Comparative Genomics Reveals Factors Associated with Phenotypic Expression of *Wolbachia*

Guilherme Costa Baião[‡], Jessin Janice[†,‡], Maria Galinou, and Lisa Klasson *

Molecular Evolution, Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[†]Present address: Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway

[‡]These authors contributed equally to this work.

*Corresponding author: E-mail: lisa.klasson@icm.uu.se.

Accepted: 17 May 2021

Abstract

Wolbachia is a widespread, vertically transmitted bacterial endosymbiont known for manipulating arthropod reproduction. Its most common form of reproductive manipulation is cytoplasmic incompatibility (CI), observed when a modification in the male sperm leads to embryonic lethality unless a compatible rescue factor is present in the female egg. CI attracts scientific attention due to its implications for host speciation and in the use of *Wolbachia* for controlling vector-borne diseases. However, our understanding of CI is complicated by the complexity of the phenotype, whose expression depends on both symbiont and host factors. In the present study, we perform a comparative analysis of nine complete *Wolbachia* genomes with known CI properties in the same genetic host background, *Drosophila simulans* STC. We describe genetic differences between closely related strains and uncover evidence that phages and other mobile elements contribute to the rapid evolution of both genomes and phenotypes of *Wolbachia*. Additionally, we identify both known and novel genes associated with the modification and rescue functions of CI. We combine our observations with published phenotypic information and discuss how variability in *cif* genes, novel CI-associated genes, and *Wolbachia* titer might contribute to poorly understood aspects of CI such as strength and bidirectional incompatibility. We speculate that high titer CI strains could be better at invading new hosts already infected with a CI *Wolbachia*, due to a higher rescue potential, and suggest that titer might thus be a relevant parameter to consider for future strategies using CI *Wolbachia* in biological control.

Key words: Wolbachia, cytoplasmic incompatibility, Drosophila, genomics, comparative genomics, symbiosis.

Significance statement

The bacterium Wolbachia infects many different arthropods and affects their reproduction. The male sterility phenotype called cytoplasmic incompatibility (CI) is one of the most common forms of reproductive manipulation by *Wolbachia*. Although the main *Wolbachia* genes causing CI were discovered a few years ago, some aspects of CI might not be explained by these genes alone. In this article we compare *Wolbachia* genomes with known CI properties in a single insect species, *Drosophila simulans*, to exclude the host contribution to phenotypic expression. We find both known and new *Wolbachia* genes associated with CI and conclude that infection titer might be an important aspect to consider when using the *Wolbachia* CI phenotype for biological control of insects.

Introduction

Endosymbiotic bacteria are associated with most insects and contribute to the biology and evolution of their hosts. Among the most well studied of these endosymbionts is *Wolbachia*, which infects 40% of all arthropods and several filarial nematodes (Zug and Hammerstein 2012). Although only one species is currently recognized, *Wolbachia* strains show considerable diversity and are organized in several

supergroups (Lo et al. 2007; Lefoulon et al. 2020). *Wolbachia* can affect their hosts in different ways and are, for example, known to increase fecundity, longevity, fertility and provide protection against viruses (Hedges et al. 2008; Teixeira et al. 2008; Fast et al. 2011; Martinez et al. 2015). However, it is as a reproductive parasite that *Wolbachia* is best known, and its evolutionary success is often attributed to its efficacy in manipulating the host reproductive system to increase its own spread.

The most common and well-studied phenotype, cytoplasmic incompatibility (CI) is a form of sterility that results in embryonic mortality when an infected male mates with an uninfected female (unidirectional CI) or when a female and male carrying different and incompatible *Wolbachia* strains mate (bidirectional CI) (Werren et al. 2008). CI results in a reproductive advantage for infected females over uninfected females, which leads to an effective spread of the symbiont in host populations. These characteristics have made CI applicable for biological control of vector and pest insects (Flores and O'Neill 2018; Zheng et al. 2019). From an evolutionary perspective, bidirectional CI is implicated in host speciation, as it creates a reproductive barrier between individuals that are infected with incompatible strains (Bordenstein et al. 2001).

The phenotypic expression of CI is often described in terms of modification (*mod*) and rescue (*resc*) (Werren 1997). Modification occurs in the sperm of infected males before *Wolbachia* is shed. For offspring to be produced, the modified sperm have to fuse with a *Wolbachia*-infected egg containing a rescue factor. If the egg does not contain the correct rescue factor, the development will halt when the embryo enters the first mitotic division as a result of asynchrony between the paternal and maternal chromosomes (Tram and Sullivan 2002). The *mod* and *resc* functions are independent, since some strains can rescue but not modify, and *Wolbachia* strains can be classified based on their ability to exert them. Although all variations exist, most strains can either modify sperm and rescue their own modification (mod^{+} $resc^{+}$) or neither modify nor rescue (mod^{-} $resc^{-}$) (Poinsot et al. 2003; Zabalou et al. 2008).

Recently, the phage-associated genes *cifA* and *cifB* were shown to play a major role in the CI phenotype, although their exact functions in terms of *mod* and *resc* are still debated. While *cifB* is undoubtedly linked to *mod*, it is not clear if *cifA* is involved only in *resc* or both *mod* and *resc* (Beckmann et al. 2019; Shropshire and Bordenstein 2019). In the latter hypothesis, known as the two-by-one genetic model of CI, both *cifA* and *cifB* are required for causing modification, while *cifA* alone performs rescue when expressed at an appropriate level (Shropshire et al. 2018; Shropshire and Bordenstein 2019). Homologs of *cifA* and *cifB* have been identified in various *Wolbachia* strains as well as in a few other Rickettsiaceae species, and are classified into five Types (I–V) based on phylogenetic analyses (Lindsey et al. 2018; Martinez et al. 2021). The different *cif* Types show considerable variation in length and predicted protein domains, but are all expected to perform CI-associated functions (LePage et al. 2017; Martinez et al. 2021). Currently, experimental evidence for the ability of Type I and IV *cifAB* to induce and rescue CI exists (LePage et al. 2017; Chen et al. 2019). Interestingly, the *mod* function of Type I is associated with a deubiquitylase domain while that of Type IV is linked to a nuclease domain (Chen et al. 2019; LePage et al. 2017), suggesting that several distinct molecular mechanisms of CI might exist (Lindsey et al. 2018; Martinez et al. 2021). Recent evidence also imply that multiple domains of CifAB are likely involved in both *mod* and *resc* functions (Shropshire, Kalra et al. 2020). A strong correlation exists between strains carrying *cif* genes and those known to induce and rescue CI (LePage et al. 2017; Martinez et al. 2021) and generally strains carrying phylogenetically related *cif* genes also tend to be compatible with each other (Shropshire, Leigh et al. 2020). However, the *cif* genes do not explain all phenotypic variations of CI, especially not strength and bidirectional incompatibility between strains (Shropshire, Leigh et al. 2020). For example, the *w*Mel strain that only carries Type I *cifAB* genes can partially rescue the modification of *w*Ri, which only has a *cifB* gene of Type II (Charlat et al. 2004; Zabalou et al. 2008). Such cases suggest that CI phenotypic expression is also modulated by other genes and factors (Shropshire, Leigh et al. 2020).

Several mechanistic models have been proposed to explain CI *mod* and *resc* (Poinsot et al. 2003; Bossan et al. 2011; Beckmann et al. 2019; Shropshire et al. 2019). Poinsot et al. (2003) evaluated three different models and concluded that the "lock-and-key" best fit the knowledge at the time. This model suggests that the *mod* factor puts a lock on the paternal chromosome and a matching key, the *resc* factor, has to be present in the egg in order for the paternal chromosome to enter mitosis. The model requires that the *mod* and *resc* functions are unique and encoded by separate bacterial genes. Later, Bossan et al. (2011) combined the qualitative lock-and-key model with added quantitative parameters such as timing and expression, making the model fit better with observations. Currently, the Toxin-Antidote (TA) and Host-Modification (HM) models are the main mechanistic hypotheses for CI (Beckmann et al. 2019; Shropshire et al. 2019). The TA model is similar to lock-and-key and suggests that *Wolbachia* releases a toxin in the male sperm which must be counteracted by an appropriate antidote in the female egg (Beckmann et al. 2019; Hurst 1991). The HM model, on the other hand, postulates that *Wolbachia* modifies a host product in the male sperm which leads to embryonic mortality unless it is reversed by a *Wolbachia* factor in the egg (Shropshire et al. 2019).

Independently of the mechanistic model, it is clear that the phenotypic expression of CI as well as of other *Wolbachia* phenotypes not only depends on symbiont factors but also on the host. The same *Wolbachia* strain can, for example, cause male-killing in one host and CI in another (Sasaki et

al. 2005; Jaenike 2007) or induce a different strength of CI when transferred to a new host species. The latter is seen when the two strains *w*Mel and *w*Ri are transferred to each other's natural host. In its natural host *Drosophila melanogaster*, *w*Mel induces up to 30% embryonic mortality, whereas it causes almost 100% CI in *D. simulans* (Poinsot et al. 1998). The opposite effect can be seen for *w*Ri, which causes almost 100% embryonic mortality in its natural host *D. simulans* but only around 30% in *D. melanogaster* (Boyle et al. 1993). Similarly, the strains *w*Tei and *w*MelPop induce no or weak CI in their natural hosts (*D. teissieri* and *D. melanogaster*, respectively), but almost 100% embryonic mortality when transferred into *D. simulans* (McGraw et al. 2001; Zabalou et al. 2008). These examples also show that *D. simulans* is a host where many *Wolbachia* strains induce stronger CI than in other *Drosophila* species. The permissiveness of *D. simulans* is also reflected in the variety of *Wolbachia* strains that naturally infect this species, at least five, and in the many successful experimental transfers of *Wolbachia* from other hosts into *D. simulans* (Merçot and Charlat 2004). As a result, *D. simulans* is an important model for CI studies and phenotypic comparisons between *Wolbachia* strains (Merçot and Charlat 2004; Zabalou et al. 2008; Martinez et al. 2015).

In this article, we investigate *Wolbachia* genome evolution with a focus on CI-associated genes by using five newly sequenced (*w*San, *w*Yak, *w*Tei, *w*Au, and *w*Ma) and four previously available (*w*Ri, *w*No, *w*Ha, and *w*Mel) complete *Wolbachia* genomes. All nine strains have known *mod* and *resc* phenotypes in the *D. simulans* STC host background (Zabalou et al. 2008), and five of them naturally infect *D. simulans.* Among these five, three are mod^{+} $resc^{+}$ (*w*Ri, *w*Ha, and *w*No) and show variable CI strength, while two (*w*Ma and *w*Au) do not induce CI (mod^{-}). The non-CI inducers differ in their rescue properties, with *w*Au incapable of rescue ($resc^{-}$) while *w*Ma is able to rescue the modification of *w*No ($resc^{+}$). Three other strains, *w*San, *w*Yak, and *w*Tei (hereafter referred to as *w*SYT when mentioned collectively), naturally infect the species of the *Drosophila yakuba* group, *D. santomea*, *D. yakuba*, and *D. teissieri*, respectively. These are closely related and cause no to low CI in their natural hosts but show different CI strength after being transferred to *D. simulans.* In the new host, *w*San and *w*Yak continue to cause no or low CI while *w*Tei induces a strong incompatibility (Zabalou et al. 2004, 2008; Martinez et al. 2015; Cooper et al. 2017). The *w*SYT strains also differ in infection titer and compatibility with other strains, as *w*Tei has a higher titer than *w*SY (Martinez et al. 2015) and is capable of rescuing the modification of *w*Mel while *w*SY are not (Zabalou et al. 2008).

Thus, our data set focuses on a single host and includes closely related *Wolbachia* strains with distinct phenotypes, which creates a unique opportunity to identify *Wolbachia* factors associated with the specific traits of each strain.

Our results identify unique genetic features of our strains and highlight the importance of mobile elements for *Wolbachia* evolution, uncovering lateral gene transfers between *Wolbachia* strains as well as between *Wolbachia* and other organisms. By screening for CI-associated genes we recover the *cif* genes and identify novel candidate genes potentially associated with *mod* and *resc.* We discuss how CI-associated genes as well as symbiont titer may influence CI strength and compatibility between *Wolbachia* strains. Overall, this study contributes to further understanding of the CI phenotype and the genomic flexibility that allows *Wolbachia* to accumulate genetic changes with potentially major effects on both host and symbiont within short time scales.

Results

Genome Features and Strain Relationship

The five *Wolbachia* genomes that were completely sequenced in this study, *w*San, *w*Yak, *w*Tei, *w*Au, and *w*Ma, are circular and range in size between 1.27 and 1.41 Mbp (table 1). Their genome size and features are similar to the four previously sequenced *Wolbachia* genomes used in our comparative analyses (Wu et al. 2004; Klasson et al. 2009; Ellegaard et al. 2013; Sutton et al. 2014) as well as to many other sequenced *Wolbachia* genomes (Klasson et al. 2008; Newton et al. 2016; Sinha et al. 2019).

