

1 **Metagenomic sequencing for rapid identification of *Xylella fastidiosa* from leaf**
2 **samples.**

3
4 Veronica Roman-Reyna^{1,2}, Enora Dupas^{3,4}, Sophie Cesbron³, Guido Marchi⁵, Sara
5 Campigli⁵, Mary Ann Hansen⁶, Elizabeth Bush⁶, Melanie Prarat⁷, Katherine Shiplett⁷,
6 Melanie L. Lewis Ivey⁸, Joy Pierzynski⁹, Sally A. Miller^{2,8}, Francesca Peduto Hand¹,
7 Marie-Agnes Jacques³, Jonathan M. Jacobs^{1,2,*}.

8
9 ¹ Department of Plant Pathology, The Ohio State University, Columbus, OH, USA

10 ² Infectious Disease Institute, The Ohio State University, Columbus, OH, USA

11 ³ Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France

12 ⁴ French Agency for Food, Environmental and Occupational Health & Safety, Plant
13 Health Laboratory, Angers, France

14 ⁵ Department of Agriculture, Food, Environment and Forestry, University of Florence,
15 Italy

16 ⁶ School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA

17 ⁷ Animal Disease Diagnostic Laboratory, Ohio Department of Agriculture, Reynoldsburg,
18 OH, USA

19 ⁸ Department of Plant Pathology, The Ohio State University, Wooster, OH, USA

20 ⁹ C. Wayne Ellett Plant and Pest Diagnostic Clinic, Department of Plant Pathology, The
21 Ohio State University Reynoldsburg, OH, USA

22
23 ***Correspondence: jacobs.1080@osu.edu**

24 **ABSTRACT**

25

26 *Xylella fastidiosa* (*Xf*) is a globally distributed plant pathogenic bacterium. The primary
27 control strategy for *Xf* diseases is eradicating infected plants; therefore, timely and
28 accurate detection is necessary to prevent crop losses and further pathogen dispersal.
29 Conventional *Xf* diagnostics primarily relies on quantitative PCR (qPCR) assays.
30 However, these methods do not consider new or emerging variants due to pathogen
31 genetic recombination and sensitivity limitations. We developed and tested a
32 metagenomics pipeline using in-house short-read sequencing as a complementary
33 approach for affordable, fast, and highly accurate *Xf* detection. We used metagenomics
34 to identify *Xf* to strain level in single and mixed infected plant samples at concentrations
35 as low as one picogram of bacterial DNA per gram of tissue. We also tested naturally
36 infected samples from various plant species originating from Europe and the United
37 States. We identified *Xf* subspecies in samples previously considered inconclusive with
38 real-time PCR ($C_q > 35$). Overall, we showed the versatility of the pipeline by using
39 different plant hosts and DNA extraction methods. Our pipeline provides taxonomic and
40 functional information for *Xf* diagnostics without extensive knowledge of the disease.
41 We hope this pipeline can be used for early detection of *Xf* and incorporated as a tool to
42 inform disease management strategies.

43

44 **IMPORTANCE**

45

46 *Xylella fastidiosa* (*Xf*) destructive outbreaks in Europe highlight this pathogen's capacity
47 to expand its host range and geographical distribution. The current disease diagnostic

48 approaches are limited by a multiple-step process, biases to known sequences, and
49 detection limits. We developed a low-cost, user-friendly metagenomic sequencing tool
50 for *Xf* detection. In less than three days, we were able to identify *Xf* subspecies and
51 strains in field-collected samples. Overall, our pipeline is a diagnostics tool that could be
52 easily extended to other plant-pathogen interactions and implemented for emerging
53 plant threat surveillance.

54

55 **KEYWORDS**

56 *Xylella fastidiosa*, metagenomics, diagnostics, short-read sequencing

57

58 **INTRODUCTION**

59

60 *Xylella fastidiosa* (*Xf*), is a globally distributed insect-transmitted plant pathogenic
61 bacterium, causing diseases on a large hosts range. To date, 595 plant species
62 grouped belonging to 85 botanical families have been reported as *Xf* hosts (1), some of
63 which are of major socio-economic interest, such as grapevine, olive, citrus, coffee and
64 almond (2). *Xf* colonizes the xylem vessels of plants where it forms biofilms (3) that,
65 together with tyloses and gums produced by the plant in response to the infection (4),
66 limit water translocation. Infected hosts display symptoms of leaf scorch and plant
67 dieback finally followed by plant death (3).

68

69 *Xf* was first described in and limited to the Americas but recently emerged in Europe,
70 highlighting the pathogen's capacity to expand its host range and geographical

71 distribution (2, 5). The pathogen was reported in Italy in 2013, where is currently
72 devastating Apulian olive production, then detected in France in 2015, Spain in 2016
73 and Portugal in 2018, both on cultivated as well as spontaneous Mediterranean plant
74 species (2). The primary control strategy for *Xf* diseases includes eradication of hosts;
75 therefore, fast and accurate detection is necessary to prevent major losses to growers
76 and further pathogen dispersal.

77
78 The diagnostic of diseases caused by fastidious pathogens like *Xf* is difficult. This
79 difficulty is increased as infected plants may remain asymptomatic for very long periods
80 of time, which is associated with low bacterial concentrations, and by an irregular
81 distribution of the pathogens in the plants (6). It is of major interest to develop reliable
82 and highly sensitive tools for detection and detailed identification that can be used
83 directly on plant extracts. Current standards for *Xf* diagnostics primarily rely on
84 quantitative real-time PCR (qPCR) assays to detect and sometimes identify the
85 bacterium (7–12), followed by the amplification and sequencing of two, for subspecies
86 identification, to seven housekeeping genes (*cysG*, *gltT*, *holC*, *leuA*, *malF*, *nuoL* and
87 *petC*) for is Sequence Type (ST) determination and phylogeny reconstruction (2) (Fig.
88 1A). Five subspecies are proposed in *X. fastidiosa*, ie. *fastidiosa*, *multiplex*, *pauca*,
89 *morus*, and *sandyi* (13–15). However, whole genome analyses revealed similarities of
90 the subspecies *fastidiosa*, *morus* and *sandyi*, which cluster into one clade. Moreover,
91 genome analysis indicated high frequency of horizontal gene transfer and
92 recombination among *Xf* subspecies (14–16).

93

94 Plant samples infected by more than one *Xf* strain belonging to several subspecies are
95 not uncommon and are not easy to detect, (17, 18). Nevertheless, current methods do
96 not consider new or emerging variants resulting from pathogen genetic recombination
97 (14). For example, qPCR with high Cq values (>35) are considered inconclusive (2),
98 making decisions about disease control difficult. A complementary tool for diagnostics is
99 the use of Next-Generation Sequencing (19) (Fig. 1). Because this approach can be
100 directly used on plant extracts, it is not biased towards known sequences and provides
101 more information about the pathogen genome, such as virulence traits. Metagenomics,
102 the study of genetic material from environmental samples, beyond whole genome
103 sequencing allows for the detection of strains from several subspecies and ST at the
104 same time from the host (20). Recently, the use of long-read sequencing as diagnostic
105 tool identified *Xf* subspecies and ST from infected samples (12, 21).

106
107 In this study, we developed and tested a metagenomics pipeline using in-house short-
108 read sequencing as a complementary approach for affordable and accurate *Xf*
109 detection. We were able to use metagenomics to identify *Xf* to strain level in single and
110 mixed infected plant samples, at concentrations as low as one picogram of bacterial
111 DNA per gram of tissue. In addition, we tested naturally infected field samples from
112 Europe and the United States. We identified *Xf* subspecies in samples with Cq values
113 equal to and greater than 37, which is beyond the threshold of detection for the
114 standard and certified qPCR methods (2). Overall, we developed a robust diagnostics
115 pipeline that could be easily extended to other pathogens and implemented for
116 surveillance of emerging agricultural threats.

117 **RESULTS**

118

119 **Metagenomics for diagnostics pipeline**

120 We developed and tested a metagenomics pipeline for *Xf* detection and subspecies
121 identification (Fig. 1). We tested this pipeline based on three types of DNA samples:
122 from bacterial colonies in culture, spiked plant samples, and naturally infected plant
123 samples (Fig. 1B). To recover and identify *Xf* subspecies and compare it to the already
124 sequenced genomes, we developed a pipeline that uses six different tools and custom-
125 made databases (22) (Fig. 1C). The pipeline recovers *Xf* reads with the software
126 Kraken2 and a custom-made database (22). The database has user-specified genomes
127 for *Xf* reads identification. The user-specified genomes belonged to *Xylella* (n=81),
128 *Xanthomonas* sp. (n=10), *E. coli* (n=1) and several plant sequences from NCBI (Table
129 S1). The database had plant sequences because some *Xf* genomes from NCBI
130 contained plant genomic DNA sequences. We could not clean all 18S sequences, plant
131 plastids, or chloroplast reads from the NCBI *Xf* genomes. Therefore, the plant reads in
132 the database serve as a filter to ensure plant reads were not misidentifying as *Xf* reads.
133

134 After Kraken2, the pipeline *de-novo* assembled the recovered *Xf* reads into contigs with
135 the program SPAdes (23). The pipeline used the *Xf* contigs for four different analyses:
136 1) subspecies identification, 2) phylogeny reconstruction, 3) identification of the already
137 sequenced genetically closest strains, and 4) alleles for MLST profile and virulence-
138 related genes determination. The pipeline used *Xf* contigs, and the tool SendSketch
139 from the BMAP software to identify subspecies. Then it used Pyani and LINbase

140 software to reconstruct phylogeny by ANI (24, 25). Next, it assigned *Xf* strains to each
141 *Xf* contig to identify the closest strains with the tool BBSplit from BMap software.
142 Finally, to identify specific genes or alleles, the pipeline used local BLAST with two type
143 of subjects: a) one subject was the reported MLST allele genes and b) the other subject
144 was the protein sequences from genes associated with virulence.

