Species boundaries and molecular markers for classification among 16SrI phytoplasmas informed 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 diverse plant-pathogenic bacteria that greatly impact agriculture worldwide. The current classification system for these uncultivated bacteria 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 that belong to the 16SrI group within the genus '*Candidatus* Phytoplasma' and investigated the possible species boundaries. We found 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 equivalent to species. Strains assigned to the same species share >97% genome-wide average nucleotide identity (ANI) and >78% of their protein-coding genes. In comparison, strains assigned to different species share <94% ANI and <75% of their genes. These findings suggested the existence of barrier(s) against homologous recombination between species, and supported the proposal that 95% ANI could serve as a cutoff for distinguishing species in bacteria. Critical examination of these results and the raw sequencing reads also allowed us to identify one genome that was possibly mis-assembled by mixing two sequencing libraries containing different phytoplasmas. 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 four 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 us to further expand the 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 *Mycoplasma* (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 *Mycoplasma* and several other Mollicutes lineages, axenic cultivation of phytoplasmas 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 plant disease diagnosis 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 phytoplasmas. In addition, to infer the evolutionary history and biodiversity of phytoplasmas, defining species or other equivalent biological units are of fundamental importance. The current 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 iPhyClassifier (Zhao et al., 2009), this system has been well-adopted by the phytoplasma research community.

However, despite the popularity, the current 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. Whereas the use of whole genome sequences 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), whole genome analysis is still challenging for uncultivated 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 is critical for two reasons. First, genome-based classification provides accurate assignments for the strains being considered, which serves as a reference for the MLSA results. Second, genome sequences provide a complete catalog of the genomic regions that may be used as markers, such that the relative performance of each potential marker may be evaluated. In this regard, phytoplasmas belonging to the 16SrI group, also often known as '*Candidatus* Phytoplasma asteris' (i.e.,

defined as encompassing all known subgroups within the 16SrI group) (Lee et al., 2004), represent the best system to test the concept of genome-assisted MLSA marker development. Among the >30 16Sr groups that have been described (Zhao and Davis, 2016), only eight have genome sequences available (Cho et al., 2019b). Moreover, the genome sequencing efforts have been highly focused on the 16SrI group (Table 1), which has a worldwide geographical distribution, wide range of plant hosts, and great impact on agriculture (Lee et al., 2004). Finally, this group is also the one that received most research attention on the molecular mechanisms of plant-microbe interactions; the four phytoplasma effectors that have been characterized were all initially studied in 16SrI phytoplasmas (Sugio et al., 2011b; Tomkins et al., 2018; Gamboa et al., 2019; Huang and Hogenhout, 2019). Our preliminary study of this group identified extensive levels of genome divergence and suggested that multiple species exist (Cho et al., 2019a). In this work, we expanded the sampling of available genome sequences to perform in-depth analysis of genome comparisons. Specifically, we aimed to provide quantitative guidelines for defining species boundaries, which could better inform phytoplasma taxonomy. Additionally, we aimed to develop MLSA markers that could facilitate future genetic characterization of the phytoplasmas in this important group.

Materials and Methods

Data Sets and Analysis Methods

The 11 genome sequences included in this study are listed in Table 1. These data sets represent all phytoplasma genomes available from GenBank (Benson et al., 2018) as of January 2020 and are recognized as belonging to the 16SrI group or '*Ca*. P. asteris' (Lee et al., 2004). Two other '*Ca*. Phytoplasma' species affiliated with the 16SrI group have been described, including '*Ca*. P. japonicum' (Sawayanagi et al., 1999) and '*Ca*. P. lycopersici' (Arocha et al., 2007). However, only few partial gene sequences are available for these two '*Ca*.' species so they were not included in this study.

To confirm the 16Sr group and subgroup assignments, we extracted the 16S rRNA gene sequences from the genomes and analyzed those sequences using the iPhyClassifier (Zhao et al., 2009). Most phytoplasma genomes contain two copies of 16S rRNA genes and each copy was processed individually. The strain OY-V lacks any 16S rRNA gene in the current assembly and therefore was excluded from this analysis. The strain NJAY has four partial sequences for its 16S rRNA genes and all of these four sequences were examined.

All bioinformatic tools for processing these data sets were used with the default settings unless stated otherwise. For correlation tests, the cor.test function implemented in R v3.4.4 (R Core Team, 2019) was used to examine the Pearson's product moment correlation coefficient.

For strain TW1 (GenBank genome accession GCA_003181115.1), we downloaded the raw sequencing results from the NCBI Sequence Read Archive (SRA) to examine the phytoplasma sequences present in the sequencing libraries. The raw reads were mapped to two other reference genomes (i.e., AYWB and OY-M; see Table 1) to detect the presence of 16SrI-A and 16SrI-B type phytoplasmas, respectively. For the Oxford Nanopore MinION reads (SRA accession SRR7548026), the mapping was performed using Minimap2 v2.15 (Li, 2018). For the Illumina MiSeq paired-end reads (SRA accession SRR7548027), the mapping was performed using the Burrows-Wheeler Alignment (BWA) tool v0.7.17 (Li and Durbin, 2009). The mapping results were programmatically checked using SAMtools v1.9 (Li et al., 2009) to identify the sequence variations.

Genome Comparisons

For whole-genome comparison, FastANI v1.1 (Jain et al., 2018) was used to calculate the proportion of genomic segments mapped and the average nucleotide identity (ANI) of those segments in each genome-pair. The genome alignment of representative strains with high quality assemblies were performed using genoPlotR v0.8.9 (Guy et al., 2010). The syntenic regions were identified using BLASTN v2.6.0 (Camacho et al., 2009), the cutoff values for sequence similarity were set to e-value = $1e^{-15}$ and alignment length = 2,000 bp.

For gene-centric investigation, the homologous gene clusters among all genomes were identified using OrthoMCL v1.3 (Li et al., 2003) based on the procedures described in our previous studies (Lo et al., 2013; Chung et al., 2013). Briefly, all coding sequences (CDS), including those annotated as pseudogenes, were included in the all-against-all similarity searches using BLASTN v2.6.0 (Camacho et al., 2009) with an e-value cutoff of $1e^{-15}$. The BLASTN results were used as the input for OrthoMCL, which normalized the within- and between-genome comparisons and identified homologs according to the principle of reciprocal best hits. Similar to the procedures for whole-genome comparisons, the proportion of homologs shared and the sequence identity of shared homologs were considered as two separate metrics of genome similarities. In addition to the pairwise comparisons, the gene content of all genomes was compared using a procedure based on principal coordinates analysis (Lo et al., 2018). For this procedure, the homologous gene clustering result was converted into a matrix of 11 genomes by 1,233 gene clusters; the value in each cell indicates the gene copy number. This matrix was converted into a Jaccard distance matrix among genomes using the VEGAN package in R, then processed using the PCOA function in the APE package (Popescu et al., 2012).

To compare the genes for putative secreted proteins among these phytoplasmas, we examined all coding sequences using a procedure described in a previous study (Cho et al., 2019b). Briefly, SignalP v5.0 (Armenteros et al., 2019) was used to check for the presence of signal peptide based on the Gram-positive bacteria model. The program TMHMM v2.0 (Krogh et al., 2001) was used to verify that these putative secreted proteins have no transmembrane domain.

