Using genomes and evolutionary analyses to screen for host-specificity and arms-race dynamics in the plant pathogen *Xylella 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. It is not clear, however, whether *X. fastidiosa* isolates display specialization to plant hosts and, if so, what genetic determinants underlie the associations. 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. We first investigated evolutionary relationships among isolates, recapitulating the three recognized *X. fastidiosa* subspecies based on information from both core and accessory genes. We used these phylogenies as the basis to test associations between genetic divergence and host species, yielding significant evidence for specialization of *X. fastidiosa* to plant host species for core, but not accessory, genes. This observation places constraint on the hypothesis that host specificity is conferred by shifts in accessory gene content. We also analyzed the ratio of nonsynonymous to synonymous substitutions in both core and accessory genes to identify positively selected genes that are potentially involved in arms-race dynamics. Arms-race genes may contribute to interactions with the plant hosts or insect vectors, and they are thus candidates to act as virulence or pathogenicity factors. We uncovered evidence of positive selection across codons of 19.5% (200 of 1,205) core genes and 7.7% (181 of 2,349) accessory genes. 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 biofilm formation *Neisseria gonorrhoeae*, and *cya*, which modulates gene expression in pathogenic bacteria.

36 **INTRODUCTION**

37 Bacteria exhibit extensive intraspecific variation in genome content. This variation is the
38 raw material for evolutionary adaptation, including the evolution of pathogenicity and virulence
39 (Furuya and Lowy 2006; Yacoubi et al. 2007; Juhas 2015; Chen et al. 2018). One example of
40 genome variation comes from an early study of *Escherichia coli* that compared two pathogenic
41 strains and one non-pathogenic laboratory strain (Welch et al. 2002). Of the entire set of protein
42 coding genes annotated by the three genomes, only 39.2% were shared among the three isolates.
43 Intriguingly, the two pathogenic strains each had 1,300 unique genes, while the laboratory strain
44 had only 585, suggesting that genes that vary across accessions (i.e., accessory genes) contribute
45 to virulence. Similar patterns have been illustrated for plant pathogens (Badet and Croll 2020;
46 Kim et al. 2020). In *Xanthomonas*, for example, horizontal gene transfer (HGT) has shuffled
47 virulent accessory genes from pathogenic strains to previously non-pathogenic strains (Chen et
48 al. 2018), facilitating the infection of common bean (*Phaseolus vulgaris* L.). In short, accessory
49 genes contribute to host-pathogen interactions, making them a critical focus for comparative
50 analyses of genome evolution and function.

51 Here we investigate variation in the genome content of another plant pathogen. *Xylella*
52 *fastidiosa* is endemic to the Americas and was first identified as the causal agent of Pierce's
53 Disease (PD), an economically devastating disease in grapevines (*Vitis vinifera* ssp. *vinifera*)
54 (Sicard et al. 2018; Burbank and Roper 2021). *X fastidiosa* causes additional economically and
55 ecologically impactful diseases like citrus variegated chlorosis, coffee leaf scorch, oak leaf
56 scorch and elm leaf scorch among others. Historically, the geographic distribution of *X.*
57 *fastidiosa* was limited to the Americas, but it has recently been introduced to the European
58 continent by anthropogenic transmission, which has further expanded its host range and led to

59 emerging diseases like olive quick decline syndrome (OQDS) in Italy (Schuenzel et al. 2005;
60 Loconsole et al. 2016). Since its detection in Italy, many olive trees have succumbed to OQDS
61 and *X. fastidiosa* has been detected in various plants species in Europe including France, Spain,
62 and Portugal (Chatterjee et al. 2008; Rapicavoli et al. 2018). In susceptible hosts, *X. fastidiosa*
63 can often be lethal, leading to significant crop losses, and it continues to threaten crops globally,
64 which has been further compounded by its recent spread in Europe (Tumber et al. 2014; Alston
65 et al. 2015).

66 For each of these diseases, *X. fastidiosa* is transmitted by xylem-feeding insect vectors to
67 the plant host, where it then utilizes cell wall degrading enzymes to systemically colonize the
68 xylem. In the xylem, it forms biofilms that are thought to be integral to virulence (Koo et al.
69 2017; Castro et al. 2021). Colonization is also governed, in part, by virulence and pathogenicity
70 factors that influence a wide range of bacterial functions – e.g., biofilm formation, host cell wall
71 degradation, regulatory systems, stress responses and bacterial membrane composition --
72 although other abiotic factors (like plant drought stress) likely also contribute to disease
73 progression (Rapicavoli et al. 2018). Given its economic impact, the effects and mechanisms of
74 *X. fastidiosa* infection have been studied widely, especially in grapevine (Roper and Lindow
75 2016). However, many pathogenicity factors likely remain undiscovered, and crucial questions
76 remain unanswered about the genetic factors that govern host-pathogen interactions and potential
77 host specialization (Rapicavoli et al. 2018).

78 In this context, it is helpful to recognize that *X. fastidiosa* consists of three commonly
79 recognized subspecies that form distinct phylogenetic clades: ssp. *fastidiosa*, *multiplex*, and
80 *pauca.* Each subspecies has unique phenotypic characteristics and DNA markers (Marcelletti and
81 Scortichini 2016). Two other subspecies, *morus* and *sandyi*, have also been suggested, though

82 they are not recognized as broadly (Burbank and Roper 2021); indeed, *morus* is believed to be a
83 product of a recombination event between *fastidiosa* and *multiplex* isolates (Sicard et al. 2018).
84 The recognition of subspecies is critical, because they loosely correlate with specific plant hosts
85 (Nunney et al. 2013). For example, cross-inoculation experiments have shown some almond
86 isolates cannot cause disease in grapes, a mulberry isolate does not cause disease in grapes and
87 that strains infecting citrus plants do not cause disease in coffee plants (and *vice versa*). These
88 observations suggest that genetic differences among strains can facilitate host-plant
89 specialization (Almeida and Purcell 2003; Hernandez-Martinez et al. 2006; Almeida et al. 2008;
90 Roper and Lindow 2016). However, there is not a clear one-to-one correspondence between
91 pathogen and host, because some strains can infect more than one host species, as demonstrated
92 by a strain that causes PD in grapevines and also leaf scorch in almonds (Almeida and Purcell
93 2003).

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. Most 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* (Castillo et al. 2021). Still other studies have focused on populations. For

105 example, Vanhove et al. (2020) isolated and sequenced *X. fastidiosa* subsp. *fastidiosa* from
106 symptomatic grapevines from five different California locations (Vanhove et al. 2020). One
107 common theme of these studies is that they identified the set of genes that were present in all of
108 their samples (i.e., core genes) and used those genes as the basis to perform phylogenetic
109 inference. However, most of these studies ignored the evolutionary properties of accessory
110 genes.

111 Like previous work, our study begins by identifying the core and accessory genes across
112 genomes, but it differs from previous work in at least three ways. First, we have sequenced and
113 assembled 20 new *X. fastidiosa* genomes, which we combined with publicly available data to
114 build a dataset allows us to investigate patterns of host specificity in a phylogenetic context.
115 Second, we focus on information from accessory genes to assess their phylogenetic signals,
116 patterns of their gene gain and loss across the *Xylella* phylogeny, and the potential associations
117 of genetic divergence with host species. We focus on these features under the hypothesis that
118 host switching is associated with accelerated turnover of accessory genes. Finally, we perform
119 extensive analyses of the ratio of nonsynonymous to synonymous (dN/dS or ω) substitutions in
120 both core and accessory genes. We study ω because genes involved in arms-race (or Red-Queen)
121 dynamics between pathogens and hosts often have ω values > 1.0, indicating a history of positive
122 selection (Daugherty and Malik 2012; Aleru and Barber 2020). In other systems, these
123 approaches have identified molecular functions that contribute to host defense and identified
124 entirely new genes and pathways involved in pathogen-host interactions (Mitchell et al. 2012; Ng
125 et al. 2015; Daugherty et al. 2016). Here we apply tests of positive selection to test the
126 hypothesis that accessory genes are particularly prone to arms race dynamics. By doing so, we
127 also hope to gain insights into genes that act as virulence or pathogenicity factors.

128

129 **METHODS**

130 **Novel *X. fastidiosa* genomes**

131 Fully extracted DNA from 20 *X. fastidiosa* isolates were provided by the French
132 Collection of Plant-Associated Bacteria (CIRM-CFBP; http://www6.inra.fr/cirm_eng/CFBP-
133 Plant-Associated-Bacteria) and from the University of California, Riverside. Genomic DNA was
134 prepared for Illumina sequencing using the Illumina Nextera DNA Flex Library Prep kit,
135 following the manufacturer's recommendations and for Pacific Biosciences (PacBio) sequencing
136 with the SMRTbell Express Template Prep Kit 2.0. SMRTbell libraries had 10kb DNA target
137 insert size (Pacific BioSciences, Menlo Park, CA) using 360ng of sheared DNA as input. DNA
138 libraries were sequenced with both Illumina and PacBio technologies at the University of
139 California, Irvine Genomics High Throughput Facility (https://ghtf.biochem.uci.edu). The
140 Illumina sequencing reads were quality assessed using FastQC, and reads were trimmed using
141 Trimmomatic v. 0.32 (Andrews 2010; Bolger et al. 2014) using default options. PacBio
142 sequencing reads were corrected and trimmed using Canu v. 1.5 (Koren et al. 2017). The long
143 and short reads were used for genome assembly with Unicycler v. 0.4.8 in hybrid assembly mode
144 (Wick et al. 2017). Genome assembly statistics were calculated using Quast v. 5.0.2 (Gurevich et
145 al. 2013). As is common practice (Chase et al. 2018), short contigs (<500 bp) were removed
146 from the assemblies using Seqkit v. 0.13.2 (Shen et al. 2016).

147

148 **Genome Assembly of public data and sample set curation**

149 We complemented our set of novel genomes with publicly available data. To do so, we
150 downloaded all available whole genome assemblies of *X. fastidiosa* and *X. taiwanensis* (as an

151 outgroup) from the National Center for Biotechnology Information (NCBI) and Sequence Read
152 Archive (SRA) databases on July 9, 2020. In addition, we downloaded the raw, short-read
153 sequences for an additional 20 isolates (Castillo et al. 2020; Vanhove et al. 2020). For each
154 isolate, we gathered information about its geographic origin and its host plant from NCBI and
155 from the Pathosystems Resource Integration Center (PATRIC) database. To assemble the raw
156 reads from the 20 unassembled accessions into genomes, we assessed quality and trimmed the
157 reads and applied SPAdes v. 3.14.0 (Bankevich et al. 2012) with the *--careful* option, following
158 Vanhove et al. (2020). If long reads were also available, as they were for 5 isolates from the
159 work of Castillo et al. 2020, then whole genome assembly was performed with Unicycler v. 0.4.8
160 in hybrid assembly mode (Wick et al. 2017).