145

146 ***Xf* identification from *in-silico* prepared samples**

147 To test the pipeline sensitivity, we used *in-silico* samples with target (e.g. *Xf*) and non-
148 target DNAs (e.g. non-host plant). The samples included variable amounts of non-host
149 barley (*Hordeum vulgare*) sequenced reads *in silico* spiked with *Xff* CFBP 7970 reads.
150 We obtained a strong linear correlation between the Kraken2 results and the proportion
151 of spiked *Xf* sequence reads ($y = 103.21x - 0.0127$; $R^2 = 1$) (Fig. S1). We
152 recovered *Xf* reads and assembled them as contigs using SPAdes. With the *Xf* contigs,
153 we performed BLAST analysis to identify MLST alleles and virulence genes. We were
154 able to identify one to four MLST-related gene for samples spiked with 0.5 to 2.4%
155 *Xf* reads (Table S3). This result indicated that we cannot capture the full MLST gene set
156 for ST identification with less than 2.4% *Xf* reads (Table S3). We calculated, for all
157 samples, the percentage of gene similarity to the virulence-related genes (Table S4).
158 The percentage of gene similarity increased with the higher number of spiked *Xf* reads.
159 Samples with a lower number of *Xf* reads had a low genome coverage to recover and
160 analyze complete gene sequences (Table S4, S7).

161

162 We then identified *Xf* subspecies using the *Xf* contigs. Since the *in-silico* samples only
163 identified *Xff* reads, we expected that SendSketch assigned all contigs to *Xff*. However,
164 we found that 9 to 15% of *Xff* contigs were instead assigned to *Xfm*. Based on these
165 results, we did two additional analyses to determine the best approach to analyze the *Xf*
166 subspecies composition. For the first analysis, we hypothesized that complete
167 assembled genomes would reduce the percentage of reads assigned to other
168 subspecies. To test this, we created a smaller Kraken2 database with 30 genomes
169 instead of 81. These 30 *Xf* genomes had complete assemblies. We recovered 1 to 2%
170 fewer *Xf* reads with the new database, and the subspecies distribution remained the
171 same (data not shown). The results indicated that *Xf* subspecies classification is not
172 related to the level of genome assembly. Therefore, the original Kraken2 database with
173 81 NCBI *Xf* genomes was retained for all further analyses.

174
175 For the second analysis, we manually assessed the SendSketch sensitivity to mixed
176 infections with new *in-silico* samples. The new samples included a set amount of barley
177 reads *in silico* spiked with variable amounts of *Xff* CFBP 7970 and *Xfm* CFBP8418
178 reads (Table S2). For these new *in-silico* samples, we recovered *Xf* reads with Kraken2,
179 assembled the reads as contigs and run SendSketch to identify subspecies. When
180 using BLAST, a certain number of *Xf* contigs mapped equally (100% identity) to *Xff* and
181 *Xfm* (*Xf* core contigs). We observed that *Xf* core contigs are directly proportional to the
182 total of *Xf* recovered reads and samples with a higher *Xff* to *Xfm* spiked reads ratio
183 (Table S2, Fig. S2). Moreover, the tool SendSketch randomly assigned the subspecies
184 to *Xf* contigs with 100% identity to *Xff* and *Xfm*. Consequently, we developed a manual

185 correction to separate single from mixed infections. We only used samples with either
186 *Xfm* or *Xff*; *Xfp* was not part of the analysis. The correction consists of calculating the
187 logarithm of the *Xfm* : *Xff* contigs ratio. Corrected log-ratios from 0.081 to 0.4 are
188 considered a mixed infection. Log-ratios below 0.08 will be a single *Xff* infection, and
189 higher than 0.43 will be *Xfm* single infection (Fig. S2).

190
191 To evaluate the pipeline with samples free of *Xf*, we used extracted DNAs of two
192 healthy plant samples and a non-*Xylella* controls (i.e., barley leaves infiltrated with
193 *Xanthomonas*). For the artificially inoculate barley samples, Kraken2 software recovered
194 20 to 30% of the total reads as *Xanthomonas*. For all four *Xf*-free samples, Kraken2
195 recovered six to 19 *Xf* reads (Table S2). All these *Xf* reads corresponded to plant reads
196 based on the BLAST webtool from NCBI. Based on these results, the pipeline considers
197 a sample *Xf*-free when it cannot recover more than 19 *Xf* reads (Fig. S3).

198

199 ***Xf* identification from isolated bacteria**

200 To test the capacity of iSeq 100 sequencing, we used six gDNAs from isolated bacteria
201 and two known *Xf* gDNAs as control. The six *Xf* gDNAs were isolated from Italian field
202 samples (See Material and Methods). The two control *Xf* gDNA samples were *Xff* CFBP
203 7970 (CFBP 7970 iSeq100) and *Xfm* CFBP 8418 (CFBP 8418 iSeq100). All eight gDNA
204 samples were sequenced with the iSeq 100 System. For all eight samples, Kraken2
205 recovered 99% of the total reads as *Xf* reads (Table S2). We assembled the *Xf* reads as
206 contigs and classified them into subspecies. For CFBP 7970 and CFBP 8418 for which
207 a genome was already available, 54-78% of the contigs corresponded to *Xf* core

208 contigs, 44-65% to their respective subspecies, and 2% to the closest subspecies. On
209 average, within the six Italian samples, 20% of the contigs corresponded to *Xf* core
210 contigs, 25% to *Xfm*, and 2% to *Xff*.

211
212 The ANI values were consistent with the *Xf* contig abundance (Fig. 2, Table S2). All six
213 Italian samples and CFBP 8418 iSeq100 had 99 to 100% identity to *Xfm* and less than
214 97% identity to *Xff* and *Xfp*. The control sample, CFBP 7970 iSeq100, had 100% identity
215 to *Xff* and less than 98% identity to *Xfm* and *Xfp* (Fig. 2, Table S2).

216
217 For strain identification, we used the program BBSplit and the Harvest suite. For each
218 sample, we selected the top three closest strains based on the program BBSplit output.
219 Then, we used these closest strains and the sample to compare the number of single
220 nucleotide polymorphisms (SNPs) with the Harvest suite. For the sample CFBP 8418
221 iSeq100, the closest strain with 30 SNPs was *Xfm* CFBP 8418 (Table S5, S6). For the
222 sample CFBP 7970 iSeq100, the closest strain with fewer SNPs was *Xff* CFBP 7970.
223 The six Italian samples had the same three closest *Xfm* strains, TOS5, TOS4, and
224 TOS14. All six samples had fewer SNPs when compared against the strain *Xfm* TOS4.
225 The three TOS strains and the Italian samples were isolated from the outbreak area of
226 Monte Argentario, Tuscany, Italy (26).

227
228 We performed BLAST analysis to identify MLST alleles and virulence genes for all eight
229 isolated bacteria with the assembled *Xf* contigs. For virulence genes, the sample CFBP
230 7970 iSeq100 had 100% similarity to all *Xff* virulence genes except for *rpfE* (96.4%)

231 (Table S4). The sample CFBP 8418 iSeq100 had 100% similarity to all *Xfm* virulence
232 genes except for pilB (99.8%). The six Italian samples had the same similarity
233 percentages for all *Xfm* virulence genes except for hemagglutinin (95.7 to 100%). We
234 were able to identify all virulence genes and to complete the allelic profiles for ST
235 identification (Table S3). As expected, the ST identified for the sample CFBP 7970
236 iSeq100 was ST2, and the one for CFBP 8418 iSeq100 was ST6. The six Italian
237 samples had the same ST87 number.

238

239 ***Xf* identification from spiked plant samples**

240 We tested the pipeline with DNA extracted from grapevine petioles and midribs
241 artificially inoculated with known bacterial concentrations of the strain *Xff* CFBP 7970,
242 *Xfm* CFBP 8418, or an equal mix of both strains. Kraken2 output recovered 0.01% to
243 74.8% of the total sequences as *Xf* reads (Table S2). The percentage of
244 recovered *Xf* reads had a positive correlation with Log₁₀ CFU values ($R^2= 0.9876$) (Fig.
245 3A) and Cq values ($R^2= 0.9141$) (data not shown). The pipeline detected *Xf* with lowest
246 bacterial concentration tested in this study (1×10^4 CFU/ml), equivalent to a Cq 28.85
247 and $1.62 \text{ pg} \cdot \mu\text{L}^{-1}$.

248

249 After *Xf* contig assembly, we were able to identify the subspecies for all samples, and
250 the log-ratio separated single from mixed infections (Fig. 3B, 3C). The log-ratio for *Xff*-
251 single infections varied between -1.56 and -0.23. The log-ratio for *Xfm*-single infection
252 was 1.15, while for mixed-strains infection was 0.08. The ANI values confirmed the
253 *Xff* and *Xfm* subspecies for single infected samples (Fig. S4, Table S2). The mixed

254 sample (*Xff* CFBP 7970 + *Xfm* CFBP 8418) had a higher ANI value for *Xfm*. This result
255 was consistent with a higher number of *Xfm* contigs for the mixed-strain sample (Fig.
256 3C, Table S2).

257
258 Based on BBSplit results, the genetically closest strains sequenced for most of the *Xff*-
259 single infections, was CFBP 7970, followed by ATCC 35879, GV230, and TPD4 (Table
260 S5). For *Xfm*-single infection, the genetically closest *Xfm* strains were Dixon and CFBP
261 8418. For the mixed infection, 90% of *Xf* contigs were assigned to *Xfm* Dixon and 6%
262 to *Xff* strains.

