Species boundaries and molecular markers for the classification of 16SrI phytoplasmas inferred by genome analysis

Shu-Ting Cho[1], Hung-Jui Kung[1], Weijie Huang[2], Saskia A. Hogenhout[2], Chih-Horng Kuo[1*]

[1]Institute of Plant and Microbial Biology, Academia Sinica, Taipei 11529, Taiwan
[2]Department of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, Norfolk, NR4 7UH, UK

*** Correspondence:**
Chih-Horng Kuo; chk@gate.sinica.edu.tw

Keywords: plant pathogen, comparative genomics, average nucleotide diversity (ANI), taxonomy, multilocus sequence analysis (MLSA), effector.

Abstract

Phytoplasmas are plant-pathogenic bacteria that impact agriculture worldwide. The commonly adopted classification system for phytoplasmas is based on the restriction fragment length polymorphism (RFLP) analysis of their 16S rRNA genes. With the increased availability of phytoplasma genome sequences, the classification system can now be refined. This work examined 11 strains in the 16SrI group within the genus '*Candidatus* Phytoplasma' and investigated the possible species boundaries. We confirmed that the RFLP classification method is problematic due to intragenomic variation of the 16S rRNA genes and uneven weighing of different nucleotide positions. Importantly, our results based on the molecular phylogeny, differentiations in chromosomal segments and gene content, and divergence in homologous sequences, all supported that these strains may be classified into multiple operational taxonomic units (OTUs) equivalent to species. Strains assigned to the same OTU share >97% genome-wide average nucleotide identity (ANI) and >78% of their protein-coding genes. In comparison, strains assigned to different OTUs share <94% ANI and <75% of their genes. Reduction in homologous recombination between OTUs is one possible explanation for the discontinuity in genome similarities, and these findings supported the proposal that 95% ANI could serve as a cutoff for distinguishing species in bacteria. Additionally, critical examination of these results and the raw sequencing reads led to the identification of one genome that was presumably mis-assembled by combining two sequencing libraries built from phytoplasmas belonging to different OTUs. This finding provided a cautionary tale for working on uncultivated bacteria. Based on the new understanding of phytoplasma divergence and the current genome availability, we developed five molecular markers that could be used for multilocus sequence analysis (MLSA). By selecting markers that are short yet highly informative, and are distributed evenly across the chromosome, these markers provided a cost-effective system that is robust against recombination. Finally, examination of the effector gene distribution further confirmed the rapid gains and losses of these genes, as well as the involvement of potential mobile units (PMUs) in their molecular evolution. Future improvements on the taxon sampling of phytoplasma genomes will allow further expansions of similar analysis, and thus contribute to phytoplasma taxonomy and diagnostics.

Introduction

Phytoplasmas are a group of insect-transmitted plant-pathogenic bacteria that reduce yields of diverse crops worldwide (Lee et al., 2000; Hogenhout et al., 2008; Bertaccini and Lee, 2018). These obligate parasites are related to the animal-pathogenic mycoplasmas (Chen et al., 2012), and both groups of these wall-less bacteria were assigned to the class Mollicutes under the phylum Tenericutes (Brown, 2010). However, unlike mycoplasmas and several other Mollicutes lineages, pure culture of phytoplasmas outside of their hosts has not been achieved yet. Due to this reason, all phytoplasmas were assigned to a '*Candidatus*' ('*Ca*.') genus (The IRPCM Phytoplasma/Spiroplasma Working Team, 2004).

For the purposes of pathogen detection or research on their basic biology, a reliable system for strain identification and classification is critical. For example, it is important to understand if the diseases affecting different plants, transmitted by different vectors, and/or occurring in different locations are caused by the same or different phytoplasma strains. In addition, to infer the evolutionary history and biodiversity of phytoplasmas, defining species or other equivalent biological units are of fundamental importance. The commonly used system for phytoplasma classification was first established in the 1990s (Lee et al., 1993; Lee et al., 1998). This system utilizes a defined set of 17 restriction enzymes to perform restriction fragment length polymorphism (RFLP) analysis on a 1.25-kb PCR product derived from their 16S rRNA gene (Gundersen and Lee, 1996). Under this system, a set of reference strains were chosen, and a new strain could be assigned to the same 16S rRNA gene RFLP (16Sr) group and subgroup as a reference when the similarity coefficient is higher than 0.85 and 0.97, respectively (Zhao et al., 2009; Zhao and Davis, 2016). With the development of an easy-to-use web-based tool *i*PhyClassifier (Zhao et al., 2009), this system has been well-adopted by the phytoplasma research community.

However, despite the popularity, this 16Sr classification system has several shortcomings. To begin with, the RFLP approach considers only those nucleotides located within the restriction sites, rather than the entire sequence. Additionally, the similarity coefficients of the RFLP patterns do not provide information about phylogenetic relationships. Furthermore, intragenomic sequence variations of 16S rRNA genes may cause problems in classification. Most importantly, although 16S rRNA gene is considered as a universal marker for bacterial classification, the use of a single locus is inherently unreliable. Rather, adoption of multilocus sequence analysis (MLSA) is strongly recommended (Glaeser and Kämpfer, 2015). Unfortunately, selection of suitable markers for MLSA is not straightforward. Due to the highly variable evolutionary rates among different bacteria and among different genes (Kuo and Ochman, 2009), the markers that work well for one group do not necessary work well for others. Although extensive efforts have been devoted to the development of MLSA markers for phytoplasmas (Martini et al., 2019), genome-level assessment for the relative performance of these markers is lacking. Moreover, while the use of genome analysis to replace MLSA provides a potential solution and is gaining popularity (Konstantinidis and Tiedje, 2005; Konstantinidis et al., 2017; Jain et al., 2018; Parks et al., 2018), genome analysis is still challenging for uncultured bacteria such as phytoplasmas. For these reasons, development and evaluation of MLSA markers are still important.

To develop and evaluate MLSA markers, the availability of genome sequences provides a comprehensive guide for PCR primer design. For example, based on the complete genome sequence of '*Candidatus* Phytoplasma asteris' OY-M (Oshima et al, 2004), 18 PCR primer sets were designed to cover different regions of the chromosome (Kakizawa and Kamagata, 2014).

95 Moreover, by developing a multiplex-PCR method for these primer sets and establishing the
96 amplification patterns in a collection of reference strains, this previous study provided a fast and
97 cost-effective method for strain identification (Kakizawa and Kamagata, 2014). As the number
98 of available genome sequences increases, the genome-enabled MLSA maker development could
99 be further improved in two critical aspects. First, the genome-scale molecular phylogeny
100 provides a reliable reference for evaluating the performance of each candidate markers, such that
101 the best-performing ones could be selected. Second, with the comprehensive genomic
102 information of all reference strains, the chromosomal location of each candidate markers and the
103 exact primer sequences could be evaluated using bioinformatic methods. In this regard,
104 phytoplasmas belonging to the 16SrI group, also often known as the aster yellows (AY) group or
105 '*Ca.* P. asteris' (i.e., a provisional species defined as encompassing all known subgroups within
106 the 16SrI group) (Lee et al., 2004), represent the best system to test the concept of genome-
107 assisted MLSA marker development. Among the >30 16Sr groups that have been described
108 (Zhao and Davis, 2016), only eight have genome sequences available (Cho et al., 2019b).
109 Moreover, the genome sequencing efforts have been highly focused on the 16SrI group (Table
110 1), which has a worldwide geographical distribution, wide range of plant hosts, and great impact
111 on agriculture (Lee et al., 2004). Finally, this group is also the one that received most research
112 attention on the molecular mechanisms of plant-microbe interactions; most of the phytoplasma
113 effectors that have been characterized were initially studied in 16SrI phytoplasmas (Bai et al.,
114 2009; Hoshi et al., 2009; MacLean et al., 2011; Wang et al. 2018b; Huang and Hogenhout,
115 2019). Previous studies of this group identified extensive levels of genome divergence and
116 suggested that '*Ca.* P. asteris' may be classified into multiple operational taxonomic units
117 (OTUs) equivalent to species (Firrao et al., 2013; Cho et al., 2019a). In this work, we expanded
118 the sampling of available genome sequences to perform in-depth analysis of genome
119 comparisons. Specifically, we aimed to provide quantitative guidelines for defining putative
120 species boundaries, which could better inform phytoplasma taxonomy. Additionally, we aimed to
121 develop MLSA markers that could facilitate future genetic characterization of the phytoplasmas
122 belonging to the 16SrI or AY group.
123
124 **Materials and Methods**
125 Data Sets and Analysis Methods
126 The 11 genome sequences included in this study are listed in Table 1. These data sets
127 represent all phytoplasma genomes available from GenBank (Benson et al., 2018) as of January
128 2020 and are recognized as belonging to the 16SrI group, AY group, or '*Ca*. P. asteris' (Lee et
129 al., 2004). Two other '*Ca*. Phytoplasma' species have been described as being affiliated with the
130 16SrI group, including '*Ca*. P. japonicum' (Sawayanagi et al., 1999) and '*Ca*. P. lycopersici'
131 (Arocha et al., 2007), but were not included in this study due to the lack of genome sequence
132 availability. Notably, '*Ca*. P. japonicum' was first reported as a member of the 16SrI-D subgroup
133 (Sawayanagi et al., 1999) and later listed under the 16SrI group in the current taxonomy (The
134 IRPCM Phytoplasma/Spiroplasma Working Team, 2004). However, the *i*PhyClassifier (Zhao et
135 al., 2009) analysis of its 16S rRNA gene sequence (GenBank accession AB010425) assigned it
136 to the 16SrXII-D subgroup, suggesting that its classification remained to be investigated.
137 To confirm the 16Sr group and subgroup assignments, we extracted the 16S rRNA gene
138 sequences from the genomes and analyzed those sequences using the *i*PhyClassifier (Zhao et al.,
139 2009). Most phytoplasma genomes contain two copies of 16S rRNA genes and each copy was
140 processed individually. The strain OY-V lacks any 16S rRNA gene in the current assembly and

141 therefore was excluded from this analysis. The strain NJAY has four partial sequences for its 16S
142 rRNA genes and all of these four sequences were examined.
143 All bioinformatic tools for processing these data sets were used with the default settings
144 unless stated otherwise. For correlation tests, the cor.test function implemented in R v3.4.4 (R
145 Core Team, 2019) was used to examine the Pearson's product moment correlation coefficient.
146 For strain TW1 (GenBank genome accession GCA_003181115.1), we downloaded the
147 raw sequencing results from the NCBI Sequence Read Archive (SRA) to examine the
148 phytoplasma sequences present in the sequencing libraries. The raw reads were mapped to two
149 other reference genomes (i.e., AYWB and OY-M; see Table 1) to detect the presence of 16SrI-A
150 and 16SrI-B type phytoplasmas, respectively. For the Oxford Nanopore MinION reads (SRA
151 accession SRR7548026), the mapping was performed using Minimap2 v2.15 (Li, 2018). For the
152 Illumina MiSeq paired-end reads (SRA accession SRR7548027), the mapping was performed
153 using the Burrows-Wheeler Alignment (BWA) tool v0.7.17 (Li and Durbin, 2009). The mapping
154 results were programmatically checked using SAMtools v1.9 (Li et al., 2009) to identify the
155 sequence variations.
156
157 Genome Comparisons
158 For whole-genome comparison, FastANI v1.1 (Jain et al., 2018) was used to calculate the
159 proportion of genomic segments mapped and the average nucleotide identity (ANI) of those
160 segments in each genome-pair. For each pair, reciprocal comparisons between the two genomes
161 were conducted (i.e., X as the query against Y as the reference and Y as the query against X as
162 the reference). To calculate the proportion of genomic segments mapped, the numbers of
163 segments mapped from those two reciprocal comparisons were combined as the numerator, and
164 the total numbers of segments from those two genomes were combined as the denominator. The
165 reciprocal ANI values were averaged. For example, in the AYWB-NJAY comparison, 183 out of
166 the 239 segments of AYWB could be mapped to NJAY, and all 187 segments of NJAY could be
167 mapped to AYWB. Based on these results, these two strains share (183 + 187) / (239 + 187) =
168 370 / 426 = 86.9% of their genomic segments. The ANI values were 99.5% and 99.8% in the two
169 reciprocal comparisons, and an average ANI of 99.7% was reported.
170 For gene-centric investigation, the homologous gene clusters among all genomes were
171 identified using OrthoMCL v1.3 (Li et al., 2003) based on the procedures described in our
172 previous studies (Lo et al., 2013; Chung et al., 2013). Briefly, all coding sequences (CDS),
173 including those annotated as pseudogenes, were included in the all-against-all similarity searches
174 using BLASTN v2.6.0 (Camacho et al., 2009) with an e-value cutoff of $1e^{-15}$. The BLASTN
175 results were used as the input for OrthoMCL, which normalized the within- and between-genome
176 comparisons and identified homologs according to the principle of reciprocal best hits. Similar to
177 the procedures for whole-genome comparisons, the proportion of homologs shared and the
178 sequence identity of shared homologs were considered as two separate metrics of genome
179 similarities. The proportion of homologs shared was calculated based on averaging (i.e., the
180 number of homologs shared was used as the numerator, and the average number of homologs
181 present in those two genomes was used as the denominator). The sequence identities of coding
182 regions were calculated based on multiple sequence alignment of those single-copy genes present
183 in all of the genomes compared. The homologs were aligned using MUSCLE (Edgar, 2004) and
184 the sequence identities were calculated using the DNADIST program in PHYLIP v3.697
185 (Felsenstein, 1989); as expected, reciprocal comparisons all produced the same result for
186 sequence identities.

