Comparative Genome Analysis of '*Candidatus* Phytoplasma luffae' Reveals the Influential Roles of Potential Mobile Units in Phytoplasma Evolution

Running Title: PMU and Phytoplasma Genome Evolution

Ching-Ting Huang[1], Shu-Ting Cho[1], Choon-Meng Tan[2], Yi-Ching Chiu[2,3], Jun-Yi Yang[2,3,4]*, Chih-Horng Kuo[1,3,5,6]*

[1] Institute of Plant and Microbial Biology, Academia Sinica, Taipei 115, Taiwan

[2] Graduate Institute of Biochemistry, National Chung Hsing University, Taichung 402, Taiwan

[3] PhD Program in Microbial Genomics, National Chung Hsing University and Academia Sinica, Taichung 402, Taiwan

[4] Advanced Plant Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan

[5] Molecular and Biological Agricultural Sciences Program, Taiwan International Graduate Program, National Chung-Hsing University and Academia Sinica, Taipei 115, Taiwan

[6] Biotechnology Center, National Chung Hsing University, Taichung 402, Taiwan

* Correspondence

Jun-Yi Yang: jyang@nchu.edu.tw

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

Abstract

Phytoplasmas are insect-transmitted plant pathogens that cause substantial losses in agriculture. In addition to economic impact, phytoplasmas induce distinct disease symptoms in infected plants, thus attracting attention for research on molecular plant-microbe interactions and plant developmental processes. Due to the difficulty of establishing an axenic culture of these bacteria, culture-independent genome characterization is a crucial tool for phytoplasma research. However, phytoplasma genomes have strong nucleotide composition biases and are repetitive, which make it challenging to produce complete assemblies. In this study, we utilized Illumina and Oxford Nanopore sequencing technologies to obtain the complete genome sequence of '*Candidatus* Phytoplasma luffae' strain NCHU2019 that is associated with witches' broom disease of loofah (*Luffa aegyptiaca*) in Taiwan. The fully assembled circular chromosome is 769 kb in size and is the first representative genome sequence of group 16SrVIII phytoplasmas. Comparative analysis with other phytoplasmas revealed that NCHU2019 has an exceptionally repetitive genome, possessing a pair of 75 kb repeats and at least 13 potential mobile units (PMUs) that account for ~25% of its chromosome. This level of genome repetitiveness is exceptional for bacteria, particularly among obligate pathogens with reduced genomes. Our genus-level analysis of PMUs demonstrated that these phytoplasma-specific mobile genetic elements can be classified into three major types that differ in gene organization and phylogenetic distribution. Notably, PMU abundance explains nearly 80% of the variance in phytoplasma genome sizes, a finding that provides a quantitative estimate for the importance of PMUs in phytoplasma genome variability. Finally, our investigation found that in addition to horizontal gene transfer, PMUs also contribute to intra-genomic duplications of effector genes, which may provide redundancy for neofunctionalization or subfunctionalization.

Taken together, this work improves the taxon sampling for phytoplasma genome research and provides novel information regarding the roles of mobile genetic elements in phytoplasma evolution.

Keywords

Plant pathogen, phytoplasma, genomics, molecular evolution, mobile genetic element, effector

56 **Introduction**

57 Phytoplasmas are uncultivated bacteria associated with plant diseases in several
58 hundred species (Lee et al., 2000; Hogenhout et al., 2008; Bertaccini et al., 2014; Namba,
59 2019). In infected plants, phytoplasma cells are restricted to phloem tissues and can secrete
60 effector proteins that cause developmental abnormalities of the hosts (Sugio et al., 2011b).
61 Typical symptoms of phytoplasma infections include stunting, dwarfism, virescence (i.e.,
62 greening of flowers), phyllody (i.e., abnormal development of floral parts into leaf-like
63 tissues), and witches' broom (i.e., proliferation of stems and leaves), which result in
64 substantial agricultural losses.

65 For classification of these uncultivated bacteria, a system based on restriction
66 fragment length polymorphism (RFLP) analysis of their 16S rRNA genes was developed in
67 the 1990s (Lee et al., 1993, 1998; Gundersen and Lee, 1996) and at least 33 16S rRNA gene
68 RFLP (16Sr) groups have been described (Zhao et al., 2009; Zhao and Davis, 2016). Later, a
69 provisional genus-level taxon '*Candidatus* Phytoplasma' was proposed to accommodate
70 these bacteria (The IRPCM Phytoplasma/Spiroplasma Working Team - Phytoplasma
71 taxonomy group, 2004) and at least 41 '*Ca*. P.' species have been described or proposed
72 (Bertaccini and Lee, 2018). Based on analysis of 16S rRNA and other conserved genes,
73 phytoplasmas are divided into three major phylogenetic clusters (Hogenhout and Seruga
74 Music, 2009; Chung et al., 2013; Seruga Music et al., 2019). Early genomics studies were
75 mainly conducted for clusters I (i.e., '*Ca*. P. asteris' of group 16SrI and '*Ca*. P. australiense' of
76 16SrXII) (Oshima et al., 2004; Bai et al., 2006; Tran-Nguyen et al., 2008) and II (i.e., '*Ca*. P.
77 mali' of 16SrX) (Kube et al., 2008). In comparison, cluster III harbors the highest level of
78 diversity, yet has received limited attention for comparative genomics studies (Chung et al.,
79 2013; Wang et al., 2018a).

80 To improve our understanding of phytoplasma genomes, we conducted whole
81 genome sequencing of a '*Ca*. P. luffae' strain collected in Taiwan. The species-level taxon '*Ca*.
82 P luffae' belongs to group 16SrVIII in cluster III and is associated with witches' broom
83 disease of loofah (*Luffa aegyptiaca*) (Davis et al., 2017). The availability of a complete
84 genome sequence from this taxon provides a complete view of its gene content, which can
85 facilitate the study of its pathogenesis mechanisms and other aspects of biology. More
86 importantly, with increased availability of genome sequences from diverse phytoplasmas
87 (**Table 1**), we performed genus-level comparative analysis to obtain a more comprehensive
88 picture of their genomic diversity. This improves upon previous works that are limited to
89 comparisons of closely related taxa or have sparse sampling (Bai et al., 2006; Tran-Nguyen
90 et al., 2008; Kube et al., 2012; Saccardo et al., 2012; Andersen et al., 2013; Chung et al.,
91 2013; Orlovskis et al., 2017; Wang et al., 2018a; Cho et al., 2019, 2020a; Seruga Music et al.,
92 2019; Kirdat et al., 2021; Zhao et al., 2021). Furthermore, our focused analysis of the
93 potential mobile units (PMUs) (Bai et al., 2006) revealed the influential roles of these mobile
94 genetic elements in the evolution of phytoplasma genome organization and effector gene
95 content.
96
97 **Materials and Methods**
98 Biological Materials
99 The strain NCHU2019 was collected from a naturally infected loofah plant found in
100 Dacheng Township (Changhua County, Taiwan; 23.861860 N, 120.291322 E) on July 4th,
101 2019 (**Figure 1A**). After initial collection, the bacterium was transferred to lab-grown loofah
102 plants (cultivar A-Jun, Known-You Seed Co., Kaohsiung, Taiwan) through grafting and
103 maintained in a plant growth facility in the National Chung Hsing University (Taichung,

104 Taiwan) (**Figure 1B**). To confirm the presence and identity of this phytoplasma strain, a
105 partial sequence of the rRNA operon was PCR amplified using the phytoplasma-specific
106 primer set P1/P7 for Sanger sequencing as described (Chung et al., 2013).
107
108 Genome Sequencing
109 The procedures for genome sequencing and analysis were based on those described
110 in our previous work on phytoplasma genomes (Chung et al., 2013; Cho et al., 2019, 2020a,
111 2020b). All kits were used according to the manufacturer's protocols and all bioinformatics
112 tools were used with the default settings unless stated otherwise.
113 For whole genome shotgun sequencing, leaves from one artificially infected plant
114 exhibiting typical symptoms (i.e., small leaves and witches' broom) (**Figure 1B**) were
115 collected for total genomic DNA extraction using the Wizard Genomic DNA Purification Kit
116 (A1120; Promega, USA). For Illumina sequencing, the DNA sample was processed using the
117 KAPA Library Preparation Kit (KK8234) and the Invitrogen SizeSelect Gels (G6610-02) to
118 obtain ~550 bp fragments, followed by MiSeq 2x300 bp paired-end sequencing (v3
119 chemistry). For Oxford Nanopore Technologies (ONT) sequencing, the library was prepared
120 using the ONT Ligation Kit (SQK-LSK109) without shearing or size selection, followed by
121 MinION sequencing (R9.4.1 chemistry) and Guppy v3.3.3 base-calling.
122 The first round of *de novo* assembly was performed using Velvet v1.2.10 (Zerbino
123 and Birney, 2008) with only the Illumina reads. The draft assembly was queried against the
124 NCBI non-redundant protein database (Benson et al., 2018) using BLASTX v2.10.0 (Camacho
125 et al., 2009) to identify putative phytoplasma contigs. These contigs were used as the
126 reference for extracting phytoplasma reads from the Illumina data set using BWA v0.7.17 (Li
127 and Durbin, 2009) with an alignment score cutoff of 30 and from the ONT data set using

128 Minimap2 v2.15 (Li, 2018) with an alignment score cutoff of 1000. The extracted reads from
129 both data sets were processed together using Unicycler v0.4.9b for the second round of *de*
130 *novo* assembly. The putative phytoplasma contigs were identified by BLAST searches and
131 used as the starting point for an iterative process of assembly improvement. In each
132 iteration, all Illumina and ONT raw reads were mapped to the draft assembly as described.
133 The mapping results were programmatically checked using SAMTOOLS v1.9 (Li et al., 2009)
134 and manually inspected using IGV v2.5.0 (Robinson et al., 2011) to identify assembly errors
135 and to close gaps.
136 After complete sequence of the circular chromosome was obtained, gene prediction
137 was performed using RNAmmer v1.2 (Lagesen et al., 2007), tRNAscan-SE v1.3.1 (Lowe and
138 Eddy, 1997), and Prodigal v2.6.3 (Hyatt et al., 2010). The annotation was based on the
139 homologous gene clusters present in other phytoplasma genomes (**Table 1**) as identified by
140 OrthoMCL v1.3 (Li et al., 2003), followed by manual curation based on information obtained
141 from GenBank (Benson et al., 2018), KEGG (Kanehisa et al., 2010), and COG (Tatusov et al.,
142 2003) databases. Additionally, putative secreted proteins were predicted using SignalP v5.0
143 (Armenteros et al., 2019) based on the Gram-positive bacteria model. Those candidates
144 with transmembrane domains as identified by TMHMM v2.0 (Krogh et al., 2001) were
145 removed and the remaining ones were required to have a signal peptide length in the range
146 of 21 to 52 amino acids. For visualization, the Circos v0.69-6 (Krzywinski et al., 2009) was
147 used to draw the genome map.
148
149 Comparative Analysis
150 For comparative analysis with other representative phytoplasma genomes (**Table 1**),
151 homologous gene clusters were identified using OrthoMCL (Li et al., 2003). Multiple

152 sequence alignments of homologous genes were prepared using MUSCLE v3.8.31 (Edgar,
153 2004) and visualized using JalView v2.11 (Waterhouse et al., 2009). Maximum likelihood
154 phylogenies were inferred using PhyML v3.3 (Guindon and Gascuel, 2003) and visualized
155 using FigTree v1.4.4. PHYLIP v3.697 (Felsenstein, 1989) was used for bootstrap analysis.
156 Classification of the 16S RFLP group was performed using *i*PhyClassifier (Zhao et al., 2009).
157 For whole-genome comparison, FastANI v1.1 (Jain et al., 2018) was used to calculate
158 the proportion of genomic segments mapped and the average nucleotide identity (ANI). For
159 pairwise genome alignments, Mummer v3.23 (Kurtz et al., 2004) was used with the options
160 "--maxmatch --mincluster 30" and the results were visualized using genoPlotR v0.8.9 (Guy et
161 al., 2010).
162 The PMU analysis was based on the eight core genes (i.e., *tra5*, *dnaB*, *dnaG*, *tmk*,
163 *hflB*, *himA*, *ssb*, and *rpoD*) defined previously (Bai et al., 2006). To ensure uniform
164 annotation and to include possible pseudogenes, all genome assemblies were examined by
165 using available PMU genes sequences as queries to run TBLASTN searches. All statistical
166 tests were performed in the R statistical environment; correlation coefficients were
167 calculated using the 'cor.test' function, linear regression was performed using the 'lm'
168 function and visualized using the 'plot' and 'abline' functions.
169
170 **Results and Discussion**
171 Genomic Characterization of NCHU2019
172 The shotgun sequencing generated ~4.7 Gb of Illumina raw reads and ~3.5 Gb of
173 ONT raw reads. As expected, most of the reads are derived from the plant host. By mapping
174 to the finalized phytoplasma genome assembly, ~1.0% of the Illumina reads and ~5.1% of
175 the ONT reads are derived from phytoplasma DNA, which corresponds to ~51- and ~80-fold

176 sequencing depth, respectively. The *de novo* genome assembly produced one complete
177 circular chromosome that is 769,143 bp in size with 23.3% G+C content; no plasmid was
178 found (**Figure 2**).

