

1 **Evolutionary genetics of cytoplasmic incompatibility genes**

2 ***cifA* and *cifB* in prophage WO of *Wolbachia***

3
4 Amelia R. I. Lindsey^{1,7}, Danny W. Rice², Sarah R. Bordenstein³, Andrew W. Brooks^{3,4}, Seth R.
5 Bordenstein^{*3,4,5,6}, Irene L. G. Newton^{*2}

6
7 ¹Department of Entomology, University of California Riverside, Riverside, California, USA

8 ²Department of Biology, Indiana University, Bloomington, Indiana, USA

9 ³Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, USA

10 ⁴Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville,
11 Tennessee, USA

12 ⁵Vanderbilt Institute of Infection, Immunology, and Inflammation, Vanderbilt University,
13 Nashville, Tennessee, USA

14 ⁶Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, USA

15 ⁷Current Address: Department of Biology, Indiana University, Bloomington, Indiana, USA

16
17 *Authors for Correspondence:

18 Irene L.G. Newton, Department of Biology, Indiana University, Bloomington, Indiana, USA,

19 812-855-3883, irnewton@indiana.edu, and Seth R. Bordenstein, Department of Biological

20 Sciences, Vanderbilt University, Nashville, Tennessee, USA, 615-322-9087,

21 s.bordenstein@vanderbilt.edu

24 **Abstract**

25 The bacterial endosymbiont *Wolbachia* manipulates arthropod reproduction to facilitate its
26 maternal spread through populations. The most common manipulation is cytoplasmic
27 incompatibility (CI): *Wolbachia*-infected males produce modified sperm that cause embryonic
28 mortality, unless rescued by eggs harboring the same *Wolbachia*. The genes underlying CI, *cifA*
29 and *cifB*, were recently identified in the eukaryotic association module of *Wolbachia*'s prophage
30 WO. Here, we use transcriptomic and genomic approaches to address three important
31 evolutionary facets of these genes. First, we assess whether or not *cifA* and *cifB* comprise a
32 classic toxin-antitoxin operon, and show they do not form an operon in strain *wMel*. They
33 coevolve but exhibit strikingly distinct expression across host development. Second, we provide
34 new domain and functional predictions across homologs within *Wolbachia*, and we show amino
35 acid sequences vary substantially across the genus. Lastly, we investigate conservation of *cifA*
36 and *cifB* and find degradation and loss of the genes is common in strains that no longer induce
37 CI. Taken together, we find no evidence for the operon hypothesis in *wMel*, provide functional
38 annotations that broaden the potential mechanisms of CI induction, illuminate recurrent erosion
39 of *cifA* and *cifB* in non-CI strains, and advance an understanding of the most widespread form of
40 reproductive parasitism.

41

42 **Key words**

43 symbiosis, reproductive manipulation, gene loss, bacteriophage

44

45 **Introduction**

46 The genus *Wolbachia* is the most widespread group of maternally transmitted endosymbiotic
47 bacteria (Zug and Hammerstein 2012). They occur worldwide in numerous arthropods and
48 nematodes and can selfishly manipulate reproduction (Werren, et al. 2008), confer antiviral
49 defense (Bian, et al. 2010; Teixeira, et al. 2008), and assist reproduction and development of
50 their hosts (Dedeine, et al. 2001; Hoerauf, et al. 1999; Hosokawa, et al. 2010). The most
51 common parasitic manipulation is cytoplasmic incompatibility (CI), whereby *Wolbachia*-
52 infected males produce modified sperm that can only be rescued by eggs infected with the same
53 *Wolbachia* strain (Yen and Barr 1971). If the modified sperm fertilize eggs infected with no
54 *Wolbachia* (unidirectional CI) or a genetically-incompatible *Wolbachia* strain (bidirectional CI),
55 then delayed histone deposition, improper chromosome condensation and cell division
56 abnormalities result in embryonic mortality (Landmann, et al. 2009; Lassy and Karr 1996;
57 Serbus, et al. 2008; Tram and Sullivan 2002). Other described reproductive manipulations
58 include parthenogenesis (Stouthamer, et al. 1990), male-killing (Hurst, et al. 1999), and
59 feminization (Rousset, et al. 1992), all of which give a fitness advantage to *Wolbachia*-infected
60 females and thus assist the spread of the infected matriline through a population. These
61 manipulations, once sustained, can also impact host evolution including speciation (Bordenstein,
62 et al. 2001; Brucker and Bordenstein 2013; Jaenike, et al. 2006) and mating behaviors (Miller, et
63 al. 2010; Moreau, et al. 2001; Randerson, et al. 2000; Shropshire and Bordenstein 2016).

64

65 In addition to the aforementioned reproductive manipulations, *Wolbachia* strains affect host
66 biology by provisioning nutrients (Hosokawa, et al. 2010), altering host survivorship (Min and
67 Benzer 1997) and fecundity (Dedeine, et al. 2001; Stouthamer and Luck 1993), and importantly,

68 protecting the host against pathogens (Bian, et al. 2010; Hughes, et al. 2011; Kambris, et al.
69 2009; Moreira, et al. 2009; Teixeira, et al. 2008; Walker, et al. 2011). The combination of
70 reproductive manipulations that enable *Wolbachia* to spread in a population, and the ability to
71 reduce vector competence through pathogen protection, have placed *Wolbachia* in the forefront
72 of efforts to control target arthropod populations (Bourtzis, et al. 2014; Hoffmann, et al. 2011;
73 Turelli and Hoffmann 1991; Walker, et al. 2011; Zabalou, et al. 2004). Despite these important
74 applications, the widespread prevalence of *Wolbachia* across arthropod taxa (Hilgenboecker, et
75 al. 2008; Werren and Windsor 2000; Zug and Hammerstein 2012), and decades of research, only
76 recently have the genes underlying CI been determined (Beckmann, et al. 2017; LePage, et al.
77 2017).

78
79 Two studies converged on the same central finding: coexpression of a pair of syntenic genes
80 recapitulates the CI phenotype (Beckmann, et al. 2017; LePage, et al. 2017). Uninfected
81 *Drosophila melanogaster* males transgenically expressing the two genes from *wMel Wolbachia*
82 caused CI-like embryonic lethality when crossed with uninfected females that was notably
83 rescued by *wMel*-infected females (LePage, et al. 2017). Additionally, the two *wMel* genes
84 separately enhanced *Wolbachia*-induced CI in a dose dependent manner when expressed in
85 *Wolbachia*-infected males (LePage, et al. 2017). In a separate study, CI-like embryonic lethality
86 was also recapitulated through transgenic coexpression in *D. melanogaster* males of homologous
87 transgenes encoded by the *Wolbachia wPip* strain (which infects *Culex* mosquitoes) (Beckmann,
88 et al. 2017). These two genes occur in the recently discovered eukaryotic association module of
89 temperate phage WO (Bordenstein and Bordenstein 2016), which was previously implicated in
90 influencing CI (Bordenstein, et al. 2006; Duron, et al. 2006; Masui, et al. 2000; Sinkins, et al.

91 2005). The presence of these genes within prophage WO has implications for the transmission of
92 these genes, namely vertical transmission in the *Wolbachia* genome versus horizontal transfer of
93 phage WO. The genes were proposed as candidate CI effectors due to the presence of one of the
94 protein products in the spermathecae of infected female mosquitoes (Beckmann and Fallon 2013)
95 and their absence in the *wAu Wolbachia* strain that lost CI function (Sutton, et al. 2014).

96
97 The *wMel* homologs of these genes are designated cytoplasmic incompatibility factors *cifA*
98 (locus WD0631) and *cifB* (locus WD0632), with *cifA* always encoded directly upstream of *cifB*
99 (LePage, et al. 2017). The gene set occurs in varying copy number across eleven total CI-
100 inducing strains that correlates with CI levels. Core sequence changes of the two genes exhibit a
101 pattern of codivergence and in turn closely match bidirectional incompatibility patterns between
102 *Wolbachia* strains. Homologs of CifA and CifB protein sequences belong to four distinct
103 phylogenetic types (designated Types I – IV) that do not correlate with various phylogenies of
104 *Wolbachia* housekeeping genes or phage WO *gpW* (locus WD0640) (LePage, et al. 2017). The
105 homologous sequences in *wPip* also cluster in Type I, though they are 66% and 76% different
106 from *wMel*'s, respectively (Beckmann, et al. 2017). Hereinafter we use *cifA* and *cifB* to refer to
107 these genes, unless specifically referring to analyses of the *wPip* homologs, *cidA* and *cidB*. *In*
108 *vitro* functional analyses revealed that *cidB* can encode deubiquitylase activity, and *cidA* encodes
109 a protein that binds CidB (Beckmann, et al. 2017). Mutating the catalytic residue in the
110 deubiquitylating domain of CidB results in a loss of the CI-like function in transgenic flies
111 (Beckmann, et al. 2017). Whether these genes have additional enzymatic or regulatory roles and
112 which other residues are important for function remain open questions.

113

114 There are important considerations for the location, organization, and characterization of these
115 genes. Whether or not *cifA* and *cifB* form a strict, toxin-antitoxin operon is debatable, and
116 likewise has important implications for how gene expression is regulated by *Wolbachia* during
117 host infection. Support for the operon hypothesis is based on weak transcription across the
118 junction between *cidA* and *cidB*, inferred to be due to the presence of polycistronic mRNA
119 (Beckmann and Fallon 2013; Beckmann, et al. 2017); an alternative explanation is transcriptional
120 slippage. Quantitative transcription analyses and computational predictions of operon structure
121 do not support the operon hypothesis (LePage, et al. 2017). Moreover and importantly,
122 transgenic studies show that both *cifA* and *cifB* are required for induction of CI and thus cannot
123 form a strict toxin (*cifB*) - antitoxin (*cifA*) system. As both genes encode CI function and can
124 individually enhance *Wolbachia*-induced CI, and there is mixed evidence for classification as an
125 operon, it does not appear that characterization as a strict toxin-antitoxin operon is warranted
126 (LePage, et al. 2017). However, like toxin-antitoxin systems, CidA binds CidB *in vitro* and
127 expression of *cidA* rescues temperature-sensitive growth inhibition induced by *cidB* expression
128 in *Saccharomyces*, via an as-yet-unknown mechanism (Beckmann, et al. 2017).

129
130 As it stands now, the genes remain largely unannotated with the exception of a few small
131 domains. If other predicted protein domains occur in *cifA* and *cifB*, they would provide new
132 hypotheses for the mechanism of CI. Finally, the sequence diversity and/or loss of *cif* genes
133 across the *Wolbachia* tree may give insights into the selective conditions that maintain the *cif*
134 genes versus those that do not. Exploration of *cif* gene regulation, expression, and function thus
135 can provide a framework for more targeted investigations of *Wolbachia*-host interactions, and
136 potentially inform the deployment of *Wolbachia*-based arthropod control.

137

138 **Materials and Methods**

139 **Expression**

140 For analysis of RNAseq data we used our published approach (Gutzwiller, et al. 2015). Briefly,
141 fastq sequences for 1 day old male and female flies were mapped against the *Wolbachia wMel*
142 reference genome (Ensembl Genomes Release 24,
143 *Wolbachia_endosymbiont_of_drosophila_melanogaster.GCA_000008025.1.24*) using bwa mem
144 v. 0.7.5a with default parameters in paired-end mode. Mapped reads were sorted and converted
145 to BAM format using samtools v0.1.19 after which BAM files were used as input to Bedtools
146 (bedcov) to generate pileups and count coverage at each position. For expression correlations
147 between genes, the raw RNAseq counts were divided by (gene length + 99), where 99
148 corresponds to read length (100) – 1. Within a growth stage these values were multiplied by $1e^6 /$
149 (sum of values in stage) (Li and Dewey 2011). A pairwise distance between all genes was
150 defined as $(1 - R)$, where the R is the Pearson correlation coefficient between the normalized
151 expression values of two genes. Possible negative correlations would be “penalized” here,
152 resulting in a larger distance. Distances were clustered using the Kitsch program of PHYLIP
153 (Felsenstein 1989).

154

155 **Operon Prediction *in silico***

156 We used the dynamic profile of the transcriptome above to identify operons within the *wMel*
157 genome using two different approaches. We used the program Rockhopper (McClure, et al.
158 2013), using default parameters, in conjunction with the BAM files generated above to delineate
159 likely operons across the entire genome. In addition, we took a fine-scale approach, focusing on

160 the junction between *cifA* and *cifB* (Fortino, et al. 2014), using the pileup files generated above
161 and identifying drops in gene expression correlated to genomic position using a sliding window
162 analysis.