The *w*Au genome sequenced here differs by five SNPs and five indels from the *w*Au genome published by Sutton et al. (2014). All five SNPs are present in intergenic regions, of which four are in repeats. The five indels are all present in repeat regions, two of which cause pseudogenization of mobile elements in the genome published by Sutton et al. (2014). Even though the sequences of the two *w*Au genomes themselves are very similar, the annotation differs considerably, as we have used a different annotation pipeline followed by manual curation. For consistency, all of our analyses were done using the *w*Au genome sequence and annotation presented in this study.

In order to further increase the consistency of the annotations between our compared genomes, we also manually curated the *w*Mel annotation (numbers in parenthesis in table 1). Although *w*Mel is closely related to *w*SYT and *w*Au, the coding density in the original annotation of wMel is higher and the average gene length is shorter than in the *w*SYT and *w*Au genomes (table 1). These differences are mostly due to dissimilarities in pseudogene annotation (6% vs ~10%) and were alleviated by our manual curation. Furthermore, since one of our goals with the study is to identify candidates associated with phenotypic differences in a controlled host background, we also changed the gene sequences of *w*Mel in accordance with a *w*Mel strain that we sequenced after transinfection to *D. simulans* (see next section).

To establish a robust phylogeny between the nine *Wolbachia* strains (table 1), we clustered their proteomes and used the resulting 714 single-copy orthologous genes

Table 1

Genome Features of the Nine *Wolbachia* Genomes in This Study

Genomes	*w*Mel	*w*Au	*w*San	*w*Yak	*w*Tei	*w*Ri	*w*Ha	*w*Ma	*w*No
Supergroup	A	A	A	A	A	A	A	B	B
Genome size (Mbp)	1.27	1.27	1.41	1.39	1.35	1.45	1.30	1.27	1.30
GC (%)	35.2	35.2	35.2	35.2	35.2	35.2	35.1	34.0	34.0
Genes	1,199 (1,011)	996	1,120	1,102	1,069	1,150	1,009	1,006	1,042
Coding density	0.80 (0.76)	0.76	0.77	0.76	0.77	0.80	0.78	0.80	0.81
Avg. gene length (bp)	851 (958)	963	963	965	968	976	1,001	1,015	1,012
Pseudogenes	74 (111)	122	118	115	116	114	96	89	91
Phage WO (%)	8	9	14	13	10	8	8	6	8
ANK[a] (%)	3	3	3	3	2	3	3	6	7
Pseudogenes (%)	6	10	9	9	11	8	7	7	7
Repeats[b] (%)	9	10	17	15	11	23	10	4	4

[a]Ankyrin repeat domain proteins.

[b]Repeats longer than 300 bp with higher than 95% identity.

Numbers in parenthesis were obtained after manual curation of the *w*Mel original annotation.

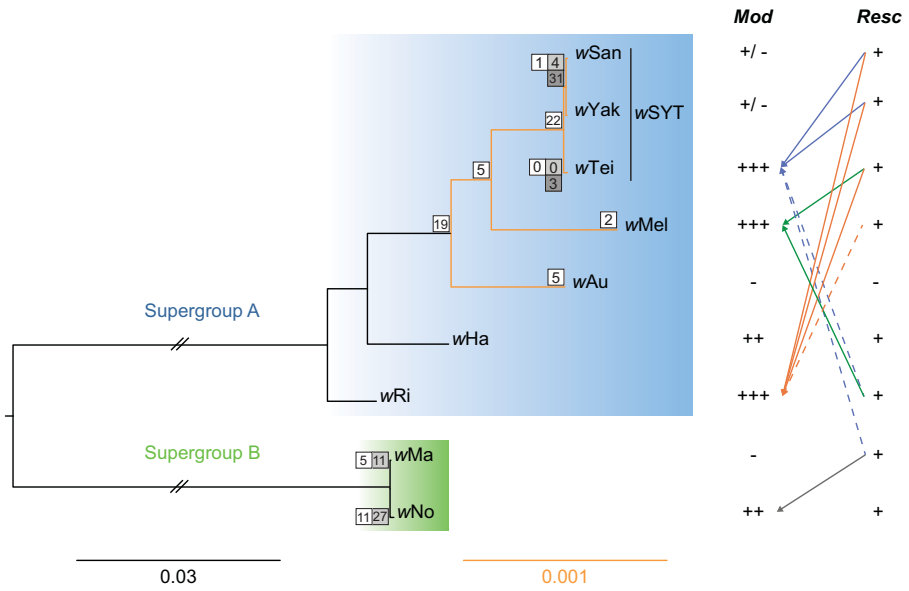

Fig. 1.—Phylogenetic relationship between the nine *Wolbachia* genomes in this study. Maximum likelihood tree based on the concatenated alignment of 714 orthologous single copy genes. All nodes have bootstrap values of 100. Branches of the SYTMA clade (in orange) were lengthened to more clearly show their internal relationships; they follow the orange scale bar. The branch connecting supergroups A and B was halved. The numbers shown on each node indicate the number of protein clusters that are unique to each subclade or node (white boxes), unique between the close relatives *w*SY and *w*Tei or *w*Ma and *w*No in comparison to each other (light gray boxes), or duplicated in *w*SY or *w*Tei in comparison to each other (dark gray boxes). The "Mod" and "Resc" columns show if a strain is capable (+) or uncapable (−) of *mod* and *resc* in *D. simulans*. CI strength is indicated as low (+), medium (++) or strong (+++). Colored arrows connecting the "Resc" and "Mod" columns show the ability of a strain to fully (full lines) or partially (dashed lines) rescue the modification induced by *w*Tei (blue), *w*Mel (green), *w*Ri (orange), and *w*No (gray). Data for CI strength, inter-strain compatibility, *resc* and *mod* phenotypes are summarized from Martinez et al. (2015) and Zabalou et al. (2008).

for phylogenetic reconstruction. A maximum likelihood tree based on the concatenated alignment of these genes showed 100% bootstrap support for all nodes (fig. 1). Although the branch lengths are very short, it is clear that *w*San and *w*Yak are most closely related followed by *w*Tei and that the *w*SYT genomes group together with *w*Mel to the exclusion of *w*Au. This result is in agreement with Cooper et al. (2019) and in contrast to Zabalou et al. (2008), who found *w*Au branching closest to *w*SYT.

Mutations after Transfer to D. simulans

In order to investigate what mutations might have occurred after transfer to a new host, we sequenced DNA from

multiple independent *Drosophila* lines which carried our supergroup A *Wolbachia* strains (table 1). Three separate *Drosophila* lines infected with *w*Tei and *w*Yak and two lines infected with *w*Au and *w*San (supplementary table S1, Supplementary Material online) were sequenced, the reads were mapped against their respective closed genome and variants were called. Additionally, one *D. simulans* line transinfected with *w*Mel was sequenced and compared to the publicly available *w*Mel genome (Wu et al. 2004). As a control, we also ran the pipeline with reads from the same line used to produce the reference genome.

In *w*Au, *w*Tei and *w*Yak, there were SNPs called between the reference and the Illumina reads used to create it (supplementary table S2, Supplementary Material online). In those positions, we found discrepancies between the PacBio and Illumina reads and we chose to call the sequence according to the PacBio reads. However, all such SNPs are present in intergenic regions, so their impact on our analyses is minimal.

In the comparisons between *Wolbachia* genomes from the same strain but different *Drosophila* lines, we found a few SNPs located mostly in intergenic regions (supplementary table S2, Supplementary Material online). Only in two of the comparisons did we observe mutations that would likely alter the function of a protein. First, the *w*Yak strain sequenced from its natural host, *D. yakuba*, had an indel that causes a frameshift in a gene encoding a permease. Since it is a loss-of-function mutation that is not present in the transinfected line nor in the published draft assembly of *w*Yak from *D. yakuba* (GCA_005862115.1), it is most likely that this mutation occurred in our sequenced *D. yakuba* line and not after *w*Yak was transferred to *D. simulans*.

Second, in the *w*Mel strain sequenced from *D. simulans*, we found indels in four genes, all coding for hypothetical proteins. We believe that all four might represent errors or possibly mutations that occurred in the published *w*Mel genome (Wu et al. 2004) rather than after wMel was transferred to *D. simulans*. Three of the indels restore the frame so that two short ORFs become one long (WD1043–WD1044, WD1215–WD1216, and WD1231–WD1232), possibly leading to functional restoration of the affected proteins. The last indel puts WD1155–WD1156 in the same frame, creating a new long putative gene that contains an in-frame stop codon. The resulting sequence is similar to other supergroup A genomes sequenced in this study, which also contain the same in-frame stop codon (*w*SYT and *w*Au). Hence, we believe that this might also be an error in the published *w*Mel genome or a mutation in the *w*Mel strain used.

Overall, we did not identify any parallel mutations between the genomes that have been transinfected into *D. simulans*, indicating that there is no strong selection on any particular protein as a result of the transfer to the new host background.

Genomic Variation between Close Relatives

Among the nine genomes compared in this study, there are two clades of very closely related *Wolbachia* strains, *w*SYT plus *w*Mel and *w*Au (hereafter SYTMA), and *w*No and *w*Ma (hereafter NoMa). To estimate the overall level of divergence between the *Wolbachia* strains within each of these clades, Illumina reads from each strain within SYTMA and NoMa were mapped against each genome within the clade and variants were called. Variants for each pair of strains were calculated twice, since the numbers vary slightly depending on which of the two genomes was used as reference (supplementary table S3, Supplementary Material online). Using the resulting SNP variants, we calculated the number of synonymous and nonsynonymous mutations and compared them to the frequency of nonsynonymous sites in the genomes, which was estimated to 76% in the 714 single-copy orthologs between all nine genomes. We classified the variants, both SNPs and indels, into three categories—genic, phage, and intergenic (supplementary table S3, Supplementary Material online)—based on their genomic location. Additionally, we analyzed gene content differences between the most closely related genomes, *w*SYT and NoMa, and proteins that were uniquely present in the genomes of the SYTMA clade.

Variation between the wSYT Strains

We found the three *w*SYT genomes to be extremely similar, differing only by 32–68 SNPs and 4–12 indels, thus making them 99,995% identical to each other in sequence (supplementary table S3, Supplementary Material online). We observed that mutations were slightly underrepresented in the prophage WO regions, with only ca 5% of the total number of SNPs even though phage WO regions make up ca 10% of the genomes. Additionally, the genic SNP pattern indicated that purifying selection might not have had enough time to act, as the frequency of non-synonymous mutations (75–80%) was close to neutrality (76%). Even so, there is an apparent overrepresentation of substitutions in intergenic regions (50–60%).

When analyzing our protein clusters, we did not find any cluster that was unique to either one of the three *w*SYT genomes, further emphasizing the close relationship between these three *Wolbachia* strains. Only five protein clusters were present in *w*SY to the exclusion of *w*Tei, even though the genomes of *w*SY are clearly larger. Three of the five clusters contain phage WO proteins (supplementary table S4, Supplementary Material online). Only one protein is unique to *w*SY among our nine clustered proteomes (fig. 1, supplementary table S4, Supplementary Material online) and it is found as a pseudogene in *w*Tei (located in *Dozen Island* described below). A total of 31 clusters contain more copies in *w*SY than in *w*Tei (fig. 1, supplementary table S4, Supplementary Material online), of which 29 are associated

with phage WO and one is a putative non-WO phage terminase. The higher number of prophage proteins in the *w*SY genomes agrees with the larger proportion of phage sequences in their genomes (table 1) and also explains the larger genome sizes of *w*SY compared to *w*Tei. Only three protein clusters have more copies in *w*Tei than in *w*SY (fig. 1, supplementary table S4, Supplementary Material online), a transposase, a Group II intron, and CifB, which is discussed in more detail in the next section.

Additional variation between the three *w*SYT genomes exists in the copy number of an IS-element as well as in reverse transcriptase and the phage WO associated major tail sheath protein.

Finally, in contrast to the very low number of mutations and few gene content differences, we observed that gene order is highly variable between the *w*SYT genomes (supplementary fig. S1, Supplementary Material online).

Variation between the NoMa Strains

In the comparison between the NoMa genomes, we called approximately 1,000 SNPs and 90 indels, making them 99.925% identical in sequence (supplementary table S3, Supplementary Material online). Looking at the distribution of SNPs across the genomes, we found that the phage WO SNPs were very slightly overrepresented (10%).

We identified 16 protein clusters present in *w*Ma to the exclusion of *w*No and 38 clusters in *w*No that were absent from the *w*Ma genome (fig. 1, supplementary table S4, Supplementary Material online). These clusters are largely made up of Ankyrin repeat containing proteins (9), hypothetical proteins (21) and phage WO proteins (12) (supplementary table S4, Supplementary Material online). Very few of the genes that differ between the NoMa genomes are unique to either *w*Ma or *w*No. Instead, they are also present in other *Wolbachia* genomes, either in our set of genomes or in others.