263
264 With the assembled *Xf* contigs, we performed BLAST analysis to identify MLST alleles
265 and virulence genes. We identified the ST number for three of the seven artificially
266 inoculated samples. The mixed infected sample of grapevine inoculated with the strain
267 CFBP 8418 were identified as ST6 (Tables S4). The grapevine sample inoculated with
268 CFBP 7970 (10^8 CFU/ml) was ST2. We could not assign an ST the sample for
269 grapevine sample inoculated with the strain CFBP 7970 (10^7 CFU/ml) because we only
270 identified six of the seven MLST alleles. In contrast to MLST analysis, we detected at
271 least two virulence genes per sample (Table S4). The sample with the lowest CFU
272 values, CFBP 7970 (10^4 CFU/ml), had 42% similarity to *Xff* hemagglutinin and 41%
273 similarity to *Xfm* pilQ. For the remaining *Xff*-single infected samples, the percentage of
274 similarity to a single subspecies increased with the higher CFU number, which is also
275 associated with higher genome coverage (Table S7). For the *Xfm*-single infection, all

276 the virulence genes had 100% similarity to *Xfm*. For the mixed infection, all the virulence
277 genes had 100% similarity to *Xfm* and, on average, 98% to *Xff*.

278

279 ***Xf* identification from field-collected samples**

280 Finally, we tested the iSeq 100 sequencing capacity with European and American field
281 samples (Table S2). We used 24 samples with Cq values ranging from 21 to 40 based
282 on Harper's qPCR assay. We used three samples that were negative based on the
283 same qPCR assay. The DNAs from the 27 samples were extracted from six different
284 hosts: *Olea europaea*, *Polygala myrtifolia* (France and Italy), *Quercus ilex*, *Spartium*
285 *junceum*, *Rhamnus alaternus*, and *Vitis vinifera*.

286

287 Kraken2 recovered 0.004% to 1.43% of the total reads as *Xf* (Table S2). We assembled
288 the *Xf* reads into one to 2896 contigs. We found all samples had at least one contig with
289 at least 400 bp. The limit of *Xf* detection with Harper and tetraplex qPCR correspond to
290 30-37 Cq values (17). Therefore, we evaluated 16 samples that either had less than 30
291 *Xf* contigs, were classified as *Xf*-negative, or had Cq higher than 30 (2). We used each
292 *Xf* contig from the 16 samples as query for a Nucleotide BLAST search using webtool
293 from NCBI. Eleven of the 16 samples gave 100% identity to *Xf* genomes. Hence, these
294 11 samples were considered *Xf*-positive. All contigs from the other five samples FR1-
295 Pm, FR1-Oe, IT6-Sj, IT11-Sj, and US1-Vv, had 100% identity to chloroplast and 18S
296 plant sequences but none to *Xf*. Therefore, these five samples were considered *Xf*-
297 negative (Fig. 4). With our pipeline, we were able to detect *Xf* in samples considered
298 inconclusive by qPCR according to Harper's (EPPO 2019).

299
300 We then used the contigs from the 22 *Xf*-positive samples for subspecies classification.
301 Overall, the samples had 50 to 100% of contigs classified as *Xf* core contigs (Fig. 4).
302 Three French samples (FR2-Qi, FR2-Pm, FR4-Oe) and five Italian samples (IT5-Sj, IT7-
303 Sj, IT8-Sj, IT9-Sj, IT6-Ra) had 1 to 6 contigs assigned as *Xfm*. The French sample FR3-
304 Oe, seven Italian samples (IT2-Sj, IT3-Sj, IT4-Sj, IT12-Sj, IT4-Ra, IT2-Pm, IT3-Pm) and
305 the USA sample US2-Vv, were *Xfm*-single infected based on the manual log correction
306 (log-ratio > 0.43) (Table S2). The sample FR2-Oe had 17% contigs assigned to *Xfp* and
307 the sample US3-Vv was *Xff*-single infected (log-ratio < 0.08).

308
309 Then, we determined the ANI values and *Xf* strain composition for the 22 *Xf*-positive
310 samples. Both results were consistent with the subspecies identification. The Italian
311 samples had 99-100% ANI to *Xfm*. Five French samples (FR1-Qi, FR2-Pm, FR3-Oe,
312 FR4-Oe, FR2-Qi) had 99-100% ANI to *Xfm* and FR2-Oe 98% ANI to *Xfp* (Fig. 5A, Table
313 S2). The sample US2-Vv had 99% ANI to *Xfm*, and US3-Vv had 100% to *Xff* (Fig. 5A).
314 For strain distribution, all the French, USA, and three Italian samples (IT2-Sj, IT4-Sj and
315 IT5-Sj) had more than 40% *Xf* contigs had 100% identity to one strain (Table S5, Fig.
316 5B). Except for IT2-Sj, Italian samples had most of the contigs assigned to the three
317 strains TOS4, TOS5, and TOS14. The sample IT2-Sj had more reads assigned to *Xfm*
318 RAAR14. Overall, the tools SendSketch, Pyani and Bbsplit validated the qPCR
319 subspecies results for field samples.

320

321 We performed BLAST analysis to identify MLST alleles and virulence genes for the 22
322 infected samples. For 16 of 22 *Xf*-positive samples, we found the percentage of gene
323 similarity to be 26 to 100% for at least one virulence gene (Table S4). Distinct from
324 single gene analysis, we only identified some MLST-related alleles for four samples
325 (US3-Vv, FR2-Oe, IT2-Pm, IT3-Pm); consequently, we could not identify the ST number
326 (Table S3).

327
328 To compare some of our results with a high-performance, deep-sequencing Illumina
329 platform as a control, we selected nine samples for re-sequencing with the MiSeq
330 platform using the same iSeq 100 libraries from this study. The nine samples were IT3-
331 Pm, IT5-Sj, FR2-Oe, FR3-Oe, US1-Vv, US2-Vv, FR1-Pm, FR2-Pm, IT9-Sj (Table S8,
332 Fig. S5). We analyzed the MiSeq sequences with our pipeline and recovered 0.005%-
333 0.792% of total reads as *Xf* reads with Kraken2. We assembled the *Xf* reads into
334 contigs and manually assessed all samples with less than 30 *Xf* contigs. The NCBI
335 Blastn analysis indicated that the samples US1-Vv and FR1-Pm, had contigs with 100%
336 identity to plant reads; therefore, we confirmed they were *Xf*-negative samples. The
337 other seven samples were considered *Xf*-positive. We followed the pipeline to identify
338 and determine subspecies, phylogeny, genetically closest sequenced genome strains,
339 MLST profile, and virulence-related genes. The results for subspecies and phylogeny
340 identification were the same between MiSeq and iSeq100 sequencers (Fig. S5), but
341 there were some differences for the other three analysis results (Table S2, Table S8).
342 For the genetically closest sequenced genome strains analysis, all samples gave the
343 same strain distribution as iSeq100 results, except FR2-Oe which showed *Xfp* OLS0478

344 instead of *Xfp* COF0407 as the most abundant strain. These two *Xfp* strains are
345 phylogenetically close. For MLST analysis, we identified four more alleles for the
346 sample IT3-Pm with the MiSeq platform than with iSeq100 (Table S3), while we only
347 detected two alleles in the sample US2-Vv sequenced with Miseq. We were only able to
348 detect MLST alleles for the sample FR2-Oe with the iSeq100 platform. We were able to
349 calculate the percentage of gene similarity for more virulence genes with the MiSeq
350 platform than with the iSeq100 platform. The variation between Illumina platforms was
351 not consistent.

352

353 **DISCUSSION**

354

355 In this study, we developed a user- friendly metagenomic pipeline to identify and
356 determine *Xylella fastidiosa* subspecies from field-collected samples without the need
357 for pathogen isolation. We demonstrated the flexibility of the pipeline by using seven
358 different plant hosts and three DNA extraction methods. We recovered and
359 assembled *Xf* reads into contigs from total DNA samples. We used percentage of
360 similarity to a single subspecies to identify *Xf* subspecies and validated the results
361 through phylogeny and strain proximity. Finally, we examined potential virulence-related
362 genes among all sequenced samples.

363

364 To recover *Xf* reads from field samples with the tool Kraken2, we used *Xf* genomes
365 available on NCBI. We found plant plastid reads in all genomes obtained from pure *Xf*
366 cultures, except for *Xff* CFBP7970. We decided to add plant reads in the database to

367 filter out potential plant contamination. We still recovered *Xf* reads for some plant
368 samples reported as *Xf*-negative. Therefore, we manually examined the contigs of
369 samples with less than 30 *Xf* contigs to determine if they have a low *Xf* concentration or
370 are negative samples. More than the Cq values, the contig evaluation will be necessary
371 to determine if a sample is truly *Xf*-negative.

372
373 We observed that the Kraken2 tool not only recovered *Xf* reads but also
374 identified *Xf* subspecies. We decided not to use Kraken2 to identify subspecies because
375 we found that the recovery is affected by incomplete NCBI *Xf* genomes subspecies
376 information. Kraken2 uses by default the NCBI taxonomy to classify reads, if the
377 genomes used to build the database does not have a subspecies information it will keep
378 most of the reads at the species level. To improve the subspecies resolution, we
379 decided to use contigs instead of reads and the tool SendSketch (BBMap tool). Contigs
380 or assembled reads increases the coverage and reduces false-positive reads (27). We
381 used Sendskecth because it uses MinHash algorithm to be fast and it takes into account
382 whole genomes and do not uses taxonomy.

