional genes (genetic distance calculated from Figure 2: 1.15 and 1.28 respectively).

Discussion

CI is the most commonly observed reproductive manipulation induced by *Wolbachia*, and its evolution has been investigated for over sixty years through phenotypic experiments (Laven 1957; Sinkins et al. 1995; Charlat et al. 2003; Zabalou et al. 2008) and evolutionary models

(Caspari and Watson 1959; Turelli 1994; Hurst and McVean 1996; Vautrin et al. 2007). The discovery of the genes underlying CI now makes it possible to reconstruct the trait's evolution at the molecular level and infer the selection pressures acting on CI using the tools of molecular evolution. Here we analysed 71 *Wolbachia* genome sequences to investigate the importance of recombination, pseudogenization and horizontal gene transfer in the evolution of CI.

The *cif* genes are widespread across the *Rickettsiales*. Our large dataset supported the observation of Lindsey et al. (2018) that *cif* genes are common in *Wolbachia* supergroups A and B. These supergroups contain most of the *Wolbachia* strains that have been described, and here the genes are tightly linked to the CI phenotype. The genes were absent from a small sample of *Wolbachia* strains from other supergroups. This may explain why, to our knowledge, there are no reports of strains from these supergroups inducing CI. However, divergent Type V *cif* genes are found in other *Rickettsiales*, and here their phenotypic effects are uncertain (Gillespie et al. 2018). Gillespie et al., 2018 proposed that *cif* homologs found in *Rickettsia* may induce other reproductive phenotypes. For instance, *Rickettsia felis* strain LSU-Lb, which carries Type V *cif* genes on its plasmid, infects a parthenogenetic insect. Accurate rooting of the *cif* gene tree together with functional characterisation of these genes will be needed to confirm whether the ancestral function of these genes was to induce CI. However, the presence of a CI-inducing *Wolbachia* strain that has only Type V genes suggests that CI evolved very early in the evolution of *cif* genes.

It has long been observed that crosses between insects carrying different *Wolbachia* strains are frequently incompatible, suggesting that modification and rescue factors must be matched to produce viable offspring. A possible molecular cause of this phenomenon comes from the observation that *cifA* and *cifB* bind each other (Beckmann et al. 2017; Chen et al. 2019), and so binding affinities may be greatest between coevolved genes. If this is the case, then there would be strong selection against recombination between the genes as this could generate symbionts that are unable to rescue crosses with infected females (Charlat et al. 2001). Furthermore, even if the symbiont retained the ability to rescue the cross, for example if genes are swapped among paralogous pairs within the genome, the pair of genes might be eliminated in the long term if it generated self-incompatibility when transferred to a new genome. Despite this, clear evidence of recombination has been observed among *cif* genes from *w*Pip, which infects *C. pipiens* mosquitoes (Bonneau et al. 2018). *w*Pip strains can carry multiple copies of Type I *cif* genes, and recombination between closely related paralogs correlates with different CI crossing types. Whether recombination itself created these incompatibilities or whether crossing types arose due to sequence divergence following recombination is unknown. We found that recombination is frequent across the *cif*

gene phylogeny, but it almost exclusively occurs between related syntenic *cifA*-*cifB* gene pairs within the same Type. Since *Wolbachia* genomes often carry divergent *cif* homologs, the absence of recombination between these genes is compatible with recombination being constrained by selection. This can be explained by the coevolution of binding affinities between the proteins encoded by *cifA* and their cognate *cifB* gene.

The evolution of sperm modification by *Wolbachia* poses an evolutionary puzzle—the trait is only expressed in males and yet symbionts in males are not passed on to the next generation. In randomly mating populations it will be at best selectively neutral (Prout 1994; Turelli 1994). While kin selection can act to maintain CI in structured populations (Frank 1997), this is a weak force that operates only under specific circumstances (Haygood and Turelli 2009). Therefore, theory predicts that a gene involved exclusively in sperm modification, such as *cifB,* will accumulate loss of function mutations. Once these have been fixed within a population, the gene required for the rescue function (*cifA*) will then degenerate by mutation. It has already been reported that *cif* genes often carry putative loss-of-function mutations (Asselin et al. 2018; Lindsey et al. 2018; Meany et al. 2018; Cooper et al. 2019). We found *cifB* carries loss-of-function mutations more often than *cifA*, and that *cifA* generally acquires such mutations after *cifB*. This confirms a key prediction of theory, and supports the hypothesis that lack of selection to maintain sperm modification may lead to the loss of CI in some populations (Turelli 1994; Hurst and McVean 1996). While evolutionary predictions were based on a mechanistic model that assumes sperm modification and rescue were encoded by different genes, the same pattern of gene loss would be expected if both *cifA* and *cifB* are required to modify sperm (Shropshire and Bordenstein 2019). This is because a mutant that lost *cifA*, and therefore the ability to both induce and rescue CI, would initially be rare in a population of CI-inducing symbionts and should be counter-selected. By contrast, mutating *cifB* will only cause the loss of sperm modification and will therefore not be exposed to selection. Therefore, even in the two-by-one mechanistic model, the spread of a symbiont carrying a pseudogenized *cifA* and a functional *cifB* is unlikely. Nonetheless, how the two-by-one model affects details of the dynamics, such as the effects of population structure, remain to be explored theoretically.

The degeneration of *cif* genes could also occur if the CI phenotype is lost first, for instance due to the host evolving to suppress the trait (Koehcke et al. 2009), leaving the genes free to degenerate by mutation. This is plausible as artificial transfers of *Wolbachia* between species have shown that host genetic background can affect the expression of CI (Poinsot et al. 1998; Jaenike 2007; Zabalou et al. 2008). As this model makes no predictions about the order of gene loss, it may explain why we found at least one *Wolbachia* strain that does not induce any reproductive incompatibility and whose genome contains a *cifA* pseudogene

alongside an intact *cifB.* As *cifB* is typically longer than *cifA*, even under this model of evolution *cifB* may tend to acquire loss-of-function mutations first. Indeed, the number of loss-of-function mutations per amino acid is similar between *cifA* and *cifB,* which is compatible with the two genes evolving under similar selective pressures. However, this is a rather weak test as after a loss of function mutation in *cifB,* under any evolutionary model both genes are free to degenerate at the same rate. Therefore, to separate these models it is necessary to know whether strains carrying *cifB* pseudogenes alongside functional *cifA* genes are found in hosts that can express CI. An example is *w*Ri. This strain induces CI, and its genome contains an intact pair of *cif* genes alongside two distantly related pairs of genes where *cifB* is a pseudogene. Another apparent example of this is *w*Yak, which has two copies of *cifB* in its genome, both of which contain internal stop codons (Figure 1; Cooper et al. 2019). When its natural host *D. yakuba* was experimentally infected with a different *Wolbachia* strain it induced strong CI (Zabalou et al. 2004). However, the recent discovery that *w*Yak itself induces weak CI means that it is unclear whether the mutations in the two *cifB* genes reduced the ability of *w*Yak to modify sperm (Cooper et al. 2017). Together these examples provide tentative support for the model that sperm modification may degenerate by mutation, even in hosts expressing CI.