Variation in the SYTMA Clade

The more distant genomes of the SYTMA clade have about 99.8% overall sequence identity, with a total of 2,108–3,202 SNPs and 234–457 indels (supplementary table S3, Supplementary Material online). When classifying the SNPs based on the different genomic regions, it is clear that they are not randomly distributed in the genomes. We observed a strong overrepresentation of SNPs in the phage WO regions, which contain approximately 40–50% of all SNPs even though they represent only 10–15% of the genomes. This overrepresentation reflects the nonorthologous nature of several of the phage WO regions between the genomes (supplementary fig. S2, Supplementary Material online). Additionally, we observed a lower frequency of nonsynonymous mutations in genes located in phage WO regions than in genes outside phage WO regions. The frequency of nonsynonymous substitutions is only 35–40% in genes located in the phage WO regions, but around 65–69% in genes outside. Thus, the frequency of nonsynonymous substitutions is much lower in genes from the phage WO regions compared to the overall estimated frequency of nonsynonymous sites in single copy orthologs (76%). Such result indicates that selection might have acted during a longer time on the divergent and nonorthologous phage WO sequences (when they were present in other genomes), resulting in a lower ratio of nonsynonymous to synonymous substitutions. Taken together, our analysis of SNPs suggests that the genes outside phage WO regions likely represent the "true" divergence between the genomes, making the overall similarity between them much higher than 99.8%.

We found 19 protein clusters that were exclusive to the SYTMA genomes (fig. 1, supplementary table S4, Supplementary Material online). Twelve of them contain hypothetical proteins and include the *w*Mel proteins WD0353 and WD0811, which were both seen to affect the growth of yeast cells (Rice et al. 2017). However, none of the proteins in the 19 clusters were unique to this clade when compared to other *Wolbachia* genomes.

Among the SYTM genomes, we identified five unique clusters (fig. 1, supplementary table S4, Supplementary Material online), two of which were not found in any other *Wolbachia* genome. Three of these proteins are located in the "Octomom" region of *w*Mel (Chrostek et al. 2013), which is further analyzed below.

Finally, we identified 22 protein clusters that were unique to *w*SYT (fig. 1, supplementary table S4, Supplementary Material online). A majority of these were located in two regions of the genome. One is a phage WO copy that is divergent from the other genomes in our clustering (supplementary fig. S2, Supplementary Material online) and the other is a region with mostly hypothetical proteins that we call "*Dozen Island*" (described below).

The Octomom Region

The Octomom region contains eight genes in *w*Mel and is involved in over-replication and pathogenicity of the *w*MelPop strain (Chrostek and Teixeira 2015). It was previously noted as missing from the *w*Au genome (Iturbe-Ormaetxe et al. 2005) and two of the proteins were shown to have been laterally transferred between *Wolbachia* and mosquitoes (Klasson, Kambris et al. 2009; Woolfit et al. 2009). In *w*SYT, the Octomom genes are directly flanking one of the phage WO regions (fig. 2*A*), similar to what was seen in the supergroup B strain *w*Pip from *Culex* mosquitoes (Klasson, Kambris et al. 2009). Phylogenetic reconstruction of one of the proteins (WD0513 in *w*Mel) shows that the *w*SYT proteins are most closely related to *w*Mel (fig. 2*B*), although the divergence of this gene between *w*SYT and *w*Mel is much

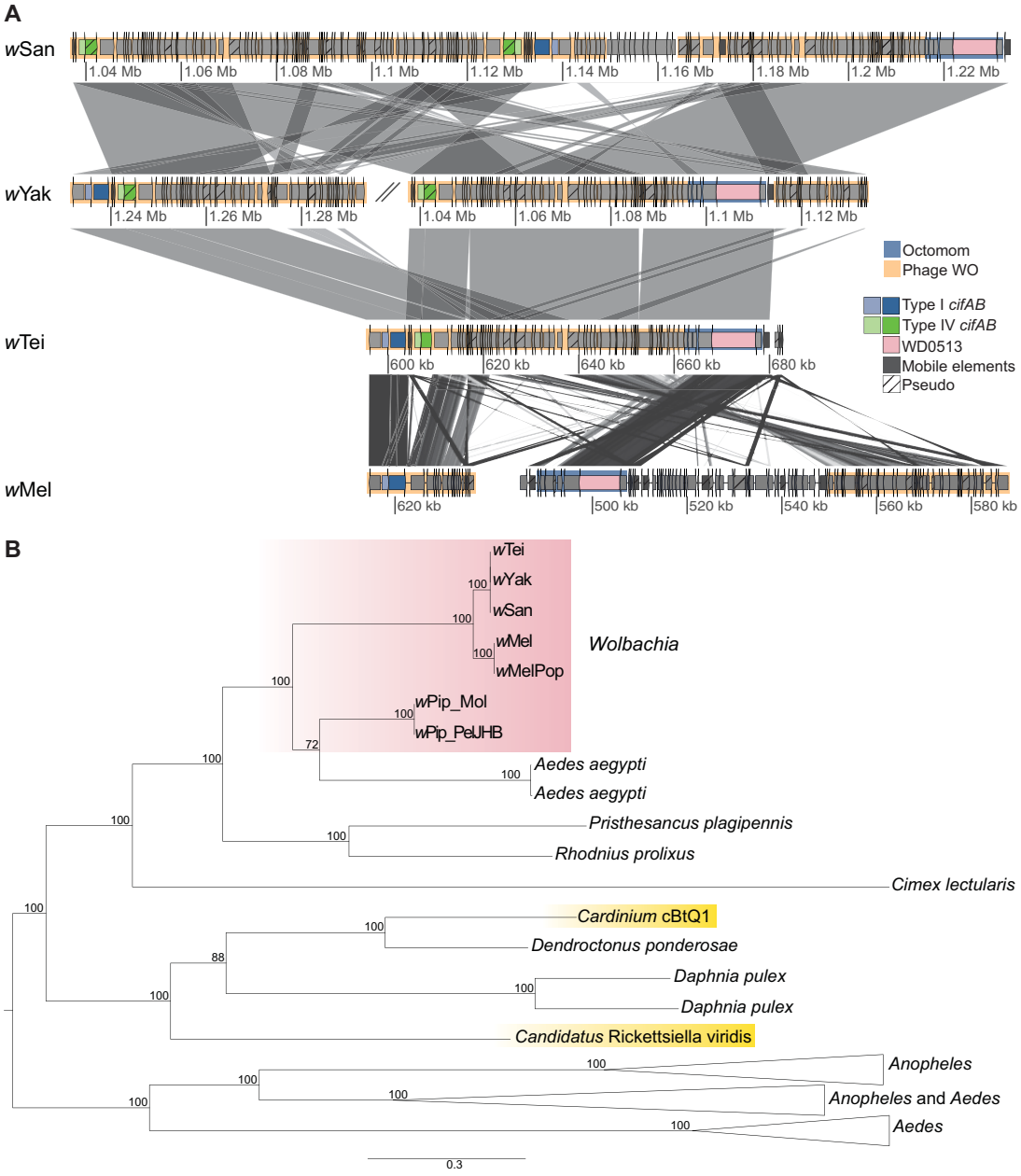

Fig. 2.—The Octomom and neighboring phage regions. (A) Comparison of the Octomom (blue) and neighboring phage regions (orange) in the SYTM genomes. The *w*Mel WD0513 gene and its homologs in *w*SYT are shown in pink. Other genes are represented in light gray and mobile elements in dark gray. Pseudogenes are marked by diagonal lines. Similarity between sequences is indicated by gray lines, where darker is more similar. Blastn was used for comparisons between *w*SYT genomes and tblastx was used for the comparison between *w*Tei and *w*Mel. (B) Maximum likelihood tree of the WD0513 protein of wMel and homologs from the *w*SYT genomes as well as other species identified through blast searches in the nr database. Bootstrap values are shown on nodes. The tree was midpoint-rooted in Figtree. Accession numbers for the proteins featured in the tree are available in supplementary table S5.

higher than most other parts of the genomes (ca. 15% of all SNPs and 6% of all indels).

We also identified homologs of WD0513 in two other bacterial symbionts, the reproductive manipulator *Cardinium* (Zchori-Fein and Perlman 2004; Schön et al. 2019), and the aphid endosymbiont "*Candidatus* Rickettsiella viridis" (Tsuchida et al. 2010) as well as in several Hemiptera. These new Hemiptera homologs branch outside of the *Wolbachia* clade and sit on long branches (fig. 2*B*), but are still closer to *Wolbachia* than to any of the other symbionts. The only non-*Wolbachia* protein that goes inside the *Wolbachia* clade is one from *Aedes aegypti* (fig. 2*B*), which was previously described by Klasson, Kambris et al. (2009). The phylogenetic position of the WD0513 homologs from *Cardinium* and "*Candidatus*

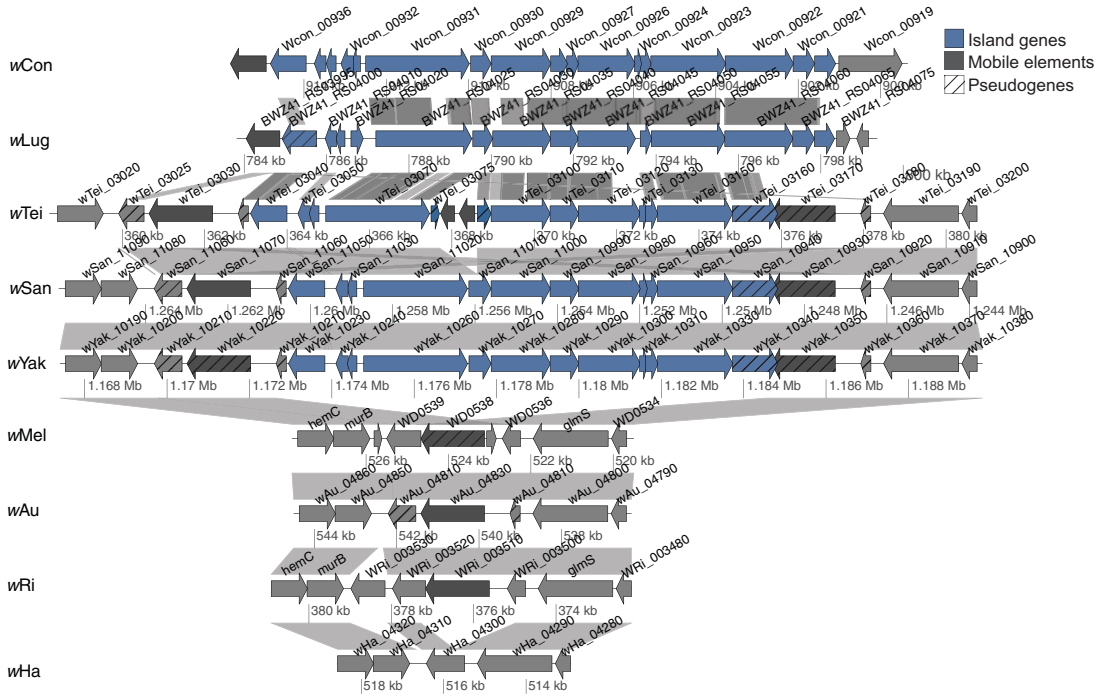

Fig. 3.—The *Dozen Island*. The *w*SYT *Dozen Island* genes and their homologs in *w*Con and *w*Lug are shown in blue. Other genes are shown in light gray and mobile elements in dark gray. Pseudogenes are marked by diagonal lines. Similarity between sequences is indicated by gray lines, where darker is more similar.

Rickettsiella viridis" suggests lateral transfers between these symbionts and their putative eukaryotic hosts.

The Dozen Island

The Dozen Island region in *w*SYT contains twelve genes (fig. 3) and is flanked by a 2.8 kbp repeat that includes a Group II intron and a degraded transposase. Only one protein of the 12, a putative addiction module toxin, clusters together with proteins from the other genomes. One additional protein contains a known protein domain, the C-terminal domain of DnaB-like helicase (PF03796). The remaining 10 proteins have no hits to known protein domains or to any non-*Wolbachia* genome.

We identified five other *Wolbachia* genomes that contain proteins with significant similarity to Dozen Island. The genomes of two supergroup B strains, *w*Con and *w*Lug, both contain a region that is highly similar in content to the *w*SYT Dozen Island (fig. 3). Additionally, three other *Wolbachia* genomes, *w*Cle of supergroup F, *w*DacA of supergroup A and *w*Stri of supergroup B, have regions with significant similarity (supplementary fig. S3, Supplementary Material online) but with many pseudogenized proteins. Dozen Island is immediately flanked by genes with similarity to mobile elements on at least one side in all genomes (fig. 3), and one of the genes in *w*Con (Wcon_09220) and *w*Lug (BWZ41_RS04065) (pseudogenized in *w*SYT) has low similarity to a phage portal protein.

Dozen Island is missing in the genomes of all of our other supergroup A strains (fig. 3). However, the gene order flanking the region is conserved between them except one of the ends in *w*Tei (fig. 3). Hence, the most parsimonious explanation is that Dozen Island entered the *w*SYT genomes through lateral gene transfer.