163

164 **Nucleic Acid Extractions and Quantitative PCR**

165 To identify *Wolbachia* gene expression in adult male and female *D. melanogaster*, RNA was
166 extracted from individual, age-matched flies (1-3 days old, stock 145) using a modified Trizol
167 extraction protocol. Briefly, 500 uL of Trizol was added to individual flies and samples
168 homogenized using a pestle. After a 5-minute incubation at room temperature, a 12,000 rcf
169 centrifugation (at 4C for 10 min) was followed by a chloroform extraction. Aqueous phase
170 containing RNA was extracted a second time with phenol:chloroform before isopropanol
171 precipitation of RNA. This RNA pellet was washed and resuspended in THE RNA Storage
172 Solution (Ambion). To detect the number of *cifA* and *cifB* transcripts as well as RNA levels
173 across the junction between *cifA* and *cifB*, we utilized the RNA extracted from these flies and the
174 SensiFAST SYBER Hi-ROX One-step RT mix (Bioline) and the Applied Biosystems StepOne
175 Real-time PCR system with the following primer sets: *cifAF*: ATAAAGGCGTTTCAGCAGGA,
176 *cifAR*: TCAATGAGGCGCTTCTAGGT; *cifBF*: TACGGGAAGTTTCATGCACA,
177 *cifBR*: TTGCCAGCCATCATTCTATAA; *cifA_endF*:
178 TCTGGTTCTCATAAGAAAAGAAGAATC, *cifB_begR*: AACCATCAAGATCTCCATCCA.
179 As a reference for transcription activity of the core *Wolbachia* genome, we utilized the
180 *Wolbachia ftsZ* gene (Forward: TTTTGTTGTCGCAAATACCG; Reverse:
181 AGCAAAGCGTTCACATTTCC). We designed primers to *ftsZ* because as a core protein
182 involved in cell division, the quantities of *ftsZ* would better correlate with bacterial numbers and

183 activity. Reactions were performed in duplicate or triplicate in a 96-well plate and CT values
184 generated by the machine, were used to calculate the relative amounts of *Wolbachia* using the
185 $\Delta\Delta\text{Ct}$ (Livak) method.

186

187 **Correlated Cif Trees and Distance Matrices**

188 Quantifying congruence scores between the CifA and CifB trees was carried out with Matching
189 Cluster (MC) and Robinson Foulds (RF) metrics using a custom python script previously
190 described (Brooks, et al. 2016) and the TreeCmp program (Bogdanowicz, et al. 2012). MC
191 weights topological congruency of trees, similar to the widely used RF metric. However, MC
192 takes into account sections of subtree congruence and therefore is a more refined evaluation of
193 small topological changes that affect incongruence. Significance in the MC and RF analyses was
194 determined by the probability of 100,000 randomized bifurcating dendrogram topologies
195 yielding equivalent or more congruent trees than the actual tree. Normalized scores were
196 calculated as the MC and RF congruency score of the two topologies divided by the maximum
197 congruency score obtained from random topologies. The number of trees that had an equivalent
198 or better score than the actual tree was used to calculate the significance of observing that
199 topology. Mantel tests were also performed on the CifA and CifB patristic distance matrices
200 calculated in Geneious v8.1.9 (Kearse, et al. 2012). A custom Jupyter notebook (Pérez and
201 Granger 2007) running python v3.5.2 (<http://python.org>) was written in the QIIME2 (Caporaso,
202 et al. 2010) anaconda environment, and the Mantel test (Mantel 1967) utilized the scikit-bio
203 v0.5.1 (scikit-bio.org) Mantel function run using scikit-bio distance matrix objects for each gene.
204 The Mantel test was run with 100,000 permutations to calculate significance of the Pearson
205 correlation coefficient between the two matrices using a two-sided correlation hypothesis.

206

207 **Genomes Used in Comparative Analyses**

208 In order to identify *cif* homologs across the *Wolbachia* genomes, we defined orthologs across
209 existing, sequenced genomes using reciprocal best blastp. We included *Wolbachia* genomes
210 across five Supergroups: monophyletic clades of *Wolbachia* based on housekeeping genes,
211 denoted by uppercase letters (O'Neill, et al. 1992; Werren, et al. 1995). Supergroups A and B are
212 the major arthropod infecting lineages, while C and D infect nematodes (Bandi, et al. 1998).
213 Supergroup F *Wolbachia* infect a variety of hosts (Lo, et al. 2002). Included in this analysis were
214 11 type A strains (*w*Ri, *w*Ana, *w*Suzi, *w*Ha, *w*Mel, *w*MelPop, *w*Au, *w*Rec, *w*Gmm, *w*Uni,
215 *w*VitA), 10 type B strains (*w*PipJHB, *w*PipPel, *w*PipMol, *w*Bol1-b, *w*Bru, *w*CauB, *w*No, *w*Tpre,
216 *w*AlbB, *w*Di), 2 type C strains (*w*Ov, *w*Oo), and one each type D (*w*Bm) and type F (*w*Cle). We
217 included all genomic data available for each strain such that if multiple assemblies existed for
218 each *Wolbachia* variant (such as in the case of *w*Uni) we included the union of all available
219 contigs for that strain. *Wolbachia* orthologs were defined based on reciprocal best blast hits
220 between amino acid sequences in *Wolbachia* genomes. An orthologous group of genes was
221 defined by complete linkage such that all members of the group had to be the reciprocal best hit
222 of all other members of the group. *w*Ana, *w*Gmm, *w*PipMol, *w*Bru, and *w*CauB were not used in
223 subsequent analyses due to problematic assemblies. Information on strain phenotypes, hosts, and
224 accession numbers can be found in Table 1.

225

226 **Cif Phylogenetics**

227 CifA and CifB protein sequences were identified using BLASTp searches of WOMelB WD0631
228 (NCBI accession number AAS14330.1) and WD0632 (AAS14331.1), respectively. Homologs

229 were selected based on: 1) $E \leq 10^{-30}$, 2) query coverage greater than 70%, and 3) presence
230 in fully sequenced *Wolbachia* genomes. All sequences were intact with the exception of a partial
231 WOSuziC CifA (WP_044471252.1) protein. The missing N-terminus was translated from the
232 end of contig accession number CAOU02000024.1 and concatenated with partial protein
233 WP_044471252.1 for analyses, resulting in 100% amino acid identity to WORiC CifA
234 (WP_012673228.1). In addition, two previously identified sequences (LePage, et al. 2017),
235 WOREcB CifB and WORiB CifB, were not available in NCBI's database and translated from
236 nucleotide accession numbers JQAM01000018.1 and CP001391.1, respectively. The previously
237 identified WOSol homologs (CifA: AGK87106 and CifB: AGK87078) (LePage, et al. 2017)
238 were also included in our analyses. All protein sequences were aligned with the MUSCLE
239 (Edgar 2004) plugin in Geneious Pro version 8.1.7 (Kearse, et al. 2012); the best models of
240 evolution, according to corrected Akaike (Hurvich and Tsai 1993) information criteria, were
241 estimated to be JTT-G using the ProtTest server (Abascal, et al. 2005); and phylogenetic trees
242 were built using the MrBayes (Ronquist, et al. 2012) plugin in Geneious.

243

244 **Protein Structure**

245 All candidate CI gene protein sequences were assessed for the presence of domain structure
246 using HHpred (<https://toolkit.tuebingen.mpg.de/hhpred/>) (Söding, et al. 2005)) with default
247 parameters and the following databases: SCOPe95_2.06, SCOPe70_2.06, cdd_04Jul16,
248 pfamA_30.0, smart_04Jul16, COG_04Jul16, KOG_04_Jul16, pfam_04Jul16, and cd_04Jul16.
249 Schematics were created in inkscape (<https://inkscape.org/>), to show regions with significant
250 structural hits, at a corrected p-value of $p < 0.05$. Modules were defined based on the presence of
251 multiple highly significant hits within a region.

252

253 **Protein Conservation**

254 Protein conservation was determined with the Protein Residue Conservation Prediction tool
255 (<http://compbio.cs.princeton.edu/conservation/index.html> (Capra and Singh 2007)), using
256 aligned amino acid sequences, Shannon entropy scores, a window size of zero, and sequence
257 weighting set to “false”. Conservation was subsequently plotted in R version 3.3.2, and module
258 regions were delineated according to coordinates of the WOMelB modules within the alignment.
259 CI gene conservation scores were calculated separately for Type I sequences, and for all types
260 together. For CifB Type I sequences, the WOVitA4 ortholog was left out, due to the extended C-
261 terminus of that protein. Conservation scores were also calculated for “control proteins”: Wsp
262 (*Wolbachia* surface protein), known to be affected by frequent recombination events (Baldo, et
263 al. 2005), and FtsZ, which is relatively unaffected by recombination (Baldo, et al. 2006b; Ros, et
264 al. 2009). Variation in amino acid conservation between modules and non-module regions was
265 assessed in R version 3.3.2 with a one-way ANOVA including “region” (either the unique
266 module number, or “non-module”) as a fixed effect, and followed by Tukey Honest Significant
267 Difference for post hoc testing.

268

269 **Cif Modules**

270 The WOMelB structural regions delineated by HHpred were used to search for the presence of
271 Cifs or remnants of Cifs across the *Wolbachia* phylogeny. Amino acid sequences of the
272 WOMelB modules were queried against complete genome sequences (Table 1) using tblastn.
273 Any hit that was at least 50% of the length and 30% identity, or at least 90% of the length and
274 20% identity of the WOMelB module was considered a positive match. Module presence was

275 plotted across a *Wolbachia* phylogeny constructed using the five Multi Locus Sequence Typing
276 (MLST) genes defined by Baldo *et al.* (Baldo, et al. 2006b). Nucleotide sequences were aligned
277 with MAFFT version 7.271 (Kato and Standley 2013), and concatenated prior to phylogenetic
278 reconstruction with RAxML version 8.2.8 (Stamatakis 2014), the GTRGAMMA substitution
279 model, and 1000 bootstrap replicates.

280

281 **Hidden Markov Model Searches**

282 To identify *cif* homologs in draft *Wolbachia* genome assemblies we used the program suite
283 HMMER (Eddy 2011). We defined *cif* types based on our phylogenetic trees (Figure 4) and used
284 aligned amino acids from these types as input to HMMBUILD, using default parameters. We
285 then searched six *Wolbachia* WGS assemblies (NCBI project numbers PRJNA310358,
286 PRJNA279175, PRJNA322628) using HMMSEARCH with $-F3\ 1e-20$ $-cut_nc$ and $-domE\ 1e-$
287 10. Regardless of thresholds used, or *cif* type of HMM, resulting hits did not differ.

288

289 **Results**

290 ***cifA* and *cifB* are Not Co-transcribed or Co-regulated and Do Not Comprise an Operon in** 291 ***wMel***

292 To assess the operon hypothesis, we reasoned that genes which are co-transcribed and co-
293 regulated will exhibit the following properties: similar total expression levels in whole animals
294 and correlated gene expression across host development. We therefore utilized an existing
295 RNAseq dataset for *Wolbachia* in *Drosophila melanogaster*, covering 24 life cycle stages and 3
296 time samplings each for adult males and females (Gutzwiller, et al. 2015). We mapped reads to
297 the existing *wMel* assembly (see methods), and calculated Pearson correlation coefficients for

298 normalized expression values for each pairwise comparison across host development. In adult
299 males and females, *cifA* and *cifB* in *wMel* are not expressed at similar levels (Figure 1), with *cifA*
300 expressed at significantly higher levels compared to *cifB* (eight-fold higher based on RPKM
301 values across both genes).

302
303 To further explore expression of the *cif* genes in *wMel* and assess whether or not polycistronic
304 mRNA is produced, we performed a quantitative PCR analysis of gene expression from three-
305 day old male and female flies (Figure 2). We observed transcripts covering the junction between
306 *cifA* and *cifB*. However, transcripts covering this junction were much more similar to expression
307 levels in *cifA*, while expression of *cifB* was nine-fold less. Therefore, as *cifA* and *cifB* are
308 separated by only 76 bp, distinguishing between 3' UTRs from *cifA* and full *cifA-cifB* transcripts
309 is not possible.

310
311 We next used two computational methods to test for a potential operon between *cifA* and *cifB*
312 using our RNAseq analyses. After mapping reads to the *wMel* assembly, we used the resulting
313 BAM files as input to Rockhopper (McClure, et al. 2013). The program was able to correctly
314 identify known operons in *wMel* (such as the T4SS WD0004-WD0008 and the ribosomal protein
315 operon) but it did not identify *cifA* and *cifB* as an operon. We also used a sliding-window
316 approach, using pileup files generated as part of the mapping, to identify correlations between
317 genomic position and gene expression drops in the RNAseq data, as in (Fortino, et al. 2014). The
318 two open reading frames for *cifA* and *cifB* span positions 617223-618647 and 618723-622223,
319 respectively. From positions 618600 to 618700, we observe a significant positive correlation
320 between coverage and genomic location (Pearson Correlation = 0.99, $p < 0.001$). However,

321 across the junction between *cifA* and *cifB* (position 618700), we saw a very large drop in gene
322 expression in both males and females (from an average coverage of 4616 to 38 per position).
323 This result suggests that *cifA* and *cifB* are not co-transcribed.