In the absence of other phenotypic effects on the host, theory predicts that the loss of sperm modification will lead to the loss of *Wolbachia* from the host population (Turelli 1994; Hurst and McVean 1996). We found that the majority of loss-of-function mutations appear to have occurred recently. First, these mutations tend to be on terminal branches of the gene tree. Second, pseudogenes rarely accumulate many loss-of-function mutations. While this suggests that *cifA* and *cifB* pseudogenes rarely co-diverge with the *Wolbachia* genome for long periods, it is unclear whether this results from the loss of the *Wolbachia* infection or through the genes being deleted from the genome, for instance by phage excision. Since the publication of theoretical studies on CI evolution (Turelli 1994; Hurst and McVean 1996), it has become increasingly clear that many *Wolbachia* strains can persist in host populations without inducing any reproductive manipulation by providing fitness benefits (Kriesner & Hoffmann, 2018; Kriesner, Hoffmann, Lee, Turelli, & Weeks, 2013; Dedeine et al., 2001; Hosokawa, Koga, Kikuchi, Meng, & Fukatsu, 2010; Taylor, Bandi, & Hoerauf, 2005 Martinez et al., 2014). Therefore, it remains to be demonstrated whether the degeneration of *cif* genes causes the loss of *Wolbachia* from populations.

The frequent loss of *cif* genes raises a paradox when trying to explain the high prevalence of CI across *Wolbachia* strains. Hurst & McVean (1996) argued that CI-inducing *Wolbachia* infections were more likely to transfer horizontally between host species, so symbiont strains that induce CI are more likely to persist over an evolutionary timescale. Consistent with this

process of clade selection*, Wolbachia* frequently jumps between host species (Zhou et al. 1998; Vavre et al. 1999; Baldo et al. 2008). Our observations suggest that clade selection may also be acting at the level of the *cif* genes as *Wolbachia* genomes appear to be frequently recolonised by *cif* genes from other symbionts. The pervasive horizontal transfer of *cif* genes may allow them to evade inevitable extinction within symbiont lineages by escaping into a new symbiont population. This process is analogous to the evolution of transposable elements, which frequently go extinct within host species, but persist long term by jumping into new species (Schaack et al. 2010).

Horizontal transfer of *cif* genes occurs frequently and sometimes over large phylogenetic distances, sometimes even crossing the bacterial genus boundaries. High rates of horizontal transfer likely result from the *cif* genes being associated with mobile genetic elements, such as WO prophage sequences, transposons and plasmids (Gillespie et al. 2018; Lindsey et al. 2018; Cooper et al. 2019). Interestingly, Lepage et al. (2017) found no significant association between the phylogenies of *cif* genes and genes found in the structural module of phages. Prophage regions often rearrange (Bordenstein and Wernegreen 2004) which likely explains this pattern. However, this does not mean that phages are not a major route of horizontal transmission of *cif* genes, and the association between *cif* genes and mobile genetic elements supports the idea that the ability to move horizontally between genomes may be an important adaptation of these elements. Indeed, *cif* genes that lose their association with mobile elements may ultimately go extinct as they would be less likely to undergo horizontal transfer. As argued by Beckmann, Bonneau, et al. (2019), in some cases CI may be best viewed as an adaptation of mobile genetic elements to spread within *Wolbachia* populations.

Wolbachia genomes often carry multiple *cifA-cifB* gene pairs. This is analogous to the frequent occurrence of multiple *Wolbachia* strains within the same host individuals. Both experimental and theoretical studies have demonstrated that such multiple infections can invade host populations provided the strains carry different modification and rescue factors (Sinkins et al. 1995; Rousset et al. 1999). This is because females harbouring additional *Wolbachia* strains will be compatible with all males in the population. An equivalent process will promote the invasion and maintenance of *cif* paralogs within the same genome, provided that they encode bidirectionally incompatible modification and rescue factors. If bidirectionally incompatibility evolves gradually, this hypothesis predicts that paralogous *cif* genes within a genome might be distantly related. However, we found no evidence for this, and frequently paralogs within the same genome are closely related. We would argue that this does not mean that we should discount the hypothesis that *cif* paralogs accumulate within the same genome because they encode bidirectionally incompatible CI factors. In particular, our analysis assumes that genetic divergence can be used as a proxy for the

545 divergence of crossing types and this may not be the case. For example, the evolution of
546 new crossing types could occur following *cif* gene duplication events (Beckmann, Bonneau,
547 et al. 2019), meaning that a single genome could harbour closely related paralogs that
548 encode incompatible crossing types, as observed in the *w*Pip-*Culex* system described above
549 (Bonneau et al. 2018). While sequence divergence following duplication is thought to have
550 led to new compatibility types (Beckmann, Bonneau, et al. 2019), an open question is
551 whether carrying multiple identical copies of a *cif* gene pair might be sufficient to produce
552 new crossing types, perhaps by "dose effects".

553 There is both direct and indirect evidence that homologs from all the main *cif* gene Types
554 can induce CI, and analysis of protein domains suggests that the molecular basis of CI is
555 conserved. The *cifA* domain structure varies little across the gene family. All *cifB* genes
556 associated with CI had one of the domains that has been experimentally linked to CI—a
557 functional PD-(D/E)XK nuclease domain or a deubiquitinase domain (Beckmann et al. 2017;
558 Chen et al. 2019). As previously reported, Type I *cifB* genes lack the catalytic residues in
559 their PD-(D/E)XK nuclease domains, and instead have the deubiquitinase domain
560 (Beckmann et al. 2017). Unexpectedly, this domain is also present in some divergent Type V
561 sequences, making its evolutionary origins unclear. Alongside these conserved core
562 domains, we found a diverse range of other *cifB* domains, notably in the long Type V genes.
563 It is of interest whether these additional *cifB* domains are associated with CI. Given the
564 relative homogeneity of *cifA* domains and the similarities of these additional *cifB* domains
565 with known toxins and eukaryotic-like ankyrin proteins, it is tempting to hypothesize that they
566 are linked to the modification function by interacting with the host rather than the binding to
567 *cifA*. It would be interesting to test whether these domains allow manipulation of reproduction
568 across a broader host range. Interestingly, one of these domains, the ovarian tumor domain
569 (OTU), is also found in a toxin involved in male-killing induced by the symbiont *Spiroplasma*
570 *poulsonii*, raising the possibility that some domains could be involved in multiple forms of
571 reproductive manipulations.

572 In conclusion, our study illustrates the dynamic evolution of CI genes and highlights the high
573 rates of gene loss and horizontal gene transfer. Further functional analysis will open new
574 avenues of research and allow us to reconstruct the full evolutionary history of CI. In
575 particular, the identification of the insect factors targeted by the sperm modification may
576 soon allow us to study the coevolution of *cif* genes with the insect reproductive system
577 (Beckmann, Sharma, et al. 2019). Finally, a deeper analysis of divergent *cif* homologs found
578 outside *Wolbachia* should allow us to address questions around the deep evolutionary origin
579 of CI and may reveal novel functions of these genes.

Methods

Sequencing of new *Wolbachia* genomes

New *Wolbachia* genomes were obtained from eleven *Drosophila* species (listed in Table 1) using the protocol described in Ellegaard et al. (2013). Briefly, *Wolbachia* cells were purified from 20-30 fly embryos that were dechorionated in bleach and homogenized in phosphate-buffered saline (PBS). The homogenate was then centrifuged, passed through 5 and 2.7 µm pore size filters and multiple-displacement amplification (MDA) was performed directly on the bacterial pellet using the Repli-g midi kit (Qiagen) as in Ellegaard et al., 2013. The amplified DNA was finally cleaned using the QIAamp DNA mini kit prior to sequencing. From each DNA sample, 3 kb paired-end and 50 bp paired-end DNA libraries were prepared. These were multiplexed and sequenced on one plate of 454 Roche FLX (University of Cambridge, Department of Biochemistry, UK) and one lane of Illumina HiSeq2000 instruments (The Genome Analysis Centre, Norwich, UK) respectively.

