Using genomes and evolutionary analyses to screen for host-specificity and positive selection in the plant pathogen *Xylella fastidiosa*

Running title: Screening for host-specificity in *X. fastidiosa*

Tiffany N. Batarseh[1], Abraham Morales-Cruz[1], Brian Ingel[2], M. Caroline Roper[2] and Brandon S. Gaut[1*]

[1]Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA, USA 92697

[2] Department of Plant Pathology, UC Riverside, Riverside, CA 92502

*Author for Correspondence: 321 Steinhaus Hall,

University of California Irvine, Irvine, CA, 92617

(949) 824-2564

bgaut@uci.edu

ABSTRACT

Xylella fastidiosa infects several economically important crops in the Americas, and it also recently emerged in Europe. Here, using a set of *Xylella* genomes reflective of the genus-wide diversity, we performed a pan-genome analysis based on both core and accessory genes, for two purposes: i) to test associations between genetic divergence and plant host species and ii) to identify positively selected genes that are potentially involved in arms-race dynamics. For the former, tests yielded significant evidence for specialization of *X. fastidiosa* to plant host species. This observation contributes to a growing literature suggesting that the phylogenetic history of *X. fastidiosa* lineages affects host range. For the latter, our analyses uncovered evidence of positive selection across codons for 5.3% (67 of 1,257) of core genes and 5.4% (201 of 3,691) of accessory genes; these genes are candidates to encode interacting factors with plant and insect hosts. Most of these genes had unknown functions, but we identified some tractable candidates including *nagZ_2*, which encodes a beta-glucosidase that is important for *Neisseria gonorrhoeae* biofilm formation; *cya*, which modulates gene expression in pathogenic bacteria; and *barA*, a membrane associated histidine kinase that has roles in cell division, metabolism, and pili formation.

KEY WORDS

Evolution, virulence, genomics, phylogenetic analysis, positive selection

ABSTRACT IMPORTANCE

Xylella fastidiosa causes devasting diseases to several critical crops. Because *X. fastidiosa* colonizes and infects many plant species, it is important to understand whether the genome of *X.*

38 *fastidiosa* has genetic determinants that underlie specialization to specific host plants. We
39 analyzed genome sequences of *X. fastidiosa* to investigate evolutionary relationships and to test
40 for evidence of positive selection on specific genes. We found a significant signal between
41 genome diversity and host plants, consistent with bacterial specialization to specific plant hosts.
42 By screening for positive selection, we identified both core and accessory genes that may affect
43 pathogenicity, including genes involved in biofilm formation.
44

45 **INTRODUCTION**

46 Bacteria exhibit extensive intraspecific variation in genome content. This variation is the
47 raw material for evolutionary adaptation, including the evolution of pathogenicity and virulence
48 (1–4). One example of genome variation comes from an early study of *Escherichia coli* that
49 compared two pathogenic strains and one non-pathogenic laboratory strain (5). Of the entire set
50 of protein coding genes annotated by the three genomes, only 39.2% were shared among the
51 three isolates. Intriguingly, the two pathogenic strains each had 1,300 unique genes, while the
52 laboratory strain had only 585, suggesting that genes that vary across accessions (i.e., accessory
53 genes) contribute to virulence. Similar patterns have been illustrated for plant pathogens (6, 7).
54 In *Xanthomonas*, for example, horizontal gene transfer (HGT) has shuffled virulent accessory
55 genes from pathogenic strains to previously non-pathogenic strains (4), facilitating the infection
56 of common bean (*Phaseolus vulgaris* L.). In short, accessory genes contribute to host-pathogen
57 interactions, making them a critical focus for comparative analyses of genome evolution and
58 function.

59 Here we investigate variation in the genome content of another plant pathogen. *Xylella*
60 *fastidiosa* is endemic to the Americas and was first identified as the causal agent of Pierce's

61 Disease (PD), an economically devastating disease in grapevines (*Vitis vinifera* ssp. *vinifera*) (8,
62 9). *X fastidiosa* causes additional economically and ecologically impactful diseases like citrus
63 variegated chlorosis, coffee leaf scorch, oak leaf scorch and elm leaf scorch, among others.
64 Historically, the geographic distribution of *X. fastidiosa* was limited to the Americas, but it was
65 recently introduced to the European continent by anthropogenic transmission, which has further
66 expanded its host range and led to emerging diseases like olive quick decline syndrome (OQDS)
67 in Italy (10, 11). *X. fastidiosa* has since been detected in various plants species across locations
68 in Europe including France, Spain, and Portugal (12, 13). In susceptible hosts, *X. fastidiosa* can
69 lead to significant crop losses, and it continues to threaten crops globally (14, 15).
70 For each of these diseases, *X. fastidiosa* is transmitted by xylem-feeding insect vectors
71 into the plant host, where it then utilizes cell wall degrading enzymes to systemically colonize
72 the xylem. In the xylem, it forms biofilms that are thought to be integral to pathogenicity (16,
73 17). Colonization is also governed, in part, by virulence and pathogenicity factors that influence
74 a wide range of bacterial functions – e.g., biofilm formation, host cell wall degradation,
75 regulatory systems, stress responses and bacterial membrane composition -- although other
76 abiotic factors (like plant drought stress) likely also contribute to disease progression (13). Given
77 its economic impact, the effects and mechanisms of *X. fastidiosa* infection have been studied
78 widely, especially in grapevine (18). However, many pathogenicity factors likely remain
79 undiscovered, and crucial questions remain unanswered about the genetic factors that govern
80 host-pathogen interactions and potential host specialization (13).
81 In this context, it is helpful to recognize that *X. fastidiosa* consists of three commonly
82 recognized subspecies that form distinct phylogenetic clades: ssp. *fastidiosa*, *multiplex*, and
83 *pauca.* Each subspecies has unique phenotypic characteristics and DNA markers (19). Two other

84 subspecies, *morus* and *sandyi*, have also been suggested, though they are not recognized as
85 broadly (9); indeed, *morus* is believed to be a product of a recombination event between
86 *fastidiosa* and *multiplex* isolates (8). The recognition of subspecies is critical, because initial
87 work suggested that subspecies correlate with specific plant hosts (20). While it has long been
88 known that that genetic differences among strains facilitate host-plant specialization (18, 21–23),
89 there is not a clear one-to-one correspondence between pathogen and host. For example, some
90 strains can infect more than one host species, as demonstrated by a strain that causes PD in
91 grapevines and also leaf scorch in almonds (21). Consequently, the questions of the evolution of
92 and determinants of host specificity are still central for understanding the distribution and effects
93 of this pathogen.

94 In this study, we analyze *X. fastidiosa* genome evolution among isolates from different
95 plant hosts. Our study is not unique in some respects, because numerous comparative genomic
96 studies of *X. fastidiosa* have been published already. Many of these studies have focused on
97 clarifying phylogenetic relationships. For example, Marcelletti and Scortichini (2016) studied 21
98 genomes to resolve taxonomic relationships among subspecies; Giampetruzzi et al. 2017
99 extended sampling to 27 genomes, in part to place a novel strain (ST53) in the broader *X.*
100 *fastidiosa* phylogeny; and Denancé et al. (2019) used kmers from 46 genomes to untangle
101 species and subspecies relationships. Another recent study compared *X. fastidiosa* populations
102 from Central/South America (Costa Rica, Brazil), North America (California, Southeastern US),
103 Europe (Spain, Italy), and Asia (Taiwan) to elucidate the evolutionary origins of the subsp.
104 *fastidiosa* and *pauca* (24). Still other studies have focused on populations. For example, Vanhove
105 et al. (2020) isolated and sequenced *X. fastidiosa* subsp. *fastidiosa* from symptomatic grapevines
106 from five different California locations (25).

107 One common theme of genomic studies is that they identify the set of genes that are
108 present in most samples (i.e., core genes) and used those genes as the basis to perform
109 phylogenetic inference. These phylogenies have been used for various purposes. For example,
110 two recent papers have used phylogenies to explore the question of host specificity. In one,
111 Uceda-Campos et al. (2022) found that *X. fastidiosa* isolates grouped on the phylogeny by
112 geography, but not by plant host species, suggesting host specificity is not correlated with
113 phylogenetic relationships and genetic divergence (26). In contrast, Kahn and Almeida (2022)
114 used the phylogeny to infer the ancestral character states of plant hosts and found that the
115 ancestral host plant could be inferred for most ancestral nodes (27). They concluded that genetic
116 history affects host range and also identified ~30 genes whose presence/absence correlated with
117 specific plant hosts.
118 In this study, we combined 20 new *X. fastidiosa* genomes with publicly available data to
119 build a dataset for molecular evolutionary analysis and to investigate patterns of host specificity
120 in a phylogenetic context. For host-specificity analyses, we focused on core genes, but we also
121 assessed the phylogenetic signal, pattern of gene gain and loss, and potential host associations of
122 accessory (i.e., non-core) genes. Our goals for these analyses were to add to the growing
123 literature about genetic correlations between phylogenetic history and host specificity but also to
124 further consider the dynamic evolution of accessory genes in this context (27). In addition, we
125 performed extensive analyses of the ratio of nonsynonymous to synonymous (dN/dS or ω)
126 substitutions to identify genes under positive selection ($\omega > 1.0$). Genes under positive selection
127 may be involved in arms-race (or Red-Queen) dynamics between pathogens and hosts (28, 29).
128 In other systems, ω analyses have identified genes with functions that contribute to host defense
129 and also discovered entirely new sets of genes and pathways involved in pathogen-host

130 interactions (30–32). Here we apply tests for positive selection in the hope of gaining insights
131 into the sets of genes that may affects host-pathogen interactions.
132
133 **METHODS**
134 **Novel *X. fastidiosa* genomes**
135 Fully extracted DNA from 20 *X. fastidiosa* isolates were provided by the French
136 Collection of Plant-Associated Bacteria (CIRM-CFBP; http://www6.inra.fr/cirm_eng/CFBP-
137 Plant-Associated-Bacteria) and from the University of California, Riverside. Genomic DNA was
138 prepared for Illumina sequencing using the Illumina Nextera DNA Flex Library Prep kit,
139 following the manufacturer's recommendations and for Pacific Biosciences (PacBio) sequencing
140 with the SMRTbell Express Template Prep Kit 2.0. SMRTbell libraries had 10kb DNA target
141 insert size (Pacific BioSciences, Menlo Park, CA) using 360ng of sheared DNA as input. DNA
142 libraries were sequenced with both Illumina and PacBio technologies at the University of
143 California, Irvine Genomics High Throughput Facility (https://ghtf.biochem.uci.edu). The
144 Illumina sequencing reads were quality assessed using FastQC, and reads were trimmed using
145 Trimmomatic v. 0.32 (33, 34) using default options. PacBio sequencing reads were corrected and
146 trimmed using Canu v. 1.5 (35). The long and short reads were used for genome assembly with
147 Unicycler v. 0.4.8 in hybrid assembly mode (36). Genome assembly statistics were calculated
148 using Quast v. 5.0.2 (37). As is common practice (38), short contigs (<500 bp) were removed
149 from the assemblies using Seqkit v. 0.13.2 (39).
150
151 **Genome Assembly of public data and sample set curation**

152 We complemented our set of novel genomes with publicly available data. To do so, we
153 downloaded all available whole genome assemblies of *X. fastidiosa* and *X. taiwanensis* (as an
154 outgroup) from the National Center for Biotechnology Information (NCBI) and Sequence Read
155 Archive (SRA) databases on July 9, 2020 (Supplementary Table 1). In addition, we downloaded
156 the raw, short-read sequences for an additional 20 isolates (25, 40). For each isolate, we gathered
157 information about its geographic origin and its host plant from NCBI and from the Pathosystems
158 Resource Integration Center (PATRIC) database. To assemble the raw reads from the 20
159 unassembled accessions into genomes, we assessed quality and trimmed the reads and applied
160 SPAdes v. 3.14.0 (41) with the *--careful* option, following Vanhove et al. (Supplementary Table
161 2; 2020). If long reads were also available, as they were for 5 isolates from the work of Castillo
162 et al. 2020, then whole genome assembly was performed with Unicycler v. 0.4.8 in hybrid
163 assembly mode (36).
164 In total, we gathered and generated 148 *Xylella* genome assemblies. From this set, we
165 removed isolates that did not have information about their host isolation source or were lab-
166 derived recombinant strains. The remaining 129 genomes were re-annotated by the same method
167 - based on Prokka v. 1.14.6 analysis – that we applied to the new genomes, to ensure
168 homogeneity. The Prokka analyses were then input into Roary v. 3.13.0 with options *-i 80 -cd*
169 *100 -e -n -z* to obtain a core gene alignment for initial comparisons among isolates (42, 43); we
170 defined core genes as those that were detectable in 100% of the samples. This core set was
171 aligned with MAFFT and polished using gBlocks v. 0.91b (44–46). The polished alignment was
172 used as input for RAxML v. 8.2.12 to build a preliminary phylogenetic tree (47), which we used
173 to evaluate and curate the isolates (Supplementary Figure 1).