Molecular Phylogenetics

For phylogenetic inference, MUSCLE v3.8.31 (Edgar, 2004) was used to generate multiple sequence alignments and PhyML v3.3 (Guindon and Gascuel, 2003) was used to infer maximum likelihood phylogenies. The visualization was performed using JalView v2.11 (Waterhouse et al., 2009) and FigTree v1.4.4. In each inference, PHYLIP v3.697 (Felsenstein, 1989) was used to generate 1,000 replicates for bootstrap analysis.

Development of Molecular Markers

To identify molecular markers that may be suitable for classification, those multiple sequence alignments of single-copy genes shared by all genomes were examined manually. Genes were ranked by the number of variable sites that could be used for distinguishing different 16SrI subgroups. Then the list was prioritized based on the criterion of having multiple markers located in distinct regions of those phytoplasma chromosomes. The PCR primers (Table 2) were designed manually with the aim of producing ~700-800 bp products to cover those variable regions, such that each product could be sequenced in one single Sanger sequencing reaction. To validate the primer design, a DNA sample derived from a periwinkle infected by the phytoplasma strain DY2014 (Cho et al., 2019b) was used as the template for PCR tests using the

KAPA Taq ReadyMix (Roche, USA) following the manufacturer's instructions. The cycling conditions were: (1) an initial 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 60 sec; and (3) a final extension step of 72°C for 5 min. The annealing temperature was lowered to 53°C for the newly designed primers for *secY* (Table 2), while the other three primer sets used 55°C. Agarose gel electrophoresis was conducted to confirm that all PCR tests produced one single band of expected size. For further confirmation, the PCR products were purified using the QIAquick Gel Extraction Kit (Qiagen, Germany) according to the manufacturer's instructions, then sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit on an Applied Biosystems 3730XL DNA Analyzer (Thermo Fisher Scientific, USA). The sequencing results were compared to the genome sequence for final confirmation.

The procedure for calculating the substitution rates was based on that described in our previous studies (Kuo and Ochman, 2009; Kuo et al. 2009). Briefly, the homologous protein sequences were aligned by using MUSCLE v3.8.31 (Edgar, 2004). The protein alignments were converted into codon-based nucleotide alignment using PAL2NAL v14 (Suyama et al., 2006), which were then processed using the YN00 method (Yang and Nielsen, 2000) implemented in PAML v4.9h (Yang 2007).

Results and Discussion

Genome Characteristics

Among the 11 phytoplasma genomes included in this study, four have the complete chromosomal sequence available while the other seven are draft assemblies with varying levels of completeness (Table 1). For those incomplete genomes, the true genome sizes are difficult to determine. Nonetheless, the largest (i.e., OY-M; 853 kb) and the smallest (i.e., MBSP-M3; 576 kb) genomes in this data set are both completed assemblies, indicating the high level of genome size variation among these closely related strains. The total number of CDS is strongly correlated with the genome size ($r = 0.92$, $p = 2.7e\text{-}5$). However, the number and proportion of pseudogenes varied widely among these genomes. For example, the strains AYWB and NJAY have a genome-wide ANI of 99.7%, yet the proportion of pseudogenes are 17% and 41%, respectively. In another case, DY2014 and OY-V have a genome-wide ANI of 99.4%, yet the proportion of pseudogenes are 8% and 25%, respectively. Although these differences may reflect true biological processes, such as some strains have experienced genome degradation, it is also possible that these differences may be explained by artifacts. Regarding the sequencing methods, AYWB was based on Sanger sequencing (Bai et al., 2006) and DY2014 was based on Illumina (Cho et al., 2019b), while NJAY was based on one single Roche/454 library (Sparks et al., 2018) and OY-V was based on one Illumina paired-end library and one 454 GS FLX mate-pair library (Kakizawa et al., 2014). Due to the low GC-content of these genomes (i.e., ~27-28%), it is plausible that sequencing errors in homopolymeric regions, particularly those errors originated from the 454 sequencing technology, may have contributed to the high numbers of pseudogenes with frameshift mutations in NJAY and OY-V. Moreover, assemblies that are more fragmented may also have more partial genes located at the ends of contigs. Due to these concerns, all pseudogenes were included in the gene-centric investigation of this study.

Classification of the 16SrI Phytoplasmas Analyzed

The 16SrI subgroup assignments of these strains were mostly as expected but there were few surprises (Figure 1A). First, WBD was reported as a 16SrI-B strain (Chen et al., 2014).

However, our examination of the sequence record deposited in GenBank revealed that the two copies of 16S rRNA genes in this genome differ by 4-bp, and iPhyClassifier (Zhao et al., 2009) assigned these two sequences to 16SrI-R and 16SrI-S, respectively. Results from molecular phylogeny also support that these two sequences form a clade that is distinct from other 16SrI-B sequences. Second, DY2014 has two 16S rRNA genes that differ by 1-bp. While one was assigned to 16SrI-B, the other was identified as being a new subgroup with a RFLP pattern most similar to 16SrI-B. Finally, TW1 has two sequences that differ by 12-bp, with one assigned to 16SrI-A and the other assigned to 16SrI-B (Town et al., 2018).

These results highlighted several issues of this RFLP-based classification system for phytoplasmas. First, the intra-genomic variation of 16S rRNA genes may result in a strain being assigned to different subgroups depending on which homolog was examined. This issue has been reported for other phytoplasmas and a three-letter subgroup designation system has been proposed (Wei et al., 2008). For example, a paulownia witches'-broom (PaWB) phytoplasma was first classified as a 16SrI-D strain (Lee et al., 1998) but later found to harbor a 16SrI-B type sequence as well. To accommodate these findings, the subgroup status of this PaWB was redesignated as 16SrI-(B/D)D, with the first two letters in parentheses denoting the types of those two 16S rRNA genes, and the third letter denoting its 16Sr subgroup assignment (Wei et al., 2008; Zhao et al., 2009). Although this three-letter system provides comprehensive information, it may be unnecessarily complex and potentially confusing. Moreover, it is unclear which one of the two types should be chosen for the formal subgroup assignment. In the case of PaWB, it was assigned to 16SrI-D based on precedence. However, in the case of WBD (i.e., originally described as a 16SrI-B strain), it is unclear if it should be redesignated as 16SrI-(R/S)R or 16SrI-(R/S)S. Second, this RFLP-based system considers only the sequence variations in restriction sites. By ignoring the sequence variations in other positions, this approach could alleviate the problem of intra-genomic variations to a certain degree (e.g., the two sequences in AYWB differ by 2-bp, yet both were assigned to 16SrI-A). However, this approach wasted much of the information contained in the full sequence and the uneven weighing of nucleotide positions could introduce strong biases. The strain DY2014 represents one extreme example to illustrate such biases, with a single-bp difference that is sufficient to result in different subgroup assignments between the two 16S rRNA genes. Finally, a classification system that depends solely on the 16S rRNA genes while ignoring other regions of the genome does not provide sufficient resolution.