454 and Illumina reads were used to perform hybrid de novo assemblies in Newbler v2.6 (454 Life Sciences Corp., Roche, Branford, CT 06405, US). Non-*Wolbachia* contigs were then removed from each assembly by aligning the contigs to the *w*Mel reference genome (Genbank accession number: NC_002978.6) using Mauve v2.3.1 (Darling et al. 2004) and visual comparisons in the Artemis Comparison Tool (Carver et al. 2005). A blastn analysis of the discarded contigs revealed positive matches with *Drosophila* mitochondrial and nuclear genomes, as well as with *Saccharomyces cerevisiae* (yeast used to collect the fly embryos), suggesting low levels of contamination during the DNA extraction process. Scaffolding was refined using SSPACE v2 (Boetzer et al. 2011) and gaps were filled with Gapfiller v1.11 (Boetzer and Pirovano 2012). Additionally, for strains *w*Stv, *w*Ara and *w*Bor, Illumina reads were assembled separately using Abyss v1.3.5 (Simpson et al. 2009) and the contigs generated as well as the Illumina reads were mapped onto the hybrid assemblies using Consed (Gordon et al. 1998) in order to manually edit the scaffolds. Final assemblies were annotated as in Ellegaard *et al.* (2013) using a custom pipeline. In brief, gene and pseudogene predictions were performed using Prodigal (Hyatt et al. 2010) and GenePrimp (Pati et al. 2010) respectively. Domain prediction was done using hmmsearch implemented in pfam_scan.pl with the PFAM database (Bateman et al. 2002). Finally, annotations were manually edited through visual inspection in Artemis (Carver et al. 2012). The completeness of the assemblies was assessed using BUSCO v3 by searching the genomes against the near-universal, single-copy genes of the proteobacteria database (Simão et al. 2015).

Blast search of CI genes and annotation

Previously identified *cif* gene homologs have been categorized into four phylogenetic clusters denominated as Type I to IV, as well as an uncharacterized type of more divergent homologs found in the *Wolbachia* strain *w*Stri (Lindsey et al. 2018). The presence of *cif* gene homologs in publicly available and newly sequenced *Wolbachia* genomes (Table S1) was searched with TBLASTN (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using *cifA* and *cifB* amino acid sequences representative of each CI type as queries (Table S2). Some of the genomes assembled in Pascar & Chandler (2018) were excluded from our analysis as they showed signs of multiple infections based on the assembly size, the sequencing coverage and/or the presence of duplicated BUSCO reference genes. Additionally, when more than one genome were sequenced from the same host species and we found no evidence in the literature that they correspond to different *Wolbachia* strains, only one of them was included in the analysis. Default parameters and an e-value threshold of 0.05 were used. TBLASTN hits across the *Wolbachia* genomes were then visually inspected in Artemis. Hits that were at least 40% of the length of the smallest query sequence and/or hits displaying the typical *cifA*-*cifB* synteny were considered as positive matches. Where sequences did not span the entirety of an open reading frame (ORFs), they were manually extended to include the closest start and stop codons. The presence of stop codons and frameshifts within the reannotated sequences were interpreted as indicative of a putative pseudogenization event. DNA sequences were aligned with Clustal Omega using default parameters (https://www.ebi.ac.uk/Tools/msa/clustalo/) and putative “loss-of-function” mutations were examined by comparing closely related sequences (see below phylogenetic reconstruction) and were defined as unique mutational events or, when found in more than one homolog, as co-inherited.

Using BLASTP online tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with $cifA^{wMel}$ and $cifB^{wMel}$ amino acid sequences and the more divergent sequences found in *w*Stri as queries we also found additional homologs in prophage WOSol (AGK87106 and AGK87078) and in non-*Wolbachia* taxa (Rickettsial plasmid genes pLbAR_36/38: WP_039595309.1/WP_081996388.1; *Rickettsia gravesii*: NZ_AWXL00000000.1 ; *Rickettsia amblyommatis*: GCA_000964995.1 ; *Occidentia massiliensis*: CANJ01000001). These *cif* homologs were searched again and manually annotated as above from the original nucleotide sequences present in Genbank (https://www.ncbi.nlm.nih.gov/genbank/).

CI genes phylogeny and recombination

Following the manual reannotation, *cifA* and *cifB* DNA sequences were aligned separately based on their amino acid translations using TranslatorX (Abascal et al. 2010). The alignment program MAFFT implemented in the TranslatorX pipeline was used along with GBlocks in order to filter out weakly conserved regions from the alignment. The same was done for concatenated *cifA* and *cifB* sequences. PhyML v3.0 (Guindon et al. 2010) was used with the GTR GAMMA substitution model of evolution and 1,000 bootstrap replicates. All phylogenies were first reconstructed with all sequences, including partial ones (located at the end of genome contigs or showing bases called as Ns within their ORF) to classify the genes into phylogenetic types. Phylogenies were then rerun without partial sequences for the remaining analyses.

Recombination was analysed using the recombination detection program RDP4 (Martin et al. 2015). The alignment of the concatenated *cifA* and *cifB* linear sequences were used to detect putative recombination events with the default parameters of the six methods implemented in RDP4 (RDP, GENECONV, Bootscan, MaxChi, Chimaera and SiScan). Recombination events with a phylogenetic support and a Bonferroni-corrected *p*-value < 0.05 with at least four of the detection methods were interpreted as reliable evidence of recombination. Out of 83 significant events, twenty-two were randomly selected and visually inspected to ensure they were genuine. This was done by realigning the amino acid sequences of the putative recombinant and the two inferred parents.

Prediction of functional domains

Protein domains were predicted for fully sequenced *cifA* and *cifB* homologs using the HHpred webserver (https://toolkit.tuebingen.mpg.de/#/tools/hhpred/ ; Söding, Biegert, & Lupas, 2005) with defaults parameters as in Lindsey et al., 2018. Premature stop codons were first removed manually from putative pseudogenes and DNA sequences translated into amino acid sequences. Amino acid sequences were then queried individually against the following databases: SCOPe70 (v.2.07), Pfam (v.32.0), SMART (v6.0), and COG/KOG (v1.0). Only hits with probabilities >75% in at least one *cif* homolog were considered as putative functional domains. In many cases, predicted domains on one sequence were not detected by HHpred on closely related homologs, although their presence could be suspected by visual inspection of the sequence alignments. In order to refine our domain search, we used representative amino acid sequences of each domain to build HMM profiles using hmmbuild implemented in HHMMER v3.2.1 (Eddy 2011). Representative domain

selection was conducted by choosing domains distributed across the CI gene phylogenies, encompassing the maximum genetic divergence between reference homologs. Basically, one domain copy per phylogenetic type as well as all copies from non-*Wolbachia* taxa were chosen as references where available. When domains were missing from some types, a maximum of five domain copies were selected across the different types or all copies if there were fewer than five copies in total.

