Current Opinion

A critical re-evaluation of multilocus sequence typing (MLST) efforts in *Wolbachia*

Christoph Bleidorn[1,2*] & Michael Gerth[3]

[1]Museo Nacional de Ciencias Naturales, Spanish National Research Council (CSIC), Madrid, Spain.

[2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany.

[3]Institute for Integrative Biology, University of Liverpool, Liverpool, UK.

[*]Correspondence to:

Christoph Bleidorn

Museo Nacional de Ciencias Naturales

Spanish National Research Council (CSIC)

Department of Biodiversity and Evolutionary Biology

José Gutiérrez Abascal 2

Madrid 28006

Spain

Email: christoph.bleidorn@gmail.com

Keywords: phylogenetic informativeness, phylogeny, MLST, recombination, strain typing, *Wolbachia*

Abstract

Wolbachia (Alphaproteobacteria, Rickettsiales) is the most common, and arguably one of the most important inherited symbionts. Molecular differentiation of *Wolbachia* strains is routinely performed with a set of five multilocus sequence typing (MLST) markers. However, since its inception in 2006, the performance of MLST in *Wolbachia* strain typing has not been assessed objectively. Here, we evaluate the properties of *Wolbachia* MLST markers and compare it to 252 other single copy loci present in the genome of most *Wolbachia* strains. Specifically, we investigated how well MLST performs at strain differentiation, at reflecting genetic diversity of strains, and as phylogenetic marker. We find that MLST loci are outperformed by other loci at all tasks they are currently employed for, and thus that they do not reflect the properties of a *Wolbachia* strain very well. We argue that whole genome typing approaches should be used for *Wolbachia* typing in the future. Alternatively, if few-loci-approaches are necessary, we provide a characterization of 252 single copy loci for a number a criteria, which may assist in designing specific typing systems or phylogenetic studies.

41 **Introduction**

42 *Wolbachia* is a genus of maternally inherited intracellular Alphaproteobacteria
43 that is found in arthropod and nematode hosts (Werren et al. 2008). Meta-analyses
44 suggest that between 40% and 52% of all terrestrial arthropods are infected, making
45 these bacteria the most common animal endosymbiont on earth (Zug & Hammerstein
46 2012; Weinert et al. 2015). Host specificity and type of symbiosis differs between major
47 lineages of *Wolbachia*, which are currently classified into 16 supergroups named with
48 capital letters from A–F and H–Q, consecutively in the order of their description
49 (Glowska et al. 2015; Gerth 2016). Supergroups A and B are found in arthropods,
50 representing the vast majority of described *Wolbachia* lineages. Many different types of
51 symbioses, including reproductive parasitism, facultative mutualism, and obligate
52 mutualism have been found for these lineages (Zug & Hammerstein 2015). In contrast,
53 supergroups C and D are restricted to filarial nematodes, with which they share a close
54 relationship that can be described as obligate mutualism (Makepeace & Gill 2016).
55 Supergroup F has been found in both nematodes and arthropods and all other
56 supergroups are rather rare, limited to a single or few hosts (Gerth et al. 2014).

57 Several host manipulations have been described for *Wolbachia,* and it is thought
58 that those accelerate their spread in host populations, such as male-killing, feminization,
59 induction of parthogenesis, and cytoplasmic incompatibility (Werren et al. 2008). These
60 manipulations are considered to have a predominantly negative effect on their hosts.
61 However, several positive aspects for hosts have been reported as well. These include
62 provision of the host with amino acids or vitamins, or protection against viruses
63 (Hedges et al. 2008; Teixeira et al. 2008; Zug & Hammerstein 2015). It appears likely
64 that positive fitness effects drive the establishment of novel *Wolbachia* infections in host
65 populations (Kriesner et al. 2013). Recently, field studies demonstrated that mosquito

66 populations can be artificially infected with fast spreading *Wolbachia* lineages which
67 confer virus resistance to their hosts, thereby suppressing the transmission of the human
68 pathogen Dengue (Hoffmann et al. 2011). However, not all strains of *Wolbachia* are able
69 to confer virus resistance or to manipulate their host's reproduction (Makepeace & Gill
70 2016).

71 The growing interest in the peculiar biology of *Wolbachia*, and its almost
72 universal distribution among arthropods have necessitated means to differentiate strains
73 by using molecular methods. Initially, genetic characterization of *Wolbachia* diversity
74 was based on the 16S rRNA gene (O'Neill et al. 1992) or the more variable *wsp* gene
75 (Zhou et al. 1998). However, in 2006, a multilocus sequence typing (MLST) system was
76 established, and this subsequently became a standard in the community of *Wolbachia*
77 researchers (Baldo et al. 2006).

78 The MLST approach was developed to provide a reproducible and portable
79 method for the molecular characterization of bacterial pathogens. Originally designed to
80 monitor local and global *Neisseria meningitides* outbreaks (Maiden et al. 1998), MLST
81 schemes have since been published for many other bacterial species (Maiden 2006). For
82 strain typing, five to ten loci (usually conserved housekeeping genes) from different
83 regions of the genome are sequenced and each unique allele is assigned a unique
84 number. Thus, a universal nomenclature based on a code of numbers referring to the
85 sequenced loci is assembled. MLST genes are selected under the assumption that they
86 underlie purifying selection, resulting in sequence variation that is mostly neutral. In the
87 absence of recombination, substitutions should accumulate approximately linearly with
88 time (Francisco et al. 2009) and therefore, genetic distances between strains at MLST
89 loci would be proportional to their divergence time. MLST data are usually provided in
90 a curated form in a freely accessible database (Jolley et al. 2004). Based on MLST

91 profiles, relationships between (or diversity of) typed strains can either be analysed
92 using the designated numbers from coding the alleles (i.e., MLST profiles), or by
93 analysing the allelic nucleotide sequence data directly.

94 For *Wolbachia* MLST, fragments of five housekeeping genes (*gatB*, *coxA*, *hcpA*,
95 *fbpA*, and *ftsZ*) are sequenced, and primers that amplify these loci across the major
96 *Wolbachia* supergroups in arthropods are available (Baldo et al. 2006). According to the
97 high number of citations for the original publication (Baldo et al. 2006, 343 citations in
98 ISI Web of Science accessed August 17th, 2017), the approach is well-established and
99 frequently used in the community of *Wolbachia* researchers. Since its original
100 description more than 10 years ago, 2355 sequences and 472 unique MLST profiles
101 have been added to the database (https://pubmlst.org/wolbachia/, accessed August 17th,
102 2017). When MLST was conceived, only two *Wolbachia* strains were represented by a
103 fully annotated genome, and therefore, it was not possible to test how well MLST
104 reflects the true *Wolbachia* strain diversity. Now, with a plethora of strains characterized
105 by MLST, and several complete or draft genomic sequences of *Wolbachia* strains
106 available (>30 strains in public repositories), the efficiency and performance of
107 *Wolbachia* MLST can be evaluated objectively.

108 In this article, we aim to do so by first identifying the most common tasks
109 *Wolbachia* MLST has been employed for by the research community. Using whole-
110 genome as well as MLST data, we next assess how well MLST performs in these tasks
111 in comparison to other single copy loci. We will argue that there is not a single locus or
112 a single set of loci that performs well in all questions that are commonly addressed by
113 *Wolbachia* researchers. Although the MLST scheme is convenient in that it provides a
114 readily employable set of molecular markers, its information content is critically
115 dependent on the research objective and the set of strains analysed. We therefore

116 advocate that molecular markers for *Wolbachia* should be chosen very carefully for each
117 particular research question, ideally based on whole-genome information.

118 **Usage of *Wolbachia* MLST in theory and research praxis**

119 Originally, MLST was aimed to provide "a reliable system for typing and
120 quantifying strain diversity" that allows "tracing the movement of *Wolbachia* globally
121 and within insect communities and for associating *Wolbachia* strains with geographic
122 regions, host features (e.g., ecology and phylogeny), and phenotypic effects on hosts"
123 (Baldo et al. 2006). In other words, ideally each *Wolbachia* strain in the MLST database
124 would not only be represented by a MLST profile, but also be linked with taxonomic
125 information about its host, geographic origin, and phenotypic effects. This would then
126 enable comparative analyses. However, out of 1828 strains ("isolates") currently listed
127 in the MLST database, only 603 (~34%) are associated with host taxonomy on the level
128 of host order, and even fewer are associated with a host species (542, ~30%). Similarly,
129 only 577 isolates (~31%) have geographic information and a phenotype is only known
130 from 92 strains (~5%). Thus, the majority of *Wolbachia* strains in the database are
131 defined by their MLST profiles alone, which further are in most cases incomplete
132 (~60% of strains lack one or more alleles). Although this likely impedes comparative
133 analyses, the lack of metadata associated with *Wolbachia* MLST isolates is not a
134 problem for strain definition as such. However, if MLST is the only definition for a
135 *Wolbachia* strain, it is crucial to understand how appropriate this definition is and to
136 ascertain that the MLST profile is not isolated from the biological properties of the
137 typed strains.

