Systematic and Applied Microbiology 42 (2019) 117–127

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: http://www.elsevier.com/locate/syapm

The genome of '*Candidatus* Phytoplasma solani' strain SA-1 is highly dynamic and prone to adopting foreign sequences

Martina Seruga Music[a,*], Ivana Samarzija[a], Saskia A. Hogenhout[b,*], Mindia Haryono[c], Shu-Ting Cho[c], Chih-Horng Kuo[c,*]

[a] *Department of Biology, Faculty of Science, University of Zagreb, Marulicev trg 9A, HR-10000 Zagreb, Croatia*
[b] *Department of Crop Genetics, John Innes Centre, Norwich Research Park, Colney Ln, Norwich NR4 7UH, UK*
[c] *Institute of Plant and Microbial Biology, Academia Sinica, 128 Sec. 2, Academia Rd., Taipei 11529, Taiwan*

ARTICLE INFO

Keywords:
Genome instability
Effector
Horizontal gene transfer
Potential mobile unit
Plant pathogen
Sequencing

ABSTRACT

Bacteria of the genus '*Candidatus* Phytoplasma' are uncultivated intracellular plant pathogens transmitted by phloem-feeding insects. They have small genomes lacking genes for essential metabolites, which they acquire from either plant or insect hosts. Nonetheless, some phytoplasmas, such as '*Ca.* P. solani', have broad plant host range and are transmitted by several polyphagous insect species. To understand better how these obligate symbionts can colonize such a wide range of hosts, the genome of '*Ca.* P. solani' strain SA-1 was sequenced from infected periwinkle via a metagenomics approach. The de novo assembly generated a draft genome with 19 contigs totalling 821,322 bp, which corresponded to more than 80% of the estimated genome size. Further completion of the genome was challenging due to the high occurrence of repetitive sequences. The majority of repeats consisted of gene arrangements characteristic of phytoplasma potential mobile units (PMUs). These regions showed variation in gene orders intermixed with genes of unknown functions and lack of similarity to other phytoplasma genes, suggesting that they were prone to rearrangements and acquisition of new sequences via recombination. The availability of this high-quality draft genome also provided a foundation for genome-scale genotypic analysis (e.g., average nucleotide identity and average amino acid identity) and molecular phylogenetic analysis. Phylogenetic analyses provided evidence of horizontal transfer for PMU-like elements from various phytoplasmas, including distantly related ones. The '*Ca.* P. solani' SA-1 genome also contained putative secreted protein/effector genes, including a homologue of SAP11, found in many other phytoplasma species.

Introduction

Phytoplasmas comprise a diverse group of wall-less pleiomorphic bacteria assigned to the genus '*Candidatus* Phytoplasma' within the class Mollicutes. Unlike their phylogenetically related species from the genus *Mycoplasma*, which are human and animal pathogens, phytoplasmas are plant pathogens transmitted by phloem-feeding insects. For most phytoplasma species, one of the hallmarks of their behaviour is the adaptability to different plant and insect hosts [25]. Worldwide, phytoplasma infections are causing significant economical losses in agriculture with phytoplasma-associated diseases being reported for hundreds of monocot and dicot plant species [42]. Since phytoplasmas cannot be successfully grown outside their hosts as in vitro cultures, these bacteria have been assigned a '*Candidatus* (*Ca.*)' taxonomic status [18] and they encompass more than thirty 16Sr groups and numerous subgroups based on 16S rRNA gene sequences [76]. However, based on their distinctive genomic, phylogenetical, ecological and nutritional characteristics, phytoplasma could be assigned to their own order and family [77].

In the past decade, genome sequence data have provided insights into phytoplasma biology [1,4,38,55,57,71]. Despite their small size compared to other bacteria, the repeat-rich nature of phytoplasma genomes has often complicated whole genome assemblies, leading to incomplete genome sequences [11,12,19,43,51,60,64,74]. Given that phytoplasmas cannot be obtained in pure culture, a range of approaches have been used to isolate phytoplasma DNA from plant and insect host tissues,

* Corresponding authors.
E-mail addresses: martina.seruga.music@biol.pmf.hr (M.S. Music), ivana.samarzija@biol.pmf.hr (I. Samarzija), Saskia.Hogenhout@jic.ac.uk (S.A. Hogenhout), mindia.adinda@gmail.com (M. Haryono), vivianlily6@hotmail.com (S.-T. Cho), chk@gate.sinica.edu.tw (C.-H. Kuo).

https://doi.org/10.1016/j.syapm.2018.10.008

for example, enrichment of phytoplasma DNA via cesium chloride density gradient centrifugation and pulse-field gel electrophoresis, which has led to the completion of whole-genome assemblies of '*Ca.* P. asteris' strains OY-M and AY-WB [4,57], as well as strains of '*Ca.* P. australiense' [1,71] and '*Ca.* P. mali' [38]. With sequencing becoming cheaper, it has also become possible to use metagenomics approaches by sequencing entire phytoplasma-carrier insect vectors or infected plant tissues [12]. For example, a complete genome assembly of '*Ca.* P. asteris' MBSP strain M3 and resequencing of additional strains was achieved by sequencing the phytoplasma-carrier leafhoppers *Dalbulus maidis* [55]. However, insect vectors for the majority of phytoplasmas are unknown or may be difficult to rear in captivity. Fortunately, many phytoplasmas successfully infect Madagascar periwinkle (*Catharanthus roseus* L. Don) [32] which can easily be grafted and grown in in vitro, enabling maintenance of phytoplasma without the insect vector [7,23]. This approach has been particularly beneficial for maintaining phytoplasma pathogens of grapevine, including '*Ca.* P. solani' SA-1, the subject of this study.

'*Ca.* P. solani' belongs to the 16SrXII ribosomal group and is closely related to '*Ca.* P. graminis' (16SrXVI) and '*Ca.* P. caricae' (16SrXVII) (Supplementary material Fig. S1 in the online version at DOI: 10.1016/j.syapm.2018.10.008). The most closely related phytoplasma with an available genome sequence is '*Ca.* P. australiense' [1,71]. '*Ca.* P. solani' is the causal agent of 'bois noir' (black wood) of grapevine and stolbur disease in wild and cultivated herbaceous and woody plants, and several other diseases of, for example, maize, tomato, pepper and strawberry in Europe and Mediterranean basin. This phytoplasma has a wide plant host range and is transmitted by several phloem-feeding insect species. However, different insect life stages have differential preference for specific plant parts [49], which complicates the ability to rear these insects in captivity, certainly on grapevines.

Recently, it has been shown that a number of putative virulence factors, named effectors, are encoded in the phytoplasma genome [3,26] which modulate processes in both plant hosts and insect vectors [48]. Phytoplasmas, being intracellular parasites, secrete these effectors via the Sec-dependent translocation system directly into the cell's cytoplasm, including phloem sieve cells [29]. Presently, there is a great interest in elucidating how phytoplasma effectors modulate plant processes and among candidate effector proteins identified in phytoplasma genomes, several have been very well studied. For instance, SAP11 of AY-WB and its homologues identified in other phytoplasmas interact with plant class II CIN-TCP transcription factors. This leads to the changes in leaf development, stem proliferation and downregulation of plant defence responses. SAP54 of AY-WB binds and degrades MADS-box transcription factors, which play fundamental roles in flowering [47]. These activities of the effectors are likely potentially important, because phytoplasmas require insects for spread.

The present study reports a draft genome assembly of '*Ca.* Phytoplasma solani' strain SA-1, which originated from infected grapevines and was transferred to and maintained in Madagascar periwinkle. The main goal of this study was to provide a better understanding of the genomes of '*Ca.* P. solani' species and gather new knowledge on how these obligate and uncultivated symbionts can colonize such a wide range of eukaryotic hosts.

Materials and methods

Plant samples

Total genomic DNA was extracted by modified CTAB procedure as described previously [65] from 0.5 g of Madagascar periwinkle (*C. roseus* L. Don.) tissue infected with '*Ca.* Phytoplasma solani' strain SA-1 originating from grapevine. Periwinkle was maintained by micropropagation in vitro and was obtained from the phytoplasma micropropagation collection at the DiSTA, University of Bologna, Italy (http://www.ipwgnet.org/doc/phyto_collection/collection-august2010.pdf). The concentration and purity of DNA was determined by using a Qubit® 1.0. fluorometer (ThermoFisher Scientific, USA).