All CI homologs were then scanned using hmmscan and the hmm profile created from each functional domain with an e-value inclusion threshold of 0.0001. The amino acid sequences of hmmscan hits were extracted and re-used to build new HMM profiles that were used to scan the CI genes again. Three search iterations were performed in this way which allowed us to retrieve many domains that HHpred failed to detect. Finally, domain coordinates were extracted and, in the case of putative pseudogenes, they were manually edited to take into account the presence of loss-of-function mutations in the original DNA sequences.

Wolbachia **strain phylogeny**

Twenty eight single copy genes that were present in >95% of the bacterial strains were identified using Phylosift v1.0.1 (Darling et al. 2014) (Table S3). The genomes of *Anaplasma marginale* (NC_012026.1) and *Ehrlichia muris* (NC_023063.1) were included as outgroups. For each genome, DNA sequences were concatenated and aligned with MAFFT v7. A bacterial phylogeny was then built with PhyML v3.0 using the GTR GAMMA substitution model with 1,000 bootstrap replicates.

Data visualization and statistical analysis

The visualization of phylogenies and their related information (host taxonomy, bacterial strain, *Wolbachia* supergroup, protein domains) was done using the online tool iTOL v4.3.3 (Letunic & Bork, 2007, https://itol.embl.de/). All statistical analyses were performed in the R software (R Core Team 2013). The congruence between the phylogenies of *cifA* and *cifB* homologs as well as their concatenation, and their respective *Wolbachia* strains were tested using Mantel tests on the pairwise patristic distances between sequences (1,000 permutations). Additionally, ParaFit with 9,999 permutations implemented in CopyCat v2.0 (Meier-Kolthoff et al. 2007) was used to examine cophylogenetic signals between trees and visualize the contribution of individual links between them. Finally, we compared the observed mean phylogenetic distances between *cif* gene pairs occurring within the same

genome to a random distribution of mean distances generated by randomly permuting *cif* genes between genomes (1,000 permutations).

Acknowledgements

We thank Prof John Jaenike, Prof Greg Hurst and Dr Bryant McAllister for providing some of the *Drosophila* lines used for sequencing. This work was supported by the Wellcome Trust (grant number WT094664MA and WT202888/Z/16/Z) and the European Research Council (grant number 281668, DrosophilaInfection). Raw sequencing reads are available at the Sequence Read Archive (Bioproject PRJNA603900) and the genome assemblies at Genbank (CP050530, CP050531, JAATKZ000000000, JAATLA000000000, JAATLB000000000, JAATLC000000000, JAATLD000000000, JAATLE000000000, JAATLF000000000, JAATLG000000000, JAATLH000000000).

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:W7-13.

Asselin AK, Villegas-Ospina S, Hoffmann AA, Brownlie JC, Johnson KN. 2018. Contrasting patterns of virus protection and functional incompatibility genes in two conspecific Wolbachia strains from Drosophila pandora. Appl. Environ. Microbiol. 85:1–14.

Atyame CM, Labbé P, Dumas E, Milesi P, Charlat S, Fort P, Weill M. 2014. Wolbachia divergence and the evolution of cytoplasmic incompatibility in Culex pipiens. PLoS One 9:21–26.

Baldo L, Ayoub NA, Hayashi CY, Russell JA, Stahlhut JK, Werren JH. 2008. Insight into the routes of Wolbachia invasion: high levels of horizontal transfer in the spider genus Agelenopsis revealed by Wolbachia strain and mitochondrial DNA diversity. Mol. Ecol. 17:557–569.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL. 2002. The Pfam protein families database. Nucleic Acids Res. 30:276–280.

Beckmann J, Ronau J, Hochstrasser M. 2017. A wolbachia deubiquitylating enzyme induces cytoplasmic incompatibility. Nat. Microbiol. 2:17007.

Beckmann JF, Bonneau M, Chen H, Hochstrasser M, Poinsot D, Merçot H, Weill M, Sicard M, Charlat S. 2019. The Toxin–Antidote Model of Cytoplasmic Incompatibility: Genetics and Evolutionary Implications. Trends Genet. 35:175–185.

Beckmann JF, Fallon AM. 2013. Detection of the Wolbachia protein WPIP0282 in mosquito spermathecae: Implications for cytoplasmic incompatibility. Insect Biochem. Mol. Biol.

752 43:867–878.

753 Beckmann JF, Sharma GD, Mendez L, Chen H, Hochstrasser M. 2019. Wolbachia
754 Cytoplasmic Incompatibility Enzyme CidB Targets Nuclear Import and Protamine-
755 Histone Exchange Factors. Elife 8:e50026.

756 Boetzer M, Henkel C V, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled
757 contigs using SSPACE. Bioinformatics 27:578–579.

758 Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol.
759 13:R56.

760 Bonneau M, Atyame C, Beji M, Justy F, Cohen-gonsaud M, Sicard M, Weill M. 2018. Culex
761 pipiens crossing type diversity is governed by an amplified and polymorphic operon in
762 Wolbachia genome. Nat. Commun. 9:319.

763 Bordenstein SR, O'Hara FP, Werren JH. 2001. Wolbachia-induced incompatibility precedes
764 other hybrid incompatibilities in Nasonia. Nature 409:707–710.

765 Bordenstein SR, Wernegreen JJ. 2004. Bacteriophage flux in endosymbionts (Wolbachia):
766 Infection frequency, lateral transfer, and recombination rates. Mol. Biol. Evol. 21:1981–
767 1991.

768 Callaini G, Dallai R, Riparbelli MG. 1997. Wolbachia-induced delay of paternal chromatin
769 condensation does not prevent maternal chromosomes from entering anaphase in
770 incompatible crosses of Drosophila simulans. J. Cell Sci. 110:271–280.

771 Cameron AC, Trivedi PK. 1990. Regression-based tests for overdispersion in the Poisson
772 model. J. Econom. 46:347–364.

773 Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated
774 platform for visualization and analysis of high-throughput sequence-based experimental
775 data. Bioinformatics 28:464—469.

776 Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. 2005. ACT:
777 the Artemis comparison tool. Bioinformatics 21:3422–3423.

778 Caspari E, Watson GS. 1959. On the Evolutionary Importance of Cytoplasmic Sterility in
779 Mosquitoes. Evolution 13:568–570.

780 Charlat S, Calmet C, Merçot H. 2001. On the mod resc model and the evolution of
781 Wolbachia compatibility types. Genetics 159:1415–1422.

782 Charlat S, Le Chat L, Merçot H. 2003. Characterization of non-cytoplasmic incompatibility
783 inducing Wolbachia in two continental African populations of Drosophila simulans.
784 Heredity 90:49–55.

785 Chen H, Ronau JA, Beckmann JF, Hochstrasser M. 2019. A Wolbachia Nuclease and Its
786 Binding Partner Comprise a Novel Mechanism for Induction of Cytoplasmic
787 Incompatibility. PNAS 116:22314–22321.

788 Cooper BS, Ginsberg PS, Turelli M, Matute DR. 2017. Wolbachia in the Drosophila yakuba
789 complex: Pervasive frequency variation and weak cytoplasmic incompatibility, but no
790 apparent effect on reproductive isolation. Genetics 205:333–351.

791 Cooper BS, Vanderpool D, Conner WR, Matute DR, Turelli M. 2019. Wolbachia acquisition
792 by drosophila yakuba-clade hosts and transfer of incompatibility loci between distantly
793 related wolbachia. Genetics 212:1399–1419.

794 Darling A, Mau B, Blattner F, Perna N. 2004. Mauve: multiple alignment of conserved
795 genomic sequence with rearrangements. Genome Res. 14:1394–1403.