174 To curate the dataset, we created a distance matrix from the RAxML phylogenetic tree,
175 using the Tree and reticulogram REConstruction (T-REX) server (48). Many of the genomes –
176 most of which were gathered for population genomic analyses - were sampled from the same
177 plant host and were nearly identical genetically. To limit sampling biases for our species-wide
178 study, we removed clones and near-clones based on the distance matrix. That is, if two or more
179 isolates had a pairwise distance ≤ 0.0001 and came from the same host, we retained the isolate
180 with the more contiguous assembly. We also used CheckM (49) to assess genome completeness
181 based on a set of conserved single copy genes (Supplementary Table 3). After applying these
182 filters, our final dataset consisted of 63 *X. fastidiosa* genomes and one outgroup genome (*X.*
183 *taiwanensis* PLS229) that were isolated from 23 distinct plant host species (Supplementary Table
184 1).
185
186 **Pan-genome analysis**
187 To perform a pan-genome analysis, we applied Roary to the 64 *Xylella* genomes using *gff*
188 files from Prokka as input. Roary was applied with the option *-i 80,* as used in previous microbial
189 studies (38, 50), to lower the blastp sequence identity to 80% from the default 95%. We defined
190 a core gene as a gene present in 95% of the isolates used in the analysis (i.e., a core gene was
191 present in at least 60 of 63 *X. fastidiosa* accessions). From the Roary output, we extracted a
192 representative nucleotide sequence of each core and accessory gene using cdbfasta
193 (https://github.com/gpertea/cdbfasta) and translated the nucleotide sequence to amino acids using
194 the transeq command from EMBL-EBI (51). The representative sequences were the basis for
195 functional categorization -- using the eggNOG-mapper v. 2 (52, 53) -- of both core and accessory
196 genes. Gene Ontology (GO) enrichment analyses were performed online at

197 (http://geneontology.org) using *Xanthomonas campestris* as the reference list (54). To explore
198 function further, we also used the Conserved Domain Database online tool
199 (https://www.ncbi.nlm.nih.gov/cdd/) to identify protein domains.
200

201 **Phylogenetic Tree Construction**

202 We used the core gene alignment from Roary to build a phylogenetic tree, based on a
203 subset of genes that were present in all 63 *X. fastidiosa* samples and the *X. taiwenensis* outgroup.
204 To do so, we curated the alignments with gBlocks v. 0.91b (44), used the polished alignment as
205 input for IQtree v. 2.0.3, and selected the best nucleotide model for phylogenetic tree
206 construction (55, 56). We ultimately constructed an unrooted tree using the GTR+F+R8 model
207 with RAxML (Stamatakis 2014), using the 'best tree' option. Phylogenetic trees were visualized
208 and annotated using the ape package v. 5.5 in R v. 4.0.2 (57, 58). We used the most likely
209 phylogeny to test associations between phylogenetic relatedness, geography, and host isolation
210 source (plant taxonomic order information taken from https://www.itis.gov/) with ANOSIM
211 implemented in the vegan package v. 2.5-7 in R (59).

212 *X. fastidiosa* is naturally transformant and undergoes homologous recombination (9, 60),
213 but recombined genomic regions can obscure vertical phylogenetic relationships. To account for
214 potential recombination among *X. fastidiosa* genomes, we applied Gubbins v. 3.2.1 (61, 62),
215 using again the subset of genes that were found in all 64 samples. From this input, Gubbins
216 identified regions that were likely to have undergone recombination and removed them from the
217 alignment. We then built a phylogeny from this recombination-adjusted core gene alignment
218 using RAxML, as described above. We assessed the congruence between the two phylogenetic

219 trees (i.e., with and without removal of potentially recombining regions) using phytools v. 1.0-1
220 in R (63).
221 Finally, we also built a Neighbor-Joining (NJ) tree based on the presence-absence matrix
222 of accessory genes. We first calculated the Euclidean distances from the presence-absence matrix
223 of the accessory genes using the *dist* function in R (64). We then built an NJ tree from the
224 Euclidean distances using the ape package in R (57). We also utilized the ANOSIM and Mantel
225 test (in the *vegan* package) to measure the correlation between accessory gene content and
226 phylogenetic relatedness. The Mantel test required two distance matrices, which were the
227 Euclidean distances estimated from the accessory gene presence-absence matrix and the
228 distances from the RAxML core gene phylogeny generated by the reticulogram REConstruction
229 (T-REX) server (48).
230
231 **Gain and Loss of Accessory Genes**
232 We utilized GLOOME to investigate gene gain and loss dynamics along the core
233 phylogenetic tree of *X. fastidiosa* (68). GLOOME uses a mixture-model approach, coupled with
234 maximum-likelihood inference, to infer rates of gain and loss of genes along the branches of a
235 phylogeny. It takes as input the phylogenetic topology, in this case the phylogenetic topology
236 based on core genes, and a presence-absence matrix of genes. The pattern of genic presence and
237 absence was obtained through M1CR0B1AL1Z3R, as recommended by the GLOOME authors,
238 and then directly input into GLOOME using default settings (69). The default settings included a
239 fixed rate of gene gains and losses with gamma distributed rates across genes (or sites). Among
240 the output, GLOOME returned two phylogenetic trees with branch lengths representing either
241 the number of expected gain events or the number loss events on each branch. As recommended

242 (68), branch lengths representing relative gain and loss rates were extracted from the
243 phylogenetic trees using FigTree v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). To
244 normalize expected gain (or loss) events with sequence divergence, we calculated the ratio of
245 inferred gain (or loss) against the branch lengths of the sequence-based core phylogeny. Outlier
246 branches with excess normalized gains or losses were identified using the interquartile range
247 criterion.
248
249 **Positive selection analyses**
250 We employed codeml from PAML v. 4.9 to calculate ω, the ratio of nonsynonymous to
251 synonymous rates (70, 71). We performed *codeml* analysis on nucleotide alignments of the
252 single-copy core genes, single-copy accessory genes, and multicopy genes (defined as genes with
253 two or more copies in a single accession). For all tests, we required at least four sequences, as the
254 minimum number suggested for *codeml* analysis
255 (http://abacus.gene.ucl.ac.uk/software/pamlFAQs.pdf). For each gene and sequence set, we ran
256 analyses by generating an unrooted maximum-likelihood tree for each gene based on the DNA
257 alignment, using RAxML v. 8.2.12. This approach recognizes that the phylogeny of a single
258 gene may not follow the consensus phylogeny due to a history of recombination. For
259 completeness, however, we also performed codeml analyses by assuming the global phylogeny
260 for the subset of genes that were present in all 64 samples. The outcomes of the two approaches
261 were highly correlated (Supplementary Figure 2), and so for simplicity we focused on results
262 based on phylogenies inferred separately for each gene.
263 Given the input phylogenies, we performed *codeml* analyses that relied on calculating
264 likelihood ratios (LRs) under various models (Yang 2007). Briefly, we used the models to test

265 the null hypothesis that ω = 1.0 against the alternative of positive selection (ω > 1.0) in two
266 different ways. The first was a global test across the entirely phylogeny of a gene – i.e., across all
267 branches and all sites. This test requires the comparison of two models: one (Model = 0 with
268 Fix_omega = 1 and Omega = 1 in the *codeml* control file) that estimates a single ω from the data
269 and another that sets ω=1.0 (Model = 0 with Fix_omega = 0 in the *codeml* control file). The two
270 models yielded evidence for positive selection when the initial ω estimate was >1.0 and when the
271 likelihood of the two models differed significantly, based on $P < 0.01$ after FDR correction. The
272 second set of analyses was across sites – i.e., testing for genes with variable selection pressure
273 across sites. For each gene, we first compared models M0 and M3 to test for heterogeneity in
274 evolutionary rates across codons. If that test was significant, we then compared sites models M1a
275 and M2a from *codeml* to test for specific codons with evidence of positive selection (ω > 1.0).
276 For all summary statistics of ω, we excluded estimates of ω that were greater than 10 as
277 potentially unreliable due either to low d_s or poorly resolved optimization. Individual codon
278 residues under positive selection were identified using the Empirical Bayes analysis in codeml.
279

280 **DATA AVAILABILITY STATEMENT**

281 All high-throughput sequence data generated in this study have been submitted to the NCBI
282 Sequence Read Archive database at https://www.ncbi.nlm.nih.gov/sra and can be accessed with
283 project number PRJNA833428.
284

285 **RESULTS**

286 **Core and Accessory Genes in *Xylella***

287 To investigate genome evolution in *X. fastidiosa*, we sequenced 20 novel *X. fastidiosa*
288 genomes using hybrid approaches and retrieved publicly available genomes and raw sequencing
289 data (Supplementary Tables 1 and 2). After filtering for isolation source and genetic distance, we
290 retained a sample of 63 genomes that were broadly distributed among the subspecies. All our
291 analyses were performed on this final set of 63 *X. fastidiosa* genomes with the *X. taiwanensis*
292 outgroup. The *X. fastidiosa* genomes ranged in size from 2.42 Mb to 2.96 Mb, with an average
293 length of 2.61 Mb (Figure 1A) and an average of 2,478 predicted genes (Figure 1B). The
294 samples were extracted from 22 plant hosts representing 12 botanical orders (Figure 1C).
295 We categorized each gene as either core (present in 95% or more of *X. fastidiosa*
296 samples) or accessory (43). Across all 64 genomes, we identified 10,477 genes within the pan-
297 genome; of those, 1,257 were core genes and 9,220 were accessory genes, with nearly 4,000
298 genes found in only a single isolate (Supplementary Table 4; Figure 1D). We performed
299 functional analyses on both the core and accessory gene sets by grouping protein coding
300 sequences into clusters of orthologous gene (COG) (Figure 1E-F). We compared COG category
301 rankings between the core and accessory gene sets, with significant differences (Wilcoxon rank
302 sum test, paired; P = 0.0001). After excluding genes with unknown function, the largest COG
303 categories in the core gene set were 'translation, ribosomal structure, and biogenesis' (123
304 genes), 'cell wall/membrane/envelope biogenesis' (119 genes), and 'amino acid transport and
305 metabolism' (92 genes). In contrast, the largest categories for accessory genes were 'replication,
306 recombination and repair' (547 genes), 'intracellular trafficking, secretion, and vesicular
307 transport' (364 genes), and 'transcription' (298 genes). Additionally, we investigated the core
308 and accessory gene lists for significant GO based enrichment of specific biological processes
309 (Supplementary Tables 5 and 6).

310

311 **Phylogenetic Patterns of Core Genes, Accessory Genes and Hosts**

312 *Phylogenetic Relationships* We constructed a maximum likelihood phylogeny based on a
313 subset of 1,024 genes that were present in all 64 isolates. The topology was highly supported; it
314 had a mean bootstrap support of 93.75% across all nodes, with a median of 100% (Figure 2). The
315 lowest bootstrap supports were primarily found at nodes separating *X. fastidiosa* strains isolated
316 predominantly from grapevines, reflecting relatively low evolutionary divergence among these
317 samples. As expected (72), isolates clustered into three distinct clades representing the three
318 main subspecies (ssps. *fastidiosa*, *multiplex*, and *pauca*), with 27, 23 and 13 isolates in each
319 clade, respectively. To account for the possibility that homologous recombination impacted the
320 resolution of the core phylogeny, we extracted regions of the core gene alignment that had an
321 apparent history of recombination (61), ultimately removing 85.1% of the alignment. The
322 phylogeny inferred from this alignment was nonetheless highly congruent with the phylogeny
323 that did not consider recombination. Only five accessions had altered positions between the
324 recombination-adjusted and non-adjusted trees (Supplementary Figure 3).