179 The annotation of this phytoplasma genome contains two complete sets of 16S-23S-
180 5S rRNA gene clusters, 31 tRNA genes, 725 coding sequences (CDSs), and 13 pseudogenes.
181 Both copies of the 16S rRNA gene are 100% identical to the reference sequence of '*Ca*. P.
182 luffae' LfWB (GenBank accession AF248956). Among the CDSs, 317 (44%) lacked any COG
183 functional category assignments. Among the CDSs that were assigned to specific functional
184 categories, those assigned to information storage and processing (e.g., replication,
185 transcription, and translation) represent the largest group that account for 32% of all CDSs.
186 In comparison, those assigned to cell process and metabolism account for 9% and 13%,
187 respectively. These observations are consistent with findings from characterization of other
188 phytoplasmas and more distantly related Mollicutes (e.g., *Spiroplasma*, *Entomoplasma*, and
189 *Mycoplasma*) (Chen et al., 2012; Kube et al., 2012; Lo et al., 2018). The observation that a
190 large fraction of genes lack specific functional annotation may be attributed to the elevated
191 evolutionary rates of Mollicutes and their distant phylogenetic relationships from model
192 organisms (Wu and Eisen, 2008; Wu et al., 2009; McCutcheon and Moran, 2012). The
193 observation that genes related to information storage and processing genes are relatively
194 abundant compared to those involved in metabolism is common among symbiotic bacteria
195 with small genomes (McCutcheon and Moran, 2012; Lo et al., 2016).

196 Other than the low G+C content and reduced gene content, phytoplasma genomes
197 are generally known to be repetitive, partly due to the presence of PMUs (Sugio and
198 Hogenhout, 2012). Interestingly, the genome of this strain is far more repetitive than other
199 phytoplasmas that have been characterized. On average, PMU-associated genes account for

200 ~4.7% (Std. Dev. = 2.7%) of the genome size among those 19 representative phytoplasma
201 genomes analyzed (**Table 1**). For strain NCHU2019, there are 117 PMU-associated genes
202 that are organized into at least 13 distinct PMU regions (**Figure 2**) and account for 11% of
203 the chromosome length. Additionally, a pair of chromosomal segments, each is ~75 kb in
204 size, were found to be duplicated (positions: 315,975-391,140 and 391,446-466,612). These
205 parts of the assembly were verified based on manual inspection of the ONT long reads
206 mapping result. Together, these two duplicated segments and the 13 PMUs account for 25%
207 of the chromosome length. The explanation for the high genome repetitiveness of strain
208 NCHU2019 compared to other phytoplasmas is unclear.
209
210 Comparisons with Closely Related Phytoplasmas
211 Based on the established classification scheme and a previous study of 16S rRNA
212 gene phylogeny (Davis et al., 2017), '*Ca*. P. luffae' belongs to group 16SrVIII and is most
213 closely related to '*Ca*. P. malaysianum' (group 16SrXXXII; GenBank accession EU371934)
214 (Nejat et al., 2013). These two species-level taxa have 95.9% sequence identity (i.e.,
215 1,463/1,526 aligned nucleotides) in their 16S rRNA genes. However, no genome sequence is
216 available for '*Ca*. P. malaysianum' for comparative analysis.
217 Other than '*Ca*. P. malaysianum', the next closest relatives of '*Ca*. P. luffae' include
218 phytoplasmas belonging to groups 16SrV ('*Ca*. P. vitis' and '*Ca*. P. ziziphi'), 16SrVI ('*Ca*. P.
219 sudamericanum' and '*Ca*. P. trifolii'), and 16SrVII ('*Ca*. P. fraxini'). Among these, one
220 complete genome sequence (GenBank accession CP025121) is available for '*Ca*. P. ziziphi'
221 strain Jwb-nky (Wang et al., 2018a), which represents the most closely related lineage for
222 comparative analysis (**Figure 3**). Comparison based on the 16S rRNA gene sequences
223 indicated that '*Ca*. P. luffae' NCHU2019 and '*Ca*. P. ziziphi' Jwb-nky have 94.9% sequence

224 identity (i.e., 1,447/1,524 aligned nucleotides). For whole-genome comparison, only 49% of
225 the chromosomal segments can be mapped between these two strains and these segments
226 have < 80% ANI. Pairwise genome alignment indicated that the most of the conserved
227 regions correspond to PMUs (**Figure 4**), which further supports that the sequence
228 divergence between these two genomes is too high for nucleotide-level comparisons. The
229 lack of chromosome-level synteny conservation was commonly reported in previous
230 comparisons between complete genome sequences of phytoplasmas (Bai et al., 2006; Tran-
231 Nguyen et al., 2008; Andersen et al., 2013; Orlovskis et al., 2017; Wang et al., 2018a), even
232 for '*Ca*. P. asteris' strains sharing > 99.9% 16S rRNA gene sequence identity and > 98.1%
233 genome-wide ANI (Cho et al., 2020a). These observations may be attributed to the strong
234 nucleotide composition biases, the high mutation accumulation rates, and the influence of
235 PMUs (Cho et al., 2020a).
236
237 Detailed Characterization of PMUs
238 To better understand the roles of PMUs in phytoplasma genome evolution, we
239 conducted detailed characterization of these mobile genetic elements. Among the 13 PMUs
240 identified in '*Ca*. P. luffae' NCHU2019 (**Figure 5A**), 11 are considered as complete ones and
241 range from 14 to 18 kb in size. The remaining two (i.e., #6 and #7) are both 12 kb in size and
242 appear to be truncated. These PMUs are interspersed across the entire chromosome and
243 there is no obvious pattern of clustering (**Figures 2 and 4**). Two sets of PMUs (i.e., #4-6 and
244 #7-9) are associated with the two 75-kb repeat regions. Notably, one set of junctions for
245 these large repeats (chromosomal positions 315,975 and 466,612) are located inside PMU
246 #4 (positions 308,032..323,983) and PMU #9 (positions 458,604..475,923). This finding
247 suggests that homologous recombination facilitated by PMUs may have facilitated the

248 segmental duplication of this chromosome, which in turn further increased the PMU copy
249 number.

250 Based on initial characterization of the PMUs in '*Ca*. P. asteris' AYWB genome, eight
251 core genes were identified (Bai et al., 2006). As more genome sequences become available
252 from more diverse phytoplasmas, we are able to include representatives of 13 '*Ca*. P.'
253 species from all three phylogenetic clusters in this analysis (**Table 1**). Compared to other
254 phytoplasmas, '*Ca*. P. luffae' NCHU2019 has the highest number of PMUs and is distinctive
255 in that all of its PMUs are similar (**Figure 5A**). Among the 11 complete PMUs in this genome,
256 the eight PMU core genes and *rad50* are all organized in the same order. The minor
257 variation in gene organization is mainly in between *rad50* and *tmk*, where genes encode
258 different putative secreted proteins and hypothetical proteins are found. For comparison, in
259 the closely related '*Ca*. P. ziziphi' Jwb-nky of cluster III, the four PMUs exhibit much higher
260 levels of diversity in gene organization between *hflB* and *tmk* (**Figure 5A**). For the distantly
261 related '*Ca*. P. asteris' AYWB and '*Ca*. P. australiense' PAa of cluster I, high levels of intra-
262 genomic PMU diversity are also observed (**Figure 5B**).

263 Based on the presence/absence pattern and order of eight PMU core genes defined
264 previously (Bai et al., 2006), we found that the PMUs in these representative phytoplasma
265 genomes can be classified into three major types (**Figure 5B**). Type A PMUs, in which *tmk* is
266 upstream of *dnaB*, are the most common ones that include PMUs found in phylogenetic
267 clusters I (e.g., '*Ca*. P. asteris' and '*Ca*. P. australiense') and III (e.g., '*Ca*. P. luffae', '*Ca*. P.
268 ziziphi', and '*Ca*. P. aurantifolia'). Type B PMUs are shorter, have *tmk* downstream of *dnaB*,
269 and are found only in '*Ca*. P. australiense' that belongs to cluster I. Type C is the rarest type,
270 with only one representative found in '*Ca*. P. mali' that belongs to cluster II, and has *hflB*
271 and *dnaG* located in between *tmk* and *dnaB*. In addition to the gene organization, molecular

272 phylogenies of *dnaB* and *tmk* also revealed divergence among homologs from different
273 PMU types (**Figure 6**). These patterns provide further support for our PMU classification
274 scheme. It is interesting to note that regardless of PMU copy numbers, most of the
275 phytoplasmas with genome sequences available harbor only one single type of PMUs. '*Ca*. P.
276 australiense' is the only exception that harbors both type A and B PMUs. Future
277 improvements in sampling more diverse phytoplasma genomes, particularly cluster II
278 lineages, are necessary to provide a more comprehensive understanding of PMU diversity.
279

280 PMU and Phytoplasma Genome Size Variation

281 One notable observation regarding phytoplasma genomes is the extensive size
282 variation at both between- and within-species levels (**Table 1**). Previous within-species
283 comparisons suggested that PMU abundance is an important factor (Bai et al., 2006;
284 Andersen et al., 2013). To further test if this pattern holds true for genus-level analysis, we
285 performed regression analysis to examine the correlation between the combined length of
286 PMU core genes and genome size. Strikingly, when all 19 representative phytoplasma
287 genome assemblies were examined together, PMU gene length explains 79% of the variance
288 in genome sizes (r = 0.89, p = 3.7e-07) (**Figure 7**). Due to the concern that draft assemblies
289 cannot provide accurate information regarding these two metrics, we also performed
290 regression analysis with only the 10 complete assemblies and obtained similar results (r =
291 0.87, p = 9.7e-04). Given that the evolution of bacterial genome sizes may be affected by
292 multiple factors (Konstantinidis and Tiedje, 2004; Ochman and Davalos, 2006; Kuo et al.,
293 2009; Novichkov et al., 2009; McCutcheon and Moran, 2012; Lo et al., 2016; Sabater-Muñoz
294 et al., 2017; Weinert and Welch, 2017), the strong correlation between PMU abundance
295 and phytoplasma genome size is surprising. Notably, as genome reduction appears to be a