Regardless of the group/subgroup assignments of these 16S rRNA genes, all these sequences share >99% identity (Supplementary Table S1), which is above the general recommended thresholds for defining species. These thresholds include: (1) 97% identity that is commonly used for defining operational taxonomic units (OTUs) corresponding to species in microbiota surveys (Schloss and Handelsman, 2005), (2) 97.5% identity recommended for '*Ca*. phytoplasma' species (The IRPCM Phytoplasma/Spiroplasma Working Team, 2004), and (3) 98.6% identity recommended for uncultivated bacteria (Konstantinidis et al., 2017). In other words, these strains may be considered as all belonging to '*Ca*. Phytoplasma asteris' (Lee et al., 2004) based on their 16S rRNA gene sequences. However, delineating species based on one single gene may be problematic; more in-depth investigation of their genomic divergence is necessary.

<u>Genomic Divergence among 16SrI Phytoplasmas and Putative Species Boundaries</u>

To further investigate the divergence of these phytoplasmas, we utilized two approaches to quantify their genomic similarities. In the first approach, the whole genome sequences were used to calculate the proportion of genomic segments shared, as well as the ANI values of these segments. A previous study that examined >90,000 prokaryotic genomes found that the ANI values exhibit a bimodal distribution with the within-species comparisons almost always have >95% ANI (Jain et al., 2018). We found that based on this criterion, the 11 strains examined could be separated into three clusters that roughly correspond to their 16SrI subgroup assignments (Figure 2A). The two 16SrI-A strains, AYWB and NJAY, share 86.9% of their genomes and these shared regions are nearly identical (i.e., ANI = 99.7%). Because the NJAY assembly is highly fragmented and ~10% smaller compared to the complete genome of AYWB, the low proportion of genomic segments shared may be explained by missing data in the NJAY assembly. The strain WBD shares only ~93% ANI with all other 16SrI phytoplasmas compared and represents a distinct lineage. All of the remaining strains, including those assigned to 16SrI-B and TW1, share >97% ANI and could considered as a third cluster. This clustering result is consistent with the maximum likelihood phylogeny inferred using 303 single-copy genes shared by all strains (Figure 1B and Supplementary Table S2).

In the second approach, we considered only the putative coding regions (i.e., intact CDS and putative pseudogenes) of these genomes for calculating genomic similarities (Figure 2B). Compared to the genome-wide ANI approach, this gene-centric method is more laborious and could potentially be affected by annotation quality. Nevertheless, this approach has a higher confidence in inferring the homology prior to calculating sequence identities and provides a complementary method to the ANI approach. Although the exact genome similarities values differ slightly, the patterns found by using these two approaches are consistent.

Based on these results regarding genomic similarities and molecular phylogeny, it appeared that these 16SrI phytoplasmas could be classified into three taxonomic units corresponding to species. Similar to a previous study that examined diverse prokaryotes (Jain et al., 2018), the sequence identity values among these phytoplasmas also exhibited a bimodal distribution (Figure 3). Excluding the comparisons involving TW1, the within-species comparisons all have >97% genome-wide ANI, while between-species comparisons all have <94% ANI (Figures 2A and 3A). For sequence identities that were calculated based on only the genic regions, in which the homology could be inferred more confidently, a similar separation was found (i.e., within-species: >98.9%, between-species: <96%; see Figures 2B and 3B). These results suggest that those strains assigned to the same species may have on-going homologous recombination that maintained high sequence similarities in their shared genomic regions. In contrast, certain genetic barriers exist to lower homologous recombination between different species. In other words, these results provide further support to the hypothesis that the biological species concept, which is based on the barriers to homologous gene exchange, could be applied to bacteria as well (Bobay and Ochman, 2017). Moreover, these results agree with a previous study in that the conventional 97% identity threshold for delineating species based on 16S rRNA gene is too low (Edgar, 2018).

In addition to sequence identity, another important measurement for genetic similarity is the proportion of genomic regions shared. For this second measurement, a wide range was observed for both within- and between-species comparisons (Figures 2 and 3). The wide spread was mostly explained by the inclusion of two highly fragmented draft genomes (i.e., OY-V and CYP; see Table 1), which resulted in low estimates that are artifacts stemming from missing data. Similarly, the inclusion of WBD produced high proportions of genomic regions shared between

species (i.e., 82% with AYWB and 84% with MBSP-M3). If the complete genome sequence of WBD is available, the denominator would change and these estimates could change in either direction. Due to these uncertainties, we restricted the comparisons to those four strains with the complete genome sequences available (i.e., AYWB, OY-M, MBSP-M3, and De Villa). For this reduced data set with high confidence, strains assigned to the same species share ~78-93% of their chromosomal segments and ~79-89% of their protein-coding genes, whereas strains assigned to different species share ~74-78% of their chromosomal segments and ~70-75% of their protein-coding genes.

For an alternative approach of visualization, we used the principal coordinates analysis to compare the gene content divergence among these phytoplasmas (Figure 4). With the exception of TW1, the results were consistent with the molecular phylogeny (Figure 1B) and pairwise genome similarities (Figures 2 and 3). The first coordinate explained ~40% of the variance and showed clear separation for the three putative species. The second coordinate explained ~23% of the variance and showed that those 16SrI-B strains could be separated into three subgroups.

Taken together, these results supported the existence of multiple operational taxonomic units equivalent to species within the previously described '*Ca*. Phytoplasma asteris' (Lee et al., 2004) and provided quantitative guidelines for establishing the species boundaries. Future work to improve the sampling of genomic diversity among these important plant pathogens are necessary to inform the taxonomy revisions.

<u>The Special Case of the TW1 Genome</u>

Based on the aforementioned framework for phytoplasma classification, the strain TW1 represents a strange and difficult case. The high level of intra-genomic variation between its two copies of 16S rRNA genes has been noted in the initial characterization (Town et al., 2018). In our whole-genome analysis, this strain shares a higher proportion of its chromosomal segments and protein-coding genes with those 16SrI-A strains, yet has higher sequence identities with those 16SrI-B strains (Figures 2 and 3). Because of these properties, this strain was clustered with those 16SrI-A strains based on gene content (Figure 4), while having closer phylogenetic relationships with those 16SrI-B strains (Figure 1B).

In the absence of an intuitive hypothesis that could explain the evolutionary processes leading to these conflicting findings, we examined the possibilities that these findings were results of artifacts. This TW1 genome was co-assembled using one long-read library (based on Oxford Nanopore MinION) and one short-read library (based on Illumina MiSeq) (Town et al., 2018). When the raw reads from these two sequencing libraries were examined separately, we found that the long-read library likely contains a 16SrI-A type phytoplasma. When the raw reads were mapped to the AYWB chromosome, all regions were covered and 3,725 sequence polymorphisms were found; when mapped to OY-M, 52,649 bp had no coverage and 31,401 sequence polymorphisms were found. In contrast, the short-read library likely contains a 16SrI-B type phytoplasma. When the raw reads were mapped to OY-M, 103,922 bp had no coverage and 9,489 sequence polymorphisms were found; when mapped to AYWB, 110,458 had no coverage and 29,323 sequence polymorphisms were found.