161 In total, we gathered and generated 148 *Xylella* genome assemblies. From this set, we
162 removed isolates that did not have information about their host isolation source or were lab-
163 derived recombinant strains. The remaining 129 genomes were re-annotated by the same method
164 - based on Prokka v. 1.14.6 analysis – that we applied to the new genomes, to ensure
165 homogeneity. The Prokka analyses were then input into Roary v. 3.13.0 with options *-i 80 -cd*
166 *100 -e -n -z* to obtain a core gene alignment for initial comparisons among isolates (Seemann
167 2014; Page et al. 2015); we defined core genes as those that were detectable in 100% of the
168 samples. This core set was aligned with MAFFT and polished using gBlocks v. 0.91b
169 (Castresana 2000; Katoh et al. 2002; Katoh and Standley 2013). The polished alignment was
170 used as input for RAxML v. 8.2.12 to build a preliminary phylogenetic tree (Stamatakis 2014),
171 which we used to evaluate and curate the isolates (Supplementary Figure 1).

172 To curate the dataset, we created a distance matrix from the RAxML phylogenetic tree,
173 using the Tree and reticulogram REConstruction (T-REX) server (Boc et al. 2012). Many of the

174 genomes – most of which were gathered for population genomic analyses - were sampled from
175 the same plant host and were nearly identical genetically. To limit sampling biases for our
176 species-wide study, we removed clones and near-clones based on the distance matrix. That is, if
177 two or more isolates had a pairwise distance ≤ 0.0001 and came from the same host, we retained
178 the isolate with the more contiguous assembly. After applying these filters, our final dataset
179 consisted of 63 *X. fastidiosa* genomes and one outgroup genome (*X. taiwanensis* PLS229) that
180 were isolated from 23 distinct plant host species (Supplementary Table 1).
181
182 **Pan-genome analysis**
183 To perform a pan-genome analysis, we applied Roary to the 64 *Xylella* genomes using *gff*
184 files from Prokka as input. Roary was applied with the options *-i 80 -cd 100 -e -z*; these options
185 lower blastp sequence identity to 80% from the default 95% and define a core gene as a gene
186 present in 100% of the isolates used in the analysis. From the Roary output, we extracted a
187 representative nucleotide sequence of each core and accessory gene using cdbfasta
188 (https://github.com/gpertea/cdbfasta) and translated the nucleotide sequence to amino acids using
189 the transeq command from EMBL-EBI (Madeira et al. 2019). The representative sequences were
190 the basis for functional categorization -- using the eggNOG-mapper v. 2 (Huerta-Cepas et al.
191 2017; Huerta-Cepas et al. 2019) -- of both core and accessory genes. Gene Ontology (GO)
192 enrichment analyses were performed online at (http://geneontology.org) using *Xanthomonas*
193 *campestris* as the reference list (Ashburner et al. 2000).
194
195 **Phylogenetic Tree Construction and Hierarchical Clustering Analysis**

196 We used the core gene alignment from Roary to build a phylogenetic tree. To do so, we
197 curated the core gene alignment with gBlocks v. 0.91b (Castresana 2000), used the polished
198 alignment as input for IQtree v. 2.0.3, and selected the best nucleotide model for phylogenetic
199 tree construction (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017). We ultimately constructed
200 an unrooted tree using the GTR+F+R8 model with RAxML (Stamatakis 2014), using the 'best
201 tree' option. Phylogenetic trees were visualized and annotated using the ape package v. 5.5 in R
202 v. 4.0.2 (Paradis and Schliep 2019; R Core Team 2019). We used the most likely phylogeny to
203 test associations between phylogenetic relatedness, geography, and host isolation source (plant
204 taxonomic order information taken from https://www.itis.gov/) with ANOSIM implemented in
205 the *vegan* package v. 2.5-7 in R (Oksanen et al. 2020).
206 In addition to a phylogeny based on the core genes, we also built a phylogeny based on
207 the presence and absence matrix of accessory genes. We applied hierarchical clustering to the
208 presence-absence matrix using the *pvclust* package v. 2.2 in R with 1000 bootstrap replicates
209 (Suzuki and Shimodaira 2006). We also utilized the Mantel test (in the *vegan* package) to
210 measure the correlation between accessory gene content and phylogenetic relatedness. The
211 Mantel test required two distance matrices, which were generated from the core and accessory
212 trees by the reticulogram REConstruction (T-REX) server (Boc et al. 2012). We also performed
213 a DensiTree analysis to assess the overall species tree topology against the topology of individual
214 gene trees (Bouckaert 2010). Single gene trees were based on alignments from PRANK
215 (Löytynoja 2014). These core gene alignments were used as input for RAxML to construct gene
216 trees. DensiTree analyses was performed using the Phangorn package v. 2.5.5 in R (Schliep
217 2011).
218

219 **Gain and Loss of Accessory Genes**

220 We utilized GLOOME to investigate gene gain and loss dynamics along the core
221 phylogenetic tree of *X. fastidiosa* (Cohen et al. 2010). GLOOME uses a mixture-model
222 approach, coupled with maximum-likelihood inference, to infer rates of gain and loss of genes
223 along the branches of a phylogeny. It takes as input the phylogenetic topology, in this case the
224 phylogenetic topology based on core genes, and a presence-absence matrix of accessory genes.
225 The phyletic pattern of presence and absence was obtained through M1CR0B1AL1Z3R, as
226 recommended by the GLOOME authors, and then directly input into GLOOME using default
227 settings (Avram et al. 2019). The default settings included a fixed rate of gene gains and losses
228 with gamma distributed rates across genes (or sites). Among the output, GLOOME returned two
229 phylogenetic trees with branch lengths representing either the number of expected gain events or
230 the number loss events on each branch. As recommended (Cohen et al. 2010), branch lengths
231 representing relative gain and loss rates were extracted from the phylogenetic trees using FigTree
232 v. 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). To normalize expected gain (or loss) events
233 with sequence divergence, we calculated the ratio of inferred gain (or loss) against the branch
234 lengths of the sequence-based core phylogeny. Outlier branches with excess normalized gains or
235 losses were identified using the interquartile range criterion.

236

237 **Positive selection analyses**

238 We employed codeml from PAML v. 4.9 to calculate ω, the ratio of nonsynonymous to
239 synonymous rates (Yang 1997; Yang 2007). We performed *codeml* analysis on nucleotide
240 alignments of the single-copy core genes and on single-copy accessory genes that were present in
241 10 or more isolates. For the analysis of individual core genes, *codeml* was performed assuming

242 the phylogeny inferred from the core gene alignments. For each accessory gene, we generated an
243 unrooted maximum-likelihood tree based on the DNA alignment of that gene, using RAxML v.
244 8.2.12.

245 Given the input phylogenies, we performed *codeml* analyses that relied on calculating
246 likelihood ratios (LRs) under various models (Yang 2007). Briefly, we used the models to test
247 the null hypothesis that $\omega = 1.0$ against the alternative of positive selection ($\omega > 1.0$) in two
248 different ways. The first was a global test across the entirely phylogeny of a gene – i.e., across all
249 branches and all sites. This test requires the comparison of two models: one (M0) that estimates a
250 single ω from the data and another that sets ω=1.0 (M1). The two models yielded evidence for
251 positive selection when the initial ω estimate was >1.0 and when the likelihood of the two
252 models differed significantly, based on $P < 0.01$ after FDR correction. The second test was
253 across sites – i.e., testing for specific codons with $\omega > 1.0$ – based on a comparison of models
254 M1a and M2a from *codeml*.

255

256 **RESULTS**

257 **Core and Accessory Genes in *Xylella***

258 To investigate genome evolution in *X. fastidiosa*, we sequenced 20 novel *X. fastidiosa*
259 genomes using hybrid approaches, and we also retrieved publicly available genomes and raw
260 sequencing data (Supplementary Tables 1 and 2). This led to an initial dataset of 148 assembled
261 *Xylella* genomes, which we further reduced to 63 isolates based on two criteria. First, we
262 removed any genomes that did not include the isolation source (i.e., host species) or geographic
263 coordinates, leading to 129 remaining genomes. Second, we culled genetic synonyms with a
264 pairwise genetic distance < 0.0001 when the two isolates came from the same plant host. This

265 second set removed many genomes that were produced for population genomic studies, leaving a
266 sample of 63 genomes that were broadly distributed among the subspecies. All our analyses were
267 performed on this final set of 63 *X. fastidiosa* genomes with the *X. taiwanensis* outgroup.

268 The *X. fastidiosa* genomes in our dataset ranged in size from 2.42 Mb to 2.96 Mb, with
269 an average length of 2.61 Mb (Figure 1A). Annotation bast on the same method (to help ensure
270 fair comparisons) resulted in an average of 2,478 predicted genes per genome, with a range from
271 2,275 to 2,993 genes (Figure 1B). The dataset represents 22 plant hosts representing 12 botanical
272 orders, including the most (16) from coffee, followed by 7 from almond and grapes (Figure 1C).
273 The annotated genomes were the raw material for a pan-genome analysis that categorized each
274 gene as either core (present in all 64 samples, including *X. taiwanensis*) or accessory genes that
275 were present in one or more (but not all) genomes (Page et al. 2015). Across all 64 genomes, we
276 identified 10,428 genes within the pan-genome; of those, 1,025 were core genes and 9,406 were
277 accessory genes, with nearly 4,000 genes found in only a single isolate (Supplementary Table 3;
278 Figure 1D). As is common in the literature (Abby and Daubin 2007; Vernikos et al. 2015), we
279 further delineated accessory genes into 306 soft core genes present in > 95% of isolates (95% <=
280 isolates < 100%), 2,172 shell genes (15% <= strains < 95%), and 6,974 cloud genes (0% <=
281 strains < 15%; Supplementary Figure 2). Here, however, we focus on the entire accessory set
282 without relying on these additional arbitrary categories.