296 common and recurring theme of symbiont evolution, the roles of PMU in genome expansion
297 of some phytoplasmas require further investigation. Because PMUs are known to be
298 associated with effector genes and likely can transfer horizontally between closely- (Cho et
299 al., 2019; Seruga Music et al., 2019) or distantly-related phytoplasmas (Chung et al., 2013;
300 Ku et al., 2013), the involvement of PMUs in phytoplasma effector gene content evolution is
301 particularly important.
302
303 Effector Genes
304 An important feature of phytoplasmas is their ability to modulate host plant
305 development through effectors, which are small secreted proteins (Sugio et al., 2011b;
306 Sugio and Hogenhout, 2012). To date, four phytoplasma effectors have been experimentally
307 characterized, including SAP05 (Gamboa et al., 2019; Huang and Hogenhout, 2019; Huang et
308 al., 2021), SAP11/SWP1 (Bai et al., 2009; Sugio et al., 2011a; Lu et al., 2014; Chang et al.,
309 2018; Wang et al., 2018c, 2018b), SAP54/PHYL1 (MacLean et al., 2011; Maejima et al., 2014;
310 Orlovskis and Hogenhout, 2016), and TENGU (Hoshi et al., 2009; Sugawara et al., 2013;
311 Minato et al., 2014). The expanded genome sequence availability allowed us to investigate
312 the phylogenetic distribution of homologous effector genes among diverse phytoplasmas.
313 The highly variable distribution patterns suggest that the effector gene content may have
314 rapid turnover during the diversification of phytoplasmas (**Figure 8**). For example, at within-
315 species level comparison among '*Ca*. P. asteris' strains, the patterns of presence/absence
316 and gene copy number are variable for three of these effectors as we reported previously
317 (Cho et al., 2020a). At genus-level, TENGU appears to be restricted to and conserved among
318 16SrI lineages in cluster I (i.e., '*Ca*. P. asteris' and '*Ca*. P. tritici'), while the other three
319 effectors are variable. Based on these patterns, it is likely that TENGU originated in the most

320 recent common ancestor (MRCA) of 16SrI phytoplasmas. However, for the other three
321 effectors, it is unclear if the MRCA of all extant phytoplasmas harbor these genes or not. If
322 yes, then multiple independent gene losses are required to explain the distribution of these
323 genes among extant phytoplasmas. Alternatively, multiple independent origins, likely
324 mediated by PMU-mediated horizontal gene transfer (Chung et al., 2013; Cho et al., 2019;
325 Seruga Music et al., 2019), are required to explain the gene distribution pattern. Another
326 interesting observation is that several phytoplasmas do not possess any of these four
327 effector genes. It is likely that these phytoplasmas harbor novel effector genes that are yet
328 to be characterized and further investigation is required to obtain a more complete picture
329 of phytoplasma effector diversity.

330 For closer inspection of these four known effectors, we performed multiple
331 sequence alignments to examine the protein sequence divergence among homologs (**Figure**
332 **9**). Consistent with the expectation derived from their phylogenetic distribution patterns,
333 the three effector genes found among diverse phytoplasmas have higher levels of sequence
334 divergence compared to phylogenetically restricted TENGU homologs.

335 For copy number variation, '*Ca*. P. luffae' NCHU2019 stands out as having the highest
336 copy numbers for SAP11 and SAP54 (**Figure 8**). These homologous genes are all located
337 within PMUs (**Figure 5**) and have nearly identical sequences (**Figure 9**), which suggest that
338 recent intra-genomic PMU duplications are responsible for expansions in effector gene copy
339 numbers. Similar patterns were observed for SAP11 and SAP54 homologs in '*Ca*. P. ziziphi'
340 Jwb-nky. Experimental characterization of the two SAP11 homologs in '*Ca*. P. ziziphi'
341 demonstrated that both can stimulate lateral bud outgrowth for witches' broom symptoms
342 when expressed in *Nicotiana bentha*miana (Zhou et al., 2021). Intriguingly, in phytoplasma-
343 infected jujube plants, these two SAP11 homologs have different expression patterns

344 among tissue types. These findings suggest that such gene duplication events may lead to
345 neofunctionalization or subfunctionalization, thus promoting the genetic diversity of
346 effectors.
347
348 **Conclusion**
349 In this work, we determined the complete genome sequence of an uncultivated '*Ca*.
350 P. luffae' strain associated with the witches' broom disease of loofah. This assembly
351 provides the first representative genome sequence for the 16SrVIII group of phytoplasmas
352 and improves the taxon sampling of these diverse plant-pathogenic bacteria. For
353 comparative genomics analysis conducted at genus level, we provided a global view of the
354 PMUs (i.e., phytoplasma-specific mobile genetic elements) and identified three major PMU
355 types that differ in gene organization and phylogenetic distribution. Importantly, statistical
356 analysis revealed that PMU abundance explains nearly 80% of the variance in phytoplasma
357 genome sizes, thus providing a quantitative estimate on the importance of these elements.
358 Finally, our investigation of effector genes highlighted the genetic diversity associated with
359 phytoplasma virulence and established the roles of PMUs in shaping such diversity.
360
361 **Data Availability**
362 This genome sequencing project was deposited in the NCBI under BioProject
363 PRJNA636624. The raw reads were deposited in the Sequence Read Archive (SRA) under the
364 accessions SRR11921288 (Illumina MiSeq paired-end reads) and SRR11921289 (ONT MinION
365 reads). The complete and annotated genome sequence of '*Ca*. P. luffae' NCHU2019 was
366 deposited in GenBank under the accession CP054393.
367

Author Contribution

JYY and CHK conceptualized the study, acquired funding, and supervised the project. JYY, YCC, and CMT provided the biological materials. CMT and STC coordinated the sequencing. CMT conducted the initial genome assembly. STC completed the assembly. STC and CTH conducted the comparative analysis and prepared the figures. CTH and CHK wrote the draft. All authors approved the submitted version.

Funding

Work in the Yang lab was supported by grants-in-aid from the Ministry of Science and Technology (110-2628-B-005-002) and the Advanced Plant Biotechnology Center from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan. Work in the Kuo lab was supported by Academia Sinica and the Ministry of Science and Technology (106-2311-B-001-028-MY3) of Taiwan. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Acknowledgments

We thank Ai-Ping Chen and Shu-Jen Chou for technical assistance. The Illumina sequencing library preparation service was provided by the Genomic Technology Core Facility (Institute of Plant and Microbial Biology, Academia Sinica, Taipei. Taiwan). The Illumina paired-end sequencing service was provided by the Genomics Core Facility (Institute of Molecular Biology, Academia Sinica, Taipei. Taiwan).

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.

397 **References**

398 Andersen, M. T., Liefting, L. W., Havukkala, I., and Beever, R. E. (2013). Comparison of the
399 complete genome sequence of two closely related isolates of '*Candidatus*
400 Phytoplasma australiense' reveals genome plasticity. *BMC Genomics* 14, 529.
401 doi:10.1186/1471-2164-14-529.

402 Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S.,
403 et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural
404 networks. *Nat. Biotechnol.* 37, 420–423. doi:10.1038/s41587-019-0036-z.

405 Bai, X., Correa, V. R., Toruño, T. Y., Ammar, E.-D., Kamoun, S., and Hogenhout, S. A. (2009).
406 AY-WB phytoplasma secretes a protein that targets plant cell nuclei. *Mol. Plant.*
407 *Microbe Interact.* 22, 18–30. doi:10.1094/MPMI-22-1-0018.

408 Bai, X., Zhang, J., Ewing, A., Miller, S. A., Jancso Radek, A., Shevchenko, D. V., et al. (2006).
409 Living with genome instability: the adaptation of phytoplasmas to diverse
410 environments of their insect and plant hosts. *J. Bacteriol.* 188, 3682–3696.
411 doi:10.1128/JB.188.10.3682-3696.2006.

412 Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., et al.
413 (2018). GenBank. *Nucleic Acids Res.* 46, D41–D47. doi:10.1093/nar/gkx1094.

414 Bertaccini, A., Duduk, B., Paltrinieri, S., and Contaldo, N. (2014). Phytoplasmas and
415 phytoplasma diseases: a severe threat to agriculture. *Am. J. Plant Sci.* 05, 1763–1788.
416 doi:10.4236/ajps.2014.512191.

417 Bertaccini, A., and Lee, I.-M. (2018). "Phytoplasmas: An update," in *Phytoplasmas: Plant*
418 *Pathogenic Bacteria - I: Characterisation and Epidemiology of Phytoplasma -*
419 *Associated Diseases*, eds. G. P. Rao, A. Bertaccini, N. Fiore, and L. W. Liefting
420 (Singapore: Springer Singapore), 1–29. doi:10.1007/978-981-13-0119-3_1.

421 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009).
422 BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.
423 doi:10.1186/1471-2105-10-421.

424 Chang, S. H., Tan, C. M., Wu, C.-T., Lin, T.-H., Jiang, S.-Y., Liu, R.-C., et al. (2018). Alterations
425 of plant architecture and phase transition by the phytoplasma virulence factor SAP11.
426 *J. Exp. Bot.* 69, 5389–5401. doi:10.1093/jxb/ery318.

427 Chen, L.-L., Chung, W.-C., Lin, C.-P., and Kuo, C.-H. (2012). Comparative analysis of gene
428 content evolution in phytoplasmas and mycoplasmas. *PLOS ONE* 7, e34407.
429 doi:10.1371/journal.pone.0034407.

430 Cho, S.-T., Kung, H.-J., Huang, W., Hogenhout, S. A., and Kuo, C.-H. (2020a). Species
431 boundaries and molecular markers for the classification of 16SrI phytoplasmas
432 inferred by genome analysis. *Front. Microbiol.* 11, 1531.
433 doi:10.3389/fmicb.2020.01531.

434 Cho, S.-T., Lin, C.-P., and Kuo, C.-H. (2019). Genomic characterization of the periwinkle leaf
435 yellowing (PLY) phytoplasmas in Taiwan. *Front. Microbiol.* 10, 2194.
436 doi:10.3389/fmicb.2019.02194.

437 Cho, S.-T., Zwolińska, A., Huang, W., Wouters, R. H. M., Mugford, S. T., Hogenhout, S. A., et
438 al. (2020b). Complete genome sequence of "*Candidatus* Phytoplasma asteris" RP166,
439 a plant pathogen associated with rapeseed phyllody disease in Poland. *Microbiol.*
440 *Resour. Announc.* 9. doi:10.1128/MRA.00760-20.

441 Chung, W.-C., Chen, L.-L., Lo, W.-S., Lin, C.-P., and Kuo, C.-H. (2013). Comparative analysis of
442 the peanut witches'-broom phytoplasma genome reveals horizontal transfer of
443 potential mobile units and effectors. *PLOS ONE* 8, e62770.
444 doi:10.1371/journal.pone.0062770.

445 Davis, R. E., Zhao, Y., Wei, W., Dally, E. L., and Lee, I.-M. (2017). '*Candidatus* Phytoplasma
446 luffae', a novel taxon associated with witches' broom disease of loofah, Luffa
447 aegyptica Mill. *Int. J. Syst. Evol. Microbiol.* 67, 3127–3133.
448 doi:10.1099/ijsem.0.001980.

449 Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high
450 throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340.

451 Felsenstein, J. (1989). PHYLIP - phylogeny inference package (version 3.2). *Cladistics* 5, 164–
452 166.

453 Gamboa, C., Cui, W., Quiroga, N., Fernández, C., Fiore, N., and Zamorano, A. (2019).
454 Identification of 16SrIII-J phytoplasma effectors using a viral vector. *Phytopathogenic*
455 *Mollicutes* 9, 229. doi:10.5958/2249-4677.2019.00115.4.

456 Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large
457 phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704.
458 doi:10.1080/10635150390235520.

459 Gundersen, D. E., and Lee, I.-M. (1996). Ultrasensitive detection of phytoplasmas by nested-
460 PCR assays using two universal primer pairs. *Phytopathol. Mediterr.* 35, 144–151.

461 Guy, L., Roat Kultima, J., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and
462 genome visualization in R. *Bioinformatics* 26, 2334–2335.
463 doi:10.1093/bioinformatics/btq413.

464 Hogenhout, S. A., Oshima, K., Ammar, E.-D., Kakizawa, S., Kingdom, H. N., and Namba, S.
465 (2008). Phytoplasmas: bacteria that manipulate plants and insects. *Mol. Plant Pathol.*
466 9, 403–423. doi:10.1111/j.1364-3703.2008.00472.x.

467 Hogenhout, S. A., and Seruga Music, M. (2009). "Phytoplasma genomics, from sequencing to
468 comparative and functional genomics - what have we learnt?," in *Phytoplasmas:*
469 *genomes, plant hosts and vectors* (CABI), 19–36.

470 Hoshi, A., Oshima, K., Kakizawa, S., Ishii, Y., Ozeki, J., Hashimoto, M., et al. (2009). A unique
471 virulence factor for proliferation and dwarfism in plants identified from a
472 phytopathogenic bacterium. *Proc. Natl. Acad. Sci.* 106, 6416–6421.
473 doi:10.1073/pnas.0813038106.