187 In addition to the pairwise comparisons, the gene content of all genomes was compared
188 using a procedure based on principal coordinates analysis (Lo et al., 2018). For this procedure,
189 the homologous gene clustering result was converted into a matrix of 11 genomes by 1,233 gene
190 clusters; the value in each cell indicates the gene copy number. This matrix was converted into a
191 Jaccard distance matrix among genomes using the VEGAN package in R, then processed using
192 the PCOA function in the APE package (Popescu et al., 2012).
193 To compare the genes for putative secreted proteins among these phytoplasmas, we
194 examined all coding sequences using a procedure described in a previous study (Cho et al.,
195 2019b). Briefly, SignalP v5.0 (Armenteros et al., 2019) was used to check the presence of signal
196 peptide based on the Gram-positive bacteria model. The program TMHMM v2.0 (Krogh et al.,
197 2001) was used to verify that these putative secreted proteins have no transmembrane domain.
198 For those four experimentally characterized effectors (i.e., SAP05, SAP11/SWP1,
199 SAP54/PHYL1, and TENGU), all putative homologs (i.e., those belong to the same homologous
200 gene cluster as defined by OrthoMCL or have a BLASTN e-value of lower than $1e^{-100}$) were
201 manually added to the list of regardless of the SignalP prediction results.
202 The genome alignment of representative strains with high quality assemblies were
203 performed using genoPlotR v0.8.9 (Guy et al., 2010). The syntenic regions were identified using
204 BLASTN v2.6.0 (Camacho et al., 2009), the cutoff values for sequence similarity were set to e-
205 value = $1e^{-15}$ and alignment length = 2,000 bp.
206
207 Molecular Phylogenetics
208 For phylogenetic inference, MUSCLE v3.8.31 (Edgar, 2004) was used to generate
209 multiple sequence alignments and PhyML v3.3 (Guindon and Gascuel, 2003) was used to infer
210 maximum likelihood phylogenies. The visualization was performed using JalView v2.11
211 (Waterhouse et al., 2009) and FigTree v1.4.4. In each inference, PHYLIP v3.697 (Felsenstein,
212 1989) was used to generate 1,000 replicates for bootstrap analysis.
213
214 Development of Molecular Markers
215 To identify molecular markers that may be suitable for classification, those multiple
216 sequence alignments of single-copy genes shared by all genomes were examined manually.
217 Genes were ranked by the number of variable sites that could be used for distinguishing different
218 OTUs. Subsequently, the list was prioritized based on the criterion of having multiple markers
219 located in distinct regions of those phytoplasma chromosomes. The PCR primers (Table 2) were
220 designed manually with the aim of producing ~700-800 bp products to cover those variable
221 regions, such that each product could be sequenced in one single Sanger sequencing reaction.
222 To validate the primer design, a DNA sample derived from a periwinkle infected by the
223 phytoplasma strain DY2014 (Cho et al., 2019b) was used as the template for PCR tests. All kits
224 were used according to the manufacturer's instructions. The PCR were performed using the
225 KAPA HiFi HotStart ReadyMix (Roche, USA). The cycling conditions were: (1) an initial
226 activation step at 95°C for 3 min; (2) 35 cycles of 98°C for 20 sec, 55°C for 30 sec, and 72°C for
227 60 sec; and (3) a final extension step of 72°C for 5 min. The annealing temperature was lowered
228 to 53°C for the newly designed primers for *secY* (Table 2), while the other four primer sets used
229 55°C. Agarose gel electrophoresis was conducted to confirm that all PCR tests produced one
230 single band of expected size. For further confirmation, the PCR products were purified using the
231 QIAquick Gel Extraction Kit (Qiagen, Germany), then sequenced using the BigDye Terminator
232 v3.1 Cycle Sequencing Kit on an Applied Biosystems 3730XL DNA Analyzer (Thermo Fisher

233 Scientific, USA). The sequencing results were compared to the published genome sequence for
234 final confirmation.
235 The procedure for calculating the substitution rates was based on that described in our
236 previous studies (Kuo and Ochman, 2009; Kuo et al., 2009). Briefly, the homologous protein
237 sequences were aligned by using MUSCLE v3.8.31 (Edgar, 2004). The protein alignments were
238 converted into codon-based nucleotide alignment using PAL2NAL v14 (Suyama et al., 2006),
239 which were then processed using the YN00 method (Yang and Nielsen, 2000) implemented in
240 PAML v4.9h (Yang, 2007). To identify the homologs of selected marker genes in other more
241 divergent phytoplasmas, the full gene sequences were used as queries to run BLASTN searches
242 against available genomes. After the homolog identification, multiple sequence alignments were
243 performed using MUSCLE (Edgar, 2004) and the newly designed primers (Table 2) were
244 examined manually to check if these primers could work on other phytoplasmas.
245
246 **Results and Discussion**
247 Genome Characteristics
248 Among the 11 phytoplasma genomes included in this study, four have the complete
249 chromosomal sequence available while the other seven are draft assemblies with varying levels
250 of completeness (Table 1 and Figure 1). For those incomplete genomes, the true genome sizes
251 are difficult to determine. Nonetheless, the largest (i.e., OY-M; 853 kb) and the smallest (i.e.,
252 MBSP-M3; 576 kb) genomes in this data set are both completed assemblies, confirming the high
253 level of genome size variation among these closely related strains. The total number of CDS is
254 strongly correlated with the genome size ($r = 0.92$, $p = 2.7e-5$). However, the number and
255 proportion of pseudogenes varied widely among these genomes. For example, the strains AYWB
256 and NJAY have a genome-wide ANI of 99.7% (Figure 2A), yet the proportion of pseudogenes
257 are 17% and 41% (Table 1), respectively. In another case, DY2014 and OY-V have a genome-
258 wide ANI of 99.4%, yet the proportion of pseudogenes are 8% and 25%, respectively. Although
259 these differences may reflect true biological processes, such as elevated mutation accumulations
260 that have occurred in the evolutionary history of some strains, it is also possible that these
261 differences may be explained by artifacts. Regarding the sequencing methods, AYWB was based
262 on Sanger sequencing (Bai et al., 2006) and DY2014 was based on Illumina (Cho et al., 2019b),
263 while NJAY was based on one single Roche/454 library (Sparks et al., 2018) and OY-V was
264 based on one Illumina paired-end library and one Roche/454 GS FLX mate-pair library
265 (Kakizawa et al., 2014). Due to the low GC-content of these genomes (i.e., ~27-28%), it is
266 plausible that sequencing errors in homopolymeric regions, particularly those errors originated
267 from the 454 sequencing technology, may have contributed to the high numbers of pseudogenes
268 with frameshift mutations in NJAY and OY-V. Moreover, assemblies that are more fragmented
269 may also have more partial genes located at the ends of contigs. Due to these concerns, all
270 pseudogenes were included in the gene-centric investigation of this study.
271
272 Classification of the 16SrI Phytoplasmas Analyzed
273 The 16SrI subgroup assignments of these strains were mostly consistent with those
274 reported in literature but there were a few surprises (Figure 1A). First, WBD was reported as a
275 16SrI-B strain (Chen et al., 2014). However, our examination of the sequence record deposited in
276 GenBank revealed that the two copies of 16S rRNA genes in this genome differ by 4-bp, and
277 *i*PhyClassifier (Zhao et al., 2009) assigned these two sequences to 16SrI-R and 16SrI-S,
278 respectively. Results from molecular phylogeny also support that these two sequences form a

279 clade that is distinct from other 16SrI-B sequences. Second, DY2014 has two 16S rRNA genes
280 that differ by 1-bp. While one was assigned to 16SrI-B, the other was identified as being a new
281 subgroup with a RFLP pattern most similar to 16SrI-B. Finally, TW1 has two sequences that
282 differ by 12-bp, with one assigned to 16SrI-A and the other assigned to 16SrI-B (Town et al.,
283 2018).
284 These results highlighted several issues of this RFLP-based classification system for
285 phytoplasmas. First, the intra-genomic variation of 16S rRNA genes may result in a strain being
286 assigned to different subgroups depending on which homolog was examined. This issue has been
287 reported for other phytoplasmas (Liefting et al., 1996) and a three-letter subgroup designation
288 system has been proposed (Wei et al., 2008). For example, a paulownia witches'-broom (PaWB)
289 phytoplasma was first classified as a 16SrI-D strain (Lee et al., 1998) but later found to harbor a
290 16SrI-B type sequence as well. To accommodate these findings, the subgroup status of this
291 PaWB was redesignated as 16SrI-(B/D)D, with the first two letters in parentheses denoting the
292 types of those two 16S rRNA genes, and the third letter denoting its 16Sr subgroup assignment
293 (Wei et al., 2008; Zhao et al., 2009). Although this three-letter system provides comprehensive
294 information, it may be unnecessarily complex and confusing. Moreover, it is unclear which one
295 of the two types should be chosen for the formal subgroup assignment. In the case of PaWB, it
296 was assigned to 16SrI-D based on precedence. However, in the case of WBD (i.e., originally
297 described as a 16SrI-B strain), it is unclear if it should be redesignated as 16SrI-(R/S)R or 16SrI-
298 (R/S)S. Second, this RFLP-based system considers only the sequence variations in restriction
299 sites. By ignoring the sequence variations in other positions, this approach could alleviate the
300 problem of intra-genomic variations to a certain degree (e.g., the two sequences in AYWB differ
301 by 2-bp, yet both were assigned to 16SrI-A). However, this approach wasted much of the
302 information contained in the full sequence and the uneven weighing of nucleotide positions
303 could introduce strong biases. The strain DY2014 represents one extreme example to illustrate
304 such biases, with a single-bp difference that is sufficient to result in different subgroup
305 assignments between the two 16S rRNA genes. Finally, a classification system that depends
306 solely on the 16S rRNA genes while ignoring other regions of the genome does not provide
307 sufficient resolution.
308 Regardless of the group/subgroup assignments of these 16S rRNA genes, all these
309 sequences share >99% identity (Supplementary Table S1), which is above the general
310 recommended thresholds for defining species. These thresholds include: (1) 97% identity that is
311 commonly used for defining OTUs corresponding to species in microbiota surveys (Schloss and
312 Handelsman, 2005), (2) 97.5% identity recommended for '*Ca*. Phytoplasma' species (The
313 IRPCM Phytoplasma/Spiroplasma Working Team, 2004), and (3) 98.6% identity recommended
314 for uncultivated bacteria (Konstantinidis et al., 2017). In other words, these strains may be
315 considered as all belonging to '*Ca*. P. asteris' (Lee et al., 2004) based on their 16S rRNA gene
316 sequences. However, delineating species-level OTUs based on one single gene may be
317 problematic; more in-depth investigation of their genomic divergence is necessary.
318
319 <u>Genomic Divergence among 16SrI Phytoplasmas and Putative Species Boundaries</u>
320 To further investigate the divergence of these phytoplasmas, we utilized two approaches
321 to quantify their genomic similarities. In the first approach, the whole genome sequences were
322 used to calculate the proportion of genomic segments shared, as well as the ANI values of these
323 segments. A previous study that examined >90,000 prokaryotic genomes found that the ANI
324 values exhibit a bimodal distribution, and the within-species comparisons almost always have