283 We performed functional analyses on both the core and accessory gene sets by grouping
284 protein coding sequences into clusters of orthologous gene (COG) (Figure 1E-F). Many genes
285 had unknown functions: 150 core genes and 1,584 accessory genes had unknown functions. We
286 compared COG category rankings between the core and accessory gene sets, and there were
287 notable differences between the two gene sets (even though difference was not statistically

288 significant based on the Wilcoxon non-parametric rank sum test; P = 0.779). After excluding
289 genes with unknown function, the largest COG categories in the core gene set were 'translation,
290 ribosomal structure, and biogenesis' (117 genes), 'cell wall/membrane/envelope biogenesis' (99
291 genes), and 'amino acid transport and metabolism' (82 genes). In contrast, the largest categories
292 for accessory genes were 'replication, recombination and repair' (538 genes), 'intracellular
293 trafficking, secretion, and vesicular transport' (365 genes), and 'transcription' (303 genes).
294 We also investigated the core and accessory gene lists for significant GO-based
295 enrichment for specific biological processes. GO analysis on the core gene set revealed
296 enrichment for pyrimidine nucleotide biosynthetic processes, translation, aromatic amino acid
297 family biosynthetic processes, and RNA methylation (Supplementary Table 4). GO analysis of
298 the accessory set identified enrichment for glutamine family amino acid biosynthetic processes,
299 tRNA aminoacylation, ncRNA processing, nucleotide biosynthetic processes, and cellular
300 component organization or biogenesis (Supplementary Table 5).
301
302 **Phylogenetic Patterns of Core Genes, Accessory Genes and Hosts**
303 *Phylogenetic Relationships* We used the core gene alignment to build a maximum
304 likelihood phylogeny. The topology was highly supported; it had a mean bootstrap support of
305 93.75% across all nodes, with a median of 100% (Figure 2). The lowest bootstrap supports were
306 primarily found at nodes separating *X. fastidiosa* strains isolated predominantly from grapevines,
307 reflecting relatively low evolutionary divergence among these samples. As expected (Denancé et
308 al. 2019), isolates clustered into three distinct clades representing the three main subspecies
309 (ssps. *fastidiosa*, *multiplex*, and *pauca*), with 27, 23 and 13 isolates in each clade, respectively.
310 While the core-gene phylogeny illustrated strong support for three clusters, it is worth noting that

311 individual core gene tree phylogenies displayed "clouds" around the three clusters, with many
312 single-gene phylogenies visually discordant (Supplementary Figure 3). Such discordance can
313 have many potential causes, including limitations in the extent and quality of phylogenetic
314 information captured within a single gene but also potential HGT events.

315 To investigate general evolutionary patterns of the accessory gene complement, we
316 compared the core gene phylogeny against a phylogeny based on accessory gene composition
317 (Supplementary Figure 4). Both the core gene phylogeny and the accessory gene dendrogram
318 clustered into three groups, and all members of the groups were consistent between phylogenetic
319 treatments (Figure 3). This pattern broadly suggests that accessory genes, while defined by their
320 inconstancy, are not exchanged *en masse* to a sufficient degree to alter phylogenetic signal
321 among subspecies. Within subspecies, however, relationships at the tips of the phylogeny often
322 differed between core and accessory trees. As an example, the cluster corresponding to *multiplex*
323 displayed the most discordance between the core and accessory trees, with all OTUs contributing
324 to phylogenetic incongruence (Figure 3). Interestingly, our *multiplex* sample also had more plant
325 host species than our *fastidiosa* and *pauca* samples, suggesting the possibility (but by no means
326 proving) that host factors may affect or moderate genome content. Nonetheless, we found a
327 significant correlation between distance matrices based on the core and accessory phylogenies
328 (Mantel test, $R = 0.8411$, $P < 0.001$), which is consistent with the fact that the two trees had the
329 same three major clades. The overarching impression of these analyses is that accessory gene
330 composition does not turn-over so rapidly, due to HGT or other mechanisms, to erase
331 phylogenetic and historical signals of subspecies diversification within *X. fastidiosa*.

332 We used the phylogenies to test associations between *X. fastidiosa* and their isolation
333 sources (i.e., geographic location or host plant information) using ANOSIM (see Methods). We

334 found a weak but significant phylogenetic association (ANOSIM R = 0.08178, P = 0.042)
335 between the core phylogeny and geographic information, suggesting (not surprisingly) that
336 strains isolated from similar geographic regions were more genetically similar to one another
337 than those isolated from different regions. Similarly, comparisons between strains isolated from
338 different host species revealed a weak significant phylogenetic signal (ANOSIM R = 0.1381, P =
339 0.047). Since *X. fastidiosa* infects a wide range of plants, we also retrieved the taxonomic order
340 of each plant host to test for a phylogenetic signal at a deeper taxonomic level. There was a
341 stronger association between plant order and phylogenetic signal (ANOSIM R = 0.3152, P <
342 0.0001); in other words, strains isolated from plants within the same taxonomic order were more
343 phylogenetically similar to one another than isolates taken from unrelated plants.
344 We hypothesized that accessory genes are crucial in pathogen-host interaction and
345 therefore repeated ANOSIM analyses with a distance matrix based on the presence and absence
346 of accessory genes (Figure 3). We found no significant association between accessory gene
347 content and geographic isolation source (ANOSIM R = 0, P = 0.5307). In contrast, we detected a
348 weakly significant association between accessory gene content and host species (ANOSIM R =
349 0.153, P = 0.042), but the association was less pronounced at the level of plant order (ANOSIM
350 R = 0.08424, P = 0.0784). Overall, associations were less evident based on accessory gene
351 content vs. the core-gene phylogeny.
352 *Gene Gain and Loss.* The sheer number of accessory genes indicate that the genome
353 content of *X. fastidiosa* is, like other microbes (Bolotin and Hershberg 2015; Iranzo et al. 2019),
354 shaped by extensive gene gain and loss events that are probably mediated by HGT (Firrao et al.
355 2021). We were interested in assessing the pattern of gene gain and loss across the phylogenetic
356 tree, hypothesizing that both could be enhanced on branches that lead to host shifts. We used

357 GLOOME to estimate the number of gain and losses of accessory genes across the *X. fastidiosa*
358 phylogeny and represented those estimates phylogenetically (Figure 4). Ignoring the branch
359 leading to the *X. taiwanensis* outgroup (PLS229), the internal branches discriminating the *X.*
360 *fastidiosa* subspecies were estimated to average ~550 separate gene gain and gene loss events.
361 The remainder of the tip and ingroup branches averaged ~100 gene gain and loss events (average
362 gains/branch = 92.8 genes; average losses/branch = 100.0 genes; Figure 4A&B).
363 While it is useful to estimate the number of gains and losses on each branch, we thought
364 it more helpful to normalize the number of estimated gain and loss events by branch lengths,
365 estimated from the sequence analysis of core genes. This normalization by branch length
366 converted the number of gene gains and losses to *rates* of gene gain (or loss) relative to sequence
367 divergence. We then sought to identify branches with aberrantly high rates of gene gain or loss
368 (Figure 4C&D), which reflect branches with especially notable turnover of accessory genes.
369 Somewhat surprisingly, we found that most of the phylogenetic lineages with outlier rates were
370 located at the tips of the phylogenetic tree. For example, of the 21 branches with high rates of
371 gene gain, 19 were at the tips of the phylogeny (Figure 4A). Similarly, 18 of 21 branches with
372 high rates of gene loss were external branches. These observations suggest features about the
373 evolutionary dynamics of genic turnover (see Discussion).
374
375 **Characterizing selection with ω**
376 We characterized selection on individual genes by estimating the dN/dS ratio (ω); we
377 especially sought to identify genes that experienced positive selection (i.e., $\omega > 1.0$), because it
378 is a potential signal of genes that contribute to arms-race dynamics between the pathogen and its
379 hosts. To do so, we applied a series of nucleotide substitution models to individual genes,

380 ultimately resulting in tests for positive selection on two levels: globally across a phylogeny and
381 across codon sites (see Methods). For these tests, we examined the full complement of 1,025
382 core genes and a subset of 2,349 accessory genes. The accessory subset was defined by presence
383 at least 10 of the 64 *Xylella* genomes; we chose this subset based on both a practical
384 consideration (to limit computational burden) and also in recognition that statistical power
385 decreases with the number of sequences in a dataset.

386 *Testing selection globally for each gene:* We first estimated a single ω value for each
387 gene with the codeml model M0, which assumes that ω is constant across all branches of the
388 entire gene tree and across all codons in the nucleotide alignment. Applied to the core genes, the
389 model estimated ω values that ranged from 0.01018 to 0.88905 with an average of 0.15415
390 (Figure 5A). None of these core genes had a ω estimates ($\widehat{\omega}$) higher than 1.0. The range of
391 $\widehat{\omega}$ was substantially broader for accessory genes, from $\widehat{\omega} = 0.0001$ to 32.1334, with an average
392 of 0.51443 (Figure 5B). Despite the wide range of $\widehat{\omega}$ values in the accessory set, only 222 genes
393 (or 9.5%) had a global estimate of ω > 1. The average $\widehat{\omega}$ was significantly higher in the
394 accessory gene set compared to the core genes (unpaired T-test, P < 0.0001), reflecting either
395 lower purifying selection against these genes, more positive selection, or both.

396 We then tested whether genes have ω significantly different from 1.0. To do so, we
397 compared the log likelihood from M0 to a model with ω fixed to 1.0. The vast majority (1,021 of
398 1,025) of core genes had $\widehat{\omega}$ significantly < 1.0 (P < 0.01, FDR correction; Figure 5A), reflecting
399 pervasive purifying selection. The trend for accessory genes was not as severe, because 1,040
400 genes had ω significantly < 1.0 (P < 0.01, FDR correction) but 1,282 genes had ω values that
401 were not significantly different from 1.0 (Figure 5B). Of the 222 genes with ω estimates > 1.0,
402 only three had statistically significant evidence for positive selection. These three genes are

403 strong candidates to encode proteins involved in arms-race dynamics, but all three were
404 annotated as hypothetical genes *(group_1159, group_1160*, and *group_4304*; Supplementary
405 Table 3).