324
325 Finally, we clustered the *wMel* *cif* genes based on their similarity in expression across
326 *Drosophila* development (Supplemental Figure S1). *cifA* did not group with *cifB* in *wMel* (Figure
327 3), suggesting that these two genes are not co-regulated. Indeed, the pattern of *cifA* expression
328 differs strikingly from that of *cifB*. *cifB* is expressed during embryogenesis and generally down-
329 regulated in pupae and adults, while *cifA* is highly expressed in adult males and females and late
330 time points during embryogenesis (Figure 1). Curiously, the expression profile of *cifA* in flies
331 during development is most closely correlated with the *wsp* locus WD1063 (Figure 3).

332

333 **New Protein Domain Predictions are Variable Across the Cif Phylogeny**

334 We recovered the four previously identified phylogenetic types (LePage, et al. 2017). Here, our
335 analyses include additional strains that cause reproductive parasitism beyond CI
336 (parthenogenesis and male-killing, Table 1), and the more divergent Type IV paralogs for *cifA*,
337 so far identified in B-Supergroup *Wolbachia*. We recover a set of Type III alleles from *wUni*, a
338 strain that induces parthenogenesis in the parasitoid wasp, *Muscidifurax uniraptor* (Stouthamer,
339 et al. 1993). The *wBol1-b* strain, a male-killer that has retained CI capabilities (Hornett, et al.
340 2008), has alleles belonging to both Type I and Type IV.

341

342 Homologs and predicted protein domains of CifA and CifB for all four phylogenetic types
343 (LePage, et al. 2017) from *Wolbachia* strains that cause CI, parthenogenesis, male-killing, or no

344 reproductive phenotype were characterized by HHpred homology and domain structure
345 prediction software (Söding, et al. 2005). Search parameters are described in the methods.
346 Several new prominent protein domains (as determined by the presence of multiple highly
347 significant structural predictions within a region), herein referred to as “modules”, were
348 identified for each CifA and CifB protein sequence. In Table 2 we list the prominent module
349 annotations identified across CifA and CifB Types. Multiple structural hits within a region can
350 be explained by the homology of the significant domains predictions to each other.

351
352 For CifA, three main modules were annotated (Figure 4A, Table 2). First, the most N-terminal
353 module (ModA-1) in Type I, II and III variants shows homology to Catalase-rel ($p = 0.001$ -
354 0.003), which is predicted to catalyze the breakdown of hydrogen peroxide (Chelikani, et al.
355 2004) (Type I) and protect the cell from toxic effects, or VirJ ($p = 0.002$ - 0.003), a bacterial
356 virulence protein and component of T4SS secretion systems (Pantoja, et al. 2002) (Types II and
357 III). The second CifA module in the central region (ModA-2) has homology to a caspase
358 recruitment domain ($p = 0.005$ - 0.009), venom and toxin-related domains ($p \leq 0.001$), and a
359 thermal regulator protein ($p = 0.002$). The very significant homology to a toxin is interesting,
360 given that CifA was hypothesized to act as an antitoxin. Notably, CifA is required for and
361 enhances the induction of CI (LePage, et al. 2017), which contradicts its proposed function as
362 simply an antitoxin (Beckmann, et al. 2017). The last CifA module in the C-terminal region
363 (ModA-3) has multiple strong hits to a STE-like transcription factor ($p \leq 0.001$). There were
364 additional annotations that emerged due to weak or singular matches. In Type IV variants, there
365 is a separate N-terminal region that shares homology with a conserved eukaryotic family with
366 potential methyltransferase activity, FAM86 ($p = 0.003$). Most Type I alleles have C-terminal

367 homology to a nuclear cap-binding protein that binds RNA ($p = 0.010 - 0.020$). WOHa1,
368 WOBol1b, and WOSol have an additional N-terminal region containing a conserved domain of
369 unknown function ($p = <0.001 - 0.005$). Type IV genes have a yeast-like salt tolerance down-
370 regulator domain NST1 ($p = 0.003$). Lastly, WOVit4 and wUni lack the most N-terminal CifA
371 homology region, ModA-1.
372
373 For CifB, three main modules were defined (Figure 4B, Table 2). The first (ModB-1) and second
374 (ModB-2) most N-terminal regions both have matches to the PDDEXK nuclease family ($p <$
375 0.001), the HSDR_N restriction enzyme ($p = <0.001-0.010$), and domains of unknown function
376 (DUF1052, DUF91). The third module, found only in the Type I C-terminus (ModB-3), has very
377 strong homology to a number of ubiquitin-modification and peptidase domains ($p < 0.001$), as
378 well as YopJ, which in *Yersinia*, aids in infecting a eukaryotic host (Paquette, et al. 2012) ($p =$
379 $<0.001-0.020$). ModB-3 contains the catalytic residue associated with toxicity/CI function in
380 CidB (Beckmann, et al. 2017). In addition to the annotated modules, all Type I alleles except
381 WOBol1b and WORiB have a single hit to a conserved domain of unknown function in the N-
382 terminus ($p = 0.001 - 0.005$), and Type III alleles (except for wAlbB) have a region of homology
383 to a methyltransferase domain (MTS) ($p < 0.001$). Both Type II and III alleles have a single short
384 hit in the N-terminus to a SecA regulator. WOVitA4 (Type 1) has an extended C-terminus not
385 present in any other alleles, and within that extended C-terminus is an additional
386 peptidase/YopT-like region, similar to ModB-3. CifB Type IV alleles (WOAlbB, WOPip2, and
387 wBol1-b) were not included in the phylogenetic reconstruction, as they are highly divergent and
388 not reciprocal blasts of WOMelB *cifB*. Despite their divergence, these Type IV CifB alleles have
389 similar structures to Type II and III alleles: two PDDEXK-like modules, and no Ulp-1-like

390 module three (Supplemental Figure S3). Full structural schematics with exact coordinates and
391 homology regions for each allele are available in the supplemental material (Supplemental
392 Figures S2 and S3), as are all significant domain hits with associated p-values and extended
393 descriptions (Supplemental Tables S1 and S2).

394

395 **CifA and CifB Codiverge**

396 Initial phylogenetic trees based on core amino acid sequences of Type I-III variants of CifA and
397 CifB exhibited similar trees (LePage, et al. 2017). Here we statistically ground the inference of
398 codivergence using the largest set of *Wolbachia* homologs to date. We quantified congruence
399 between the CifA and CifB phylogenetic trees for Types I-III (Supplemental File S1) using
400 Matching Cluster (MC) and Robinson–Foulds (RF) tree metrics (Bogdanowicz and Giaro 2013;
401 Bogdanowicz, et al. 2012; Robinson and Foulds 1981), with normalized distances ranging from
402 0.0 (complete congruence) to 1.0 (complete incongruence). Results show strong levels of
403 congruence between CifA and CifB ($p < 0.00001$ for both, normalized MC = 0.06 and
404 normalized RF = 0.125). To further statistically validate the inference of codivergence, we
405 measured the correlation between patristic distance matrices for CifA and CifB using the Mantel
406 test (Mantel 1967). Results demonstrate a high degree of correlation between patristic distance
407 matrices, and through permutation show that independent evolution of CifA and CifB is highly
408 unlikely (Pearson correlation coefficient = 0.905, $p = 0.00001$).

409

410 **Cif Proteins Evolve Rapidly**

411 Amino acid sequence conservation across the full length of the Cif proteins was determined and
412 compared to *Wolbachia* amino acid sequences of genes that either have signatures of

413 recombination and directional selection (Wsp, *Wolbachia* surface protein) or have not undergone
414 extensive recombination and directional selection (FtsZ, cell division protein). Wsp protein
415 sequences exhibit considerable divergence (mean conservation = 0.85), with very few sites in a
416 row being completely conserved (Figure 5A). In contrast, FtsZ is relatively conserved (mean
417 conservation = 0.94), and most of the divergence is clustered at the C-terminus (Figure 5B).
418 Mean conservation for the Cif protein sequences were lower than Wsp - 0.83 for Type I CifA
419 alleles (Figure 5C) and 0.82 for Type I CifB alleles (Figure 5E, Table 3). When all Cif alleles
420 were considered, mean conservation was even further reduced - 0.58 for CifA (Figure 5D) and
421 0.43 for CifB (Figure 5F). The lower average conservation of CifB genes is in part due to the
422 many insertions and deletions in the alignment, and the missing C-terminal deubiquitylase
423 region, ModB-3, of the Type II and III alleles. Thus, several CifB proteins apparently lack this
424 activity, and whether these variants cause CI remains to be determined. Importantly, although the
425 CifB proteins are highly divergent, the catalytic residue (red dot in Figures 5E and 5F) in the
426 deubiquitylating module of CifB is unique to and completely conserved for the Type I alleles.
427 The Cif proteins have extensive amounts of diversity, with completely conserved amino acids
428 distributed across the length of the protein, and not confined to any particular regions (Figure
429 5C-F, Supplemental Tables S3-S6). There were significant differences in the level of
430 conservation between modules and non-module regions for the Type I alignments of both CifA
431 ($F_{3,495} = 11.75$, $p = 0.0021$) and CifB ($F_{3,1195} = 11.75$, $p = 1.38e-07$) (Table 3). The only module
432 that had significantly higher conservation than the non-module regions of the alignment was
433 ModB-1 ($p = 0.0173$). The *w*Mel strain contains the (P)D-(D/E)XK motif (blue dots in Figures
434 5E and 5F) (Kosinski, et al. 2005), but it is less than 80% conserved across strains despite the
435 higher average conservation of this module. In contrast, ModA-3 and ModB-3 are significantly

436 less conserved than the non-module regions of the corresponding proteins (CifA, $p = 0.0400$;
437 CifB, $p = 0.0001$).

438

439 **Cif Module Presence Generally Predicts Reproductive Phenotype**

440 We used the *wMel* predicted Cif modules as a seed to search for the presence of homologous
441 modules across *Wolbachia* genome sequences using tblastn (Figure 6). In strains with more
442 divergent Cif Types, we report modules that were expected based on the HHpred results, but not
443 recovered with tblastn due to sequence divergence from WOMelB. For example, the WOSuziC
444 and WORiC ModA-1 (Catalase-rel in *wMel* and other Type I, VirJ in Type II and III) was not
445 recovered. Additionally, we recover homologous modules outside of the annotated *cif* open
446 reading frames, such as the chromosomal region with a ModB-3 (Ulp-1-like) region in *wNo*. The
447 high number of modules in *wSuzi* and *wRi* are due to the presence of a duplicated set of Type I
448 *cifs*. All arthropod-infecting strains, with the exception of *wAu* (a non-CI inducing strain),
449 contained at least one recovered module. This includes the bed-bug mutualist *wCle*, found in
450 Supergroup-F, and two strains that have lost CI abilities, *wUni* and *wTpre*. Importantly, all
451 strains that are known to be capable of inducing or rescuing CI have four or more recovered
452 modules, though they do not necessarily have ModB-3, which contains the catalytic residue
453 implicated in CI function (Beckmann, et al. 2017). The non-CI strains have fewer recovered
454 modules: ModB-1 in *wTpre*, ModB-1 and -2 in *wUni*, ModA-3 in *wCle*. and no modules in *wAu*
455 and the nematode-infecting strains. *wUni* is a unique case, where we identified *cif* alleles in the
456 genome, but recovered relatively few modules. The CifA modules are either missing (Figure 4A)
457 or divergent enough from WOMelB that they were not considered a positive match. The two N-
458 terminal *wUni* CifB modules, ModB-1 and ModB-2, are relatively more conserved, and the

459 ModB-3 is missing due to the truncated C-terminus present in all non-Type I CifB alleles (Figure
460 4B). *wAlbB* and *wNo*, both CI-inducing strains with Type III and IV alleles, have fewer
461 recovered modules, but this is congruent with the more divergent nature of those Cif types. We
462 recovered many modules in *wSuzi*, which is a strain not known to induce CI (Cattel, et al. 2016;
463 Hamm, et al. 2014). This discrepancy between *cif* presence and absence of a reproductive
464 phenotype might be explained by the disrupted Type II *cifA* in *wSuzi*. The split WOSuziC
465 sequenced was concatenated to allow for a more robust phylogenetic reconstruction (Figure 4),
466 but it is in fact disrupted by a transposase (Conner, et al. 2017). However, having a functional set
467 of Type I *cif* alleles appears to be sufficient for CI-induction in other strains (Beckmann, et al.
468 2017; LePage, et al. 2017), so it is not clear how inactivation of the Type II alleles here may
469 affect the final CI phenotype. Strain *wDi*, infecting the Asian citrus psyllid *Diaphorina citri*, has
470 no identified reproductive phenotype, but only contains a single module: ModB-1.