474 Huang, W., and Hogenhout, S. A. (2019). Phytoplasma effectors have converged onto
475 degrading plant transcription factors with fundamental roles in plant development
476 and defense to insects. in *IS-MPMI XVIII Congress Abstracts of Poster Presentations*
477 (Glasgow, UK: IS-MPMI), 158. doi:https://doi.org/10.1094/MPMI-32-10-S1.1.

478 Huang, W., MacLean, A. M., Sugio, A., Maqbool, A., Busscher, M., Cho, S.-T., et al. (2021).
479 Parasite co-opts a ubiquitin receptor to induce a plethora of developmental changes.
480 *bioRxiv*, 2021.02.15.430920. doi:10.1101/2021.02.15.430920.

481 Hyatt, D., Chen, G.-L., LoCascio, P., Land, M., Larimer, F., and Hauser, L. (2010). Prodigal:
482 prokaryotic gene recognition and translation initiation site identification. *BMC*
483 *Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119.

484 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High
485 throughput ANI analysis of 90K prokaryotic genomes reveals clear species
486 boundaries. *Nat. Commun.* 9, 5114. doi:10.1038/s41467-018-07641-9.

487 Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for
488 representation and analysis of molecular networks involving diseases and drugs.
489 *Nucleic Acids Res.* 38, D355–D360. doi:10.1093/nar/gkp896.

490 Kirdat, K., Tiwarekar, B., Thorat, V., Sathe, S., Shouche, Y., and Yadav, A. (2021). "*Candidatus*
491 Phytoplasma sacchari", a novel taxon - associated with Sugarcane Grassy Shoot
492 (SCGS) disease. *Int. J. Syst. Evol. Microbiol.* 71, 004591. doi:10.1099/ijsem.0.004591.

493 Konstantinidis, K. T., and Tiedje, J. M. (2004). Trends between gene content and genome
494 size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci.* 101, 3160–
495 3165. doi:10.1073/pnas.0308653100.

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

499 Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., et al. (2009).
500 Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–
501 1645. doi:10.1101/gr.092759.109.

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

504 Kube, M., Mitrovic, J., Duduk, B., Rabus, R., and Seemüller, E. (2012). Current view on
505 phytoplasma genomes and encoded metabolism. *Sci. World J.* 2012, 1–25.
506 doi:10.1100/2012/185942.

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

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

513 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004).
514 Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12.
515 doi:10.1186/gb-2004-5-2-r12.

516 Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H.-H., Rognes, T., and Ussery, D. W. (2007).
517 RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids*
518 *Res.* 35, 3100–3108. doi:10.1093/nar/gkm160.

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

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

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

528 Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34,
529 3094–3100. doi:10.1093/bioinformatics/bty191.

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

532 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The
533 Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
534 doi:10.1093/bioinformatics/btp352.

535 Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for
536 eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503.

537 Lo, W.-S., Gasparich, G. E., and Kuo, C.-H. (2018). Convergent evolution among ruminant-
538 pathogenic *Mycoplasma* involved extensive gene content changes. *Genome Biol.*
539 *Evol.* 10, 2130–2139. doi:10.1093/gbe/evy172.

540 Lo, W.-S., Huang, Y.-Y., and Kuo, C.-H. (2016). Winding paths to simplicity: genome evolution
541 in facultative insect symbionts. *FEMS Microbiol. Rev.* 40, 855–874.
542 doi:10.1093/femsre/fuw028.

543 Lowe, T., and Eddy, S. (1997). tRNAscan-SE: a program for improved detection of transfer
544 RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964.
545 doi:10.1093/nar/25.5.955.

546 Lu, Y.-T., Li, M.-Y., Cheng, K.-T., Tan, C. M., Su, L.-W., Lin, W.-Y., et al. (2014). Transgenic
547 plants that express the phytoplasma effector SAP11 show altered phosphate
548 starvation and defense responses. *Plant Physiol.* 164, 1456–1469.
549 doi:10.1104/pp.113.229740.

550 MacLean, A. M., Sugio, A., Makarova, O. V., Findlay, K. C., Grieve, V. M., Tóth, R., et al.
551 (2011). Phytoplasma effector SAP54 induces indeterminate leaf-like flower
552 development in Arabidopsis plants. *Plant Physiol.* 157, 831–841.
553 doi:10.1104/pp.111.181586.

554 Maejima, K., Iwai, R., Himeno, M., Komatsu, K., Kitazawa, Y., Fujita, N., et al. (2014).
555 Recognition of floral homeotic MADS domain transcription factors by a phytoplasmal
556 effector, phyllogen, induces phyllody. *Plant J.* 78, 541–554. doi:10.1111/tpj.12495.

557 McCutcheon, J. P., and Moran, N. A. (2012). Extreme genome reduction in symbiotic
558 bacteria. *Nat. Rev. Microbiol.* 10, 13–26. doi:10.1038/nrmicro2670.

559 Minato, N., Himeno, M., Hoshi, A., Maejima, K., Komatsu, K., Takebayashi, Y., et al. (2014).
560 The phytoplasmal virulence factor TENGU causes plant sterility by downregulating of
561 the jasmonic acid and auxin pathways. *Sci. Rep.* 4, 7399. doi:10.1038/srep07399.

562 Namba, S. (2019). Molecular and biological properties of phytoplasmas. *Proc. Jpn. Acad. Ser.*
563 *B* 95, 401–418. doi:10.2183/pjab.95.028.

564 Nejat, N., Vadamalai, G., Davis, R. E., Harrison, N. A., Sijam, K., Dickinson, M., et al. (2013).
565 '*Candidatus* Phytoplasma malaysianum', a novel taxon associated with virescence
566 and phyllody of Madagascar periwinkle (*Catharanthus roseus*). *Int. J. Syst. Evol.*
567 *Microbiol.* 63, 540–548. doi:10.1099/ijs.0.041467-0.

568 Novichkov, P. S., Wolf, Y. I., Dubchak, I., and Koonin, E. V. (2009). Trends in prokaryotic
569 evolution revealed by comparison of closely related bacterial and archaeal genomes.
570 *J. Bacteriol.* 191, 65–73. doi:10.1128/JB.01237-08.

571 Ochman, H., and Davalos, L. M. (2006). The nature and dynamics of bacterial genomes.
572 *Science* 311, 1730–1733. doi:10.1126/science.1119966.

573 Orlovskis, Z., Canale, M. C., Haryono, M., Lopes, J. R. S., Kuo, C.-H., and Hogenhout, S. A.
574 (2017). A few sequence polymorphisms among isolates of maize bushy stunt
575 phytoplasma associate with organ proliferation symptoms of infected maize plants.
576 *Ann. Bot.* 119, 869–884. doi:10.1093/aob/mcw213.

577 Orlovskis, Z., and Hogenhout, S. A. (2016). A bacterial parasite effector mediates insect
578 vector attraction in host plants independently of developmental changes. *Front.*
579 *Plant Sci.* 7, 885. doi:10.3389/fpls.2016.00885.

580 Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.-Y., Wei, W., Suzuki, S., et al. (2004).
581 Reductive evolution suggested from the complete genome sequence of a plant-
582 pathogenic phytoplasma. *Nat. Genet.* 36, 27–29. doi:10.1038/ng1277.

583 Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al.
584 (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26.
585 doi:10.1038/nbt.1754.

586 Sabater-Muñoz, B., Toft, C., Alvarez-Ponce, D., and Fares, M. A. (2017). Chance and
587 necessity in the genome evolution of endosymbiotic bacteria of insects. *ISME J.* 11,
588 1291–1304. doi:10.1038/ismej.2017.18.

589 Saccardo, F., Martini, M., Palmano, S., Ermacora, P., Scortichini, M., Loi, N., et al. (2012).
590 Genome drafts of four phytoplasma strains of the ribosomal group 16SrIII.
591 *Microbiology* 158, 2805–2814. doi:10.1099/mic.0.061432-0.

592 Seruga Music, M., Samarzija, I., Hogenhout, S. A., Haryono, M., Cho, S.-T., and Kuo, C.-H.
593 (2019). The genome of '*Candidatus* Phytoplasma solani' strain SA-1 is highly dynamic
594 and prone to adopting foreign sequences. *Syst. Appl. Microbiol.* 42, 117–127.
595 doi:10.1016/j.syapm.2018.10.008.

596 Sugawara, K., Honma, Y., Komatsu, K., Himeno, M., Oshima, K., and Namba, S. (2013). The
597 alteration of plant morphology by small peptides released from the proteolytic
598 processing of the bacterial peptide TENGU. *Plant Physiol.* 162, 2005–2014.
599 doi:10.1104/pp.113.218586.

600 Sugio, A., and Hogenhout, S. A. (2012). The genome biology of phytoplasma: modulators of
601 plants and insects. *Curr. Opin. Microbiol.* 15, 247–254.
602 doi:10.1016/j.mib.2012.04.002.

603 Sugio, A., Kingdom, H. N., MacLean, A. M., Grieve, V. M., and Hogenhout, S. A. (2011a).
604 Phytoplasma protein effector SAP11 enhances insect vector reproduction by
605 manipulating plant development and defense hormone biosynthesis. *Proc. Natl.*
606 *Acad. Sci.* 108, E1254–E1263. doi:10.1073/pnas.1105664108.

607 Sugio, A., MacLean, A. M., Kingdom, H. N., Grieve, V. M., Manimekalai, R., and Hogenhout, S.
608 A. (2011b). Diverse targets of phytoplasma effectors: from plant development to
609 defense against insects. *Annu. Rev. Phytopathol.* 49, 175–195. doi:10.1146/annurev-
610 phyto-072910-095323.

611 Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., et al. (2003). The
612 COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.
613 doi:10.1186/1471-2105-4-41.

614 The IRPCM Phytoplasma/Spiroplasma Working Team - Phytoplasma taxonomy group (2004).
615 "*Candidatus* Phytoplasma", a taxon for the wall-less, non-helical prokaryotes that
616 colonize plant phloem and insects. *Int. J. Syst. Evol. Microbiol.* 54, 1243–1255.
617 doi:10.1099/ijs.0.02854-0.

618 Tran-Nguyen, L. T. T., Kube, M., Schneider, B., Reinhardt, R., and Gibb, K. S. (2008).
619 Comparative genome analysis of "*Candidatus* Phytoplasma australiense" (subgroup
620 *tuf*-Australia I; *rp*-A) and "*Ca*. Phytoplasma asteris" strains OY-M and AY-WB. *J.*
621 *Bacteriol.* 190, 3979–3991. doi:10.1128/JB.01301-07.

622 Wang, J., Song, L., Jiao, Q., Yang, S., Gao, R., Lu, X., et al. (2018a). Comparative genome
623 analysis of jujube witches'-broom phytoplasma, an obligate pathogen that causes
624 jujube witches'-broom disease. *BMC Genomics* 19, 689. doi:10.1186/s12864-018-
625 5075-1.

626 Wang, N., Li, Y., Chen, W., Yang, H. Z., Zhang, P. H., and Wu, Y. F. (2018b). Identification of
627 wheat blue dwarf phytoplasma effectors targeting plant proliferation and defence
628 responses. *Plant Pathol.* 67, 603–609. doi:10.1111/ppa.12786.

629 Wang, N., Yang, H., Yin, Z., Liu, W., Sun, L., and Wu, Y. (2018c). Phytoplasma effector SWP1
630 induces witches' broom symptom by destabilizing the TCP transcription factor
631 BRANCHED1. *Mol. Plant Pathol.* 19, 2623–2634. doi:10.1111/mpp.12733.

632 Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009).
633 Jalview version 2—a multiple sequence alignment editor and analysis workbench.
634 *Bioinformatics* 25, 1189–1191. doi:10.1093/bioinformatics/btp033.

635 Weinert, L. A., and Welch, J. J. (2017). Why might bacterial pathogens have small genomes?
636 *Trends Ecol. Evol.* 32, 936–947. doi:10.1016/j.tree.2017.09.006.

637 Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A
638 phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462,
639 1056–1060. doi:10.1038/nature08656.