796 Darling AE, Jospin G, Lowe E, Matsen IV FA, Bik HM, Eisen JA. 2014. PhyloSift:
797 phylogenetic analysis of genomes and metagenomes. PeerJ 2:e243.

798 Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

799 Ellegaard KM, Klasson L, Näslund K, Bourtzis K, Andersson SGE. 2013. Comparative
800 genomics of Wolbachia and the bacterial species concept. PLoS Genet. 9:e1003381.

801 Frank SA. 1997. Cytoplasmic Incompatibility and Population Structure. J. Theor. Biol.
802 184:327–330.

803 Frank SA. 1998. Dynamics of cytoplasmic incompatibility with multiple Wolbachia infections.
804 J. Theor. Biol. 192:213–218.

805 Gillespie JJ, Driscoll TP, Verhoeve VI, Rahman MS, Kevin R, Azad AF. 2018. A Tangled
806 Web : Origins of Reproductive Parasitism. Genome Biol. Evol. 10:2292–2309.

807 Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing.
808 Genome Res. 8:195–202.

809 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
810 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
811 performance of PhyML 3.0. Syst. Biol. 59:307–321.

812 Haygood R, Turelli M. 2009. Evolution of incompatibility-inducing microbes in subdivided
813 host populations. Evolution 63:432–447.

814 Hurst L, McVean GT. 1996. Clade Selection, Reversible Evolution and the Persistence of
815 Selfish Elements: The Evolutionary Dynamics of Cytoplasmic Incompatibility. Proc. R.
816 Soc. London Ser. B-Biological Sci. 263:97–104.

817 Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
818 prokaryotic gene recognition and translation initiation site identification. BMC
819 Bioinformatics 11:119.

820 Jaenike J. 2007. Spontaneous emergence of a new Wolbachia phenotype. Evolution
821 61:2244–2252.

822 Knizewski L, Kinch LN, Grishin N V., Rychlewski L, Ginalski K. 2007. Realm of PD-(D/E)XK
823 nuclease superfamily revisited: Detection of novel families with modified transitive meta
824 profile searches. BMC Struct. Biol. 7:1–9.

825 Koehcke A, Telschow A, Werren JH, Hammerstein P. 2009. Life and death of an influential
826 passenger: Wolbachia and the evolution of CI-modifiers by their hosts. PLoS One
827 4:e4425.

828 Kriesner P, Hoffmann AA. 2018. Rapid spread of a Wolbachia infection that does not affect
829 host reproduction in Drosophila simulans cage populations. Evolution 72:1475–1487.

830 Kriesner P, Hoffmann AA, Lee SF, Turelli M, Weeks AR. 2013. Rapid Sequential Spread of
831 Two Wolbachia Variants in Drosophila simulans. PLoS Pathog. 9:e1003607.

832 Laven H. 1957. Vererbung durch Kerngene und das Problem der ausserkaryotischen
833 Vererbung bei Culex pipiens. Z. Indukt. Abstamm. Vererbungsl. 88:443–477.

834 Lepage DP, Metcalf JA, Bordenstein Sarah R, On J, Perlmutter JI, Shropshire JD, Layton
835 EM, Funkhouser-Jones LJ, Beckmann JF, Bordenstein Seth R. 2017. Prophage WO
836 genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. Nature
837 9:243–247.

838 Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree
839 display and annotation. Bioinformatics 23:127–128.

840 Lindsey ARI, Rice D, Bordenstein S, Brooks A, Bordenstein S, Newton ILG. 2018.
841 Evolutionary Genetics of Cytoplasmic Incompatibility Genes cifA and cifB in
842 ProphageWO of Wolbachia. Genome Biol. Evol. 10:434–451.

843 Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis
844 of recombination patterns in virus genomes. Virus Evol. 1:1–5.

845 Mateos M, Castrezana SJ, Nankivell BJ, Estes AM, Markow TA, Moran NA. 2006. Heritable
846 Endosymbionts of Drosophila. Genetics 174:363–376.

847 Meany MK, Conner WR, Richter S V., Bailey JA, Turelli M, Cooper BS. 2018. Loss of
848 cytoplasmic incompatibility and minimal fecundity effects explain relatively low
849 Wolbachia frequencies in Drosophila mauritiana. Evolution:461574.

850 Meier-Kolthoff JP, Auch AF, Huson DH, Göker M. 2007. CopyCat: Cophylogenetic analysis
851 tool. Bioinformatics 23:898–900.

852 Miyake H, Watada M. 2007. Molecular phylogeny of the Drosophila auraria species complex
853 and allied species of Japan based on nuclear and mitochondrial DNA sequences.
854 Genes Genet. Syst. 82:77–88.

855 O'Neill SL, Karr TL. 1990. Bidirectional incompatibility between conspecific populations of
856 Drosophila simulans. Nature 348:178–180.

857 Pascar J, Chandler CH. 2018. A bioinformatics approach to identifying *Wolbachia* infections
858 in arthropods. PeerJ 6:e5486.

859 Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC.
860 2010. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes.
861 Nat. Methods 7:455–457.

862 Poinsot D, Bourtzis K, Markakis G, Savakis C. 1998. Wolbachia Transfer from Drosophila
863 melanogaster into D. simulans: Host Effect and Cytoplasmic Incompatibility
864 Relationships. Genetics 150:227–237.

865 Poinsot D, Charlat S, Merçot H. 2003. On the mechanism of Wolbachia- induced
866 cytoplasmic incompatibility : confronting the models with the facts. BioEssays 25:259–
867 265.

868 Prout T. 1994. Some evolutionary possibilities for a microbe that causes incompatibility in its
869 host. Evolution 48:909–911.

870 R Core Team. 2013. R: A Language and Environment for Statistical Computing.Pimenta PF,
871 editor.

872 Rousset F, Braig HR, O'Neill SL. 1999. A stable triple Wolbachia infection in Drosophila with
873 nearly additive incompatibility effects. Heredity 82:620–627.

874 Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: horizontal transfer of
875 transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol.

876 25:537–546.

877 Shropshire JD, Bordenstein SR. 2019. Two-By-One model of cytoplasmic incompatibility:
878 Synthetic recapitulation by transgenic expression of cifA and cifB in Drosophila. PLOS
879 Genet. 15:e1008221.

880 Shropshire JD, Kalra M, Bordenstein SR. 2020. Evolution-guided mutagenesis of the
881 cytoplasmic incompatibility proteins: Identifying CifA's complex functional repertoire and
882 new essential regions in CifB. bioRxiv:2020.05.13.093815.

883 Shropshire JD, Leigh B, Bordenstein Sarah R, Duplouy A, Riegler M, Brownlie JC,
884 Bordenstein Seth R. 2019. Models and Nomenclature for Cytoplasmic Incompatibility:
885 Caution over Premature Conclusions. Trends Genet. 35:397–399.

886 Shropshire JD, On J, Layton EM, Zhou H, Bordenstein SR. 2018. One prophage WO gene
887 rescues cytoplasmic incompatibility in Drosophila melanogaster. PNAS 115:4987–4991.

888 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO:
889 assessing genome assembly and annotation completeness with single-copy orthologs.
890 Bioinformatics 31:3210–3212.

891 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: A parallel
892 assembler for short read sequence data. Genome Res 19:1117-1123.