471
472 The lack of evidence for homologous *cif* genes in the nematode-infecting *Wolbachia* agrees with
473 previous findings (LePage, et al. 2017) that CI-function is restricted to the A+B-Supergroup
474 clade (likely due to WO phage activity), and the absence of WO phages for the nematode-
475 infecting strains (Gavotte, et al. 2007). The loss of CI within the A and B Supergroups is likely a
476 derived trait due to the rapid evolution of prophage WO (Ishmael, et al. 2009; Kent, et al.
477 2011b), and relaxed selection after transition to a new reproductive phenotype. The low number
478 of modules identified in such strains is consistent with gene degradation and loss.

479
480 To further explore the conservation of the *cif* genes across the sequenced *Wolbachia*, and to
481 uncover diversity that may be present in other genomes, we searched the WGS databases for

482 recently sequenced genomic scaffolds from *Wolbachia* infecting the *Nomada* bees (*wNleu*, *wNla*,
483 *wNpa*, *wNfe*) (Gerth and Bleidorn 2016), *Drosophila inocompta* (*wInc_Cu*)(Wallau, et al. 2016),
484 and *Laodelphax striatellus* (*wStri*) (GenBank Accession Number NZ_LRUH00000000.1) using
485 HMMER. Only for *wStri* do we have direct evidence of CI induction (Noda, et al. 2001) yet the
486 *wStri* and *wInc_Cu* WGS projects each contain only one *cif* locus, with distant homology to *cifA*
487 (~25% identify across 60% of the *wMel* protein). Based on HHpred analyses, the *wStri* homolog
488 (WP_063631193.1) contains none of the domain modules associated with *cifA*. The *wInc_Cu*
489 homolog (WP_070356873.1) contains three modules: an N-terminal Catalase-rel domain and an
490 internal Ectatomin domain, followed by the STE like transcriptional factor domain. Because
491 these are incomplete genome projects, it is possible that other *cif* homologs have been missed
492 due to the current sequencing coverage. Alternatively, it is possible that other, as yet
493 undiscovered, mechanisms of reproductive manipulation exist in these strains. In contrast, the
494 *Nomada*-associated *Wolbachia* contain a large repertoire of *cif* homologs, including Type I, II,
495 III, IV and several homologs with variations on the Type IV domain architecture for *cifA*
496 (Supplemental Figure S4). The *Nomada Wolbachia* all harbor Type II *cifB* homologs and each of
497 the strains harbors either duplicates of this *cifB* type or novel domain architectures for *cifB*
498 including an N-terminal Oleosin domain and a C-terminal Ulp-1 domain (Supplemental Figure
499 S4).

500

501 **Discussion**

502 We explored three key features of *cif* evolution: (i) the toxin-antitoxin operon hypothesis, (ii)
503 potential enzymatic and regulatory functions across the *cifA* and *cifB* phylogenies, and (iii) the
504 conservation and diversity of *cif* genes across strains with different host-manipulation

505 phenotypes. We provide multiple lines of evidence that *cifA* and *cifB* do not comprise an operon
506 in *wMel*, including quantifications of transcription and *in silico* operon predictors. Moreover,
507 expression of *cifA* and *cifB* across host development are not correlated with each other. In fact,
508 *cifB* expression does not significantly correlate with any other *Wolbachia* locus. Combined with
509 the drastic drop off in expression across the short junction between *cifA* and *cifB*, and negative
510 results from the operon prediction software, we conclude that *cifA* and *cifB* are not co-transcribed
511 or co-regulated as an operon in *wMel*, the *Wolbachia* strain currently used in mosquito control
512 programs. While we think it unlikely that the *cif* genes are regulated and transcribed in
513 drastically different ways across closely related *cif* Types, more detailed analyses from a variety
514 of strains would be beneficial for developing a comprehensive understanding of the factors
515 regulating expression of these genes. It is especially interesting that synteny has generally been
516 maintained across prophage WO regions, despite the high level of recombination and
517 rearrangements in prophage WO and *Wolbachia* genomes (Baldo, et al. 2006a; Ellegaard, et al.
518 2013; Kent, et al. 2011a). It is not clear if there is an advantage (and what the advantage may be)
519 to maintaining syntenic orientation of these two genes; perhaps this feature can be attributed to
520 their location within prophage WO and/or functions associated with the ability of *cifA* and *cifB* to
521 act synergistically to induce CI (LePage, et al. 2017). Since type IV secretion system genes and
522 their predicted effectors are scattered across the *Wolbachia* genome (Rice, et al. 2017; Wu, et al.
523 2004) gene products involved in *Wolbachia*-host interactions can function together even when
524 the genes encoding them are not syntenic. We conclude that *cifA* and *cifB* do not comprise an
525 operon, and do not act strictly as a toxin-antitoxin system due to the requirement of both proteins
526 for the induction of CI in the insect host. Determining how *cifA* and *cifB* expression is regulated
527 in the insect host will greatly benefit vector control programs that use *Wolbachia*-mediated CI.

528
529 Despite the conservation of gene order, Cif proteins showed extensive amounts of divergence
530 and differences in domain structure as previously reported (LePage, et al. 2017). Here, the levels
531 of amino acid conservation in the Cifs are lower than FtsZ and Wsp, the latter of which is known
532 to recombine and be subject to directional selection. The conservation of the catalytic residue in
533 the C-terminal deubiquitylase domain is an important feature of CidB (Beckmann, et al. 2017).
534 However, only Type I of the four identified Types has this domain. Additionally, strains known
535 to induce CI, such as *wAlbB* and *wNo* have no Type I alleles, implying that the Ulp-1 region
536 may not be essential for inducing CI. The complete, functional capacity of Types I-IV has yet to
537 be explored *in vivo*, but is a promising direction for understanding the evolution of *Wolbachia*-
538 host associations.

539
540 Based on what is known about *Wolbachia* biology, some of the protein domains may be
541 especially good candidates for further study and *in vivo* functional characterization. Predicted
542 PDDEXK-like domains are present in all four CifB types. Given the predicted interaction of
543 these domains with DNA (Kosinski, et al. 2005), and the presence of these domains across CifB
544 proteins, determining if and how these regions interact with host (*Wolbachia* or insect) DNA,
545 and whether or not they contribute to CI function would be useful in understanding the consistent
546 presence of this module. Another good candidate for further exploration is the predicted
547 methyltransferase domain in several Type III CifB proteins, as *Wolbachia* infection has been
548 linked to changes in host genome methylation in several insects (LePage, et al. 2014; Negri, et al.
549 2009; Ye, et al. 2013), though knockout of *Drosophila* methyltransferases does not alter CI
550 levels (LePage, et al. 2014). Likewise, the antioxidant catalase domain is noteworthy as these

551 domains decompose hydrogen peroxide into water and oxygen and thus protect cells from its
552 toxic effects, which are present in *Wolbachia*-infected spermatocytes (Brennan, et al. 2012).
553
554 *Wolbachia* strains that have lost CI have a strong signature of *cif* gene degradation and loss. The
555 two parthenogenesis-inducing strains (*wTpre* and *wUni*) appear to be at different places in this
556 process of gene loss, with divergent Cif amino acid sequences recovered for *wUni*, but only one
557 PDDEXK module identified in *wTpre*. There are several explanations for this. *wUni* is likely a
558 more recent transition to parthenogenesis, as it is closely related to a CI strain (*wVitA*) (Baldo, et
559 al. 2006b; Newton, et al. 2016). In comparison, *wTpre* is part of a unique clade of *Wolbachia* that
560 all induce parthenogenesis in *Trichogramma* wasps (Rousset, et al. 1992; Schilthuizen and
561 Stouthamer 1997; Werren, et al. 1995). This strain has lost its WO phage association and only
562 has relics of WO phage genes (Gavotte, et al. 2007; Lindsey, et al. 2016). Additionally, the two
563 strains that independently transitioned to the parthenogenesis phenotype have evolved separate
564 mechanisms for doing so (Gottlieb, et al. 2002; Stouthamer and Kazmer 1994). Differences in
565 time since transition to the parthenogenesis phenotype, phage WO associations, and mechanisms
566 of parthenogenesis induction likely all play a role in the rate of *cif* gene degradation.
567
568 Based on our analyses, we propose three avenues of research on the function of the Cif proteins.
569 First, functional confirmation of the newly annotated modules will be important to understanding
570 how these genes function enzymatically. In total, we predict six modules in the Cif protein
571 sequence homologs, with varying degrees of confidence (Supplemental Tables S1 and S2
572 Tables). For some of these modules, straightforward experiments can be designed in model
573 systems (such as *Saccharomyces*) to determine if their predicted function is correct, as has been

574 done for CidB (Beckmann, et al. 2017) and countless other bacterial effectors (Archuleta, et al.
575 2011; Kramer, et al. 2007; Siggers and Lesser 2008). Second, necessity and importance of these
576 modules to the CI phenotype can be assessed in the *Drosophila* model, where the induction of
577 the phenotype and rescue is straightforward (LePage, et al. 2017). Finally, we suggest that
578 although the discovery of these genes is fundamental, it is clear from this analysis that we have
579 not comprehensively evaluated or identified the mechanisms behind CI and other reproductive
580 manipulations. The gene characterization analyses described here reveal new and relevant
581 annotations, substantial unknown sequence regions across all of the phylogenetic types, missing
582 deubiquitylase domains in particular CI strains, and a coevolving, phylogenetic relationship
583 across the Cif trees. Importantly, the locus and mechanism behind rescuing CI are still unknown,
584 as is the exact mechanism by which all Cif proteins induce CI. Therefore, the recent discovery of
585 these genes, and the gene characterization analyses described here, pave the most comprehensive
586 road to date for investigating key mechanisms of the *Wolbachia*-host symbiosis.

587

588 **Acknowledgments**

589 This work was supported by the National Science Foundation (DEB 1501227 to A.R.I.L., IOS
590 1456545 to I.L.G.N., and IOS 1456778 to S.R.B); the United States Department of Agriculture
591 (NIFA 2016-67011-24778 to A.R.I.L.); the National Institutes of Health (R21 HD086833 and
592 R01 AI132581 to S.R.B.); and Robert and Peggy van den Bosch Memorial Scholarships to
593 A.R.I.L. We thank J. Dylan Shropshire for feedback on an earlier draft of the manuscript.

594 **References**

- 595 Abascal F, Zardoya R, Posada D 2005. ProtTest: selection of best-fit models of protein
596 evolution. *Bioinformatics* 21: 2104-2105.
- 597 Archuleta TL, et al. 2011. The *Chlamydia* effector chlamydial outer protein N (CopN) sequesters
598 tubulin and prevents microtubule assembly. *J Biol Chem* 286: 33992-33998.
- 599 Baldo L, Bordenstein S, Wernegreen JJ, Werren JH 2006a. Widespread recombination
600 throughout *Wolbachia* genomes. *Mol Biol Evol* 23: 437-449.
- 601 Baldo L, et al. 2006b. Multilocus sequence typing system for the endosymbiont *Wolbachia*
602 *pipientis*. *Appl Environ Microbiol* 72: 7098-7110. doi: 10.1128/aem.00731-06
- 603 Baldo L, Lo N, Werren JH 2005. Mosaic nature of the *Wolbachia* surface protein. *J Bacteriol*
604 187: 5406-5418.
- 605 Bandi C, Anderson TJC, Genchi C, Blaxter ML 1998. Phylogeny of *Wolbachia* in filarial
606 nematodes. *Proc R Soc Lond B* 265: 2407-2413.
- 607 Beckmann JF, Fallon AM 2013. Detection of the *Wolbachia* protein WPIP0282 in mosquito
608 spermathecae: implications for cytoplasmic incompatibility. *Insect Biochem Mol Biol* 43: 867-
609 878. doi: 10.1016/j.ibmb.2013.07.002
- 610 Beckmann JF, Ronau JA, Hochstrasser M 2017. A *Wolbachia* deubiquitylating enzyme induces
611 cytoplasmic incompatibility. *Nat Micro* 2: 17007. doi: 10.1038/nmicrobiol.2017.7
- 612 Bian GW, Xu Y, Lu P, Xie Y, Xi ZY 2010. The endosymbiotic bacterium *Wolbachia* induces
613 resistance to dengue virus in *Aedes aegypti*. *PLoS Path* 6. doi: 10.1371/journal.ppat.1000833

- 614 Bogdanowicz D, Giaro K 2013. On a matching distance between rooted phylogenetic trees. Intl J
615 Appl Math Comp Sci 23: 669-684.
- 616 Bogdanowicz D, Giaro K, Wróbel B 2012. TreeCmp: Comparison of trees in polynomial time.
617 Evol Bioinf Online 8: 475.
- 618 Bordenstein SR, Bordenstein SR 2016. Eukaryotic association module in phage WO genomes
619 from *Wolbachia*. Nat Commun 7.
- 620 Bordenstein SR, Marshall ML, Fry AJ, Kim U, Wernegreen JJ 2006. The tripartite associations
621 between bacteriophage, *Wolbachia*, and arthropods. PLoS Path 2: e43. doi:
622 10.1371/journal.ppat.0020043
- 623 Bordenstein SR, O'hara FP, Werren JH 2001. *Wolbachia*-induced incompatibility precedes other
624 hybrid incompatibilities in *Nasonia*. Nature 409: 707-710.
- 625 Bourtzis K, et al. 2014. Harnessing mosquito-*Wolbachia* symbiosis for vector and disease
626 control. Acta Trop 132, Supplement: S150-S163. doi: 10.1016/j.actatropica.2013.11.004
- 627 Brennan L, Haukedal J, Earle J, Keddie B, Harris H 2012. Disruption of redox homeostasis leads
628 to oxidative DNA damage in spermatocytes of *Wolbachia* - infected *Drosophila simulans*. Insect
629 Mol Biol 21: 510-520.
- 630 Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR 2016. Phylosymbiosis:
631 Relationships and functional effects of microbial communities across host evolutionary history.
632 PLoS Biol 14: e2000225.