325 >95% ANI (Jain et al., 2018). We found that based on this criterion, the 11 strains examined
326 could be separated into three clusters that roughly correspond to their 16SrI subgroup
327 assignments (Figure 2A). The two 16SrI-A strains, AYWB and NJAY, share 86.9% of their
328 genomes and these shared regions are nearly identical (i.e., ANI = 99.7%). Because the NJAY
329 assembly is highly fragmented and ~10% smaller compared to the complete genome of AYWB,
330 the low proportion of genomic segments shared may be explained by missing data in the NJAY
331 assembly. The strain WBD shares only ~93% ANI with all other 16SrI phytoplasmas compared
332 and represents a distinct lineage. All of the remaining strains, including those assigned to 16SrI-
333 B and TW1, share >97% ANI and could considered as a third cluster. This clustering result is
334 consistent with the maximum likelihood phylogeny inferred using 303 single-copy genes shared
335 by all strains (Figure 1B and Supplementary Table S2).
336 In the second approach, we considered only the putative coding regions (i.e., intact CDS
337 and putative pseudogenes) of these genomes for calculating genomic similarities (Figure 2B).
338 Compared to the genome-wide ANI approach, this gene-centric method is more laborious and
339 could potentially be affected by annotation quality. Nevertheless, this approach has a higher
340 confidence in inferring the homology prior to calculating sequence identities and provides a
341 complementary method to the ANI approach. Although the exact genome similarities values
342 differ slightly, the patterns found by using these two approaches are consistent.
343 Based on these results regarding genomic similarities and molecular phylogeny, it
344 appeared that these 16SrI phytoplasmas could be classified into three species-level OTUs. These
345 findings are consistent with a previous study that the conventional 97% identity threshold for
346 delineating species in bacteria based on 16S rRNA gene is too low (Edgar, 2018). Similar to a
347 previous study that examined diverse prokaryotes (Jain et al., 2018), the sequence identity values
348 among these phytoplasmas also exhibited a bimodal distribution (Figure 3). Excluding the
349 comparisons involving TW1, the within-OTU comparisons all have >97% genome-wide ANI,
350 while between-OTU comparisons all have <94% ANI (Figures 2A and 3A). For sequence
351 identities that were calculated based on only the genic regions, in which the homology could be
352 inferred more confidently, a similar separation was found (i.e., within-OTU: >98.9%, between-
353 OTU: <96%; see Figures 2B and 3B). These results suggest that those strains assigned to the
354 same species-level OTU may have on-going homologous recombination that maintained high
355 sequence similarities in their shared genomic regions. In contrast, certain genetic barriers exist to
356 lower homologous recombination between different OTUs. In other words, these results provide
357 further support to the hypothesis that the biological species concept, which is based on the
358 barriers to homologous gene exchange, could be applied to bacteria as well (Bobay and Ochman,
359 2017). Alternatively, other hypotheses that may explain the discontinuity in genome similarities
360 include ecological sweeps or stochastic neutral processes (Jain et al., 2018), and the exact
361 biological mechanisms remained to be investigated.
362 In addition to sequence identity, another important measurement for genetic similarity is
363 the proportion of genomic regions shared. For this second measurement, a wide range was
364 observed for both within- and between-OTU comparisons (Figures 2 and 3). The wide spread
365 was mostly explained by the inclusion of two highly fragmented draft genomes (i.e., OY-V and
366 CYP; see Table 1), which resulted in underestimates that are artifacts stemming from missing
367 data. However, because the non-redundant chromosomal regions that are more conserved have a
368 higher probability of being included in a draft assembly, the inclusion of draft genomes may also
369 result in overestimates of the genomic regions shared. Due to these uncertainties, we restricted
370 the comparisons to those four strains with the complete genome sequences available (i.e.,

371 AYWB, OY-M, MBSP-M3, and De Villa). For this reduced data set with high confidence,
372 strains assigned to the same species-level OTU share ~78-93% of their chromosomal segments
373 and ~79-89% of their protein-coding genes, whereas strains assigned to different OTUs share
374 ~74-78% of their chromosomal segments and ~70-75% of their protein-coding genes.
375 For an alternative approach of comparing the gene content divergence among these
376 phytoplasmas, we conducted a principal coordinates analysis that considered gene copy numbers
377 in addition to the patterns of gene presence or absence (Figure 4). With the exception of TW1,
378 the results were consistent with the molecular phylogeny (Figure 1B) and pairwise genome
379 similarities (Figures 2 and 3). The first coordinate explained ~40% of the variance and showed
380 clear separation for the three species-level OTUs. The second coordinate explained ~23% of the
381 variance and showed that those 16SrI-B strains could be separated into three subgroups.
382 Taken together, these results supported the existence of multiple species-level OTUs
383 within the previously described '*Ca*. P. asteris' (Lee et al., 2004) and provided quantitative
384 guidelines for inferring the putative species boundaries based on overall genome similarities.
385 Based on these findings, the critical questions to ask are the genotypic, phenotypic, and
386 ecological differentiations among these OTUs. Unfortunately, given the limited number of
387 strains with genome sequence available, it is not possible to obtain meaningful answers yet. For
388 example, other 16SrI-A strains were known to infect different dicots or monocots, and were
389 found in North America and Europe (Lee et al., 2004). However, no genomic information is
390 available for these strains with different hosts and/or geographic distributions. It is important to
391 characterize these strains and validate their OTU assignment before meaningful comparisons
392 could be made between different OTUs. Similarly, more strains that belong to the same OTU
393 with WBD must be sampled and characterized before the biological characteristics of this OTU
394 could be described appropriately. These improvements in the understanding of genotypic and
395 phenotypic differentiations among phytoplasmas are critical in informing future taxonomy
396 revisions.
397
398 <u>The Special Case of the TW1 Genome</u>
399 Based on the aforementioned framework for phytoplasma classification, the strain TW1
400 represents a strange and difficult case. The high level of intra-genomic variation between its two
401 copies of 16S rRNA genes has been noted in the initial characterization (Town et al., 2018). In
402 our whole-genome analysis, this strain shares a higher proportion of its chromosomal segments
403 and protein-coding genes with those 16SrI-A strains, yet has higher sequence identities with
404 those 16SrI-B strains (Figures 2 and 3). Because of these properties, this strain was clustered
405 with those 16SrI-A strains based on gene content (Figure 4), while having closer phylogenetic
406 relationships with those 16SrI-B strains (Figure 1B).
407 In the absence of an intuitive hypothesis that could explain the evolutionary processes
408 leading to these conflicting findings, we examined the possibilities that these findings were
409 results of artifacts. This TW1 genome was co-assembled using one long-read library (based on
410 Oxford Nanopore MinION) and one short-read library (based on Illumina MiSeq) (Town et al.,
411 2018). When the raw reads from these two sequencing libraries were examined separately, we
412 found that the long-read library likely contains a 16SrI-A type phytoplasma. When the raw reads
413 were mapped to the AYWB chromosome, all regions were covered and 3,725 sequence
414 polymorphisms were found; when mapped to OY-M, 52,649 bp had no coverage and 31,401
415 sequence polymorphisms were found. In contrast, the short-read library likely contains a 16SrI-B
416 type phytoplasma. When the raw reads were mapped to OY-M, 103,922 bp had no coverage and

417 9,489 sequence polymorphisms were found; when mapped to AYWB, 110,458 had no coverage
418 and 29,323 sequence polymorphisms were found.
419 Based on these findings, a possible explanation is that during the assembly process, the
420 long reads produced several 16SrI-A type scaffolds, which resulted in the high similarities of the
421 TW1 genome content to 16SrI-A as observed (Figures 2-4). Subsequently, during the polishing
422 stage, the exact sequences of these scaffolds were modified based on the short-read library, thus
423 explaining the high sequence similarity of this TW1 genome to those 16SrI-B strains. In other
424 words, in its current form, this TW1 genome assembly is likely an artifact that was incorrectly
425 assembled by combining two sequencing libraries with phytoplasmas belonging to different
426 OTUs, rather than a true representative of an existing phytoplasma strain. Due to the concern that
427 the inclusion of this genome may have introduced biases, we repeated all phylogenetic inferences
428 and excluded the TW1 sequences prior to multiple sequence alignment. The resulting tree
429 topologies were consistent in terms of OTU assignments; the minor differences were related to
430 the placement of CYP (Figure 1 and Supplementary Figure 1).
431
432 Development of Molecular Markers for Multilocus Sequence Analysis
433 To develop molecular markers that may be useful for the MLSA of these phytoplasmas
434 and their relatives, we examined the multiple sequence alignments of those 303 single-copy
435 genes shared by all of the genomes analyzed. The selection criteria were based on: (1) the
436 number of informative sites that may be used to distinguish the three species-level OTUs
437 identified, (2) the possibility of designing PCR primers for a product in the size range of ~700-
438 800 bp and covers a large number of informative sites, and (3) the chromosomal locations. The
439 first two criteria were aimed to make the Sanger sequencing of these markers as cost-effective as
440 possible, while the third criterion was aimed to make the MLSA robust against recombination.
441 Based on these criteria, we selected five markers to supplement the 16S rRNA gene.
442 These five makers all have high densities of informative sites, particularly when compared to
443 those markers that have been developed previously (Table 2 and Supplementary Figure S1). The
444 substitute rates of these marker genes are relatively high but do not deviate strongly when
445 compared to other shared single-copy genes (Supplementary Figure S2).
446 Based on the functional annotation, two of these selected markers (i.e., DegV family
447 protein and TIGR00282 family metallophosphoesterase) are not generally considered as house-
448 keeping genes. These results reflected our philosophy of marker selection (i.e., prioritize the
449 candidates with the highest resolving power for the target taxa regardless of other irrelevant
450 attributes). Similar to our criteria and results, a pioneer study on genome-enable MLSA marker
451 design in phytoplasma selected several genes that were annotated as hypothetical proteins
452 (Kakizawa and Kamagata, 2014). In terms of phylogenetic distribution, we found that four of
453 these five selected genes are present as single-copy genes in other phytoplasmas, including '*Ca*.
454 P. australiense' of the 16SrXII group (Tran-Nguyen et al., 2008), '*Ca*. P. aurantifolia' of the
455 16SrII group (Chung et al., 2013), '*Ca*. P. pruni' of the 16SrIII group (Lee et al., 2015), '*Ca*. P.
456 ziziphi' of the 16SrV group (Wang et al., 2018a), '*Ca*. P. oryzae' of the 16SrXI group (GenBank
457 accession NZ_JHUK00000000), and '*Ca*. P. mali' of the 16SrX group (Kube et al., 2008). The
458 only exception is *degV*, which is absent in two of the draft genomes (i.e., '*Ca*. P. aurantifolia'
459 and '*Ca*. P. oryzae'; may be located in the unassembled regions) and has two tandem copies in
460 the complete genome sequence of '*Ca*. P. mali'. Nevertheless, the sequence divergence levels are
461 too high for all these genes, such that the primers designed in this study (Table 2) may not work
462 for these more divergent phytoplasmas outside of the 16SrI group. These results are consistent

463 with our expectations that different MLSA markers are necessary for different phylogenetic
464 depths.
465 In terms of chromosomal locations, these markers are well separated in all genomes
466 examined despite the extensive genome rearrangements observed (Figure 5). Although the extent
467 of recombination in phytoplasmas is not well understood, recent studies on *Xylella fastidiosa*
468 (i.e., another insect-transmitted plant pathogen that is restricted to vascular tissues) revealed that
469 recombination affected ~0.5-15% of the genome, depending on the exact subspecies (Potnis et
470 al., 2019; Vanhove et al., 2019). The recombined fragments could have sizes up to 31 kb, with an
471 average of 1 kb. Based on these estimates, it is unlikely that one single recombination event
472 would affect more than one of these five markers. Thus, the results of MLSA should be robust
473 against the interference of recombination.
474 When we used these five newly developed markers for molecular phylogenetic inference,
475 the concatenated alignment (together with the 16S rRNA gene, which is expected to be
476 sequenced in the initial characterization of newly collected strains) produced a phylogeny
477 (Figure 1C and Supplementary Figure 1C) that is comparable to the result from genome-scale
478 analysis (Figure 1B and Supplementary Figure 1B). This result demonstrated the usefulness of
479 these markers for future genotyping work. With better sampling of phytoplasma strains and more
480 accurate classification of those strains using these markers, informed decision could be made to
481 select representatives for genome sequencing efforts, which could contribute to the study of
482 phytoplasma genetic diversity and evolutionary history, as well as improve the taxonomy.
483 Moreover, as the sampling focus changes in the future (e.g., a higher resolution is desired for one
484 of the OTUs), the list of conserved single-copy genes and the substitution rates of those genes
485 (Supplementary Table S2) could provide a guide for developing more suitable MLSA markers.
486
487 Phylogenetic Distribution of Phytoplasma Effector Genes
488 With the well-resolved phylogeny (Figure 1B) as a framework, we investigated the
489 distribution of putative effector genes among these phytoplasmas. The results indicated a high
490 level of variation in the gene counts (Figure 6 and Supplementary Table S3). One clade (i.e.,
491 MBSP-M3, De Villa, and LD1) have only ~10-14 genes that encode putative secreted proteins.
492 These numbers are much lower than the ~36-45 found in their sister clade (i.e., OY-M, OY-V,
493 and DY2014). The low numbers are not artifacts of incomplete genome sequences. In fact, the
494 two strains with the lowest numbers (i.e., MBSP-M3 and De Villa) both have the complete
495 chromosomal sequences available. These findings suggest that there may have been lineage-
496 specific reduction or expansion in effector gene counts, although the underlying ecological
497 factors are unclear.
498 For those 16SrI phytoplasma effectors that have been characterized experimentally,
499 namely SAP05 (Gamboa et al., 2019; Huang and Hogenhout, 2019), SAP11/SWP1 (Bai et al.,
500 2009; Sugio et al., 2011a; Lu et al., 2014; Chang et al., 2018; Wang et al. 2018b; Wang et al.
501 2018c), SAP54/PHYL1 (MacLean et al., 2011; Maejima et al., 2014; Orlovskis and Hogenhout,
502 2016), and TENGU (Hoshi et al., 2009; Sugawara et al., 2013; Minato et al., 2014), only
503 TENGU is conserved among all strains. Although some cases of gene absence may be artifacts
504 of incomplete genome assemblies, the strain De Villa (which has the complete chromosomal
505 sequence available) lacks all other three effector genes, indicating that none of those three
506 effectors is essential.
507 The variable pattern of effector gene presence and absence may be linked to the
508 molecular evolution of potential mobile units (PMU) in these phytoplasma genomes (Bai et al.,