Based on these findings, a possible explanation is that during the assembly process, the long reads produced several 16SrI-A type scaffolds, which resulted in the high similarities of the TW1 genome content to 16SrI-A as observed (Figures 2-4). Subsequently, during the polishing stage, the exact sequences of these scaffolds were modified based on the short-read library, thus explaining the high sequence similarity of this TW1 genome to those 16SrI-B strains. In other

words, in its current form, this TW1 genome assembly represents an artifact due to mis-assembly, rather than a true representative of an existing phytoplasma strain.

Development of Molecular Markers for Multilocus Sequence Analysis

To develop molecular markers that may be useful for the MLSA of these phytoplasmas and their relatives, we examined the multiple sequence alignments of those 303 single-copy genes shared by all of the genomes analyzed. The selection criteria were based on: (1) the number of informative sites that may be used to distinguish the three species identified, (2) the possibility of designing PCR primers for a product in the size range of ~700-800 bp and covers a large number of informative sites, and (3) the chromosomal locations. The first two criteria were aimed to make the Sanger sequencing of these markers as cost-effective as possible, while the third criterion was aimed to make the MLSA robust against recombination.

Based on these criteria, we selected four markers to supplement the 16S rRNA gene. These four makers all have high numbers of informative sites, particularly when compared to those markers that have been developed previously (Table 2 and Supplementary Figure S1). The substitute rates of these marker genes are relatively high but do not deviate strongly when compared to other shared single-copy genes (Supplementary Figure S2).

In terms of chromosomal locations, these markers are well separated in all genomes examined despite the extensive genome rearrangements observed (Figure 5). Although the extent of recombination in phytoplasmas is not well understood, recent studies on *Xylella fastidiosa* (i.e., another insect-transmitted plant pathogen that is restricted to vascular tissues) revealed that recombination affected ~0.5-15% of the genome, depending on the exact subspecies (Potnis et al., 2019; Vanhove et al., 2019). The recombined fragments could have sizes up to 31 kb, with an average of 1 kb. Based on these estimates, it is unlikely that one single recombination event would affect more than one of these four markers. Thus, the results of MLSA should be robust against the interference of recombination.

When we used these four newly developed markers for molecular phylogenetic inference, the concatenated alignment (together with the 16S rRNA gene, which is expected to be sequenced in the initial characterization of newly collected strains) produced a phylogeny (Figure 1C) that is comparable to the result from genome-scale analysis (Figure 1B). This result demonstrated the utility of these markers for future genotyping work. With better sampling of phytoplasma strains and more accurate classification of those strains using these markers, informed decision could be made to select representatives for genome sequencing efforts, which could contribute to the study of phytoplasma genetic diversity and evolutionary history, as well as improve the taxonomy.

Phylogenetic Distribution of Phytoplasma Effector Genes

With the well-resolved phylogeny (Figure 1B) as a framework, we investigated the distribution of putative effector genes among these phytoplasmas. The results indicated a high level of variation in the gene counts (Figure 6). One clade (i.e., MBSP-M3, De Villa, and LD1) have only ~10-14 genes that encode putative secreted proteins. These numbers are much lower than the ~34-44 found in their sister clade (i.e., OY-M, OY-V, and DY2014). The low numbers are not artifacts of incomplete genome sequences. In fact, the two strains with the lowest numbers (i.e., MBSP-M3 and De Villa) both have the complete chromosomal sequences available. These findings suggest that there may have been lineage-specific reduction in effector gene counts, although the underlying ecological factors are unclear.

For those four phytoplasma effectors that have been characterized experimentally, namely SAP05 (Gamboa et al., 2019; Huang and Hogenhout, 2019), SAP11 (Bai et al., 2009; Sugio et al., 2011a; Lu et al., 2014; Sugio et al., 2014; Chang et al., 2018), SAP54/PHYL1 (MacLean et al., 2011; MacLean et al., 2014; Maejima et al., 2014; Orlovskis and Hogenhout, 2016), and TENGU (Hoshi et al., 2009; Sugawara et al., 2013; Minato et al., 2014), only TENGU is conserved among all strains. Although some cases of gene absence may be artifacts of incomplete genome assemblies, the strain De Villa (which has the complete chromosomal sequence available) lacks all other three effector genes, indicating that none of those three effectors is essential.

The variable pattern of effector gene presence and absence may be linked to the 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 genome similarities. The results suggested that these phytoplasmas could be classified into multiple distinct taxonomic units equivalent to species. Importantly, the current classification system for phytoplasmas that is based on the RFLP analysis of their 16S rRNA genes does not provide sufficient resolution or accuracy. Although using 16Sr subgroup for species may appears to be a simple solution (e.g., in the cases of 16SrI-A and 16SrI-B), this does not always work (e.g., several strains have two 16S rRNA genes that were assigned to different subgroups). As such, 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.

Acknowledgments

We thank Ms. Mei-Jane Fang in the Genomic Technology Core Facility of our institute for providing Sanger sequencing service.

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. 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.

References

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

Arocha, Y., Antesana, O., Montellano, E., Franco, P., Plata, G., and Jones, P. (2007). '*Candidatus* Phytoplasma lycopersici', a phytoplasma associated with 'hoja de perejil' disease in Bolivia. *Int J Syst Evol Microbiol* 57, 1704–1710. doi:10.1099/ijs.0.64851-0.

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

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

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

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

Bobay, L.-M., and Ochman, H. (2017). Biological species are universal across life's domains. *Genome Biol Evol* 9, 491–501. doi:10.1093/gbe/evx026.

Brown, D. R. (2010). "Phylum XVI. Tenericutes Murray 1984a, 356VP (Effective publication: Murray 1984b, 33.)," in *Bergey's Manual of Systematic Bacteriology*, eds. N. R. Krieg, J. T. Staley, D. R. Brown, B. P. Hedlund, B. J. Paster, N. L. Ward, et al. (Springer New York), 567–723. doi:10.1007/978-0-387-68572-4_5.

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

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

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

Chen, W., Li, Y., Wang, Q., Wang, N., and Wu, Y. (2014). Comparative genome analysis of wheat blue dwarf phytoplasma, an obligate pathogen that causes wheat blue dwarf disease in China. *PLOS ONE* 9, e96436. doi:10.1371/journal.pone.0096436.

Cho, S.-T., Kung, H.-J., and Kuo, C.-H. (2019a). Genomic similarities among 16SrI phytoplasmas and implications on species boundaries. *Phytopathogenic Mollicutes* 9, 73. doi:10.5958/2249-4677.2019.00037.9.

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

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

Coetzee, B., Douglas-Smit, N., Maree, H. J., Burger, J. T., Krüger, K., and Pietersen, G. (2019). Draft genome sequence of a "*Candidatus* Phytoplasma asteris"-related strain (aster yellows, subgroup 16SrI-B) from South Africa. *Microbiol Resour Announc* 8, e00148-19. doi:10.1128/MRA.00148-19.

Dumonceaux, T. J., Green, M., Hammond, C., Perez, E., and Olivier, C. (2014). Molecular diagnostic tools for detection and differentiation of phytoplasmas based on chaperonin-60 reveal differences in host plant infection patterns. *PLOS ONE* 9, e116039. doi:10.1371/journal.pone.0116039.