325 To investigate general evolutionary patterns of the accessory gene complement, we
326 compared the core gene phylogeny against a phylogeny based on accessory gene composition
327 (Figure 3). Both the core gene and the accessory gene phylogenies clustered into three groups,
328 and all members of the groups were consistent between phylogenetic treatments. This pattern
329 broadly suggests that accessory genes, while defined by their inconstancy, are not exchanged *en*
330 *masse* to a sufficient degree to alter phylogenetic signal among subspecies. Within subspecies,
331 however, relationships at the tips of the phylogeny often differed between core and accessory
332 trees. As an example, the cluster corresponding to *multiplex* displayed the most discordance

333 between the core and accessory trees, with all OTUs contributing to phylogenetic incongruence
334 (Figure 3). Interestingly, our *multiplex* sample also had more plant host species than our
335 *fastidiosa* and *pauca* samples, suggesting the possibility (but by no means proving) that host
336 factors may affect or moderate genome content (27). Nonetheless, we found a significant
337 correlation between distance matrices based on the core and accessory phylogenies (Mantel test,
338 R = 0.1144, P = 0.019), which is consistent with the fact that the two trees had the same three
339 major clades. The overarching impression of these analyses is that accessory gene composition
340 does not turn-over so rapidly, due to HGT or other mechanisms, to erase phylogenetic and
341 historical signals of subspecies diversification within *X. fastidiosa*.
342 We used both species phylogenies (based on alignments with and without putative
343 recombinant regions) to test for associations between *X. fastidiosa* and their isolation sources
344 (i.e., geographic location or host plant information) using ANOSIM (see Methods). There was a
345 weakly significant phylogenetic association (ANOSIM R = 0.08178, P = 0.042) between the
346 geographic location and the phylogeny built from the full core gene alignment (ANOSIM R =
347 0.08178, P = 0.042) but not with the phylogeny built from non-recombinant regions (ANOSIM R
348 = -0.004147, P = 0.4895). Applying the same approach to host species revealed a significant
349 phylogenetic signal for both phylogenies (ANOSIM R = 0.1381, P = 0.047; non-recombining
350 regions only, ANOSIM R = 0.6698, $P < 1 \times 10^{-4}$). Since *X. fastidiosa* infects a wide range of
351 plants, we also retrieved the taxonomic order of each plant host to test for a phylogenetic signal
352 at a deeper taxonomic level, recapitulating the significant association with both phylogenies
353 (ANOSIM R = 0.3152, P < 0.0001; non-recombining regions only, ANOSIM R = 0.1226, P =
354 0.0198). In other words, strains isolated from plants within the same taxonomic order were more
355 phylogenetically similar to one another than isolates taken from unrelated plants.

356 We hypothesized that accessory genes are crucial in pathogen-host interaction and
357 therefore repeated ANOSIM analyses with a distance matrix based on the presence and absence
358 of accessory genes (Figure 3). We found a significant association between accessory gene
359 content and geographic isolation source (ANOSIM R = 0.4553, P = 0.5307) and a weakly
360 significant association between accessory gene content and host species (ANOSIM R = 0.1503,
361 P = 0.0372). The association was lost, however, at the level of plant order (ANOSIM R =
362 0.02367, P = 0.3033). Overall, associations were less evident based on accessory gene content
363 vs. the core-gene phylogeny.

364 *Gene Gain and Loss.* The sheer number of accessory genes indicate that the genome
365 content of *X. fastidiosa* is, like other microbes (73, 74), shaped by extensive gene gain and loss
366 events that are probably mediated by HGT (75). We were interested in assessing the pattern of
367 gene gain and loss across the phylogenetic tree, hypothesizing that both could be enhanced on
368 branches that lead to host shifts. We used GLOOME to estimate the number of gain and losses of
369 accessory genes across the *X. fastidiosa* phylogeny and represented those estimates
370 phylogenetically (Figure 4). Ignoring the branch leading to the *X. taiwanensis* outgroup
371 (PLS229), the internal branches discriminating the *X. fastidiosa* subspecies were estimated to
372 average ~550 separate gene gain and gene loss events. The remainder of the tip and ingroup
373 branches averaged ~100 gene gain and loss events (average gains/branch = 92.8 genes; average
374 losses/branch = 100.0 genes; Figure 4A&B).

375 While it is useful to estimate the number of gains and losses on each branch, we thought
376 it more helpful to normalize the number of estimated gain and loss events by branch lengths,
377 estimated from the sequence analysis of core genes. This normalization by branch length
378 converted the number of gene gains and losses to *rates* of gene gain (or loss) relative to sequence

379 divergence. We then sought to identify branches with aberrantly high rates of gene gain or loss
380 (Figure 4C&D), which reflect branches with especially notable turnover of accessory genes. Like
381 a previous microbial study (76), we found that most of the phylogenetic lineages with outlier
382 rates were located at the tips of the phylogenetic tree. For example, of the 21 branches with high
383 rates of gene gain, 19 were at the tips of the phylogeny (Figure 4A). Similarly, 18 of 21 branches
384 with high rates of gene loss were external branches. These observations suggest features about
385 the evolutionary dynamics of genic turnover (see Discussion).
386
387 **Characterizing selection with ω**
388 We characterized selection on individual genes by estimating the dN/dS ratio (ω); we
389 especially sought to identify genes that experienced positive selection (i.e., $\omega > 1.0$), as a
390 potential signal of genes that contribute to dynamics between the pathogen and its hosts. To do
391 so, we applied a series of nucleotide substitution models to individual genes, ultimately resulting
392 in tests for positive selection on two levels: globally across a phylogeny and across codon sites
393 (see Methods). For these tests, we examined the full complement of 1,257 core genes, a subset of
394 3,691 accessory genes, and a set of 187 multicopy genes.
395 *Testing selection globally for each gene:* We first estimated a single ω value for each
396 gene, using a method that assumes ω is constant across all branches of the entire gene tree and
397 across all codons in the nucleotide alignment. Applied to the core genes, ω estimates ($\widehat{\omega}$) ranged
398 from 0.01048 to 2.92803 with an average of 0.21973 (Figure 5A). Nineteen core genes had $\widehat{\omega}$
399 higher than 1.0, but none of these were significantly > 1.0 (P > 0.01, FDR correction). In fact, the
400 vast majority (1,144 of 1,257) of core genes had $\widehat{\omega}$ significantly < 1.0 (P < 0.01, FDR correction;
401 Figure 5A), reflecting pervasive purifying selection. The range of $\widehat{\omega}$ was substantially broader

402 for accessory genes, from $\hat{\omega}$ = 0.0001 to 9.60069, with an average of 0.51443 (Figure 5B).
403 Among accessory genes, 367 (9.9%) had a global estimate of ω > 1.0, but only eight had
404 statistically significant evidence for positive selection. These eight genes were candidates to
405 encode proteins involved in host-pathogen interactions, but seven of eight were annotated as
406 hypothetical genes (Supplementary Table 7). Overall, the average $\hat{\omega}$ was significantly higher in
407 the accessory gene set compared to the core genes (Welch's T-test, $P < 2.2 \times 10^{-16}$), reflecting
408 either lower purifying selection against these genes, more positive selection, or both.
409 We also identified 187 genes that had 2 or more copies within a single accession in a
410 syntenic context, but that were single copy in other accessions. We performed *codeml* analysis to
411 estimate ω for each multicopy gene; $\hat{\omega}$ ranged from 0.02272 to 4.26800, with an average of
412 0.51129 (Figure 5D). Over half of the genes had $\hat{\omega}$ significantly < 1.0 (59.4%; P < 0.01, FDR
413 correction); only one, a hypothetical gene (*group_1109*) had $\hat{\omega}$ significantly higher than 1.0 ($\hat{\omega}$ =
414 1.85845, P < 0.01, FDR correction).
415 *Positive selection in codon sites:* The global test is a conservative criterion to search of
416 positive selection, perhaps overly so. Accordingly, we turned to an alternative method that tests
417 for variation in ω among codon sites and identifies whether sites are under positive selection. To
418 do so, we ran the sites models in *codeml*, which are a group of nested models. For completeness,
419 we first compared sites model M0, which represents the null hypothesis that there is a single ω
420 value for all sites, against sites model M3, which permits ω to vary among sites. In the core
421 genes, the likelihood ratio test was significant for 501 genes (P < 0.01, FDR correction). We then
422 took this set to compare to test for positive selection using sites models. A total of 67 core genes
423 had evidence of positive selection among sites (P < 0.01, FDR correction). We also tested for
424 positive selection on codon sets within the 3,691 accessory genes using the same approach. Of

425 the total, 895 displayed evidence of variable ω among sites (P < 0.01, FDR correction) and 201
426 yielded evidence of positive selection (P < 0.01, FDR correction). Finally, we applied the sites
427 models to the set of 187 multicopy genes, yielding another 33 genes with evidence for positive
428 selection. To sum, 5.3% (i.e., 67 of 1,257) of core genes, 5.4% (201/3,691) of accessory genes
429 and 17.6% of multicopy genes had significant evidence of at least one codon with an apparent
430 history of positive selection. Among the 201 accessory genes, four (*cya, group_454,*
431 *group_1057*, and *group_3542*) also had evidence for positive selection via the global test.
432

433 **DISCUSSION**

434 Host-pathogen interactions can drive rapid evolution of pathogenic bacteria, particularly
435 for genes involved in arms-race dynamics (28, 77). Here, we investigated the genomic evolution
436 of the plant pathogen, *X. fastidiosa*, through comparative genomic analysis of genomes that
437 represent diversity across the species, based on a sample set of 64 genomes. The sample was
438 isolated from 23 different plant hosts (Figure 1C) from throughout the world (Supplementary
439 Figure 1). With these data, we constructed a pangenome that contained 1,257 core genes and
440 9,220 accessory genes, similar to previous studies (40, 78). Of the core genes, the majority were,
441 as expected (79), involved in essential cellular processes -- such as translation, cell wall
442 biogenesis, and amino acid metabolism (Figure 1E). We used the set of core genes to infer a
443 maximum likelihood phylogeny, either with or without adjusting to putatively recombining
444 regions of the genome (Figure 2; Supplementary Figure 3). As with previous systematic
445 treatments of *X. fastidiosa* (19, 72, 80), both phylogenies identified three clades corresponding to
446 the three main subspecies (*fastidiosa, multiplex*, and *pauca*).