138 In current practise, it is generally assumed that MLST markers are a good
139 approximation of genome-wide characteristics of *Wolbachia* strains. As such, they have
140 been used to describe and analyse the *Wolbachia* diversity, phylogeny, or

141 phylogeography of particular host taxa (Russell et al. 2009; Watanabe et al. 2012;
142 Schuler et al. 2013; Zhang et al. 2013a; Sontowski et al. 2015), taxa from a particular
143 ecological background/community (Stahlhut et al. 2010; Zhang et al. 2013b), and to
144 explore horizontal movements of *Wolbachia* strains (Baldo et al. 2008; Gerth et al.
145 2013; Ahmed et al. 2016). All of these research questions entail a number of implicit
146 assumptions about the performance of *Wolbachia* MLST. We will in the following
147 examine three of these assumptions that we consider most important in this regard:
148 1) *Wolbachia* MLST can differentiate *Wolbachia* strains.
149 2) Genetic divergence at *Wolbachia* MLST genes corresponds to genome–wide
150 divergence levels
151 3) *Wolbachia* MLST gene phylogeny reflects the phylogeny of the core genome.

152 **Differentiating *Wolbachia* strains with MLST markers**

153 One common task for MLST in *Wolbachia* research is the discrimination (or
154 "quantification" as in Baldo et al. 2006) of *Wolbachia* strains, i.e., to answer if two (or
155 any other number) of strains are genetically different. For this task, the level of
156 resolution depends on the number and type of genes used, length of the sequences and
157 the genetic diversity of chosen loci (Cooper & Feil 2004). The limits of MLST schemes
158 were pointed out for genetically monomorphic bacteria such as *Mycobacterium*
159 *tuberculosis* or *Bacillus anthracis* (Achtman 2008; Achtman 2012). *Wolbachia* MLST
160 diversity within supergroups is far from being monomorphic, as evident from the large
161 number of available profiles in the database (see above). However, the actual
162 evolutionary pace of *Wolbachia* genes and genomes was an open question. Recently,
163 based on a time-calibrated phylogenomic analyses it was hypothesized that *Wolbachia*
164 lineages are much older than previously assumed – and therefore that genetic change
165 due to substitutions or recombination accumulate slower than expected (Gerth &

166 Bleidorn 2016). In accordance with this estimate, it was repeatedly reported that
167 *Wolbachia* MLST is not suited to discriminate between closely related strains (Ishmael
168 et al. 2009; Atyame et al. 2011; Riegler et al. 2012; Siozios et al. 2013a; Conner et al.
169 2017). This does not come as a surprise, as per definition, MLST genes are of conserved
170 nature, and thus slowly evolving. They are therefore inherently unsuited to trace very
171 recent evolutionary events.

172 Comparing the ability of MLST loci to differentiate *Wolbachia* strains with that
173 of 252 other single copy loci employed in a recent phylogenomic study of *Wolbachia*
174 evolution (Gerth & Bleidorn 2016) shows that MLST loci are not ideal for this task
175 (Supplementary Table 1). MLST loci are able to differentiate 42–63% of the 19
176 analysed *Wolbachia* genomes, whereas other conserved single copy loci may
177 differentiate up to 84% of the strains (16/19). Strikingly, none of the 252 loci that were
178 originally selected as phylogenetic markers can discriminate between all strains. This is
179 because they were chosen to be present in a single copy in all of the analysed *Wolbachia*
180 genomes (Gerth & Bleidorn 2016) and thus also represent mostly conserved
181 housekeeping loci (Supplementary Table 1). In summary, conserved single copy genes
182 are generally unsuited marker to differentiate for closely related *Wolbachia* genomes,
183 and among those, MLST loci do not perform particularly well.

184 Therefore, when designing an experiment with the main or foremost goal of
185 differentiating *Wolbachia* strains, one should employ fast evolving markers such as
186 ankyrin repeats, insertion sequences, or other mobile elements that have been shown to
187 be the fastest evolving genomic features of *Wolbachia* (Wu et al. 2004; Tanaka et al.
188 2009; Newton et al. 2016). As these will likely be very different between distantly
189 related strains (Cerveau et al. 2011) a universal set of markers suitable across the
190 breadth of *Wolbachia* diversity does not exist. As a consequence, in many cases it will

191 be inevitable to identify suitable markers for *Wolbachia* differentiation through
192 comparative genomics of a representative sample of the strains to be investigated.
193 Furthermore, we advocate to adjust not only the type, but also the number of loci
194 employed for strain differentiation. Random sampling MLST profiles from the known
195 diversity of *Wolbachia* MLST profiles illustrates that in many cases, two or three MLST
196 loci provide similar resolution to all five MLST genes (Fig. 1). For example, when
197 analysing 20 *Wolbachia* strains and using only the two most variable MLST genes *hcpA*
198 and *gatB*, one would on average be able to differentiate at least 19 of these strains. For
199 40 strains, three loci provide a similar resolution (Fig. 1). Although this comes with the
200 caveat that not all systems will show the same *Wolbachia* MLST profile frequencies as
201 the MLST database, it demonstrates that careful adjustment of loci to the study system
202 can save time and money. Instead of typing all *Wolbachia* samples with five MLST loci,
203 we therefore recommend to maximise the number of detectable *Wolbachia* strains by
204 first typing with the fastest evolving marker available (ideally, this would have been
205 identified *a prior*i through comparative genomics), and then continue with additional
206 markers as the number of samples increases.
207

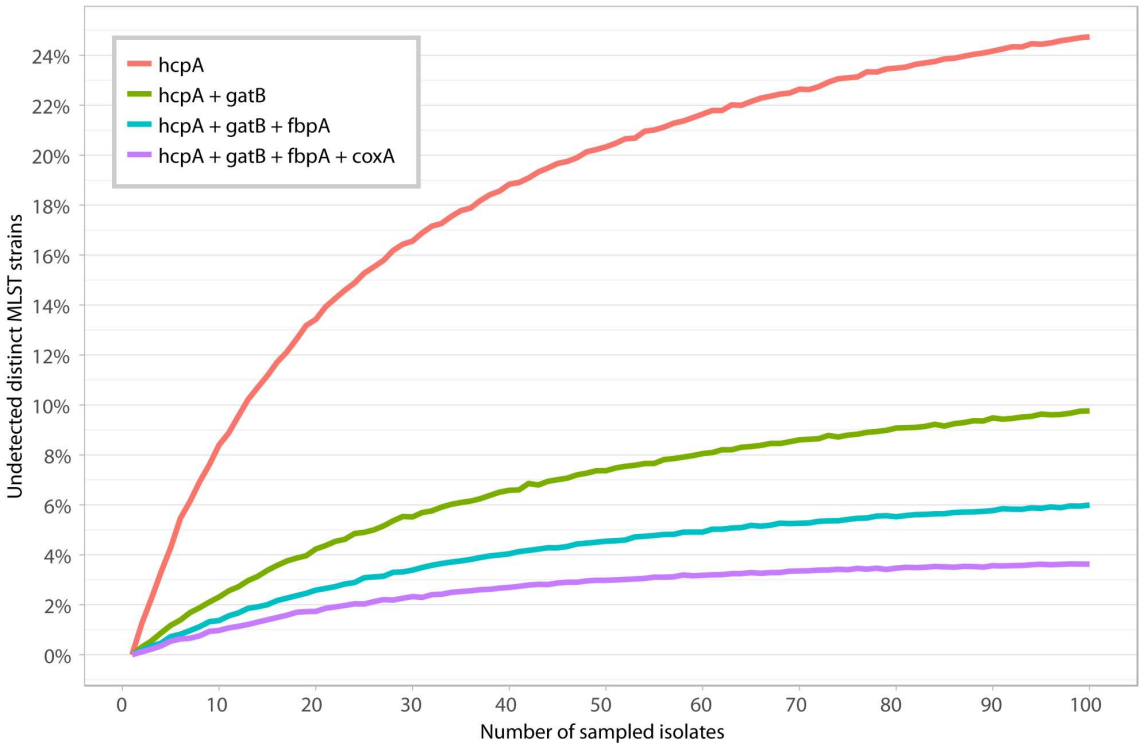


Fig. 1 Ability of MLST markers in differentiating *Wolbachia* strains. Average proportion of undetected distinct *Wolbachia* MLST profiles when using only one, two, three or four MLST genes is displayed in relation to the number of analysed strains. Figure is based on all complete *Wolbachia* MLST profiles (740 in total, 472 of which are unique) currently available from the pubMLST database (https://pubmlst.org/wolbachia/, last accessed August 17th, 2017). See methods for details.