640 Wu, M., and Eisen, J. (2008). A simple, fast, and accurate method of phylogenomic inference.
641 *Genome Biol.* 9, R151. doi:10.1186/gb-2008-9-10-r151.

642 Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly
643 using de Bruijn graphs. *Genome Res.* 18, 821–829. doi:10.1101/gr.074492.107.

644 Zhao, Y., and Davis, R. E. (2016). Criteria for phytoplasma 16Sr group/subgroup delineation
645 and the need of a platform for proper registration of new groups and subgroups. *Int.*
646 *J. Syst. Evol. Microbiol.* 66, 2121–2123. doi:10.1099/ijsem.0.000999.

647 Zhao, Y., Wei, W., Davis, R. E., Lee, I.-M., and Bottner-Parker, K. D. (2021). The agent
648 associated with blue dwarf disease in wheat represents a new phytoplasma taxon,
649 '*Candidatus* Phytoplasma tritici.' *Int. J. Syst. Evol. Microbiol.*
650 doi:10.1099/ijsem.0.004604.

651 Zhao, Y., Wei, W., Lee, I.-M., Shao, J., Suo, X., and Davis, R. E. (2009). Construction of an
652 interactive online phytoplasma classification tool, *i*PhyClassifier, and its application
653 in analysis of the peach X-disease phytoplasma group (16SrIII). *Int. J. Syst. Evol.*
654 *Microbiol.* 59, 2582–2593. doi:10.1099/ijs.0.010249-0.

Zhou, J., Ma, F., Yao, Y., Deng, M., Chen, M., Zhang, S., et al. (2021). JWB phytoplasma effectors SJP1 and SJP2 induce lateral bud outgrowth by repressing the ZjBRC1-controlled auxin efflux channel. *Plant Cell Environ*. n/a. doi:10.1111/pce.14141.

662 **Table**

663 **Table 1.** List of the genome assemblies analyzed. For each strain, information regarding the
664 16S rRNA gene (16Sr) group, genome accession number, assembly status, genome size,
665 coding sequence (CDS) count, potential mobile unit (PMU) gene count, and combined length
666 of PMU genes are provided. The PMU gene information is based on homologs of eight core
667 genes (i.e., *tra5*, *dnaB*, *dnaG*, *tmk*, *hflB*, *himA*, *ssb*, and *rpoD*); other genes such as those
668 encode hypothetical proteins or putative secreted proteins are excluded. *Acholeplasma*
669 *laidlawii* is included as an outgroup.

670

671 **Figure Legends**

672 **Figure 1.** Infection symptoms. (A) The loofah plant exhibiting phytoplasma infection
673 symptoms that was collected in Changhua, Taiwan. (B) Loofah plants grown in the lab. Left,
674 healthy control; right, grafted with the infected plant shown in panel A.

675

676 **Figure 2.** Genome map of '*Candidatus* Phytoplasma luffae' NCHU2019. Rings from outside in:
677 (1) Scale marks (kb). (2 and 3) Coding sequences on the forward and reverse strand,
678 respectively. Color-coded by functional categories. (4) Genes associated with potential
679 mobile units (PMUs), color-coded by annotation (see **Figure 5**). Gene clusters that represent
680 individual PMUs are labeled by orange lines. Genes encode putative secreted proteins are in
681 red. (5) GC skew (positive: dark gray; negative: light gray); (6) GC content (above average:
682 dark gray; below average; light gray). One high GC peak located in the 531 to 544 kb region
683 corresponds to two adjacent rRNA operons.

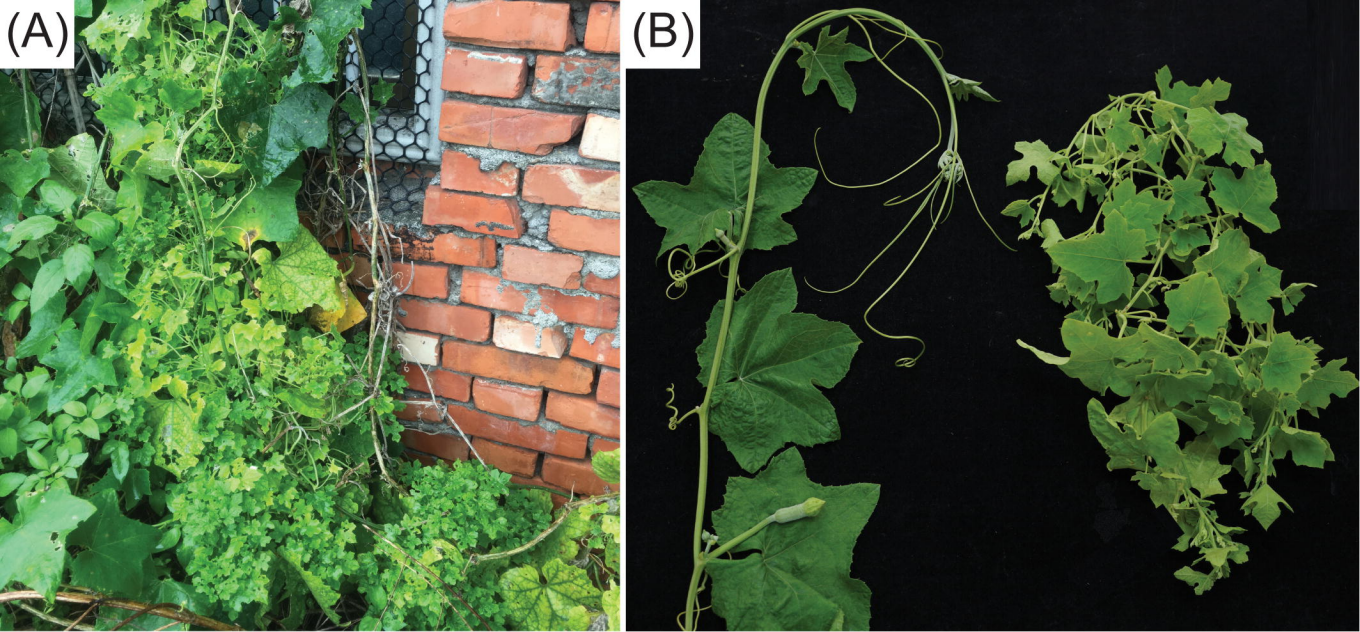
684

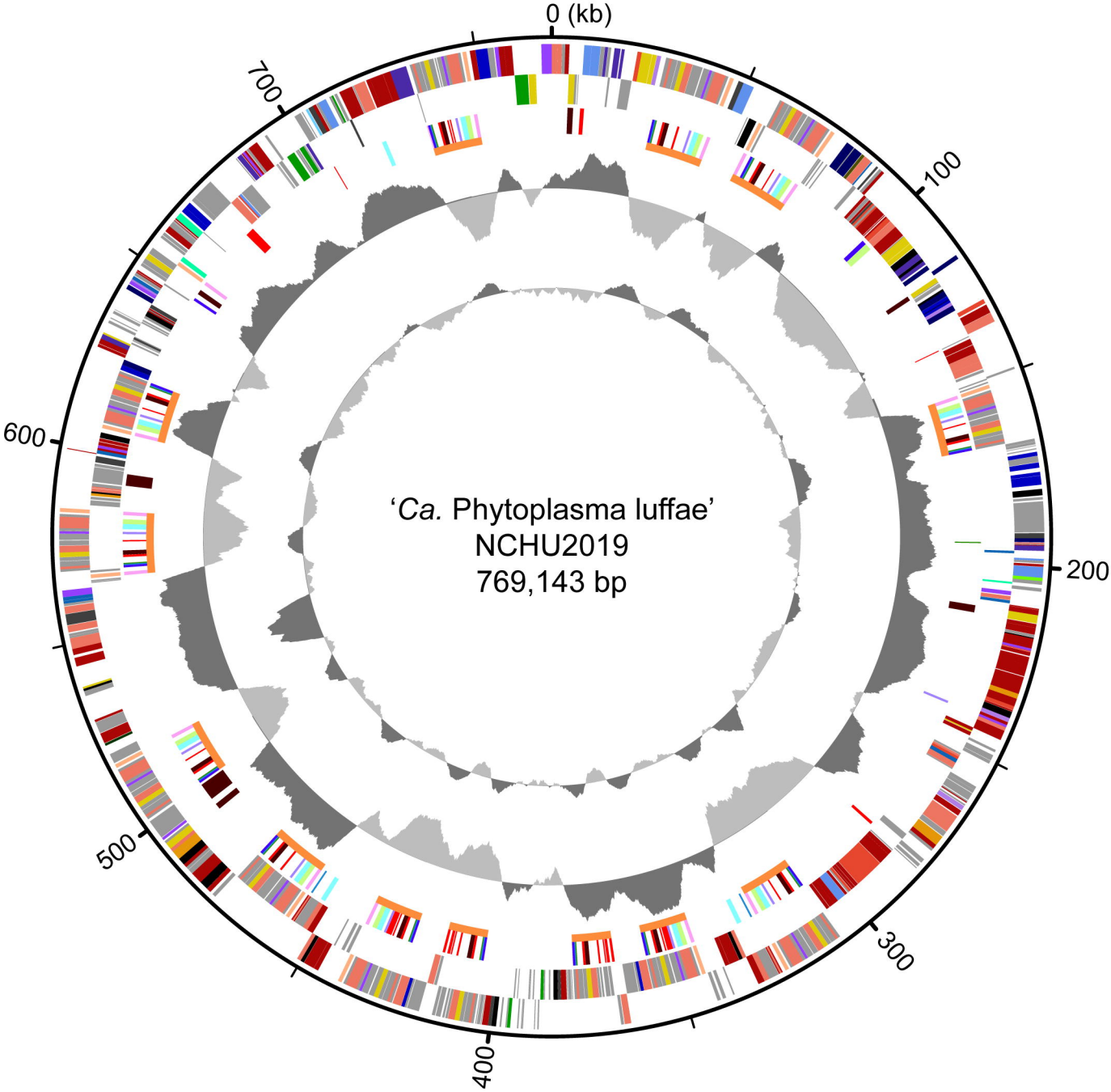
685 **Figure 3.** Evolutionary relationships among representative phytoplasmas with genome
686 sequences available. The maximum likelihood phylogeny was inferred using a concatenated
687 alignment of 80 conserved single-copy genes (25,340 aligned amino acid sites). All internal
688 nodes received > 80% bootstrap support based on 1,000 re-sampling. Strains with complete
689 genome assemblies available are highlighted in bold. The number in parentheses following
690 the strain name indicates the genome size (in kb). The 16S rRNA gene (16Sr) group
691 assignments are labeled on the right. The three major phylogenetic clusters are indicated by
692 colored backgrounds (I: blue; II; green; III: yellow). *Acholeplasma laidlawii* is included as the
693 outgroup.
694
695 **Figure 4.** Pairwise genome alignments. The genome of '*Ca*. P. luffae' NCHU2019 was used as
696 the reference for comparison with (A) itself and (B) '*Ca*. P. ziziphi' Jwb-nky. Matches on the
697 same strand and the opposite strand are indicated in red and blue, respectively. Potential
698 mobile units (PMUs) are illustrated as orange arrows on both axes.
699
700 **Figure 5.** Gene organization of potential mobile units (PMUs). Each individual PMU is
701 labeled by the strain name and a numerical identifier. Background colors for the PMU
702 identifiers indicate the three phylogenetic clusters of phytoplasmas (I: blue; II; green; III:
703 yellow). Genes are drawn to scale and color-coded according to annotation. Homologs of
704 putative effectors identified in the '*Ca*. P. asteris' AYWB genomes are labeled, those with
705 experimental evidence (i.e., SAP05, SAP11, and SAP54) are highlighted in bold. (A) All PMUs
706 in '*Ca*. P. luffae' NCHU2019 and '*Ca*. P. ziziphi' Jwb-nky. Grouped by genomes. (B)
707 Representative PMUs from selected phytoplasmas. Grouped by PMU types.
708

709 **Figure 6.** Maximum likelihood phylogenies of PMU core genes (A) *tmk* (228 aligned amino
710 acid sites). (B) *dnaB* (531 aligned amino acid sites). In both panels, a non-PMU homolog is
711 included as the outgroup. Numbers next to internal branches indicate the bootstrap support
712 levels based on 1,000 re-sampling. Background colors for the gene identifiers indicate the
713 three phylogenetic clusters of phytoplasmas (I: blue; II; green; III: yellow).
714
715 **Figure 7.** Correlation between the combined length of PMU core genes and genome size.
716 Strains with complete and draft genome assemblies are indicated by filled and open circles,
717 respectively. The linear regression line was based on all available assemblies.
718
719 **Figure 8.** Phylogenetic distribution of known effector genes. Gene presence and absence are
720 indicated by filled (with copy numbers) and empty circles, respectively. The cladogram on
721 the left is based on **Figure 3**. Strains with complete genome assemblies available, which
722 have a higher confidence level for inferring gene presence and absence, are highlighted in
723 bold.
724
725 **Figure 9.** Protein sequence alignments of four characterized phytoplasma effectors. Shaded
726 background colors indicate the levels of sequence conservation. (A) SAP05. (B) SAP11/SWP1.
727 (C) SAP54/PHYL1. (D) TENGU.
728

Table 1. List of the genome assemblies analyzed. For each strain, information regarding the 16S rRNA gene (16Sr) group, genome accession number, assembly status, genome size, coding sequence (CDS) count, potential mobile unit (PMU) gene count, and combined length of PMU genes are provided. The PMU gene information is based on homologs of eight core genes (i.e., *tra5*, *dnaB*, *dnaG*, *tmk*, *hflB*, *himA*, *ssb*, and *rpoD*); other genes such as those encode hypothetical proteins or putative secreted proteins are excluded. *Acholeplasma laidlawii* is included as an outgroup.