447 We employed both phylogenies to investigate the relationship between the *X. fastidiosa*
448 phylogeny and the plant host. The question of host specialization was first addressed using
449 phylogenetic approaches with multilocus sequencing typing (MLST) data. In this work, Sicard et
450 al. (2018) generated MLST data from 7 housekeeping genes from 50 *X. fastidiosa* genotypes.
451 After building a phylogeny, they tested coevolutionary relationships between host species and *X.*
452 *fastidiosa* MLST types but found no significant evidence of coevolution, implying a lack of host
453 specialization. This topic was recently revisited with full genome data (26, 27), but the results
454 were inconsistent between studies. Uceda-Campos et al. (2022) found no evidence that plant host
455 species clustered on their *X. fastidiosa* phylogeny, but the samples did cluster by geography. In
456 contrast, Kahn and Almeida (2022) inferred ancestral character states of plant hosts on the *X.*
457 *fastidiosa* phylogeny and were able to resolve the character state of some deep nodes. They
458 inferred, for example, that coffee plants were the ancestral host species for the node separating *X.*
459 *fastidiosa* ssp. *fastidiosa* from other subspecies. These patterns suggest that phylogenetic history
460 is associated with specific plant hosts and host ranges.
461 The disagreement among previous studies, and the fact that all such analyses are
462 properties of the sampled isolates, makes the issue worthy of further assessment. In our study, we
463 have found a significant, non-random association between phylogenetic relationships and both
464 the species and taxonomic order of plant hosts ($P < 0.0001$) based on core phylogenies. These
465 results are consistent with some level of specialization of *X. fastidiosa* to plant hosts and with the
466 recent analysis of Kahn and Almeida (2022). Moreover, these results were robust to phylogenetic
467 treatment – i.e., the inclusion or exclusion of genomic regions inferred to have histories of
468 recombination. Although it is difficult to quantitatively compare ANOSIM results across studies,
469 it is worth noting that the association of *X. fastidiosa* to plant order is similar in magnitude to the

470 association between a gut colonizing bacterium (*Bifidobacterium*) and the host species from
471 which it was isolated (50).
472 Given some evidence for host specialization, we hypothesized that it is driven in part by
473 accessory gene content. Under this hypothesis, we predicted an association between genes and
474 hosts should be as (or more) pronounced for accessory genes than for core genes. Instead, we
475 found no significant association between accessory gene complement and taxonomic order and
476 only a weak association with plant species. Our results are unlike, for example, the case of
477 bifidobacteria, where the association with host species was nearly as strong for accessory genes
478 as for host genes (50). We cannot be sure why we do not detect a signal for host specialization of
479 accessory genes, but we can think of three explanations. One is that that host associations, to the
480 extent they exist, are not driven by accessory genes but rather by evolutionary divergence in core
481 genes. Another is statistical power: because there are many more sequence changes among the
482 core genes than there are changes in accessory gene content, the distance matrix for core genes
483 likely has a higher signal-to-noise ratio than accessory gene content. Finally, if accessory genes
484 do mediate host shifts, it is possible – and even likely - that only a subset of accessory genes
485 drive these shifts. Under this scenario, there may be significant associations for a small subset of
486 accessory genes, but the signal of this association is weak across the entire accessory gene set.
487 This conjecture seems reasonable given that Kahn and Almeida (2002) found that the
488 presence/absence of a subset of only ~30 accessory genes correlated with the plant host. In
489 addition, it is worth emphasizing that *X. fastidiosa* interacts not only with plants but also insect
490 vectors and microbial communities, so that some subset of accessory genes likely contribute to
491 these interactions instead of those with plant hosts.
492

493 **The pattern of gene gain and loss events**

494 Another potential tool to study adaptation to specific hosts is by examining shifts in gene
495 composition through gene duplication, deletion, or HGT events (81, 82). We estimated the
496 number of gene loss and gain events along the core-gene phylogeny, and then normalized those
497 numbers relative to sequence divergence. Using this approach, we found that most branches
498 followed a consistent rate of gene gain or loss relative to sequence divergence. The fact that the
499 accessory gene phylogeny recapitulates the three subspecies (Figure 3) suggests, along with
500 previous evidence, that *X. fastidiosa* evolves predominantly through vertical inheritance and
501 intraspecific recombination, rather than HGT from other bacterial species (20, 83).

502 We have, however, identified 19 and 18 lineages with enriched gain or loss events,
503 respectively, and most of these branches were at the tip of the phylogeny. Again, a potential
504 explanation for these gain and loss dynamics is that they reflect host shifts. There are some
505 isolated examples that are consistent with this hypothesis. For example, isolates XF6c, Pr8x,
506 RAAR17, and OLS0478 in *pauca* have branches with enriched gene gains (Figure 4A). Two of
507 these (OLS0478 and Pr8x) were isolated from oleander and plum, respectively, and are the only
508 isolates associated with those plant hosts in their clades, suggesting a host shift. More globally,
509 however, the evidence for this hypothesis is unconvincing. When we, for example, contrast gene
510 gains between pairs of sister taxa with the same plant host, three of the 16 sister pairs had
511 enriched rates of gene gain. This proportion of enriched branches was not significantly lower
512 than the reminder of the tree ($P > 0.05$; Fisher's Exact Test), despite the fact that the sister taxa
513 did not experience a host shift. All of these inferences are of course dependent on our sample and
514 ignore the vector component of the *X. fastidiosa* lifecycle, so there are limitations to our

515 conclusions. At present, however, the evidence for an association between host shifts and
516 enhanced gene gain and loss events is weak.
517 This leaves unexplained the pattern of enriched rates of gene gain and loss at the tips of
518 the tree. We suspect this pattern is analogous to patterns of mutations in populations, as
519 suggested previously (76). New mutations begin as rare, low frequency variants in single
520 individuals. Eventually most of these mutations are removed by the processes of genetic drift and
521 natural selection, so that there are more new mutations in populations than old mutations. In a
522 phylogenetic context, these new mutations would be evident at the tip of the trees, so it may be
523 reasonable to expect higher effective rates of gene gain and loss in the 'newest' phylogenetic
524 branches. This explanation only has credence, however, if the observed gain and loss events are
525 both recent – i.e., newer than the sequence mutations that define the tip branches – and frequent.
526
527 **The identification of positively selected genes**
528 Many previous studies have implicated genes and their protein products in ongoing arms-
529 races between pathogens and their hosts (84, 85). One way to approach this question is agnostic
530 to function, which is to screen for genes with a history of positive selection. Ours is not the first
531 attempt to detect selection in *X. fastidiosa* genomes. Previous studies have searched for selection
532 by comparing levels of polymorphism or rates of synonymous and nonsynonymous mutations in
533 the core genome using Tajima's D and the McDonald-Kreitman test (24, 40). Other work has
534 measured ω in core genes but without statistically testing for positive selection (83) or by
535 applying the global test for ω for > 1.0 (25). To our knowledge no other study of *X. fastidiosa*
536 has tested for positive selection in accessory genes nor applied codon sites models. The set of
537 positively selected *X. fastidiosa* genes represents candidate pathogenicity factors to mediate

538 interactions with the environment, including plant host, insect vectors, or members of the
539 microbial community.

540 To study positive selection, we estimated ω, or the ratio of nonsynonymous to
541 synonymous mutations, for each core gene and for each accessory gene found in four or more
542 isolates. In total, this exercise encompassed 5,135 genes: 1,257 core genes, 3,691 accessory
543 genes, and 187 multicopy genes. We began by applying a global test that estimates ω over all
544 sites and phylogenetic lineages. This approach can be overly conservative, because a significant
545 test of $\omega > 1.0$ requires that positive selection is very strong, acts across many sites in a gene, is
546 present in most of the branches of the phylogeny, or all of the above. As expected, we found only
547 a few genes (eight accessory genes in total) that were significant for positive selection with this
548 test. Unfortunately, the annotations of 7 of 8 of these genes yielded few insights into their
549 functions. To explore gene function further, we identified protein domains using the Conserved
550 Domain Database. We found, for example, that the gene *group_7848* contains a VirB3 protein
551 domain, which is part of the Type IV secretory pathway and is commonly associated with the
552 membranes of the bacterial cell. The gene *cya* was also implicated using this test, which encodes
553 adenylate cyclase and plays essential roles in regulation of cellular metabolism (86).
554 Interestingly, the *cya* protein is involved in the cyclic AMP system, which is a global regulator in
555 gram-negative bacteria and has been shown to modulate gene expression in pathogenic bacteria
556 (87, 88).

557 The global test did allow, however, for two broad generalizations about patterns of
558 selection in *X. fastidiosa*. First, as a group the core genes are under strong purifying selection
559 with most (>90%) having ω estimates significantly < 1.0. Second, accessory genes generally
560 have lower levels of purifying selection, as evidenced by a lower proportion (45%) of significant

561 tests for $\omega < 1.0$ and by much higher average $\hat{\omega}$ values (0.21973 vs. 0.51443; Figure 5A). The
562 proportion of significant tests must be compared between genic sets with caution, because the
563 smaller sample sizes (n=4 to 59) for accessory genes likely reduce statistical power relative to
564 the minimum of 60 samples for all core genes, as do any differences in gene lengths.
565 Nonetheless, the contrasting pattern of ω is consistent with the ideas that core genes have
566 conserved biological functions and that accessory genes are more amenable to evolutionary
567 change due to their nonessential, but still potentially biological relevant, cellular roles (89).
568 Accessory genes may also experience higher variation in their selection dynamics because
569 recombination affects them more than core genes (83).
570 Given few signals of positive selection with the global test, we turned to codon site
571 models. To our surprise, the proportion of positively selected genes was similar for core genes
572 (5.3%) and accessory genes (5.4%). The salient question is whether these genes give some clue
573 to function. Of the 67 core genes with evidence for positive selection at the codon level, 40%
574 were unannotated. We performed a functional analysis by grouping the protein coding sequences
575 of these 67 core genes into COG categories to infer cellular functions. Excluding the category of
576 unknown function, the largest category was 'cell wall/membrane/envelope biogenesis,' followed
577 by the 'amino acid metabolism and transport,' 'carbohydrate metabolism and transport,'
578 'translation,' and 'intracellular trafficking and secretion' (Supplementary Figure 4A).
579 Of the 201 accessory genes with evidence for positive selection at the codon level, 82%
580 were not annotated for function. The remaining set of 36 genes was enriched for GO categories
581 related to protein secretion by the type IV secretion system (Supplementary Table 8). To better
582 infer function, we performed a COG analysis and found that the largest categories (excluding the
583 category of unknown function) were 'intracellular trafficking and secretion,' 'replication,

584 recombination and repair,' and 'secondary metabolites biosynthesis, transport and catabolism'
585 (Supplementary Figure 4B). Intriguingly, of this set of 201 genes, 50 overlapped with the set of
586 367 genes that had a gene-wide estimate of $\hat{\omega} > 1$. While these are especially strong candidates
587 for having a history of positive selection, a disappointing 94% of them were unannotated for
588 function. The three genes with annotations were: *cya*, *nagZ_2*, and *bacterial adaptive response A*
589 (*barA)*. The gene *nagZ_2* encodes a beta-glucosidase that is important for biofilm formation in
590 *Neisseria gonorrhoeae*, suggesting it could play a similar role in *X. fastidiosa*. It merits further
591 functional analysis, since biofilms are important to the infection cycle (90). *barA* encodes a
592 membrane associated histidine kinase that has a regulatory role in cell division, metabolism, and
593 pili formation, and it has been implicated in regulating the virulence response of uropathogenic
594 *E. coli* (91, 92). Finally, the multicopy genes also yielded evidence of positive selection,
595 including *cdiA1*, which is part of the secretory contact-dependent growth inhibition (CDI) system
596 that modulates biofilm formation in *Acinetobacter baumannii* (93).
597 As a final exercise, we cataloged the incidence of positive selection in a set of 35 genes
598 that have been listed as virulence and pathogenicity factors in *X. fastidiosa* (13). Of the 35, we
599 could identify 29 in our database based on the PD number annotation and reference sequence
600 (http://www.microbesonline.org/operons/gnc183190.html; Table 1). We expected that this set of
601 29 genes would be enriched for evidence of positive selection relative to the genomic
602 background, because these genes are putatively involved in arms-race interactions. The trend for
603 these genes was in the expected direction, because 4 of 29 (= 13.9%) were significant vs. 301 of
604 5,135 (= 5.8%) in the rest of the genome. However, the difference in proportions was not
605 significant (Fisher Exact Test, P = 0.1091). Nonetheless, this set of experimental genes is
606 interesting. All four genes with evidence of positive selection encode proteins associated with the

607 membrane of gram-negative bacteria and are involved in membrane transport or adhesin.
608 Specifically, the genes *fimF, xadA* and *xatA* encode proteins involved in fimbrial adhesion, non-
609 fimbrial adhesion, and biofilm formation, respectively, and the gene PD1311 encodes a protein
610 involved in membrane transport (94–98). Because there is a resolved protein structure for fimF
611 (99), we investigated the location of positively selected codons. Of the four positively selected
612 codons (N80, D87, F137, and D142), one (D87) was in a flexible loop and a second (D142)
613 comprised part of the second β-sheet of the protein (99). Together this suggests that changes in
614 the amino acid sequence of *fimF* may be impacting its function
615 We must caution that positive selection analyses are subject to false positives, and they
616 are also dependent on specific analysis features, like the sample set, the criteria for determining
617 homology, and the sequence alignments. We have nonetheless found several genes with some
618 evidence of positive selection that may also contribute to functions relevant to infection. We
619 believe they represent suitable candidates for further functional analyses to elucidate their role in
620 host-pathogen interactions and perhaps even host specificity.
621

622 **ACKNOWLEDGEMENTS**

623 We thank three anonymous reviewers for comments. We also thank R. Gaut, INRA-CFBP, and
624 the UC Irvine Genomics High-Throughput Facility for contributing to data generation. We also
625 thank J.J. Emerson, A. Martiny, E. Solares, and C.I. Rodriguez for helpful input on methodology.
626 T.N.B. was supported by the National Science Foundation Graduate Research Fellowship
627 Program and the University of California Irvine President's Dissertation Year Fellowship. This
628 work was supported by National Science Foundation grant no. 1741627 to B.S.G. and California

Department of Food and Agriculture agreement numbers 15-0218-SA and 18-0328-000-SA to M.C.R.