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

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34, 2371–2375. doi:10.1093/bioinformatics/bty113.

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

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

Glaeser, S. P., and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst Appl Microbiol* 38, 237–245. doi:10.1016/j.syapm.2015.03.007.

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

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

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

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

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

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

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

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.

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., 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.

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

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

Lim, P.-O., and Sears, B. B. (1991). DNA sequence of the ribosomal protein genes *rp12* and *rps19* from a plant-pathogenic mycoplasma-like organism. *FEMS Microbiol Lett* 84, 71–74. doi:10.1111/j.1574-6968.1991.tb04571.x.

Lo, W.-S., Chen, L.-L., Chung, W.-C., Gasparich, G. E., and Kuo, C.-H. (2013). Comparative genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium. *BMC Genomics* 14, 22. doi:10.1186/1471-2164-14-22.

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

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

MacLean, A. M., Orlovskis, Z., Kowitwanich, K., Zdziarska, A. M., Angenent, G. C., Immink, R. G. H., et al. (2014). Phytoplasma effector SAP54 hijacks plant reproduction by degrading MADS-box proteins and promotes insect colonization in a RAD23-dependent manner. *PLOS Biol* 12, e1001835. doi:10.1371/journal.pbio.1001835.

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

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

Martini, M., Lee, I.-M., Bottner, K. D., Zhao, Y., Botti, S., Bertaccini, A., et al. (2007). Ribosomal protein gene-based phylogeny for finer differentiation and classification of phytoplasmas. *Int J Syst Evol Microbiol* 57, 2037–2051. doi:10.1099/ijs.0.65013-0.

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

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

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

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

Pacifico, D., Galetto, L., Rashidi, M., Abbà, S., Palmano, S., Firrao, G., et al. (2015). Decreasing global transcript levels over time suggest that phytoplasma cells enter stationary phase during plant and insect colonization. *Appl Environ Microbiol* 81, 2591–2602. doi:10.1128/AEM.03096-14.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. doi:10.1038/nbt.4229.

Potnis, N., Kandel, P. P., Merfa, M. V., Retchless, A. C., Parker, J. K., Stenger, D. C., et al. (2019). Patterns of inter- and intrasubspecific homologous recombination inform eco-evolutionary dynamics of *Xylella fastidiosa*. *ISME J*, 1. doi:10.1038/s41396-019-0423-y.

Popescu, A.-A., Huber, K. T., and Paradis, E. (2012). ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* 28, 1536–1537. doi:10.1093/bioinformatics/bts184.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.r-project.org/ [Accessed August 1, 2019].

Sawayanagi, T., Horikoshi, N., Kanehira, T., Shinohara, M., Bertaccini, A., Cousin, M.-T., et al. (1999). '*Candidatus* Phytoplasma japonicum', a new phytoplasma taxon associated with Japanese *Hydrangea* phyllody. *Int J Syst Evol Microbiol* 49, 1275–1285. doi:10.1099/00207713-49-3-1275.

Schloss, P. D., and Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71, 1501–1506. doi:10.1128/AEM.71.3.1501-1506.2005.

Schneider, B., and Gibb, K. S. (1997). Sequence and RFLP analysis of the elongation factor Tu gene used in differentiation and classification of phytoplasmas. *Microbiology*, 143, 3381–3389. doi:10.1099/00221287-143-10-3381.

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

Sparks, M. E., Bottner-Parker, K. D., Gundersen-Rindal, D. E., and Lee, I.-M. (2018). Draft genome sequence of the New Jersey aster yellows strain of '*Candidatus* Phytoplasma asteris.' *PLOS ONE* 13, e0192379. doi:10.1371/journal.pone.0192379.

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

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

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

Sugio, A., MacLean, A. M., and Hogenhout, S. A. (2014). The small phytoplasma virulence effector SAP11 contains distinct domains required for nuclear targeting and CIN-TCP binding and destabilization. *New Phytol* 202, 838–848. doi:10.1111/nph.12721.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucl Acids Res* 34, W609-612. doi:10.1093/nar/gkl315.

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

Tomkins, M., Kliot, A., Marée, A. F., and Hogenhout, S. A. (2018). A multi-layered mechanistic modelling approach to understand how effector genes extend beyond phytoplasma to modulate plant hosts, insect vectors and the environment. *Curr Opin Plant Biol* 44, 39–48. doi:10.1016/j.pbi.2018.02.002.

Town, J. R., Wist, T., Perez-Lopez, E., Olivier, C. Y., and Dumonceaux, T. J. (2018). Genome sequence of a plant-pathogenic bacterium, "*Candidatus* Phytoplasma asteris" strain TW1. *Microbiol Resour Announc* 7, e01109-18. doi:10.1128/MRA.01109-18.

Vanhove, M., Retchless, A. C., Sicard, A., Rieux, A., Coletta-Filho, H. D., Fuente, L. D. L., et al. (2019). Genomic diversity and recombination among *Xylella fastidiosa* subspecies. *Appl Environ Microbiol* 85, e02972-18. doi:10.1128/AEM.02972-18.

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

Wei, W., Lee, I.-M., Davis, R. E., Suo, X., and Zhao, Y. (2008). Automated RFLP pattern comparison and similarity coefficient calculation for rapid delineation of new and distinct phytoplasma 16Sr subgroup lineages. *Int J Syst Evol Microbiol* 58, 2368–2377. doi:10.1099/ijs.0.65868-0.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24, 1586–1591. doi:10.1093/molbev/msm088.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17, 32–43.

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

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

Zhu, Y., He, Y., Zheng, Z., Chen, J., Wang, Z., and Zhou, G. (2017). Draft genome sequence of rice orange leaf phytoplasma from Guangdong, China. *Genome Announc.* 5, e00430-17. doi:10.1128/genomeA.00430-17.

Figure Legends

Figure 1. Maximum likelihood phylogenies. The numbers on the internal branches indicate the level of bootstrap support based on 1,000 resampling; only values >=60% are shown. (A) 16S rRNA genes. The alignment contains 1,541 aligned nucleotide sites. The strain name, locus tag of the 16S rRNA gene (in parentheses), and the 16SrI subgroup classification (in square brackets) are labeled. A question mark '?' in the subgroup classification indicates that the sequence is classified as a new subgroup, the existing subgroup with the highest similarity is provided. (B) Single-copy coding genes shared by all strains. The concatenated alignment contains 303 genes and 291,990 aligned nucleotide sites. (C) Five molecular markers selected in this study. The concatenated alignment contains 4,228 aligned nucleotide sites.

Figure 2. Pairwise genome similarities. The strains with the complete genome sequences available are highlighted in bold. (A) Similarity scores based on whole genome comparisons. Above diagonal: average nucleotide identity (%); diagonal: genome size (kb); below diagonal: genomic segments mapped (%). (B) Similarity scores based on genic regions. Above diagonal: genic sequence identity (%); diagonal: number of homologous gene clusters; below diagonal: homologs shared (%).