Library preparation and sequencing

Three Illumina shotgun paired-end libraries with a target insert size of approx. 550 bp were constructed from total genomic DNA extracts and sequenced using the Illumina MiSeq platform (Illumina). The Illumina sequencing library preparation and MiSeq sequencing services were provided by The Genome Analysis Centre (Norwich, UK) and the core facilities in Academia Sinica (Taipei, Taiwan).

Assembly and annotation

The approach and procedures used for the assembly and annotation of the genome were based on those described previously [11,12] with some adjustments. All bioinformatics tools were used with the default settings unless stated otherwise. In summary, the *de novo* assembly was performed by using Velvet package version 1.2.10 [75]. For identification of putative phytoplasma contigs, BLASTx [8] searches against nonredundant NCBI database [6] were performed, followed by mapping of raw reads using the Burrows-Wheeler Alignment (BWA) tool version 0.7.12 [44] and visual inspection using the Integrative Genomics Viewer (IGV) version 2.3.57 [63]. Scaffolding and gap closure by PCR was attempted but unsuccessful. To estimate the completeness of this draft assembly, the contigs without annotation were used as the input for CheckM [58], which was executed with the "–reduced_tree" option to reduce memory usage.

The prediction of rRNA, tRNA and protein-coding genes was done by using RNAmmer [40], tRNAscan-SE [46] and Prodigal [27], respectively. Homologous gene identification and annotation was based on the genes from other phytoplasma genomes such as PnWB [12] and AY-WB [4], as identified by OrthoMCL [45]. Additional manual curation was performed based on BLASTp searches against the NCBI nonredundant database and the KEGG database [30].

The putative secreted proteins were predicted based on the procedure described in Bai et al. [3]. Briefly, SignalP v3.0 [5] was used to predict the presence of signal peptide. Based on the HMM method, the positive candidates were defined as those with a signal peptide presence probability of >0.5 and a N-terminal signal peptide between 20 and 50 amino acids in length. For each candidate, the mature protein sequence without the signal peptide was processed using TMHMM v2.0 [35] to identify the presence of transmembrane domains. Those without any transmembrane domain were manually examined to remove candidates with other well-annotated functions (e.g., ABC transporter) and they were annotated as putative secreted proteins in the final annotation. In a search for protein domains of effectors and PMU-associated genes, the MOTIF Search web tool (http://www.genome.jp/tools/motif/) was used.

This '*Ca.* P. solani' SA-1 whole genome shotgun sequencing project has been deposited at DDBJ/ENA/GenBank under the accession MPBG00000000. The version described in this paper is version MPBG01000000.

For visualization, the GC content, GC skew, and general features of SA-1 chromosome (Fig. 1) were depicted using DNAplotter software [9]. The genome alignment (Supplementary material Fig. S2 in the online version at DOI: 10.1016/j.syapm.2018.10.008) was generated using the Artemis Comparison Tool (ACT) [10].

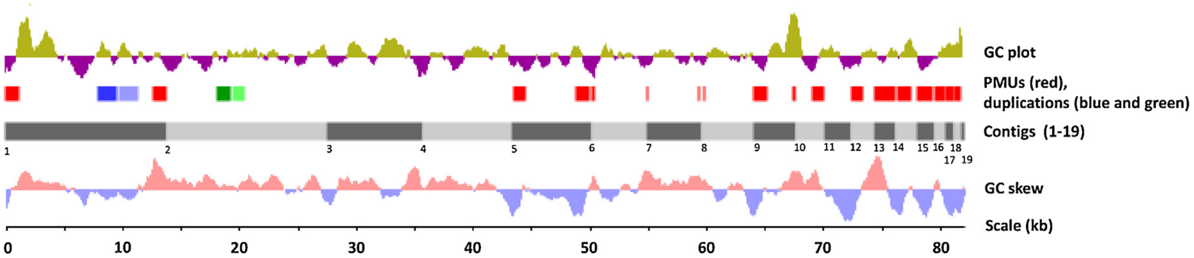

Fig. 1. Schematic view of '*Ca.* P. solani' SA-1 genome draft assembly of 821,322 bp.

Functional characterization of protein–coding genes and comparison of orthologous gene clusters

For functional characterization of all protein-coding genes, annotation by using KAAS tool [53] supplied by the KEGG database [30] was performed, followed by mapping to the COG functional category assignment [68] in order to produce summary statistics (Supplementary material Fig. S3 in the online version at DOI: 10.1016/j.syapm.2018.10.008). Genes without any COG functional category assignment were assigned to a custom category X.

OrthoVenn software (http://aegilops.wheat.ucdavis.edu/OrthoVenn/) [73] based on a modified OrthoMCL algorithm was used to compare orthologous clusters between representative phytoplasma strains.

Phylogenetic analyses, average nucleotide identity (ANI) and average amino acid identity (AAI) calculation

Multiple alignment of nucleotide and amino-acid sequences was done by using ClustalX [41]. Phylogenetic analyses were performed and phylogenetic trees inferred by maximum-likelihood method with different models and parameters (for example Tamura-Nei model and Gamma-distribution with invariant sites) using MEGA7 [67] and PhyML [21] using Geneious® 10.1.3. software (Biomatters Ltd., Auckland, New Zealand). Bootstrapping with 500 replicates was used to test the reliability of inferred trees.

In the genome-wide approach, for ANI and AAI calculation, the homologous genes were identified using OrthoMCL [45] and the multiple sequence alignment was based on MUSCLE [15]. The sequence identities were calculated using PHYLIP [17]. The subsequent whole-genome and 16S rDNA phylogeny from Fig. 2 was based on PhyML [22].

Results and discussion

Genome assembly and general features of 'Ca. *P. solani' strain SA-1 genome*

A general characteristic of cultivated members of the class Mollicutes, such as mycoplasmas, spiroplasmas and acholeplasmas is a considerable variation in genome size [62]. Uncultivated phyoplasmas also seem to posses the same characteristics, as shown by the estimation of chromosome size by PFGE [50]. Among the phytoplasmas, a significant chromosome size heterogeneity ranging from 860 to 1350 kbp has also been observed for strains of the 16SrXII group (stolbur; '*Ca.* P solani') [50]. In this study, the Illumina shotgun sequencing at approximately 1,6 Gb raw reads provided ~110-fold coverage of the '*Ca.* P. solani' SA-1 genome, which was assembled into 19 contigs with a total size of 821,322 bp (Table 1, Supplementary Table S1 in the online version at DOI: 10.1016/j.syapm.2018.10.008, Fig. 1). Based on a previous estimation by PFGE, the chromosome size of this bacterium is 1020 kbp [50]. Hence, this draft assembly probably covers ~81% of the chromosome. However, this completeness is likely an underestimate because PFGE typically overestimates the genome size by 10–15% [72]. Using an alternative method for assessing the assembly quality based on the presence of conserved marker genes, CheckM [58] estimated the completeness of this draft assembly at 97.8% and possible contamination at 5.8%. However, due to the inherent bias of using a small set of marker genes that do not necessarily have a uniform distribution across the chromosome and the under-representation of complete phytoplasma genomes in the current database, these estimates are most likely overestimates, therefore, these numbers must be considered with caution.