2006). These mobile genetic elements were known to facilitate horizontal gene transfers between either closely or distantly related strains and may be gained or lost rapidly (Chung et al., 2013; Ku et al., 2013; Orlovskis et al., 2017; Cho et al., 2019b; Seruga Music et al., 2019). When we examined the chromosomal locations of those four effector genes, TENGU is not associated with PMUs, while the other three are. Finally, as noted in our previous study (Cho et al., 2019b), DY2014 and OY-V both have two copies of SAP11 homologs. One copy is located within a 16SrI-B type PMU, suggesting vertical inheritance, while the other copy is located within a 16SrI-A type PMU, suggesting horizontal acquisition. These observations provided further support to the importance of PMUs in phytoplasma effector evolution.

Conclusion

By quantifying the genetic divergence among a group of closely-related phytoplasmas, this work demonstrated a discontinuity in their similarities of genomic content and homologous sequences. The results suggested that these phytoplasmas could be classified into multiple distinct taxonomic units equivalent to species in other bacteria (Jain et al., 2018). Importantly, the widely used classification system for phytoplasmas that is based on the sequencing and RFLP analysis of their 16S rRNA genes does not provide sufficient resolution or accuracy. Rather, it is important to incorporate genomic information in the future revisions of phytoplasma taxonomy. A previous proposal of using 95% ANI as a cutoff to delineate bacterial species (Jain et al., 2018) appears to work well and warrants further consideration to be adopted. In recognition of the difficulties involved in genomic studies of uncultivated bacteria, this work also demonstrated the feasibility of using genome analysis to develop cost-effective MLSA markers, which will be useful for practical purposes. For future directions, it is critical to improve the taxon sampling of phytoplasma genomes strategically, such that similar approaches may be expand to genus-wide analysis to facilitate the studies of these important plant pathogens. Moreover, continuing effort in establishing axenic culture is critical for the study of these important plant pathogens. In the absence of axenic culture, it is important that the reference strains, particularly those with provisional species status, are made available to the scientific community, preferably in the form of a centralized micropropagation collection as suggested by the international phytoplasma working team (The IRPCM Phytoplasma/Spiroplasma Working Team, 2004). If live culture within plant or insect hosts is not possible, then the DNA samples should be made available. For genome sequencing projects, the raw sequencing results should be made publicly available together with the assembled genome sequences at the time of publication, such that the assemblies could be validated independently. The TW1 genome (Town et al., 2018) provided a good example to illustrate the importance of this point. Although the publication of an erroneous genome sequence is unfortunate, the raw reads made available by those authors allowed us to investigate and identify the issues; such openness is appreciated and important for science to move forward.

Acknowledgments

We thank Ms. Mei-Jane Fang in the Genomic Technology Core Facility (Institute of Plant and Microbial Biology, Academia Sinica) for providing Sanger sequencing service. This manuscript was posted as a preprint at bioRxiv (https://doi.org/10.1101/2020.01.31.928135). Comments received during the peer review process greatly improved this work. This manuscript is not meant to be a comprehensive review of all the available literature on phytoplasma

classification; as such we apologize if we did not cite all publications on this topic in the manuscript.

Funding

The funding was provided by Academia Sinica and the Ministry of Science and Technology of Taiwan (MOST 106-2311-B-001-028-MY3) to C-HK. SAH and WH were supported by the Human Frontier Science Program grant RGP0024/2105 with additional support from Biotechnology and Biological Sciences Research Council grant BB/P012574/1 and the John Innes Foundation. The funders have no role in study design, data collection and interpretation, or the decision to submit this work for publication.

Author Contribution

S-TC, H-JK, and C-HK performed the experiments and analyzed the data. WH and SAH provided intellectual input. S-TC prepared the figures and supplementary materials. S-TC and C-HK wrote the manuscript. SAH edited the manuscript. C-HK designed the experiments, acquired the funding, and supervised the project.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

575 **References**

576 Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S.,
577 et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks.
578 *Nat Biotechnol* 37, 420–423. doi:10.1038/s41587-019-0036-z.
579 Arocha, Y., Antesana, O., Montellano, E., Franco, P., Plata, G., and Jones, P. (2007).
580 '*Candidatus* Phytoplasma lycopersici', a phytoplasma associated with 'hoja de perejil'
581 disease in Bolivia. *Int J Syst Evol Microbiol* 57, 1704–1710. doi:10.1099/ijs.0.64851-0.
582 Bai, X., Zhang, J., Ewing, A., Miller, S. A., Jancso Radek, A., Shevchenko, D. V., et al. (2006).
583 Living with genome instability: the adaptation of phytoplasmas to diverse environments of
584 their insect and plant hosts. *J Bacteriol* 188, 3682–3696. doi:10.1128/JB.188.10.3682-
585 3696.2006.
586 Bai, X., Correa, V. R., Toruño, T. Y., Ammar, E.-D., Kamoun, S., and Hogenhout, S. A. (2009).
587 AY-WB phytoplasma secretes a protein that targets plant cell nuclei. *Mol Plant Microbe*
588 *Interact* 22, 18–30. doi:10.1094/MPMI-22-1-0018.
589 Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., et al.
590 (2018). GenBank. *Nucleic Acids Res* 46, D41–D47. doi:10.1093/nar/gkx1094.
591 Bertaccini, A., and Lee, I.-M. (2018). "Phytoplasmas: An Update," in *Phytoplasmas: Plant*
592 *Pathogenic Bacteria - I: Characterisation and Epidemiology of Phytoplasma - Associated*
593 *Diseases*, eds. G. P. Rao, A. Bertaccini, N. Fiore, and L. W. Liefting (Singapore: Springer
594 Singapore), 1–29. doi:10.1007/978-981-13-0119-3_1.
595 Bobay, L.-M., and Ochman, H. (2017). Biological species are universal across life's domains.
596 *Genome Biol Evol* 9, 491–501. doi:10.1093/gbe/evx026.
597 Brown, D. R. (2010). "Phylum XVI. Tenericutes Murray 1984a, 356VP (Effective publication:
598 Murray 1984b, 33.)," in *Bergey's Manual of Systematic Bacteriology*, eds. N. R. Krieg, J. T.
599 Staley, D. R. Brown, B. P. Hedlund, B. J. Paster, N. L. Ward, et al. (Springer New York),
600 567–723. doi:10.1007/978-0-387-68572-4_5.
601 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009).
602 BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-
603 2105-10-421.
604 Chang, S. H., Tan, C. M., Wu, C.-T., Lin, T.-H., Jiang, S.-Y., Liu, R.-C., et al. (2018).
605 Alterations of plant architecture and phase transition by the phytoplasma virulence factor
606 SAP11. *J Exp Bot* 69, 5389–5401. doi:10.1093/jxb/ery318.
607 Chen, L.-L., Chung, W.-C., Lin, C.-P., and Kuo, C.-H. (2012). Comparative analysis of gene
608 content evolution in phytoplasmas and mycoplasmas. *PLOS ONE* 7, e34407.
609 doi:10.1371/journal.pone.0034407.
610 Chen, W., Li, Y., Wang, Q., Wang, N., and Wu, Y. (2014). Comparative genome analysis of
611 wheat blue dwarf phytoplasma, an obligate pathogen that causes wheat blue dwarf disease in
612 China. *PLOS ONE* 9, e96436. doi:10.1371/journal.pone.0096436.
613 Cho, S.-T., Kung, H.-J., and Kuo, C.-H. (2019a). Genomic similarities among 16SrI
614 phytoplasmas and implications on species boundaries. *Phytopathogenic Mollicutes* 9, 73.
615 doi:10.5958/2249-4677.2019.00037.9.
616 Cho, S.-T., Lin, C.-P., and Kuo, C.-H. (2019b). Genomic characterization of the periwinkle leaf
617 yellowing (PLY) phytoplasmas in Taiwan. *Front Microbiol* 10, 2194.
618 doi:10.3389/fmicb.2019.02194.

619 Chung, W.-C., Chen, L.-L., Lo, W.-S., Lin, C.-P., and Kuo, C.-H. (2013). Comparative analysis
620 of the peanut witches'-broom phytoplasma genome reveals horizontal transfer of potential
621 mobile units and effectors. *PLOS ONE* 8, e62770. doi:10.1371/journal.pone.0062770.
622 Coetzee, B., Douglas-Smit, N., Maree, H. J., Burger, J. T., Krüger, K., and Pietersen, G. (2019).
623 Draft genome sequence of a "*Candidatus* Phytoplasma asteris"-related strain (aster yellows,
624 subgroup 16SrI-B) from South Africa. *Microbiol Resour Announc* 8, e00148-19.
625 doi:10.1128/MRA.00148-19.
626 Dumonceaux, T. J., Green, M., Hammond, C., Perez, E., and Olivier, C. (2014). Molecular
627 diagnostic tools for detection and differentiation of phytoplasmas based on chaperonin-60
628 reveal differences in host plant infection patterns. *PLOS ONE* 9, e116039.
629 doi:10.1371/journal.pone.0116039.
630 Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
631 throughput. *Nucl Acids Res* 32, 1792–1797. doi:10.1093/nar/gkh340.
632 Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs.
633 *Bioinformatics* 34, 2371–2375. doi:10.1093/bioinformatics/bty113.
634 Felsenstein, J. (1989). PHYLIP - phylogeny inference package (version 3.2). *Cladistics* 5, 164–
635 166.
636 Firrao, G., Martini, M., Ermacora, P., Loi, N., Torelli, E., Foissac, X., et al. (2013). Genome
637 wide sequence analysis grants unbiased definition of species boundaries in "*Candidatus*
638 Phytoplasma." *Syst Appl Microbiol* 36, 539–548. doi:10.1016/j.syapm.2013.07.003.
639 Gamboa, C., Cui, W., Quiroga, N., Fernández, C., Fiore, N., and Zamorano, A. (2019).
640 Identification of 16SrIII-J phytoplasma effectors using a viral vector. *Phytopathogenic*
641 *Mollicutes* 9, 229. doi:10.5958/2249-4677.2019.00115.4.
642 Glaeser, S. P., and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic
643 taxonomy. *Syst Appl Microbiol* 38, 237–245. doi:10.1016/j.syapm.2015.03.007.
644 Gundersen, D. E., and Lee, I.-M. (1996). Ultrasensitive detection of phytoplasmas by nested-
645 PCR assays using two universal primer pairs. *Phytopathol Mediterr* 35, 144–151.
646 Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large
647 phylogenies by maximum likelihood. *Syst Biol* 52, 696–704.
648 doi:10.1080/10635150390235520.
649 Guy, L., Roat Kultima, J., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and
650 genome visualization in R. *Bioinformatics* 26, 2334–2335.
651 doi:10.1093/bioinformatics/btq413.
652 Hogenhout, S. A., Oshima, K., Ammar, E.-D., Kakizawa, S., Kingdom, H. N., and Namba, S.
653 (2008). Phytoplasmas: bacteria that manipulate plants and insects. *Mol Plant Pathol* 9, 403–
654 423. doi:10.1111/j.1364-3703.2008.00472.x.
655 Hoshi, A., Oshima, K., Kakizawa, S., Ishii, Y., Ozeki, J., Hashimoto, M., et al. (2009). A unique
656 virulence factor for proliferation and dwarfism in plants identified from a phytopathogenic
657 bacterium. *Proc Natl Acad Sci U S A* 106, 6416–6421. doi:10.1073/pnas.0813038106.
658 Huang, W., and Hogenhout, S. A. (2019). Phytoplasma effectors have converged onto degrading
659 plant transcription factors with fundamental roles in plant development and defense to
660 insects. in *IS-MPMI XVIII Congress Abstracts of Poster Presentations* (Glasgow, UK: IS-
661 MPMI), 158. doi:https://doi.org/10.1094/MPMI-32-10-S1.1.
662 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High
663 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat*
664 *Commun* 9, 5114. doi:10.1038/s41467-018-07641-9.