- 633 Brucker RM, Bordenstein SR 2013. The hologenomic basis of speciation: gut bacteria cause
634 hybrid lethality in the genus *Nasonia*. *Science* 341: 667-669.
- 635 Caporaso JG, et al. 2010. QIIME allows analysis of high-throughput community sequencing
636 data. *Nat Methods* 7: 335.
- 637 Capra JA, Singh M 2007. Predicting functionally important residues from sequence
638 conservation. *Bioinformatics* 23: 1875-1882.
- 639 Cattel J, et al. 2016. *Wolbachia* in European populations of the invasive pest *Drosophila suzukii*:
640 regional variation in infection frequencies. *PLoS One* 11: e0147766.
- 641 Chelikani P, Fita I, Loewen PC 2004. Diversity of structures and properties among catalases.
642 *Cell Mol Life Sci* 61: 192-208.
- 643 Conner WR, et al. 2017. Genome comparisons indicate recent transfer of wRi-like *Wolbachia*
644 between sister species *Drosophila suzukii* And *D. subpulchrella*. bioRxiv: 135475.
- 645 Dedeine F, et al. 2001. Removing symbiotic *Wolbachia* bacteria specifically inhibits oogenesis
646 in a parasitic wasp. *Proc Natl Acad Sci* 98: 6247-6252. doi: 10.1073/pnas.101304298
- 647 Duron O, Fort P, Weill M 2006. Hypervariable prophage WO sequences describe an unexpected
648 high number of *Wolbachia* variants in the mosquito *Culex pipiens*. *Proc R Soc Lond B* 273: 495-
649 502.
- 650 Eddy SR 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7: e1002195.
- 651 Edgar RC 2004. MUSCLE: multiple sequence alignment with high accuracy and high
652 throughput. *Nucleic Acids Res* 32: 1792-1797.

- 653 Ellegaard KM, Klasson L, Naslund K, Bourtzis K, Andersson SGE 2013. Comparative genomics
654 of *Wolbachia* and the bacterial species concept. PLoS Genet 9: e1003381. doi:
655 10.1371/journal.pgen.1003381
- 656 Felsenstein J 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics 5: 6.
- 657 Fortino V, Smolander O-P, Auvinen P, Tagliaferri R, Greco D 2014. Transcriptome dynamics-
658 based operon prediction in prokaryotes. BMC Bioinformatics 15: 145.
- 659 Gavotte L, et al. 2007. A survey of the bacteriophage WO in the endosymbiotic bacteria
660 *Wolbachia*. Mol Biol Evol 24: 427-435. doi: 10.1093/molbev/msl171
- 661 Gerth M, Bleidorn C 2016. Comparative genomics provides a timeframe for *Wolbachia*
662 evolution and exposes a recent biotin synthesis operon transfer. Nat Micro 2: 16241.
- 663 Gottlieb Y, Zchori-Fein E, Werren JH, Karr TL 2002. Diploidy restoration in *Wolbachia*-
664 infected *Muscidifurax uniraptor* (Hymenoptera: Pteromalidae). J Invertebr Pathol 81: 166-174.
- 665 Gutzwiller F, et al. 2015. Dynamics of *Wolbachia pipientis* gene expression across the
666 *Drosophila melanogaster* life cycle. G3: Genes| Genomes| Genetics 5: 2843-2856.
- 667 Hamm CA, et al. 2014. *Wolbachia* do not live by reproductive manipulation alone: infection
668 polymorphism in *Drosophila suzukii* and *D. subpulchrella*. Mol Ecol 23: 4871-4885.
- 669 Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH 2008. How many
670 species are infected with *Wolbachia*? - A statistical analysis of current data. FEMS Microbiol
671 Lett 281: 215-220. doi: 10.1111/j.1574-6968.2008.01110.x

- 672 Hoerauf A, et al. 1999. Tetracycline therapy targets intracellular bacteria in the filarial nematode
673 *Litomosoides sigmodontis* and results in filarial infertility. J Clin Invest 103: 11-18.
- 674 Hoffmann AA, et al. 2011. Successful establishment of *Wolbachia* in *Aedes* populations to
675 suppress dengue transmission. Nature 476: 454-U107. doi: 10.1038/nature10356
- 676 Hornett EA, et al. 2008. You can't keep a good parasite down: evolution of a male-killer
677 suppressor uncovers cytoplasmic incompatibility. Evolution 62: 1258-1263. doi: 10.1111/j.1558-
678 5646.2008.00353.x
- 679 Hosokawa T, Koga R, Kikuchi Y, Meng XY, Fukatsu T 2010. *Wolbachia* as a bacteriocyte-
680 associated nutritional mutualist. Proc Natl Acad Sci 107: 769-774. doi:
681 10.1073/pnas.0911476107
- 682 <http://python.org>. Python Language Reference, version 2.7.
- 683 Hughes GL, Koga R, Xue P, Fukatsu T, Rasgon JL 2011. *Wolbachia* infections are virulent and
684 inhibit the human malaria parasite *Plasmodium falciparum* in *Anopheles Gambiae*. PLoS Path 7:
685 e1002043. doi: 10.1371/journal.ppat.1002043
- 686 Hurst GD, et al. 1999. Male-killing *Wolbachia* in two species of insect. Proc R Soc Lond B 266:
687 735-740.
- 688 Hurvich CM, Tsai CL 1993. A corrected Akaike information criterion for vector autoregressive
689 model selection. Journal of Time Series Analysis 14: 271-279.
- 690 Ishmael N, et al. 2009. Extensive genomic diversity of closely related *Wolbachia* strains.
691 Microbiology 155: 2211-2222. doi: 10.1099/mic.0.027581-0

- 692 Jaenike J, Dyer KA, Cornish C, Minhas MS 2006. Asymmetrical reinforcement and *Wolbachia*
693 infection in *Drosophila*. PLoS Biol 4: e325.
- 694 Kambris Z, Cook PE, Phuc HK, Sinkins SP 2009. Immune activation by life-shortening
695 *Wolbachia* and reduced filarial competence in mosquitoes. Science 326: 134-136. doi:
696 10.1126/science.1177531
- 697 Katoh K, Standley DM 2013. MAFFT multiple sequence alignment software version 7:
698 improvements in performance and usability. Mol Biol Evol 30: 772-780. doi:
699 10.1093/molbev/mst010
- 700 Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform
701 for the organization and analysis of sequence data. Bioinformatics 28: 1647-1649.
- 702 Kent BN, Funkhouser LJ, Setia S, Bordenstein SR 2011a. Evolutionary genomics of a temperate
703 bacteriophage in an obligate intracellular bacteria (*Wolbachia*). PLoS One 6: e24984.
- 704 Kent BN, et al. 2011b. Complete bacteriophage transfer in a bacterial endosymbiont (*Wolbachia*)
705 determined by targeted genome capture. Genome Biol Evol 3: 209-218. doi: 10.1093/gbe/evr007
- 706 Kosinski J, Feder M, Bujnicki JM 2005. The PD-(D/E) XK superfamily revisited: identification
707 of new members among proteins involved in DNA metabolism and functional predictions for
708 domains of (hitherto) unknown function. BMC Bioinformatics 6: 172.
- 709 Kramer RW, et al. 2007. Yeast functional genomic screens lead to identification of a role for a
710 bacterial effector in innate immunity regulation. PLoS Pathog 3: e21.

- 711 Landmann F, Orsi GA, Loppin B, Sullivan W 2009. *Wolbachia*-mediated Cytoplasmic
712 Incompatibility is associated with impaired histone deposition in the male pronucleus. PLoS Path
713 5. doi: 10.1371/journal.ppat.1000343
- 714 Lassy CW, Karr TL 1996. Cytological analysis of fertilization and early embryonic development
715 in incompatible crosses of *Drosophila simulans*. *Mechanisms of Development* 57: 47-58.
- 716 LePage DP, Jernigan KK, Bordenstein SR 2014. The relative importance of DNA methylation
717 and Dnmt2-mediated epigenetic regulation on *Wolbachia* densities and cytoplasmic
718 incompatibility. *PeerJ* 2: e678. doi: 10.7717/peerj.678
- 719 LePage DP, et al. 2017. Prophage WO genes recapitulate and enhance *Wolbachia*-induced
720 cytoplasmic incompatibility. *Nature* 543: 243-247. doi: 10.1038/nature21391
- 721 Li B, Dewey CN 2011. RSEM: accurate transcript quantification from RNA-Seq data with or
722 without a reference genome. *BMC Bioinformatics* 12: 323.
- 723 Lindsey ARI, Werren JH, Richards S, Stouthamer R 2016. Comparative genomics of a
724 parthenogenesis-inducing *Wolbachia* symbiont. *G3: Genes|Genomes|Genetics*. doi:
725 10.1534/g3.116.028449
- 726 Lo N, Casiraghi M, Salati E, Bazzocchi C, Bandi C 2002. How many *Wolbachia* supergroups
727 exist? *Mol Biol Evol* 19: 341-346.
- 728 Mantel N 1967. The detection of disease clustering and a generalized regression approach.
729 *Cancer Res* 27: 209-220.

- 730 Masui S, Kamoda S, Sasaki T, Ishikawa H 2000. Distribution and evolution of bacteriophage
731 WO in *Wolbachia*, the endosymbiont causing sexual alterations in arthropods. J Mol Evol 51:
732 491-497.
- 733 McClure R, et al. 2013. Computational analysis of bacterial RNA-Seq data. Nucleic Acids Res
734 41: e140-e140.
- 735 Miller WJ, Ehrman L, Schneider D 2010. Infectious speciation revisited: impact of symbiont-
736 depletion on female fitness and mating behavior of *Drosophila paulistorum*. PLoS Pathog 6:
737 e1001214.
- 738 Min K-T, Benzer S 1997. *Wolbachia*, normally a symbiont of *Drosophila*, can be virulent,
739 causing degeneration and early death. Proc Natl Acad Sci 94: 10792-10796. doi:
740 10.1073/pnas.94.20.10792
- 741 Moreau J, Bertin A, Caubet Y, Rigaud T 2001. Sexual selection in an isopod with *Wolbachia* -
742 induced sex reversal: males prefer real females. J Evol Biol 14: 388-394.
- 743 Moreira LA, et al. 2009. A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue,
744 chikungunya, and *Plasmodium*. Cell 139: 1268-1278. doi: 10.1016/j.cell.2009.11.042
- 745 Negri H, et al. 2009. Unravelling the *Wolbachia* evolutionary role: the reprogramming of the
746 host genomic imprinting. Proc R Soc Lond B 276: 2485-2491. doi: 10.1098/rspb.2009.0324
- 747 Newton IL, et al. 2016. Comparative genomics of two closely related *Wolbachia* with different
748 reproductive effects on hosts. Genome Biol Evol 8: 1526-1542. doi: 10.1093/gbe/evw096

- 749 Noda H, Koizumi Y, Zhang Q, Deng K 2001. Infection density of *Wolbachia* and incompatibility
750 level in two planthopper species, *Laodelphax striatellus* and *Sogatella furcifera*. Insect Biochem
751 Mol Biol 31: 727-737.
- 752 O'Neill SL, Giordano R, Colbert AME, Karr TL, Robertson HM 1992. 16S ribosomal-RNA
753 phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic
754 incompatibility in insects. Proc Natl Acad Sci 89: 2699-2702. doi: 10.1073/pnas.89.7.2699
- 755 Pantoja M, Chen L, Chen Y, Nester EW 2002. *Agrobacterium* type IV secretion is a two - step
756 process in which export substrates associate with the virulence protein VirJ in the periplasm. Mol
757 Microbiol 45: 1325-1335.
- 758 Paquette N, et al. 2012. Serine/threonine acetylation of TGF β -activated kinase (TAK1) by
759 *Yersinia pestis* YopJ inhibits innate immune signaling. Proc Natl Acad Sci 109: 12710-12715.
- 760 Pérez F, Granger BE 2007. IPython: a system for interactive scientific computing. Computing in
761 Science & Engineering 9.
- 762 Randerson JP, Jiggins FM, Hurst LD 2000. Male killing can select for male mate choice: a novel
763 solution to the paradox of the lek. Proc R Soc Lond B 267: 867-874.
- 764 Rice DW, Sheehan KB, Newton IL 2017. Large-scale identification of *Wolbachia pipientis*
765 effectors. Genome Biol Evol 9: 1925-1937.
- 766 Robinson DF, Foulds LR 1981. Comparison of phylogenetic trees. Math Biosci 53: 131-147.
- 767 Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model
768 choice across a large model space. Syst Biol 61: 539-542.