The contig length ranged from 1801 to 138,876 bp and the N50 value of SA-1 draft assembly was 76,256 bp, as compared to 9757 and 4036 bp in other available draft genomes of '*Ca.* P. solani' strains 284/09 and 231/09, respectively (Table 1). A total of 709 full-length coding sequences (CDSs) were annotated, with 452 of them having an assigned function and 257 annotated as hypothetical proteins (Table 1). The number of CDSs from this SA-1 draft assembly was comparable to those of fully sequenced phytoplasma genomes such as closely related '*Ca.* P. australiense' PAa strain (839 CDSs; Table 1), as well as distantly related '*Ca.* P. asteris' strains AY-WB (675 CDSs; [4]) and OY-M (754 CDSs; [57]). A considerably higher number of tRNA genes and both rRNA operons were found in SA-1 genome, whereas in 284/09 and 231/09 '*Ca.* P. solani' strains no rRNA operons were annotated (Table 1) [51]. A high proportion of the CDSs (18.5% of total genome length) in the SA-1 strain draft genome consisted of multicopy genes lying within PMU-like regions (Fig. 1) which were higher compared to other phytoplasma genomes, including '*Ca.* P. asteris' strains OY-M (14,1%) [57] and AY-WB (10.2%) [4], as well '*Ca.* P. australiense' PAa (12.1%) [71]. In comparison, '*Ca.* P. mali' and MBSP have lower numbers of CDSs (497 and 531, respectively) and smaller genomes (601,943 bp and 576,118 bp) with fewer repeats [38,55]. Moreover, '*Ca.* P. mali' and MBSP are specialists (have only one plant host and are transmitted by one insect genus), while '*Ca.* P. solani' and '*Ca.* P. australiense' are generalists (have a broad host range and are transmitted by polyphagous insects). Therefore, to date, the genome size and numbers of PMU-like regions are positively correlated to the numbers of plant hosts and insect vector species of phytoplasma.

As also shown for some of the sequenced phytoplasma genomes such as AY-WB, OY-M and MBSP [4,55,57], the '*Ca.* P. solani' SA-1 genome sequence had an irregular GC skew pattern (Fig. 1). This feature, together with the presence of insertion sequences or repeats, such as PMUs, appeared to be characteristic of high genome plasticity. Furthermore, the SA-1 genome had a low GC content (28.3%; Table 1 and Fig. 1), which is a common feature among most members of the class Mollicutes [62]. The GC content was comparable between three '*Ca.* P. solani' strains and related '*Ca.* P. australiense' (Table 1), but higher than in more divergent phytoplasmas such as PnWB and '*Ca.* P. mali' [12,38].

In order to compare the genome organization of SA-1 and related phytoplasmas, genome alignments were created, which revealed little synteny conservation with both '*Ca.* P. australiense' PAa and

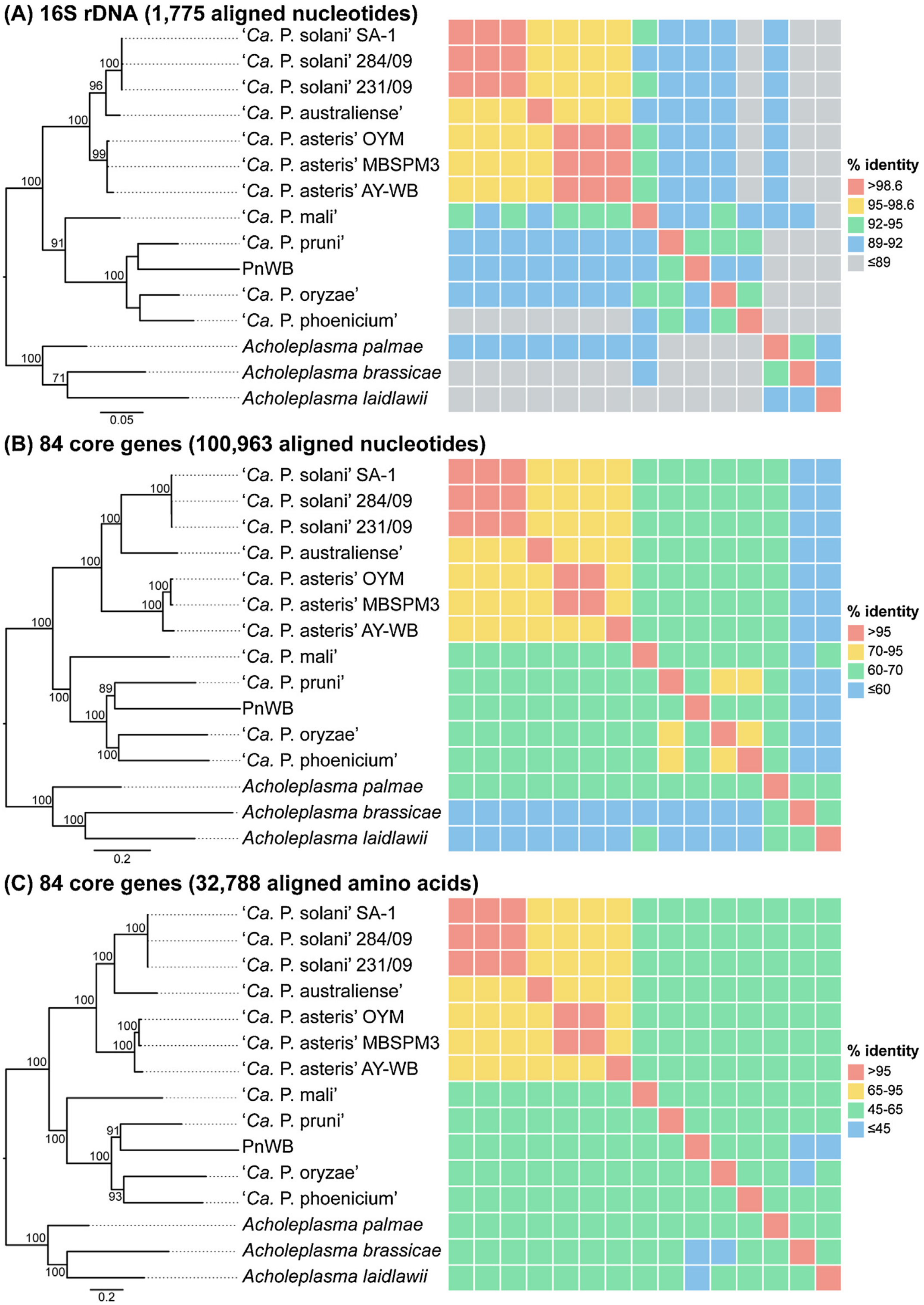

Fig. 2. Molecular phylogeny based on 16S rDNA (A) nucleotide sequences of the core genes (B) and amino acid sequences of the core genes (C). The *Acholeplasma* species are included as outgroups to root the tree. The numbers on branches indicate the level of bootstrap support (based on 1000 replicates; only values above 70% are shown). The scale bar represents the number of substitutions per site. The heatmaps on the right-hand side are colored based sequence identity. The thresholds are based on those suggested for taxonomic assignment to the same species (red), genus (yellow), family (green), or higher ranks (blue and grey) [33]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

'*Ca.* P. solani' 284/09 assemblies (Supplementary material Fig. S2 in the online version at DOI: 10.1016/j.syapm.2018.10.008). However, since both SA-1 and 284/09 are draft assemblies organized into 19 and 128 contigs [51], respectively, this could have considerably influenced the results of genome alignments. Nevertheless, genome alignment results indicated that there were significant genome rearrangements and biological divergence between '*Ca.* P. solani' and '*Ca.* P. australiense' (Supplementary material Fig. S2 in the online version at DOI: 10.1016/j.syapm.2018.10.008) even though these two phytoplasmas both belong to the 16Sr group XII

Table 1
General features of ‘*Ca.* P. solani’ SA-1 genome and comparison to related phytoplasma genomes.

Characteristic	’*Ca.* P. solani’			’*Ca.* P. australiense‘ PAa
	SA-1	284/09	231/09	
No. of contigs	19	128	298	1
Combined length (bp)	821,322	557,538	515,758	879,324
Average contig length (bp)	43,227	4,356	1,731	N/A
N50 (bp)	76,256	9,757	4,036	N/A
G + C content (%)	28.3	28.2	28.6	27
No. of CDSs:				
Complete CDSs	709	448	346	839
Partial CDSs	0	72	227	N/A
CDSs with assigned function	452	366	404	502
CDSs annotated as hypothetical proteins	257	154	169	337
No. of tRNA genes	32	27	8	35
No. of rRNA operons	2	Not found	Not found	2
GenBank accession number	MPBG01000000	FO393427	FO393428	AM422018

N/A = not applicable.