Interesting to note is that the plasmid pWCP, found in some *w*Pip strains, also contains putative toxin–antitoxin systems, a protein with the C-terminal domain of DnaB and a transposon (Reveillaud et al. 2019). However, we did not find any other homologous proteins between Dozen Island and the pWCP plasmid. We observe that in the draft genome of *w*Con, Dozen Island is located on a relatively small contig containing the same transposase at both ends. This suggests that the contig could represent an extrachromosomal circular DNA molecule, such as a plasmid.

Genetic Variation Associated with CI

The cif *Genes*

To investigate how the *cif* genes correlate with the CI properties of our strains (fig. 1), we identified the Cif proteins in our genomes and performed phylogenetic reconstructions. We included the Cif proteins from the incomplete genomes of three additional strains with known CI phenotypes in *D. simulans* (*w*Ara, *w*Stv, and *w*Tri-2) (Martinez et al. 2015). Additionally, to get a good representation of the different

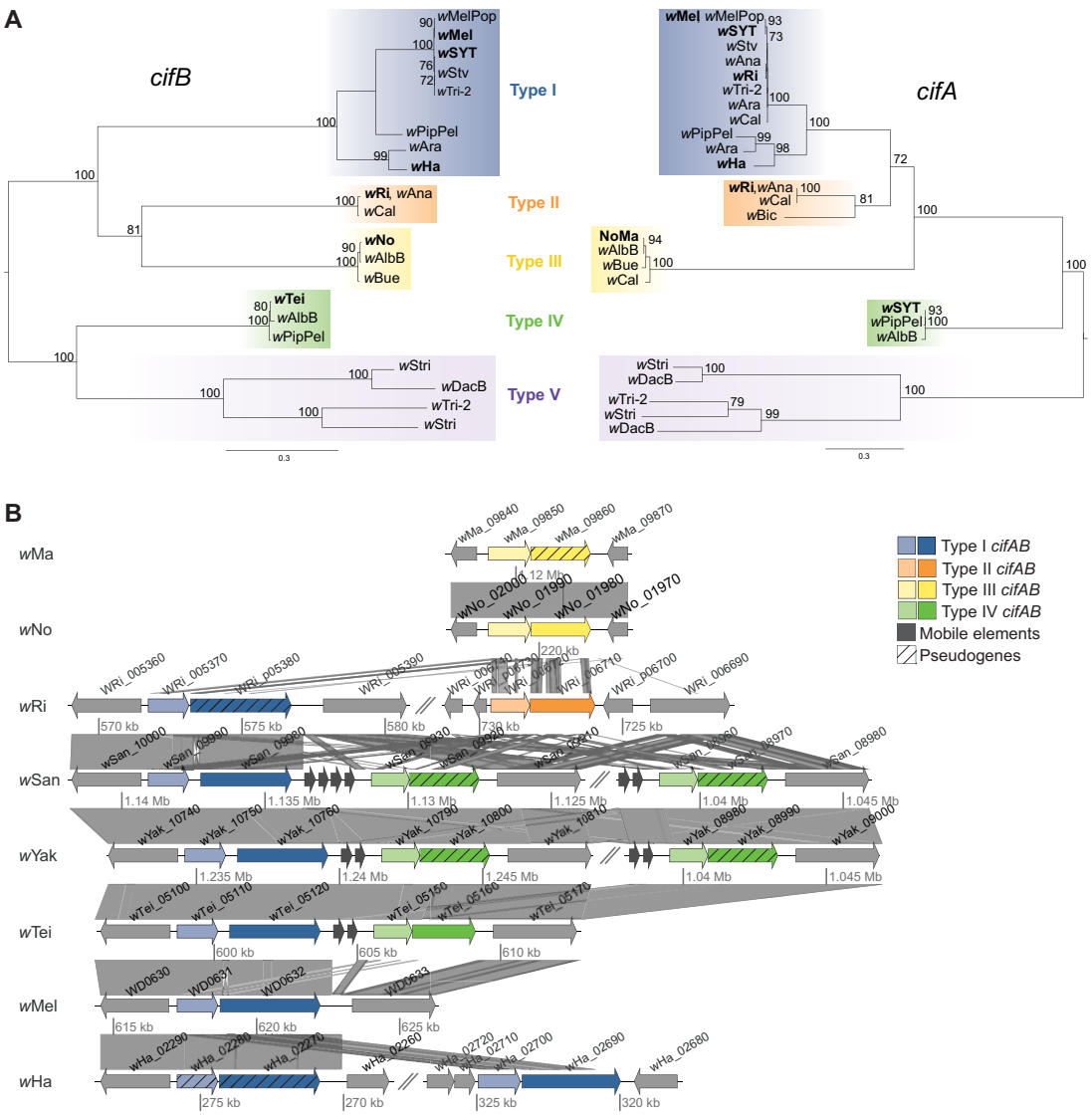

Fig. 4.—The CI-associated *cifA and cifB* genes. (A) Maximum likelihood trees of CifB and CifA homologs showing five clades that correspond to Types I–V, as indicated by labels and colors. Homologs found in our nine genomes are shown in bold. Trees were midpoint-rooted in Figtree. Bootstrap values below 70 are not shown. (B) Comparison of *cifAB* homologs in our nine *Wolbachia* genomes. Distinct colors identify *cifAB* homologs of different Types, with Type I in blue, Type II in orange, Type III in yellow, and Type IV in green. For each *cifAB* pair, *cifA* is shown in a lighter tone than *cifB*. Other genes are represented in light gray and mobile elements in dark gray. Pseudogenes are marked by diagonal lines. Similarity between sequences is indicated by gray lines, where darker is more similar.

Cif Types in the tree, we included Cif proteins from other complete *Wolbachia* genomes.

We found CifB proteins in the genomes of all *mod*$^+$ strains. They belonged to four of the different types (Types I–IV), with Type I proteins found in *w*SYT, *w*Mel and *w*Ha, Type II in *w*Ri, Type III in *w*No, and Type IV in *w*Tei (fig. 4). The predicted catalytic sites of the Type I deubiquitylase and Types II–IV nuclease domains were found to be preserved in all copies (Kosinski et al. 2005; Beckmann et al. 2017). Among the genomes of the *mod*$^-$ strains, the *cifB* gene is completely absent from *w*Au and pseudogenized by a point mutation in *w*Ma.

CifA homologs were found in all strains except in the *resc*$^-$ strain *w*Au. The phylogenies of CifA and CifB were highly congruent (fig. 4*A*) and genomes that contain a *mod* factor of one type also contain the *resc* factor of the same type. Additional CifA proteins that did not have a corresponding CifB protein, due to pseudogenization, were also found in some genomes, for example in *w*SY and *w*Ri (fig. 4).

Among our genomes, *w*Ma is the only strain that is unquestionably *mod*$^-$*resc*$^+$, as it can rescue the modification of *w*No but not itself induce CI (fig. 1). Hence, the presence of a CifA protein in *w*Ma that is identical to the CifA of *w*No makes sense.

The *cif* Genes of wSYT

The very closely related *w*SYT strains differ in their CI properties (fig. 1), with *w*Tei inducing stronger CI and being able to rescue more strains than *w*SY. Hence, we expect differences between the Cif proteins in their genomes if these are the only determinants of CI.

We observed that *w*Tei contains both a Type I and a Type IV CifB protein, while the *w*San and *w*Yak genomes encode one Type I CifB protein plus two Type IV *cifB* genes that are both pseudogenized by frameshift mutations (figs. 2*A* and 4*B*). Since the Type I CifB proteins are identical between the wSYT genomes and the Type IV CifB is probably nonfunctional in *w*SY, it is most likely the Type IV CifB protein in *w*Tei that causes strong CI in *D. simulans*. Additionally, the Type I CifB proteins in *w*SYT are 112 amino acids shorter than the CifB protein of *w*Mel. This N-terminal truncation is due to an inversion (supplementary fig. S4, Supplementary Material online), also noted by Cooper et al. (2019) and Martinez et al. (2021), that might have occurred via homologous recombination of a small inverted repeat. An AAA-ATPase-like domain was previously predicted in the truncated part of the protein in *w*Yak, as well as in all other Type I–IV CifB proteins (Martinez et al. 2021). This domain might thus be important for CifB function given that the truncation has rendered the Type I CifB proteins of *w*SYT either nonfunctional, based on the lack of CI induction by *w*SY reported in Martinez et al. (2015), or not very effective in inducing CI, based on the results of Cooper et al. (2017) and Zabalou et al. (2008).

For CifA, each *w*SYT genome has one Type I protein plus either one Type IV protein (*w*Tei), or two identical Type IV proteins (*w*SY) (figs. 2*A* and 4*B*). The results from Zabalou et al. (2008) suggest that the rescue properties of *w*Tei and *w*SY are different, with *w*Tei being able to rescue the modification of *w*Mel while the *w*SY strains are not. Hence, even though these CifA proteins are most likely involved in rescue, they cannot explain the differences in rescue potential between the *w*SYT genomes.

Origin and Movement of *cif* Genes in wSYT

Using draft genomes, Cooper et al. (2019) suggested that the Type IV *cifAB* genes of *w*SYT might have been transferred laterally by the aid of flanking IS-elements (ISWpi1). Similar to Cooper et al. (2019) we found that *w*Yak, as well as *w*Tei and *w*San, have ISWpi1 elements between the Type I and Type IV *cifAB* loci (figs. 2*A* and 4*B*). However, when comparing the *w*SYT genomes, we did not find ISWpi1 elements in the same location at the other end. Additionally, the *w*SYT genomes are not syntenic through this phage WO region. At least one duplication followed by several rearrangements of this region must have occurred, which makes it hard to infer a detailed scenario (figs. 2*A* and 4*B*). However, we observe that the phage WO copy associated with the two Cif loci appears complete in *w*Tei, and that the phage WO copy connected to the single Type IV *cifAB* locus in *w*Yak has the same content and gene order as *w*Tei (fig. 2*A*). The same phage WO region is also flanked by Octomom at its other end, after which another ISWpi1 copy exists in *w*Tei (fig. 2*A*). Given that the phage WO copy found in connection with the Type IV *cifAB* genes appears complete in *w*Tei and *w*Yak, and that the proteins have a relatively consistent phylogenetic position throughout its full extent (figs. 2 and 4), it is most likely that the Type IV *cifAB* genes as well as the Octomom region entered the *w*SYT genomes via a WO phage rather than via recombination of a DNA segment flanked by ISWpi1 elements. The close relationship between *w*SYT and *w*Pip for both the WD0513 homologs and Type IV CifAB proteins make this hypothesis highly plausible and suggests a supergroup B origin of the Type IV *cifAB* locus in *w*SYT. The presence of Octomom in *w*Mel might indicate that the same WO phage was present in the ancestor of SYTM. If so, the Type IV *cifAB* genes and most phage WO genes must have been lost from *w*Mel.

Other Genes Associated with mod *and* resc

Mod Candidates

To identify additional proteins associated with modification, we looked for clusters of proteins that were present in all CI-inducing strains but absent from non-CI-inducing strains (fig. 1). No protein clusters were identified by treating *w*SY as *mod*$^-$ together with *w*Au and *w*Ma. However, when treating *w*SY as *mod*$^+$, we identified two clusters (table 2), one with the CifB proteins and one containing hypothetical proteins homologous to *w*Mel WD0462. The latter cluster contains one protein from each CI strain but not from *w*Ma and *w*Au, where the gene is pseudogenized (fig. 5). In several strains, this protein contains the HAUS Augmin-like complex subunit 3, N-terminal domain (PF14932) (fig. 5).

We further checked the link between WD0462 homologs and the CI phenotype by analyzing the status of this gene in the draft genomes of *Wolbachia* strains *w*Ara, *w*Stv, *w*Tri-2, and *w*Tro, all of which have known CI phenotypes in *D. simulans* (Martinez et al. 2015). Our predictions are met in the *mod*$^+$ strains *w*Ara and *w*Stv, which have complete WD0462 homologs, and in the *mod*$^-$ *w*Tro, which has a pseudogenized copy. Only *w*Tri-2 does not follow our prediction, as this strain is *mod*$^+$ but has a truncated and split WD0462 gene. Interestingly, the neighboring gene WD0463 is a distant homolog of WD0462 that also varies significantly between strains (fig. 5) and occasionally contains an AAA-ATPase domain (PF00004) (fig. 5).

To identify more potential *mod* candidates, we searched for genes that are divergent between the CI and non-CI (or low CI) genomes. We primarily considered genes that contained substitutions that separated the most closely related strains with different phenotypes, *w*SYT and NoMa. Out of our 714 single copy protein clusters, we identified three genes

Table 2
Proteins Associated with the CI Phenotype.