Assessing genetic differentiation of *Wolbachia* strains with MLST genes

In addition to differentiating strains, a strain typing system should also be able to characterize the genetic diversity of a set of strains to be analysed. For this to be as accurate as possible, the molecular divergence of investigated strains at their MLST loci would have to be identical or very similar to genome wide divergence rates, or correlate with genome wide rates very well. If the assumptions underlying the choice of MLST

223 loci (mostly neutral selection, see above) are correct, one might expect these two
224 characteristics to be met. However, an analysis of MLST vs. core genome divergence
225 rates shows that this is not true for the currently employed *Wolbachia* MLST loci (Fig.
226 2). As expected from the previous observations (see above) core genome divergence of
227 lower than ~0.2% cannot be detected with any of the MLST loci (Fig. 2). For *ftsZ*, even
228 strains that are genetically divergent by more than 1% may appear identical.
229 Furthermore, a number of strains that are diverged by 1.5–2% appear similarly
230 divergent at their *ftsZ* and *fbpA* loci (Fig. 2). This may indicate nucleotide substitution
231 saturation, which would impede genetic comparison of distantly related *Wolbachia*
232 strains with MLST.
233

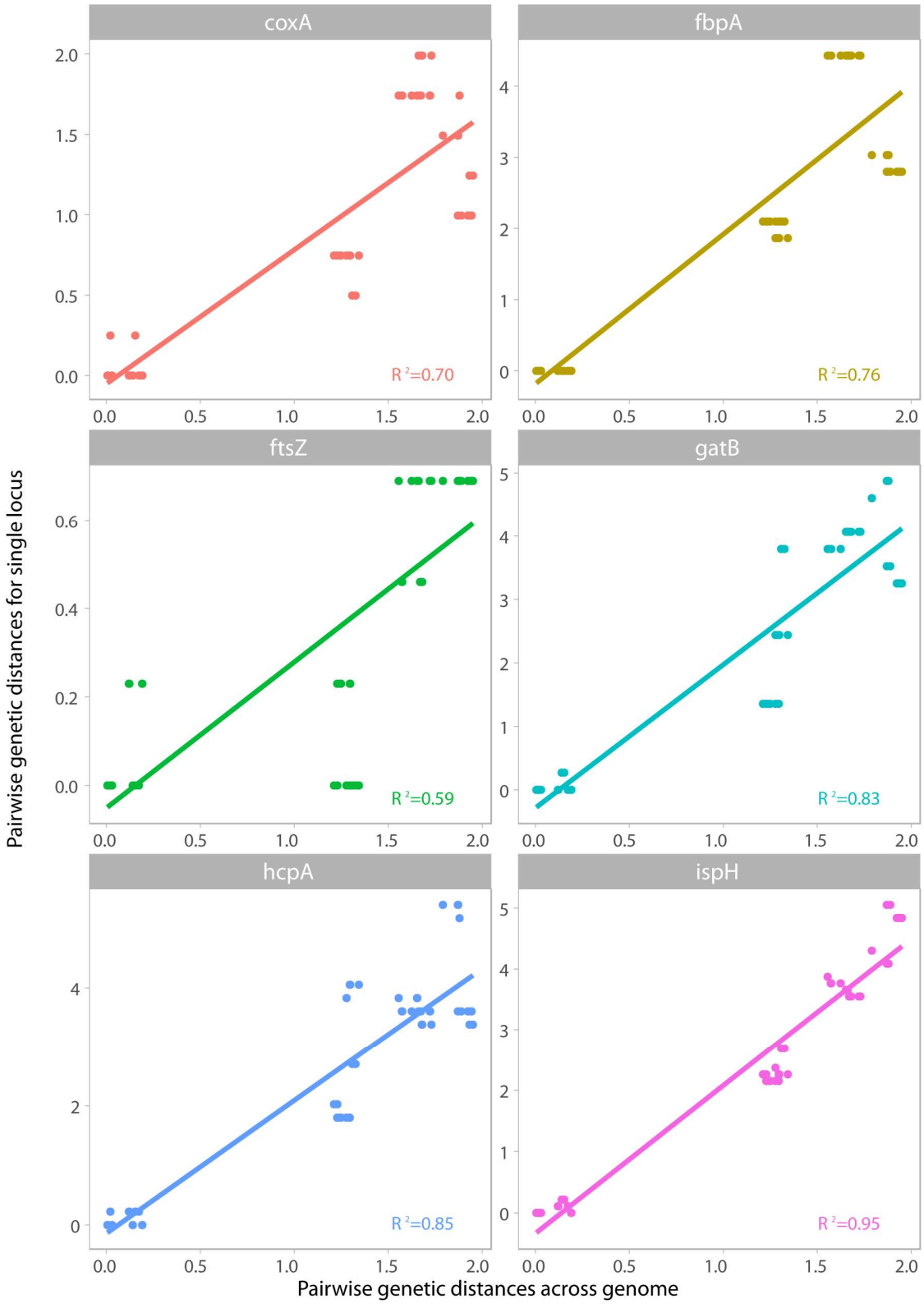


Fig. 2 Correlation of genetic distances of *Wolbachia* strains at MLST loci to genome-wide distances. Each data point corresponds to a single pair of *Wolbachia* strains, and shows the divergence between these two strains at MLST loci (y-axis) and genome-