(Supplementary material Fig. S1 in the online version at DOI: 10.1016/j.syapm.2018.10.008) and share more than 97,5% of 16S rRNA gene sequence identity [61]. This is an interesting point regarding phytoplasma evolution since both species have a broad host range and are associated with similar diseases; however, ‘*Ca.* P. australiense’ is restricted to Australia and New Zealand while ‘*Ca.* P. solani’ mainly resides in Europe (it is sporadically found in Asia with no defined vectors).

In spite of our additional efforts aimed to complete the SA-1 genome, it was not possible to perform the gap closure successfully. Therefore, it is speculated that several reasons were responsible for the incompleteness of this genome, one of which was certainly obvious, since there was a very high presence of large repeats that impeded the assembly. Another limitation for the ‘*Ca.* P. solani’ SA-1 strain assembly pipeline was the fact that there were no other more closely related phytoplasmas with complete genome sequence available. Hence, only contigs with protein-coding genes that had identifiable sequence similarities with known phytoplasma proteins were considered. Consequently, some regions containing ‘*Ca.* P. solani’-specific protein-coding genes that are absent in all other phytoplasmas might have been missed. Moreover, due to the fact that for Madagascar periwinkle, only plastid genome is sequenced [36], some phytoplasma reads might have also been excluded as possible contamination from the plant host. Heterogeneous populations of the SA-1 phytoplasma consisting of several genotypes present in an infected periwinkle host, each of them possibly having a different genome rearrangement, could have also prevented the successful assembly of the genome into a single circular contig. It should also be taken into account that a long-term micropropagation (more then 10 years) and maintenance in periwinkle would also contribute to diversification of the phytoplasma population within the host. A genetically more uniform phytoplasma populations may be obtained by generating a bottleneck, such as by conducting insect vector transmission using short acquisition and inoculation times. Hence, this strategy should be considered in future efforts towards the completion of the ‘*Ca.* P. solani’ genome assembly.

Molecular phylogeny and genomic approach to taxonomy

The genomics era has provided new opportunities for determining the taxonomy of phytoplasmas. The number of studies on a genome-wide based bacterial taxonomy are increasing [69] and the whole-genome taxonomic aspect has been shown to be useful for characterising complex evolutionary relationships for different bacteria [69]. The approach based on genome-wide average nucleotide identity (ANI) and average amino-acid identity (AAI) values has been suggested as a concept that could contribute significantly to a genome-based taxonomy for microbial organisms [34], particularly for uncultivated ones [33]. However, the sequence divergence rates are known to vary extensively among bacteria [39].

With the availability of genome sequences in the current study, the molecular phylogeny based on 16S rDNA was compared to that inferred using a set of 84 core genes conserved among all characterized phytoplasmas (Fig. 2). The core-gene phylogeny provided better resolution for resolving close relationships (e.g., among the ‘*Ca.* P. asteris’ strains), as well as stronger bootstrap support for all branches. Based on the sequence identity thresholds proposed previously for taxonomic assignments [33], all of the three strains of ‘*Ca.* P. solani’ could be confidently assigned to the same species using either the 16S rDNA sequence or ANI/AAI values (Fig. 2 and Supplementary Table S2 in the online version at DOI: 10.1016/j.syapm.2018.10.008). In contrast, although the three strains of ‘*Ca.* P. asteris’ all had >98.6% nucleotide sequence identity for their 16S rDNA sequences, the ‘*Ca.* P. asteris’ AY-WB had ANI and AAI values below the species-level threshold. Moreover, based on these proposed thresholds, only the species within the phytoplasmas subclade I [12,24] (i.e., ‘*Ca.* P. asteris’, ‘*Ca.* P. australiense’, and ‘*Ca.* P. solani’) would be assigned to the same genus, whereas those belonging to subclade II (i.e., ‘*Ca.* P. mali) and subclade III would be far too divergent to be considered as belonging to the same genus (Fig. 2, Supplementary material Fig. S1 in the online version at DOI: 10.1016/j.syapm.2018.10.008). Thus, in this study, it was demonstrated that the sequence divergence rates were extremely high among phytoplasmas and their relatives (i.e., *Acholeplasma*). These findings clearly illustrated the difficulties of proposing a set of universal guidelines for taxonomy based on molecular sequences. Nevertheless, the increasing availability of genome sequences could certainly better inform taxonomy updates. For example, a recent paper raised the question whether the genus “*Ca.* Phytoplasma” should be retained in the order Acholeplasmatales or moved to a novel provisional order and family [77]. The results reported in this study, as well as those from future genomic characterizations that would further improve the taxon sampling, could provide quantative data sets for such taxonomy reassessments.

Functional classification of genes and comparative analysis of gene content

Functional classification of annotated genes into COG categories revealed that the highest number of annotated genes from SA-1 genome belong to information storage and processing (subcategories (J) translation, ribosomal structure and modification; (K)

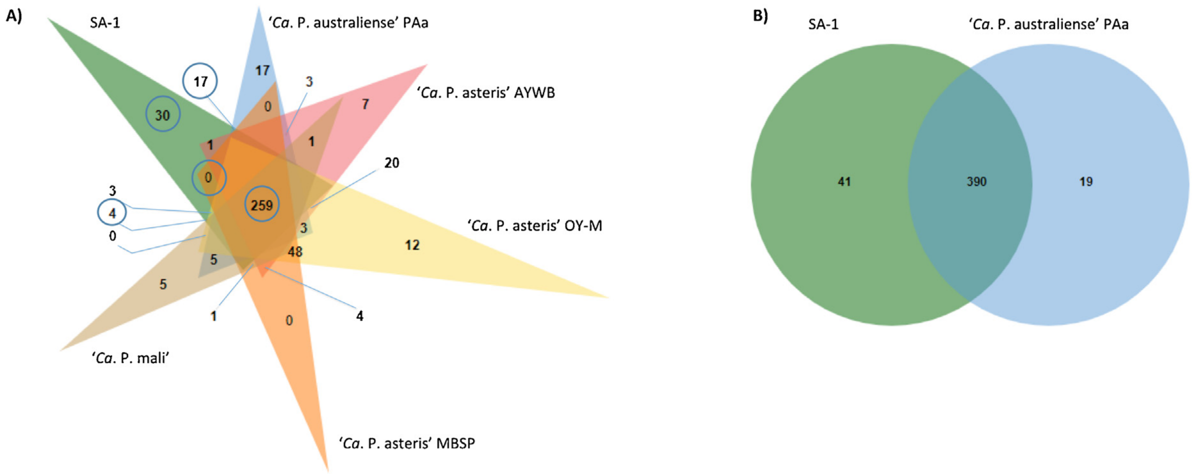

Fig. 3. Intersection of orthologous gene clusters in '*Ca.* P. solani' SA-1, '*Ca.* P. australiense' PAa, '*Ca.* P. asteris' AY-WB, OY-M, MBSP and '*Ca.* P. mali' (A) and SA-1 and '*Ca.* P. australiense' PAa (B) genomes. Numbers discussed in manuscript are encircled on Fig. 3A.

transcription and (L) replication, recombination and repair) and metabolism (subcategories (C) energy production and conversion; (G) carbohydrate transport and metabolism; (E) aminoacid transport and metabolism; (F) nucleotide transport and metabolism, (H) coenzyme transport and metabolism; (I) lipid transport and metabolism; (P) inorganic ion transport and metabolism and (Q) secondary metabolites biosynthesis, transport and catabolism) (Supplementary material Fig. S3 in the online version at DOI: 10.1016/j.syapm.2018.10.008). The cell process and signalling category was represented by a lower number of annotated genes, while a considerably higher number of genes belong to the poorly characterized category (49,72%; Supplementary material Fig. S3 in the online version at DOI: 10.1016/j.syapm.2018.10.008). This is in agreement with another published phytoplasma genome analyses [12], although pseudogenes were not included in this analysis.