383
384 After we identified samples as *Xf*-positive, we defined *Xf* subspecies. We observed that
385 some samples had mapped contigs to both *Xff* and *Xfm*. These are contigs most likely
386 associated with core sequences as only 3% of *Xff* and *Xfm* genomes are different.
387 However, it is also possible to have a percentage of annotation error due to sequencing
388 contamination (28, 29). To determine if the samples were single- or mixed-infected, we
389 corrected the results by calculating the Log *Xfm*: *Xff* contigs ratio (see Materials and

390 methods). Once we defined the Log values for *Xff*, *Xfm*, and *Xfm+Xff* (mixed) infections,
391 we validated the presence of single *Xfm* infections in all the tested European samples.
392 Identifying *Xf* subspecies in ornamental and crop plants is essential for the correct
393 application of eradication measures or for plant movements within the EU territory
394 according to regulation (EU) 2020/1201 (30).

395
396 The number of reads generated by the iSeq100 sequencer highlighted some limitations
397 of the pipeline. For example, the ST were not determined for any field sample because
398 we could not recover the complete sequences of the seven genes. This is probably
399 associated with low genome coverage. The low number of reads also hampers deep
400 SNP diversity and intersubspecific homologous recombination analyses. Other
401 sequencing systems, with higher reads output than the iseq100, can also be used with
402 this pipeline as the input is fastq files. With a high number of reads, the pipeline will
403 provide better resolution to recover the MSLT genes.

404
405 Some of the diagnostic tools for *Xf* diagnostics are qPCR and Sanger sequencing.
406 These tools require amplification of known *Xf* genome regions, but as is the case of
407 MLST, do not consider new or emerging variants. Moreover, these tools introduce bias
408 due to primer design, have unresolved results with high Cq values, and may take longer
409 since they follow a multistep process. Our pipeline complements these conventional
410 tools by obtaining metagenomic data directly from symptomatic or asymptomatic
411 samples and increasing the detection power. We found some discrepancies between
412 the number of recovered *Xf* reads and Cq values. These differences could be caused by

413 PCR inhibitors or the genomic target region for qPCR that underestimates the bacterial
414 concentration (31, 32). Metagenomics sequencing is becoming a more affordable and
415 faster approach for diagnostics. For example, the whole detection/identification with
416 qPCR and MLST scheme could have an estimated cost of 52-54USD per sample and
417 takes three to four days to detect one to seven genes. With the iSeq100, it could cost
418 50-70USD (when having 12 samples in the same run) and take two days but while also
419 allowing to get a complete genomic analysis of the plant and its pathogenic and
420 commensal microbiota.

421
422 In conclusion, our pipeline provides *Xf* taxonomy and functional information for
423 diagnostics without extensive knowledge of the host or pathogen. The pipeline
424 databases used for the analysis can be public repositories or privately collected gDNA
425 and could be adapted by the user and tailored to different plant pathogens. The
426 sequencing can be adapted to be an in-house system as the library preparation and
427 sequencing are user-friendly and not limited by the DNA quality or quantity. The
428 analysis can be adjusted to detect several pathogens simultaneously. We hope this
429 pipeline can be used for early detection of *Xf* or other crop pathogens and incorporated
430 as part of management strategies.

431

432 **MATERIAL AND METHODS**

433 ***Xylella fastidiosa* strains**

434 The strain *Xf* subsp. *fastidiosa* (*Xff*) CFBP 7970 isolated in the United States (Florida) in
435 2013 from *Vitis vinifera* and *Xf* subsp. *multiplex* (*Xfm*) CFBP 8418 isolated in France in

436 2015 from *Spartium junceum* were provided by the French Collection of Plant-
437 Associated Bacteria (CIRM-CFBP (CIRM-CFBP [https://www6.inra.fr/cirm_eng/CFBP_-](https://www6.inra.fr/cirm_eng/CFBP_-Plant-Associated-Bacteria)
438 Plant-Associated-Bacteria) and used as controls for whole-genome sequencing. Both
439 strains were cultivated on modified PWG medium (Gelrite 12 g. L⁻¹; soytone 4 g. L⁻¹;
440 bacto tryptone 1 g. L⁻¹; MgSO₄. 7H₂O 0, 4 g. L⁻¹; K₂HPO₄ 1. 2 g. L⁻¹; KH₂PO₄ 1 g. L⁻¹;
441 hemin chloride (0. 1% in NaOH 0. 05 M) 10 ml. L⁻¹; BSA (7. 5%) 24 ml. L⁻¹ ; L-glutamine
442 4 g. L⁻¹) at 28°C for one week.

443

444 **Artificially inoculated plant samples**

445 For artificial inoculations, 10 ml of a calibrated CFBP 7970 or CFBP 8418 strain
446 suspension was spiked in 2 g of detached *V. vinifera* leaves. Sterile water was used for
447 negative controls. The DNA extraction was performed using a CTAB-based extraction
448 protocol (2) with slight modifications in order to concentrate bacterial DNA. After a 20-
449 min centrifugation at 20,000 g of the sample, the pellet was resuspended in 1 ml of
450 CTAB buffer. At the end of the extraction protocol, the pellet was resuspended in 50 µl
451 of sterile demineralized water. *Xf* presence in the infected samples was checked using
452 Harper's qPCR assay (8).

453

454 **Plant material and bacterial gDNA**

455 Healthy plant material of *Vitis vinifera* (2 g of leaf petioles) was spiked with 10 ml of a
456 calibrated CFBP 7970 or CFBP 8418 strains suspension. Sterile water was used as
457 negative control. The DNA extraction was performed using a CTAB-based extraction
458 protocol (2) with slight modifications in order to concentrate bacterial DNA. After a 20-

459 min centrifugation at 20,000 g of the plant macerate sample was resuspended in 1 ml of
460 CTAB buffer. At the end of the extraction protocol, the pellet was resuspended in 50 µl
461 of sterile demineralized water. *Xf* presence in the infected samples was checked using
462 Harper's qPCR assay (8).

463

464 Naturally infected samples were collected in Europe and the USA. Symptomatic
465 samples of *Olea europaea*, *Polygala myrtifolia* and *Quercus ilex* were collected in
466 October 2018 in Corsica (France) and in September 2019 in the French Riviera. *Xf*
467 detection and DNA extraction of whole infected-plant tissue were performed as
468 mentioned above. *Xf* subspecies were identified using the tetraplex qPCR (17).

469

470 Twig tissues, leaf petioles or green shoots of *Rhamnus alaternus*, *Spartium junceum*
471 and *P. myrtifolia* growing in the *Xf* outbreak zone of Monte Argentario (Grosseto,
472 Tuscany, Italy) were collected during 2019 and 2020 (33). *Xf* was detected using
473 Harper's qPCR assay (8). For samples with a Cq value lower than 30, *Xf* isolation was
474 attempted on Buffered Charcoal Yeast Extract (BCYE) agar according to PM 7/24-4 (2).
475 Bacterial isolates that became visible to the unaided eye within three days of incubation
476 at 28°C were discarded; those that became visible thereafter were streaked twice for
477 purity on BCYE agar and identified as *Xf* based on qPCR results (8). Reactions were
478 carried out after boiling bacterial suspension for 10 min. The DNA of one isolate among
479 those that tested positive by qPCR from each plant, was extracted using the CTAB
480 based protocol and further characterized to subspecies and ST level following the Multi
481 Locus Sequence Typing (MLST) approach (34). The GoTaq probe qPCR Master Mix

482 (Promega, A6102) and GoTaq G2 (Promega, M784B) polymerase were used for qPCR
483 and conventional PCR experiments, respectively. The bacterial DNA of three isolates
484 from *R. alaternus* (IT1-Ra to IT3-Ra), one from *S. junceum* (IT1-Sj), one from *P.*
485 *myrtifolia* (IT1-Pm) and one from *Prunus dulcis* (IT1-Pd), were sequenced in this study.
486 The assembled genomes of the six isolates were deposited on NCBI under the
487 Bioproject PRJNA728043.

488
489 *Vitis vinifera* DNA samples were received from the Virginia Tech Plant Disease Clinic
490 (VA, USA). All samples were collected from Virginia vineyards in 2019. US2-Vv sample
491 was collected from the vineyard in Greene County and US3-Vv sample was from a
492 vineyard in Isle of Wight County. The DNA extraction and *Xf* detection protocols are
493 based on work instructions from the Virginia Tech Plant Disease Clinic (VTPDC).
494 Approximately 50-100 mg of grape leaf or petiole tissue was excised from each sample
495 using a razor blade. The excised tissue was transferred to lysing Matrix A tubes (MP,
496 6910-500) and ground using a FastPrep 24 (MP, 116004500). DNA was extracted using
497 ISOLATE II Plant DNA extraction kit (Bioline, 52070) following manufacturer's
498 recommendations and CTAB lysis buffer. For *Xf* detection, qPCR Harper's was
499 performed on the StepOnePlus™ system (Life Technologies, 4376600) with Sensi-
500 FAST Probe Hi-ROX qPCR kit (Bioline, 82005) (8).

501

502 **Pipeline controls**

503 To test the pipeline, seven samples were used as negative controls. The controls were
504 two DNA samples from healthy barley and wild grass leaves that were grown in a

505 greenhouse, two DNA samples from barley leaves were infiltrated with a bacterial
506 suspension (10^8 CFU/ml) of *Xanthomonas translucens* pv. *translucens* UPB886. Three
507 healthy samples were collected in France in 2020 and in the USA in 2019. Petioles and
508 midribs were collected from healthy *Olea europaea* plants in a non-*Xf* infected area
509 (Angers, France) and from *Polygala myrtifolia* plants that were purchased from a local
510 nursery. *V. vinifera* leaves were collected from the vineyard in Greene County (Va,
511 USA). The DNA from these samples was extracted using CTAB method as mentioned
512 above. The absence of *Xf* in the healthy plants was confirmed using Harper's qPCR
513 assay (8).