Strain	16Sr Group	Accession	Assembly	Genome Size (bp)	CDS Count	PMU Gene Count	PMU Gene Length (bp)
'*Ca*. P. asteris' AYWB	I	GCF_000012225.1	Complete	706,569	514	112	49,188
'*Ca*. P. asteris' De Villa	I	GCF_004214875.1	Complete	600,116	521	21	11,958
'*Ca*. P. asteris' DY2014	I	GCA_005093185.1	8 contigs	824,596	775	89	48,761
'*Ca*. P. asteris' LD1	I	GCF_001866375.1	8 contigs	599,264	513	29	15,441
'*Ca*. P. asteris' M3	I	GCF_001712875.1	Complete	576,118	490	27	14,541
'*Ca*. P. asteris' OY-M	I	GCF_000009845.1	Complete	853,092	708	137	72,399
'*Ca*. P. tritici' WBD	I	GCF_000495255.1	6 contigs	611,462	471	45	19,549
'*Ca*. P. aurantifolia' NCHU2014	II	GCA_001307505.2	Complete	635,584	471	23	16,378
'*Ca*. P. pruni' CX	III	GCF_001277135.1	46 contigs	598,511	434	31	22,474
'*Ca*. P. ziziphi' Jwb-nky	V	GCF_003640545.1	Complete	750,803	641	69	49,824
'*Ca*. P. luffae' NCHU2019	VIII	GCA_018024475.1	Complete	769,143	725	117	84,321
'*Ca*. P. mali' AT	X	GCF_000026205.1	Complete	601,943	495	20	15,084
'*Ca*. P. oryzae' Mbita	XI	GCF_001578535.1	28 contigs	533,195	432	17	14,259
'*Ca*. P. sacchari' SCGS	XI	GCF_009268105.1	28 contigs	502,197	402	15	13,620
'*Ca*. P. australiense' NZSb11	XII	GCF_000397185.1	Complete	959,779	828	131	78,642
'*Ca*. P. australiense' PAa	XII	GCF_000069925.1	Complete	879,959	699	108	55,565
'*Ca*. P. solani' SA-1	XII	GCA_003698095.1	19 contigs	821,322	709	66	45,234
'*Ca*. P. cynodontis' LW01	XIV	GCF_009268075.1	23 contigs	483,935	425	13	11,856
'*Ca*. P. pini' MDPP	XXI	GCF_007821455.1	16 contigs	474,136	392	15	11,685
Acholeplasma laidlawii PG-8A	N/A	GCF_000018785.1	Complete	1,496,992	1,378	0	0


(A)
(B)



INFORMATION

- [J] Translation
- [K] Transcription
- [L] Replication

POORLY CHARACTERIZED

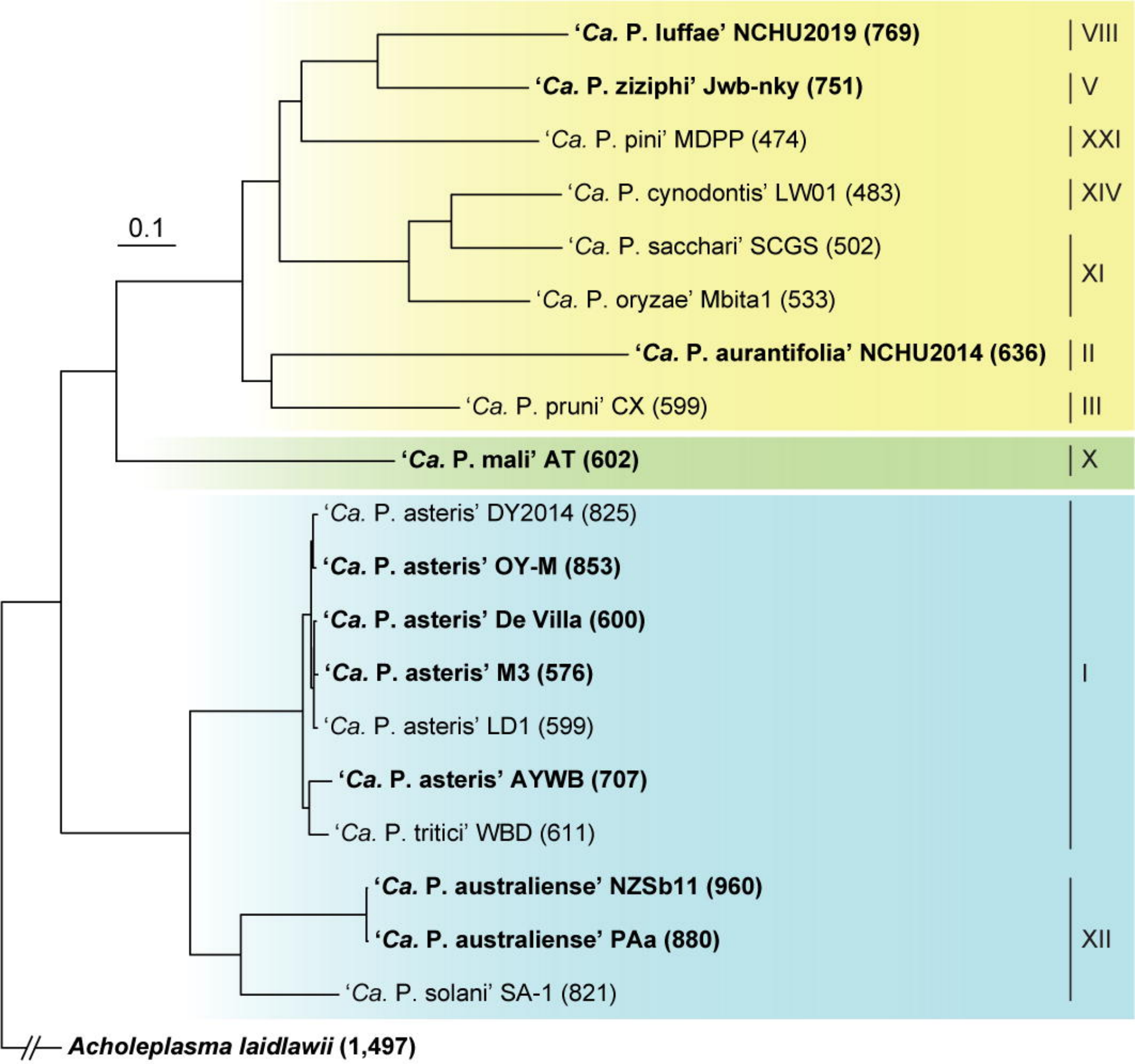
- [R] General prediction only
- [S] Function unknown
- [?] No COG assignment