406 *Positive selection in codon sites:* The global test is a conservative criterion to search of
407 positive selection, perhaps overly so. Accordingly, we turned to an alternative method that tests
408 for variation in ω among codon sites and identifies whether sites are under positive selection. To
409 do so, we ran the sites models in *codeml*, which are a group of nested models. For completeness,
410 we first compared sites model M0, which represents the null hypothesis that there is a single ω
411 value for all sites, against sites model M3, which permits ω to vary among sites. In the core
412 genes, the likelihood ratio test was significant for 726 genes ($P < 0.01$, FDR correction),
413 implying that ω varies among sites. Next, we compared sites model M2a against model M1a to
414 test for positive selection among sites in the gene. We found that 200 core genes had evidence of
415 positive selection among sites ($P < 0.01$, FDR correction); of these, 199 were also significant for
416 the M0-M3 comparison. We also tested for positive selection on codon sets within the 2,349
417 accessory genes by comparing sites model M2a against M1a. Of those, 181 genes yielded
418 evidence of positive selection among sites ($P < 0.01$, FDR correction, 7.7% of accessory genes),
419 of which 173 were also significant for the M0-M3 comparison. To sum: 19.5% (i.e., 200 of
420 1,024) of core genes had significant evidence for codons under positive selection, while 7.7%
421 (181/2,349) of accessory genes contained codons with an apparent history of positive selection.
422 Interestingly, of the 181 accessory genes, 41 of these had estimates of $\omega > 1.0$ across all sites and
423 phylogenetic branches, even though the global test was not significant for any of these genes. We
424 nonetheless consider these 41 to be the strong candidates to have experienced positive selection.
425

426 **DISCUSSION**

427 Host-pathogen interactions can drive rapid evolution of pathogenic bacteria, particularly
428 for genes involved in arms-race dynamics (Daugherty and Malik 2012; Sironi et al. 2015). Here,
429 we investigated the genomic evolution of the plant pathogen, *X. fastidiosa*, through comparative
430 genomic analysis of genomes that represent diversity across the species, based on a sample set of
431 64 genomes. The sample was isolated from 23 different plant hosts (**Figure 1C**) from throughout
432 the world, with the majority from the United States (n=21) followed by Costa Rica (n=14) and
433 Brazil (n=8; Supplementary Figure 1). With these data, we constructed a pangenome that
434 contained 1,024 core genes and 9,406 accessory genes, reflecting numbers similar to previous
435 studies that used fewer samples or subspecies-specific sets of *X. fastidiosa* genomes
436 (Giampetruzzi et al. 2017; Castillo et al. 2020). These values illustrate, however, that less than
437 half of the ~2,748 coding sequences (per genome average) are likely to be essential, because
438 most genes vary across isolates. Of the core genes, the majority were involved in essential
439 cellular processes -- such as translation, cell wall biogenesis, and amino acid metabolism
440 (**Figure 1E**) -- as expected (Tettelin et al. 2008).

441 We used the set of core genes to infer a maximum likelihood phylogeny. As with
442 previous systematic treatments of *X. fastidiosa* (Yuan et al. 2010; Marcelletti and Scortichini
443 2016; Denancé et al. 2019), our phylogeny identified three clades corresponding to the three
444 main subspecies (*fastidiosa, multiplex*, and *pauca*). The phylogeny suggests that *pauca* is the
445 oldest among the three, because it diverged prior to the divergence of *multiplex* and *fastidiosa*
446 (**Figure 2**). Interestingly, based on our sample, it is not clear that *pauca* has greater diversity of
447 host species despite its greater age. Our sample includes 13 *pauca* strains from 7 plant hosts (a
448 ratio of 0.54 hosts per strain), while *multiplex* has a similar ratio (23 strains from 12 different

449 plant hosts = 0.52 hosts/strain). In contrast, *fastidiosa* has a much lower ratio of 0.27
450 hosts/strains, with 27 strains from 10 plants. These ratios may certainly reflect properties of our
451 sample more than biological trends – e.g., much of the research on subspecies *fastidosa* have
452 focused on PD in grapevines and so is biased – but these observations nonetheless suggest that
453 the older age of *pauca* has not led to obviously greater host diversity than *multiplex*.
454
455 **The question of plant host specialization**
456 The phenomenon of host specialization of *X. fastidiosa* (and perhaps whether it exists at
457 all) is somewhat unclear. On the one hand, *X. fastidiosa* is a generalist known to colonize or
458 infect over 350 plant species (Sicard et al. 2018). Furthermore, some strains are known to cross-
459 infect hosts, such as the isolate that causes both PD and almond leaf scorch (Chen et al. 2010).
460 On the other hand, experimental work has shown some evidence of host specificity with respect
461 to colonization and disease progression (Almeida et al. 2008). For example, *X. fastidiosa*
462 isolates known to cause citrus variegated chlorosis and coffee leaf scorch were apparently unable
463 to colonize the alternative host (Almeida et al. 2008). Similarly, Castillo et al. 2020 demonstrated
464 that the *X. fastidiosa* ssp. *fastidiosa* isolates from coffee or periwinkle did not efficiently infect
465 grapevines, even though ssp. *fastidiosa* commonly infects grapevines.
466 In addition to experimental evidence, the question of host specialization has been
467 addressed previously using phylogenetic approaches based on multilocus sequencing typing
468 (MLST) data. In this work, Sicard et al. (2018) generated MLST data from 7 housekeeping genes
469 from 50 *X. fastidiosa* genotypes. After building a phylogeny, they tested coevolutionary
470 relationships between host species and *X. fastidiosa* MLST types but found no significant
471 evidence of coevolution, implying a lack of host specialization. We have revisited this issue

472 using whole genome sequence data and a larger sample of isolates. We found a significant, non-
473 random association between phylogenetic relationships and both the species and taxonomic order
474 of plant hosts ($P < 0.0001$). These results are consistent with some level of specialization of *X.*
475 *fastidiosa* to plant hosts. Although it is difficult to compare ANOSIM results across studies, it is
476 worth noting that the association of *X. fastidiosa* to plant order is similar in magnitude to the
477 association between a gut colonizing bacterium (bifidobacteria) and the host species from which
478 it was isolated (Rodriguez and Martiny 2020).

479 Given some evidence for host specialization, we hypothesized that it is driven in part by
480 accessory gene content. Under this hypothesis, we predicted an association between genes and
481 hosts should be as (or more) pronounced for accessory genes than for core genes. Instead, we
482 found no significant association between accessory gene complement and taxonomic order and
483 only a weak association with plant species. Our results are unlike, for example, the case of
484 bifidobacteria, where the association with host species was nearly as strong for accessory genes
485 as for host genes (Rodriguez and Martiny 2020). We do not know why we did not detect a signal
486 for host specialization of accessory genes, but we can think of three explanations. One is that our
487 hypothesis is incorrect – i.e., that host associations, to the extent they exist, are not driven by
488 accessory genes but rather evolutionary divergence in core genes. Another is statistical power:
489 because there are many more sequence changes among the core genes than there are changes in
490 accessory gene content, the distance matrix for core genes likely has a higher signal-to-noise
491 ratio than accessory gene content. Finally, if accessory genes do mediate host shifts, it is possible
492 – and even likely - that only a subset of accessory genes drive these shifts. Under this scenario,
493 there may be significant associations for a small subset of accessory genes, but the signal of this
494 association is weak across the entire accessory gene set. This conjecture seems reasonable given

495 that *X. fastidiosa* interacts not only with plants but also insect vectors and microbial
496 communities. Hence, some subset of accessory genes may contribute to these interactions, rather
497 than plant host interactions. Untangling our three explanations should be possible, but it will
498 likely require even larger datasets (to tackle potential issues of statistical power) and also a better
499 understanding of *X. fastidiosa* diversity within and among insect vectors.
500
501 **The pattern of gene gain and loss events**
502 Another potential tool to study adaptation to specific hosts is by examining shifts in gene
503 composition through gene duplication, deletion, or HGT events (Hurles 2004; Arnold et al.
504 2022). We estimated the number of gene loss and gain events along the core-gene phylogeny,
505 and then normalized those numbers relative to sequence divergence. Using this approach, we
506 found that most branches followed a consistent rate of gene gain or loss relative to sequence
507 divergence. The fact that the accessory gene phylogeny recapitulates the three subspecies (Figure
508 3) suggests, along with previous evidence, that *X. fastidiosa* evolves predominantly through
509 vertical inheritance and intraspecific recombination, rather than HGT from other bacterial
510 species (Nunney et al. 2013; Castillo and Almeida 2021).
511 We have, however, identified 19 and 18 lineages with enriched gain or loss events,
512 respectively, and most of these branches were at the tip of the phylogeny. Again, a potential
513 explanation for these gain and loss dynamics is that they reflect host shifts. There are some
514 isolated examples that are consistent with this hypothesis. For example, isolates XF6c, Pr8x,
515 RAAR17, and OLS0478 in *pauca* have branches with enriched gene gains (Figure 4A). Two of
516 these (OLS0478 and Pr8x) were isolated from oleander and plum, respectively, and are the only
517 isolates associated with those plant hosts in their clades, suggesting a host shift. More globally,

518 however, the evidence for this hypothesis is unconvincing. When we, for example, contrast gene
519 gains between pairs of sister taxa with the same plant host, three of the 16 sister pairs had
520 enriched rates of gene gain. This proportion of enriched branches was not significantly lower
521 than the reminder of the tree ($P > 0.05$; Fisher's Exact Test), despite the fact that the sister taxa
522 did not experience a host shift. All of these inferences are of course dependent on our sample and
523 ignore the vector component of the *X. fastidiosa* lifecycle, so there are limitations to our
524 conclusions. At present, however, the evidence for an association between host shifts and
525 enhanced gene gain and loss events is weak.

526 This leaves unexplained the pattern of enriched rates of gene gain and loss at the tips of
527 the tree. We suspect this pattern is analogous to patterns of mutations in populations. New
528 mutations begin as rare, low frequency variants in single individuals. Eventually most of these
529 mutations are removed by the processes of genetic drift and natural selection, so that there are
530 more new mutations in populations than old mutations. In a phylogenetic context, these new
531 mutations would be evident at the tip of the trees, so it may be reasonable to expect higher
532 effective rates of gene gain and loss in the 'newest' phylogenetic branches. This explanation only
533 has credence, however, if the observed gain and loss events are both recent – i.e., newer than the
534 sequence mutations that define the tip branches – and frequent.

535

536 **The identification of putative arms-race genes**

537 Many previous studies have implicated genes and their protein products in ongoing arms-
538 races between pathogens and their hosts (Anderson et al. 2010; Schulte et al. 2010). One way to
539 approach this question is agnostic of function, which is to screen for genes with a history of
540 positive selection. In our case, a set of positively selected *X. fastidiosa* genes represents potential

541 candidate virulence factors and also factors that mediate interactions with the environment,
542 including plant host, insect vectors, or members of the microbial community. Previous research
543 on *X. fastidiosa* has searched for signatures of selection by comparing levels of polymorphism or
544 rates of synonymous and nonsynonymous mutations in the core genome using Tajima's D and
545 the McDonald-Kreitman test (Castillo et al. 2020; Castillo et al. 2021). These population-based
546 tests generally search for a signal of local adaptation, rather than arms-race dynamics. Another
547 work has measured ω in core genes but without statistically testing for positive selection
548 (Castillo and Almeida 2021). Thus, no other work on *X. fastidiosa* has yet taken as thorough an
549 approach to search for signals of positive selection.