Figure 3. Correlations between metrics of genome similarities. (A) Similarity scores based on whole genome comparisons. (B) Similarity scores based on genic regions. Filled symbols indicate the pairwise comparisons involving two complete genomes, empty symbols indicate the pairwise comparisons involving one or two draft genomes. The strain TW1 is a special case; its similarity scores in comparison with 16SrI-A and 16SrI-B strains are indicated by red and blue, respectively.

Figure 4. Principal coordinates analysis of gene content. Each genome is represented by one dot (color-coded according to the 16Sr subgroup assignment). The distance between dots corresponds to their differences in gene content. The % variance explained by each axis is provided in parentheses. The strains with the complete genome sequences available are highlighted in bold.

Figure 5. Genomic locations of marker genes. Nucleotide sequence identity of the conserved regions between genomes are illustrated (red: match on the same strand; blue: match on the opposite strand).

Figure 6. Phylogenetic distribution of putative effector genes. The cladogram that illustrates the phylogenetic relationships among these strains is based on Figure 1B (i.e., maximum likelihood phylogeny based on 303 shared single-copy genes). The number in parentheses after the strain name indicates the number of genes that encode putative secreted proteins in each genome. The strains with the complete genome sequences available are highlighted in bold. On the right-hand side, the table lists the locus tags of homologs for those four phytoplasma effectors that have been characterized. Those locus tags with an asterisk (*) were annotated as putative pseudogenes.

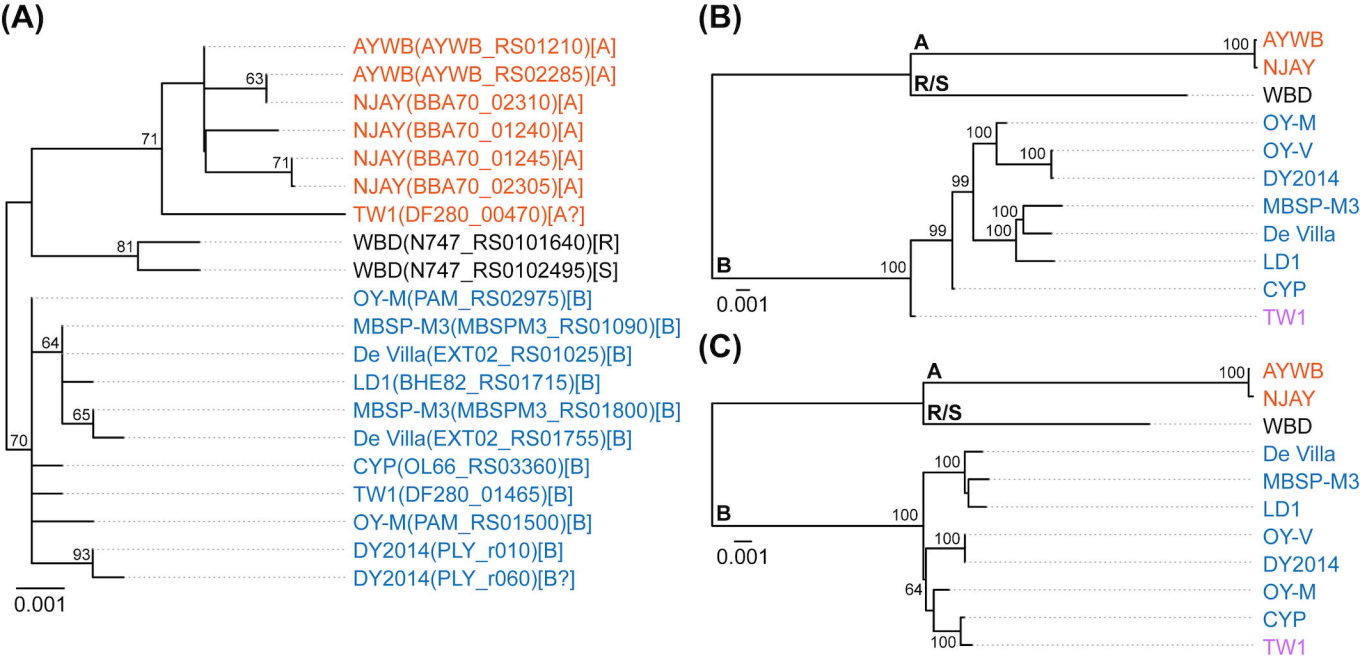
Supplementary Materials

Supplementary Figure S1. Multiple sequence alignments of selected marker genes. The primer sites are highlighted in orange. Shades of blue colors in the alignment indicate the levels of sequence conservation.

Supplementary Figure S2. Distribution of substitution rates among those single-copy genes shared by the 11 phytoplasma genomes analyzed. The synonymous substitution rate (Ks) and non-synonymous substitution rate (Ka) were calculated based on the homologs found in AYWB and MBSP-M3. Among those 303 single-copy genes, nine were annotated as pseudogenes and were omitted from the substitution rate calculation.

Supplementary Table S1. Pairwise nucleotide sequence identity of the molecular markers used in the phylogenetic inference. Numbers in the diagonal cells indicate the length of each unaligned sequence.

Supplementary Table S2. List of the 303 single-copy genes shared by all of the 11 phytoplasma genomes analyzed in this study. The gene identity, genomic location, and functional annotation of these genes are provided based on the information from AYWB. The substitution rate estimates are based on the comparison between AYWB and MBSP-M3.


(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
99
100
100
100
0.001
(C)
A
R/S
B
100
AYWB
NJAY
WBD
De Villa
MBSP-M3
LD1
OY-V
DY2014
OY-M
CYP
TW1
100
100
100
64
100
0.001

(A)