893 Sinkins SP, Braig HR, O'Neill SL. 1995. Wolbachia superinfections and the expression of
894 cytoplasmic incompatibility. Proc. R. Soc. B Biol. Sci. 261:325–330.

895 Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology
896 detection and structure prediction. Nucleic Acids Res. 33:W244–W248.

897 Tram U, Sullivan W. 2002. Role of Delayed Nuclear Envelope Breakdown and Mitosis in
898 Wolbachia-Induced Cytoplasmic Incompatibility. Science 296:1124-1126.

899 Turelli M. 1994. Evolution of incompatibility-inducing microbes and their hosts. Evolution
900 48:1500–1513.

901 Turelli M, Cooper BS, Richardson KM, Chiu JC, Conner WR, Hoffmann AA, Turelli M,
902 Cooper BS, Richardson KM, Ginsberg PS, et al. 2018. Rapid Global Spread of wRi-like
903 Wolbachia across Multiple Drosophila Report Rapid Global Spread of wRi-like
904 Wolbachia across Multiple Drosophila. Curr. Biol. 28:963–971.

905 Turelli M, Hoffmann AA, McKechnie SW. 1992. Dynamics of cytoplasmic incompatibility and
906 mtDNA variation in natural Drosophila simulans populations. Genetics 132:713–723.

907 Vautrin E, Charles S, Genieys S, Vavre F. 2007. Evolution and invasion dynamics of multiple
908 infections with Wolbachia investigated using matrix based models. J. Theor. Biol.
909 245:197–209.

910 Vavre F, Fleury F, Lepetit D, Fouillet P, Boulétreau M. 1999. Phylogenetic evidence for
911 horizontal transmission of Wolbachia in host-parasitoid associations. Mol. Biol. Evol.
912 16:1711–1723.

913 Watada M, Matsumoto M, Kondo M, Kimura MT. 2011. Taxonomic study of the Drosophila
914 auraria species complex (Diptera: Drosophilidae) with description of a new species.
915 Entomol. Sci. 14:392–398.

916 Weinert LA, Araujo-Jnr E V., Ahmed MZ, Welch JJ. 2015. The incidence of bacterial
917 endosymbionts in terrestrial arthropods. Proc. R. Soc. B Biol. Sci. 282:20150249–

20150249.

Zabalou S, Apostolaki A, Pattas S, Veneti Z, Paraskevopoulos C, Livadaras I, Markakis G, Brissac T, Merçot H, Bourtzis K. 2008. Multiple rescue factors within a Wolbachia strain. Genetics 178:2145–2160.

Zabalou S, Charlat S, Nirgianaki A, Lachaise D, Merçot H, Bourtzis K. 2004. Natural Wolbachia infections in the Drosophila yakuba species complex do not induce cytoplasmic incompatibility but fully rescue the wRi modification. Genetics 167:827–834.

Zhou W, Rousset F, O'Neill S. 1998. Phylogeny and PCR–based classification of Wolbachia strains using wsp gene sequences. Proc. R. Soc. B-Biological Sci. 265:509–515.

Figures and figure legends

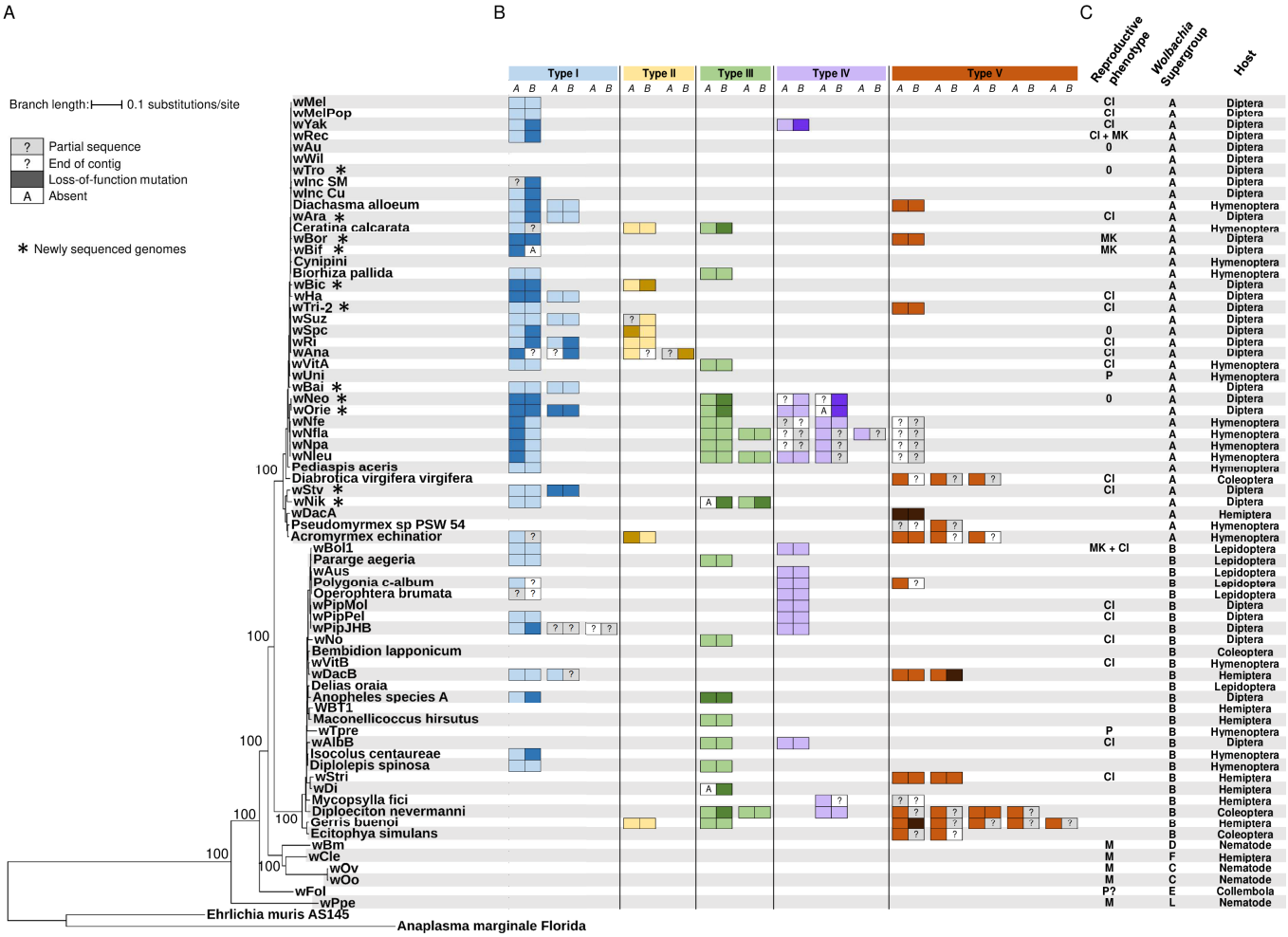
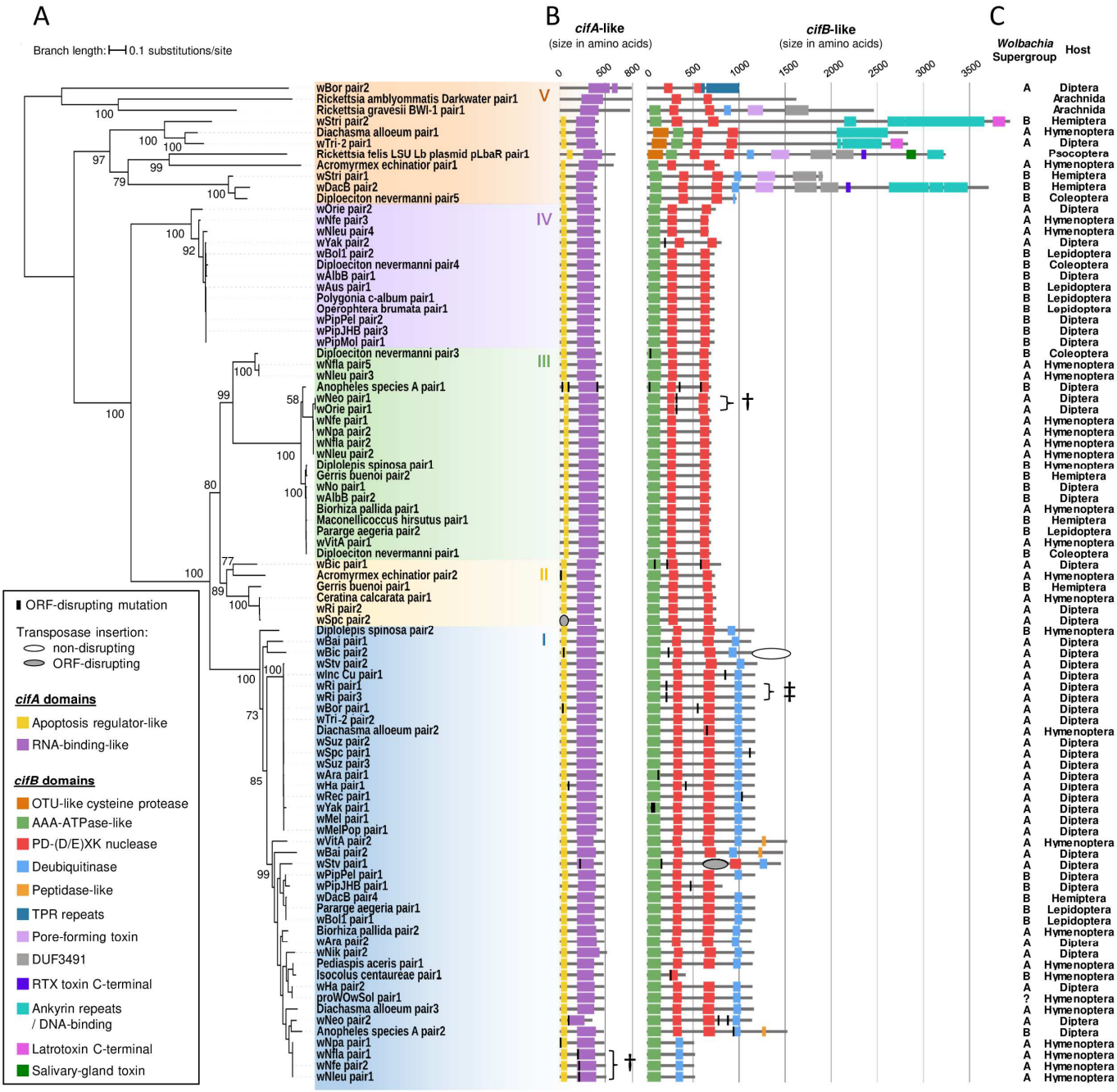