550 To study positive selection, we estimated ω, or the ratio of nonsynonymous to
551 synonymous mutations, for each core gene and for each accessory gene found in 10 or more
552 isolates. In total, this exercise encompassed 3,373 genes: 1,024 core and 2,349 accessory genes.
553 We began by applying a global test that estimates ω over all sites and phylogenetic lineages.
554 This approach can be overly conservative, because a significant test of $\omega > 1.0$ requires that
555 positive selection is very strong, acts across many sites in a gene, is present in most of the
556 branches of the phylogeny, and/or all of the above. As expected, we found only a few genes
557 (three accessory genes in total) that were significant for positive selection with this test.
558 Unfortunately, the annotations of these few genes yielded no insights into their functions. This
559 test did allow, however, for two broad generalizations about patterns of selection in *X. fastidiosa*.
560 First, as a group the core genes are under strong purifying selection with most (>99%) having ω
561 estimates significantly < 1.0. Interestingly, the highest ω value for a core gene set was ω =
562 0.889, for a hypothetical protein within unknown function but with a possible role in the cell
563 membrane (EggNOG query OG: ENOG5031E4H). Second, accessory genes generally have

564 lower levels of purifying selection, as evidenced by a lower proportion (45%) of significant tests
565 for $\omega < 1.0$ and by much higher average $\hat{\omega}$ values (0.15415 vs. 0.51443; Figure 5A). The
566 proportion of significant tests must be compared between genic sets with caution, because the
567 smaller sample sizes (n=10 to 63) for accessory genes likely reduce statistical power relative to
568 the n=64 samples for all core genes, as do any differences in gene lengths. Nonetheless, the
569 contrasting pattern of ω is consistent with the ideas that core genes have conserved biological
570 functions and that accessory genes are more amenable to evolutionary change due to their
571 nonessential, but still potentially biological relevant, cellular roles (Horesh et al. 2021).
572 Accessory genes may also experience higher variation in their selection dynamics because
573 recombination affects them more than core genes (Castillo and Almeida 2021).
574 Given few signals of positive selection, we turned to codon site models, which identify
575 evidence of particular codon sites with histories of positive selection. To our surprise, we not
576 only found evidence of positive selection in core genes, but the proportion of positively selected
577 genes was higher for core genes (at 19.5%) than for accessory genes (8.1%). These proportions
578 may again reflect differences in statistical power between the two groups, but the data
579 nonetheless suggest that a large number of core genes have some history of positive selection.
580 The salient question is whether these genes give some clue to function. Of the 200 core genes
581 with evidence for positive selection at the codon level, we found 26% were unannotated. The
582 remaining genes were enriched for GO categories related to tryptophan, glutathione, and
583 threonine biosynthetic process along with fatty acid catabolic process, lipid modification, and
584 nucleoside triphosphate biosynthetic process among others (Supplementary Table 6), none of
585 which seem especially integral to interactions with plant or insect hosts. We also performed a
586 functional analysis by grouping the protein coding sequences of these 200 core genes into COG

587 categories to infer cellular functions. Excluding the category of unknown function (which was
588 the largest with 29 instances), the two largest categories were 'amino acid transport and
589 metabolism' and 'cell wall/membrane/envelope biogenesis' followed by 'energy production and
590 conversion' and 'intracellular trafficking, secretion, and vesicular transport' (Supplementary
591 Figure 6A).

592 Of the 181 accessory genes with evidence for positive selection at the codon level, 82%
593 were not annotated for function. To infer function, we performed a similar functional analysis
594 and found that the largest categories (excluding the category of unknown function) were
595 'intracellular trafficking, secretion, and vesicular transport,' 'replication, recombination, and
596 repair,' and 'post-translational modification, protein turnover, and chaperones' (Supplementary
597 Figure 6B). Intriguingly, of this set of 181 genes, 41 overlapped with the set of 222 genes that
598 had a gene-wide estimate of $\widehat{\omega} > 1$. While these are especially strong candidates for arms-race
599 genes, a disappointing 95% of them were unannotated for function. Two genes had annotations
600 which we could explore further: *nagZ_2* with $\widehat{\omega} = 3.77387$ and *cya* with $\widehat{\omega} = 1.2424$. The gene
601 *nagZ_2* encodes a beta-glucosidase that is important for biofilm formation in *Neisseria*
602 *gonorrhoeae*, suggesting it could play a similar role in *X. fastidiosa*. It merits further functional
603 analysis, since biofilms are important to the infection cycle (Bhoopalan et al. 2016). We also
604 examined *cya*, which encodes adenylate cyclase and plays essential roles in regulation of cellular
605 metabolism (Danchin et al. 1984). Interestingly, the protein is involved in the cyclic AMP
606 system, which is a global regulator in gram-negative bacteria and has been shown to modulate
607 gene expression in pathogenic bacteria (Smith et al. 2004; Kim et al. 2005).

608 As a final exercise, we cataloged the incidence of positive selection in a set of 35 genes
609 that have been listed as virulence and pathogenicity factors in *X. fastidiosa* (Rapicavoli et al.

610 2018). Of the 35, we could identify 29 in our database (Table 1). We expected that this set of 29
611 genes would be enriched for evidence of positive selection relative to the genomic background,
612 because these genes are putatively involved in arms-race interactions. The trend for these genes
613 was in the right direction for enrichment, because 7 of 29 (= 24%) were significant vs. 352 of
614 3345 (= 10.5%) in the rest of the genome. However, the difference in proportions was only
615 borderline significant (Fisher Exact Test, P = 0.078). Nonetheless, this set of experimental genes
616 is quite interesting. Of the seven with evidence of positive selection, six encode proteins
617 associated with the membrane of gram-negative bacteria and are involved in membrane transport
618 or adhesin. Specifically, the genes *fimF, xadA* and *xatA* encode proteins involved in fimbrial
619 adhesion, non-fimbrial adhesion, and biofilm formation, respectively, and the genes PD1311,
620 *tonB1*, and *xpsE* encode proteins involved in membrane transport (Ma Rodriguez et al. 1993; Sun
621 et al. 2005; Abbas et al. 2007; Das et al. 2009; Zeiner et al. 2012). The remaining gene *lesA/lipA*
622 is an extracellular lipase that affect biofilm stability in *X. fastidiosa*, similar to other pathogens
623 like *Pseudomonas aeruginosa* (Tielen et al. 2013; Nascimento et al. 2016).
624

625 **CONCLUSIONS**

626 We performed comparative genomic analysis of *X. fastidiosa* genomes to explore the
627 evolutionary history of the core and accessory genes in this plant pathogen and to explore genetic
628 determinants that may contribute to host specificity or virulence. Using phylogenetic methods
629 and association tests, we found that *X. fastidiosa* shows evidence for adaptation to host plants as
630 reflected in their evolutionary history of core genes. We also inferred that the evolutionary
631 dynamics of *X. fastidiosa* accessory gene complement tends to retain information about vertical
632 inheritance and intraspecific recombination rather than horizontal gene transfer between

633 interspecifics. We also found evidence of positive selection in 200 and 181 core and accessory
634 genes using sites models, suggesting that *X. fastidiosa*, like other pathogens, is undergoing
635 adaptive evolution in response to its hosts.
636

637 **DATA AVAILABILITY STATEMENT**

638 All high-throughput sequence data generated in this study have been submitted to the NCBI
639 Sequence Read Archive database at https://www.ncbi.nlm.nih.gov/sra and can be accessed with
640 project number SUB11276742.
641

642 **ACKNOWLEDGEMENTS**

643 We thank R. Gaut, INRA-CFBP, and the UC Irvine Genomics High-Throughput Facility for
644 contributing to data generation. We also thank J.J. Emerson, A. Martiny, E. Solares, and C.I.
645 Rodriguez for helpful input on methodology. T.N.B. was supported by the National Science
646 Foundation Graduate Research Fellowship Program and the University of California Irvine
647 President's Dissertation Year Fellowship. This work was supported by National Science
648 Foundation grant no. 1741627 to B.S.G. and California Department of Food and Agriculture
649 agreement numbers 15-0218-SA and 18-0328-000-SA to M.C.R.
650

651 **REFERENCES**
652

653 Abbas A, Adams C, Scully N, Glennon J, O'Gara F. 2007. A role for TonB1 in biofilm
654 formation and quorum sensing in Pseudomonas aeruginosa. *FEMS Microbiol. Lett.*
655 274:269–278.

656 Abby S, Daubin V. 2007. Comparative genomics and the evolution of prokaryotes. *Trends*
657 *Microbiol.* 15:135–141.

658 Aleru O, Barber MF. 2020. Battlefronts of evolutionary conflict between bacteria and animal
659 hosts. *PLOS Pathog.* 16:e1008797.

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

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

665 Alston JM, Fuller KB, Kaplan JD, Tumber KP eds. 2015. Assessing the returns to R&D on
666 perennial crops: the costs and benefits of Pierce's disease research in the California
667 winegrape industry. *Aust. J. Agric. Resour. Econ.*

668 Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB, Anderson JP, Gleason
669 CA, Foley RC, Thrall PH, et al. 2010. Plants versus pathogens: an evolutionary arms
670 race. *Funct. Plant Biol.* 37:499–512.

671 Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.
672 Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

673 Arnold BJ, Huang I-T, Hanage WP. 2022. Horizontal gene transfer and adaptive evolution in
674 bacteria. *Nat. Rev. Microbiol.* 20:206–218.

675 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K,
676 Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat.*
677 *Genet.* 25:25–29.

678 Avram O, Rapoport D, Portugez S, Pupko T. 2019. M1CR0B1AL1Z3R—a user-friendly web
679 server for the analysis of large-scale microbial genomics data. *Nucleic Acids Res.*
680 47:W88–W92.

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

683 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko
684 SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A New Genome Assembly Algorithm
685 and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19:455–477.

686 Bhoopalan SV, Piekarowicz A, Lenz JD, Dillard JP, Stein DC. 2016. nagZ Triggers Gonococcal
687 Biofilm Disassembly. *Sci. Rep.* 6:22372.

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

690 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
691 data. *Bioinformatics* 30:2114–2120.

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

694 Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*
695 26:1372–1373.

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

698 Castillo AI, Almeida RPP. 2021. Evidence of gene nucleotide composition favoring replication
699 and growth in a fastidious plant pathogen. *G3 GenesGenomesGenetics* 11:jkab076.

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

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

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

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

710 Chase AB, Gomez-Lunar Z, Lopez AE, Li J, Allison SD, Martiny AC, Martiny JBH. 2018.
711 Emergence of soil bacterial ecotypes along a climate gradient. *Environ. Microbiol.*
712 20:4112–4126.

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

715 Chen NWG, Serres-Giardi L, Ruh M, Briand M, Bonneau S, Darrasse A, Barbe V, Gagnevin L,
716 Koebnik R, Jacques M-A. 2018. Horizontal gene transfer plays a major role in the
717 pathological convergence of Xanthomonas lineages on common bean. *BMC Genomics*
718 19:606.