- 769 Ros VID, Fleming VM, Feil EJ, Breeuwer JAJ 2009. How diverse is the genus *Wolbachia*?
770 Multiple-gene sequencing reveals a putatively new *Wolbachia* supergroup recovered from spider
771 mites (Acari: Tetranychidae). Appl Environ Microbiol 75: 1036-1043. doi: 10.1128/aem.01109-
772 08
- 773 Rousset F, Bouchon D, Pintureau B, Juchault P, Solignac M 1992. *Wolbachia* endosymbionts
774 responsible for various alterations of sexuality in arthropods. Proc R Soc Lond B 250: 91-98.
- 775 Schilthuizen M, Stouthamer R 1997. Horizontal transmission of parthenogenesis-inducing
776 microbes in *Trichogramma* wasps. Proc R Soc Lond B 264: 361-366.
- 777 scikit-bio.org. scikit-bio: core bioinformatics data structures and algorithms in Python.
778 caporasolab.us: Northern Arizona University.
- 779 Serbus LR, Casper-Lindley C, Landmann F, Sullivan W 2008. The genetics and cell biology of
780 *Wolbachia*-host interactions. Annu Rev Genet 42: 683-707. doi:
781 10.1146/annurev.genet.41.110306.130354
- 782 Shropshire JD, Bordenstein SR 2016. Speciation by symbiosis: the microbiome and behavior.
783 MBio 7. doi: 10.1128/mBio.01785-15
- 784 Siggers KA, Lesser CF 2008. The Yeast *Saccharomyces cerevisiae*: a versatile model system for
785 the identification and characterization of bacterial virulence proteins. Cell host & microbe 4: 8-
786 15.
- 787 Sinkins SP, et al. 2005. *Wolbachia* variability and host effects on crossing type in *Culex*
788 mosquitoes. Nature 436: 257-260.

- 789 Söding J, Biegert A, Lupas AN 2005. The HHpred interactive server for protein homology
790 detection and structure prediction. *Nucleic Acids Res* 33: W244-W248.
- 791 Stamatakis A 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of
792 large phylogenies. *Bioinformatics*. doi: 10.1093/bioinformatics/btu033
- 793 Stouthamer R, Breeuwer JAJ, Luck RF, Werren JH 1993. Molecular-identification of
794 microorganisms associated with parthenogenesis. *Nature* 361: 66-68. doi: 10.1038/361066a0
- 795 Stouthamer R, Kazmer DJ 1994. Cytogenetics of microbe-associated parthenogenesis and its
796 consequences for gene flow in *Trichogramma* wasps. *Heredity* 73: 317-327. doi:
797 10.1038/hdy.1994.139
- 798 Stouthamer R, Luck R 1993. Influence of microbe-associated parthenogenesis on the fecundity
799 of *Trichogramma deion* and *T. pretiosum*. *Entomol Exp Appl* 67: 183-192. doi: 10.1111/j.1570-
800 7458.1993.tb01667.x
- 801 Stouthamer R, Luck RF, Hamilton WD 1990. Antibiotics cause parthenogenetic *Trichogramma*
802 (Hymenoptera, Trichogrammatidae) to revert to sex. *Proc Natl Acad Sci* 87: 2424-2427. doi:
803 10.1073/pnas.87.7.2424
- 804 Sutton E, Harris S, Parkhill J, Sinkins S 2014. Comparative genome analysis of *Wolbachia* strain
805 *w*Au. *BMC Genomics* 15: 928.
- 806 Teixeira L, Ferreira Á, Ashburner M 2008. The bacterial symbiont *Wolbachia* induces resistance
807 to RNA viral infections in *Drosophila melanogaster*. *PLoS Biol* 6: e1000002.

- 808 Tram U, Sullivan W 2002. Role of delayed nuclear envelope breakdown and mitosis in
809 *Wolbachia*-induced cytoplasmic incompatibility. *Science* 296: 1124-1126.
- 810 Turelli M, Hoffmann AA 1991. Rapid spread of an inherited incompatibility factor in California
811 *Drosophila*. *Nature* 353: 440-442.
- 812 Walker T, et al. 2011. The wMel *Wolbachia* strain blocks dengue and invades caged *Aedes*
813 *aegypti* populations. *Nature* 476: 450-U101. doi: 10.1038/nature10355
- 814 Wallau GL, da Rosa MT, De Re FC, Loreto EL 2016. *Wolbachia* from *Drosophila incompta*:
815 just a hitchhiker shared by *Drosophila* in the New and Old World? *Insect Mol Biol* 25: 487-499.
816 doi: 10.1111/imb.12237
- 817 Werren JH, Baldo L, Clark ME 2008. *Wolbachia*: master manipulators of invertebrate biology.
818 *Nat Rev Micro* 6: 741-751. doi: 10.1038/nrmicro1969
- 819 Werren JH, Windsor DM 2000. *Wolbachia* infection frequencies in insects: Evidence of a global
820 equilibrium? *Proc R Soc Lond B* 267: 1277-1285. doi: 10.1098/rspb.2000.1139
- 821 Werren JH, Zhang W, Guo LR 1995. Evolution and phylogeny of *Wolbachia* - reproductive
822 parasites of arthropods. *Proc R Soc Lond B* 261: 55-63. doi: 10.1098/rspb.1995.0117
- 823 Wu M, et al. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A
824 streamlined genome overrun by mobile genetic elements. *PLoS Biol* 2: 327-341. doi:
825 10.1371/journal.pbio.0020069

- 826 Ye YXH, et al. 2013. Infection with a virulent strain of *Wolbachia* disrupts genome wide-
827 patterns of cytosine methylation in the mosquito *Aedes aegypti*. PLoS One 8: e66482. doi:
828 10.1371/journal.pone.0066482
- 829 Yen JH, Barr AR 1971. New hypothesis of the cause of cytoplasmic incompatibility in *Culex*
830 *pipiens*. Nature.
- 831 Zabalou S, et al. 2004. *Wolbachia*-induced cytoplasmic incompatibility as a means for insect pest
832 population control. Proc Natl Acad Sci 101: 15042-15045. doi: 10.1073/pnas.0403853101
- 833 Zug R, Hammerstein P 2012. Still a host of hosts for *Wolbachia*: Analysis of recent data suggests
834 that 40% of terrestrial arthropod species are infected. PLoS One 7. doi:
835 10.1371/journal.pone.0038544
- 836
- 837

838 **Tables**

839 **Table 1. Strains used in comparative analyses of *cifA* and *cifB*.**

Supergroup	Strain	Host	Reproductive Phenotypes ^a	Accession Number
A	wMel	<i>Drosophila melanogaster</i>	CI	NC_002978.6
	wMelPop	<i>Drosophila melanogaster</i>	CI	AQQE00000000.1
	wRec	<i>Drosophila recens</i>	CI	NZ_JQAM00000000.1
	wAu	<i>Drosophila simulans</i>	None	LK055284.1
	wHa	<i>Drosophila simulans</i>	CI	NC_021089.1
	wRi	<i>Drosophila simulans</i>	CI	NC_012416.1
	wSuzi	<i>Drosophila suzukii</i>	None	NZ_CAOU00000000.2
	wUni	<i>Muscidifurax uniraptor</i>	PI	NZ_ACFP00000000.1
wVitA	<i>Nasonia vitripennis</i>	CI	NZ_MUJM00000000.1	
B	wAlbB	<i>Aedes albopictus</i>	CI	CAGB00000000.1
	wNo	<i>Drosophila simulans</i>	CI	NC_021084.1
	wDi	<i>Diaphorina citri</i>	Undetermined	NZ_KB223540.1
	wTpre	<i>Trichogramma pretiosum</i>	PI	CM003641.1
	wVitB	<i>Nasonia vitripennis</i>	CI	AERW00000000.1
	wBol1-b	<i>Hypolimnas bolina</i>	CI, MK	NZ_CAOH00000000.1
	wPipJHB	<i>Culex quinquefasciatus</i>	CI	ABZA00000000.1
wPipPel	<i>Culex pipiens</i>	CI	NC_010981.1	
C	wOo	<i>Onchocerca ochengi</i>	OM	NC_018267.1
	wOv	<i>Onchocerca volvulus</i>	OM	NZ_HG810405.1
D	wBm	<i>Brugia malayi</i>	OM	NC_006833.1
F	wCle	<i>Cimex lectularius</i>	OM	NZ_AP013028.1

840 ^aReproductive phenotypes include: CI) cytoplasmic incompatibility, PI) parthenogenesis-
 841 inducing, MK) male-killing, OM) obligate mutualism, None) no phenotype discovered after
 842 assessment, and Undetermined) phenotype was not assayed.

843

844 **Table 2. Predicted structural modules of Cif proteins.**

Protein	Module ^a	Size Range (AA)	Homology
CifA	ModA-1 ●	24-55	<ul style="list-style-type: none"> Catalase-rel, decomposes hydrogen peroxide into water and oxygen VirJ, bacterial virulence protein and component of T4SS secretion systems
	ModA-2 ●	99-152	<ul style="list-style-type: none"> DUF3243, DUF603, domains of unknown function CARD_MDA5_r1, Caspase activation and recruitment domain Ectatomin, toxic component of ant venom Ldr, type I toxin-antitoxin system IstR, lineage-specific thermal regulator protein
	ModA-3 ●	47-68	<ul style="list-style-type: none"> STE, STE-like transcription factor
CifB	ModB-1 ●	97-127	<ul style="list-style-type: none"> PDDEXK, PD-(D/E)XK nuclease superfamily DUF91, Domain of Unknown Function HSDR_N, type I restriction enzyme R protein N terminus
	ModB-2 ●	122-155	<ul style="list-style-type: none"> PDDEXK, PD-(D/E)XK nuclease superfamily DUF1052, domain of unknown function HSDR_N, type I restriction enzyme R protein N terminus
	ModB-3 ^b ○	277	<ul style="list-style-type: none"> RE_XamI, XamI restriction endonuclease
			<ul style="list-style-type: none"> Ulp-1, ubiquitin-like proteases SUMO, small ubiquitin-related modifier YopJ, Serine/Threonine acetyltransferase (<i>Yersinia</i>) Ssel, deubiquitylase SidE, Dot/Icm substrate protein

845 ^aColors next to modules are used throughout the text

846 ^bOnly present in Type I

847

848

849 **Table 3. Average amino acid conservation of Cifs and modules.**

Protein	Region^a	Type I	All
CifA	ModA-1	0.94	0.70
	ModA-2	0.82	0.55
	ModA-3	0.77	0.53
	CifA	0.83	0.58
CifB	ModB-1	0.89	0.71
	ModB-2	0.85	0.62
	ModB-3 ^b	0.77	0.39
	CifB	0.82	0.43