632 **REFERENCES**
633

634 1. Furuya EY, Lowy FD. 2006. Antimicrobial-resistant bacteria in the community setting. 1.
635 Nat Rev Microbiol 4:36–45.

636 2. Yacoubi BE, Brunings AM, Yuan Q, Shankar S, Gabriel DW. 2007. In Planta Horizontal
637 Transfer of a Major Pathogenicity Effector Gene. Appl Environ Microbiol 73:1612–1621.

638 3. Juhas M. 2015. Horizontal gene transfer in human pathogens. Crit Rev Microbiol 41:101–
639 108.

640 4. Chen NWG, Serres-Giardi L, Ruh M, Briand M, Bonneau S, Darrasse A, Barbe V,
641 Gagnevin L, Koebnik R, Jacques M-A. 2018. Horizontal gene transfer plays a major role in
642 the pathological convergence of Xanthomonas lineages on common bean. BMC Genomics
643 19:606.

644 5. Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R,
645 Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT,
646 Mobley HLT, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by
647 the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci
648 99:17020–17024.

649 6. Badet T, Croll D. 2020. The rise and fall of genes: origins and functions of plant pathogen
650 pangenomes. Curr Opin Plant Biol 56:65–73.

651 7. Kim Y, Gu C, Kim HU, Lee SY. 2020. Current status of pan-genome analysis for
652 pathogenic bacteria. Curr Opin Biotechnol 63:54–62.

653 8. Sicard A, Zeilinger AR, Vanhove M, Schartel TE, Beal DJ, Daugherty MP, Almeida RPP.
654 2018. Xylella fastidiosa: Insights into an Emerging Plant Pathogen. Annu Rev Phytopathol
655 56:181–202.

656 9. Burbank LP, Roper MC 2021. 2021. Microbe Profile: Xylella fastidiosa – a devastating
657 agricultural pathogen with an endophytic lifestyle. Microbiology 167:001091.

658 10. Schuenzel EL, Scally M, Stouthamer R, Nunney L. 2005. A Multigene Phylogenetic Study
659 of Clonal Diversity and Divergence in North American Strains of the Plant Pathogen
660 Xylella fastidiosa. Appl Environ Microbiol 71:3832–3839.

661 11. Loconsole G, Saponari M, Boscia D, D'Attoma G, Morelli M, Martelli GP, Almeida RPP.
662 2016. Intercepted isolates of Xylella fastidiosa in Europe reveal novel genetic diversity. Eur
663 J Plant Pathol 146:85–94.

664 12. Chatterjee S, Almeida RPP, Lindow S. 2008. Living in two Worlds: The Plant and Insect
665 Lifestyles of Xylella fastidiosa. Annu Rev Phytopathol 46:243–271.

666 13. Rapicavoli J, Ingel B, Blanco-Ulate B, Cantu D, Roper C. 2018. Xylella fastidiosa: an
667 examination of a re-emerging plant pathogen. Mol Plant Pathol 19:786–800.

668 14. Tumber K, Alston J, Fuller K. 2014. Pierce's disease costs California $104 million per year.
669 Calif Agric 68:20–29.

670 15. 2015. Assessing the returns to R&D on perennial crops: the costs and benefits of Pierce's
671 disease research in the California winegrape industry. Aust J Agric Resour Econ
672 https://doi.org/10.22004/ag.econ.280230.

673 16. Koo H, Allan RN, Howlin RP, Stoodley P, Hall-Stoodley L. 2017. Targeting microbial
674 biofilms: current and prospective therapeutic strategies. 12. Nat Rev Microbiol 15:740–755.

675 17. Castro C, DiSalvo B, Roper MC. 2021. Xylella fastidiosa: A reemerging plant pathogen
676 that threatens crops globally. PLOS Pathog 17:e1009813.

677 18. Roper C, Lindow SE. 2016. CHAPTER 16: Xylella fastidiosa: Insights into the Lifestyle of
678 a Xylem-Limited Bacterium, p. 307–320. *In* Caroline, R, Steven, EL (eds.), Virulence
679 Mechanisms of Plant-Pathogenic Bacteria. The American Phytopathological Society.

680 19. Marcelletti S, Scortichini M. 2016. Genome-wide comparison and taxonomic relatedness of
681 multiple Xylella fastidiosa strains reveal the occurrence of three subspecies and a new
682 Xylella species. Arch Microbiol 198:803–812.

683 20. Nunney L, Vickerman DB, Bromley RE, Russell SA, Hartman JR, Morano LD, Stouthamer
684 R. 2013. Recent Evolutionary Radiation and Host Plant Specialization in the Xylella
685 fastidiosa Subspecies Native to the United States. Appl Environ Microbiol 79:2189–2200.

686 21. Almeida RPP, Purcell AH. 2003. Biological Traits of Xylella fastidiosa Strains from Grapes
687 and Almonds. Appl Environ Microbiol 69:7447–7452.

688 22. Hernandez-Martinez R, Costa HS, Dumenyo CK, Cooksey DA. 2006. Differentiation of
689 Strains of Xylella fastidiosa Infecting Grape, Almonds, and Oleander Using a Multiprimer
690 PCR Assay. Plant Dis 90:1382–1388.

691 23. Almeida RPP, Nascimento FE, Chau J, Prado SS, Tsai C-W, Lopes SA, Lopes JRS. 2008.
692 Genetic Structure and Biology of Xylella fastidiosa Strains Causing Disease in Citrus and
693 Coffee in Brazil. Appl Environ Microbiol 74:3690–3701.

694 24. Castillo AI, Bojanini I, Chen H, Kandel PP, De La Fuente L, Almeida RPP. 2021.
695 Allopatric Plant Pathogen Population Divergence following Disease Emergence. Appl
696 Environ Microbiol 87:e02095-20.

697 25. Vanhove M, Sicard A, Ezennia J, Leviten N, Almeida RPP. 2020. Population structure and
698 adaptation of a bacterial pathogen in California grapevines. Environ Microbiol 22:2625–
699 2638.

700 26. Uceda-Campos G, Feitosa-Junior OR, Santiago CRN, Pierry PM, Zaini PA, de Santana
701 WO, Martins-Junior J, Barbosa D, Digiampietri LA, Setubal JC, da Silva AM. 2022.
702 Comparative Genomics of Xylella fastidiosa Explores Candidate Host-Specificity
703 Determinants and Expands the Known Repertoire of Mobile Genetic Elements and
704 Immunity Systems. 5. Microorganisms 10:914.

705 27. Kahn AK, Almeida RPP. 2022. Phylogenetics of Historical Host Switches in a Bacterial
706 Plant Pathogen. Appl Environ Microbiol 88:e02356-21.

707 28. Daugherty MD, Malik HS. 2012. Rules of Engagement: Molecular Insights from Host-
708 Virus Arms Races. Annu Rev Genet 46:677–700.

709 29. Aleru O, Barber MF. 2020. Battlefronts of evolutionary conflict between bacteria and
710 animal hosts. PLOS Pathog 16:e1008797.

711 30. Mitchell PS, Patzina C, Emerman M, Haller O, Malik HS, Kochs G. 2012. Evolution-
712 Guided Identification of Antiviral Specificity Determinants in the Broadly Acting
713 Interferon-Induced Innate Immunity Factor MxA. Cell Host Microbe 12:598–604.

714 31. Ng M, Ndungo E, Kaczmarek ME, Herbert AS, Binger T, Kuehne AI, Jangra RK, Hawkins
715 JA, Gifford RJ, Biswas R, Demogines A, James RM, Yu M, Brummelkamp TR, Drosten C,
716 Wang L-F, Kuhn JH, Müller MA, Dye JM, Sawyer SL, Chandran K. 2015. Filovirus
717 receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats.
718 eLife 4:e11785.

719 32. Daugherty MD, Schaller AM, Geballe AP, Malik HS. 2016. Evolution-guided functional
720 analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. eLife.
721 eLife Sciences Publications Limited. https://elifesciences.org/articles/14228. Retrieved 2
722 April 2022.

723 33. Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.
724 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

725 34. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
726 sequence data. Bioinformatics 30:2114–2120.

727 35. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu:
728 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat
729 separation. Genome Res gr.215087.116.

730 36. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome
731 assemblies from short and long sequencing reads. PLOS Comput Biol 13:e1005595.

732 37. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for
733 genome assemblies. Bioinforma Oxf Engl 29:1072–1075.

734 38. Chase AB, Gomez-Lunar Z, Lopez AE, Li J, Allison SD, Martiny AC, Martiny JBH. 2018.
735 Emergence of soil bacterial ecotypes along a climate gradient. Environ Microbiol 20:4112–
736 4126.

737 39. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for
738 FASTA/Q File Manipulation. PLOS ONE 11:e0163962.

739 40. Castillo AI, Chacón-Díaz C, Rodríguez-Murillo N, Coletta-Filho HD, Almeida RPP. 2020.
740 Impacts of local population history and ecology on the evolution of a globally dispersed
741 pathogen. BMC Genomics 21:369.

742 41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
743 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,
744 Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its
745 Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

746 42. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–
747 2069.

748 43. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D,
749 Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis.
750 Bioinformatics 31:3691–3693.

751 44. Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use
752 in Phylogenetic Analysis. Mol Biol Evol 17:540–552.

753 45. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
754 sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.

755 46. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
756 Improvements in Performance and Usability. Mol Biol Evol 30:772–780.

757 47. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis
758 of large phylogenies. Bioinformatics 30:1312–1313.

759 48. Boc A, Diallo AB, Makarenkov V. 2012. T-REX: a web server for inferring, validating and
760 visualizing phylogenetic trees and networks. Nucleic Acids Res 40:W573–W579.

761 49. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM:
762 assessing the quality of microbial genomes recovered from isolates, single cells, and
763 metagenomes. Genome Res 25:1043–1055.

764 50. Rodriguez CI, Martiny JBH. 2020. Evolutionary relationships among bifidobacteria and
765 their hosts and environments. BMC Genomics 21:26.

766 51. Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN,
767 Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools
768 APIs in 2019. Nucleic Acids Res 47:W636–W641.

769 52. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P.
770 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by
771 eggNOG-Mapper. Mol Biol Evol 34:2115–2122.

772 53. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende
773 DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a
774 hierarchical, functionally and phylogenetically annotated orthology resource based on 5090
775 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314.

776 54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K,
777 Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese
778 JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the
779 unification of biology. Nat Genet 25:25–29.

780 55. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and
781 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol
782 Evol 32:268–274.

783 56. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
784 ModelFinder: fast model selection for accurate phylogenetic estimates. 6. Nat Methods
785 14:587–589.

786 57. Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
787 evolutionary analyses in R. Bioinformatics 35:526–528.

788 58. R Core Team. 2019. R: A language and environment for statistical computing. R
789 Foundation for Statistical Computing. https://www.R-project.org/.

790 59. Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin
791 PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2020.
792 vegan: Community Ecology Package.

793 60. Kung SH, Almeida RPP. 2011. Natural Competence and Recombination in the Plant
794 Pathogen Xylella fastidiosa. Appl Environ Microbiol 77:5278–5284.

795 61. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris
796 SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole
797 genome sequences using Gubbins. Nucleic Acids Res 43:e15.

798 62. Shikov AE, Malovichko YV, Nizhnikov AA, Antonets KS. 2022. Current Methods for
799 Recombination Detection in Bacteria. 11. Int J Mol Sci 23:6257.