514
515 For *in-silico* pipeline controls, two types of positive controls were used. To validate the
516 detection of different concentrations of *Xf*, barley fasta sequence files were *in-silico*
517 mixed with reads of the *Xff* CFBP 7970 sequenced in this study. The final proportion of
518 *Xff* reads in the sample ranged from 0.2 to 2.4% of total reads. To validate the detection
519 limits for single and mixed infections, the barley fasta sequence files were *in silico*
520 mixed with different proportion of *Xff* CFBP 7970 and *Xfm* CFBP 8418 reads, to get
521 from % CFBP7970 : 99% CFBP8418 to 99% CFBP7970 : 1% CFBP8418.

522

523 **iSeq 100 sequencing**

524 iSeq 100 sequencing libraries were prepared according to the Illumina reference guide
525 for Nextera DNA Flex Library Prep and Nextera DNA CD Indexes. In brief, 200 to 500
526 ng of DNA were quantified by spectrophotometry and used for library preparation. Then,
527 the libraries were diluted to have the same starting concentration prior to sample

528 pooling. Eight to 12 libraries were mixed together, and 1 nM of pooled library was used
529 per run. The sequencing settings were paired-ended (PE) read type, 151 read cycles
530 and eight index cycles. In the iSeq 100 System, the illumina GenerateFASTQ Analysis
531 Module for base calling and demultiplexing was selected. After sequencing for 17 h,
532 fastq paired end read files were extracted from the machine for subsequent analysis.

533

534 **Pipeline for *Xylella* sp. detection, classification and quantification via** 535 **metagenomic analysis**

536 Fastq files and the program Kraken2 were used to recover *Xylella* reads (22). The
537 Kraken2 command options were: --paired, --minimum-hit-groups 5, --report and --db.
538 The database was created with 92 NCBI genomes: 79 from *Xf*, two from *Xylella*
539 *taiwanensis*, ten from *Xanthomonas* sp. and one from *Escherichia coli* (Table S1). The
540 81 *Xylella* genomes were used to recover *Xylella* reads. The last 11 genomes were
541 added to remove reads common to Proteobacteria that might give a false positive
542 match. The tool SendSketch with nt server was used to make sure the 81 NCBI *Xf*
543 genomes were not contaminated with plant reads (BBMap – Bushnell B. –
544 sourceforge.net/projects/bbmap/; October 30, 2019). Forty-nine NCBI sequences from
545 plant 18S and chloroplast were added to the Kraken2 customized database to avoid
546 extracting reads annotated as plant reads (Table S1).

547

548 The reads classified as *Xylella* were extracted with the script `extract_kraken_reads` from
549 the KrakenTools suite (GitHub [jenniferlu717/KrakenTools](https://github.com/jenniferlu717/KrakenTools)). The extracted *Xf* reads were
550 used for downstream analysis. First *Xf* reads were *de-novo* assembled with the software

551 SPAdes (23) using defaults settings and the option --only-assembler. Second, the *Xf*
552 contigs were the query sequences in the Basic Local Alignment Search Tool website
553 (NCBI) to confirm if they were *Xf* reads, or misclassified plant reads. The Blastn
554 parameters were Nucleotide collection (nr/nt) as database and megablast program
555 selection.

556
557 The *Xf*-positive contigs were used in four different analyses: 1) to identify subspecies, 2)
558 to reconstruct phylogeny, 3) to identify the genetically closest strains already sequenced
559 and 4) to identify alleles from specific genes and the MLST profile.

560
561 To identify subspecies, the tool SendSketch was run with the parameters,
562 mode=sequence, records=2, and format=3, minani=100, minhit=1, address=ref, level=0
563 (BBMap – Bushnell B. – sourceforge.net/projects/bbmap/). Only contigs with 100%
564 average nucleotide identity (ANI) were used to identify *Xf* subspecies. The *Xf* contigs
565 with no hits were considered as core sequences. For visualization, results were plotted
566 using stacked bars.

567
568 To reconstruct phylogeny, ANI was calculated using the software Pyani (v. 0. 2. 10) and
569 LINbase (24, 25). For Pyani, the option -ANIm was set and for LINbase the “Identify
570 using a gene sequence” was set as identification method. The R package
571 ComplexHeatmap was used to visualize the ANI cluster analysis with the parameters
572 clustering_distance_rows = robust_dist and clustering_method_rows = "average".
573 Robust_dist was a function suggested by the ComplexHeatmap tutorial.

574

575 To identify the genetically closest already sequenced *Xf* strains, the tool BBSplit was
576 run with *Xf*-contigs (BBMap – Bushnell B. – sourceforge.net/projects/bbmap/). The tool
577 used 81 *Xf* genomes from NCBI and the options minratio=1 ambig=best. For
578 comparisons, each sample was normalized to their total *Xf* contigs and plotted using the
579 R package ComplexHeatmap. For the isolated bacterial genomes, the most abundant
580 strains were used as reference to identify genomics variants using Harvest suite tools
581 (35).

582

583 To determine specific gene alleles, percentage of identity was calculated using *Xf*
584 contigs as query and the Blastn algorithm (Nucleotide-Nucleotide BLAST 2. 8. 1+). The
585 database contained complete nucleotide sequences for all alleles for the seven genes
586 used for ST identification (*cysG*, *gltT*, *leuA*, *malF*, *nuoL*, *holC* and *petC*) (34). All 147
587 alleles were downloaded from the website PubMLST (36) (Last updated: 2019-03-06).

588 To determine the presence of reported *Xf* virulence-related or common to several plant
589 pathogenic bacterial genes (37, 38), percentage of similarity was calculated using *Xf*
590 contigs as query using Blastx algorithm (Nucleotide-Nucleotide BLAST 2. 8. 1+. The
591 databases contained complete amino acid sequences for gumBCDE, pilBMQTVW,
592 rpfCEFG, tolC, 6-phosphogluconolactonase (pgl), and hemagglutinin from the *Xfm* M12
593 and *Xff* M23 NCBI genomes.

594

595

596

597 **MiSeq sequencing**

598 To validate the iSeq 100 results, the same iSeq 100 libraries were used for MiSeq deep
599 sequencing. Nine samples were selected, at least one from each iSeq 100 run, making
600 sure not to use the same i5 and i7 tags. These nine libraries were sent to the Animal
601 Disease Diagnostic Laboratory (Ohio Department of Agriculture, Reynoldsburg, Ohio)
602 for sequencing. Library preparation was performed using an Illumina DNA Flex kit, and
603 2x250 sequencing was performed on the MiSeq platform using V3 chemistry. The
604 pipeline described above was used to analyze the MiSeq fastq files.

605

606 **ACKNOWLEDGEMENTS**

607 The authors are grateful for funding support from the Ohio Department of Agriculture
608 Specialty Crops Block Grant (AGR-SCG-19-03) and USDA NIFA FACT (2021-67021-
609 34343) to JMJ. Sara Campigli was financed by a grant from the Phytosanitary Service
610 of the Tuscany Region (Italy). We would like to thank The Ohio Supercomputer Center
611 for providing High Performance Computing resources.

612

613 We would also like to thank Dr. Stephen Cohen, Dr. Jeff Chang, and Dr. Alexandra
614 Weinsberg for their valuable input for the experiment development and editing.

615

616 **REFERENCES**

617

- 618 1. EFSA. 2020. Update of the *Xylella* spp. host plant database – systematic literature
619 search up to 30 June 2019. EFSA J 18:e06114.

- 620 2. EPPO. 2019. PM 7/24 (4) *Xylella fastidiosa*. EPPO Bull 49:175–227.
- 621 3. Chatterjee S, Wistrom C, Lindow SE. 2008. A cell–cell signaling sensor is required
622 for virulence and insect transmission of *Xylella fastidiosa*. Proc Natl Acad Sci U S A
623 105:2670–2675.
- 624 4. Roper C, Castro C, Ingel B. 2019. *Xylella fastidiosa*: bacterial parasitism with
625 hallmarks of commensalism. Curr Opin Plant Biol 50:140–147.
- 626 5. Pierce NB (Newton B. 1892. The California vine disease : a preliminary report of
627 investigations. Washington : G.P.O.
- 628 6. EFSA. 2018. Updated pest categorisation of *Xylella fastidiosa*. Eur Food Saf Auth
629 16:1–61.
- 630 7. Francis M, Lin H, Rosa JC-L, Doddapaneni H, Civerolo EL. 2006. Genome-based
631 PCR Primers for Specific and Sensitive Detection and Quantification of *Xylella*
632 *fastidiosa*. Eur J Plant Pathol 115:203–213.
- 633 8. Harper SJ, Ward LI, Clover GRG. 2010. Development of LAMP and real-time PCR
634 methods for the rapid detection of *Xylella fastidiosa* for quarantine and field
635 applications. Phytopathology 100:1282–1288.
- 636 9. Li W, Teixeira DC, Hartung JS, Huang Q, Duan Y, Zhou L, Chen J, Lin H, Lopes S,
637 Ayres AJ, Levy L. 2013. Development and systematic validation of qPCR assays
638 for rapid and reliable differentiation of *Xylella fastidiosa* strains causing citrus
639 variegated chlorosis. J Microbiol Methods 92:79–89.