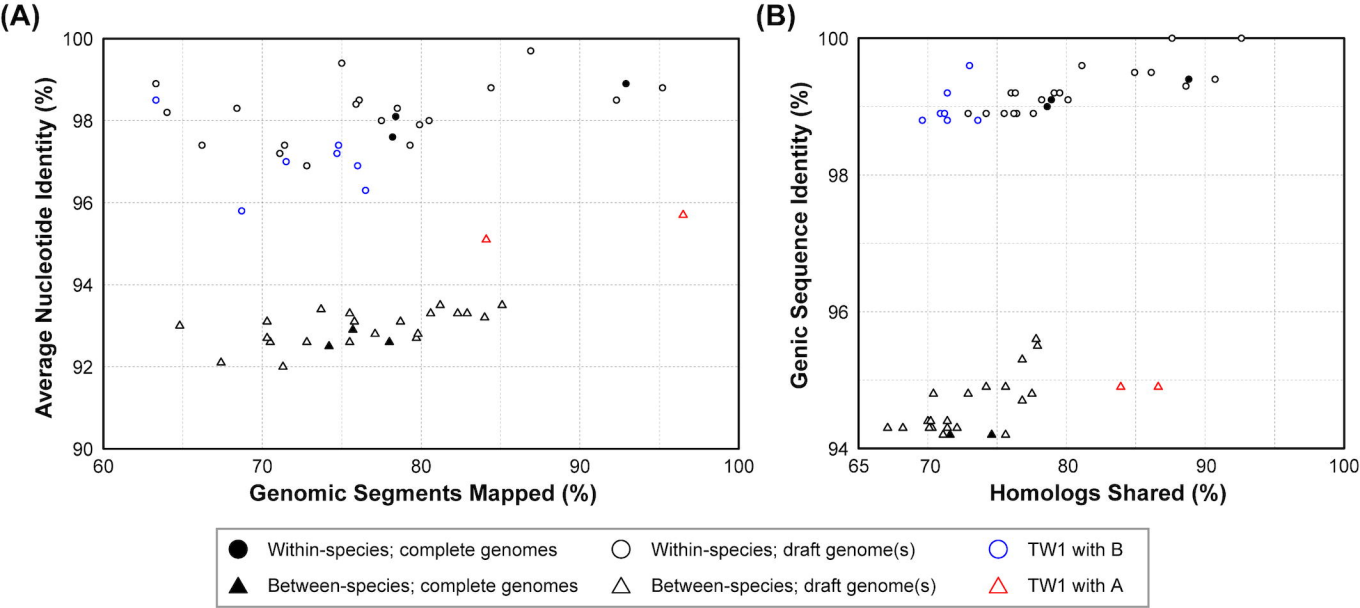
	AYWB	NJAY	WBD	OY-M	OY-V	DY2014	MBSP-M3	De Villa	LD1	CYP	TW1
AYWB	724	99.7	93.3	92.5	92.1	92.6	92.6	92.9	93.1	93.0	95.7
NJAY	86.9	652	93.5	92.6	92.0	92.6	92.7	92.8	92.8	93.1	95.1
WBD	82.3	81.2	611	93.3	92.7	93.1	93.2	93.3	93.5	93.4	93.3
OY-M	74.2	70.5	75.5	853	98.3	98.8	97.6	98.1	98.3	98.9	97.0
OY-V	67.4	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

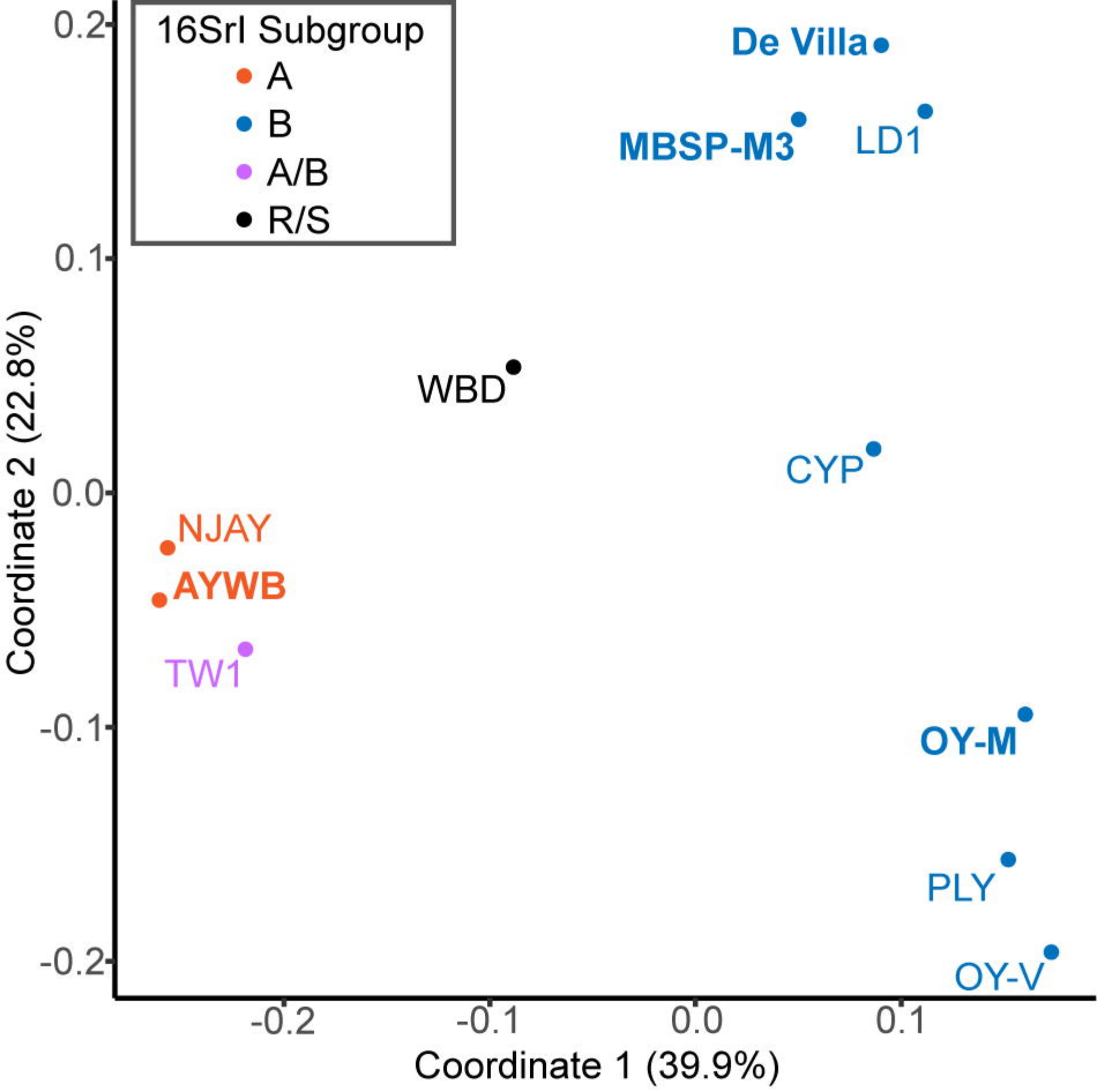
Above diagonal: Average Nucleotide Identity (%) 95 99
Diagonal: Genome Size (kb)
Below diagonal: Genomic Segments Mapped (%) 70 80