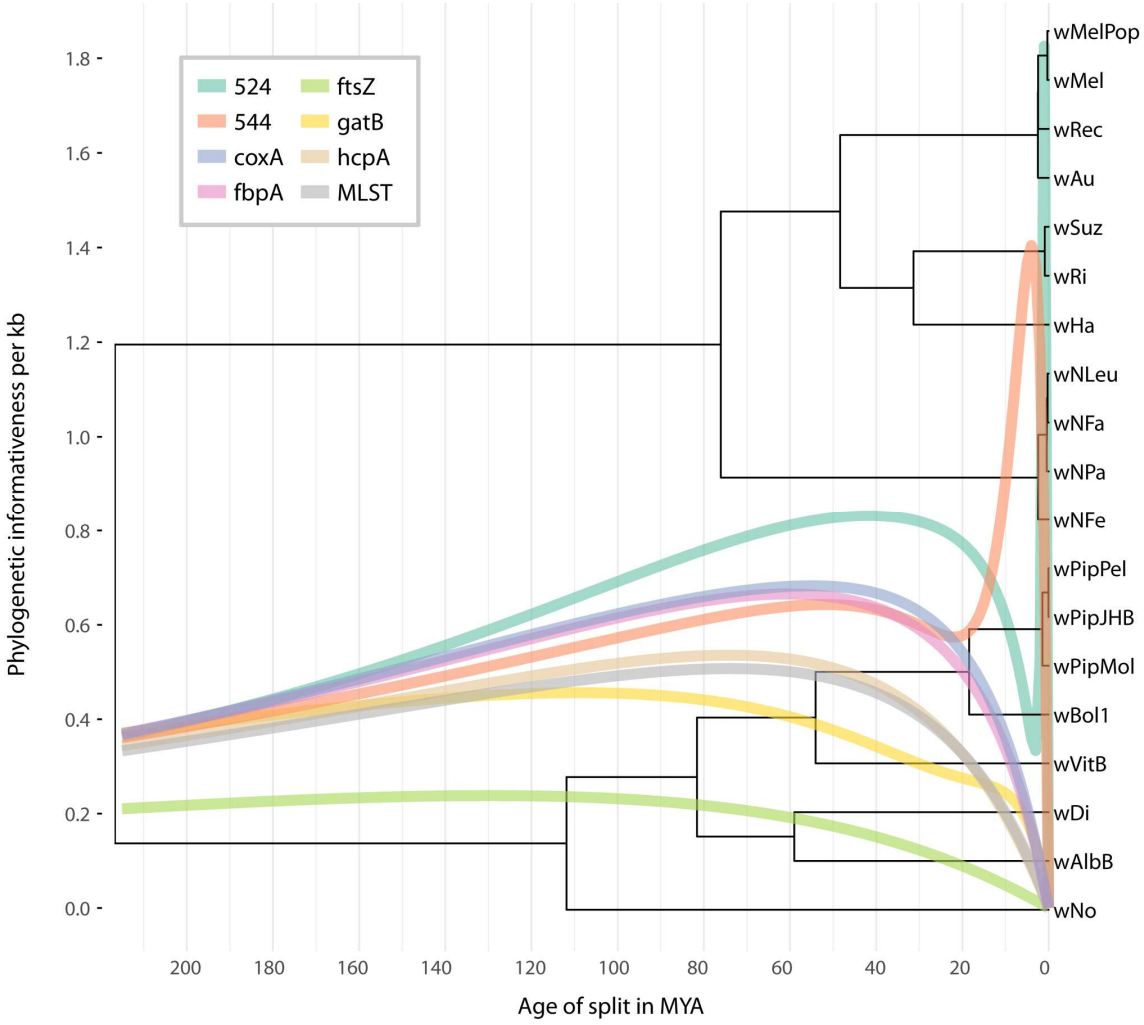
238 wide distance (x-axis, mean distance from 252 single copy orthologs). Panels
239 correspond to one of the 5 MLST loci and *ispH* (encoding 4-hydroxy-3-methylbut-2-
240 enyl diphosphate reductase) for comparison. Linear regression models were fitted using
241 the R statistical environment (R Core Team 2015). All distances are displayed as raw
242 genetic distances in percent. Please note that all pairwise distances are from supergroup
243 A strains only, as including supergroup B strains would lead to skewed distributions
244 (small distances within supergroups and large distances between supergoups) and
245 therefore biased correlation estimates. All R^2 values for all analysed loci and both
246 supergroups can be found in Supplementary Table 1.
247
248 Further to these patterns, out of the five MLST loci, only *coxA* shows genetic
249 divergence rates similar to those obtained from whole genome information, whereas
250 those of *ftsZ* are lower and the ones from *hcpA*, *fbpA* and *gatB* are higher (Fig. 2).
251 Finally, none of the divergence rates estimated from the 5 MLST loci correlate very well
252 with genome wide rates (R^2 values of regression in linear model 0.59–0.85, Fig. 2),
253 which contrasts with loci that show a very good correlation in this respect (e.g., *ispH*,
254 Fig. 2). In summary, the MLST loci are not a good approximation for genome wide
255 divergence rates of *Wolbachia* strains, and other loci may be more appropriate (Fig. 2,
256 Supplementary Table 1). This also means that genetic divergence ratios obtained from
257 MLST loci should be interpreted cautiously and other loci should be explored as
258 alternative. However, as the performance at this task differs even for a single locus
259 between supergroups (Supplementary Table 1), comparative genomics will again in
260 many cases be the only option to reliably determine divergence rates between a sample
261 *Wolbachia* strains.

262 **Phylogenetic analyses of *Wolbachia* strains using MLST**

263 Judging from the abstracts and keywords of all articles citing the original
264 *Wolbachia* MLST publication, questions that are commonly addressed with *Wolbachia*
265 MLST are phylogeny & phylogeography (102 articles with corresponding terms), and
266 horizontal transmission of strains (58 articles). Since the determination of horizontal
267 transmission with molecular methods also requires phylogenetic approaches, one can
268 summarize that phylogenies are one major field of application for *Wolbachia* MLST.
269 This is despite the authors' original assessment that MLST loci are not necessarily good
270 phylogenetic markers ("caution in interpretation of phylogenetic relationships is
271 necessary", Baldo et al. 2006), and despite the fact that an assessment of its
272 performance as phylogenetic marker is lacking.

273 The level of resolution across time for a given gene in a phylogenetic analysis
274 can be estimated by its phylogenetic informativeness (PI), which measures the relative
275 ratio of phylogenetic signal to noise across time (Townsend 2007). Analysing the PI
276 profiles of all MLST genes for a set of *Wolbachia* strains covering supergroup A and B
277 reveals that all of them show the highest phylogenetic resolution on the supergroup
278 level (Fig. 3). According to Gerth & Bleidorn (2016), the supergroups A and B have
279 diverged more than 200 million years ago. MLST genes however provide only little
280 phylogenetic information for strains that diverged much more recently (Fig. 3). As
281 *Wolbachia* likely moves between hosts at a fast rate (Gerth et al. 2013; Bailly-Bechet et
282 al. 2017), the MLST approach is not suited to infer phylogenetic relationships of closely
283 related strains, to detect recent horizontal transmissions or to assess ecological
284 timescales of *Wolbachia* movements between populations. However, a number of
285 *Wolbachia* genes evolve considerably faster than MLST loci (as measured by genetic
286 divergence or number of variable alignment sites) and also provide phylogenetic

287 information on very shallow phylogenetic levels (Fig. 3, Supplementary Table 1,
288 Supplementary Fig. 1). These loci might be good candidates for resolving very recent
289 evolutionary events.
290



291

292 **Fig. 3** Phylogenetic informativeness (PI) of five MLST-gene alignments from
293 supergroup A and B *Wolbachia* strains. For comparison, two loci displaying relatively
294 high PI for recent evolutionary events are also shown. Ultrametric tree (based on 252
295 loci available under https://github.com/gerthmicha/wolbachia-
296 mlst/tree/master/alignments) was taken from Gerth & Bleidorn (2016). See
297 Supplementary Figure 1 for PI profiles for of all analysed loci.
298

299 Moreover, as already mentioned in the original MLST publication (Baldo *et al.*,
300 2006), all MLST loci except for *ftsZ* show some level of intragenic recombination.
301 Indeed, using all alleles present in the *Wolbachia* database today, signals of
302 recombination can be detected for all five markers using the PHI test (Bruen et al.
303 2006). The presence of horizontal genetic exchange makes the interpretation of
304 phylogenetic analyses of MLST genes challenging, as the resulting tree may not reflect
305 the evolutionary relationships of the analysed strains (Holmes et al. 1999; Jiggins et al.
306 2001). When comparing the phylogenetic reconstruction for the dataset in Fig. 3 using
307 the five concatenated MLST-fragments with the original analysis based on 252
308 orthologs, several differences in the topology are apparent (Fig. 4A, B). Seven internal
309 nodes are reconstructed differently in the MLST-based analysis (Fig. 4B) and the
310 branch length differences between analyses, especially within supergroups are striking.
311 These differences are likely due to the misleading signal from recombination events.
312 ClonalFrame is a Bayesian phylogenetic framework that was developed
313 especially for MLST datasets and capable of inferring relationships despite the presence
314 of recombination (Didelot & Falush 2007). Nevertheless, analysing our dataset with
315 ClonalFrame led to a similarly high number of conflicting nodes (six) in comparison to
316 the phylogenomic dataset, and multiple polytomies (i.e., unresolved nodes, Fig 4C).
317 This shows that the usage of recombination-aware phylogenetic methods cannot
318 circumvent the inherent problems of *Wolbachia* MLST genes as phylogenetic markers.
319 As some level of conflict exists between the trees recovered from most single gene loci
320 and the one from the supermatrix (Supplementary Fig. 1), whole genome based
321 phylogenies are desirable to minimize biases.
322
323