- 640 10. Ouyang P, Arif M, Fletcher J, Melcher U, Corona FMO. 2013. Enhanced Reliability
641 and Accuracy for Field Deployable Bioforensic Detection and Discrimination of
642 *Xylella fastidiosa* subsp. *pauca*, Causal Agent of Citrus Variegated Chlorosis Using
643 Razor Ex Technology and TaqMan Quantitative PCR. PLOS ONE 8:e81647.
- 644 11. Ito T, Suzaki K. 2017. Universal detection of phytoplasmas and *Xylella* spp. by
645 TaqMan singleplex and multiplex real-time PCR with dual priming oligonucleotides.
646 PLOS ONE 12:e0185427.
- 647 12. Bonants P, Griekspoor Y, Houwers I, Krijger M, van der Zouwen P, van der Lee
648 TAJ, van der Wolf J. 2019. Development and Evaluation of a Triplex TaqMan
649 Assay and Next-Generation Sequence Analysis for Improved Detection of *Xylella* in
650 Plant Material. Plant Dis 103:645–655.
- 651 13. Schaad NW, Postnikova E, Lacy G, Fatmi M, Chang C-J. 2004. *Xylella fastidiosa*
652 subspecies: *X. fastidiosa* subsp. *piercei*, subsp. nov., *X. fastidiosa* subsp. *multiplex*
653 subsp. nov., and *X. fastidiosa* subsp. *pauca* subsp. nov. Syst Appl Microbiol
654 27:290–300.
- 655 14. Nunney L, Schuenzel EL, Scally M, Bromley RE, Stouthamer R. 2014. Large-Scale
656 Intersubspecific Recombination in the Plant-Pathogenic Bacterium *Xylella*
657 *fastidiosa* Is Associated with the Host Shift to Mulberry. Appl Environ Microbiol
658 80:3025–3033.

- 659 15. Denancé N, Briand M, Gaborieau R, Gaillard S, Jacques M-A. 2019. Identification
660 of genetic relationships and subspecies signatures in *Xylella fastidiosa*. *BMC*
661 *Genomics* 20:239.
- 662 16. Marcelletti S, Scortichini M. 2016. Genome-wide comparison and taxonomic
663 relatedness of multiple *Xylella fastidiosa* strains reveal the occurrence of three
664 subspecies and a new *Xylella* species. *Arch Microbiol* 198:803–812.
- 665 17. Dupas E, Briand M, Jacques M-A, Cesbron S. 2019. Novel Tetraplex Quantitative
666 PCR Assays for Simultaneous Detection and Identification of *Xylella fastidiosa*
667 Subspecies in Plant Tissues. *Front Plant Sci* 10.
- 668 18. Denancé N, Legendre B, Briand M, Olivier V, Boisseson C de, Poliakoff F, Jacques
669 M-A. 2017. Several subspecies and sequence types are associated with the
670 emergence of *Xylella fastidiosa* in natural settings in France. *Plant Pathol* 66:1054–
671 1064.
- 672 19. Jones S, Baizan-Edge A, MacFarlane S, Torrance L. 2017. Viral Diagnostics in
673 Plants Using Next Generation Sequencing: Computational Analysis in Practice.
674 *Front Plant Sci* 8.
- 675 20. Mechan Llontop ME, Sharma P, Aguilera Flores M, Yang S, Pollok J, Tian L,
676 Huang C, Rideout S, Heath LS, Li S, Vinatzer BA. 2019. Strain-Level Identification
677 of Bacterial Tomato Pathogens Directly from Metagenomic Sequences.
678 *Phytopathology*® 110:768–779.

- 679 21. Faino L, Scala V, Albanese A, Modesti V, Grottoli A, Pucci N, L'Aurora A, Reverberi
680 M, Loreti S. 2019. Nanopore sequencing for the detection and the identification of
681 *Xylella fastidiosa* subspecies and sequence types from naturally infected plant
682 material. bioRxiv 810648.
- 683 22. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence
684 classification using exact alignments. *Genome Biol* 15:R46.
- 685 23. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
686 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,
687 Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm
688 and its applications to single-cell sequencing. *J Comput Biol J Comput Mol Cell*
689 *Biol* 19:455–477.
- 690 24. Tian L, Huang C, Mazloom R, Heath LS, Vinatzer BA. 2020. LINbase: a web server
691 for genome-based identification of prokaryotes as members of crowdsourced taxa.
692 *Nucleic Acids Res* <https://doi.org/10.1093/nar/gkaa190>.
- 693 25. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. 2015. Genomics and
694 taxonomy in diagnostics for food security: soft-rotting enterobacterial plant
695 pathogens. *Anal Methods* 8:12–24.
- 696 26. Giampetruzzi A, D'Attoma G, Zicca S, Abou Kubaa R, Rizzo D, Boscia D, Saldarelli
697 P, Saponari M. 2019. Draft Genome Sequence Resources of Three Strains (TOS4,
698 TOS5, and TOS14) of *Xylella fastidiosa* Infecting Different Host Plants in the Newly
699 Discovered Outbreak in Tuscany, Italy. *Phytopathology*® 109:1516–1518.

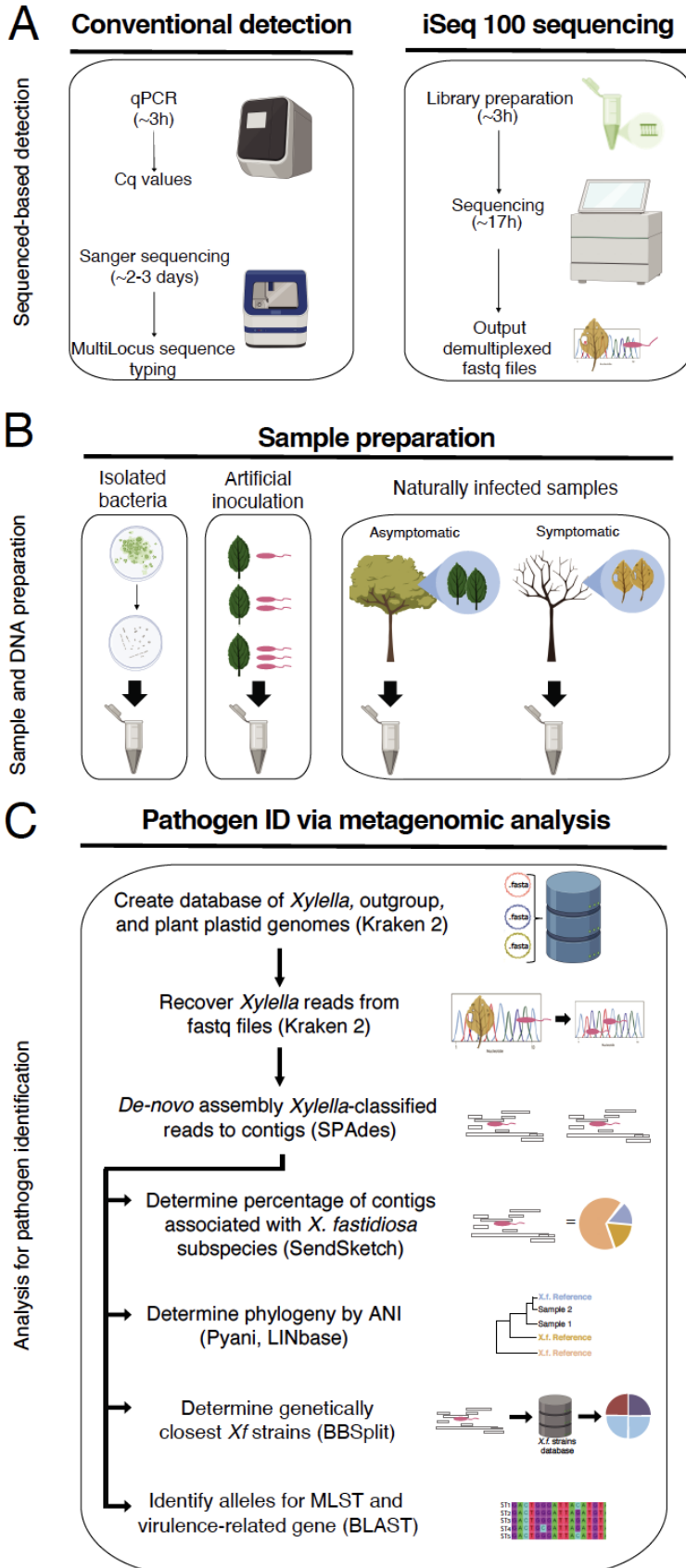
- 700 27. Ayling M, Clark MD, Leggett RM. 2020. New approaches for metagenome
701 assembly with short reads. *Brief Bioinform* 21:584–594.
- 702 28. Nunney L, Elfekih S, Stouthamer R. 2012. The Importance of Multilocus Sequence
703 Typing: Cautionary Tales from the Bacterium *Xylella fastidiosa*. *Phytopathology*®
704 102:456–460.
- 705 29. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in
706 multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40:e3.
- 707 30. Nunney L, Vickerman DB, Bromley RE, Russell SA, Hartman JR, Morano LD,
708 Stouthamer R. 2013. Recent Evolutionary Radiation and Host Plant Specialization
709 in the *Xylella fastidiosa* Subspecies Native to the United States. *Appl Environ*
710 *Microbiol* 79:2189–2200.
- 711 31. Modesti V, Pucci N, Lucchesi S, Campus L, Loreti S. 2017. Experience of the
712 Latium region (Central Italy) as a pest-free area for monitoring of *Xylella fastidiosa*:
713 distinctive features of molecular diagnostic methods. *Eur J Plant Pathol* 148:557–
714 566.
- 715 32. Schrader C, Schielke A, Ellerbroek L, Johne R. 2012. PCR inhibitors – occurrence,
716 properties and removal. *J Appl Microbiol* 113:1014–1026.
- 717 33. Marchi G, Rizzo D, Ranaldi F, Ghelardini L, Ricciolini M, Scarpelli I, Drosera L, Goti
718 E, Capretti P, Surico G. 2018. First detection of *Xylella fastidiosa* subsp. *multiplex*
719 DNA in Tuscany (Italy) | *Phytopathologia Mediterranea*. *Phytopathol Mediterr*
720 57:363–364.