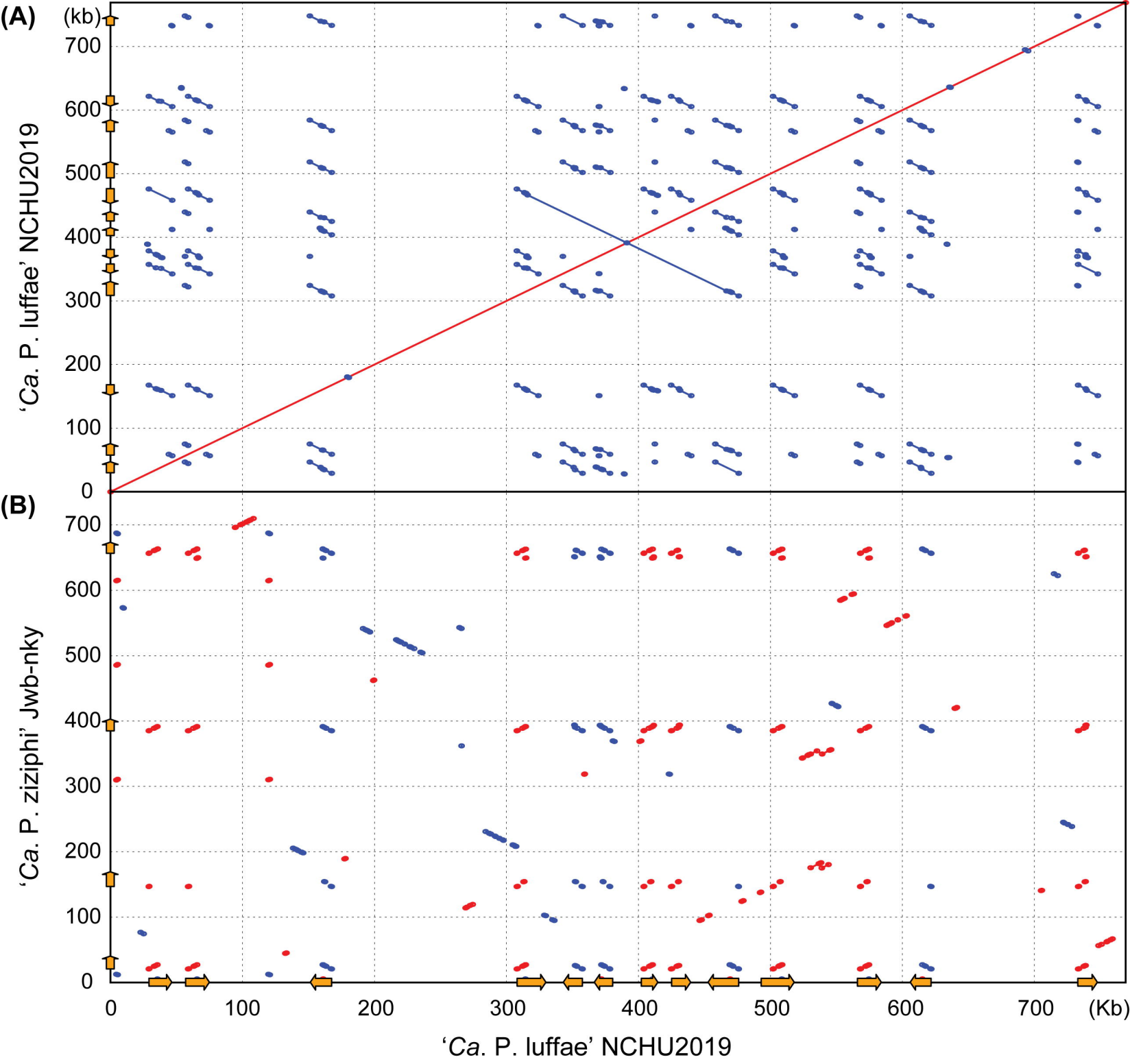
CELL PROCESS

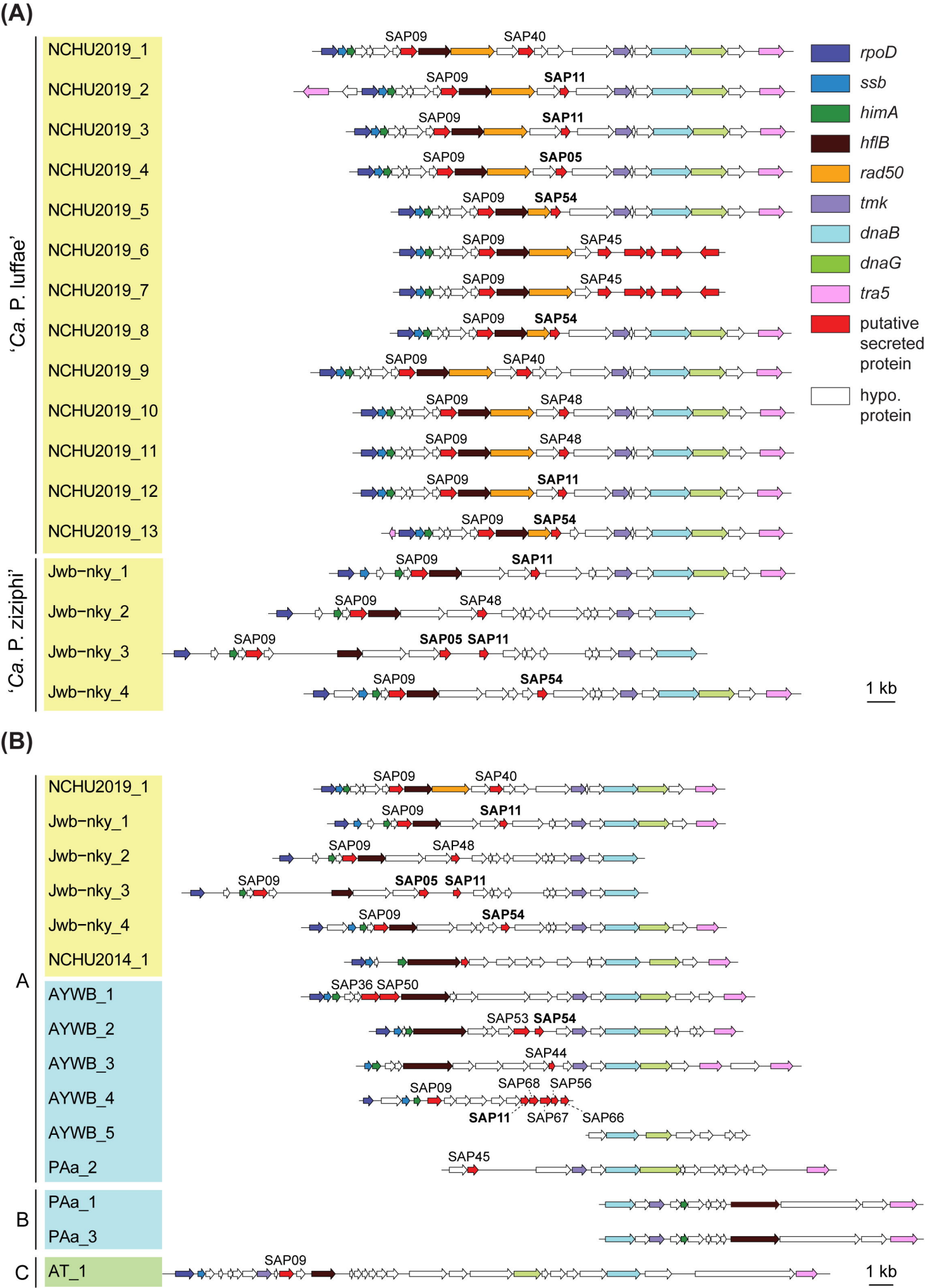
- [D] Cell cycle
- [V] Defense
- [T] Signal transduction
- [M] Cell wall, membrane, & envelope
- [N] Motility
- [W] Extracellular structure
- [U] Trafficking, secretion, & transport
- [O] Posttranslational modification
- [X] Mobilome

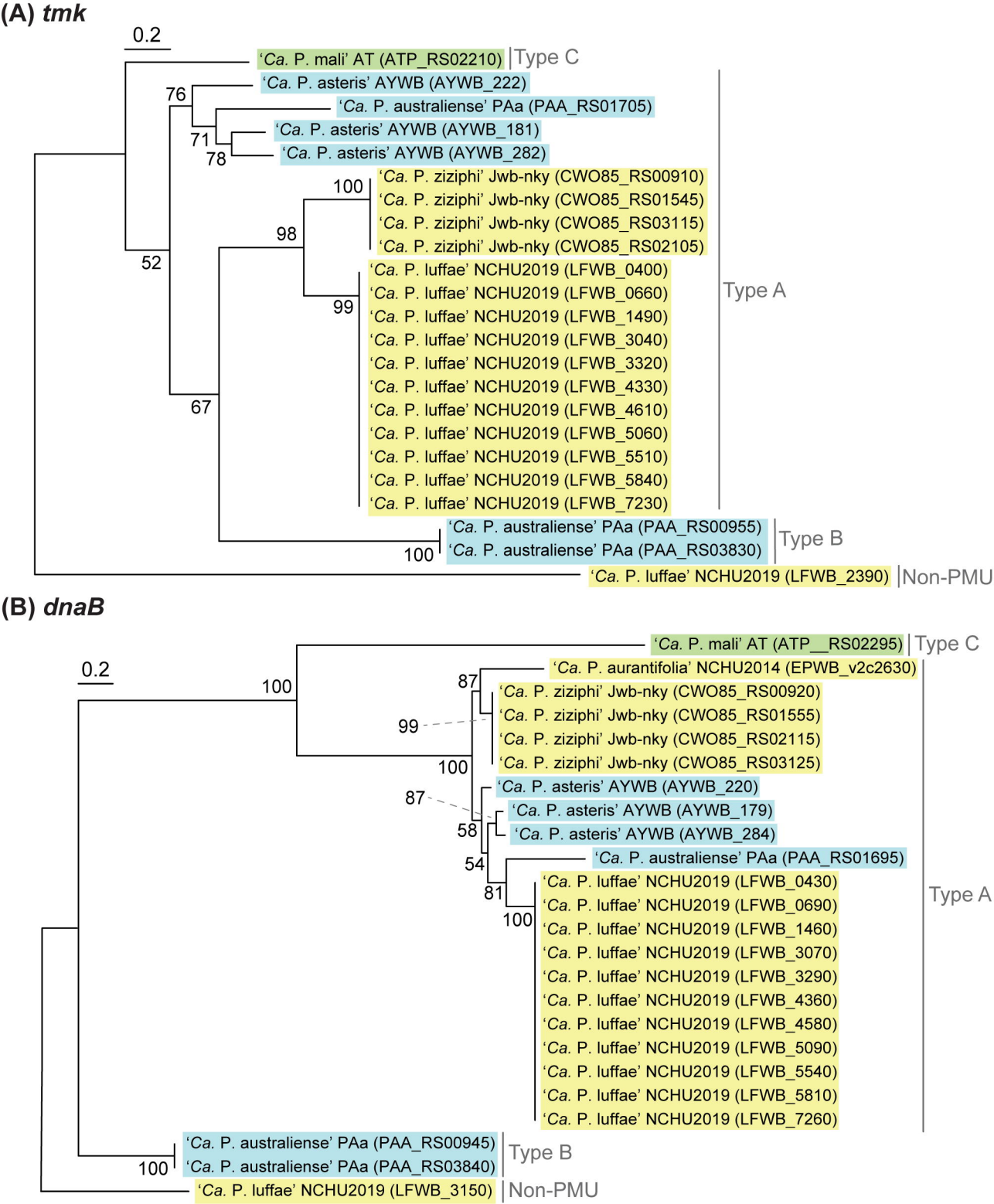
METABOLISM

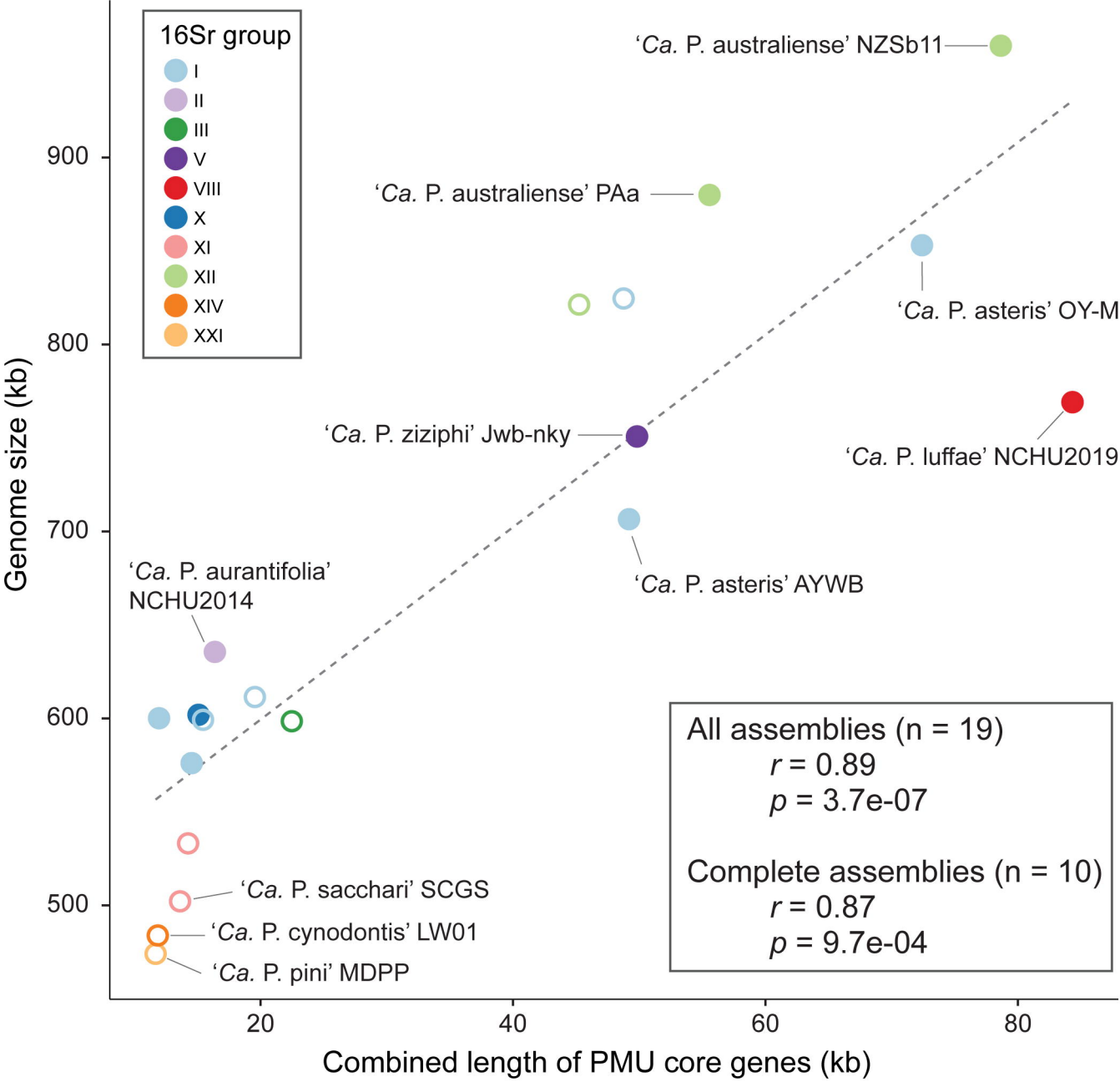
- [C] Energy
- [G] Carbohydrate
- [E] Amino acid
- [F] Nucleotide
- [H] Coenzyme
- [I] Lipid
- [P] Inorganic ion
- [Q] Secondary metabolites