Kakizawa, S., and Kamagata, Y. (2014). A multiplex-PCR method for strain identification and detailed phylogenetic analysis of AY-group phytoplasmas. *Plant Dis* 98, 299–305. doi:10.1094/PDIS-03-13-0216-RE.

Kakizawa, S., Makino, A., Ishii, Y., Tamaki, H., and Kamagata, Y. (2014). Draft genome sequence of "*Candidatus* Phytoplasma asteris" strain OY-V, an unculturable plant-pathogenic bacterium. *Genome Announc* 2, e00944-14. doi:10.1128/genomeA.00944-14.

Konstantinidis, K. T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *ISME J* 11, 2399–2406. doi:10.1038/ismej.2017.113.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102, 2567–2572. doi:10.1073/pnas.0409727102.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305, 567–580. doi:10.1006/jmbi.2000.4315.

Ku, C., Lo, W.-S., and Kuo, C.-H. (2013). Horizontal transfer of potential mobile units in phytoplasmas. *Mob Genet Elements* 3, e26145. doi:10.4161/mge.26145.

Kube, M., Schneider, B., Kuhl, H., Dandekar, T., Heitmann, K., Migdoll, A., et al. (2008). The linear chromosome of the plant-pathogenic mycoplasma "*Candidatus* Phytoplasma mali." *BMC Genomics* 9, 306. doi:10.1186/1471-2164-9-306.

Kuo, C.-H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res* 19, 1450–1454. doi:10.1101/gr.091785.109.

Kuo, C.-H., and Ochman, H. (2009). Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol Direct* 4, 35. doi:10.1186/1745-6150-4-35.

Lee, I.-M., Davis, R. E., and Gundersen-Rindal, D. E. (2000). Phytoplasma: phytopathogenic mollicutes. *Annu Rev Microbiol* 54, 221–255. doi:10.1146/annurev.micro.54.1.221.

Lee, I.-M., Gundersen-Rindal, D. E., Davis, R. E., and Bartoszyk, I. M. (1998). Revised classification scheme of phytoplasmas based on RFLP analyses of 16S rRNA and ribosomal protein gene sequences. *Int J Syst Evol Microbiol* 48, 1153–1169. doi:10.1099/00207713-48-4-1153.

Lee, I.-M., Gundersen-Rindal, D. E., Davis, R. E., Bottner, K. D., Marcone, C., and Seemüller, E. (2004). '*Candidatus* Phytoplasma asteris', a novel phytoplasma taxon associated with aster yellows and related diseases. *Int J Syst Evol Microbiol* 54, 1037–1048. doi:10.1099/ijs.0.02843-0.

Lee, I.-M., Hammond, R., Davis, R., and Gundersen, D. (1993). Universal amplification and analysis of pathogen 16S rDNA for classification and identification of mycoplasmalike organisms. *Phytopathology* 83, 834–842. doi:10.1094/phyto-83-834.

Lee, I.-M., Shao, J., Bottner-Parker, K. D., Gundersen-Rindal, D. E., Zhao, Y., and Davis, R. E. (2015). Draft genome sequence of "*Candidatus* Phytoplasma pruni" strain CX, a plant-pathogenic bacterium. *Genome Announc* 3, e01117-15. doi:10.1128/genomeA.01117-15.

Lee, I.-M., Zhao, Y., and Bottner, K. D. (2006). SecY gene sequence analysis for finer differentiation of diverse strains in the aster yellows phytoplasma group. *Molecular and Cellular Probes* 20, 87–91. doi:10.1016/j.mcp.2005.10.001.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

711 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence
712 Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
713 doi:10.1093/bioinformatics/btp352.
714 Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for
715 eukaryotic genomes. *Genome Res* 13, 2178–2189. doi:10.1101/gr.1224503.
716 Liefting, L. W., Andersen, M. T., Beever, R. E., Gardner, R. C., and Forster, R. L. (1996).
717 Sequence heterogeneity in the two 16S rRNA genes of *Phormium* yellow leaf phytoplasma.
718 *Appl Environ Microbiol* 62, 3133–3139.
719 Lim, P. O., and Sears, B. B. (1992). Evolutionary relationships of a plant-pathogenic
720 mycoplasmalike organism and *Acholeplasma laidlawii* deduced from two ribosomal protein
721 gene sequences. *J Bacteriol* 174, 2606–2611. doi:10.1128/jb.174.8.2606-2611.1992.
722 Lo, W.-S., Chen, L.-L., Chung, W.-C., Gasparich, G. E., and Kuo, C.-H. (2013). Comparative
723 genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium.
724 *BMC Genomics* 14, 22. doi:10.1186/1471-2164-14-22.
725 Lo, W.-S., Gasparich, G. E., and Kuo, C.-H. (2018). Convergent evolution among ruminant-
726 pathogenic *Mycoplasma* involved extensive gene content changes. *Genome Biol Evol* 10,
727 2130–2139. doi:10.1093/gbe/evy172.
728 Lu, Y.-T., Li, M.-Y., Cheng, K.-T., Tan, C. M., Su, L.-W., Lin, W.-Y., et al. (2014). Transgenic
729 plants that express the phytoplasma effector SAP11 show altered phosphate starvation and
730 defense responses. *Plant Physiol* 164, 1456–1469. doi:10.1104/pp.113.229740.
731 MacLean, A. M., Sugio, A., Makarova, O. V., Findlay, K. C., Grieve, V. M., Tóth, R., et al.
732 (2011). Phytoplasma effector SAP54 induces indeterminate leaf-like flower development in
733 Arabidopsis plants. *Plant Physiol* 157, 831–841. doi:10.1104/pp.111.181586.
734 Maejima, K., Iwai, R., Himeno, M., Komatsu, K., Kitazawa, Y., Fujita, N., et al. (2014).
735 Recognition of floral homeotic MADS domain transcription factors by a phytoplasmal
736 effector, phyllogen, induces phyllody. *Plant J* 78, 541–554. doi:10.1111/tpj.12495.
737 Martini, M., Lee, I.-M., Bottner, K. D., Zhao, Y., Botti, S., Bertaccini, A., et al. (2007).
738 Ribosomal protein gene-based phylogeny for finer differentiation and classification of
739 phytoplasmas. *Int J Syst Evol Microbiol* 57, 2037–2051. doi:10.1099/ijs.0.65013-0.
740 Martini, M., Quaglino, F., and Bertaccini, A. (2019). "Multilocus genetic characterization of
741 phytoplasmas," in *Phytoplasmas: Plant Pathogenic Bacteria - III: Genomics, Host Pathogen*
742 *Interactions and Diagnosis*, eds. A. Bertaccini, K. Oshima, M. Kube, and G. P. Rao
743 (Singapore: Springer), 161–200. doi:10.1007/978-981-13-9632-8_9.
744 Minato, N., Himeno, M., Hoshi, A., Maejima, K., Komatsu, K., Takebayashi, Y., et al. (2014).
745 The phytoplasmal virulence factor TENGU causes plant sterility by downregulating of the
746 jasmonic acid and auxin pathways. *Sci Rep* 4, 7399. doi:10.1038/srep07399.
747 Mitrović, J., Kakizawa, S., Duduk, B., Oshima, K., Namba, S., and Bertaccini, A. (2011). The
748 *groEL* gene as an additional marker for finer differentiation of '*Candidatus* Phytoplasma
749 asteris'-related strains. *Annals of Applied Biology* 159, 41–48. doi:10.1111/j.1744-
750 7348.2011.00472.x.
751 Orlovskis, Z., Canale, M. C., Haryono, M., Lopes, J. R. S., Kuo, C.-H., and Hogenhout, S. A.
752 (2017). A few sequence polymorphisms among isolates of maize bushy stunt phytoplasma
753 associate with organ proliferation symptoms of infected maize plants. *Ann Bot* 119, 869–
754 884. doi:10.1093/aob/mcw213.

755 Orlovskis, Z., and Hogenhout, S. A. (2016). A bacterial parasite effector mediates insect vector
756 attraction in host plants independently of developmental changes. *Front Plant Sci* 7, 885.
757 doi:10.3389/fpls.2016.00885.
758 Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.-Y., Wei, W., Suzuki, S., et al. (2004).
759 Reductive evolution suggested from the complete genome sequence of a plant-pathogenic
760 phytoplasma. *Nat Genet* 36, 27–29. doi:10.1038/ng1277.
761 Pacifico, D., Galetto, L., Rashidi, M., Abbà, S., Palmano, S., Firrao, G., et al. (2015). Decreasing
762 global transcript levels over time suggest that phytoplasma cells enter stationary phase
763 during plant and insect colonization. *Appl Environ Microbiol* 81, 2591–2602.
764 doi:10.1128/AEM.03096-14.
765 Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al.
766 (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises
767 the tree of life. *Nat Biotechnol*. doi:10.1038/nbt.4229.
768 Potnis, N., Kandel, P. P., Merfa, M. V., Retchless, A. C., Parker, J. K., Stenger, D. C., et al.
769 (2019). Patterns of inter- and intrasubspecific homologous recombination inform eco-
770 evolutionary dynamics of *Xylella fastidiosa*. *ISME J*, 1. doi:10.1038/s41396-019-0423-y.
771 Popescu, A.-A., Huber, K. T., and Paradis, E. (2012). ape 3.0: New tools for distance-based
772 phylogenetics and evolutionary analysis in R. *Bioinformatics* 28, 1536–1537.
773 doi:10.1093/bioinformatics/bts184.
774 R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria:
775 R Foundation for Statistical Computing Available at: https://www.r-project.org/ [Accessed
776 August 1, 2019].
777 Sawayanagi, T., Horikoshi, N., Kanehira, T., Shinohara, M., Bertaccini, A., Cousin, M.-T., et al.
778 (1999). '*Candidatus* Phytoplasma japonicum', a new phytoplasma taxon associated with
779 Japanese *Hydrangea* phyllody. *Int J Syst Evol Microbiol* 49, 1275–1285.
780 doi:10.1099/00207713-49-3-1275.
781 Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for
782 defining operational taxonomic units and estimating species richness. *Appl Environ*
783 *Microbiol* 71, 1501–1506. doi:10.1128/AEM.71.3.1501-1506.2005.
784 Schneider, B., and Gibb, K. S. (1997). Sequence and RFLP analysis of the elongation factor Tu
785 gene used in differentiation and classification of phytoplasmas. *Microbiology,* 143, 3381–
786 3389. doi:10.1099/00221287-143-10-3381.
787 Seruga Music, M., Samarzija, I., Hogenhout, S. A., Haryono, M., Cho, S.-T., and Kuo, C.-H.
788 (2019). The genome of '*Candidatus* Phytoplasma solani' strain SA-1 is highly dynamic and
789 prone to adopting foreign sequences. *Syst Appl Microbiol* 42, 117–127.
790 doi:10.1016/j.syapm.2018.10.008.
791 Sparks, M. E., Bottner-Parker, K. D., Gundersen-Rindal, D. E., and Lee, I.-M. (2018). Draft
792 genome sequence of the New Jersey aster yellows strain of '*Candidatus* Phytoplasma
793 asteris.' *PLOS ONE* 13, e0192379. doi:10.1371/journal.pone.0192379.
794 Sugawara, K., Honma, Y., Komatsu, K., Himeno, M., Oshima, K., and Namba, S. (2013). The
795 alteration of plant morphology by small peptides released from the proteolytic processing of
796 the bacterial peptide TENGU. *Plant Physiol* 162, 2005–2014. doi:10.1104/pp.113.218586.
797 Sugio, A., Kingdom, H. N., MacLean, A. M., Grieve, V. M., and Hogenhout, S. A. (2011a).
798 Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating
799 plant development and defense hormone biosynthesis. *Proc Natl Acad Sci U S A* 108,
800 E1254–E1263. doi:10.1073/pnas.1105664108.