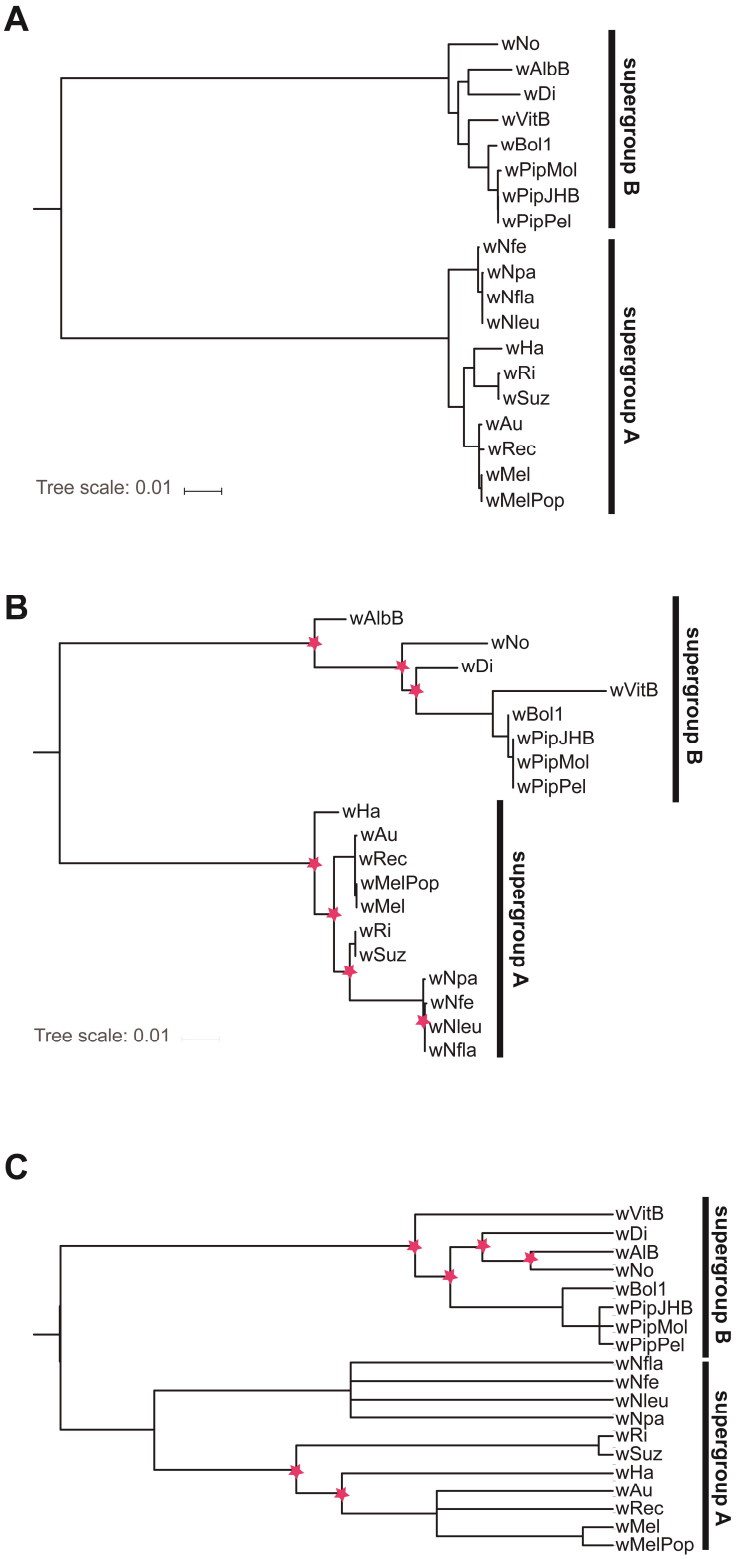


Fig. 4 Phylogenetic reconstruction of supergroup A and B strains as selected by Gerth & Bleidorn (2016). A) Maximum likelihood analysis based on optimal partitions and models as selected by IQ-TREE (Nguyen et al. 2015) for a dataset containing nucleotide data of 252 non-recombining orthologs. B) Maximum likelihood reconstruction of five

329 MLST gene fragments of the same taxa using optimal partitions and models as selected
330 by IQ-TREE. Seven conflicting nodes (red asterisks) compared to the phylogenomic
331 analyses are highlighted. C) ClonalFrame (Didelot & Falush 2007) analysis of the
332 MLST dataset, with six conflicting nodes highlighted (red asterisks).
333
334 Homologous recombination is widespread among Bacteria (Didelot & Maiden
335 2010). One way to circumvent problems in phylogenetics arising from recombination is
336 to estimate relationships between strains based on allele designations. A simple method
337 for this is to cluster strains based on their similarity, which can be visualized as a
338 dendrogram. However, strain similarity does not necessarily reflect common ancestry. A
339 popular and more sophisticated method to analyse allele-based strain data is eBURST
340 (Feil et al. 2004). This software incorporates a model of bacterial evolution in which
341 strains that are increasing in frequency diversify, thereby forming clusters of similar
342 genotypes. For MLST data, so-called clonal complexes are defined as groups that share
343 a predefined number of alleles (e.g., three of five allele designations are identical) with
344 at least one other strain type. After searching for these clonal complexes, the likely
345 founding strain type is inferred, as are evolutionary relationships within this clonal
346 complex. Simulation studies have shown that when recombination is absent or present
347 in low to moderate levels, the inferred relationships of clonal complexes are very similar
348 to the (known) true ancestry (Turner et al. 2007). However, increasing rates of the
349 frequency of recombination to mutations led to a strong decrease in the reliability of
350 eBURST analyses. In *Wolbachia,* the overall ratio of recombination to mutation events
351 to explain the generation of a substitution is ranging from 2.3 to 8.2, depending on the
352 analysed genome (Ellegaard et al. 2013). Therefore, the high recombination rates in
353 *Wolbachia* genomes make allele-based analyses unreliable.

354 In addition to problems with recombination, there are also theoretical arguments
355 against using 'eBURST'-like clustering algorithms with *Wolbachia* MLST profiles.
356 Because the only criterion for assigning a novel allele number is at least one nucleotide
357 difference compared to all described alleles, any number of substitutions in one allele is
358 weighted equally. For example, 10 different *Wolbachia* strains may be differentiated by
359 only 9 nucleotide differences in total, or by 50, and could potentially be characterised by
360 identical sets of MLST profiles. This makes comparing these profiles across systems
361 challenging. When sampling is dense and therefore the majority of the allele diversity is
362 known, this will likely not be problematic. However, this is rarely ever the case for
363 *Wolbachia.* Given the large number of infected species, it is essentially impossible to
364 know the true diversity of *Wolbachia* in any ecosystem. Furthermore, because horizontal
365 transmissions are common (Baldo et al. 2008; Zug et al. 2012; Gerth et al. 2013; Ahmed
366 et al. 2016), and exact pathways of these transmissions are still discussed (Huigens et al.
367 2004; Le Clec'h et al. 2013; Li et al. 2016), it does not make sense to define "founding"
368 and "descending" *Wolbachia* genotypes in most cases.

369 **Alternatives to MLST**

370 MLST was developed as a replacement for an earlier strain typing approach
371 called multilocus enzyme electrophoresis (MLEE), which measured genetic variation by
372 the resolution of electrophoretic variants (electromorphs) of metabolic enzymes
373 (Maiden 2006). One problem of this method was that experiments were difficult to
374 reproduce across labs. With the availability of affordable and faster Sanger sequencers it
375 was possible to directly use sequence data instead of electromorphs. Nowadays, a wide
376 array of different high-throughput sequencing techniques is available (Bleidorn 2015;
377 Goodwin et al. 2016). Due to their small size, sequencing of complete bacterial
378 genomes is affordable and routinely carried out using benchtop sequencers in

379 laboratories with standard equipment (Loman & Pallen 2015). Consequently, strain
380 typing methods based on whole genome data were proposed, e.g., rMLST, in which a
381 set of 53 ribosomal proteins is used (Jolley et al. 2012). Ribosomal proteins are
382 universally found in bacterial genomes, show a wide distribution across genomes and
383 are expected to underlie stabilizing selection, similar to the above mentioned MLST
384 genes. In the case of *Wolbachia*, ribosomal proteins have already been used successfully
385 for phylogenomic analyses (Nikoh et al. 2014). Other typing methods simply employ all
386 available genes, i.e., whole genome sequence-typing (WGST) (Pérez-Losada et al.
387 2013) or core genome MLST (cgMLST) (De Been et al. 2015).