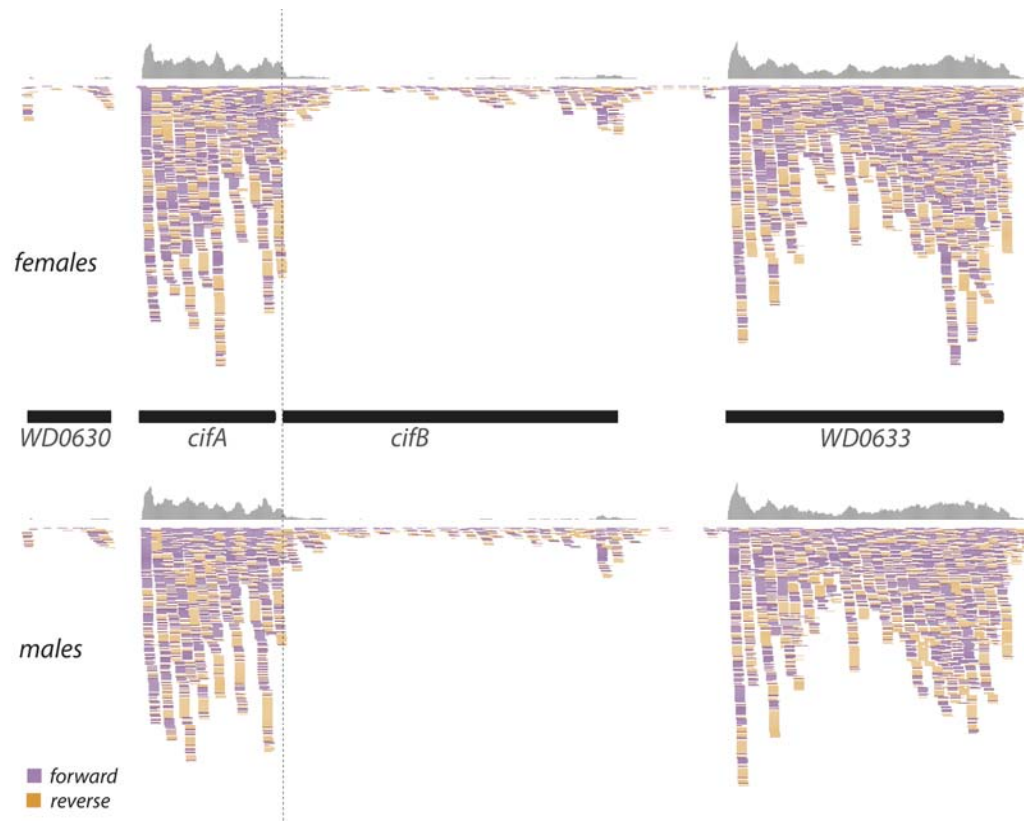
850 ^aModule number defined in Table 2

851 ^bOnly present in Type I

852

853 **Figures**

854 **Figure 1**



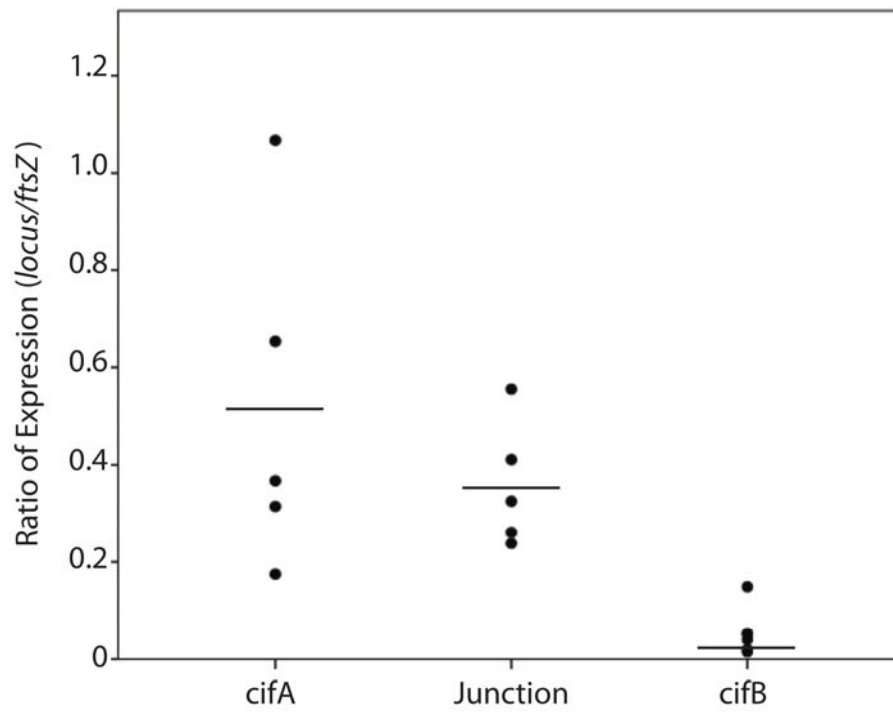
855
856

857 **Figure 2**

(A)

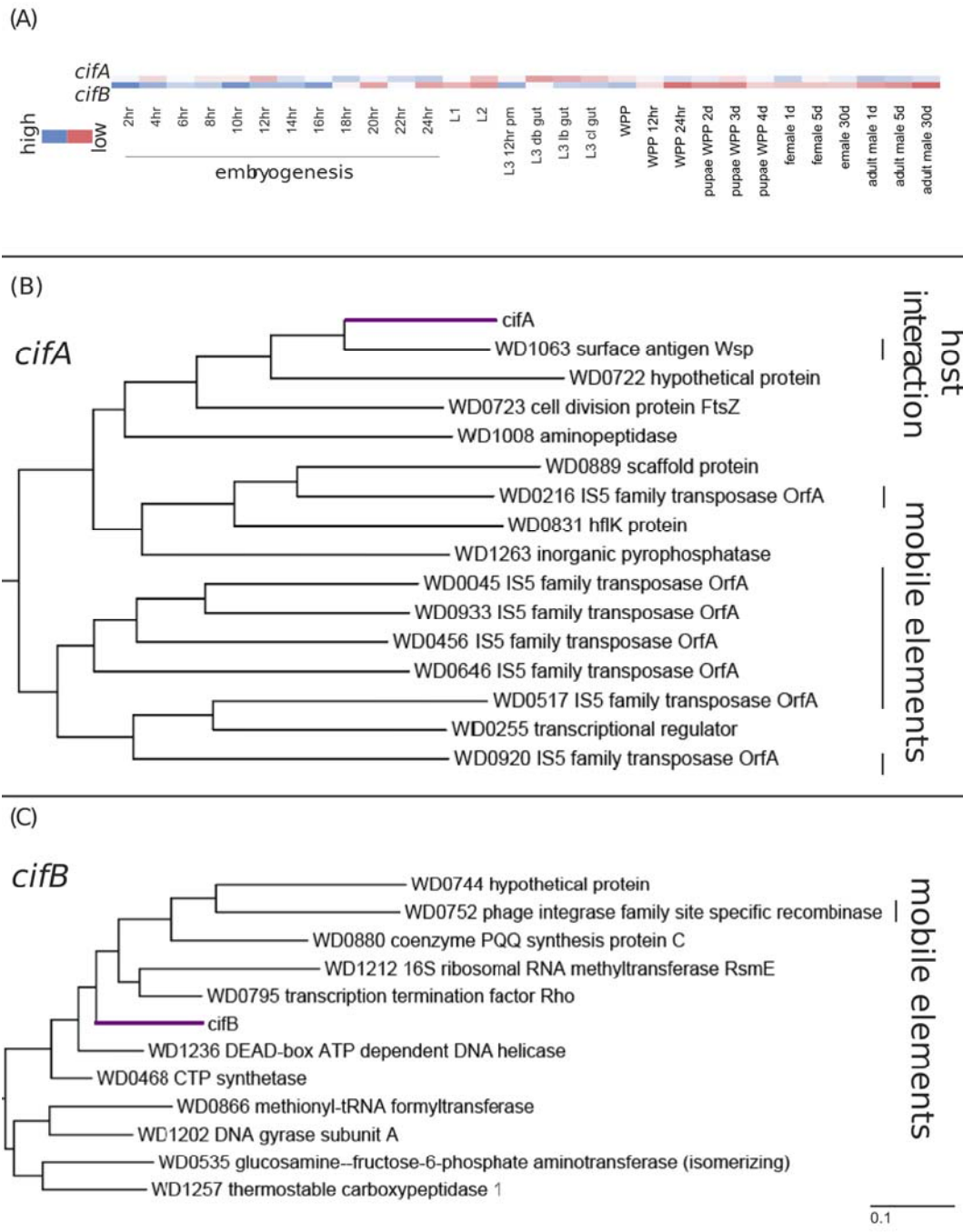


(B)