Comparative analyses of gene contents among '*Ca.* P. solani' strain SA-1 and five completely sequenced phytoplasma genomes ('*Ca.* P. australiense PAa', '*Ca.* P. mali' and '*Ca.* P. asteris' strains AY-WB, OY-M and MBSP strain M3) identified a total of 259 orthologous gene clusters shared by all analyzed phytoplasmas (Fig. 3). These conserved clusters include genes belonging mainly to metabolism and information storage and processing COG categories. However, 29 out of 259 (±10%) genes were annotated as hypothetical proteins and did not have similarities to sequences beyond phytoplasma, indicating that these shared genes might be phytoplasma-specific. The analyses also identified 30 clusters (comprising 87 genes) that were found exclusively in the SA-1 genome (Fig. 3), together with 82 singletons annotated as hypothetical proteins or putative effectors often located within PMU-like regions. A substantially high percentage of SA-1 specific genes in the SA-1 draft genome (23,84%; 169/709) was in agreement with previous studies where it was shown that genome-specific genes in PnWB accounted for 22–30% of the gene content [12]. Thus, this finding further corroborated the significance of a high number of genome-specific genes in diversification of phytoplasmas and the adaptability to different hosts.

Interestingly, four gene clusters were shared solely by SA-1 and the distantly related '*Ca.* P. mali' genome (Fig. 3). Genes lying within these clusters are annotated as hypothetical proteins including a gene for a RecA protein, which is involved in DNA repair mechanism and recombination. In addition, 17 orthologous gene clusters were found to be unique to the '*Ca.* P. australiense' PAa and SA-1 genomes (Fig. 3A). These clusters contained PMU-related *dnaB* and *tra5* genes, as well as genes for hypothetical proteins together with a gene annotated as RibF encoding riboflavin biosynthesis protein. When comparative analysis including only the '*Ca.* P. solani' SA-1 strain and '*Ca.* P. australiense' PAa was performed, it revealed that 390 orthologous clusters were shared between these genomes (Fig. 3B). These include genes involved in all cellular processes and signalling, metabolism, information storage and processing, as well as 102 gene clusters of genes encoding hypothetical proteins. At least three of these encompassed proteins annotated as putative effectors and located within PMU-like regions. Additional comparative analyses including '*Ca.* P. solani' strains 284/09 and 231/09 showed the presence of 24 gene clusters (comprising 74 genes) together with 50 singletons that were only found to be specific to the SA-1 phytoplasma (not shown). However, this number may have been overestimated due to the fact that the 284/09 and 231/09 draft genomes had a considerable amount of partial CDSs (Table 1) and covered a smaller proportion of the estimated '*Ca.* P. solani' genome size. Nevertheless, the possibility for presence of significant number of strain-specific genes within '*Ca.* P. solani' species cannot be excluded, considering its broad host range and polyphagous insect vectors.

In spite of a relatively high number of genome-specific genes in SA-1 when compared to the other '*Ca.* P. solani' strains and closely related '*Ca.* P. australiense', almost no differences were found in the presence of the main phytoplasma metabolic pathways. An overview of main metabolic pathways is shown on Supplementary material Fig. S4 in the online version at DOI: 10.1016/j.syapm.2018.10.008 and it was noticeable that all three phytoplasmas compared shared the main metabolic pathways. However, a *sucP* pseudogene (encoding only C-terminal domain of sucrose phosphorylase) sharing 100% sequence ID with those found in the 284/09 and 231/09 draft genomes was identified in the SA-1 genome, whereas '*Ca.* P. australiense' had a gene encoding the complete SucP protein of 486 amino acids. Interestingly, the *sucP* gene was absent from the genomes of '*Ca.* P. asteris' AY-WB and MBSP, as well as '*Ca.* P. mali', while in '*Ca.* P. asteris' OY-M a *sucP* pseudogene was found, which contained a frameshift mutation resulting in an early stop codon and two shorter ORFs. Whether there is another, possibly functional copy of the *sucP* gene in the SA-1 genome remains to be elucidated. Nevertheless, it is possible that the problem of entry into glycolysis is solved by uptake of phosphorylated hexose, as previously suggested [37]. Several glycolytic genes were annotated in the SA-1 genome, that belonged both to the upper part of the glycolysis (*pgi*, *pfkA*, *fbaA*, *tpiA*) and to the energy yielding part (*gapA*, *pgk*, *gpmI*, *eno*, *pyk*). Furthermore, genes related to the suggested alternative pathway from malate (*citS*, *sfcA*, *pdhA*, *pdhB*, *pdhC*, *pdhD*, *pduL* and *ackA*) were also found. Another interesting finding was that the SA-1 genome contained two duplications of 17 kbp (18 genes) and 11,5 kbp (7 genes) each, located on the first contig and the second contig, respectively (Fig. 1). Most of the genes found in these duplications were metabolic genes, together with the genes involved in information storage and processing No duplications were reported in the draft genomes of '*Ca*. P. solani' 284/09 and 231/09 and '*Ca*. P. australiense' PAa. Nonetheless, both regions duplicated in the SA-1 genome were present and partially conserved among four compared phytoplasmas (Supplementary material Fig. S5 in the online version at DOI: 10.1016/j.syapm.2018.10.008). A duplication of 30 kbp containing metabolic genes was previously reported for the '*Ca*. P. asteris' strain OY-W where the authors suggested that the more aggressive strain OY-W, unlike milder strain OY-M, used these genes for better consumption of the carbon source [56]. A similar possibility for the SA-1 phytoplasma can be envisaged since the *plsX* gene is involved downstream in glucose metabolism and the *pdh* genes are involved in pyruvate metabolism.

Comparative analysis of genes involved in replication, DNA modification and structure and DNA repair revealed the existence of these gene sets in '*Ca*. P. solani' SA-1 and 284/09 as well as closely related '*Ca*. P. australiense' [51,71]. All strains encoded five subunits of DNA polymerase III (alpha, beta, delta, delta', gamma/tau) and the excision repair complex *uvrABC*, according to KAAS annotation [51,53,71]. Like the majority of other sequenced phytoplasmas [37], the SA-1 phytoplasma did not possess genes encoding RuvA and RuvB proteins, but encoded for RuvX (Holliday junction resolvase-like protein). Surprisingly, both SA-1 and 284/09 phytoplasmas have a *recA* gene [51], while majority of other sequenced phytoplasmas lack this gene [37].

Since phytoplasmas are wall-less bacteria, their membrane proteins are in direct contact with the environment and consequently show a wide diversity as they are submitted to a positive diversifying selective pressure. The variable surface protein VMP1, the antigenic protein StAMP (AMP homologue) are highly variable among '*Ca*. P. solani' strains [13,16]. Genes for these two membrane proteins and the housekeeping genes *secY* and *tuf* are commonly used in the multilocus sequence typing (MLST) scheme to genotype and assess diversity of '*Ca*. P. solani' strains [14,54,59] and were also identified in the SA-1 genome. Moreover, all the genes of the Sec-dependent machinery for the protein export (including effectors) across the phytoplasma plasma membrane, and *yidC* involved in membrane protein configuration inside the phytoplasma membrane were also identified in the SA-1 phytoplasma genome.

Effectors

In the SA-1 draft genome, 38 putative secreted proteins/effectors were predicted (Supplementary Table 3 in the online version at DOI: 10.1016/j.syapm.2018.10.008), with 20 of these genes lying within PMU-like regions. Interestingly, as revealed by the BLAST search, for five of the SA-1 predicted putative secreted proteins, no similar nucleotide or protein sequences were found in other phytoplasma genomes. Three putative secreted proteins (PSSA1_v1c1220, PSSA1_v1c4850, and PSSA1_v1c6880) were found solely in '*Ca*. P. solani' strains 231/09 and 284/09 that shared 96–99% sequence identity. Moreover, homologues of the *SAP11* and *SAP21* genes of '*Ca*. P. asteris' AY-WB have also been identified [4]. SAP11 homologues had also been identified previously in genomes of several phytoplasmas, including '*Ca*. P. asteris' strains OY-M and MBSP, '*Ca*. P. mali', and '*Ca*.P. aurantifolia' strain PnWB [12,38,55,57], but intriguingly not in '*Ca*. P. australiense' [28,71]. The *SAP11*-like genes of SA-1, 284/09 and 231/09 were complete and 100% identical to SA-1 *SAP11* homologue. Moreover, the genomic regions were identical among the strains. However, in the 284/09 and 231/09 draft genomes they were annotated as conserved hypothetical proteins rather than putative effectors.