Protein	*w*Mel Locus tag
Proteins associated with modification	
Presence/absence	
CifB	WD0632
Hypothetical protein	WD0462
Divergence	
DNA directed RNA polymerase, beta/beta' subunits	WD0024
NADH-quinone oxidoreductase subunit H	WD0159
Acetyl/propionyl-CoA carboxylase, alpha subunit	WD0433
Genes with upstream variation in *w*SYT	
Ankyrin repeat domain protein	WD0292
Hypothetical protein	WD0403
S-adenosylmethionine synthase	WD0136
Aspartate-semialdehyde dehydrogenase	WD0954
Proteins associated with rescue	
Presence/absence	
CifA	WD0631
Phage replication protein RepA	WD0582, WD0609
Hypothetical protein	WD1187
DNA recombination-mediator protein A	WD0092
Divergence	
M16 family peptidase	WD0762
Folylpolyglutamate synthase	WD1052

that follow a pattern of divergence that could make them associated with *mod*, that is, they were identical between *w*SY but with at least one nonsynonymous substitution compared to *w*Tei and there was at least one nonsynonymous substitution between *w*Ma and *w*No (table 2). Based on parsimony, the substitution in NADH-quinone oxidoreductase subunit H occurred in *w*Ma and *w*Tei; in DNA directed RNA polymerase, beta/beta' subunits (*rpoBC*) the substitution occurred in *w*No and *w*Tei; and in acetyl/propionyl-CoA carboxylase, alpha subunit four substitutions were exclusive to *w*No, one to *w*Ma and one to *w*SY. None of these substitutions occur at the same positions in the different strains. Given the putative functions of these proteins, the very low divergence between strains with different phenotypes and the lack of parallel mutations, we believe that none of these genes are likely to be involved in *mod*. However, we note that the two CI-inducers, *w*Tei and *w*No, both have mutations in *rpoBC* and that the RpoBC protein was found in the ovaries of *Culex* infected with a CI-inducing *Wolbachia* (LePage et al. 2017).

We also used the *w*SYT genomes to look for mutations that might be involved in regulating expression levels by analyzing the upstream region of genes. We found mutations that differentiate *w*SY and *w*Tei in upstream regions of only six genes. Four of the genes are present in all CI-inducing strains (table 2). Since noncoding regions are much less conserved, we could not infer in which *Wolbachia* strains these mutations took place.

Resc Candidates

To identify potential rescue candidates, we looked for clusters that contained proteins from all genomes except *w*Au, since *w*Au is the only strain in our analysis that is unable to rescue the modification of any other strain. We identified four protein clusters that contained at least one protein from all genomes except *w*Au, where the genes were pseudogenized or lost (table 2).

One of the proteins is CifA, which was described above. Another phage WO protein that we found is the multifunctional phage replication protein RepA (Mardanov and Ravin 2006), which is present in one copy in each supergroup B strain and several copies in all supergroup A strains (supplementary fig. S5, Supplementary Material online). The third candidate, hypothetical protein WD1187, is present in one copy in each *resc*$^{+}$ genome but pseudogenized in *w*Au (supplementary fig. S6, Supplementary Material online). This protein has 3–4 transmembrane domains and we detected a very low similarity to the Endoplasmic-reticulum-associated protein degradation (ERAD)-associated E3 ubiquitin-protein ligase HRD1B from several plant species (*Brassica, Raphanus*, and *Arabidopsis*) with two rounds of PSI-blast (14% identity over 70% of the protein, with E-values above 1). However, the protein is present in most *Wolbachia* genomes, including for example the mutualistic non-CI strain *w*Bm from the nematode *Brugia malayi*.

Finally, the DNA recombination-mediator protein A (previously DNA processing chain A—DprA) is found in one copy in

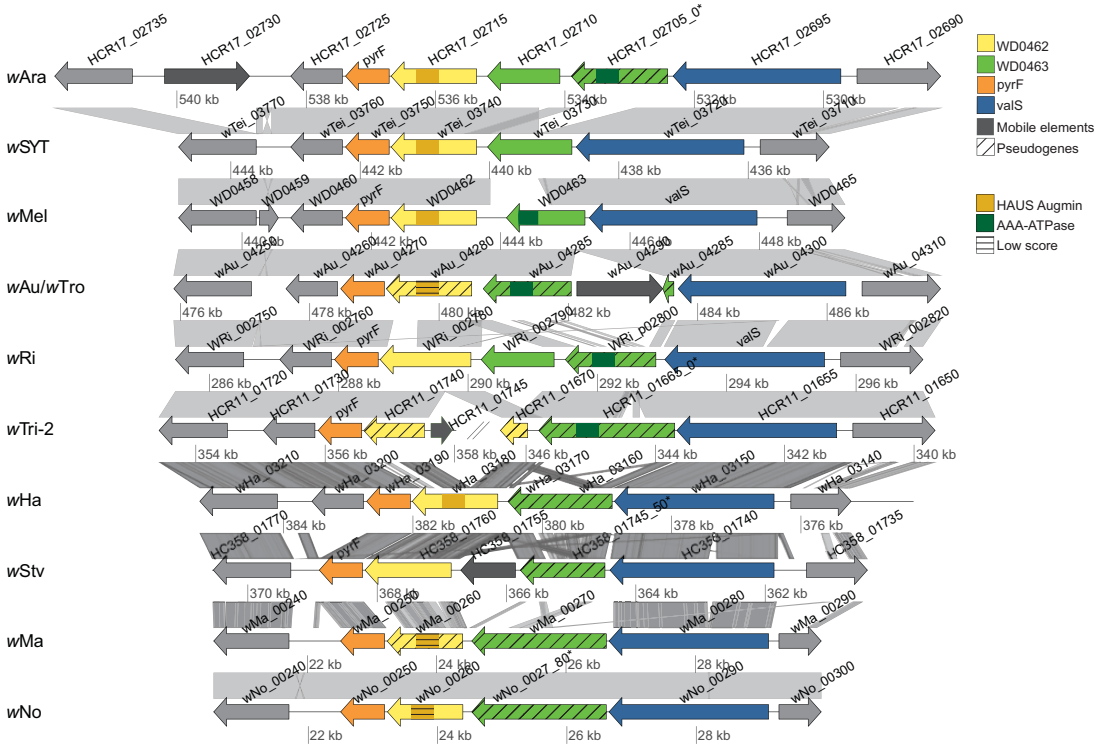

Fig. 5.—Comparison of the genomic region containing the CI-associated gene WD0462. Homologs of WD0462 are shown in yellow, with the predicted HAUS Augmin3 domain (PF14932) indicated in dark yellow. The neighboring gene WD0463 is shown in green, with the predicted AAA-ATPase domain (PF00004) indicated in dark green. The flanking genes *pyrF* and *valS* are shown in orange and blue, respectively. Other genes are represented in light gray and mobile elements in dark gray. Pseudogenes are marked by diagonal lines. Domain predictions with scores below the significance threshold are marked with horizontal lines. Similarity between sequences is indicated by gray lines, where darker is more similar.

all genomes (supplementary fig. S7, Supplementary Material online).

As observed for CifA, all three additional rescue candidates were identical between the *w*SYT genomes. Hence, we screened for divergent genes that correlate with *resc* properties, and identified two candidate genes under the assumptions that the *resc* factor has to be different between *w*SY and *w*Tei, different between *w*Au and all other, and could be identical between *w*No and *w*Ma. The first candidate encodes a putative M16 family peptidase where one mutation seems to have occurred in *w*SY, one in *w*Tei and one in *w*Au. The second candidate encodes Folylpolyglutamate synthase, where one mutation has occurred in *w*SY, one in *w*SYT and one in *w*Mel. Notably, an ortholog of the M16 family peptidase was found in ovaries of *C. pipiens* infected with a CI-inducing *Wolbachia* strain (LePage et al. 2017).

Discussion

Wolbachia participates in a remarkable variety of host phenotypes which range from mutualism to reproductive parasitism. Among these, CI stands out for its evolutionary implications as well as for the recent use in controlling insect-transmitted diseases. Despite the scientific interest, the genetic causes and mechanisms of CI are still relatively poorly understood, partly due to the multiple host and symbiont factors that influence the phenotype. Here, we perform in-depth comparative analyses of nine *Wolbachia* strains with known CI properties in the same host. By focusing on a single host background, we ensure that phenotypic variation between strains is associated with symbiont rather than host factors. We identify strain-specific genetic variation and evolutionary patterns across closely related *Wolbachia*, and effectively pinpoint *Wolbachia* genes potentially associated with *mod* and *resc* of CI.

Rapid Evolution of Wolbachia *Genomes and Phenotypes are Mediated by Mobile Elements*

Phages and other mobile elements often occupy a relatively large proportion of *Wolbachia* genomes (Wu et al. 2004; Klasson et al. 2009). They may also have a significant impact on *Wolbachia* ecology and evolution, since they frequently carry genes involved in host interaction and can be laterally transferred between strains (Bordenstein and Bordenstein 2016; Wang et al. 2016). The phage WO-associated *cif* genes, involved in CI, are prime examples of this phenomenon

(LePage et al. 2017; Madhav et al. 2020; Martinez et al. 2021).

Similar to previous studies (Ishmael et al. 2009; Ellegaard et al. 2013; Gerth and Bleidorn 2016), our comparisons show that phage WO regions contribute massively to the variation between closely related strains, as they contain a high proportion of the SNPs (SYTMA) as well as large gene content variability (all comparisons). Importantly, we observe that the Type IV *cif* genes of *w*Tei, which likely cause the strong CI of this strain, are located in a phage WO region that was potentially transferred into *w*SYT from a Supergroup B donor. This *cif* pair may have been the only fully functional *cif* locus in the ancestor of *w*SYT, as the inversion in Type I *cifB* occurs in all three genomes while the pseudogenization of Type IV *cifB* only occurs in *w*SY. Thus, the acquisition of this WO phage and consequently of the Type IV *cif* genes by an ancestor of *w*SYT may have had significant ecological importance for that *Wolbachia* lineage.

The same WO phage copy that carries the Type IV *cif* in *w*SYT is also associated with the Octomom region, implicated in titer regulation of the *w*MelPop strain (Chrostek and Teixeira 2015; Duarte et al. 2021). The location of the Octomom region next to phage WO in both *w*SYT and *w*Pip as well as its sporadic presence in *Wolbachia* genomes suggests that the region is often laterally transferred by phage WO. Furthermore, it supports the claim that the Octomom region in *w*Mel was also originally part of a WO phage (Klasson, Kambris et al. 2009). Our results show that homologs of WD0513 are present not only in *Wolbachia* but also in a variety of arthropod lineages and two other endosymbionts of arthropods, *Rickettsiella* and *Cardinium.* This suggests that lateral transfers potentially occur both between *Wolbachia* strains as well as between *Wolbachia*, other endosymbionts and their hosts. Although the mechanisms behind such transfers are unknown, the WO phage is a likely culprit in *Wolbachia* transfers (Bordenstein and Bordenstein 2016). Less is known about mobile elements in the other symbionts, but the genome of "*Candidatus* Rickettsiella viridis" contains one prophage region (Nikoh et al. 2018) and some *Cardinium* strains carry plasmids (Stouthamer et al. 2019) that potentially could facilitate lateral transfers.

The novel "*Dozen Island*" also shows evidence of lateral transfer from Supergroup B into *w*SYT. We observed a few similarities between the types of genes found in Dozen Island and those located on the pWCP plasmid of some *w*Pip strains (Reveillaud et al. 2019). Although no direct conclusion can be made, we speculate that Dozen Island could be derived from an integrated plasmid. Since both plasmid- and phage-associated genes are often implicated in the environment and host interaction in symbionts (Wernegreen and Moran 2001; Weldon et al. 2013; Harumoto and Lemaitre 2018), the Dozen Island genes could potentially carry such functions.

Our observations suggest that mobile elements are drivers of rapid evolution in *Wolbachia*, where they mediate the gain and loss of genes involved in ecologically important traits such as titer variation and CI.

Factors Associated with Induction and Rescue of CI

Wolbachia CI is a complex phenotype whose expression depends not only on the *cif* genes but also on a variety of factors (see Shropshire, Leigh et al. (2020) for a recent review). Here, we take advantage of the lack of host contribution to the phenotypic variation in our data set to generate new insight into *Wolbachia*-associated CI factors.

The cif *Genes*

The *cif* genes are the main *Wolbachia* factors implicated in CI, with *cifB* linked to *mod* and *cifA* either to *resc* or both *mod* and *resc* (Beckmann et al. 2019; Shropshire and Bordenstein 2019). These roles imply that a strain carrying a functional *cifB* also needs a functional *cifA* to be compatible with itself (Martinez et al. 2021). We observe such a pattern in our strains, in which both *cifA* and *cifB* are intact or *cifB* is pseudogenized either alone or in combination with *cifA*. However, *cifA* is never pseudogenized alone. Additionally, the association of *cifA* with *resc* is supported by the fact that all of our *resc*$^+$ strains have at least one putatively functional copy of *cifA*.