388 Although *Wolbachia* harbour small genomes (1 to 1.5mbp in size) (Makepeace
389 & Gill 2016), sequencing and assembly is more difficult than for many other Bacteria.
390 As strictly intracellular endosymbionts, *Wolbachia* cannot be cultured axenically, and
391 although maintaining them in cell cultures is possible (Dobson et al. 2002), it is very
392 laborious and often not practical. Thus, in many cases a metagenomic sequencing
393 approach is used, targeting both host and *Wolbachia* DNA. *Wolbachia* sequence data
394 can then be retrieved using BLAST-searches and read mapping (Gerth et al. 2014).
395 However, in this case a high sequencing depth per genome is needed, as typically only a
396 small proportion of the reads will be of *Wolbachia* origin. For more efficient sequencing
397 of *Wolbachia* genomes, target enrichment protocols (Lemmon & Lemmon 2012) have
398 been established (Geniez et al. 2012; Dunning-Hotopp et al. 2017), although these are
399 not yet broadly applied.

400 Another problem in *Wolbachia* genome sequencing is the high density of mobile
401 genetic elements with repetitive sequence motives (Wu et al. 2004), which may lead to
402 very fragmented assemblies. However, for analyses focussing on sequence data of
403 selected loci and not on synteny, incompletely assembled *Wolbachia* draft genomes are

404 sufficient. Working with complete (or draft) genomes has the advantage that
405 comparative analyses can be used to retrieve large sets of orthologous and
406 recombination-free loci (Comandatore et al. 2013). These datasets allow to circumvent
407 almost all problems with MLST outlined in this article, and further enable the
408 identification hypervariable regions such as tandem repeat markers (Riegler et al. 2012)
409 or ankyrin repeat domains (Siozios et al. 2013b).

410 Although whole genome approaches are the arguably the best way to address
411 *Wolbachia* strain differentiation, diversity estimates, and phylogeny, they may in some
412 cases be too cost- or time intensive, and there will be questions that must be addressed
413 with a small number of genetic marker loci. In this case we here provide a
414 characterization of 252 conserved single copy genes by a number of criteria, each of
415 which may be important in strain typing, depending on the question to be addressed
416 (Supplementary Table 1). We point out that for none of these criteria, the MLST loci
417 perform particularly well, and we therefore strongly suggest to chose marker loci based
418 on the experimental design rather than on the convenient availability

419 **Summary & conclusion**

420 MLST analyses are widely used in the community of *Wolbachia* researchers and
421 a large database for comparative studies is available. This database and the availability
422 of PCR protocols for most *Wolbachia* strains represent a convenient and valuable
423 resource. However, for most tasks routinely employed for, *Wolbachia* MLST markers
424 are unsuited. They are too conserved to allow reliable and fine-scaled strain
425 differentiation, they do not reflect genome wide divergence rates well, and they are poor
426 phylogenetic markers at shallow or deep divergence levels. Further, they are
427 outcompeted at all of these tasks by other loci. These properties make the definition of a
428 strain in the genus *Wolbachia* per MLST very problematic and we recommend that this

429 practice is discontinued. Instead, we advise to tailor adequate marker loci as required for
430 the investigated strains. Naturally, these will differ between study systems and research
431 questions, but we think that the shortcomings of MLST loci outweigh their benefit of
432 universality. Generally, we hope that the *Wolbachia* community will embrace whole
433 genome typing methods, which are already standardly employed in clinical
434 microbiology. However, efficient novel *Wolbachia* genome sequencing (or enrichment)
435 protocols are needed for this to succeed.

436 **Methods**

437 *Data acquisition*

438 Most MLST sequences, isolates and profiles described and analysed in this paper
439 were downloaded from the *Wolbachia* PubMLST database (Jolley et al. 2004; Baldo et
440 al. 2006; https://pubmlst.org/wolbachia/, last accessed 17th of August 2017). For
441 comparative analysis of 19 supergroup A and B *Wolbachia* strains, the corresponding
442 MLST gene sequences were recovered via blastn (Camacho et al. 2009) searches
443 against coding nucleotide sequences of the 19 *Wolbachia* strains, using MLST
444 sequences from the online database as a query. The hits were trimmed manually to
445 conform to the length of *Wolbachia* MLST alleles. In addition, 252 loci from complete
446 or draft *Wolbachia* genomes were acquired as described in Gerth & Bleidorn (2016).
447 Alignments are available at https://github.com/gerthmicha/wolbachia-
448 mlst/tree/master/alignments). Briefly, the 252 loci were single copy genes present in all
449 of the 19 investigated *Wolbachia* strains that did not show evidence for recombination.
450 Orthology was assessed with OrthoFinder version 0.2.8 (Emms & Kelly 2015), and
451 alignment was performed based on codons using Mafft version 7.215 (Katoh &
452 Standley 2013). In the following, the performance of *Wolbachia* MLST loci was

453 compared to that of the 252 loci with regard to their ability to differentiate strains, to
454 approximate genome-wide divergence, and to reflect core genome phylogeny.

455 *Strain differentiation*

456 Strain differentiation ability was assessed for all investigated loci by the
457 proportion of distinct alleles in all alleles. This was calculated using the function
458 'haplotype' of the R package pegas (Paradis 2010; R Core Team 2015). As additional
459 measures of strains differentiation, we calculated average pairwise genetic distances and
460 the number of variable alignment sites using the functions 'dist.dna' and 'seg.sites' of the
461 R package APE (Paradis et al. 2004), respectively. All measures can be found in
462 Supplementary Table 1.

463 To determine the resolution of the single, two, three or four most variable MLST
464 loci in comparison to all five loci, we randomly sampled MLST profiles from the
465 known diversity of MLST strains in the pubMLST database (at the time of the analysis,
466 740 complete MLST profiles, 472 of which were unique). Random sampling was
467 performed for datasets of 1–100 samples, and repeated 10,000 times in all cases. The
468 number of distinct isolates among the samples based on a single, two, three or four
469 MLST loci was counted and compared to the number of distinct isolated based on
470 complete MLST profiles.

471 *Divergence rates*

472 For all investigated loci, we aimed to assess how well genetic distances of a
473 single locus reflect the genetic distances of the core genome. To this end, we calculated
474 all possible pairwise raw genetic distances (55 pairwise distances for 19 strains
475 analysed) for each locus and for the concatenated 252 loci (as approximation of the core
476 genome) as described above. Next, the correlation of the distances from each single
477 locus with the core genome was determined by fitting a linear model within the R

478 statistical framework. All R^2 values for these models can be found in Supplementary
479 Table 1. Due to the nature of the dataset, there is a bimodal distribution of distances:
480 large distances between supergroups, and small distances within supergoups. Using this
481 biased dataset, all correlation measures for all loci were very high. Therefore, we
482 decided it would be more appropriate to perform this analysis separately for each
483 supergroup.

484 *Phylogenetic analyses*

485 Phylogenetic analyses of 19 *Wolbachia* strains was performed for a dataset of
486 five concatenated MLST genes, one dataset of 252 concatenated single copy orthologs
487 and for each of the 257 loci separately. For all analyses, a maximum likelihood tree
488 search was performed with IQ-TREE version 1.5.4 (Nguyen et al. 2015) using the
489 implemented optimal model search and, for multi-gene analyses, optimal partition
490 selection algorithms (Lanfear et al. 2012; Chernomor et al. 2016; Kalyaanamoorthy et
491 al. 2017). The MLST dataset was further analysed with ClonalFrame version 1.2
492 (Didelot & Falush 2007), using four independent runs with 1,000,000 generations each
493 and a burnin of 50% for all runs. Convergence of runs and stability of sampled
494 parameters was verified by plotting likelihood values and other parameters in R. All
495 runs converged on identical topologies. Congruence and conflict between single gene
496 analyses and core genome analysis was also assessed by calculating normalized
497 Robinson-Foulds distances (Robinson & Foulds 1981) with RAxML version 8.2.1
498 (Stamatakis 2014), and by substracting the log likelihoods from single gene analyses
499 from the one of the best core genome phylogeny.

500 Finally, phylogenetic informativeness (PI), i.e., the relative amount of
501 phylogenetic signal to noise across time was estimated for all analysed loci using
502 TAPIR (Faircloth et al. 2012), an efficient implementation of Townsend's phylogenetic

503 informativeness (Townsend 2007), which makes use of the HyPhy software package
504 (Pond & Muse 2005). To this end, an ultrametric tree of the analysed *Wolbachia* strains
505 was taken from (Gerth & Bleidorn 2016).

506 **Funding**

507 This work was supported by the Spanish Ministry of Science and Education (MEC)
508 [RYC-2014-15615 to CB]; European Molecular Biology Organization [ALTF 48-2015
509 to MG]; and Marie-Curie Actions of the European Commission [LTFCOFUND2013,
510 GA-2013-609409 to MG].