800 63. Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other
801 things). Methods Ecol Evol 3:217–223.

802 64. Mateo-Estrada V, Graña-Miraglia L, López-Leal G, Castillo-Ramírez S. 2019.
803 Phylogenomics Reveals Clear Cases of Misclassification and Genus-Wide Phylogenetic
804 Markers for Acinetobacter. Genome Biol Evol 11:2531–2541.

805 65. Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics
806 26:1372–1373.

807 66. Löytynoja A. 2014. Phylogeny-aware alignment with PRANK, p. 155–170. *In* Russell, DJ
808 (ed.), Multiple Sequence Alignment Methods. Humana Press, Totowa, NJ.

809 67. Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

810 68. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss
811 mapping engine. Bioinformatics 26:2914–2915.

812 69. Avram O, Rapoport D, Portugez S, Pupko T. 2019. M1CR0B1AL1Z3R—a user-friendly
813 web server for the analysis of large-scale microbial genomics data. Nucleic Acids Res
814 47:W88–W92.

815 70. Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum
816 likelihood. Bioinformatics 13:555–556.

817 71. Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol
818 24:1586–1591.

819 72. Denancé N, Briand M, Gaborieau R, Gaillard S, Jacques M-A. 2019. Identification of
820 genetic relationships and subspecies signatures in Xylella fastidiosa. BMC Genomics
821 20:239.

822 73. Bolotin E, Hershberg R. 2015. Gene Loss Dominates As a Source of Genetic Variation
823 within Clonal Pathogenic Bacterial Species. Genome Biol Evol 7:2173–2187.

824 74. Iranzo J, Wolf YI, Koonin EV, Sela I. 2019. Gene gain and loss push prokaryotes beyond
825 the homologous recombination barrier and accelerate genome sequence divergence. Nat
826 Commun 10:5376.

827 75. Firrao G, Scortichini M, Pagliari L. 2021. Orthology-Based Estimate of the Contribution of
828 Horizontal Gene Transfer from Distantly Related Bacteria to the Intraspecific Diversity and
829 Differentiation of Xylella fastidiosa. 1. Pathogens 10:46.

830 76. Graña-Miraglia L, Lozano LF, Velázquez C, Volkow-Fernández P, Pérez-Oseguera Á,
831 Cevallos MA, Castillo-Ramírez S. 2017. Rapid Gene Turnover as a Significant Source of
832 Genetic Variation in a Recently Seeded Population of a Healthcare-Associated Pathogen.
833 Front Microbiol 8.

834 77. Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host–pathogen
835 interactions from mammalian sequence data. 4. Nat Rev Genet 16:224–236.

836 78. Giampetruzzi A, Saponari M, Loconsole G, Boscia D, Savino VN, Almeida RPP, Zicca S,
837 Landa BB, Chacón-Diaz C, Saldarelli P. 2017. Genome-Wide Analysis Provides Evidence
838 on the Genetic Relatedness of the Emergent Xylella fastidiosa Genotype in Italy to Isolates
839 from Central America. Phytopathology® 107:816–827.

840 79. Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-
841 genome. Curr Opin Microbiol 11:472–477.

842 80. Yuan X, Morano L, Bromley R, Spring-Pearson S, Stouthamer R, Nunney L. 2010.
843 Multilocus sequence typing of Xylella fastidiosa causing Pierce's disease and oleander leaf
844 scorch in the United States. Phytopathology 100:601–611.

845 81. Hurles M. 2004. Gene Duplication: The Genomic Trade in Spare Parts. PLOS Biol 2:e206.

846 82. Arnold BJ, Huang I-T, Hanage WP. 2022. Horizontal gene transfer and adaptive evolution
847 in bacteria. 4. Nat Rev Microbiol 20:206–218.

848 83. Castillo AI, Almeida RPP. 2021. Evidence of gene nucleotide composition favoring
849 replication and growth in a fastidious plant pathogen. G3 GenesGenomesGenetics
850 11:jkab076.

851 84. Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB, Anderson JP,
852 Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB. 2010. Plants versus pathogens:
853 an evolutionary arms race. Funct Plant Biol 37:499–512.

854 85. Schulte RD, Makus C, Hasert B, Michiels NK, Schulenburg H. 2010. Multiple reciprocal
855 adaptations and rapid genetic change upon experimental coevolution of an animal host and
856 its microbial parasite. Proc Natl Acad Sci 107:7359–7364.

857 86. Danchin A, Guiso N, Roy A, Ullmann A. 1984. Identification of the Escherichia coli cya
858 gene product as authentic adenylate cyclase. J Mol Biol 175:403–408.

859 87. Smith RS, Wolfgang MC, Lory S. 2004. An Adenylate Cyclase-Controlled Signaling
860 Network Regulates Pseudomonas aeruginosa Virulence in a Mouse Model of Acute
861 Pneumonia. Infect Immun 72:1677–1684.

862 88. Kim YR, Kim SY, Kim CM, Lee SE, Rhee JH. 2005. Essential role of an adenylate cyclase
863 in regulating Vibrio vulnificus virulence. FEMS Microbiol Lett 243:497–503.

864 89. Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, Thomson NR.
865 2021. Different evolutionary trends form the twilight zone of the bacterial pan-genome.
866 Microb Genomics 7:000670.

867 90. Bhoopalan SV, Piekarowicz A, Lenz JD, Dillard JP, Stein DC. 2016. nagZ Triggers
868 Gonococcal Biofilm Disassembly. 1. Sci Rep 6:22372.

869 91. Palaniyandi S, Mitra A, Herren CD, Lockatell CV, Johnson DE, Zhu X, Mukhopadhyay S.
870 2012. BarA-UvrY Two-Component System Regulates Virulence of Uropathogenic E. coli
871 CFT073. PLOS ONE 7:e31348.

872 92. Sahu SN, Acharya S, Tuminaro H, Patel I, Dudley K, LeClerc JE, Cebula TA,
873 Mukhopadhyay S. 2003. The bacterial adaptive response gene, barA, encodes a novel
874 conserved histidine kinase regulatory switch for adaptation and modulation of metabolism
875 in Escherichia coli. Mol Cell Biochem 253:167–177.

876 93. Roussin M, Rabarioelina S, Cluzeau L, Cayron J, Lesterlin C, Salcedo SP, Bigot S. 2019.
877 Identification of a Contact-Dependent Growth Inhibition (CDI) System That Reduces
878 Biofilm Formation and Host Cell Adhesion of Acinetobacter baumannii DSM30011 Strain.
879 Front Microbiol 10:2450.

880 94. Ma Rodriguez A, Olano C, Vilches C, Méndez C, Salas JA. 1993. Streptomyces
881 antibioticus contains at least three oleandomycin-resistance determinants, one of which
882 shows similarity with proteins of the ABC-transporter superfamily. Mol Microbiol 8:571–
883 582.

884 95. Sun QH, Hu J, Huang GX, Ge C, Fang RX, He CZ. 2005. Type-II secretion pathway
885 structural gene xpsE, xylanase- and cellulase secretion and virulence in Xanthomonas
886 oryzae pv. oryzae. Plant Pathol 54:15–21.

887 96. Abbas A, Adams C, Scully N, Glennon J, O'Gara F. 2007. A role for TonB1 in biofilm
888 formation and quorum sensing in Pseudomonas aeruginosa. FEMS Microbiol Lett 274:269–
889 278.

890 97. Das A, Rangaraj N, Sonti RV. 2009. Multiple Adhesin-Like Functions of *Xanthomonas*
891 *oryzae* pv. *oryzae* Are Involved in Promoting Leaf Attachment, Entry, and Virulence on
892 Rice. Mol Plant-Microbe Interactions® 22:73–85.

893 98. Zeiner SA, Dwyer BE, Clegg S. 2012. FimA, FimF, and FimH Are Necessary for Assembly
894 of Type 1 Fimbriae on Salmonella enterica Serovar Typhimurium. Infect Immun 80:3289–
895 3296.

896 99. Gossert AD, Bettendorff P, Puorger C, Vetsch M, Herrmann T, Glockshuber R, Wüthrich
897 K. 2008. NMR Structure of the Escherichia coli Type 1 Pilus Subunit FimF and Its
898 Interactions with Other Pilus Subunits. J Mol Biol 375:752–763.

899
900

901 **Table 1:** *Codeml* results for experimentally identified virulence and pathogenicity genes, as
902 listed (13)
903

PD Number	Gene Name	Pan-genome classification	No. Genomes[1]	(M0)[2]	M2a vs. M1a p-value[3]
PD0058	*fimF*	Accessory	41	0.31555	**3.25E-08**
PD0062	*fimA*	Accessory	26	0.81255	0.247
PD0233	*rpfB*	Accessory	57	0.16832	1
PD0279	*cgsA*	Core	64	0.14404	1
PD0406	*rpfC*	Accessory	44	0.34502	1
PD0528	*xatA*	Core	64	0.43097	**1.38E-41**
PD0731	*xadA*	Accessory	58	0.39196	**0.004**
PD0732	*xpsE*	Core	64	0.05825	1
PD0814	*wzy*	Accessory	43	0.17675	1
PD0843	*tonB1*	Core	64	0.11374	0.534
PD0848	*pilL*	Core	64	0.18195	1
PD0986		Core	64	0.10828	1
PD1099	*dinJ/relE*	Accessory	25	0.10271	1
PD1100		Accessory	15	0.20708	0.731
PD1284	*algU*	Core	64	0.19261	1
PD1311		Accessory	33	0.42541	**3.47E-05**
PD1380	*csp1*	Core	64	0.15702	1
PD1391	*gumH*	Accessory	46	0.12964	1
PD1394	*gumD*	Core	63	0.11504	1
PD1485	*pglA*	Accessory	59	0.28401	0.114
PD1678	*phoQ*	Core	64	0.1086	1
PD1679	*phoP*	Core	64	0.03272	1
PD1703	*lesA/lipA*	Core	64	0.06614	1
PD1792	*hxfB*	Core	64	0.10828	1
PD1826	*chiA*	Core	64	0.11424	1
PD1856	*engXCA1*	Core	63	0.24034	1
PD1964	*tolC*	Core	64	0.10051	1
PD1984	*gacA*	Core	64	0.13444	1
PD2118	*hxfA*	Core	64	0.10828	1