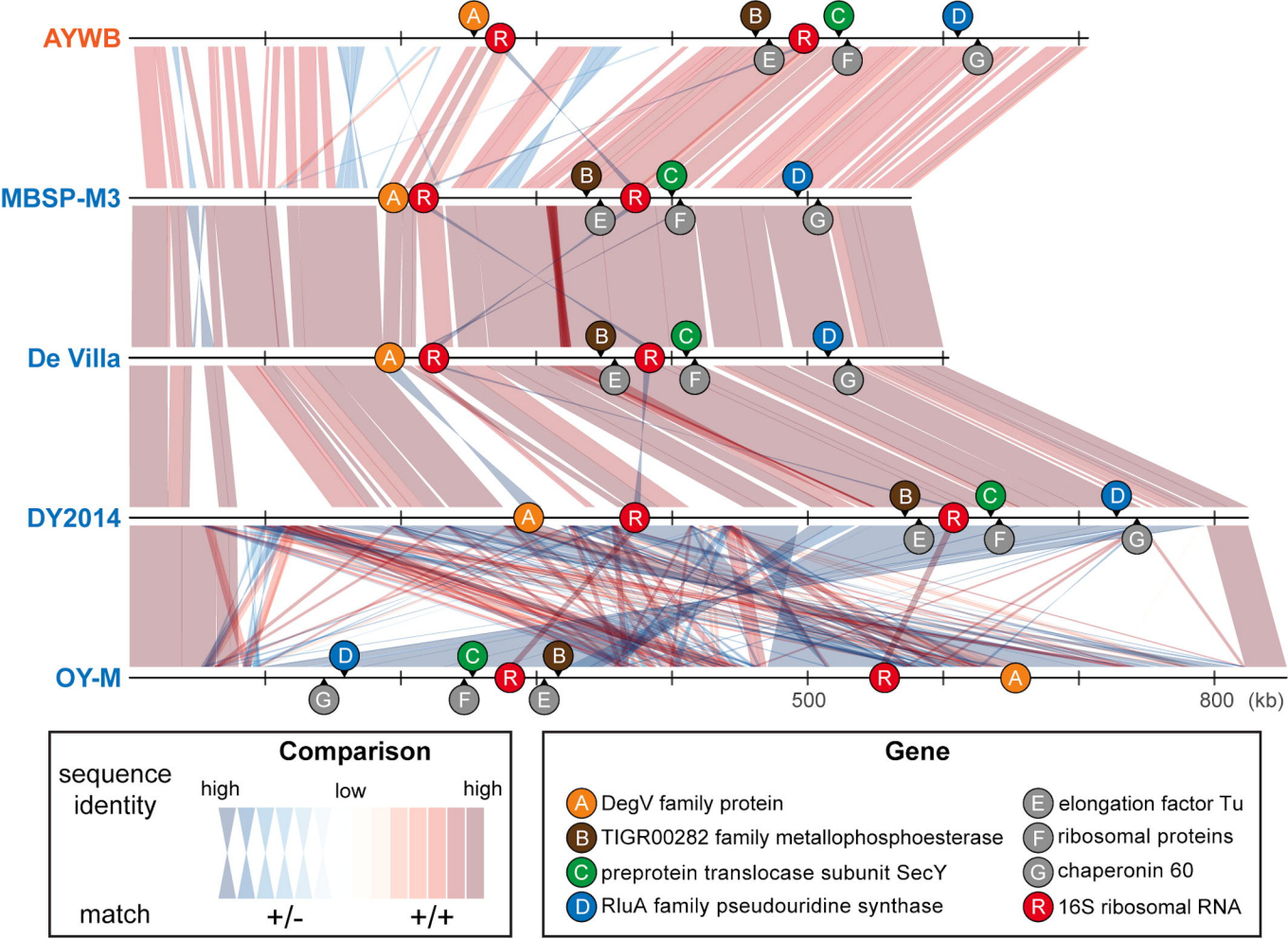
(B)

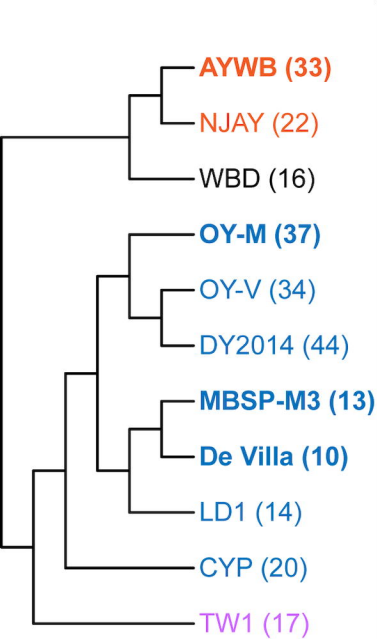
	AYWB	NJAY	WBD	OY-M	OY-V	DY2014	MBSP-M3	De Villa	LD1	CYP	TW1
AYWB	570	100.0	95.5	94.3	94.3	94.3	94.2	94.2	94.2	94.4	94.9
NJAY	92.6	551	95.6	94.4	94.3	94.3	94.2	94.3	94.3	94.4	94.9
WBD	77.9	77.8	524	94.9	94.8	94.8	94.7	94.8	94.7	94.9	95.3
OY-M	70.2	70.0	74.2	660	99.5	99.5	99.0	99.1	99.1	99.6	99.2
OY-V	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 (%)
Genomic Segments Mapped (%)
(B)
Genic Sequence Identity (%)
Homologs Shared (%)
Within-species; complete genomes
Within-species; draft genome(s)
TW1 with B
Between-species; complete genomes
Between-species; 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
PLY
OY-V
0.2
0.1
0.0
-0.1
-0.2
-0.2
-0.1
0.0
0.1
Coordinate 2 (22.8%)
Coordinate 1 (39.9%)


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



	SAP05	SAP11	SAP54/PHYL1	TENGU
AYWB (33)	AYWB_RS00160	AYWB_RS01770	AYWB_RS01055	AYWB_RS00785
NJAY (22)	BBA70_00185	BBA70_01750	BBA70_01070*	BBA70_00770
WBD (16)	N747_RS0100715	N747_RS0100020	N747_RS03525*	N747_RS0101805
OY-M (37)	PAM_RS03110	PAM_RS03115	PAM_RS00330	PAM_RS02860
OY-V (34)	OYV_RS02545	OYV_RS02510* OYV_RS02550	OYV_RS03895	OYV_RS02965*
DY2014 (44)	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 (20)	OL66_RS02400		OL66_RS03105	OL66_RS00470
TW1 (17)	DF280_02690	DF280_00985*	DF280_01620	DF280_01810

Table 1. List of the genome sequences analyzed.

Strain	**Accession**	**Assembly**	**Genome Size (bp)**	**CDS-intact**	**CDS-pseudo**	**Location**	**Host**	**Reference**
AYWB	GCF_000012225	Complete	723,970	557	116	USA	Lettuce	Bai et al., 2006
NJAY	GCA_002554195	75 contigs	652,092	371	262	USA	Lettuce	Sparks et al., 2018
WBD	GCF_000495255	6 contigs	611,462	471	76	China	Wheat	Chen et al., 2014
OY-M	GCF_000009845	Complete	853,092	728	144	Japan	Onion	Oshima et al., 2004
OY-V	GCF_000744065	170 contigs	739,592	651	217	Japan	Onion	Kakizawa et al., 2014
DY2014	GCA_005093185	8 contigs	824,596	775	63	Taiwan	Periwinkle	Cho et al., 2019b
MBSP-M3	GCF_001712875	Complete	576,118	498	34	Brazil	Maize	Orlovskis et al., 2017
De Villa	GCF_004214875	Complete	603,949	512	21	South Africa	Grapevine	Coetzee et al., 2019
LD1	GCF_001866375	8 contigs	599,264	536	29	China	Rice	Zhu et al., 2017
CYP	GCF_000803325	252 contigs	659,699	546	154	Italy	Daisy	Pacifico et al., 2015
TW1	GCA_003181115	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 (*rpl23*, *rpl2*, and *rps19*)	[rpF1] GGACATAAGTTAGGTGAATTT [rpR1] ACGATATTTAGTTCTTTTTGG	1,240	70	25	Lim and Sears 1991
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
Elongation factor Tu (*tuf*)	[fTufu] CCTGAAGAAAGAGAACGTGG [rTufu] CGGAAATAGAATTGAGGACG	842	39	23	Schneider and Gibb, 1997
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