- 721 34. Yuan X, Morano L, Bromley R, Spring-Pearson S, Stouthamer R, Nunney L. 2010.
722 Multilocus sequence typing of *Xylella fastidiosa* causing Pierce's disease and
723 oleander leaf scorch in the United States. *Phytopathology* 100:601–611.
- 724 35. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid
725 core-genome alignment and visualization of thousands of intraspecific microbial
726 genomes. *Genome Biol* 15:524.
- 727 36. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population
728 genomics: BIGSdb software, the PubMLST.org website and their applications.
729 *Wellcome Open Res* 3:124.
- 730 37. da Silva FR, Vettore AL, Kemper EL, Leite A, Arruda P. 2001. Fastidian gum: the
731 *Xylella fastidiosa* exopolysaccharide possibly involved in bacterial pathogenicity.
732 *FEMS Microbiol Lett* 203:165–171.
- 733 38. Chatterjee S, Almeida RPP, Lindow S. 2008. Living in two Worlds: The Plant and
734 Insect Lifestyles of *Xylella fastidiosa*. *Annu Rev Phytopathol* 46:243–271.

735

736

737 **Main figures**



739

740 **Fig. 1.** Metagenomics for diagnostic pipeline. **A)** Sequenced-based detection. Two

741 approaches were used for *Xf* detection: conventional detection and iSeq 100

742 sequencing. For conventional, samples were analyzed using qPCR assays, Harper's

743 test or tetraplex Dupas's test, and MLST involving Sanger sequencing of seven

744 housekeeping genes. iSeq 100 libraries were prepared according to manufacturer. After

745 17h of sequencing, demultiplexed samples were recover from the machine and used for

746 subsequent analysis. **B)** Sample preparation. The samples used for the pipeline were

747 DNAs extracted from bacterial strains in culture, spiked plant material , and naturally

748 infected samples. **C)** Pathogen identification via metagenomic analysis. Demultiplexed

749 fastq reads from all samples were then used for metagenomic analysis. We created a

750 database to recover *Xf* reads using Kraken2. The database contained *Xylella*,

751 *Xanthomonas* and *Escherichia coli* genomes. We also added plant plastid genomes to

752 remove false positive results. *Xf* reads were recovered from the fastq files. The *Xf*

753 recovered reads were de-novo assembled to obtain *Xf* contigs, using SPADes. The *Xf*

754 contigs were used in four different analyses, subspecies identification, phylogeny

755 reconstruction, identification of the genetically closest strain with a sequenced genome

756 and alleles from specific genes. To determine subspecies, we used the tool

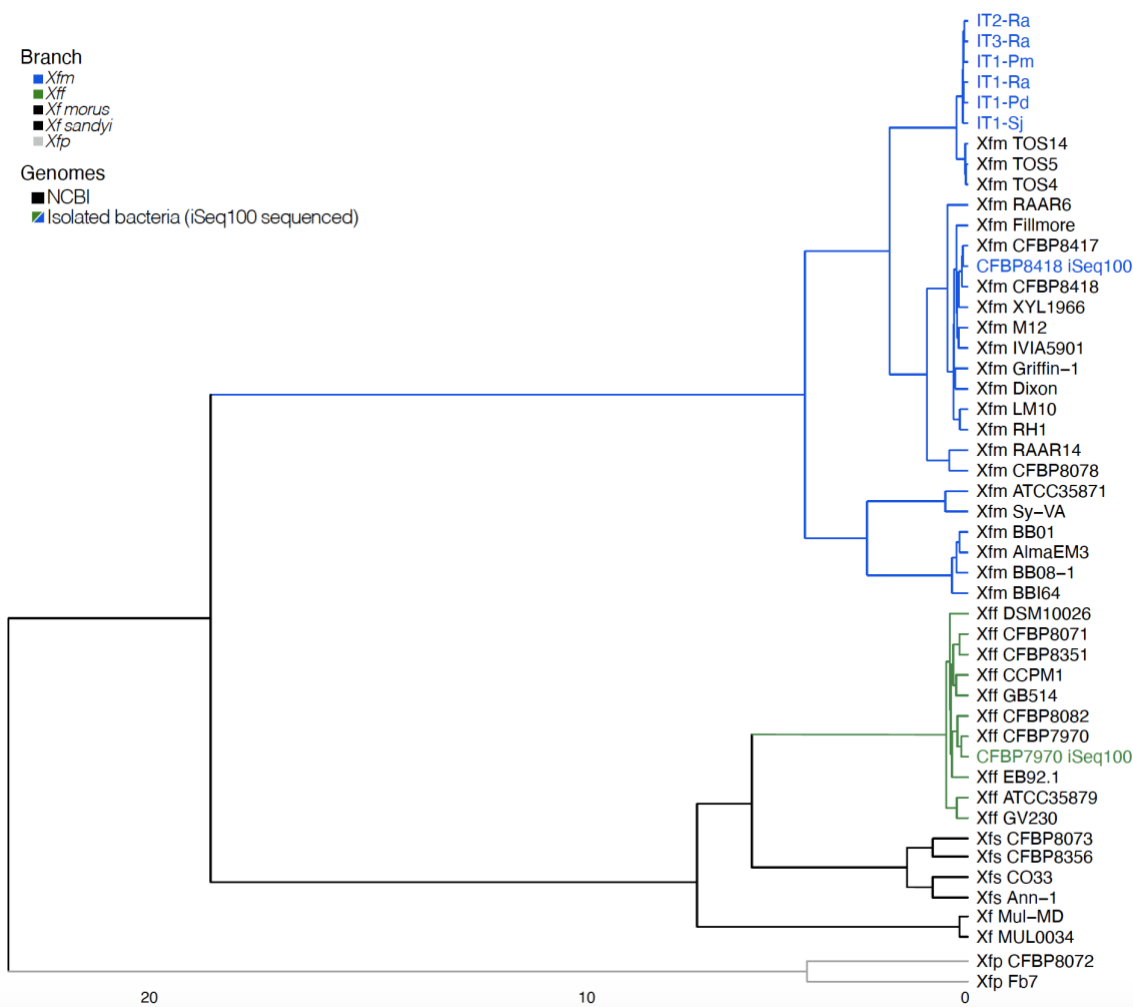
757 SendSketch. To reconstruct phylogeny, we calculated ANI using Pyani and the website

758 tool LINbase (<https://linbase.org/>). To determine the genetically closest known *Xf* strain,

759 we detected the number of hits to each *Xf* strain using BBSplit. To identify specific

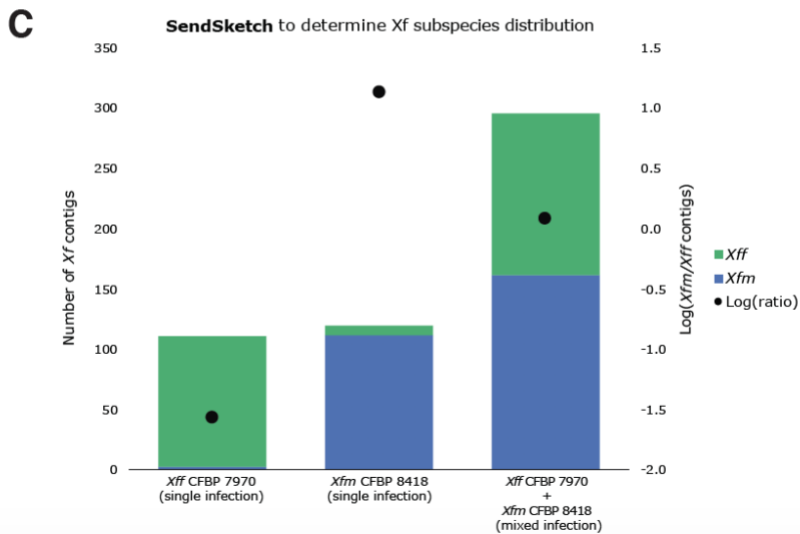
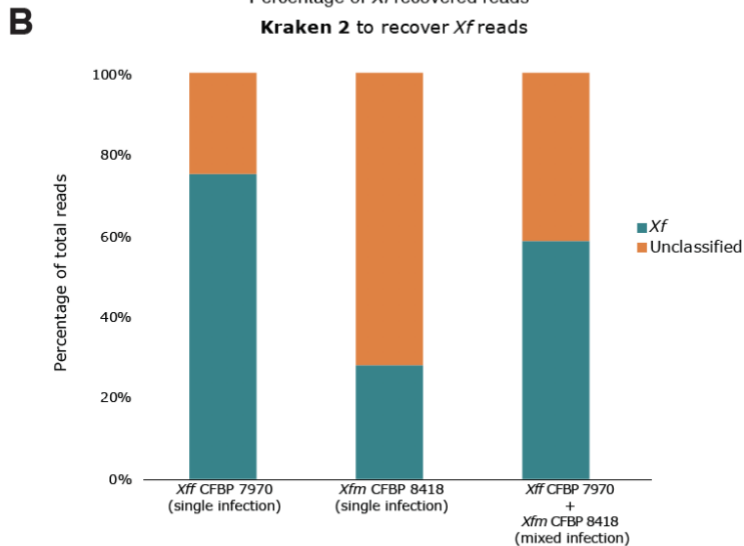
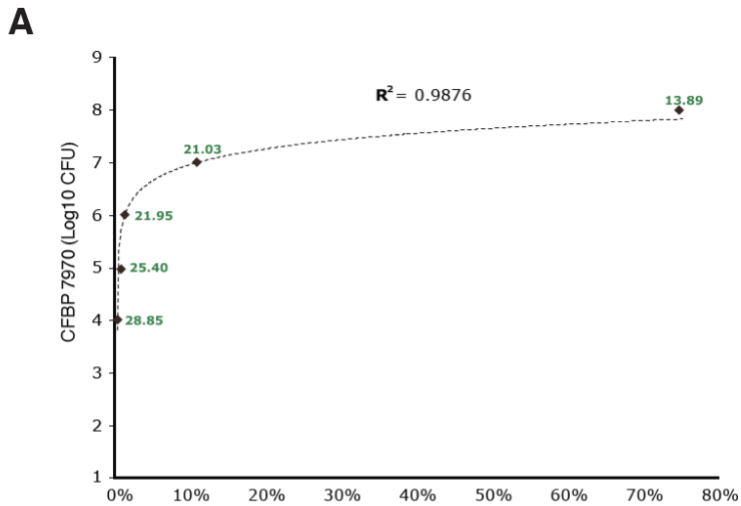
760 genes alleles, we calculated the percentage of identity to the seven MLST genes (*cysG*,