A similar association between *cifB* and *mod* implies that all *mod*$^+$ strains should carry a putatively functional copy of *cifB.* This is indeed the case for the strains *w*Ha, *w*Mel, *w*No, *w*Ri, and *w*Tei. However, the *w*SY strains are also *mod*$^+$ according to Zabalou et al. (2008) and Cooper et al. (2017) but do not carry any fully intact *cifB* genes. We must then either consider that they do not cause CI or that their truncated Type I CifB is at least partially functional. If the latter case is true, a weaker CifB function would support recent findings that mutations outside of the main described domains of the Cif proteins can affect their CI properties (Shropshire, Kalra et al. 2020). It might also suggest that the AAA-ATPase-like domain found by Martinez et al. (2021) is important for strong CI induction. Reduced CifB functionality due to truncation could, perhaps together with low infection titer (see discussion about titer below), be one of the reasons why *w*SY cause weaker CI in *D. simulans* in comparison to other strains that carry Type I CifB, such as *w*Mel and *w*Ha (Zabalou et al. 2008).

The analysis of the *cif* genes in our genomes supports previous observations that strains carrying phylogenetically related *cif* tend to be compatible with each other (Bonneau et al. 2018; Shropshire, Leigh et al. 2020). Similarity between *cif* genes can explain why the *w*SYT strains can rescue each other's modification, as the three strains have identical Type I and IV *cifA* genes, and why *w*Ma can rescue *w*No, as they have identical Type III *cifA* genes. However, several discrepancies remain regarding the observed patterns of *cif* genes in different strains and their published CI phenotypes in *D.*

simulans. First, Zabalou et al. (2008) showed that the three *w*SYT strains can rescue *w*Ri, but according to our analysis none of the *w*SYT genomes possess a Type II CifA homolog, and Type II is the only complete *cifB* gene in the *w*Ri genome. Secondly, the NoMa strains were seen to partially rescue *w*Tei (Zabalou et al. 2008) but their CifA homolog is of Type III rather than Type IV, which is the CifB type likely causing CI in *w*Tei. Additionally, we observed that all CI-inducing supergroup A genomes in our data set contain Type I *cifAB* genes, but only in *w*Mel and *w*Ha are both of genes intact, whereas *w*SYT and *w*Ri only encode an intact CifA. Thus, based on CifA and CifB being the *resc* and *mod* factors, *w*SYT, *w*Ri and *w*Ha should all be able to rescue *w*Mel. However, this is only partly in agreement with the results of Zabalou et al. (2008), as *w*Ri and *w*Tei rescue the CI induced by *w*Mel, but *w*SY do not. According to the same study, *w*SYT and *w*Ri cannot rescue the modification induced by *w*Ha even though *w*Ha only has an intact *cifB* of Type I and both *w*SYT and *w*Ri have intact *cifA* genes of Type I. In this case, it is worth noting that the Type I *cif* genes of *w*Ha are in a distinct subclade within the Type I phylogeny compared to those of *w*SYT and *w*Ri (fig. 4*A*). Hence, further experiments are necessary to investigate whether the two subclades of Type I represent distinct Types in the sense that *cif* genes from one cannot rescue modifications caused by genes from the other.

Are There More Wolbachia *Genes Involved in CI than* cif*?*

Since the *cif* genes cannot explain all variation in CI properties between our strains, we conclude that other genes must be involved in the phenotype. Our search for *Wolbachia* genes associated with *mod* and *resc* recovered a few novel CI-associated genes. Among these, homologs of WD0462 are particularly promising for having a role in *mod*, as they have high sequence variability between genomes (fig. 5), the *w*Mel protein was shown to negatively affects growth when expressed in yeast under stress conditions (Rice et al. 2017) and several of them have a Haus-Augmin3-like complex subunit 3, N-terminal domain (PF14932). This protein domain is present in the Dgt3 protein of *D. melanogaster*, where it binds to the gamma-Tubulin ring complex (gamma-TuRC) and is required for the accumulation of the gamma-TuRC to the mitotic spindle (Chen et al. 2017). The density of microtubules in the mitotic spindle is reduced without Augmin, which can lead to perturbed chromosome alignment and mitotic progression (Goshima et al. 2008; Uehara et al. 2009). Additionally, Augmin contributes to the generation of astral microtubules during mitosis, which are essential for checkpoint satisfaction and chromosome segregation (Hayward et al. 2014). Interestingly, the neighboring gene, WD0463, is highly variable between strains. Only strains encoding the WD0462 protein with a significant prediction for the Haus-Augmin3 domain also encode an intact WD0463 protein (fig. 5). Such pattern suggests possible coevolution between the two proteins. The AAA-ATPase domain (PF00004) found in several of the homologs of WD0463 is associated with a variety of cell functions including cell-cycle regulation and notably a similar domain is found in most CifB proteins. Despite these interesting characteristics, we note that WD0462 is not variable between *w*SYT and therefore cannot explain differences between them regarding CI strength or compatibility with other strains (Zabalou et al. 2008).

Other *mod* candidates that were identified due to their sequence divergence between our strains seem less likely to have a role in CI. However, potential effects on gene expression caused by mutations in RpoBC of *w*Tei and *w*No could perhaps affect the occurrence or strength of CI (see discussion about titer below). The same might be true for mutations in the upstream region of certain genes in *w*Tei in comparison to *w*SY.

Among the genes associated with *resc*, the multifunctional phage protein RepA is of interest, since it has the potential to regulate phage copy number which in turn might affect *Wolbachia* titer (Bordenstein et al. 2006). A putative *resc*-related role of RepA is also supported by its presence in the proteome data from ovaries of the mosquito *C. pipiens* infected with a CI-inducing *Wolbachia* (LePage et al. 2017). Recently, RepA was also identified as a CI candidate by Scholz et al. (2020), who observed that the protein was present in many *w*Mel and *w*Ri-like metagenomically assembled genomes (MAGs) but absent in several *w*Au-like MAGs.

One of our other *resc*-related proteins, the hypothetical protein WD1187, has low similarity to some E3 ubiquitin ligases from plants. This is interesting given that CifB Type I is a deubiquitinating enzyme able to cleave both Lysine-48 and Lysine-63 linked ubiquitin (Beckmann et al. 2017). Additionally, the concentration of E3 ligase in the cell is possibly a way to control the localization and fate of ubiquitinated proteins (Li et al. 2003), which might indicate that either the protein expression level or *Wolbachia* titer could be important if the *resc* phenotype occurs through such a mechanism.

The last *resc*-associated protein, DprA, is necessary for natural transformation in several bacterial species (Smeets et al. 2000; Takata et al. 2005; Duffin and Barber 2016) and acquisition of genes via the gene transfer agent in *Rhodobacter capsulatum* (Brimacombe et al. 2014). It has been seen to bind single-stranded DNA and interact with the RecA protein, thereby assisting in recombination (Mortier-Barriere et al. 2007). We note that although no ortholog of DprA was detected in the ovaries of *w*Pip-infected *C. pipiens*, RecA was (LePage et al. 2017). Even so, based on the known functions of this protein, we find it hard to speculate how it might be involved in the rescue of CI.

It is important to consider that the potential CI-associated effect of these genes may be indirect rather than a direct role in *mod* or *resc*. An example of this would be an effect on *Wolbachia* traits such as titer and localization which in turn influence CI.

Is Wolbachia *Titer Important for* Resc*?*

The variable ability of *w*SYT to rescue the modification of *w*Mel in *D. simulans* cannot be explained by either the *cif* genes or by our new CI gene candidates, since these are all identical in the three strains. Hence, we propose that the difference in rescue between the strains could be due to a quantitative rather than qualitative variation in the rescue factor. At least two lines of evidence support this suggestion. The rescue function of Type I CifA in *D. melanogaster* was shown to be dependent on expression level (Shropshire et al. 2018), and CI strength is correlated with bacterial titer in eggs (Martinez et al. 2015). As strong CI is clearly not caused by high *Wolbachia* titers in the egg, since modification occurs in sperm, this observation indicates that high bacterial titers are needed in eggs of *Wolbachia* strains causing strong CI. A likely interpretation of this is that high levels of the *resc* factor are needed to rescue a strong CI. Thus, one possibility is that the difference in rescue between *w*SYT is due to the higher *Wolbachia* titer of *w*Tei compared to *w*SY in the eggs of *D. simulans* (Martinez et al. 2015), where rescue occurs. The higher titer of *w*Tei would then result in enough CifA production to rescue the modification of *w*Mel, while the lower titer of *w*SY would not allow them to do the same.

Even so, the titer of *w*Tei is still much lower than that of *w*Ri or *w*Mel (Veneti et al. 2004; Martinez et al. 2015). Hence, an alternative hypothesis could be that higher levels of *cifA* in *w*Tei might be obtained independently of titer variation, for example through increased expression. In this context, it is interesting that we found a nonsynonymous mutation between *w*Tei and *w*SY in the *rpoBC* gene. Although we did not find any differences in the upstream regions of known CI genes in *w*SYT, other forms of gene regulation may exist. It is also interesting to note that the *w*SY genomes have two copies of the Type IV *cifA* genes, which might partly compensate for their low titer. Regardless of whether the quantitative effect is due to titer or expression, our reasoning leads to the testable hypothesis that the right amount of the *resc* factor as well as a good fit between *mod* and *resc* factors are both needed to rescue the modification of a strong CI inducer such as *w*Mel in *D. simulans*.

One possibility is that if the *mod* and *resc* factors fit perfectly together by having evolved under selection in the same genome, bacterial titer (or the amount of expressed *resc* factor) matters less than if *mod* and *resc* have a worse fit. With a less than perfect fit, perhaps rescue might only be possible if the *resc* factor is overexpressed compared to the *mod* factor with a perfect fit, a model of "force by numbers." If correct, this model predicts that *Wolbachia* strains with a higher amount of *resc* factor could more easily rescue the modification of other strains. This could give such strains an ecological advantage, as they would be potentially better at invading populations that are already infected with other CI-causing *Wolbachia* strains. In contrast, low titer strains, in which drift has created a worse fit between the *resc* and *mod* factors, would have difficulty to infect new host species that are more permissive to CI than their current host, since more *resc* factor might be needed to rescue the CI induced by the strain itself. This hypothesis might explain how "suicide" strains that don't fully rescue themselves, such as *w*Tei after transfer into *D. simulans* (Zabalou et al. 2008), can evolve under low CI conditions when there is low selection pressure on the *resc* function, like *w*Tei in its natural host.

Wolbachia *Factors Influencing Nonreproductive Phenotypes*

Due to the early establishment of *D. simulans* as a permissive host for a multitude of *Wolbachia* strains, several investigations of nonreproductive phenotypes have been performed.

Seven of our strains were used to investigate *Wolbachia*-associated protection against two RNA viruses (FHV and DCV) as well as female fecundity and lifespan in the *D. simulans* STC background (Martinez et al. 2014). Five of our strains were also used to investigate *Wolbachia* tropism in the germline stem cell niche (GSCN) during oogenesis and in the hub of testes during spermatogenesis (Toomey et al. 2013; Toomey and Frydman 2014).

Although the closely related *w*SYT strains have variable phenotypes in four of the five phenotypes mentioned above, none of the strains were uniquely represented in any of our protein clusters. Hence, differently from the CI phenotype, it is unlikely that these nonreproductive phenotypes occur through the action of proteins that are uniquely involved in those functions. This is perhaps not surprising, as these phenotypes are continuous rather than discrete and several of them correlate with *Wolbachia* titer in somatic tissues of *D. simulans* (Martinez et al. 2015). Thus, titer is likely also a crucial factor for the expression of *Wolbachia*-induced nonreproductive phenotypes.

Three genetic properties of *Wolbachia* have so far been seen to affect its titer. These are the number of copies of the Octomom region (Chrostek and Teixeira 2015; Duarte et al. 2021), the expression level of the *Wolbachia* actin-localizing effector 1 (Sheehan et al. 2016), and the presence of lytic WO phages (Bordenstein et al. 2006). Interestingly, one of the few things that clearly differ between the *w*SYT genomes is the number of phage WO regions, with *w*San having the largest amount of prophage DNA in its genome followed by *w*Yak and then *w*Tei. Currently, we don't know if the WO prophages in the *w*SYT genomes are expressed as lytic phages or whether they affect *Wolbachia* titer, but the correlation between titer and amount of prophage WO in the genome is intriguing. However, the titer of different *Wolbachia* strains may be controlled by several different mechanisms, which would make it more difficult to pinpoint the exact genetic component involved, especially when more divergent strains are compared.

Finally, the embryonic distribution of *Wolbachia* was tested in eight of our nine strains (Veneti et al. 2004), albeit in their natural hosts. While *w*Ri has a global distribution in the embryo, the SYTMA strains have a posterior distribution and the NoMa strains an anterior distribution. Although we cannot take advantage of our closely related genomes, it might still be interesting to investigate the 19 protein clusters specifically found in the SYTMA clade in order to elucidate if any of them could be involved in *Wolbachia* posterior localization. Notably, two of the proteins were shown to affect growth when expressed in yeast, and might thus be *Wolbachia* effectors (Rice et al. 2017).