511 **References**

512 Achtman M (2008) Evolution, population structure, and phylogeography of genetically
513 monomorphic bacterial pathogens. *Annual Review of Microbiology* **62**, 53–70.

514 Achtman M (2012) Insights from genomic comparisons of genetically monomorphic
515 bacterial pathogens. *Philosophical Transactions of the Royal Society of London B:*
516 *Biological Sciences* **367**, 860–867.

517 Ahmed MZ, Breinholt JW, Kawahara AY (2016) Evidence for common horizontal
518 transmission of *Wolbachia* among butterflies and moths. *BMC Evolutionary Biology*
519 **16**, .

520 Atyame CM, Delsuc F, Pasteur N, Weill M, Duron O (2011) Diversification of
521 *Wolbachia* endosymbiont in the *Culex pipiens* mosquito. *Molecular Biology and*
522 *Evolution* **28**, 2761–2772.

523 Bailly-Bechet M, Martins-Simões P, Szöllősi G, Mialdea G, Sagot M-F, Charlat S
524 (2017) How long does Wolbachia remain on board?. *Molecular Biology and*
525 *Evolution* **13**, 1183–1193.

526 Baldo L, Ayoub NA, Hayashi CY, Russell JA, Stahlhut JK, Werren JH (2008) Insight
527 into the routes of *Wolbachia* invasion: high levels of horizontal transfer in the spider
528 genus *Agelenopsis* revealed by *Wolbachia* strain and mitochondrial DNA diversity.
529 *Molecular Ecology* **17**, 557–569.

530 Baldo L, Dunning Hotopp JC, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR,
531 Hayashi C, Maiden MCJ, Tettelin H, Werren JH (2006) Multilocus sequence typing
532 system for the endosymbiont *Wolbachia pipientis*. *Applied and Environmental*
533 *Microbiology* **72**, 7098–7110.

534 Bleidorn C (2015) Third generation sequencing: technology and its potential impact on
535 evolutionary biodiversity research. *Systematics and Biodiversity* **14**, 1–8.

536 Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting
537 the presence of recombination. *Genetics* **172**, 2665–2681.

538 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL
539 (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

540 Cerveau N, Leclercq S, Leroy E, Bouchon D, Cordaux R (2011) Short and long-term
541 evolutionary dynamics of bacterial insertion sequences: insights from *Wolbachia*
542 endosymbionts. *Genome Biology and Evolution* **3**, 1175–1186.

543 Chernomor O, von Haeseler A, Minh BQ (2016) Terrace aware data structure for
544 phylogenomic inference from supermatrices. *Systematic Biology* **65**, 997-1008.

545 Comandatore F, Sassera D, Montagna M, Kumar S, Koutsovoulos G, Thomas G,
546 Repton C, Babayan S a, Gray N, Cordaux R, Darby AC, Makepeace B, Blaxter ML
547 (2013) Phylogenomics and analysis of shared genes suggest a single transition to
548 mutualism in *Wolbachia* of nematodes. *Genome Biology and Evolution* **5**, 1668–
549 1674.

550 Conner WR, Blaxter ML, Anfora G, Ometto L, Rota-Stabelli O, Turelli M (2017)
551 Genome comparisons indicate recent transfer of wRi-like *Wolbachia* between sister
552 species *Drosophila suzukii* and *D subpulchrella*. *bioRxiv* , 135475.

553 Cooper JE, Feil EJ (2004) Multilocus sequence typing–what is resolved?. *Trends in*
554 *Microbiology* **12**, 373–377.

555 De Been M, Pinholt M, Top J, Bletz S, Mellmann A, Van Schaik W, Brouwer E, Rogers
556 M, Kraat Y, Bonten M, others (2015) Core genome multilocus sequence typing
557 scheme for high-resolution typing of *Enterococcus faecium*. *Journal of Clinical*
558 *Microbiology* **53**, 3788–3797.

559 Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus
560 sequence data. *Genetics* **175**, 1251–1266.

561 Didelot X, Maiden MC (2010) Impact of recombination on bacterial evolution. *Trends*
562 *in Microbiology* **18**, 315–322.

563 Dobson SL, Marsland EJ, Veneti Z, Bourtzis K, O'Neill SL (2002) Characterization of
564 *Wolbachia* host cell range via the in vitro establishment of infections. *Applied and*
565 *Environmental Microbiology* **68**, 656–660.

566 Dunning-Hotopp JC, Slatko BE, Foster JM (2017) Targeted enrichment and sequencing
567 of recent endosymbiont-host lateral gene transfers. *Scientific Reports* **7**, 857.

568 Ellegaard KM, Klasson L, Näslund K, Bourtzis K, Andersson SGE (2013) Comparative
569 genomics of *Wolbachia* and the bacterial species concept. *Plos Genetics* **9**, e1003381.

570 Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome
571 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*
572 **16**, 157.

Faircloth BC, Chang J, Alfaro ME (2012) TAPIR enables high-throughput estimation and comparison of phylogenetic informativeness using locus-specific substitution models. *arXiv preprint 1202.1215* , .

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology* **186**, 1518–1530.

Francisco AP, Bugalho M, Ramirez M, Carriço JA (2009) Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* **10**, 152.

Geniez S, Foster JM, Kumar S, Moumen B, LeProust E, Hardy O, Guadalupe M, Thomas SJ, Boone B, Hendrickson C, Bouchon D, Grève P, Slatko BE (2012) Targeted genome enrichment for efficient purification of endosymbiont DNA from host DNA. *Symbiosis* **58**, 201–207.

Gerth M (2016) Classification of Wolbachia (Alphaproteobacteria, Rickettsiales): No evidence for a distinct supergroup in cave spiders. *Infection, Genetics and Evolution* **43**, 378–380.

Gerth M, Bleidorn C (2016) Comparative genomics provides a timeframe for *Wolbachia* evolution and exposes a recent biotin synthesis operon transfer. *Nature Microbiology* **2**, 16241.

Gerth M, Gansauge M-T, Weigert A, Bleidorn C (2014) Phylogenomic analyses uncover origin and spread of the *Wolbachia* pandemic. *Nature Communications* **5**, 5117.

Gerth M, Röthe J, Bleidorn C (2013) Tracing horizontal *Wolbachia* movements among bees (Anthophila): a combined approach using MLST data and host phylogeny. *Molecular Ecology* **22**, 6149–6162.

Glowska E, Dragun-Damian A, Dabert M, Gerth M (2015) New *Wolbachia* supergroups detected in quill mites (Acari: Syringophilidae). *Infection, Genetics and Evolution* **30**, 140–146.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351.

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

Hoffmann AA, Montgomery BL, Popovici J, Iturbe-Ormaetxe I, Johnson PH, Muzzi F, Greenfield M, Durkan M, Leong YS, Dong Y, Cook H, Axford J, Callahan AG, Kenny N, Omodei C, McGraw EA, Ryan PA, Ritchie SA, Turelli M, O'Neill SL (2011) Successful establishment of *Wolbachia* in Aedes populations to suppress dengue transmission. *Nature* **476**, 454–457.

Huigens ME, de Almeida RP, Boons PAH, Luck RF, Stouthamer R (2004) Natural interspecific and intraspecific horizontal transfer of parthenogenesis-inducing

611 Trichogramma wasps. *Proceedings of the Royal Society of London B-Biological*
612 *Sciences* **271**, 509–515.

613 Ishmael N, Hotopp JCD, Ioannidis P, Biber S, Sakamoto J, Siozios S, Nene V, Werren
614 JH, Bourtzis K, Bordenstein SR, Tettelin H (2009) Extensive genomic diversity of
615 closely related *Wolbachia* strains. *Microbiology* **155**, 2211–2222.

616 Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna
617 H, Harrison OB, Sheppard SK, Cody AJ, others (2012) Ribosomal multilocus
618 sequence typing: universal characterization of bacteria from domain to strain.
619 *Microbiology* **158**, 1005–1015.

620 Jolley K, Chan M-S, Maiden M (2004) mlstdbNet - distributed multi-locus sequence
621 typing (MLST) databases. *Bmc Bioinformatics* **5**, 86.

622 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017)
623 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Meth* **14**,
624 587–589.

625 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
626 improvements in performance and usability. *Molecular Biology and Evolution* **30**,
627 772–780.

628 Kriesner P, Hoffmann AA, Lee SF, Turelli M, Weeks AR (2013) Rapid sequential spread
629 of two *Wolbachia* variants in *Drosophila simulans*. *Plos Pathogens* **9**, e1003607.