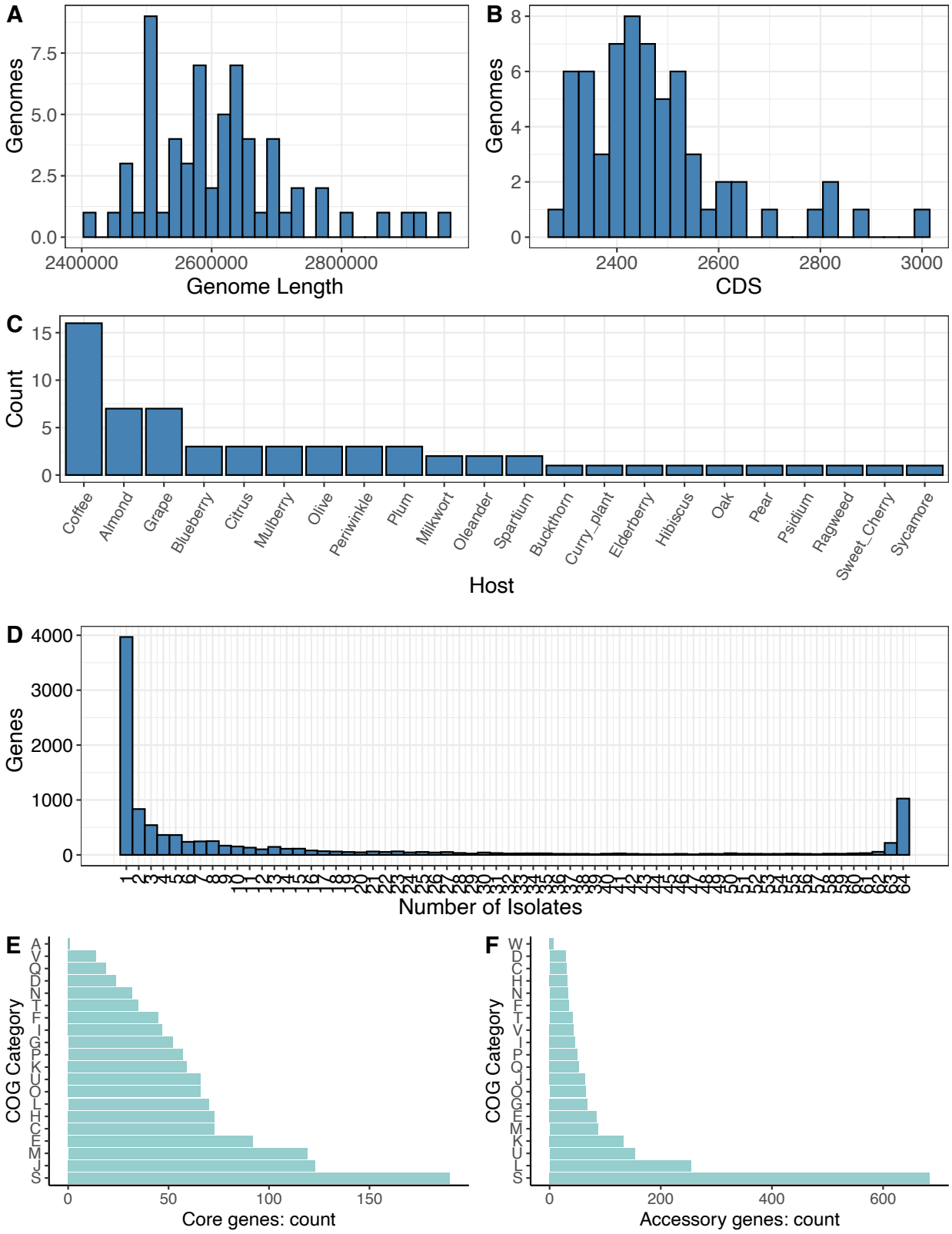
904 [1] the number of genomes, out of 64, in which the gene was detected.
905 [2] M0 estimates a single ω across the entire phylogeny of sequences.

906 [3] The p-value of tests after FDR correction. Bolded values are significant at $p < 0.01$. The
907 notation E refers to the power of ten.

908 **FIGURE LEGENDS**
909
910 **Figure 1:** Histograms reporting the characteristics of the 64 *Xylella* genomes. A) Genome
911 lengths, exhibited in base pairs. B) The number of genes within a genome. C) A histogram of the
912 plant species from which genomes were isolated. D) A histogram of the number of genes found
913 in *x* number of genomes; this histogram shows, for example, that nearly 4,000 genes are found in
914 only of one the genomes out of the entire sample of 64 genomes and that 1,024 genes are found
915 in all 64 genomes; E) the distribution of functional categories for the set of 1,257 core genes and
916 F) the distribution of functional categories for the set of 9,220 accessory genes. A key to the
917 COG categories for panels E) and F) is in Figure S4.
918
919 **Figure 2:** The inferred phylogeny of 64 *Xylella* genomes, based on maximum likelihood
920 inference of core gene alignments. Each isolate is labelled at the tips and is colored according to
921 the order of the plant isolation source (host). The common name of the host is provided to the
922 right of order information. The three *X. fastidiosa* subspecies are indicated, as are bootstrap
923 values at each node. The bootstrap values are pie charts, where black represents the percent of
924 bootstrap support, and the scale bar reflects the magnitude of sequence divergence per nucleotide
925 site.
926
927 **Figure 3:** A comparison of a NJ tree based on distances due to gene presence / absence (on the
928 left) to the likelihood tree based on the core gene alignments (from Figure 2, on the right). As in
929 Figure 2, the isolates are labelled at the tips of trees, with the colors representing plant order.
930 Both phylogenies contain three main *X. fastidiosa* clades, representing the three subspecies.
931 Lines connect the same isolate between the two trees, with angled lines representing topological

932 discordance between phylogenies. The three *Xylella* subspecies are outlined in a black box and
933 labelled.
934
935 **Figure 4:** The result of gene gain and loss analyses. A) The phylogeny of the isolates, with
936 branch lengths proportional to the number of gene gain events. The colored branches are
937 branches with outlier gene gain rates. B) The phylogeny of the isolates, with branch lengths
938 proportional to the number of gene loss events. The colored branches are branches with outlier
939 gene loss rates. C) A plot of the gene gains against sequence divergence; in the plot each dot
940 represents one of the 125 branches on the phylogeny. Outlier dots are colored red. D) As in C,
941 with gene losses plotted again sequence divergence.
942
943 **Figure 5:** Estimated values of ω under M0 (the one-ratio model) in the core and accessory genes.
944 The distribution of $\widehat{\omega}$ values is plotted for A) core genes estimated with gene trees, B) accessory
945 genes with gene trees C) core genes estimated with the core gene alignment (CGA) phylogeny,
946 and D) multicopy genes with gene trees. Histogram bars are shaded to reflect the outcome of the
947 likelihood ratio test (non-significant tests are colored red and significant tests are colored blue)
948 between a model that estimated $\widehat{\omega}$ and a model with ω fixed to 1.0. The horizontal dashed line
949 denotes $\widehat{\omega}$ for each gene set.
950

951 **Figure 1:**
952



953
954

Figure 2:

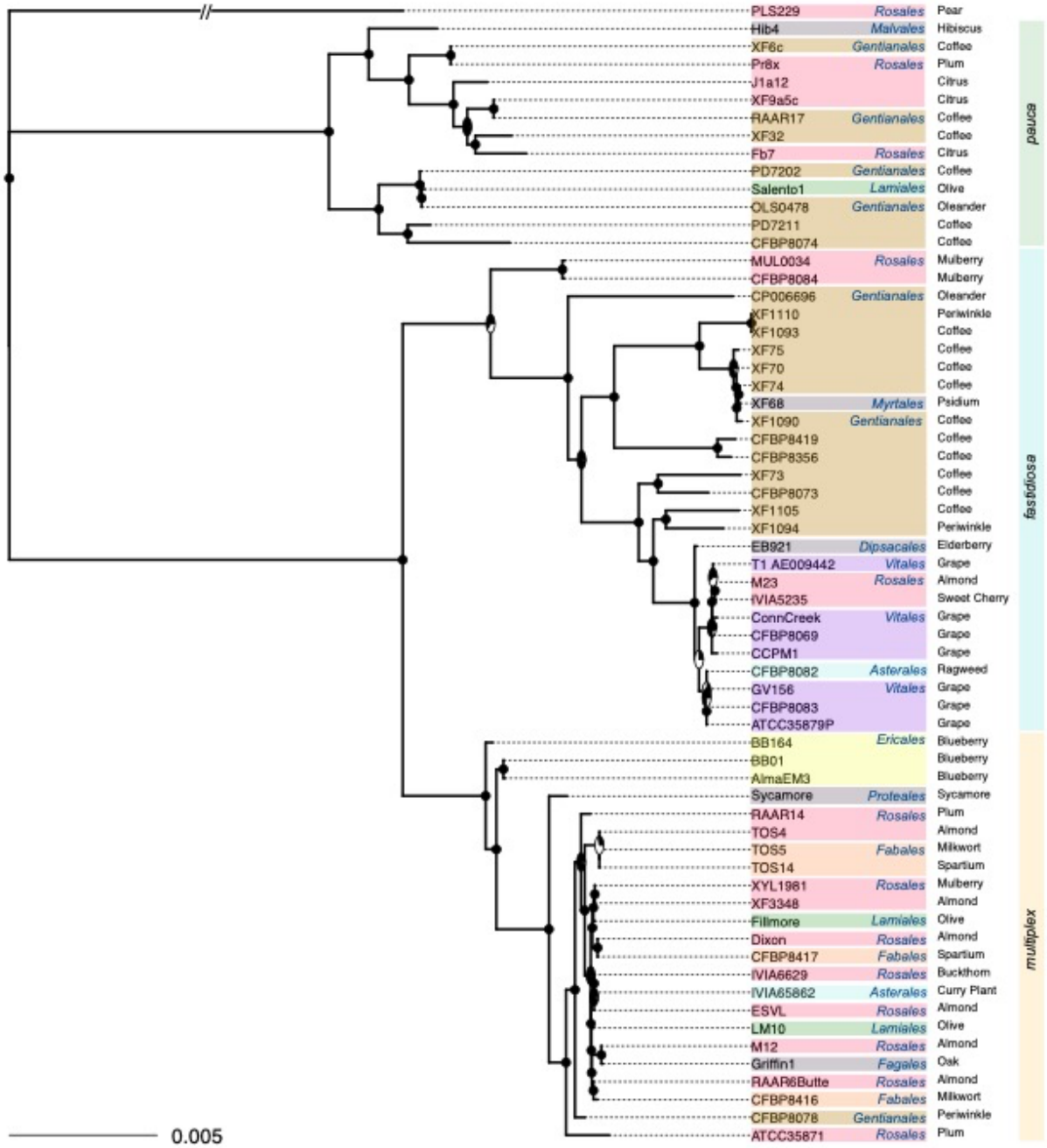

PLS229 Rosales Pear
Hib4 Malvales Hibiscus
XF6c Gentianales Coffee
Pr8x Rosales Plum
J1a12 Citrus
XF9a5c Citrus
RAAR17 Gentianales Coffee
XF32 Coffee
Fb7 Rosales Citrus
PD7202 Gentianales Coffee
Salento1 Lamiales Olive
OLS0478 Gentianales Oleander
PD7211 Coffee
CFBP8074 Coffee
MUL0034 Rosales Mulberry
CFBP8084 Mulberry
CP006696 Gentianales Oleander
XF1110 Periwinkle
XF1093 Coffee
XF75 Coffee
XF70 Coffee
XF74 Coffee
XF68 Myrtales Psidium
XF1090 Gentianales Coffee
CFBP8419 Coffee
CFBP8356 Coffee
XF73 Coffee
CFBP8073 Coffee
XF1105 Coffee
XF1094 Periwinkle
EB921 Dipsacales Elderberry
T1 AE009442 Vitales Grape
M23 Rosales Almond
IVIA5235 Sweet Cherry
ConnCreek Vitales Grape
CFBP8069 Grape
CCPM1 Grape
CFBP8082 Asterales Ragweed
GV156 Vitales Grape
CFBP8083 Grape
ATCC35879P Grape
BB164 Ericales Blueberry
BB01 Blueberry
AlmaEM3 Blueberry
Sycamore Proteales Sycamore
RAAR14 Rosales Plum
TOS4 Almond
TOS5 Fabales Milkwort
TOS14 Spartium
XYL1981 Rosales Mulberry
XF3348 Almond
Fillmore Lamiales Olive
Dixon Rosales Almond
CFBP8417 Fabales Spartium
IVIA6629 Rosales Buckthorn
IVIA65862 Asterales Curry Plant
ESVL Rosales Almond
LM10 Lamiales Olive
M12 Rosales Almond
Griffin1 Fagales Oak
RAAR6Butte Rosales Almond
CFBP8416 Fabales Milkwort
CFBP8078 Gentianales Periwinkle
ATCC35871 Rosales Plum
pauca
fastidiosa
multiplex
0.005

Figure 3:

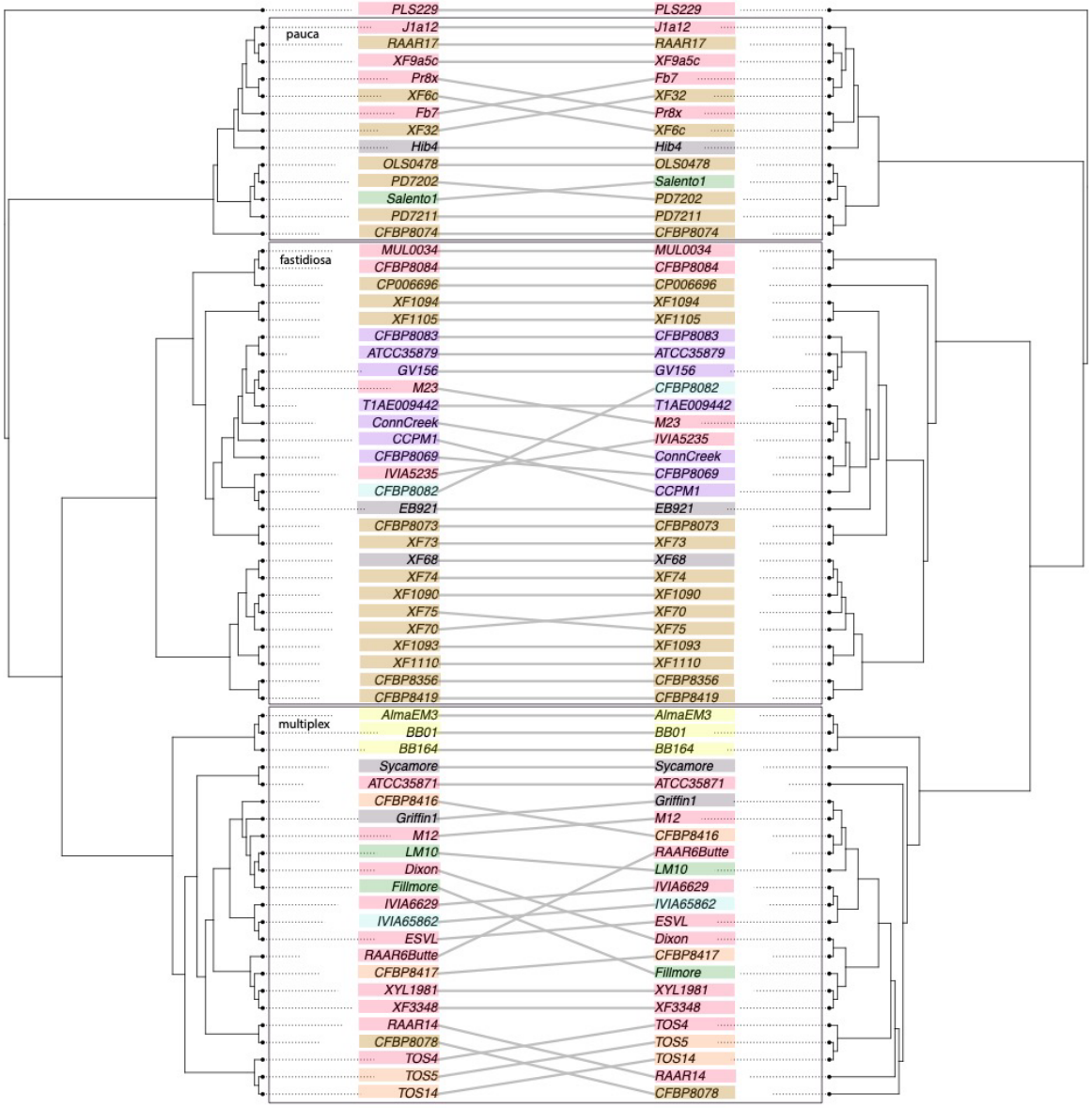
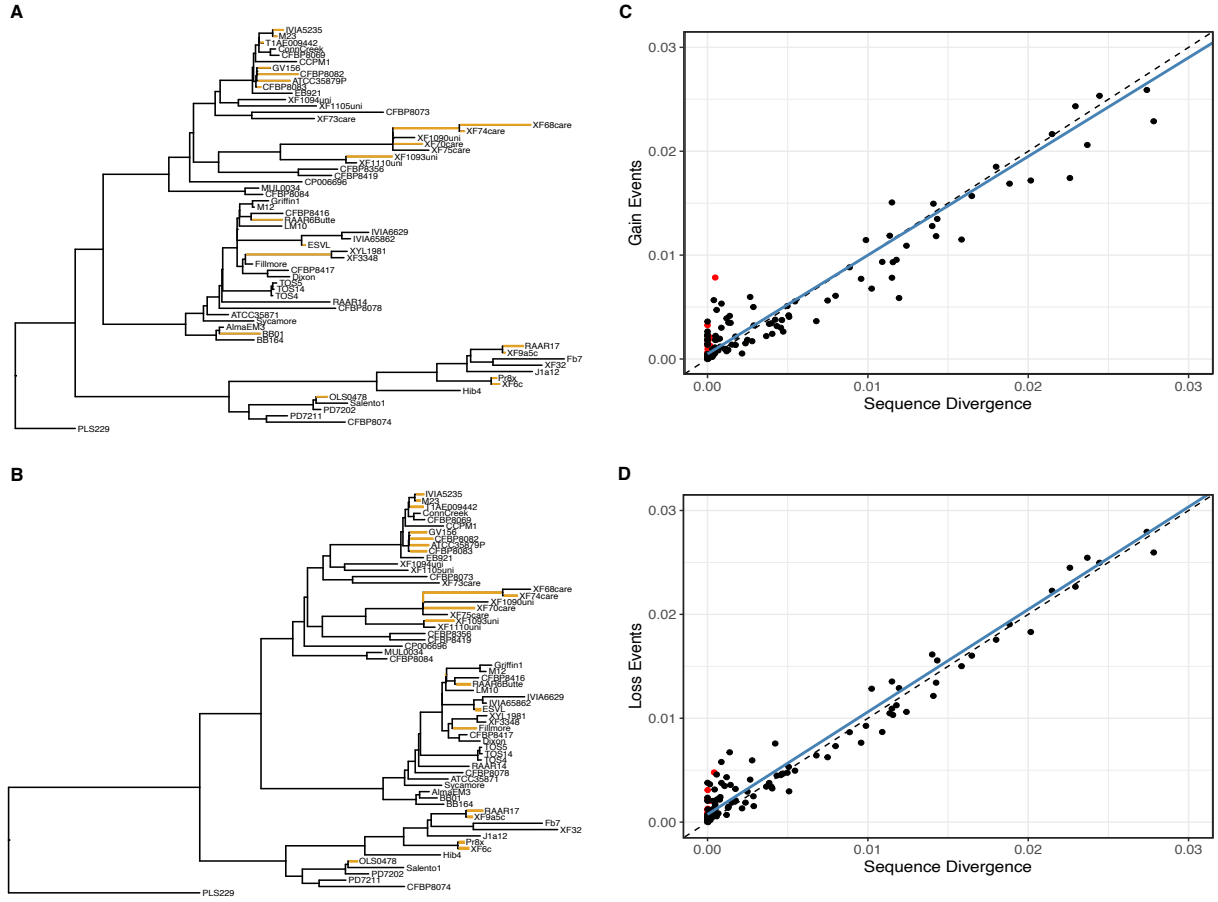
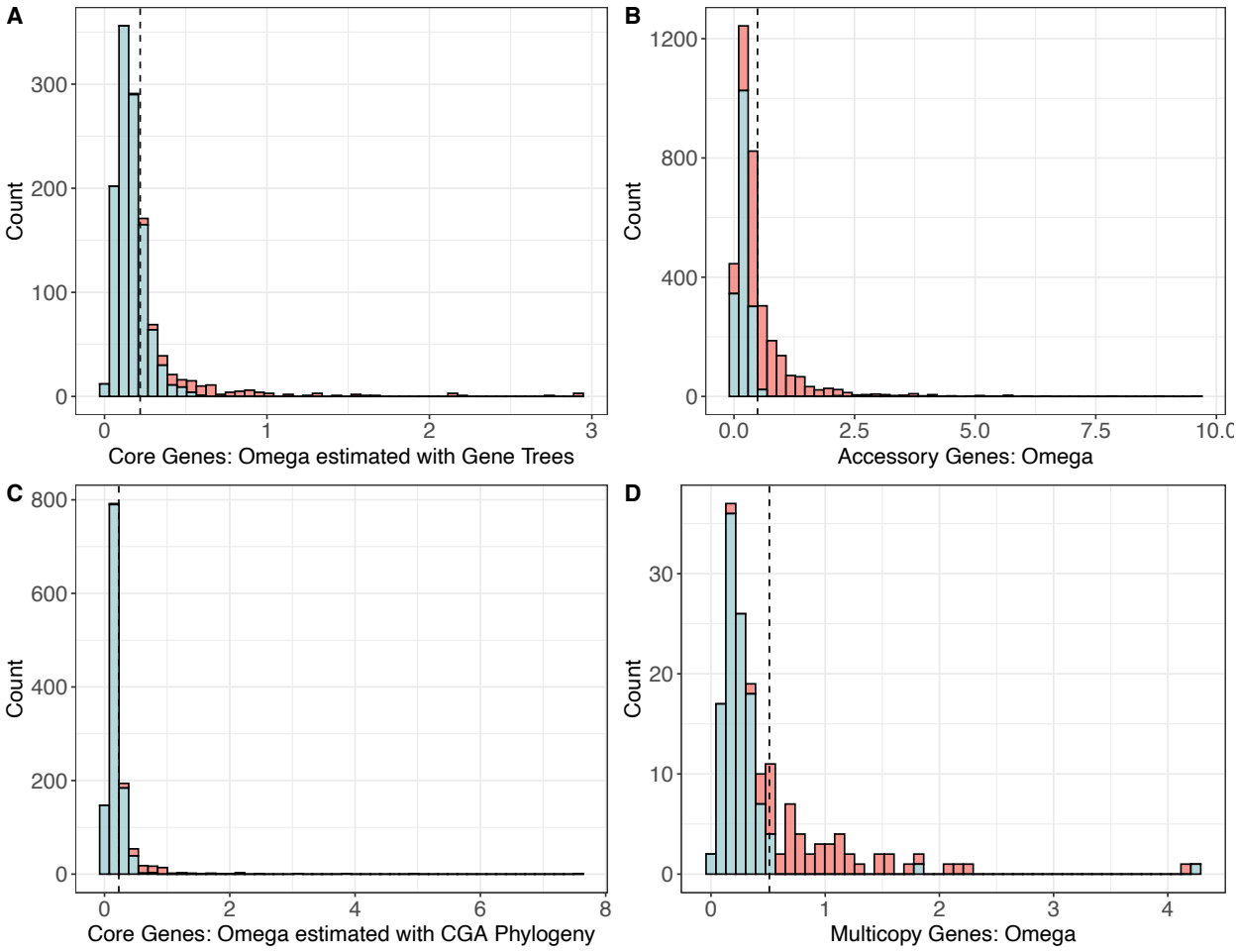

pauca
fastidiosa
multiplex
PLS229
J1a12
RAAR17
XF9a5c
Pr8x
XF6c
Fb7
XF32
Hib4
OLS0478
PD7202
Salento1
PD7211
CFBP8074
MUL0034
CFBP8084
CP006696
XF1094
XF1105
CFBP8083
ATCC35879
GV156
M23
T1AE009442
ConnCreek
CCPM1
CFBP8069
IVIA5235
CFBP8082
EB921
CFBP8073
XF73
XF68
XF74
XF1090
XF75
XF70
XF1093
XF1110
CFBP8356
CFBP8419
AlmaEM3
BB01
BB164
Sycamore
ATCC35871
CFBP8416
Griffin1
M12
LM10
Dixon
Fillmore
IVIA6629
IVIA65862
ESVL
RAAR6Butte
CFBP8417
XYL1981
XF3348
RAAR14
CFBP8078
TOS4
TOS5
TOS14
PLS229
J1a12
RAAR17
XF9a5c
Fb7
XF32
Pr8x
XF6c
Hib4
OLS0478
Salento1
PD7202
PD7211
CFBP8074
MUL0034
CFBP8084
CP006696
XF1094
XF1105
CFBP8083
ATCC35879
GV156
CFBP8082
T1AE009442
M23
IVIA5235
ConnCreek
CFBP8069
CCPM1
EB921
CFBP8073
XF73
XF68
XF74
XF1090
XF70
XF75
XF1093
XF1110
CFBP8356
CFBP8419
AlmaEM3
BB01
BB164
Sycamore
ATCC35871
Griffin1
M12
CFBP8416
RAAR6Butte
LM10
IVIA6629
IVIA65862
ESVL
Dixon
CFBP8417
Fillmore
XYL1981
XF3348
TOS4
TOS5
TOS14
RAAR14
CFBP8078

Figure 4:


A
B
C
D
Gain Events
Loss Events
Sequence Divergence
0.00
0.01
0.02
0.03

975 **Figure 5:**



976