Conclusions

In this study, we use complete genomes of closely related *Wolbachia* combined with their phenotypic data in the *D. simulans* STC host background to investigate *Wolbachia* evolution and genetic determinants of CI. Our analysis shows that transferring *Wolbachia* strains from other *Drosophila* into *D. simulans* does not seem to create significant evolutionary pressures on any particular symbiont function, which corroborates this strategy for studying *Wolbachia*.

From an evolutionary perspective, we find support for phages and mobile elements playing an important role in *Wolbachia* ecology and evolution through lateral transfers of genes implicated in phenotype expression and host interaction. We find evidence that phylogenetically related *cif* genes tend to be compatible and that other genes apart from *cif* are associated with modification and rescue of CI in our genomes. Both the TA and HM models of CI can be reconciliated with our observation, with the novel CI-associated genes potentially either affecting the affinity between CifA and CifB (TA) or influencing the host interactions that lead to modification and rescue (HM). Based on our results, we also speculate that *Wolbachia* titer could be the missing factor that explains variability in CI rescue capabilities of the *w*SYT strains as well as in other systems. A higher symbiont titer in eggs should favor rescue regardless of CI model, as it increases the probability that *resc*-associated proteins find their targets. High titer *Wolbachia* might thus be better at invading new host populations that already carry a CI-inducing *Wolbachia* strain. If true, infection titer is a highly relevant parameter to consider when designing strategies for using CI *Wolbachia* in biological control programs.

Materials and Methods

DNA Preparation and Sequencing

All DNA samples used for Illumina and PacBio sequencing were produced using the protocol described in Ellegaard et al. (2013). Briefly, *Drosophila* flies were transferred to apple juice agar plates and allowed to oviposit for 2 h, after which the eggs were collected, washed, and dechorionated in 50% bleach and manually homogenized using a plastic pestle. Following centrifugation and filtration of the homogenate to enrich for *Wolbachia* cells, the resulting cell pellet was subjected to whole genome amplification using the Repli-g® midi kit (Qiagen), after which the DNA was purified using QIAamp® DNA mini kit (Qiagen) according to the manufacturer's recommendations. Standard 350 bp fragment TruSeq libraries were constructed from DNA samples of 11 different *Wolbachia* strains (supplementary table S1, Supplementary Material online). All libraries were indexed and run together in one lane on an Illumina HiSeq 2500 machine, generating 2×100 bp sequence reads. Illumina libraries were produced and sequencing was performed at the SNP and SEQ platform, Uppsala. DNA from each of the five *Wolbachia* strains yielding complete genomes (supplementary table S1, Supplementary Material online) was used to create 5 kb fragment SMRTbell libraries. Each library was run using P6-C4 chemistry in one SMRT cell on the RSII PacBio instrument. PacBio libraries were produced and sequencing was performed at the Uppsala Genome Center, Uppsala.

Genome Assembly

Illumina reads were quality and adapter trimmed by Trimmomatic-0.22 (Bolger et al. 2014) and error corrected based on k-mer frequencies using BayesHammer in SPAdes (Bankevich et al. 2012). Corrected Illumina reads were assembled into contigs by AbySS (Simpson et al. 2009), SPAdes (Bankevich et al. 2012), IDBA (Peng et al. 2012), and Velvet (Zerbino and Birney 2008) with k-mer sizes from 63 to 95 with an interval of four. Assembly statistics were calculated using the Perl script assemblathon_stats.pl (https://github.com/ucdavis-bioinformatics/assemblathon2-analysis, last accessed May 24, 2021). The N50 value, predicted genome size, total number of contigs, and length of contigs were considered while selecting the best assembly from each assembler. To complete the draft genome, PacBio reads were obtained and assembled both independently using HGAP (Chin et al. 2013) and together with Illumina reads using Spades. The best assembly was chosen based on the same criteria used for Illumina assemblies, and thereafter overlapping contigs were merged in Consed (Gordon et al. 1998). Illumina and PacBio reads were mapped against the assemblies using BWA-mem (Li and Durbin 2009) with default parameters for Illumina reads and using PacBio settings for PacBio reads to check the correctness of the genome assembly. Gaps and inconsistencies between the assemblies and the data were tested using PCR and resolved by direct Sanger sequencing of the PCR products. In cases where repeats were too large to span with PCR products, PacBio data were extensively inspected for consistency with the assembly and PCR products that go from unique to repeat sequence were also generated

in most cases. All assemblies and read data were combined and curated using Consed.

Annotation

The genomes were annotated using an automated annotation pipeline DIYA (Stewart et al. 2009), as described in Ellegaard et al. (2013). Prodigal (Hyatt et al. 2010) was used to predict the protein-coding genes, while GenePRIMP (Pati et al. 2010) and blastx were used to identify pseudogenes. tRNAscan-SE (Lowe and Eddy 1997) and RNAmmer (Lagesen et al. 2007) were used to predict tRNA and rRNA, respectively. All predicted proteins were searched against the UniProt database and previously annotated *Wolbachia* proteomes using blastp. PFAM domains were identified using pfam_scan.pl (Li et al. 2015). Mummer was used to predict repeats (Kurtz et al. 2004). After automated annotation, all data were collected in Artemis (Rutherford et al. 2000) and used to manually curate each genome. Repeats were annotated using nucmer from the Mummer 3 package (Kurtz et al. 2004) with a minimum of 300 bp and 95% identity. Phage regions were annotated manually by comparing the gene content to previously published *Wolbachia* genomes.

Variant Calling

Quality and adapter trimmed Illumina reads were aligned against each of the other finished genomes as well as against their respective genomes with BWA-mem and subsequently sorted and marked for duplicates with the Picard toolkit (http://broadinstitute.github.io/picard/, last accessed May 24, 2021). For each set of aligned reads, reads were realigned with the IndelRealigner from GATK (McKenna et al. 2010) and SNPs and Indels were called using the Haplotypecaller from GATK with a ploidy of 1. SNPs and Indels were filtered separately using the GATK best practice settings but removing the criteria for haplotype score and increasing the QD threshold to 15. To avoid spurious variant calls, only sites with a minimum read depth of 10 were used. snpEff (Cingolani et al. 2012) was used to create a specific database for each of the five complete and annotated genomes and to identify the effect and location of each variant within them.

Clustering and Phylogenetic Analyses

The proteomes of the nine *Wolbachia* strains were clustered using OrthoMCL (Abascal and Valencia 2003) with an inflation value of 1.5. For genes found in clusters containing a single copy in each genome, nucleotide sequences were extracted, translated to proteins, aligned using mafft-linsi (Katoh and Standley 2013), and backtranslated to nucleotides. Phylogenetic trees were constructed for each of these genes individually and for a concatenated alignment of all of them using RAxML Version 8.1.16 (Stamatakis 2006) with the GTRGAMMA model and 100 rapid bootstraps. The same procedure was used for all gene alignments analyzed. Synonymous and nonsynonymous substitution rates were calculated using codeml from the PAML package (Yang 2007).

For the phylogenetic analysis of WD0513, all nonidentical *Wolbachia* homologs in our clusters were searched against the nr database using blastp. Proteins that covered at least 80% of the length of the query and had an e-value smaller than e^{-05} were aligned to the homologs in the nine *Wolbachia* genomes using mafft-linsi and trimmed using trimAl (Capella-Gutierrez et al. 2009) with the –automated1 setting. Phylogenies were inferred using RAxML Version 8.1.16 with the PROTGAMMAAUTO model and 100 rapid bootstraps. In a majority of cases, LG was the best scoring amino acid model that was used. All trees were visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree/, last accessed May 24, 2021). For the phylogenetic analysis of the *cif* genes, CifA and CifB proteins from our genomes were combined with representative Cif proteins of Type I–V chosen among those featured in Martinez et al. (2021). The resulting protein set was analyzed as described for WD0513. Domain predictions for WD462 and WD463 were made with the online implementation of pfamscan (https://www.ebi.ac.uk/Tools/pfa/pfamscan/, last accessed May 24, 2021) using default values.

All gene comparison figures were made using genoPlotR (Guy et al. 2010), after blasting each genome against every other genome with blastn for close relatives and blastp for distant relatives.

Supplementary Material

Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments

The authors thank Roel van Eijk for technical assistance running PCRs, Lina Juzokaite for DNA extractions for PacBio sequencing, and Kostas Bourtzis for providing fly lines as well as contributing with advice and support throughout. Illumina sequencing was performed at the SNP&SEQ Technology Platform and PacBio sequencing was performed at Uppsala Genomic Center in Uppsala, Sweden, both of which are part of the Swedish National Genomics Infrastructure. The data handling was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX. This work was supported by grants from Formas (2009-378) and The Swedish research council VR (2014-4353) to L.K.

Author Contributions

L.K. conceived and designed the study. J.J., L.K., and M.G. performed assemblies. J.J., G.C.B., L.K., and M.G. performed annotations. L.K., G.C.B., and J.J. performed comparative and

phylogenetic analyses. M.G. performed DNA extractions. L.K. and G.C.B. wrote the paper with contribution from J.J. All authors read and approved the manuscript.

Data Availability

The data underlying this article are available at NCBI and can be accessed from BioProject PRJNA694504. All individual accession numbers can be found in supplementary table S1, Supplementary Material online.

References

ΘAbascal F, Valencia A. 2003. Automatic annotation of protein function based on family identification. Proteins 53(3):683–692.

Beckmann JF, et al. 2019. The toxin-antidote model of cytoplasmic incompatibility: genetics and evolutionary implications. Trends Genet. 35(3):175–185.

Beckmann JF, Ronau JA, Hochstrasser M. 2017. A *Wolbachia* deubiquitylating enzyme induces cytoplasmic incompatibility. Nat Microbiol. 2:17007.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bonneau M, et al. 2018. *Culex pipiens* crossing type diversity is governed by an amplified and polymorphic operon of *Wolbachia*. Nat Commun. 9(1):319.

Bordenstein SR, Bordenstein SR. 2016. Eukaryotic association module in phage WO genomes from *Wolbachia*. Nat Commun. 7:13155.

Bordenstein SR, Marshall ML, Fry AJ, Kim U, Wernegreen JJ. 2006. The tripartite associations between bacteriophage, Wolbachia, and arthropods. PLoS Pathog. 2(5):e43.

Bordenstein SR, O'Hara FP, Werren JH. 2001. Wolbachia-induced incompatibility precedes other hybrid incompatibilities in Nasonia. Nature 409(6821):707–710.

Bossan B, Koehncke A, Hammerstein P. 2011. A new model and method for understanding *Wolbachia*-induced cytoplasmic incompatibility. PLoS One 6(5):e19757.

Boyle L, O'Neill SL, Robertson HM, Karr TL. 1993. Interspecific and intraspecific horizontal transfer of Wolbachia in Drosophila. Science 260(5115):1796–1799.

Brimacombe CA, Ding H, Beatty JT. 2014. *Rhodobacter capsulatus* DprA is essential for RecA-mediated gene transfer agent (RcGTA) recipient capability regulated by quorum-sensing and the CtrA response regulator. Mol Microbiol. 92(6):1260–1278.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Charlat S, et al. 2004. Incipient evolution of Wolbachia compatibility types. Evolution 58(9):1901–1908.

Chen H, Ronau JA, Beckmann JF, Hochstrasser M. 2019. A *Wolbachia* nuclease and its binding partner provide a distinct mechanism for cytoplasmic incompatibility. Proc Natl Acad Sci U S A. 116(44):22314–22321.

Chen JWC, et al. 2017. Cross-linking mass spectrometry identifies new interfaces of Augmin required to localise the γ-tubulin ring complex to the mitotic spindle. Biol Open. 6(5):654–663.

Chin CS, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 10(6):563–569.

Chrostek E, et al. 2013. *Wolbachia* variants induce differential protection to viruses in *Drosophila melanogaster:* a phenotypic and phylogenomic analysis. PLoS Genet. 9(12):e1003896.

Chrostek E, Teixeira L. 2015. Mutualism breakdown by amplification of Wolbachia genes. PLoS Biol. 13(2):e1002065.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 6(2):80–92.

Cooper BS, Ginsberg PS, Turelli M, Matute DR. 2017. Wolbachia in the *Drosophila yakuba* complex: pervasive frequency variation and weak cytoplasmic incompatibility, but no apparent effect on reproductive isolation. Genetics 205(1):333–351.

Cooper BS, Vanderpool D, Conner WR, Matute DR, Turelli M. 2019. *Wolbachia* acquisition by *Drosophila yakuba*-clade hosts and transfer of incompatibility loci between distantly related *Wolbachia*. Genetics 212(4):1399–1419.

Duarte EH, Carvalho A, López-Madrigal S, Costa J, Teixeira L. 2021. Forward genetics in Wolbachia: Regulation of Wolbachia proliferation by the amplification and deletion of an addictive genomic island. bioRxiv. 10.1101/2020.09.08.288217

Duffin PM, Barber DA. 2016. DprA is required for natural transformation and affects pilin variation in *Neisseria gonorrhoeae*. Microbiology (Reading). 162(9):1620–1628.

Ellegaard KM, Klasson L, Näslund K, Bourtzis K, Andersson SGE. 2013. Comparative genomics of *Wolbachia* and the bacterial species concept. PLoS Genet. 9(4):e1003381.