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

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

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

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

728 Daugherty MD, Schaller AM, Geballe AP, Malik HS. 2016. Evolution-guided functional
729 analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. *eLife*
730 [Internet]. Available from: https://elifesciences.org/articles/14228

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

733 Firrao G, Scortichini M, Pagliari L. 2021. Orthology-Based Estimate of the Contribution of
734 Horizontal Gene Transfer from Distantly Related Bacteria to the Intraspecific Diversity
735 and Differentiation of Xylella fastidiosa. *Pathogens* 10:46.

736 Furuya EY, Lowy FD. 2006. Antimicrobial-resistant bacteria in the community setting. *Nat. Rev.*
737 *Microbiol.* 4:36–45.

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

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

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

747 Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, Heinz E, Thomson NR.
748 2021. Different evolutionary trends form the twilight zone of the bacterial pan-genome.
749 *Microb. Genomics* 7:000670.

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

753 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR,
754 Letunic I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and
755 phylogenetically annotated orthology resource based on 5090 organisms and 2502
756 viruses. *Nucleic Acids Res.* 47:D309–D314.

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

758 Iranzo J, Wolf YI, Koonin EV, Sela I. 2019. Gene gain and loss push prokaryotes beyond the
759 homologous recombination barrier and accelerate genome sequence divergence. *Nat.*
760 *Commun.* 10:5376.

761 Juhas M. 2015. Horizontal gene transfer in human pathogens. *Crit. Rev. Microbiol.* 41:101–108.

762 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
763 fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.

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

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

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

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

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

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

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

780 Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Multiple
781 Sequence Alignment Methods. Methods in Molecular Biology. Totowa, NJ: Humana
782 Press. p. 155–170. Available from: https://doi.org/10.1007/978-1-62703-646-7_10

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

786 Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter
787 SC, Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in
788 2019. *Nucleic Acids Res.* 47:W636–W641.

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

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

795 Nascimento R, Gouran H, Chakraborty S, Gillespie HW, Almeida-Souza HO, Tu A, Rao BJ,
796 Feldstein PA, Bruening G, Goulart LR, et al. 2016. The Type II Secreted Lipase/Esterase
797 LesA is a Key Virulence Factor Required for Xylella fastidiosa Pathogenesis in
798 Grapevines. *Sci. Rep.* 6:18598.

799 Ng M, Ndungo E, Kaczmarek ME, Herbert AS, Binger T, Kuehne AI, Jangra RK, Hawkins JA,
800 Gifford RJ, Biswas R, et al. 2015. Filovirus receptor NPC1 contributes to species-specific
801 patterns of ebolavirus susceptibility in bats.Neher RA, editor. *eLife* 4:e11785.

802 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective
803 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*
804 32:268–274.

805 Nunney L, Vickerman DB, Bromley RE, Russell SA, Hartman JR, Morano LD, Stouthamer R.
806 2013. Recent Evolutionary Radiation and Host Plant Specialization in the Xylella
807 fastidiosa Subspecies Native to the United States. *Appl. Environ. Microbiol.* 79:2189–
808 2200.

809 Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR,
810 O'Hara RB, Simpson GL, Solymos P, et al. 2020. vegan: Community Ecology Package.
811 Available from: https://CRAN.R-project.org/package=vegan

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

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

817 R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for
818 Statistical Computing Available from: https://www.R-project.org/

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

821 Rodriguez CI, Martiny JBH. 2020. Evolutionary relationships among bifidobacteria and their
822 hosts and environments. *BMC Genomics* 21:26.

823 Roper C, Lindow SE. 2016. CHAPTER 16: Xylella fastidiosa: Insights into the Lifestyle of a
824 Xylem-Limited Bacterium. In: Caroline R, Steven EL, editors. Virulence Mechanisms of
825 Plant-Pathogenic Bacteria. Bacteriology. The American Phytopathological Society. p.
826 307–320. Available from: https://apsjournals.apsnet.org/doi/10.1094/9780890544495.016

827 Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.

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

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

834 Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069.

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

837 Sicard A, Zeilinger AR, Vanhove M, Schartel TE, Beal DJ, Daugherty MP, Almeida RPP. 2018.
838 Xylella fastidiosa: Insights into an Emerging Plant Pathogen. *Annu. Rev. Phytopathol.*
839 56:181–202.

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

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

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

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

850 Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in
851 hierarchical clustering. *Bioinformatics* 22:1540–1542.

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

854 Tielen P, Kuhn H, Rosenau F, Jaeger K-E, Flemming H-C, Wingender J. 2013. Interaction
855 between extracellular lipase LipA and the polysaccharide alginate of Pseudomonas
856 aeruginosa. *BMC Microbiol.* 13:159.

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

859 Vanhove M, Sicard A, Ezennia J, Leviten N, Almeida RPP. 2020. Population structure and
860 adaptation of a bacterial pathogen in California grapevines. *Environ. Microbiol.* 22:2625–
861 2638.

862 Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Curr.*
863 *Opin. Microbiol.* 23:148–154.

864 Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R,
865 Boutin A, Hackett J, et al. 2002. Extensive mosaic structure revealed by the complete
866 genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci.* 99:17020–
867 17024.

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

870 Yacoubi BE, Brunings AM, Yuan Q, Shankar S, Gabriel DW. 2007. In Planta Horizontal
871 Transfer of a Major Pathogenicity Effector Gene. *Appl. Environ. Microbiol.* 73:1612–
872 1621.

873 Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood.
874 *Bioinformatics* 13:555–556.

875 Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.*
876 24:1586–1591.

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

880 Zeiner SA, Dwyer BE, Clegg S. 2012. FimA, FimF, and FimH Are Necessary for Assembly of
881 Type 1 Fimbriae on Salmonella enterica Serovar Typhimurium. *Infect. Immun.* 80:3289–
882 3296.

883

884 **Table 1:** *Codeml* results for experimentally identified virulence and pathogenicity genes, as
885 listed Rapicavoli et al. (2018)
886

PD Number	Gene Name	Pan-genome classification	No. Genomes[1]	$\hat{\omega}$ (M0)[2]	M2a vs. M1a p-value[3]
PD0058	*fimF*	Accessory	41	0.31555	**2.93E-08**
PD0062	*fimA*	Accessory	26	0.81255	0.204
PD0233	*rpfB*	Accessory	57	0.16832	1.00
PD0279	*cgsA*	Core	64	0.14876	0.396
PD0406	*rpfC*	Accessory	44	0.34502	1.00
PD0528	*xatA*	Core	64	0.4568	**1.16E-117**
PD0731	*xadA*	Accessory	58	0.39197	**0.003**
PD0732	*xpsE*	Core	64	0.05966	**0.004**
PD0814	*wzy*	Accessory	43	0.17675	1.00
PD0843	*tonB1*	Core	64	0.1126	**4.17E-04**
PD0848	*pilL*	Core	64	0.1808	0.893
PD0986		Core	64	0.10876	1.00
PD1099	*dinJ/relE*	Accessory	25	0.10271	1.00
PD1100		Accessory	15	0.20708	0.648
PD1284	*algU*	Core	64	0.20339	0.979
PD1311		Accessory	33	0.42541	**2.68E-05**
PD1380	*csp1*	Core	64	0.15702	1
PD1391	*gumH*	Accessory	46	0.12964	1
PD1394	*gumD*	Accessory	63	0.12358	0.204
PD1485	*pglA*	Accessory	59	0.27857	0.271
PD1678	*phoQ*	Core	64	0.10503	0.011
PD1679	*phoP*	Core	64	0.02806	1.00
PD1703	*lesA/lipA*	Core	64	0.0676	**0.008**
PD1792	*hxfB*	Core	64	0.10876	1.00
PD1826	*chiA*	Core	64	0.12798	0.695
PD1856	*engXCA1*	Accessory	63	0.25823	0.122
PD1964	*tolC*	Core	64	0.07808	1.00
PD1984	*gacA*	Core	64	0.13444	1.00
PD2118	*hxfA*	Core	64	0.10876	1.00

887
888 [1] the number of genomes, out of 64, in which the gene was detected.
889 [2] M0 estimates a single ω across the entire phylogeny of sequences.
890 [3] The p-value of tests after FDR correction. Bolded values are significant at $p < 0.01$. The
891 notation E refers to the power of ten.

892 **FIGURE LEGENDS**
893
894 **Figure 1:** Histograms reporting the characteristics of the 64 *Xylella* genomes. A) Genome
895 lengths, exhibited in base pairs. B) The number of genes within a genome. C) A histogram of the
896 plant species from which genomes were isolated. D) A histogram of the number of genes found
897 in *x* number of genomes; this histogram shows, for example, that nearly 4,000 genes are found in
898 only of the genomes out of the entire sample of 63 genomes and that ~1,000 genes are found in
899 all 64 genomes; E) the distribution of functional categories for the set of 1,024 core genes and F)
900 the distribution of functional categories for the set of 9,258 accessory genes. A key to the COG
901 categories for panels E) and F) is in Figure S5.
902
903 **Figure 2:** The inferred phylogeny of 64 *Xylella* genomes, based on maximum likelihood
904 inference of core gene alignments. Each isolate is labelled at the tips and is colored according to
905 the order of the plant isolation source (host). The common name of the host is provided to the
906 right of order information. The three *X. fastidiosa* subspecies are indicated, as are bootstrap
907 values at each node. The bootstrap values are pie charts, where black represents the percent of
908 bootstrap support, and the scale bar reflects the magnitude of sequence divergence per nucleotide
909 site.
910
911 **Figure 3:** A comparison of the likelihood tree based on the core gene alignments (from Figure 2,
912 on the left) to a dendrogram based on distances due to gene presence / absence (on the right). As
913 in Figure 2, the isolates are labelled at the tips of trees, with the colors representing plant order.
914 Both phylogenies contain three main *X. fastidiosa* clades, representing the three subspecies.
915 Lines connect the same isolate between the two trees, with angled lines representing topological