630 Lanfear R, Calcott B, Ho SYW, Guindon S (2012) Partitionfinder: combined selection
631 of partitioning schemes and substitution models for phylogenetic analyses.
632 *Molecular Biology and Evolution* **29**, 1695–701.

633 Le Clec'h W, Chevalier FD, Genty L, Bertaux J, Bouchon D, Sicard M (2013)
634 Cannibalism and predation as paths for horizontal passage of *Wolbachia* between
635 terrestrial isopods. *Plos One* **8**, e60232.

636 Lemmon EM, Lemmon AR (2012) High-throughput genomic data in systematics and
637 phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **44**,
638 131010123440000.

639 Li S-J, Ahmed MZ, Lv N, Shi P-Q, Wang X-M, Huang J-L, Qiu B-L (2016) Plant-
640 mediated horizontal transmission of *Wolbachia* between whiteflies. *The ISME*
641 *Journal* **11**, 1019–1028.

642 Loman NJ, Pallen MJ (2015) Twenty years of bacterial genome sequencing. *Nature*
643 *Reviews Microbiology* **13**, 787.

644 Maiden MC (2006) Multilocus sequence typing of bacteria. *Annual Review of*
645 *Microbiology* **60**, 561–588.

646 Maiden MCJ, Bygraves JA, Feil E, Morelli G, Zhang Q, Zhou J, Zurth K, Feavers IM,
647 Achtman M, Spratt BG (1998) Multilocus sequence typing: A portable approach to
648 the identification of clones within populations of pathogenic microorganisms.

Proceedings of The National Academy of Sciences of The United States of America **95**, 3140–3145.

Makepeace, B.L., Gill, A.C. (2016) *Wolbachia*. In: Thomas, S. (Ed.) *Rickettsiales. Biology, Molecular Biology, Epidemiology, and Vaccine Development*, Springer Nature.

Newton IL, Clark ME, Kent BN, Bordenstein SR, Qu J, Richards S, Kelkar YD, Werren JH (2016) Comparative genomics of two closely related *Wolbachia* with different reproductive effects on hosts. *Genome Biology and Evolution* **8**, 1526–1542.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268–274.

Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T (2014) Evolutionary origin of insect-*Wolbachia* nutritional mutualism. *Proceedings of The National Academy of Sciences of The United States of America-physical Sciences* **111**, 10257–10262.

O'Neill SL, Giordano R, Colbert AM, Karr TL, Robertson HM (1992) 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. *Proceedings of The National Academy of Sciences of The United States of America-physical Sciences* **89**, 2699–2702.

Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* **26**, 419-420.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290.

Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA (2013) Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution* **16**, 38–53.

Pond SLK, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 125–181.

Riegler M, Iturbe-Ormaetxe I, Woolfit M, Miller W, O'Neill SL (2012) Tandem repeat markers as novel diagnostic tools for high resolution fingerprinting of *Wolbachia*. *BMC Microbiology* **12**, S12.

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. *Mathematical biosciences* **53**, 131–147.

Russell JA, Goldman-Huertas B, Moreau CS, Baldo L, Stahlhut JK, Werren JH, Pierce NE (2009) Specialization and geographic isolation among *Wolbachia* symbionts from ants and lycaenid butterflies. *Evolution* **63**, 624–640.

Schuler H, Bertheau C, Egan SP, Feder JL, Riegler M, Schlick-Steiner BC, Steiner FM, Johannesen J, Kern P, Tuba K, Lakatos F, Köppler K, Arthofer W, Stauffer C (2013) Evidence for a recent horizontal transmission and spatial spread of *Wolbachia* from

688 endemic *Rhagoletis cerasi* (Diptera: Tephritidae) to invasive *Rhagoletis cingulata* in
689 Europe. *Molecular Ecology* **22**, 4101–4111.

690 Siozios S, Cestaro A, Kaur R (2013a) Draft genome sequence of the *Wolbachia*
691 endosymbiont of *Drosophila suzukii*. *Genome Announcements* **1**, e00032–13.

692 Siozios S, Ioannidis P, Klasson L, Andersson SGE, Braig HR, Bourtzis K (2013b) The
693 diversity and evolution of *Wolbachia* ankyrin repeat domain genes. *Plos One* **8**,
694 e55390.

695 Sontowski R, Bernhard D, Bleidorn C, Schlegel M, Gerth M (2015) *Wolbachia*
696 distribution in selected beetle taxa characterized by PCR screens and MLST data.
697 *Ecology and Evolution* **5**, 4345–4353.

698 Stahlhut JK, Desjardins CA, Clark ME, Baldo L, Russell JA, Werren JH, Jaenike J
699 (2010) The mushroom habitat as an ecological arena for global exchange of
700 *Wolbachia*. *Molecular Ecology* **19**, 1940–1952.

701 Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-
702 analysis of large phylogenies. *Methods of Biochemical Analysis* **30**, 1312–1313.

703 Tanaka K, Furukawa S, Nikoh N, Sasaki T, Fukatsu T (2009) Complete wo phage
704 sequences reveal their dynamic evolutionary trajectories and putative functional
705 elements required for integration into the *Wolbachia* genome. *Applied and*
706 *Environmental Microbiology* **75**, 5676–5686.

707 R Core Team (2015) *R: A language and environment for statistical computing.*. R
708 Foundation for Statistical Computing, Vienna, Austria.

709 Teixeira L, Ferreira A, Ashburner M (2008) The bacterial symbiont *Wolbachia* induces
710 resistance to RNA viral infections in Drosophila melanogaster. *Plos Biology* **6**,
711 e1000002.

712 Townsend JP (2007) Profiling phylogenetic informativeness. *Systematic Biology* **56**,
713 222–231.

714 Turner KM, Hanage WP, Fraser C, Connor TR, Spratt BG (2007) Assessing the
715 reliability of eBURST using simulated populations with known ancestry. *BMC*
716 *Microbiology* **7**, 30.

717 Watanabe M, Tagami Y, Miura K, Kageyama D, Stouthamer R (2012) Distribution
718 patterns of *Wolbachia* endosymbionts in the closely related flower bugs of the genus
719 *Orius*: implications for coevolution and horizontal transfer. *Microbial Ecology* **64**,
720 537-545.

721 Weinert LA, Araujo-Jnr EV, Ahmed MZ, Welch JJ (2015) The incidence of bacterial
722 endosymbionts in terrestrial arthropods. *Proceedings of the Royal Society of London*
723 *B-Biological Sciences* **282**, 20150249.

724 Werren JH, Baldo L, Clark ME (2008) *Wolbachia*: master manipulators of invertebrate
725 biology. *Nature Reviews Microbiology* **6**, 741–751.

Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadinejad N, Wiegand C, Madupu R, Beanan MJ, Brinkac LM, Daugherty SC, Durkin AS, Kolonay JF, Nelson WC, Mohamoud Y, Lee P, Berry K, Young MB, Utterback T, Weidman J, Nierman WC, Paulsen IT, Nelson KE, Tettelin H, O'Neill SL, Eisen JA (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *Plos Biology* **2**, e69.

Zhang K-J, Han X, Hong X-Y (2013b) Various infection status and molecular evidence for horizontal transmission and recombination of *Wolbachia* and Cardinium among rice planthoppers and related species. *Insect Science* **20**, 329–44.

Zhang Y-K, Zhang K-J, Sun J-T, Yang X-M, Ge C, Hong X-Y (2013a) Diversity of *Wolbachia* in natural populations of spider mites (genus Tetranychus): evidence for complex infection history and disequilibrium distribution. *Microbial Ecology* **65**, 731–9.

Zhou WG, Rousset F, O'Neill SL (1998) Phylogeny and PCR-based classification of *Wolbachia* strains using wsp gene sequences. *Proceedings of the Royal Society of London B-Biological Sciences* **265**, 509–515.

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**, e38544.

Zug R, Hammerstein P (2015) Bad guys turned nice? A critical assessment of *Wolbachia* mutualisms in arthropod hosts. *Biological Reviews of The Cambridge Philosophical Society* **90**, 89–111.

Zug R, Koehncke A, Hammerstein P (2012) Epidemiology in evolutionary time: the case of *Wolbachia* horizontal transmission between arthropod host species. *Journal of Evolutionary Biology* **25**, 2149–2160.