801 Sugio, A., MacLean, A. M., Kingdom, H. N., Grieve, V. M., Manimekalai, R., and Hogenhout,
802 S. A. (2011b). Diverse targets of phytoplasma effectors: from plant development to defense
803 against insects. *Annu Rev Phytopathol* 49, 175–195. doi:10.1146/annurev-phyto-072910-
804 095323.
805 Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein
806 sequence alignments into the corresponding codon alignments. *Nucl Acids Res* 34, W609-
807 612. doi:10.1093/nar/gkl315.
808 The IRPCM Phytoplasma/Spiroplasma Working Team - Phytoplasma taxonomy group (2004).
809 "*Candidatus* Phytoplasma", a taxon for the wall-less, non-helical prokaryotes that colonize
810 plant phloem and insects. *Int J Syst Evol Microbiol* 54, 1243–1255. doi:10.1099/ijs.0.02854-
811 0.
812 Tomkins, M., Kliot, A., Marée, A. F., and Hogenhout, S. A. (2018). A multi-layered mechanistic
813 modelling approach to understand how effector genes extend beyond phytoplasma to
814 modulate plant hosts, insect vectors and the environment. *Curr Opin Plant Biol* 44, 39–48.
815 doi:10.1016/j.pbi.2018.02.002.
816 Town, J. R., Wist, T., Perez-Lopez, E., Olivier, C. Y., and Dumonceaux, T. J. (2018). Genome
817 sequence of a plant-pathogenic bacterium, "*Candidatus* Phytoplasma asteris" strain TW1.
818 *Microbiol Resour Announc* 7, e01109-18. doi:10.1128/MRA.01109-18.
819 Tran-Nguyen, L. T. T., Kube, M., Schneider, B., Reinhardt, R., and Gibb, K. S. (2008).
820 Comparative genome analysis of "*Candidatus* Phytoplasma australiense" (subgroup tuf-
821 Australia I; rp-A) and "*Ca*. Phytoplasma asteris" strains OY-M and AY-WB. *J Bacteriol*
822 190, 3979–3991. doi:10.1128/JB.01301-07.
823 Vanhove, M., Retchless, A. C., Sicard, A., Rieux, A., Coletta-Filho, H. D., Fuente, L. D. L., et al.
824 (2019). Genomic diversity and recombination among *Xylella fastidiosa* subspecies. *Appl*
825 *Environ Microbiol* 85, e02972-18. doi:10.1128/AEM.02972-18.
826 Wang, J., Song, L., Jiao, Q., Yang, S., Gao, R., Lu, X., et al. (2018a). Comparative genome
827 analysis of jujube witches'-broom phytoplasma, an obligate pathogen that causes jujube
828 witches'-broom disease. *BMC Genomics* 19, 689. doi:10.1186/s12864-018-5075-1.
829 Wang, N., Li, Y., Chen, W., Yang, H. Z., Zhang, P. H., and Wu, Y. F. (2018b). Identification of
830 wheat blue dwarf phytoplasma effectors targeting plant proliferation and defence responses.
831 *Plant Pathol* 67, 603–609. doi:10.1111/ppa.12786.
832 Wang, N., Yang, H., Yin, Z., Liu, W., Sun, L., and Wu, Y. (2018c). Phytoplasma effector SWP1
833 induces witches' broom symptom by destabilizing the TCP transcription factor
834 BRANCHED1. *Mol Plant Pathol* 19, 2623–2634. doi:10.1111/mpp.12733.
835 Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009).
836 Jalview Version 2—a multiple sequence alignment editor and analysis workbench.
837 *Bioinformatics* 25, 1189–1191. doi:10.1093/bioinformatics/btp033.
838 Wei, W., Lee, I.-M., Davis, R. E., Suo, X., and Zhao, Y. (2008). Automated RFLP pattern
839 comparison and similarity coefficient calculation for rapid delineation of new and distinct
840 phytoplasma 16Sr subgroup lineages. *Int J Syst Evol Microbiol* 58, 2368–2377.
841 doi:10.1099/ijs.0.65868-0.
842 Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24,
843 1586–1591. doi:10.1093/molbev/msm088.
844 Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates
845 under realistic evolutionary models. *Mol Biol Evol* 17, 32–43.

846 Zhao, Y., and Davis, R. E. (2016). Criteria for phytoplasma 16Sr group/subgroup delineation and
847 the need of a platform for proper registration of new groups and subgroups. *Int J Syst Evol*
848 *Microbiol* 66, 2121–2123. doi:10.1099/ijsem.0.000999.
849 Zhao, Y., Wei, W., Lee, I.-M., Shao, J., Suo, X., and Davis, R. E. (2009). Construction of an
850 interactive online phytoplasma classification tool, *i*PhyClassifier, and its application in
851 analysis of the peach X-disease phytoplasma group (16SrIII). *Int J Syst Evol Microbiol* 59,
852 2582–2593. doi:10.1099/ijs.0.010249-0.
853 Zhu, Y., He, Y., Zheng, Z., Chen, J., Wang, Z., and Zhou, G. (2017). Draft genome sequence of
854 rice orange leaf phytoplasma from Guangdong, China. *Genome Announc.* 5, e00430-17.
855 doi:10.1128/genomeA.00430-17.
856
857

858 **Figure Legends**
859 Figure 1. Maximum likelihood phylogenies. The numbers on the internal branches indicate the
860 level of bootstrap support based on 1,000 resampling; only values >=60% are shown. (A) 16S
861 rRNA genes. The alignment contains 1,541 aligned nucleotide sites. The strain name, locus tag
862 of the 16S rRNA gene (in parentheses), and the 16SrI subgroup classification (in square
863 brackets) are labeled. A question mark '?' in the subgroup classification indicates that the
864 sequence is classified as a new subgroup, the existing subgroup with the highest similarity is
865 provided. (B) Single-copy coding genes shared by all strains. The concatenated alignment
866 contains 303 genes and 291,990 aligned nucleotide sites. (C) 16S rRNA gene and the five
867 markers developed in this study; see Table 2 for detailed information. The concatenated
868 alignment contains 4,995 aligned nucleotide sites.
869
870 Figure 2. Pairwise genome similarities. The strains with the complete genome sequences
871 available are highlighted in bold. (A) Similarity scores based on whole genome comparisons.
872 Above diagonal: average nucleotide identity (%); diagonal: genome size (kb); below diagonal:
873 genomic segments mapped (%). (B) Similarity scores based on genic regions. Above diagonal:
874 genic sequence identity (%); diagonal: number of homologous gene clusters; below diagonal:
875 homologs shared (%).
876
877 Figure 3. Correlations between metrics of genome similarities. (A) Similarity scores based on
878 whole genome comparisons. (B) Similarity scores based on genic regions. Filled symbols
879 indicate the pairwise comparisons involving two complete genomes, empty symbols indicate the
880 pairwise comparisons involving one or two draft genomes. The strain TW1 is a special case; its
881 similarity scores in comparison with 16SrI-A and 16SrI-B strains are indicated by red and blue,
882 respectively.
883
884 Figure 4. Principal coordinates analysis of gene content. Each genome is represented by one dot
885 (color-coded according to the 16Sr subgroup assignment). The distance between dots
886 corresponds to their differences in gene content. The % variance explained by each axis is
887 provided in parentheses. The strains with the complete genome sequences available are
888 highlighted in bold.
889
890 Figure 5. Genomic locations of marker genes. Nucleotide sequence identity of the conserved
891 regions between genomes are illustrated (red: match on the same strand; blue: match on the
892 opposite strand).
893
894 Figure 6. Phylogenetic distribution of putative effector genes. The cladogram that illustrates the
895 phylogenetic relationships among these strains is based on Figure 1B (i.e., maximum likelihood
896 phylogeny based on 303 shared single-copy genes). The number in parentheses after the strain
897 name indicates the number of genes that encode putative secreted proteins in each genome
898 (including putative pseudogenes). The strains with the complete genome sequences available are
899 highlighted in bold. On the right-hand side, the table lists the locus tags of homologs for those
900 four phytoplasma effectors that have been characterized. Those locus tags with an asterisk (*)
901 were annotated as putative pseudogenes.
902
903

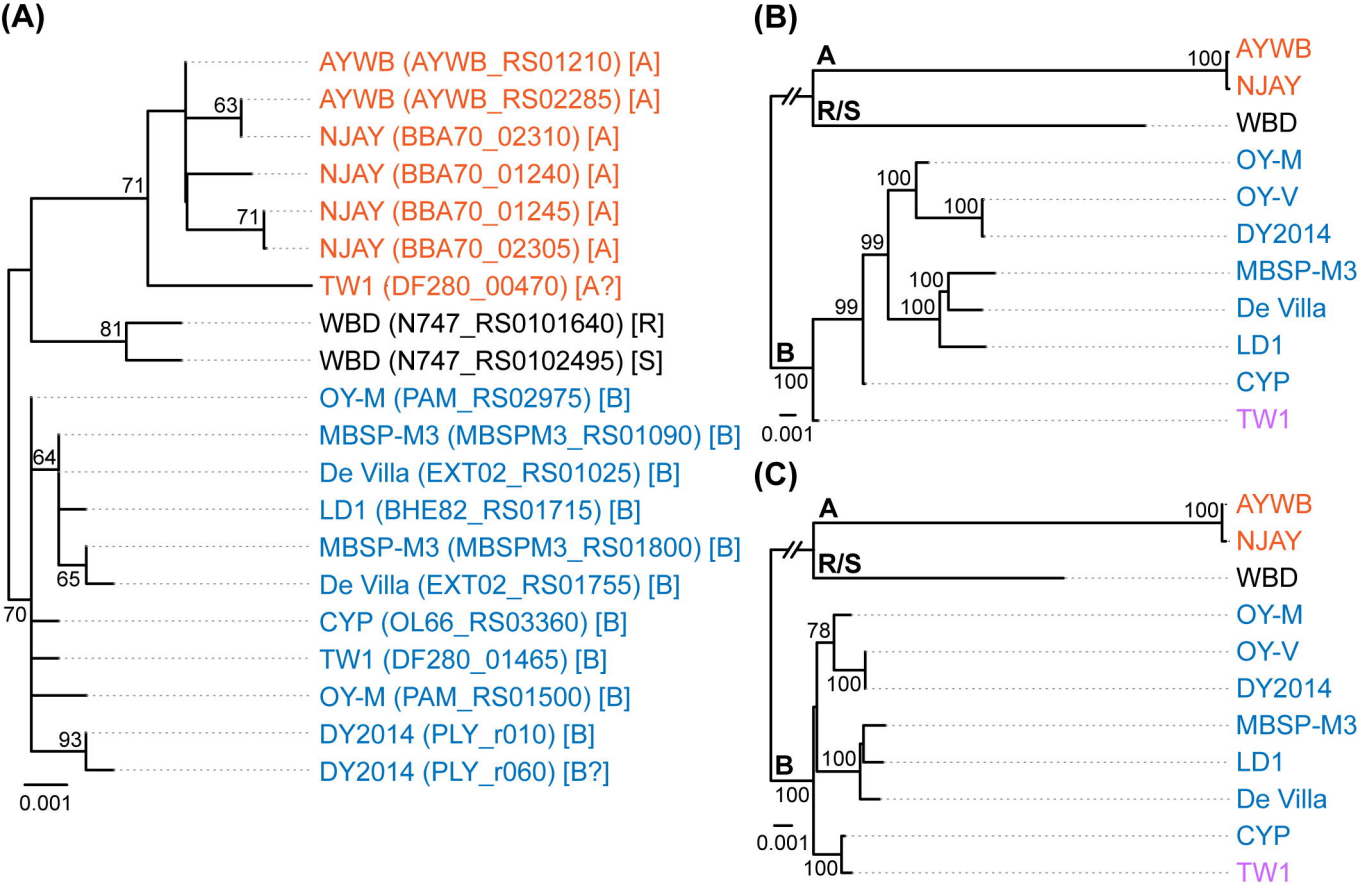
904 **Supplementary Materials**
905 Supplementary Figure S1. Maximum likelihood phylogenies. Based on Figure 1, the TW1
906 sequences were excluded prior to multiple sequence alignment. The numbers on the internal
907 branches indicate the level of bootstrap support based on 1,000 resampling; only values >=60%
908 are shown. (A) 16S rRNA genes. The strain name, locus tag of the 16S rRNA gene (in
909 parentheses), and the 16SrI subgroup classification (in square brackets) are labeled. A question
910 mark '?' in the subgroup classification indicates that the sequence is classified as a new
911 subgroup, the existing subgroup with the highest similarity is provided. (B) The 303 single-copy
912 coding genes shared by all strains. (C) 16S rRNA gene and the five markers developed in this
913 study; see Table 2 for detailed information.
914
915 Supplementary Figure S2. Multiple sequence alignments of selected marker genes. The primer
916 sites are highlighted in orange. Shades of blue colors in the alignment indicate the levels of
917 sequence conservation.
918
919 Supplementary Figure S3. Distribution of substitution rates among those single-copy genes
920 shared by the 11 phytoplasma genomes analyzed. The synonymous substitution rate (Ks) and
921 non-synonymous substitution rate (Ka) were calculated based on the homologs found in AYWB
922 and MBSP-M3. Among those 303 single-copy genes, nine were annotated as pseudogenes and
923 were omitted from the substitution rate calculation.
924
925 Supplementary Table S1. Pairwise nucleotide sequence identity of the molecular markers used in
926 the phylogenetic inference. Numbers in the diagonal cells indicate the length of each unaligned
927 sequence.
928
929 Supplementary Table S2. List of the 303 single-copy genes shared by all of the 11 phytoplasma
930 genomes analyzed in this study. The gene identity, genomic location, and functional annotation
931 of these genes are provided based on the information from AYWB. The substitution rate
932 estimates are based on the comparison between AYWB and MBSP-M3.
933
934 Supplementary Table S3. List of the putative effector genes. Each gene is uniquely identified by
935 its locus_tag. The "Cluster_id" field provides the assignment of homologous gene cluster; genes
936 sharing the same Cluster_id are considered as homologs. Note that SAP11/SWP1 homologs are
937 separated into two clusters (i.e., 503 and 598) due to high levels of sequence divergence.
938 Additionally, one SAP05 homolog is in a separate cluster (i.e., 1117) due to truncation. The gene
939 names and product description are based on existing GenBank annotation; additional information
940 (e.g., homology information with known effector or pseudogene status) are provided in the
941 "Note" field.
942
943