858
859

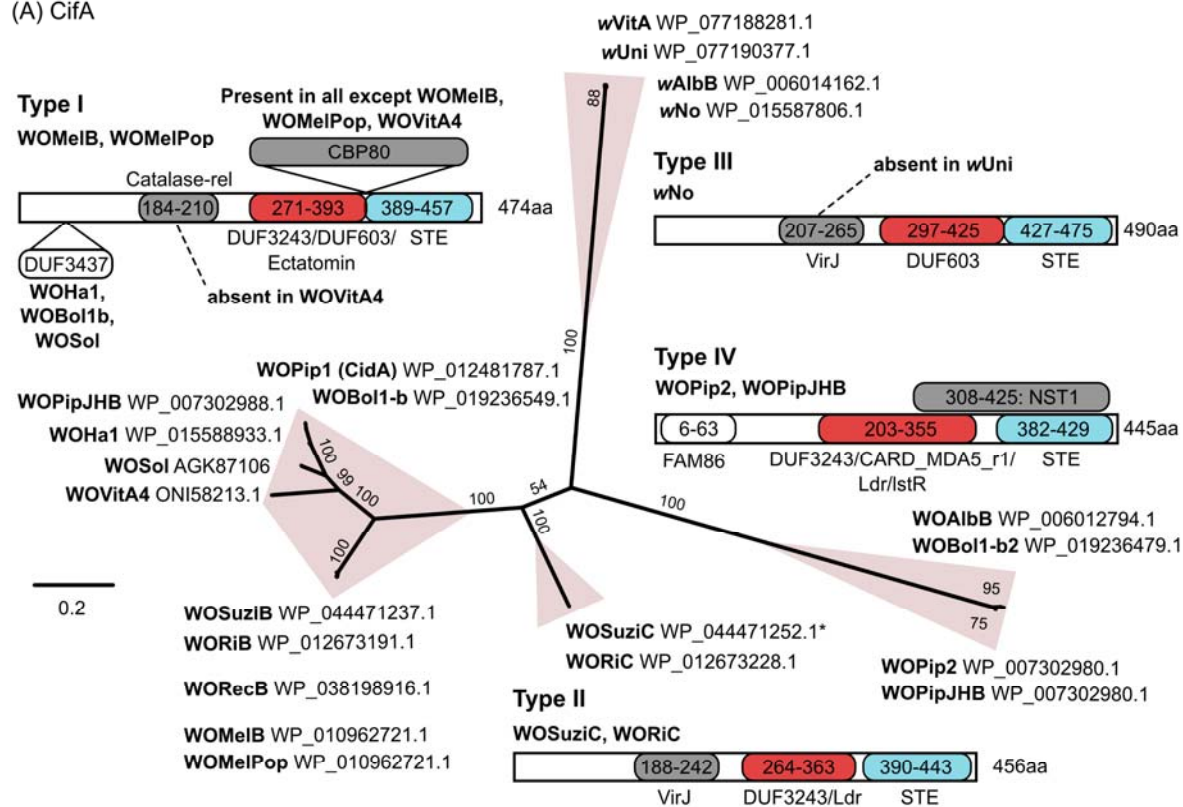
860 **Figure 3**



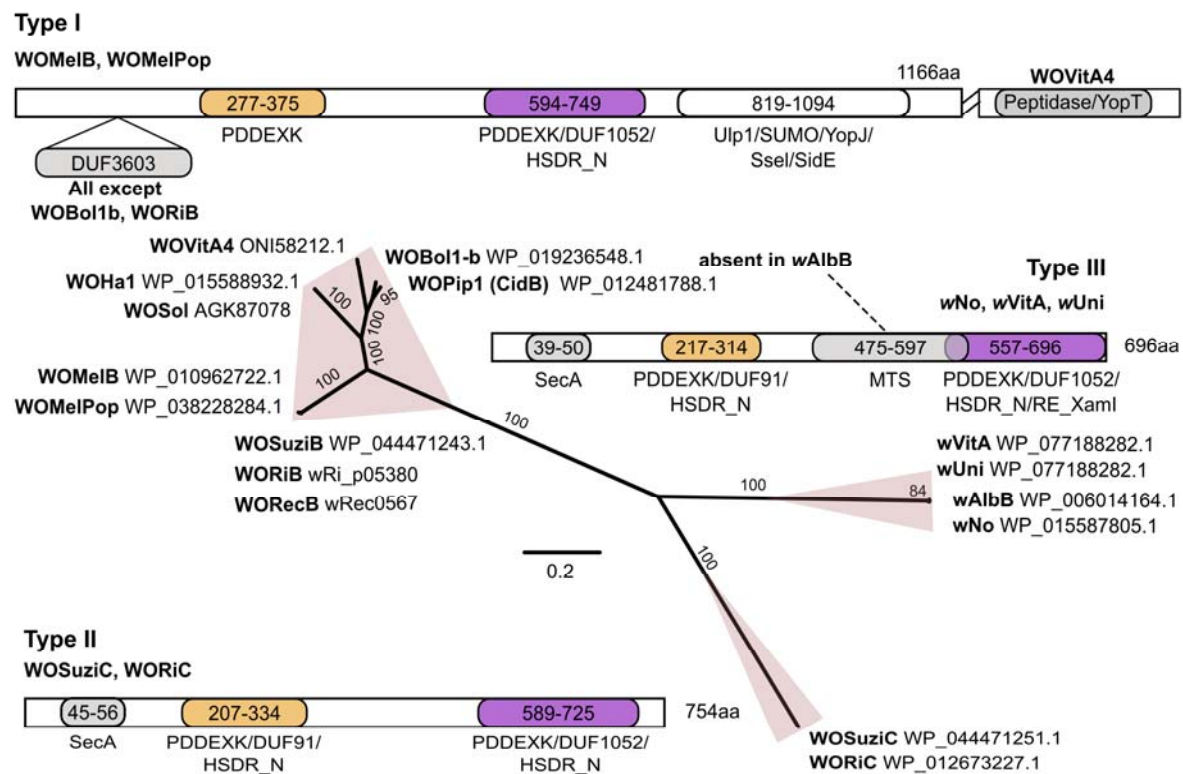
861
862

863 **Figure 4**

(A) CifA



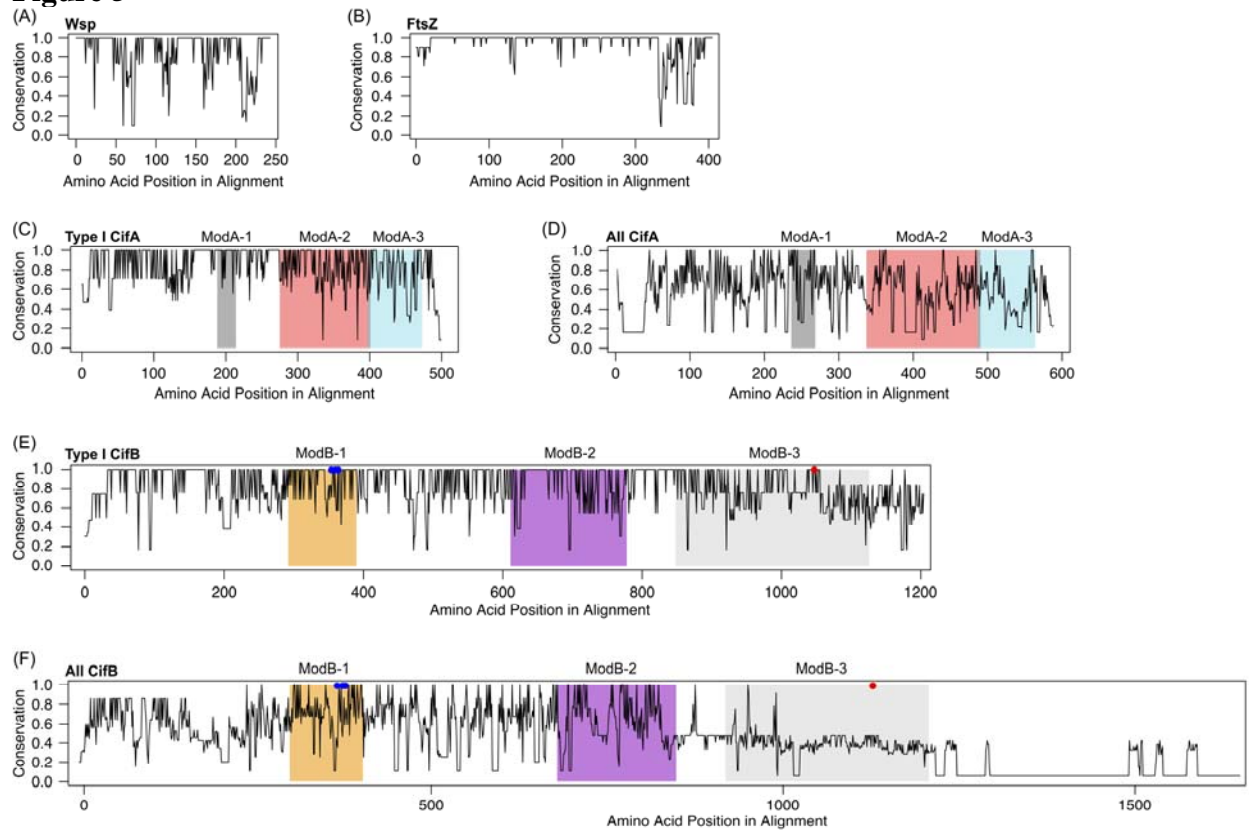
(B) CifB



864

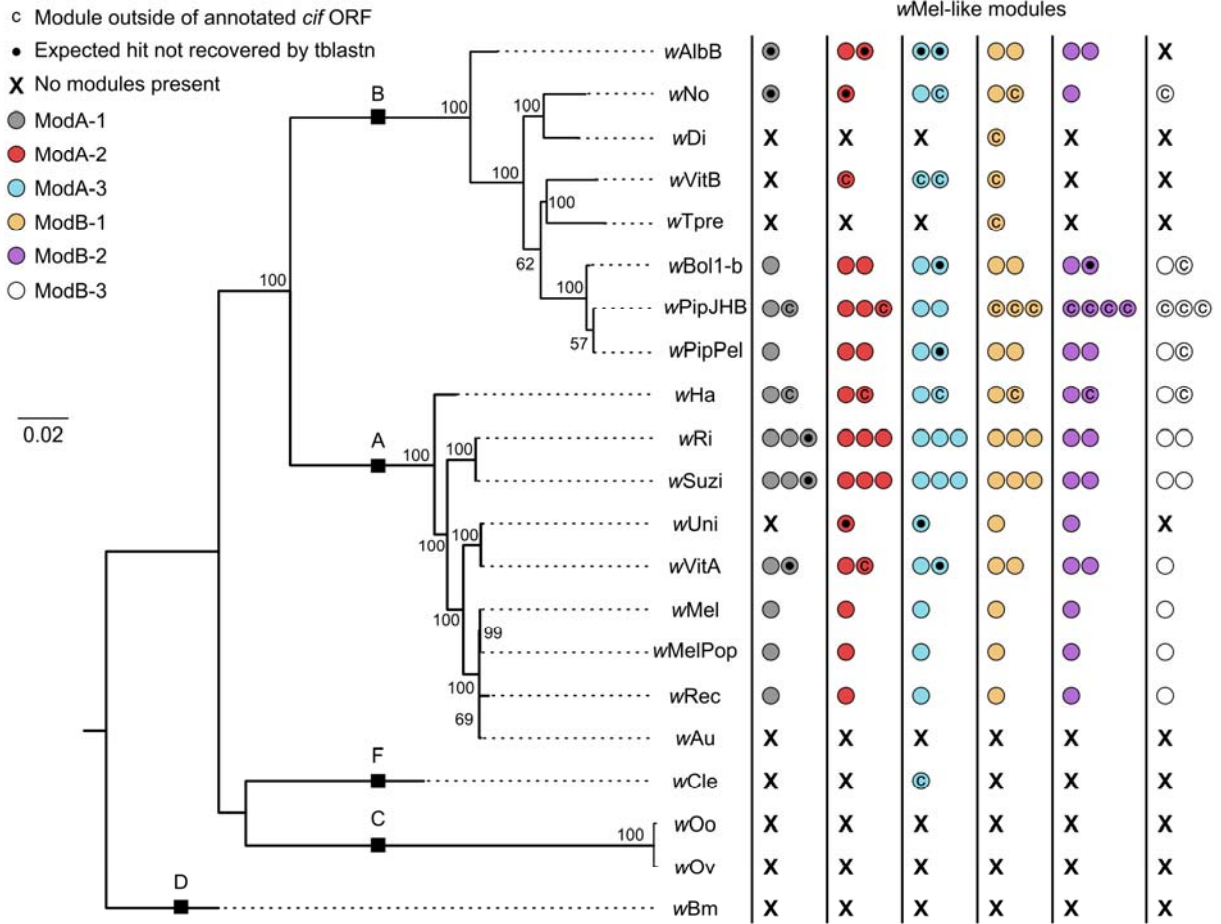
865

Figure 5



866
867

868 **Figure 6**



869

870 **Figure legends**

871 **Fig 1. RNASeq analysis of *cifA* and *cifB* gene expression in whole adult, 1 day old female**
872 **and male *Drosophila melanogaster* flies.** Raw reads were mapped to the *wMel* assembly (using
873 *bwa*) and coverage visualized using the Integrated Genomics Viewer (v2.3.77). The start of the
874 *cifB* open reading frame is denoted by a vertical, dotted line.

875
876 **Fig 2. Relative expression ratio of *cifA*, the junction between *cifA/cifB*, and *cifB* to *ftsZ*.**
877 Expression of both genes and their junction was quantified using qRT-PCR, and normalized to
878 *Wolbachia ftsZ* gene expression. *cifB* gene expression is significantly less than that of the
879 junction ($t= 3.220$, $df=16$, $p=0.005$) and less than *cifA* ($t=-3.840$, $df=17$, $p=0.001$).

880
881 **Fig 3. Gene expression of *cifA* and *cifB* during *Drosophila melanogaster* development. A)**
882 Heatmap representation of normalized transcripts per kilobase million (TPM) for both *cifA* and
883 *cifB* during *Drosophila melanogaster* development. *cifB* is highly expressed during
884 embryogenesis and downregulated after pupation while *cifA* is more highly expressed in adults
885 and pupae. Clustering of *Wolbachia* loci based on expression across fly development illustrates
886 correlated expression profiles between *wMel* loci and *cifA* (B) or *cifB* (C). Mobile elements and
887 loci involved in host interaction (*wsp*) are indicated with vertical lines on the right side of the
888 figure.

889
890 **Fig 4. Phylogenetic relationships and representative predicted protein structure of Cif**
891 **protein types. A) CifA and B) CifB.** Alleles are in bold next to their corresponding accession
892 number, and pink shapes around branches designate monophyletic “types”. Representative

893 structures are shown for each type, with the length of the protein indicated at the C-terminus.
894 Variations in within-type structure are shown. If an allele is not listed as a representative, and
895 significant structural variations are not indicated, then only the exact coordinates of the structural
896 regions differed by a few amino acids. All HHpred structural predictions are significant at a
897 corrected p-value of < 0.05 , and listed in order of ascending p-value for regions with multiple
898 structural hits. Allele names use the previously described naming convention with a WO prefix
899 referring to particular phage haplotype, and the *w* prefix indicating a phage-like island (LePage,
900 et al. 2017). The N-terminus of WOSuziC (*) was translated from the end of another contig and
901 concatenated to get the full-length protein (see methods). WOMelB and WOMelPop are identical
902 at the amino acid level, as are WOPipJHB and WOPip2.

903

904 **Fig 5. Protein conservation, as determined by Shannon entropy scores.** A) Wsp (*Wolbachia*
905 surface protein), B) Cell division protein FtsZ, C) Type I CifA, D) All CifA, E) Type I CifB
906 alleles except for WOVitA4, F) All CifB alleles. Red dots in E and F indicate the ModB-3
907 catalytic residue (Beckmann, et al. 2017), unique to and completely conserved for Type I alleles.
908 Blue dots in E and F represent the (P)D-(D/E)XK motif (Kosinski, et al. 2005) present in *w*Mel.
909 We found no (P)D-(D/E)XK putative catalytic motif in the second PDDEXK-like module of
910 CifB.

911

912 **Fig 6. Presence of *w*Mel-like Cif modules across the *Wolbachia* phylogeny.** The WOMelB
913 module sequences were used to query available *Wolbachia* genomes to look for the presence of
914 Cif-like regions beyond those within the annotated Cifs (Figure 4). Colored dots correspond to
915 the structural regions delimited by HHpred, shown in Figure 4, and listed in Table 2. A "C"

916 within a dot indicates the presence of a module outside of annotated *cif* open reading frames
917 (Figure 4 and Supplemental Figures S2 and S3). The black dot indicates a module annotated by
918 HHpred, but not identified by tblastn due to divergence from the WOMelB module. Black boxes
919 labeled with uppercase letters indicate branches leading to *Wolbachia* Supergroups. Dotted lines
920 on the phylogeny lead to taxon names and are not included in the branch length.