In a search for effector domains, several potentially interesting sequences were found that might have contributed to the putative effector action. For example, effector PSSA1_v1c0520 had a domain of a TPR/MLP1/MLP2-like protein implicated in nuclear protein import. This is of great interest, since, as mentioned previously, it is known that main phytoplasma effectors, such as SAP11 and SAP54, can be located in the nucleus of the host cell and can interact with host transcription factors [47,66]. Furthermore, SA-1 SAP11 and SAP21 homologues contained a SVM protein domain characteristic of phytoplasmas that is cleaved off when the effectors are secreted. Another interesting domain found in PSSA1_v1c5140 protein was a ryanodine receptor domain that participates in different signalling pathways involving calcium release from intracellular organelles. The same putative effector contained a sigma factor regulator N-terminal domain that interacts with sigma factor that has been suggested to be a key regulator of host switching in phytoplasmas [52]. Moreover, protein PSSA1_v1c5600 had an RtcR domain that is a sigma54-dependent enhancer binding protein directing the transcription of a wide variety of genes. Protein PSSA1_v1c6590 contained an amino acid permease domain involved in the transport of amino acids into the cell, while one of the proteins found only in SA-1 strain (PSSA1_v1c1140) had a PHAT domain involved in RNA binding. All these findings indicated that PMUs are likely to have a role in the virulence of phytoplasmas and suggested that future studies are needed in order to elucidate the functions of the potentially interesting effectors mentioned and their interaction with phytoplasma hosts.

Molecular evolution of potential mobile units

The unique characteristic of the phytoplasma genomes is the existence of potential mobile units (PMUs), putative transposon-like elements that are thought to contribute to the genome instability observed in these bacteria [4,70]. In this study, a genomic region was defined as a PMU or a PMU-like region if it harboured genes commonly associated with PMU elements such as *tra5*, *dnaB, dnaG, hflB, himA, fliA, ssb, dam, uvrD* and *tmk*. Altogether, 15 PMUs and PMU-like regions were found (Fig. 1) as well as two additional copies of a single *tra5* without PMU-associated genes, and they occurred in a genomic region rich in hypothetical protein genes. Most of the PMU-like regions detected in the SA-1 strain genome were located at the beginning or at the end of the contig, while some of the assembled contigs were entirely composed of a PMU-like region (Fig. 1). The largest PMU cluster was 18 kbp (Supplementary material Fig. S6A in the online version at DOI: 10.1016/j.syapm.2018.10.008) corresponding to an almost complete size of approximately 20 kbp of the well-described PMU1 of AY-WB phytoplasma, which has also been shown to be present as a circular extrachromosomal copy with up-regulated gene expression in insects versus plant hosts [70]. However, the corresponding SA-1 PMU region lacked the *tra5* gene, hence it as probably not a complete PMU. For comparison, in the '*Ca*. P. australiense' chromosome, the largest PMU out of five different described PMU regions was approximately 11 kb, and it has been shown to have another copy present in the genome [71]. The majority of PMU-like regions found in the SA-1 draft genome were from 10,3 to 14,3 kbp in size, while some other PMU-like regions were 2–3 kbp in length and appeared to be fragmented or degenerated PMUs (Supplementary material Fig. S6 in the online version at DOI: 10.1016/j.syapm.2018.10.008). However, the number of detected PMU-like regions might not be exhaustive, since repetitive elements are difficult to assemble and resulted in an incomplete '*Ca*. P. solani' SA-1 strain genome assembly.

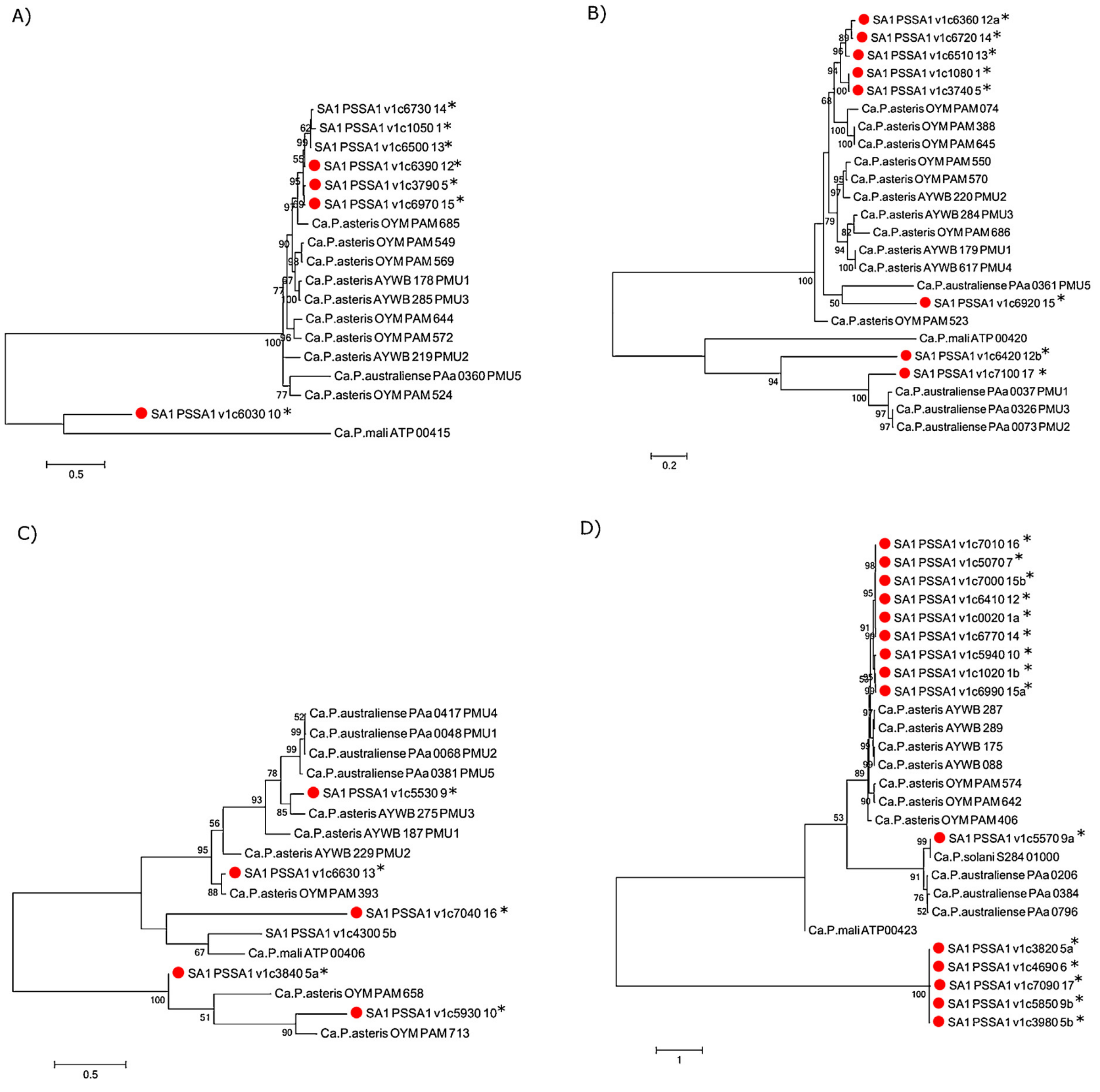

Fig. 4. Molecular phylogeny of PMU-associated genes: A) *dnaG* gene encoding DNA primase; B) *dnaB* gene encoding replicative DNA helicase; C) *hflB* gene encoding ATP-dependant Zn protease; D) *tra5* gene encoding putative transposase/phage integrase. Phylogenetic trees were inferred by maximum likelihood method with Tamura-Nei model and gamma distributed rate variation among sites. Bootstrap values based on 500 replicates are indicated with numbers next to the internal branches. The scale bar represents the number of substitutions per site. GenBank accession numbers are given next to the name of each species or strain. The positions of '*Ca.* P. solani' SA-1 strain sequences are indicated with a red dot. The numbers of SA-1 contings are marked with an asterisk.