Figure 1. Distribution of *cifA* and *cifB* homologs across *Wolbachia* strains. (A) Maximum likelihood phylogeny of *Wolbachia* reconstructed from the concatenated sequences of 28 genes. Bootstrap values were estimated from 1,000 replicates. (B) *cifA* and *cifB* homologs. Each box represents a single gene and contiguous boxes indicate adjacent genes in the genome. Genes are organised according to Type, as defined from their genealogy (**Figures 2 and S1**). (C) Reproductive phenotypes (CI: cytoplasmic incompatibility; MK: male-killing; P: parthenogenesis; M: mutualism; 0: no reproductive phenotype), *Wolbachia* supergroups and hosts in which the *Wolbachia* was found. The reproductive phenotype is not shown for strains for which conflicting phenotypic data exist (*w*Yak and *w*Suz).



969 **Figure 2. Phylogeny and protein domain architecture of *cifA* and *cifB*.** (A) Maximum
970 likelihood tree of concatenated *cifA* and *cifB* nucleotide sequences. Partially sequenced *cif*
971 homologs were excluded. The tree is midpoint rooted. Bootstrap values were estimated from
972 1,000 replicates. (B) Protein domains. Mutations that disrupt the open reading frame (ORF)
973 are indicated by a vertical bar. Symbols indicate where the same ORF-disrupting mutation is
974 found in two homologs due to speciation (†) or duplication (‡) events. All domains had a
975 HHpred probability of being true-positives >75% in at least one sequence. The suffix '-like' at
976 the end of the domain name indicates that there were no sequences where the probability of
977 the domain being true was >95%. Details of the domains are in Table S5. (C) *Wolbachia*
978 supergroups and arthropod hosts from which the *cif* sequences were isolated.