Table 1. List of the genome sequences analyzed.

Strain	Accession	Sequencing platform and depth	Assembly	Genome Size (bp)	CDS-intact	CDS-pseudo	Location	Host	Reference
AYWB	GCF_000012225	Sanger; not available	Complete	723,970	557	116	USA	Lettuce	Bai et al., 2006
NJAY	GCA_002554195	Roche/454; 200x	75 contigs	652,092	371	262	USA	Lettuce	Sparks et al., 2018
WBD	GCF_000495255	Illumina; 5,000x	6 contigs	611,462	471	76	China	Wheat	Chen et al., 2014
OY-M	GCF_000009845	Sanger; not available	Complete	853,092	728	144	Japan	Onion	Oshima et al., 2004
OY-V	GCF_000744065	Illumina and Roche/454; 300x	170 contigs	739,592	651	217	Japan	Onion	Kakizawa et al., 2014
DY2014	GCA_005093185	Illumina; 124x	8 contigs	824,596	775	63	Taiwan	Periwinkle	Cho et al., 2019b
MBSP-M3	GCF_001712875	Illumina; 42x	Complete	576,118	498	34	Brazil	Maize	Orlovskis et al., 2017
De Villa	GCF_004214875	Illumina; 50x	Complete	603,949	512	21	South Africa	Grapevine	Coetzee et al., 2019
LD1	GCF_001866375	Illumina; 200x	8 contigs	599,264	536	29	China	Rice	Zhu et al., 2017
CYP	GCF_000803325	Illumina; 47x	252 contigs	659,699	546	154	Italy	Daisy	Pacifico et al., 2015
TW1	GCA_003181115	Illumina and ONT; 146x	5 contigs	743,598	457	248	Canada	Canola	Town et al., 2018

Table 2. List of molecular markers for phytoplasma classification.

Marker	**Primers (5'-3')**	**Amplicon Size (bp)**	**Variable Sites**	**Informative Sites**	**References**
16S rRNA	[R16F2n] GAAACGACTGCTAAGACTGG [R16R2] TGACGGGCGGTGTGTACAAACCCCG	1,246	15	6	Gundersen and Lee, 1996
Ribosomal proteins (*rpl22* and *rps3*)	[rpF1] GGACATAAGTTAGGTGAATTT [rpR1] ACGATATTTAGTTCTTTTTGG	1,240	70	25	Lim and Sears, 1992
Ribosomal proteins (*rpl22* and *rps3*)	[rpF1C] ATGGTDGGDCAYAARTTAGG [rp(I)R1A] GTTCTTTTTGGCATTAACAT	1,236	70	25	Martini et al., 2007
Preprotein translocase SecY (*secY*)	[AYsecYF1] CAGCCATTTTAGCAGTTGGTGG [AYsecYR1] CAGAAGCTTGAGTGCCTTTACC	1,358	105	62	Lee et al., 2006
Chaperonin 60 (*groEL*/*cpn60*)	[H279p] GATIIIGCAGGIGATGGAACMACIAC [H280p] TGRTTITCICCAAAACCAGGIGCATT	605	21	12	Dumonceaux et al., 2014
Chaperonin 60 (*groEL*/*cpn60*)	[AYgroelF] GGCAAAGAAGCAAGAAAAG [AYgroelR] TTTAAGGGTTGTAAAAGTTG	1,396	59	34	Mitrović et al., 2011
Elongation factor Tu (*tuf*)	[fTufu] CCTGAAGAAAGAGAACGTGG [rTufu] CGGAAATAGAATTGAGGACG	842	39	23	Schneider and Gibb, 1997
Replication initiation protein DnaD (*dnaD*)	[F] CACAAGAAAAATTAGAAGCTC [R] ATAAGTTAAAAGCACATTGAC	767	77	60	This study
DegV family protein (*degV*)	[F] GTAGTTGATTCTACTTGCGG [R] ACTACAGGAGAAATAGCACC	785	57	36	This study
TIGR00282 family metallophosphoesterase	[F] AGATATTTACGGAAACCCAG [R] AGGTTTAAGAGTGACAAGTAAA	708	47	30	This study
Preprotein translocase SecY (*secY*)	[F] AATTGTTGTTTCGATGAGCC [R] TTGGCAGTAGCTTTGATGCG	711	64	36	This study
RluA family pseudouridine synthase (*rluA*)	[F] AAAGAGTTTCTTATTTCTGCCAG [R] GACCTAAAGGAGTAATATGGTG	778	80	54	This study


(A)
AYWB (AYWB_RS01210) [A]
AYWB (AYWB_RS02285) [A]
NJAY (BBA70_02310) [A]
NJAY (BBA70_01240) [A]
NJAY (BBA70_01245) [A]
NJAY (BBA70_02305) [A]
TW1 (DF280_00470) [A?]
WBD (N747_RS0101640) [R]
WBD (N747_RS0102495) [S]
OY-M (PAM_RS02975) [B]
MBSP-M3 (MBSPM3_RS01090) [B]
De Villa (EXT02_RS01025) [B]
LD1 (BHE82_RS01715) [B]
MBSP-M3 (MBSPM3_RS01800) [B]
De Villa (EXT02_RS01755) [B]
CYP (OL66_RS03360) [B]
TW1 (DF280_01465) [B]
OY-M (PAM_RS01500) [B]
DY2014 (PLY_r010) [B]
DY2014 (PLY_r060) [B?]
63
71
71
81
64
65
70
93
0.001
(B)
A
R/S
B
100
AYWB
NJAY
WBD
OY-M
OY-V
DY2014
MBSP-M3
De Villa
LD1
CYP
TW1
100
100
99
100
100
99
100
0.001
(C)
A
R/S
B
100
AYWB
NJAY
WBD
OY-M
OY-V
DY2014
MBSP-M3
LD1
De Villa
CYP
TW1
78
100
100
100
0.001
100

(A)

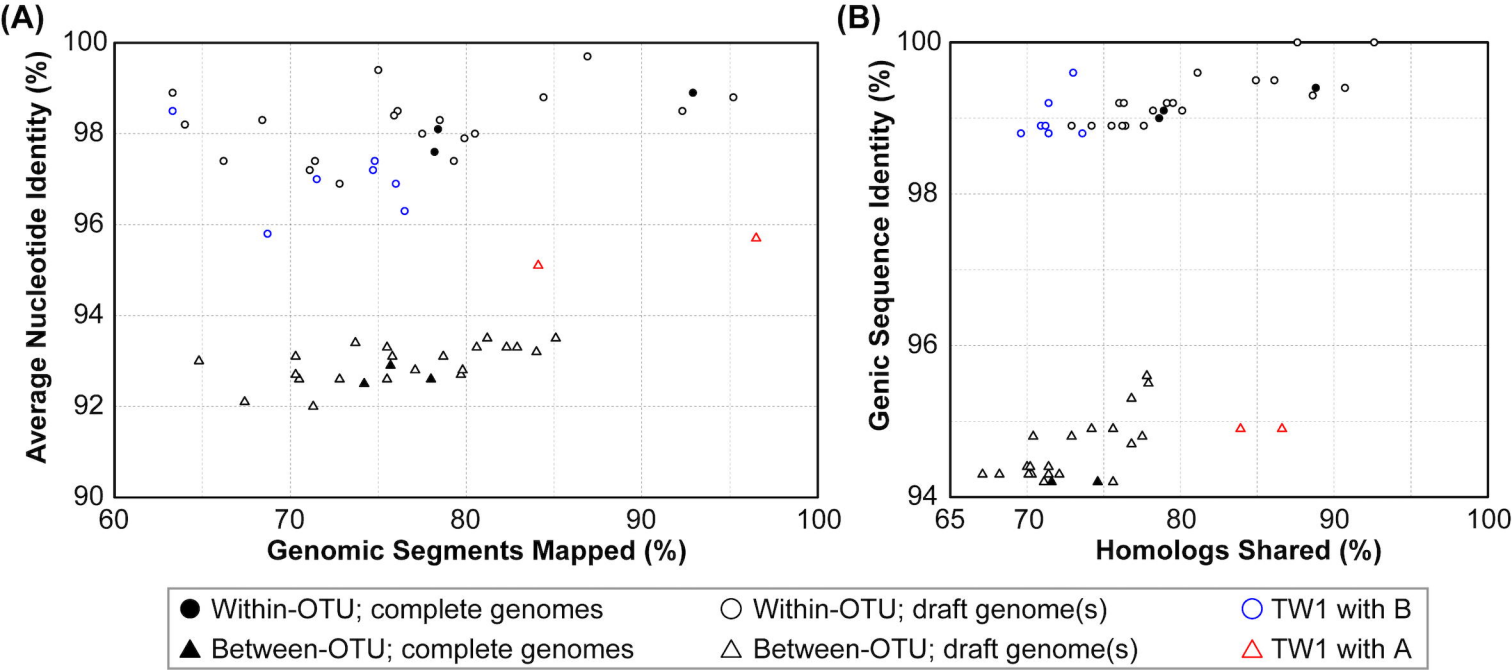
	AYWB	NJAY	WBD	OY-M	OY-V	DY2014	MBSP-M3	De Villa	LD1	CYP	TW1
AYWB	724	99.7	93.3	92.5	92.1	92.6	92.6	92.9	93.1	93.0	95.7
NJAY	86.9	652	93.5	92.6	92.0	92.6	92.7	92.8	92.8	93.1	95.1
WBD	82.3	81.2	611	93.3	92.7	93.1	93.2	93.3	93.5	93.4	93.3
OY-M	74.2	70.5	75.5	853	98.3	98.8	97.6	98.1	98.3	98.9	97.0
OY-V	67.4	71.3	70.3	68.4	740	99.4	96.9	97.2	97.4	97.4	95.8
DY2014	75.5	72.8	78.7	84.4	75.0	825	97.4	97.9	98.0	98.2	96.3
MBSP-M3	78.0	79.7	84.0	78.2	72.8	79.3	576	98.9	98.5	98.0	96.9
De Villa	75.7	77.1	82.9	78.4	71.1	79.9	92.9	604	98.8	98.5	97.2
LD1	75.8	79.8	85.1	78.5	71.4	80.5	92.3	95.2	599	98.4	97.4
CYP	64.8	70.3	73.7	63.3	66.2	64.0	77.5	76.1	75.9	660	98.5
TW1	96.5	84.1	80.6	71.5	68.7	76.5	76.0	74.7	74.8	63.3	744