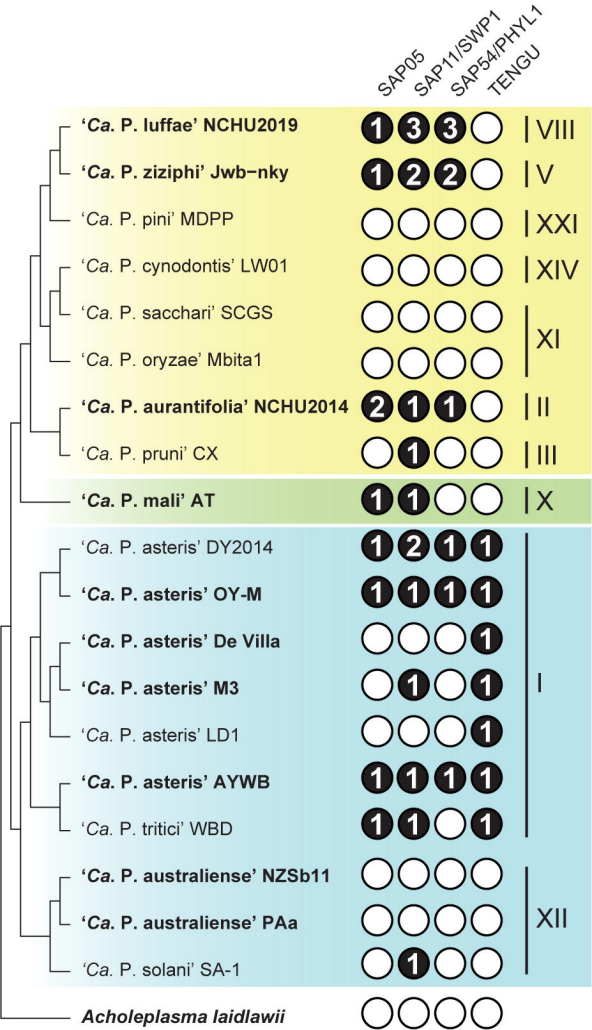

'Ca. P. luffae' NCHU2019 (769)
VIII
'Ca. P. ziziphi' Jwb-nky (751)
V
'Ca. P. pini' MDPP (474)
XXI
'Ca. P. cynodontis' LW01 (483)
XIV
0.1
'Ca. P. sacchari' SCGS (502)
XI
'Ca. P. oryzae' Mbita1 (533)
'Ca. P. aurantifolia' NCHU2014 (636)
II
'Ca. P. pruni' CX (599)
III
'Ca. P. mali' AT (602)
X
'Ca. P. asteris' DY2014 (825)
'Ca. P. asteris' OY-M (853)
'Ca. P. asteris' De Villa (600)
'Ca. P. asteris' M3 (576)
I
'Ca. P. asteris' LD1 (599)
'Ca. P. asteris' AYWB (707)
'Ca. P. tritici' WBD (611)
'Ca. P. australiense' NZSb11 (960)
'Ca. P. australiense' PAa (880)
XII
'Ca. P. solani' SA-1 (821)
Acholeplasma laidlawii (1,497)


(A)
(kb)
700
600
500
400
300
200
100
0
'Ca. P. luffae' NCHU2019
(B)
700
600
500
400
300
200
100
0
'Ca. P. ziziphi' Jwb-nky
0
100
200
300
400
500
600
700
(Kb)
'Ca. P. luffae' NCHU2019

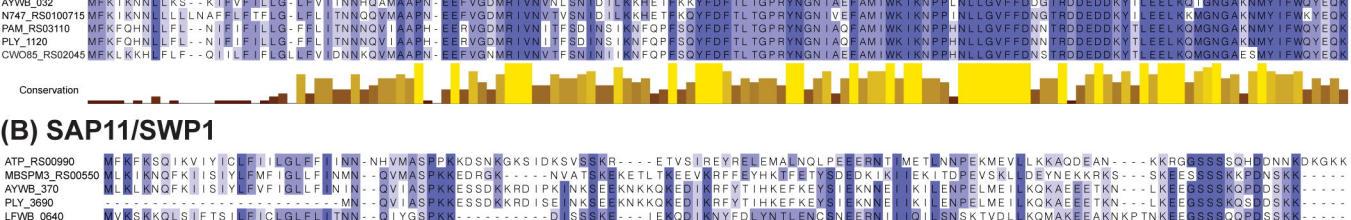

(A)
'Ca. P. luffae'
NCHU2019_1
NCHU2019_2
NCHU2019_3
NCHU2019_4
NCHU2019_5
NCHU2019_6
NCHU2019_7
NCHU2019_8
NCHU2019_9
NCHU2019_10
NCHU2019_11
NCHU2019_12
NCHU2019_13
'Ca. P. ziziphi'
Jwb-nky_1
Jwb-nky_2
Jwb-nky_3
Jwb-nky_4
SAP09
SAP40
SAP11
SAP05
SAP54
SAP45
SAP48
rpoD
ssb
himA
hflB
rad50
tmk
dnaB
dnaG
tra5
putative secreted protein
hypo. protein
1 kb
(B)
A
NCHU2019_1
Jwb-nky_1
Jwb-nky_2
Jwb-nky_3
Jwb-nky_4
NCHU2014_1
AYWB_1
AYWB_2
AYWB_3
AYWB_4
AYWB_5
PAa_2
B
PAa_1
PAa_3
C
AT_1
SAP36
SAP50
SAP53
SAP44
SAP68
SAP56
SAP67
SAP66
1 kb


(A) tmk
0.2
'Ca. P. mali' AT (ATP_RS02210)
Type C
'Ca. P. asteris' AYWB (AYWB_222)
'Ca. P. australiense' PAa (PAA_RS01705)
'Ca. P. asteris' AYWB (AYWB_181)
'Ca. P. asteris' AYWB (AYWB_282)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS00910)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS01545)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS03115)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS02105)
'Ca. P. luffae' NCHU2019 (LFWB_0400)
'Ca. P. luffae' NCHU2019 (LFWB_0660)
'Ca. P. luffae' NCHU2019 (LFWB_1490)
'Ca. P. luffae' NCHU2019 (LFWB_3040)
'Ca. P. luffae' NCHU2019 (LFWB_3320)
'Ca. P. luffae' NCHU2019 (LFWB_4330)
'Ca. P. luffae' NCHU2019 (LFWB_4610)
'Ca. P. luffae' NCHU2019 (LFWB_5060)
'Ca. P. luffae' NCHU2019 (LFWB_5510)
'Ca. P. luffae' NCHU2019 (LFWB_5840)
'Ca. P. luffae' NCHU2019 (LFWB_7230)
Type A
'Ca. P. australiense' PAa (PAA_RS00955)
'Ca. P. australiense' PAa (PAA_RS03830)
Type B
'Ca. P. luffae' NCHU2019 (LFWB_2390)
Non-PMU
76
71
78
100
98
52
99
67
100
(B) dnaB
0.2
'Ca. P. mali' AT (ATP__RS02295)
Type C
'Ca. P. aurantifolia' NCHU2014 (EPWB_v2c2630)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS00920)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS01555)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS02115)
'Ca. P. ziziphi' Jwb-nky (CWO85_RS03125)
'Ca. P. asteris' AYWB (AYWB_220)
'Ca. P. asteris' AYWB (AYWB_179)
'Ca. P. asteris' AYWB (AYWB_284)
'Ca. P. australiense' PAa (PAA_RS01695)
'Ca. P. luffae' NCHU2019 (LFWB_0430)
'Ca. P. luffae' NCHU2019 (LFWB_0690)
'Ca. P. luffae' NCHU2019 (LFWB_1460)
'Ca. P. luffae' NCHU2019 (LFWB_3070)
'Ca. P. luffae' NCHU2019 (LFWB_3290)
'Ca. P. luffae' NCHU2019 (LFWB_4360)
'Ca. P. luffae' NCHU2019 (LFWB_4580)
'Ca. P. luffae' NCHU2019 (LFWB_5090)
'Ca. P. luffae' NCHU2019 (LFWB_5540)
'Ca. P. luffae' NCHU2019 (LFWB_5810)
'Ca. P. luffae' NCHU2019 (LFWB_7260)
Type A
'Ca. P. australiense' PAa (PAA_RS00945)
'Ca. P. australiense' PAa (PAA_RS03840)
Type B
'Ca. P. luffae' NCHU2019 (LFWB_3150)
Non-PMU
100
87
99
100
87
58
54
81
100
100

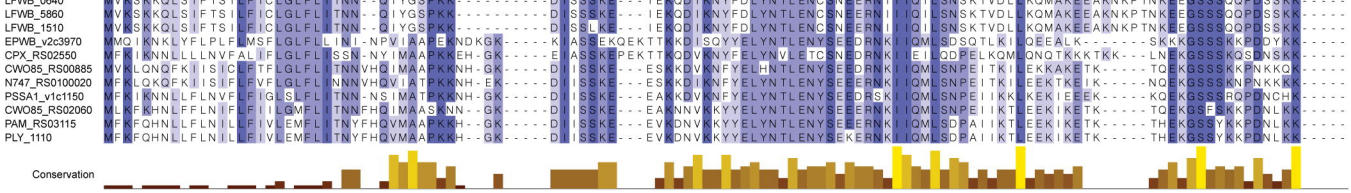

16Sr group
I
II
III
V
VIII
X
XI
XII
XIV
XXI
'Ca. P. australiense' NZSb11
'Ca. P. australiense' PAa
'Ca. P. asteris' OY-M
'Ca. P. ziziphi' Jwb-nky
'Ca. P. luffae' NCHU2019
'Ca. P. asteris' AYWB
'Ca. P. aurantifolia' NCHU2014
'Ca. P. sacchari' SCGS
'Ca. P. cynodontis' LW01
'Ca. P. pini' MDPP
All assemblies (n = 19)
r = 0.89
p = 3.7e-07
Complete assemblies (n = 10)
r = 0.87
p = 9.7e-04
Genome size (kb)
900
800
700
600
500
20
40
60
80
Combined length of PMU core genes (kb)


SAP05
SAP11/SWP1
SAP54/PHYL1
TENGU
'Ca. P. luffae' NCHU2019
1 3 3
VIII
'Ca. P. ziziphi' Jwb−nky
1 2 2
V
'Ca. P. pini' MDPP
XXI
'Ca. P. cynodontis' LW01
XIV
'Ca. P. sacchari' SCGS
'Ca. P. oryzae' Mbita1
XI
'Ca. P. aurantifolia' NCHU2014
2 1 1
II
'Ca. P. pruni' CX
1
III
'Ca. P. mali' AT
1 1
X
'Ca. P. asteris' DY2014
1 2 1 1
'Ca. P. asteris' OY-M
1 1 1 1
'Ca. P. asteris' De Villa
1
'Ca. P. asteris' M3
1 1
'Ca. P. asteris' LD1
1
'Ca. P. asteris' AYWB
1 1 1 1
'Ca. P. tritici' WBD
1 1 1
I
'Ca. P. australiense' NZSb11
'Ca. P. australiense' PAa
'Ca. P. solani' SA-1
1
XII
Acholeplasma laidlawii

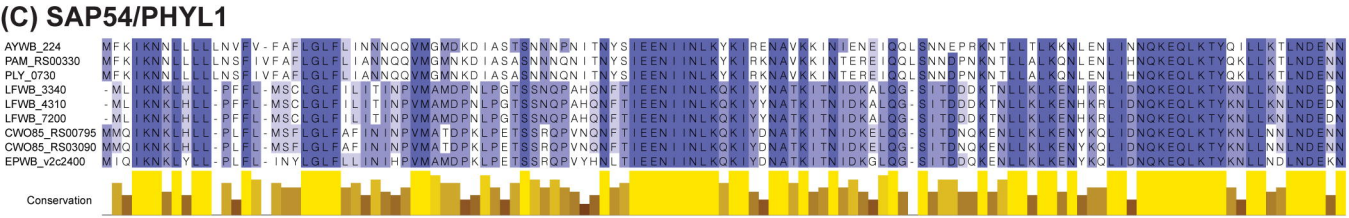
(A) SAP05



(B) SAP11/SWP1



(C) SAP54/PHYL1



(D) TENGU