979

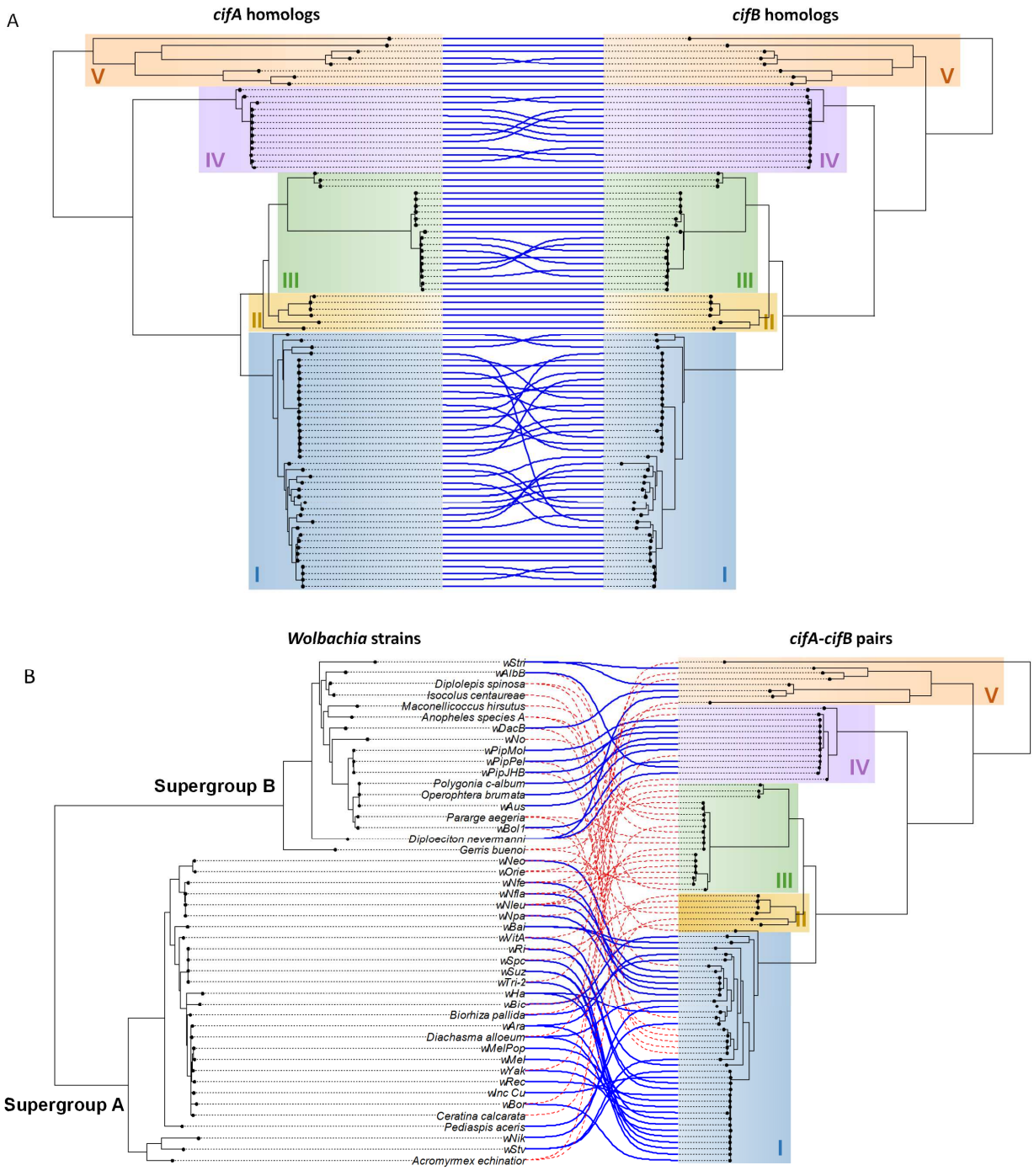


Figure 3. Phylogenetic congruence between *cif* genes and *Wolbachia* strains. (A) Cophylogeny of syntenic *cifA* and *cifB* genes. (B) Cophylogeny of *Wolbachia* genomes and the *cifA-cifB* genes. Links between phylogenies indicate (A) syntenic genes and (B) *Wolbachia* strain-*cif* gene associations respectively. Blue links have a significant contribution to the global cophylogenetic signal in the Parafit test and red links do not.

Tables

Table 1. Summary statistics of newly sequenced *Wolbachia* genomes.

Host	Strain	Mean sequencing depth	Genome size (bp)[a]	GC content (%)	*N* scaffolds	*N* contig	Scaffold N50	Contig N50	BUSCO complete	BUSCO fragment	BUSCO missing
D. arawakana	*w*Ara	30x	1,290,535	35.3	10	17	185,314	171,449	179	2	39
D. baimaii	*w*Bai	40x	1,119,646	35.3	26	123	76,530	13,921	180	2	38
D. bicornuta	*w*Bic	30x	1,177,727	35.1	24	34	71,402	49,062	182	1	37
D. bifasciata	*w*Bif	33x	1,187,580	35.1	17	26	122,861	113,299	183	3	34
D. borealis	*w*Bor	36x	1,210,092	35.3	16	39	146,013	52,110	183	2	35
D. neotestacea	*w*Neo	40x	1,353,942	35.2	19	26	124,304	93,317	182	3	35
D. nikananu	*w*Nik	27x	1,137,710	35.3	1	7	-	583,015	183	2	35
D. orientacea	*w*Orie	27x	1,359,726	35.2	19	30	103,124	73,532	181	4	34
D. sturtevanti	*w*Stv	30x	1,183,448	35.3	1	3	-	207,642	185	2	33
D. triauraria[b]	*w*Tri-2	33x	1,284,908	35.2	9	19	265,635	124,454	183	1	36
D. tropicalis	*w*Tro	30x	1,214,296	35.2	13	17	106,850	67,971	184	2	34

[a]Ungapped genome size [b] this stock was identified in (Mateos et al. 2006) and (Miyake and Watada 2007) as *D. quadraria* but (Watada et al. 2011) later concluded that *quadraria* is a junior synonym for *triauraria*.

Table 2. Association between CI and the presence of *cifA-cifB* genes in *Wolbachia* genomes.

	CI[a]	
cifA-cifB[b]	Yes	No
Present	14	1
Absent	1	10
Pseudogenes only	2	3

[a]*Wolbachia* strains for which there is no phenotypic information available or there are contradictory reports in the literature (*w*Suz) were discarded. [b] "Present" stands for *cif* genes without loss-of-function mutations; partially sequenced *cifA-cifB* pairs were discarded. *w*Ana and the strain found in *Diabrotica virgifera virgifera* induce CI but are excluded from the table as they only have partially sequenced *cif* genes, preventing us from inferring their pseudogenization status.

Supplementary Material

Figure S1. Phylogenies and protein domain architecture of individual *cifA* and *cifB*-like homologs. Maximum likelihood tree of *cifA* (A) and *cifB* (B) nucleotide sequences and their respective annotations of protein domains. Partially sequenced *cif* homologs were excluded. The trees are midpoint rooted. Bootstrap values were estimated from 1,000 replicates. †/‡: loss-of-function mutations co-inherited through putative speciation (†) or duplication (‡) events.

Figure S2. Tests of association between *cif* genes', *Wolbachia*'s phylogenies and insect orders. Phylogenetic congruence was tested using Mantel tests on pairwise genetic distances between cifA and cifB genes (A-B), *cifA*-*cifB* genes and *Wolbachia* genomes (C-D). In a similar approach, the association of *cifA*-*cifB* genes (E-F) and *Wolbachia* genomes (G-H) with insect orders was tested by random permutations (n = 1,000) of the insect orders. Significance was calculated by measuring the proportion of values in the null distributions that were higher than the observed correlation coefficients (red dotted lines).

Figure S3. Examples of sequence alignment between inferred *cifA-cifB* gene recombinants and their parental sequences. Full protein sequences of putative recombinant and their inferred parental sequences were aligned with Clustal Omega. Sites that were identical between all sequences were then removed from the alignment. Aligned variable sites are displayed with dots representing sites in minor (green) and major (blue) parental sequences that are identical to the recombinant sequence. Alternations of long green and blue stretches suggest the presence of putative recombination breakpoints (A-C), whereas no such pattern was observed in (D), indicating that the inferred recombination event could be a false-positive.

Figure S4. Null distribution of mean pairwise genetic distances between parental sequences in recombination events. Means were generated from permutation of pairwise distances between all *cifA*-*cifB* genes (n = 1,000). Significance was calculated by measuring the proportion of values in the null distribution that were lower than the observed mean genetic distance between inferred parents (red dotted lines).

Figure S5. Null distribution of mean pairwise genetic distances between *cifA-cifB* pairs occuring within the same genome. Means were generated from permutation of pairwise distances (n = 1,000). Significance was calculated by measuring the proportion of values in the null distribution that were higher than the observed mean genetic distance (red dotted lines).

Figure S6. Multiple alignments of PD-(D/E)XK nuclease domains. (A) N-terminal and (B) C-terminal amino acid sequences generated from the hmmscan search were aligned with Clustal Omega. The presence of PD-(D/E)XK catalytic residues is highlighted in red.

Figure S7. Unrooted maximum likelihood phylogeny of deubiquitinase domains.

Table S1. Summary of *Wolbachia* genomes used in this study.

Table S2. Summary of *cif* homologs.

Table S3. List of conserved genes used for *Wolbachia* phylogeny reconstruction.

Table S4. List of inferred recombination events between *cifA-cifB* gene pairs.

Table S5. List of *cif* protein domains inferred with iterative HMM search.