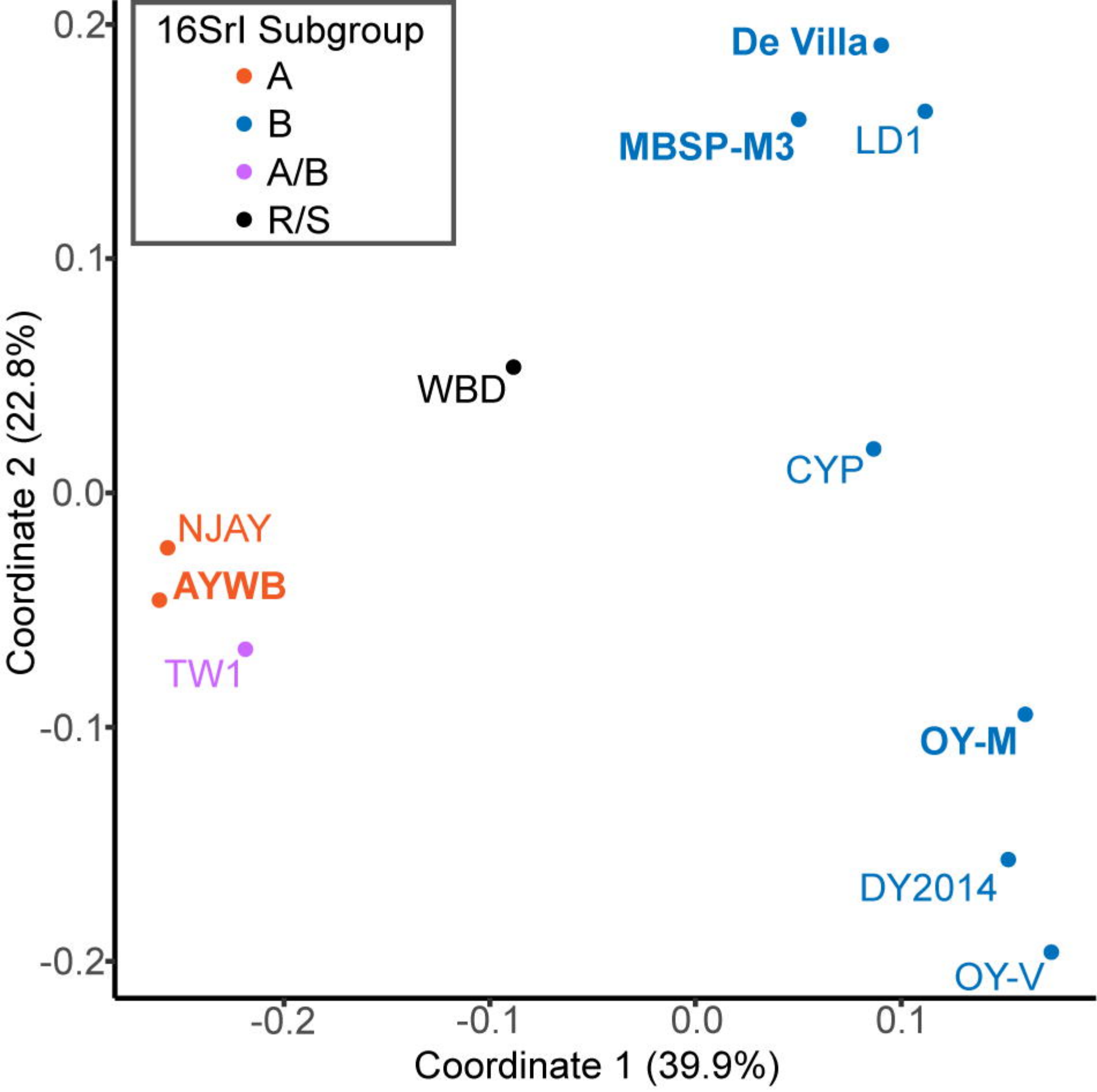
Above diagonal: Average Nucleotide Identity (%) 95 99
Diagonal: Genome Size (kb)
Below diagonal: Genomic Segments Mapped (%) 70 80

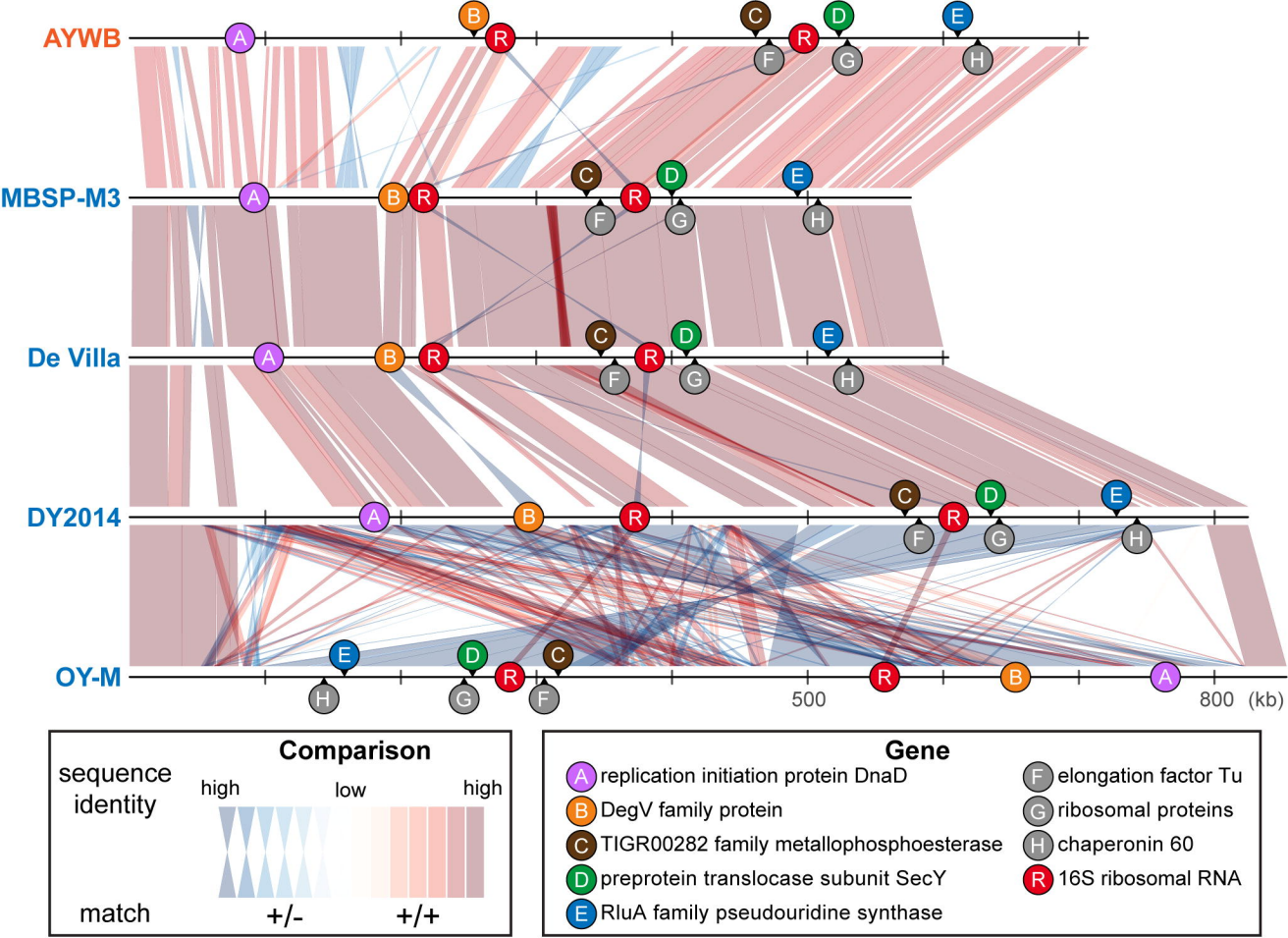
(B)

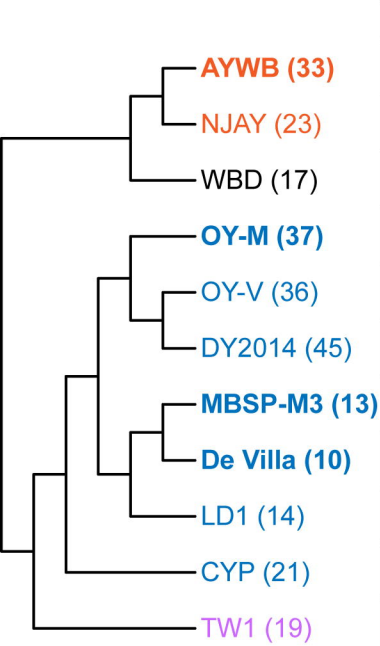
	AYWB	NJAY	WBD	OY-M	OY-V	DY2014	MBSP-M3	De Villa	LD1	CYP	TW1
AYWB	570	100.0	95.5	94.3	94.3	94.3	94.2	94.2	94.2	94.4	94.9
NJAY	92.6	551	95.6	94.4	94.3	94.3	94.2	94.3	94.3	94.4	94.9
WBD	77.9	77.8	524	94.9	94.8	94.8	94.7	94.8	94.7	94.9	95.3
OY-M	70.2	70.0	74.2	660	99.5	99.5	99.0	99.1	99.1	99.6	99.2
OY-V	68.2	67.1	70.4	84.9	718	100.0	98.9	98.9	98.9	99.2	98.8
DY2014	70.3	70.1	72.9	86.1	87.6	688	98.9	98.9	98.9	99.2	98.9
MBSP-M3	74.6	75.6	76.8	78.6	74.2	76.4	505	99.4	99.3	99.1	98.8
De Villa	71.6	72.1	77.5	78.9	72.9	76.2	88.8	506	99.4	99.2	98.9
LD1	71.1	71.4	76.8	80.1	75.5	77.6	88.6	90.7	533	99.2	98.8
CYP	71.4	70.2	75.6	81.1	76.0	76.3	78.2	79.1	79.5	635	99.6
TW1	86.6	83.9	76.8	71.4	69.6	70.9	73.6	71.2	71.4	73.0	598

Above diagonal: Genic Sequence Identity (%) 98 99.9
Diagonal: Number of Homologous Gene Clusters
Below diagonal: Homologs Shared (%) 70 80


(A)
Average Nucleotide Identity (%)
100
98
96
94
92
90
60
70
80
90
100
Genomic Segments Mapped (%)
(B)
Genic Sequence Identity (%)
100
98
96
94
65
70
80
90
100
Homologs Shared (%)
Within-OTU; complete genomes
Within-OTU; draft genome(s)
TW1 with B
Between-OTU; complete genomes
Between-OTU; draft genome(s)
TW1 with A


16SrI Subgroup
A
B
A/B
R/S
De Villa
MBSP-M3
LD1
WBD
CYP
NJAY
AYWB
TW1
OY-M
DY2014
OY-V
0.2
0.1
0.0
-0.1
-0.2
Coordinate 2 (22.8%)
-0.2
-0.1
0.0
0.1
Coordinate 1 (39.9%)


AYWB
MBSP-M3
De Villa
DY2014
OY-M
A
B
C
D
E
F
G
H
R
500
800 (kb)
Comparison
sequence identity
high
low
high
match
+/-
+/+
Gene
A replication initiation protein DnaD
B DegV family protein
C TIGR00282 family metallophosphoesterase
D preprotein translocase subunit SecY
E RluA family pseudouridine synthase
F elongation factor Tu
G ribosomal proteins
H chaperonin 60
R 16S ribosomal RNA



	SAP05	SAP11/SWP1	SAP54/PHYL1	TENGU
AYWB (33)	AYWB_RS00160	AYWB_RS01770	AYWB_RS01055	AYWB_RS00785
NJAY (23)	BBA70_00185	BBA70_01750	BBA70_01070*	BBA70_00770
WBD (17)	N747_RS0100715	N747_RS0100020	N747_RS03525*	N747_RS0101805
OY-M (37)	PAM_RS03110	PAM_RS03115	PAM_RS00330	PAM_RS02860
OY-V (36)	OYV_RS02545	OYV_RS02510* OYV_RS02550	OYV_RS03895	OYV_RS02965*
DY2014 (45)	PLY_1120	PLY_1110 PLY_3690	PLY_0730	PLY_4530
MBSP-M3 (13)		MBSPM3_RS00550		MBSPM3_RS01195
De Villa (10)				EXT02_RS01135
LD1 (14)				BHE82_RS02390
CYP (21)	OL66_RS02400 OL66_RS02415		OL66_RS03105	OL66_RS00470
TW1 (19)	DF280_02690	DF280_00985*	DF280_01620	DF280_01810