In comparison to the other phytoplasmas, PMUs and PMU-like clusters of '*Ca.* P. solani' SA-1 have even higher levels of abundance and organizational heterogeneity. Many rearrangements, as well as the presence of several species-specific PMU-related genes were identified (Supplementary material Fig. S6 in the online version at DOI: 10.1016/j.syapm.2018.10.008). Based on the organization and gene content within the mobile cluster, some of the PMU-like regions in SA-1 genome resembled '*Ca.* P. asteris' potential mobile units: the PMU1 of AY-WB phytoplasma [4] and mobile unit gene clusters (MUG) of OY-M phytoplasma [2] (Supplementary material Fig. S6A in the online version at DOI: 10.1016/j.syapm.2018.10.008). The close relatedness of these PMU-like regions was further corroborated by molecular phylogeny of PMU-associated genes. All the PMU-related genes analysed from these PMU regions clustered with the corresponding sequences originating from '*Ca.* P. asteris, strains (Fig. 4, Supplementary material Fig. S7 in the online version at DOI: 10.1016/j.syapm.2018.10.008) indicating possible horizontal transfer of these PMUs. Further evidence for the exchange of PMUs in the "*Ca.* P. solani" SA-1 genome was the finding that inverted repeats (IR) found downstream of *tra5* sequences phylogenetically closely related to '*Ca.* P. asteris' (Fig. 4D) corresponded to IR found in PMU1 of AY-WB strain. Apart from PMU-like regions that were shown to be related to '*Ca.* P. asteris' PMUs, another group of PMU-like elements was found, encompassing regions showing more complex and specific mosaic structure. Both the ordering and the molecular phylogeny of PMU-associated genes pointed to a different origin, as well as massive rearrangement and intermixing of PMUs (Supplementary material Fig. S6B in the online version at DOI: 10.1016/j.syapm.2018.10.008). In general, significant variability and heterogeneity of PMU-associated genes was found to be present in the '*Ca.* P. solani' SA-1 genome. For example, six PMU-associated *dnaG* sequences of '*Ca.* P. solani' SA-1 strain were shown to be closely related to 'Ca. P. asteris', while one clustered with '*Ca.* P. mali', a species belonging to a divergent phytoplasma lineage (Fig. 4A). Moreover, a close phylogenetic relationship to '*Ca*. P. australiense' PMUs with possibly different origins were also demonstrated for *dnaB, tra5* and *him* genes (Fig. 4, Supplementary material Fig. S7 in the online version at DOI: 10.1016/j.syapm.2018.10.008). Molecular phylogeny of PMU-associated genes such as *him, tmk* (Supplementary material Fig. S6 in the online version at DOI: 10.1016/j.syapm.2018.10.008) and *tra5* (Fig. 4D) revealed the presence of sequences specific to '*Ca*. P. solani' species, which further corroborated heterogeneity and a different origin of the PMU-associated elements. Furthermore, a BLAST search of corresponding nucleotide and amino-acid sequences demonstrated a high coverage and sequence identity of 91%–100% only with sequences originating from '*Ca.* P. solani' strains. A finding of characteristic putative transposases encoded by *tra5* genes on contigs 5, 6, 9 and 17 (Figs. 6; 7D) might be of special importance regarding the plasticity and evolution of '*Ca.* P. solani'. A domain and motif search for these proteins revealed the presence of integrase core domain Rve, as well a helix-turn helix motif (HTH). However, in a protein encoded by *tra5* gene from contig 17, a presence of the sigma-54 DNA-binding domain was revealed. Sigma-54 factor is encoded by the *rpoN* gene and is known to be involved in the alternative regulation of gene expression in a wide range of cellular processes, such as flagellar synthesis and virulence, especially in quorum sensing regulation in pathogenic bacteria [20,31].

Although horizontal gene transfer of PMUs has been demonstrated previously for the PNWB phytoplasma [12], phylogenetic analyses of PMU-related genes in the SA-1 genome revealed horizontal gene transfer from those both closely related and from distantly related (Fig. 4). Features such as mosaicism and a different origin have been described in phytoplasmas for the first time. Hence, it is likely that the SA-1 genome is prone to a high-degree of PMU-mediated recombination and rearrangement. This is in agreement with the erratic GC-skew, low synteny with genomes of related phytoplasmas such as '*Ca.* P. australiense', and the observation that SA-1 PMUs are highly diverse with intermixed sequences and distinctive gene orders (Supplemantary material Fig. S6 in the online version at DOI: 10.1016/j.syapm.2018.10.008).

Conclusions

This study reported the sequencing and draft assembly of the '*Ca*. P. solani' SA-1 strain. The functional and comparative analyses of gene content demonstrated the existence of most information processing and metabolism genes conserved among phytoplasmas, suggesting that repetitive elements could mainly account for the unassembled part of the genome.

One of the significant findings of the study was that molecular phylogeny of genes in PMUs and PMU-like regions in the SA-1 phytoplasma, and in some cases their organization, suggested horizontal transfer of PMUs from related ('*Ca*. P. australiense') and more distant ('*Ca*. P. asteris' and '*Ca*. P. mali') phytoplasmas which revealed a different origin and mosaicism of acquired elements. These rearrangements made SA-1 genome highly dynamic and prone to adopting foreign sequences, including effector genes that were very often found located within PMUs. Thus, the highly dynamic genome with many repeat-rich regions complicated whole-genome sequencing efforts for phytoplasma but may confer an evolutionary advantage on these organisms that have a wide host range. Furthermore, rearrangements of PMU clusters and the abundance of degenerated PMUs may also be taken into account because of the size of the '*Ca.* P. solani' genome that possesses one of the largest chromosomes among phytoplasmas. In agreement with this, it is hypothesized that '*Ca.* P. solani' species might have undergone specific evolution that has allowed genome adaptability and plasticity in order to be spread by different insect vector species which has led to infection of numerous plant hosts, thus enabling a cosmopolite lifestyle for this phytoplasma. Furthermore, the approach of coupling the 16S rRNA gene phylogeny and ANI/AAI results showing a very high sequence divergence rate within the '*Ca.* Phytoplasma' genus indicated that the taxonomic revision of this genus is needed.

Acknowledgments

We would like to thank Prof. Assunta Bertaccini, DiSTA, University of Bologna, Italy for providing micropropagated SA-1 strain, Dr. Xavier Foissac, INRA, University of Bordeaux, UMR1332 Fruit Biology and Pathology, Villenave d'Ornon, France for helpful advice and sharing unpublished data. The Illumina sequencing service was provided by the Genomic Technology Core Facility (Institute of Plant and Microbial Biology, Academia Sinica, Taiwan), the Genomics Core Facility (Institute of Molecular Biology, Academia Sinica, Taiwan), and The Genome Analysis Centre (Norwich, UK).

This research was supported by Academia Sinica (Taipei, Taiwan) and the Croatian Science Foundation grant no. UIP-2014-09-9744.

References

[1] Andersen, M.T., Liefting, L.W., Havukkala, I., Beever, R.E. (2013) Comparison of the complete genome sequence of two closely related isolates of "*Candidatus* Phytoplasma australiense reveals genome plasticity". BMC Genom. 14, 529.

[2] Arashida, R., Kakizawa, S., Hoshi, A., Ishii, Y., Jung, H.-Y., Kagiwada, S., Yamaji, Y., Oshima, K., Namba, S. (2008) Heterogeneic dynamics of the structures of multiple gene clusters in two pathogenetically different lines originating from the same phytoplasma. DNA Cell Biol. 27, 209–217.

[3] Bai, X., Correa, V.R., Toruño, T.Y., Ammar, E.-D., Kamoun, S., Hogenhout, S.A. (2009) AY-WB phytoplasma secretes a protein that targets plant cell nuclei. Mol. Plant Microbe Interact. 22, 18–30.

[4] Bai, X., Zhang, J., Ewing, A., Miller, S.A., Radek, A.J., Shevchenko, D.V., Tsukerman, K., Walunas, T., Lapidus, A., Campbell, J.W., Hogenhout, S.A. (2006) Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts. J. Bacteriol. 188, 3682–3696.