761 *gltT*, *holC*, *leuA*, *malF*, *nuoL*, *petC*), and the percentage of similarity to 17 virulence-
 762 related proteins using local BLAST+. Graphics were created with BioRender.
 763



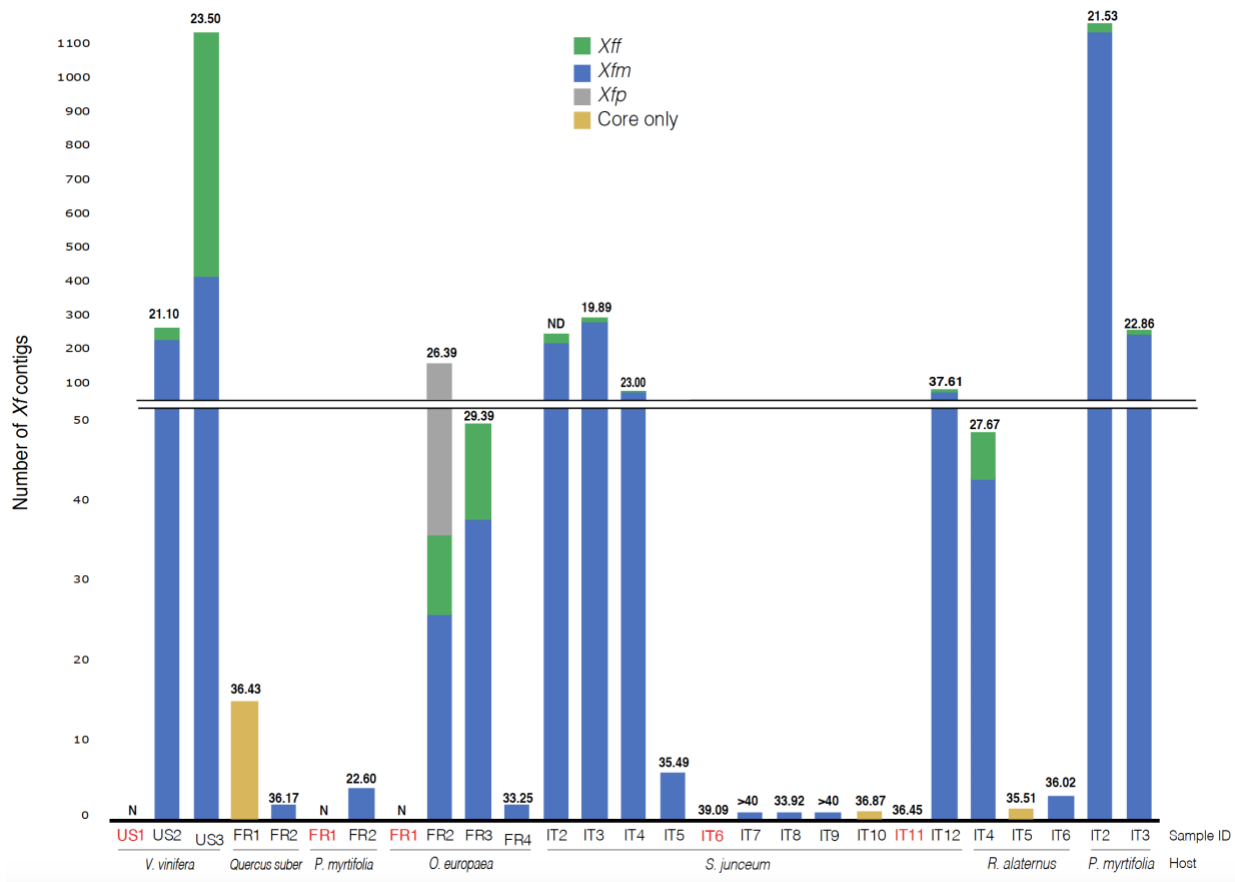
764
 765 **Fig. 2:** Phylogenetic reconstruction of isolated bacteria used in this study. The cluster
 766 analysis is based on average nucleotide identity values from Pyani. Branch colors
 767 indicate different *Xf* subspecies: *Xff* (green), *Xfm* (blue), *Xfp* (gray), *Xf* subspecies
 768 *morus* and *sandyi* (black). The sequenced gDNA from isolated bacteria are indicated in
 769 blue or green. The *Xf* genomes obtained from NCBI are indicated in black. The cluster
 770 was plotted using ComplexHeatmap R package.

771



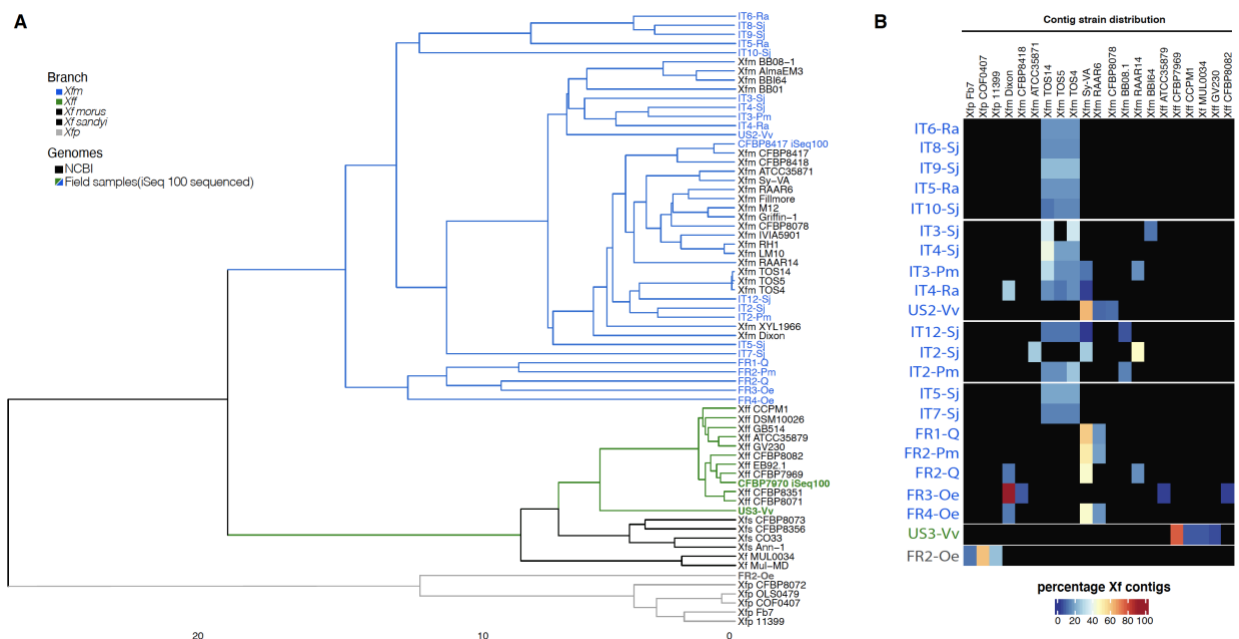
772

773 **Fig. 3:** Spiked samples with different dilutions and mixed samples. **A)** scatter plot
 774 comparing the Log10 CFU with the percentage of *Xff* CFBP7970 recovered reads from
 775 total read number. The dotted line indicates a logarithmic trendline, $y = 0.4222\ln(x) + 7.9511$; $R^2=9876$. The green numbers indicated the Cq values for each sample. **B)**
 777 Percentage of *Xf*-recovered reads by Kraken2 from single (*Xff*, 1e8; *Xfm*, 1e7 CFU)-
 778 and mixed-infected (1e7 CFU) samples. Teal bars indicate *Xf* reads, and orange bars
 779 indicate unclassified reads. Unclassified show reads with no similarity to *Xf* reads, like
 780 plant or other microorganisms reads. **C)** Proportion of *Xf* subspecies from total *Xf*
 781 contigs in single and mixed infections. The black dots indicated the log-ratio as a
 782 manual correction to detect single and mixed infection. *Xfm* and *Xff* are indicated in blue
 783 and green respectively.



784

785 **Fig. 4.** *Xf* from field-collected samples and read mapping to subspecies from database.
 786 Stacked columns indicate the number of *Xf* contigs with 100% identify to each *Xf*
 787 subspecies. Samples indicated in red were *Xf*-negative with our pipeline. *Xff*:
 788 represents the sum of *Xf* subsp. *fastidiosa*, *Xf morus*, and *Xf. sandyi*; *Xfm*: *Xf* subsp.
 789 *multiplex*; *Xfp*: *Xf* subsp. *pauca*. Only Core indicates samples that only have *Xf* core
 790 contigs. The Cq values are indicated on the top of each bar. ND is not determined. N is
 791 negative for *Xf* based on qPCR Sample code and hosts are indicated on the X-axis.
 792 Each country of origin is indicated in the sample ID: France (FR), Italy (IT) and USA
 793 (US) along with the host from which they were isolated.



794
 795 **Fig. 5:** Metagenomics analyses of field samples identifies bacterial subspecies. A) The
 796 dendrogram indicates the distance and cluster analysis based on ANI values using
 797 NCBI whole genomes and assembled *Xf* contigs. Branch colors represent each *Xf*
 798 subspecies. Blue, gray and green names indicate iSeq 100 sequenced samples. B) The

799 heatmap shows the percentage of unique contigs assigned to each *Xf* strain. The
800 cluster and strain distribution were plotted using ComplexHeatmap R package.