Fast EM, et al. 2011. Wolbachia enhance Drosophila stem cell proliferation and target the germline stem cell niche. Science 334(6058):990–992.

Flores HA, O'Neill SL. 2018. Controlling vector-borne diseases by releasing modified mosquitoes. Nat Rev Microbiol. 16(8):508–518.

Gerth M, Bleidorn C. 2016. Comparative genomics provides a timeframe for *Wolbachia* evolution and exposes a recent biotin synthesis operon transfer. Nat Microbiol. 2:16241.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8(3):195–202.

Goshima G, Mayer M, Zhang N, Stuurman N, Vale RD. 2008. Augmin: a protein complex required for centrosome-independent microtubule generation within the spindle. J Cell Biol. 181(3):421–429.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26(18):2334–2335.

Harumoto T, Lemaitre B. 2018. Male-killing toxin in a bacterial symbiont of *Drosophila*. Nature. 557(7704):252–255.

Hayward D, Metz J, Pellacani C, Wakefield JG. 2014. Synergy between multiple microtubule-generating pathways confers robustness to centrosome-driven mitotic spindle formation. Dev Cell. 28(1):81–93.

Hedges LM, Brownlie JC, O'Neill SL, Johnson KN. 2008. Wolbachia and virus protection in insects. Science 322(5902):702.

Hurst LD. 1991. The evolution of cytoplasmic incompatibility or when spite can be successful. J Theor Biol. 148(2):269–277.

Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11(1):119.

Ishmael N, et al. 2009. Extensive genomic diversity of closely related Wolbachia strains. Microbiology (Reading). 155(7):2211–2222.

Iturbe-Ormaetxe I, Burke GR, Riegler M, O'Neill SL. 2005. Distribution, expression, and motif variability of ankyrin domain genes in *Wolbachia pipientis*. J Bacteriol. 187(15):5136–5145.

Jaenike J. 2007. Spontaneous emergence of a new Wolbachia phenotype. Evolution 61(9):2244–2252.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Klasson L, Kambris Z, Cook PE, Walker T, Sinkins SP. 2009. Horizontal gene transfer between Wolbachia and the mosquito *Aedes aegypti*. BMC Genomics. 10:33.

Klasson L, et al. 2008. Genome evolution of Wolbachia strain wPip from the *Culex pipiens* group. Mol Biol Evol. 25(9):1877–1887.

Klasson L, et al. 2009. The mosaic genome structure of the *Wolbachia* wRi strain infecting *Drosophila simulans*. Proc Natl Acad Sci U S A. 106(14):5725–5730.

Kosinski J, Feder M, Bujnicki JM. 2005. The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. BMC Bioinformatics 6(1):172.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5(2):R12.

Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35(9):3100–3108.

Lefoulon E, et al. 2020. Pseudoscorpion *Wolbachia* symbionts: diversity and evidence for a new supergroup S. BMC Microbiol. 20(1):188.

LePage DP, et al. 2017. Prophage WO genes recapitulate and enhance *Wolbachia*-induced cytoplasmic incompatibility. Nature 543(7644):243–247.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Li M, et al. 2003. Mono- versus polyubiquitination: differential control of p53 fate by Mdm2. Science 302(5652):1972–1975.

Li W, et al. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 43(W1):W580–584.

Lindsey ARI, et al. 2018. Evolutionary genetics of cytoplasmic incompatibility genes *cifA* and *cifB* in prophage WO of *Wolbachia*. Genome Biol Evol. 10(2):434–451.

Lo N, et al. 2007. Taxonomic status of the intracellular bacterium *Wolbachia pipientis*. Int J Syst Evol Microbiol. 57(Pt 3):654–657.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.

Madhav M, Parry R, Morgan JAT, James P, Asgari S. 2020. *Wolbachia* endosymbiont of the horn fly (*Haematobia irritans irritans*): a supergroup A strain with multiple horizontally acquired cytoplasmic incompatibility genes. Appl Environ Microbiol. 86(6):e02589–02519.

Mardanov AV, Ravin NV. 2006. Functional characterization of the *repA* replication gene of linear plasmid prophage N15. Res Microbiol. 157(2):176–183.

Martinez J, Klasson L, Welch JJ, Jiggins FM. 2021. Life and death of selfish genes: comparative genomics reveals the dynamic evolution of cytoplasmic incompatibility. Mol Biol Evol. 38(1):2–15.

Martinez J, et al. 2014. Symbionts commonly provide broad spectrum resistance to viruses in insects: a comparative analysis of *Wolbachia* strains. PLoS Pathog. 10(9):e1004369.

Martinez J, et al. 2015. Should symbionts be nice or selfish? Antiviral effects of *Wolbachia* are costly but reproductive parasitism is not. PLoS Pathog. 11(7):e1005021.

McGraw EA, Merritt DJ, Droller JN, O'Neill SL. 2001. Wolbachia-mediated sperm modification is dependent on the host genotype in Drosophila. Proc Biol Sci. 268(1485):2565–2570.

McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.

Merçot H, Charlat S. 2004. *Wolbachia* infections in *Drosophila melanogaster* and *D. simulans*: polymorphism and levels of cytoplasmic incompatibility. Genetica 120(1-3):51–59.

Mortier-Barriere I, et al. 2007. A key presynaptic role in transformation for a widespread bacterial protein: dprA conveys incoming ssDNA to RecA. Cell 130(5):824–836.

Newton IL, et al. 2016. Comparative genomics of two closely related *Wolbachia* with different reproductive effects on hosts. Genome Biol Evol. 8(5):1526–1542.

Nikoh N, et al. 2018. Genomic insight into symbiosis-induced insect color change by a facultative bacterial endosymbiont, "*Candidatus* Rickettsiella viridis". mBio 9(3):e00890–00818.

Pati A, et al. 2010. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat Methods. 7(6):455–457.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11):1420–1428.

Poinsot D, Bourtzis K, Markakis G, Savakis C, Merçot H. 1998. Wolbachia transfer from *Drosophila melanogaster* into *D. simulans*: host effect and cytoplasmic incompatibility relationships. Genetics 150(1):227–237.

Poinsot D, Charlat S, Mercot H. 2003. On the mechanism of Wolbachia-induced cytoplasmic incompatibility: confronting the models with the facts. Bioessays 25(3):259–265.

Reveillaud J, et al. 2019. The *Wolbachia* mobilome in *Culex pipiens* includes a putative plasmid. Nat Commun. 10(1):1051.

Rice DW, Sheehan KB, Newton ILG. 2017. Large-scale identification of *Wolbachia pipientis* effectors. Genome Biol Evol. 9(7):1925–1937.

Rutherford K, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics. 16(10):944–945.

Sasaki T, Massaki N, Kubo T. 2005. Wolbachia variant that induces two distinct reproductive phenotypes in different hosts. Heredity (Edinb). 95(5):389–393.

Scholz M, et al. 2020. Large scale genome reconstructions illuminate *Wolbachia* evolution. Nat Commun. 11(1):5235.

Schön I, Kamiya T, Van den Berghe T, Van den Broecke L, Martens K. 2019. Novel *Cardinium* strains in non-marine ostracod (Crustacea) hosts from natural populations. Mol Phylogenet Evol. 130:406–415.

Sheehan KB, Martin M, Lesser CF, Isberg RR, Newton ILG. 2016. Identification and characterization of a candidate *Wolbachia pipientis* type IV effector that interacts with the actin cytoskeleton. mBio 7(4):e00622–00616.

Shropshire JD, Bordenstein SR. 2019. Two-By-One model of cytoplasmic incompatibility: synthetic recapitulation by transgenic expression of *cifA* and *cifB* in *Drosophila*. PLoS Genet. 15(6):e1008221.

Shropshire JD, Kalra M, Bordenstein SR. 2020. Evolution-guided mutagenesis of the cytoplasmic incompatibility proteins: identifying CifA's complex functional repertoire and new essential regions in CifB. PLoS Pathog. 16(8):e1008794.

Shropshire JD, Leigh B, Bordenstein SR. 2020. Symbiont-mediated cytoplasmic incompatibility: what have we learned in 50 years? eLife 9:e61989.

Shropshire JD, et al. 2019. Models and nomenclature for cytoplasmic incompatibility: caution over premature conclusions – a response to Beckmann et al. Trends Genet. 35(6):397–399.

Shropshire JD, On J, Layton EM, Zhou H, Bordenstein SR. 2018. One prophage WO gene rescues cytoplasmic incompatibility in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 115(19):4987–4991.

Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123.

Sinha A, Li Z, Sun L, Carlow CKS. 2019. Complete genome sequence of the *Wolbachia* wAlbB endosymbiont of *Aedes albopictus*. Genome Biol Evol. 11(3):706–720.

Smeets LC, Bijlsma JJ, Kuipers EJ, Vandenbroucke-Grauls CM, Kusters JG. 2000. The *dprA* gene is required for natural transformation of *Helicobacter pylori*. FEMS Immunol Med Microbiol. 27(2):99–102.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21):2688–2690.

Stewart AC, Osborne B, Read TD. 2009. DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics 25(7):962–963.

Stouthamer CM, Kelly SE, Mann E, Schmitz-Esser S, Hunter MS. 2019. Development of a multi-locus sequence typing system helps reveal the evolution of *Cardinium hertigii*, a reproductive manipulator symbiont of insects. BMC Microbiol. 19(1):266.

Sutton ER, Harris SR, Parkhill J, Sinkins SP. 2014. Comparative genome analysis of *Wolbachia* strain wAu. BMC Genomics. 15(1):928.
Takata T, Ando T, Israel DA, Wassenaar TM, Blaser MJ. 2005. Role of *dprA* in transformation of *Campylobacter jejuni*. FEMS Microbiol Lett. 252(1):161–168.
Teixeira L, Ferreira A, Ashburner M. 2008. The bacterial symbiont Wolbachia induces resistance to RNA viral infections in *Drosophila melanogaster*. PLoS Biol. 6(12):e2.
Toomey ME, Frydman HM. 2014. Extreme divergence of *Wolbachia* tropism for the stem-cell-niche in the *Drosophila* testis. PLoS Pathog. 10(12):e1004577.
Toomey ME, Panaram K, Fast EM, Beatty C, Frydman HM. 2013. Evolutionarily conserved *Wolbachia*-encoded factors control pattern of stem-cell niche tropism in *Drosophila* ovaries and favor infection. Proc Natl Acad Sci U S A. 110(26):10788–10793.
Tram U, Sullivan W. 2002. Role of delayed nuclear envelope breakdown and mitosis in Wolbachia-induced cytoplasmic incompatibility. Science 296(5570):1124–1126.
Tsuchida T, et al. 2010. Symbiotic bacterium modifies aphid body color. Science 330(6007):1102–1104.
Uehara R, et al. 2009. The augmin complex plays a critical role in spindle microtubule generation for mitotic progression and cytokinesis in human cells. Proc Natl Acad Sci U S A. 106(17):6998–7003.
Veneti Z, Clark ME, Karr TL, Savakis C, Bourtzis K. 2004. Heads or tails: host-parasite interactions in the Drosophila-Wolbachia system. Appl Environ Microbiol. 70(9):5366–5372.
Wang GH, et al. 2016. Bacteriophage WO can mediate horizontal gene transfer in endosymbiotic *Wolbachia* genomes. Front Microbiol. 7:1867.
Weldon SR, Strand MR, Oliver KM. 2013. Phage loss and the breakdown of a defensive symbiosis in aphids. Proc Biol Sci. 280(1751):20122103.
Wernegreen JJ, Moran NA. 2001. Vertical transmission of biosynthetic plasmids in aphid endosymbionts (Buchnera). J Bacteriol. 183(2):785–790.
Werren JH. 1997. Biology of Wolbachia. Annu Rev Entomol. 42: 587–609.
Werren JH, Baldo L, Clark ME. 2008. *Wolbachia*: master manipulators of invertebrate biology. Nat Rev Microbiol. 6(10):741–751.
Woolfit M, Iturbe-Ormaetxe I, McGraw EA, O'Neill SL. 2009. An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*. Mol Biol Evol. 26(2):367–374.
Wu M, et al. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. PLoS Biol. 2(3):E69.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Zabalou S, et al. 2008. Multiple rescue factors within a Wolbachia strain. Genetics 178(4):2145–2160.
Zabalou S, et al. 2004. Natural Wolbachia infections in the *Drosophila yakuba* species complex do not induce cytoplasmic incompatibility but fully rescue the wRi modification. Genetics 167(2):827–834.
Zchori-Fein E, Perlman SJ. 2004. Distribution of the bacterial symbiont Cardinium in arthropods. Mol Ecol. 13(7):2009–2016.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.
Zheng X, et al. 2019. Incompatible and sterile insect techniques combined eliminate mosquitoes. Nature 572(7767):56–61.
Zug R, Hammerstein P. 2012. Still a host of hosts for *Wolbachia*: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One 7(6):e38544.

Associate editor: Daniel Sloan