916 discordance between phylogenies. The three *Xylella* subspecies are outlined in a black box and
917 labelled.
918
919 **Figure 4:** The result of gene gain and loss analyses. A) The phylogeny of the isolates, with
920 branch lengths proportional to the number of gene gain events. The colored branches are
921 branches with outlier gene gain rates. B) The phylogeny of the isolates, with branch lengths
922 proportional to the number of gene loss events. The colored branches are branches with outlier
923 gene loss rates. C) A plot of the gene gains against sequence divergence; in the plot each dot
924 represents one of the 125 branches on the phylogeny. Outlier dots are colored red. D) As in C,
925 with gene losses plotted again sequence divergence.
926
927 **Figure 5:** Estimated values of ω under M0 (the one-ratio model) in the core and accessory genes.
928 The distribution of $\widehat{\omega}$ values is plotted for A) core genes and B) accessory genes. Histogram bars
929 are shaded to reflect the outcome of the likelihood ratio test (non-significant tests are colored red
930 and significant tests are colored blue) between a model that estimated $\widehat{\omega}$ and a model with ω
931 fixed to 1.0. The horizontal dashed line denotes $\widehat{\omega}$ for each gene set.
932
933
934
935
936
937
938
939
940
941
942
943
944
945

Figure 1:

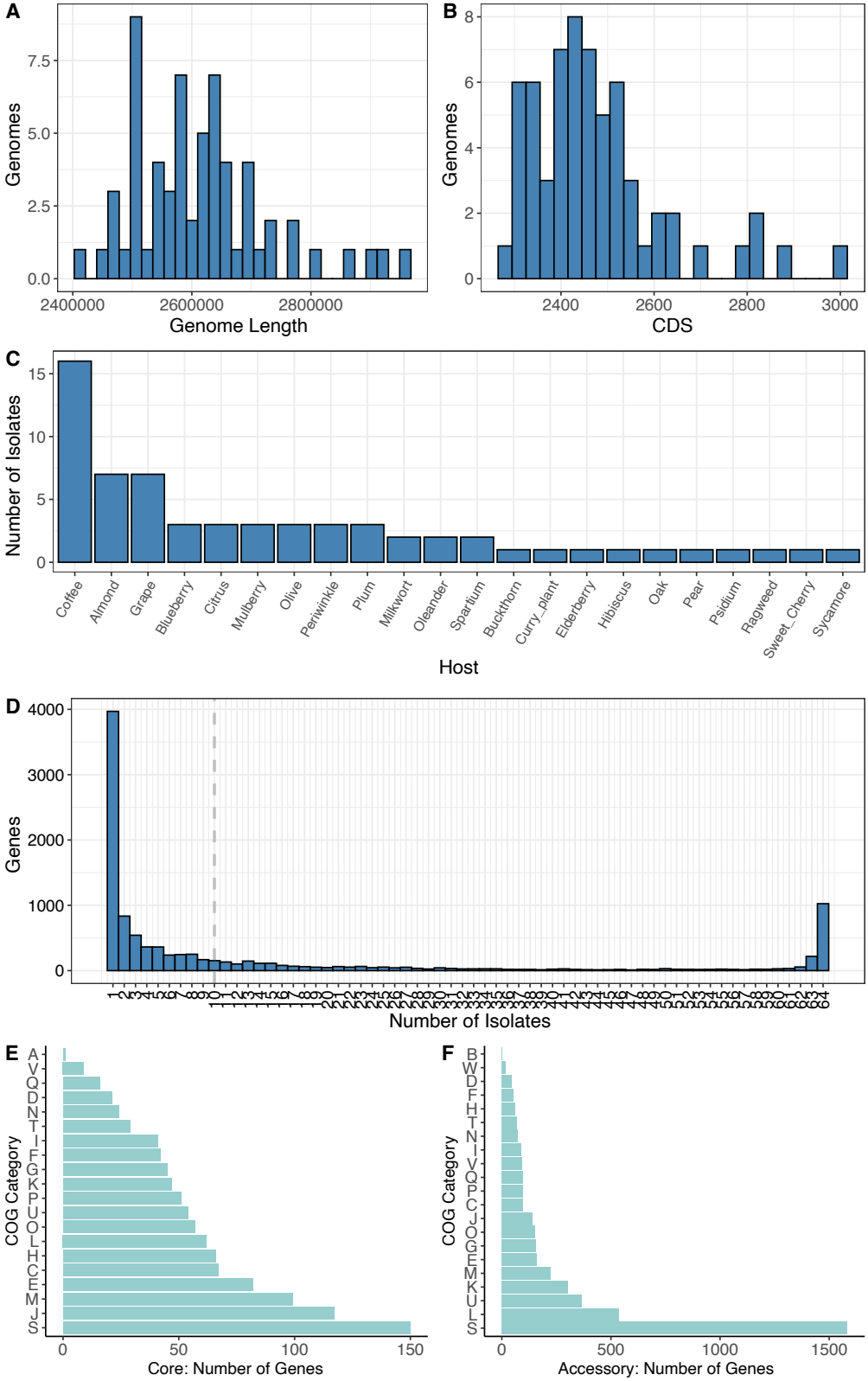

A
Genomes
Genome Length
B
Genomes
CDS
C
Number of Isolates
Host
Coffee
Almond
Grape
Blueberry
Citrus
Mulberry
Olive
Periwinkle
Plum
Milkwort
Oleander
Spartium
Buckthorn
Curry_plant
Elderberry
Hibiscus
Oak
Pear
Psidium
Ragweed
Sweet_Cherry
Sycamore
D
Genes
Number of Isolates
E
COG Category
Core: Number of Genes
F
COG Category
Accessory: Number of Genes

Figure 2:

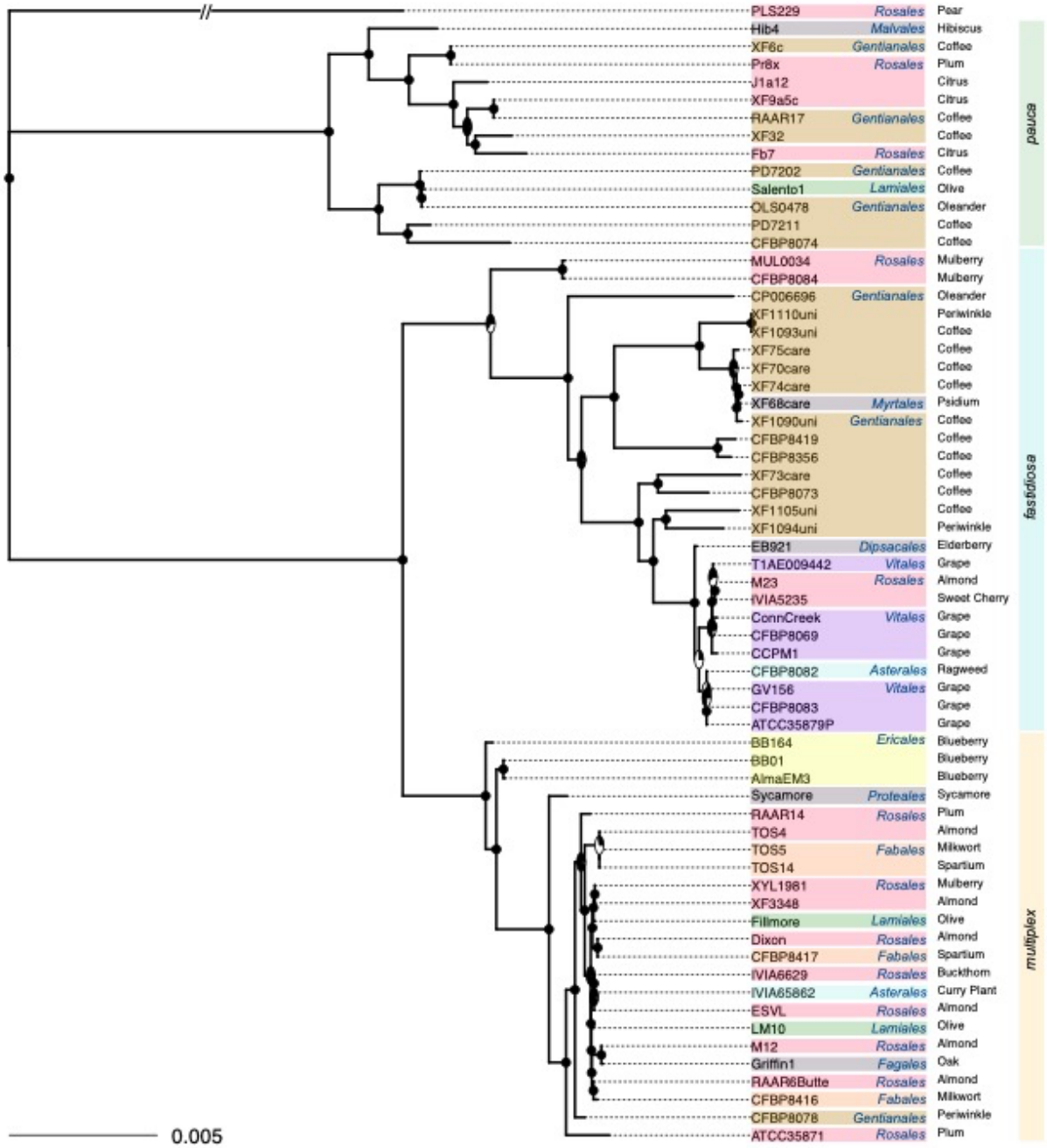

PLS229 Rosales Pear
Hib4 Malvales Hibiscus
XF6c Gentianales Coffee
Pr8x Rosales Plum
J1a12 Citrus
XF9a5c Citrus
RAAR17 Gentianales Coffee
XF32 Coffee
Fb7 Rosales Citrus
PD7202 Gentianales Coffee
Salento1 Lamiales Olive
OLS0478 Gentianales Oleander
PD7211 Coffee
CFBP8074 Coffee
MUL0034 Rosales Mulberry
CFBP8084 Mulberry
CP006696 Gentianales Oleander
XF1110uni Periwinkle
XF1093uni Coffee
XF75care Coffee
XF70care Coffee
XF74care Coffee
XF68care Myrtales Psidium
XF1090uni Gentianales Coffee
CFBP8419 Coffee
CFBP8356 Coffee
XF73care Coffee
CFBP8073 Coffee
XF1105uni Coffee
XF1094uni Periwinkle
EB921 Dipsacales Elderberry
T1AE009442 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:

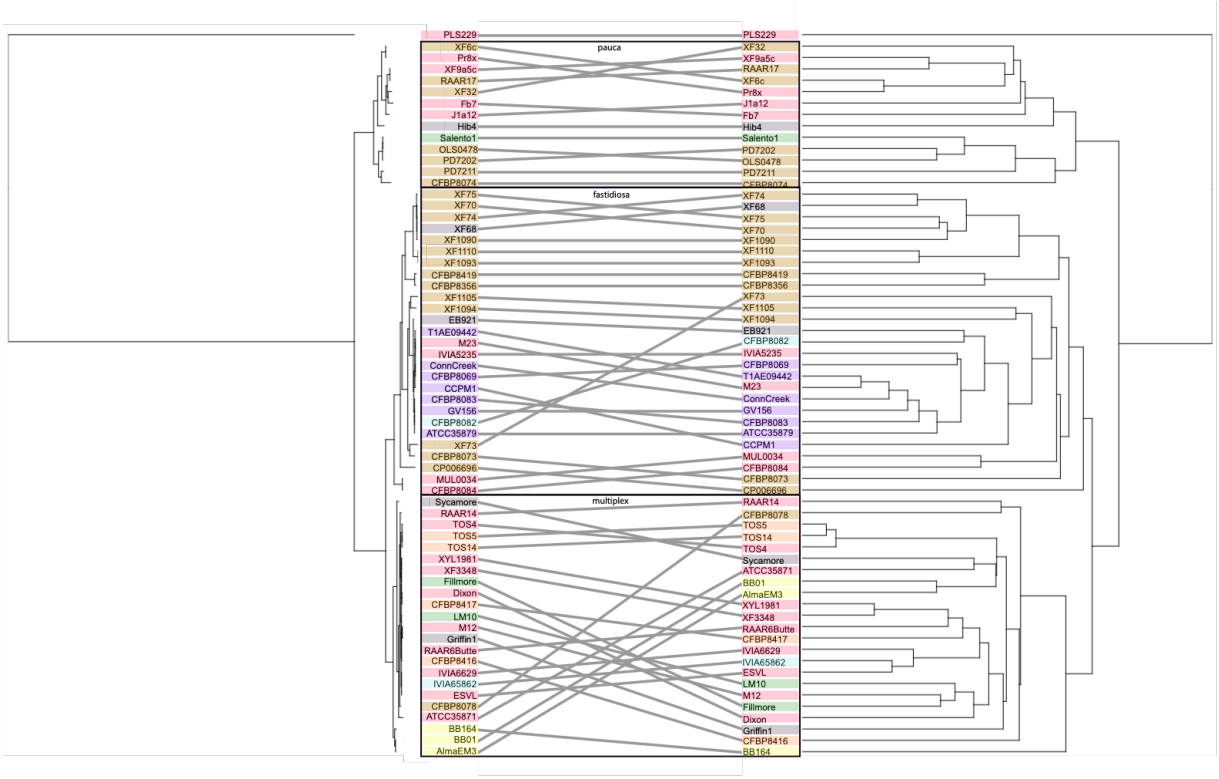
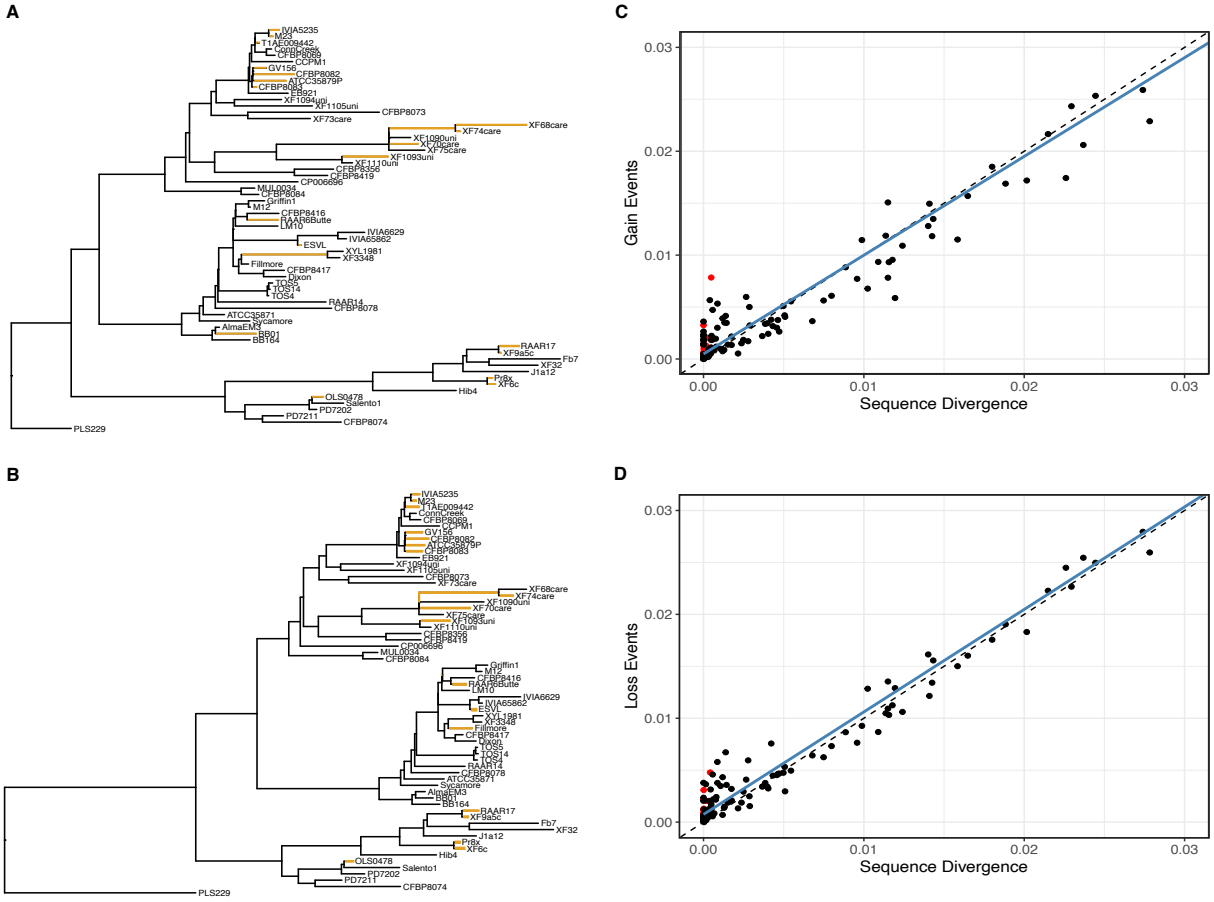
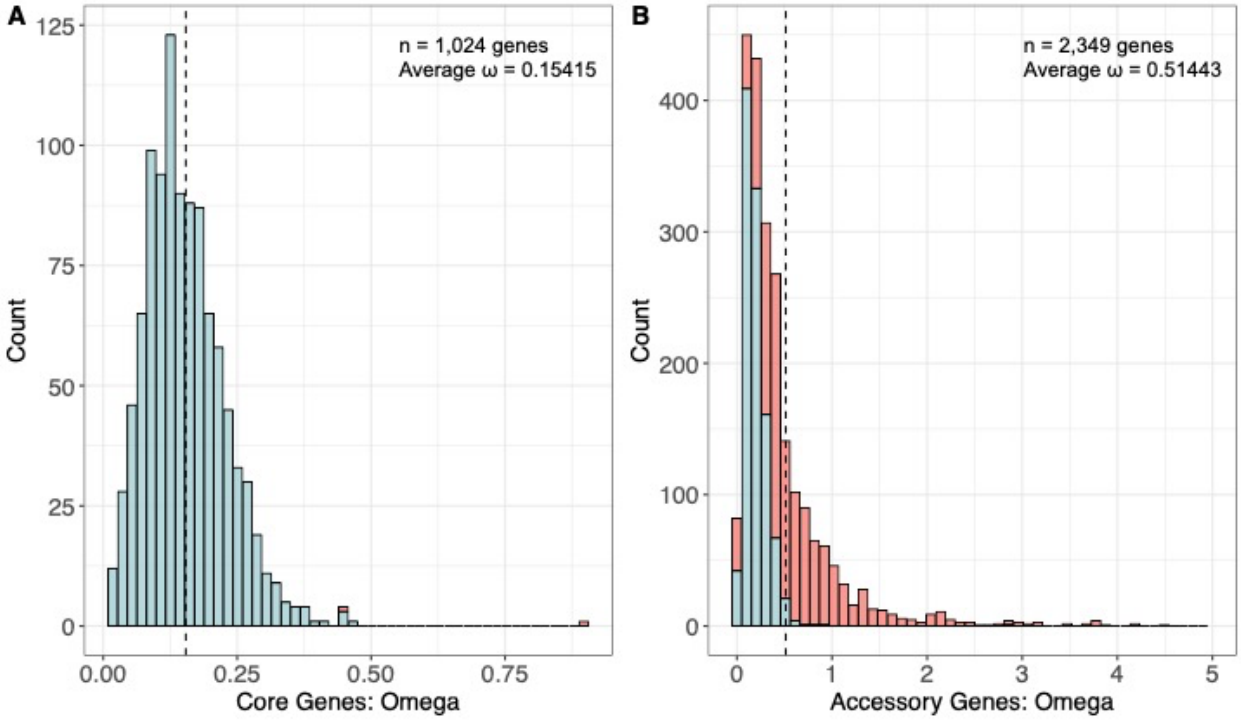

pauca
PLS229
XF6c
Pr8x
XF9a5c
RAAR17
XF32
Fb7
J1a12
Hib4
Salento1
OLS0478
PD7202
PD7211
CFBP8074
PLS229
XF32
XF9a5c
RAAR17
XF6c
Pr8x
J1a12
Fb7
Hib4
Salento1
PD7202
OLS0478
PD7211
CFBP8074
fastidiosa
XF75
XF70
XF74
XF68
XF1090
XF1110
XF1093
CFBP8419
CFBP8356
XF1105
XF1094
EB921
T1AE09442
M23
IVIA5235
ConnCreek
CFBP8069
CCPM1
CFBP8083
GV156
CFBP8082
ATCC35879
XF73
CFBP8073
CP006696
MUL0034
CFBP8084
XF74
XF68
XF75
XF70
XF1090
XF1110
XF1093
CFBP8419
CFBP8356
XF73
XF1105
XF1094
EB921
CFBP8082
IVIA5235
CFBP8069
T1AE09442
M23
ConnCreek
GV156
CFBP8083
ATCC35879
CCPM1
MUL0034
CFBP8084
CFBP8073
CP006696
multiplex
Sycamore
RAAR14
TOS4
TOS5
TOS14
XYL1981
XF3348
Fillmore
Dixon
CFBP8417
LM10
M12
Griffin1
RAAR6Butte
CFBP8416
IVIA6629
IVIA65862
ESVL
CFBP8078
ATCC35871
BB164
BB01
AlmaEM3
RAAR14
CFBP8078
TOS5
TOS14
TOS4
Sycamore
ATCC35871
BB01
AlmaEM3
XYL1981
XF3348
RAAR6Butte
CFBP8417
IVIA6629
IVIA65862
ESVL
LM10
M12
Fillmore
Dixon
Griffin1
CFBP8416
BB164

Figure 4:


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

966 **Figure 5:**
967



968
969