[5] Bendtsen, J.D., Nielsen, H., Von Heijne, G., Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340, 783–795.

[6] Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W. (2015) GenBank. Nucleic Acids Res. 43, D30–D35.

[7] Bertaccini, A., Paltrinieri, S., Martini, M., Tedeschi, M., Contaldo, N. (2013) Micropropagation and maintenance of phytoplasmas in tissue culture. Methods Mol. Biol. 938, 33–39.

[8] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinform. 10, 421.

[9] Carver, T., Thomson, N., Bleasby, A., Berriman, M., Parkhill, J. (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25, 119–120.

[10] Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., Parkhill, J. (2005) ACT: the Artemis comparison tool. Bioinformatics 21, 3422–3423.

[11] Chang, S.-H., Cho, S.-T., Chen, C.-L., Yang, J.-Y., Kuo, C.-H. (2015) Draft genome sequence of a 16SrII-A subgroup phytoplasma associated with purple coneflower (*Echinacea purpurea*) witches' broom disease in Taiwan. Genome Announc. 3, e01398-15.

[12] Chung, W.C., Chen, L.L., Lo, W.S., Lin, C.P., 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.

[13] Cimerman, A., Pacifico, D., Salar, P., Marzachi, C., Foissac, X. (2009) Striking diversity of vmp1, a variable gene encoding a putative membrane protein of the stolbur phytoplasma. Appl. Environ. Microbiol. 75, 2951–2957.

[14] Cvrković, T., Jović, J., Mitrović, M., Krstić, O., Toševski, I. (2014) Experimental and molecular evidence of *Reptalus panzeri* as a natural vector of bois noir. Plant Pathol. 63, 42–53.

[15] Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

[16] Fabre, A., Danet, J.L., Foissac, X. (2011) The stolbur phytoplasma antigenic membrane protein gene stamp is submitted to diversifying positive selection. Gene 472, 37–41.

[17] Felsenstein, J. (1989) PHYLIP-phylogeny interference package (version 3.2). Cladistics 5, 164–166.

[18] Firrao, G., Andersen, M., Bertaccini, A., Boudon, E., Bové, J.M., Daire, X., Davis, R.E., Fletcher, J., Garnier, M., Gibb, K.S., Gundersen-Rindal, D.E., Harrison, N., Hiruki, C., Kirkpatrick, B.C., Jones, P., Kuske, C.R., Lee, I.M., Liefting, L., Marcone, C., Namba, S., Schneider, B., Sears, B.B., Seemüller, E., Smart, C.D., Streten, C., Wang, K. (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.

[19] Fischer, A., Santana-Cruz, I., Wambua, L., Olds, C., Midega, C., Dickinson, M., Kawicha, P., Khan, Z., Masiga, D., Jores, J., Schneider, B. (2016) Draft genome sequence of "*Candidatus* Phytoplasma oryzae" strain Mbita1, the causative agent of Napier grass stunt disease in Kenya. Genome Announc. 4, e00297-16.

[20] Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., Siezen, R.J. (2011) Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. BMC Genom. 12, 385.

[21] Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

[22] Guindon, S., Gascuel, O. (2003) A simple, fast, and accurate method to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

[23] Hodgetts, J., Crossley, D., Dickinson, M. (2013) Techniques for the maintenance and propagation of phytoplasmas in glasshouse collections of *Catharanthus roseus*. Methods Mol. Biol. 938, 15–32.

[24] Hogenhout, S.A., Musić, M.Š. (2009) Phytoplasma genomics, from sequencing to comparative and functional genomics – what have we learnt? In: Weintraub, P., Jones, P. (Eds.), Phytoplasmas: Genomes, Plant Hosts and Vectors, Cabi, pp. 19–36.

[25] Hogenhout, S.A., Oshima, K., Ammar, E.D., Kakizawa, S., Kingdom, H.N., Namba, S. (2008) Phytoplasmas: bacteria that manipulate plants and insects Mol. Plant Pathol. 9, 403–423.

[26] Hoshi, A., Oshima, K., Kakizawa, S., Ishii, Y., Ozeki, J., Hashimoto, M., Komatsu, K., Kagiwada, S., Yamaji, Y., Namba, S. (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.

[27] Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11, 119.

[28] Janik, K., Mithöfer, A., Raffeiner, M., Stellmach, H., Hause, B., Schlink, K. (2017) An effector of apple proliferation phytoplasma targets TCP transcription factors—a generalized virulence strategy of phytoplasma? Mol. Plant Pathol. 18, 435–442.

[29] Kakizawa, S., Oshima, K., Nishigawa, H., Jung, H.Y., Wei, W., Suzuki, S., Tanaka, M., Miyata, S.I., Ugaki, M., Namba, S. (2004) Secretion of immunodominant membrane protein from onion yellows phytoplasma through the Sec protein-translocation system in *Escherichia coli*. Microbiology 150, 135–142.

[30] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M. (2009) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355–D360.

[31] Kazmierczak, M.J., Wiedmann, M., Boor, K.J. (2005) Alternative sigma factors and their roles in bacterial virulence. Microbiol. Mol. Biol. Rev. 69, 527–543.

[32] Killiny, N. (2016) Generous hosts: what makes Madagascar periwinkle (*Catharanthus roseus*) the perfect experimental host plant for fastidious bacteria? Plant Physiol. Biochem. 109, 28–35.

[33] Konstantinidis, K.T., Rosselló-Móra, R., Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. ISME J., 2399–2406.

[34] Konstantinidis, K.T., Tiedje, J.M. (2005) Towards a genome-based taxonomy for prokaryotes. J. Bacteriol. 187, 6258–6264.

[35] Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.

[36] Ku, C., Chung, W.-C., Chen, L.-L., Kuo, C.-H. (2013) The complete plastid genome sequence of Madagascar periwinkle *Catharanthus roseus* (L.) G. Don: plastid genome evolution, molecular marker identification, and phylogenetic implications in asterids. PLoS One 8, e68518.

[37] Kube, M., Mitrovic, J., Duduk, B., Rabus, R., Seemüller, E. (2012) Current view on phytoplasma genomes and encoded metabolism. Sci. World J. 2012, 1–25.

[38] Kube, M., Schneider, B., Kuhl, H., Dandekar, T., Heitmann, K., Migdoll, A.M., Reinhardt, R., Seemüller, E. (2008) The linear chromosome of the plant-pathogenic mycoplasma "*Candidatus* Phytoplasma mali". BMC Genom. 9, 306.

[39] Kuo, C.H., Ochman, H. (2009) Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biol. Direct 4, 35.

[40] Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108.

[41] Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007) Clustal W and clustal X version 2.0. Bioinformatics 23, 2947–2948.

[42] Lee, I.-M., Davis, R.E., Gundersen-Rindal, D.E. (2000) Phytoplasma: phytopathogenic mollicutes. Annu. Rev. Microbiol. 54, 221–255.

[43] Lee, I.-M., Shao, J., Bottner-Parker, K.D., Gundersen-Rindal, D.E., Zhao, Y., Davis, R.E. (2015) Draft genome sequence of "*Candidatus* Phytoplasma pruni" strain CX, a plant-pathogenic bacterium. Genome Announc. 1, e01117–15.

[44] Li, H., Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

[45] Li, L., Stoeckert, C.J.J., Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.

[46] Lowe, T.M., Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transferRNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

[47] MacLean, A.M., Orlovskis, Z., Kowitwanich, K., Zdziarska, A.M., Angenent, G.C., Immink, R.G.H., Hogenhout, S.A. (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.

[48] MacLean, A.M., Sugio, A., Makarova, O.V., Findlay, K.C., Grieve, V.M., Toth, R., Nicolaisen, M., Hogenhout, S.A. (2011) Phytoplasma effector SAP54 induces indeterminate leaf-like flower development in Arabidopsis plants. Plant Physiol. 157, 831–841.

[49] Maixner, M. (2010) Phytoplasma epidemiological systems with multiple plant hosts. In: Weintraub, P., Jones, P. (Eds.), Phytoplasmas: Genomes, Plant Hosts and Vectors, Cabi, pp. 213–233.

[50] Marcone, C., Neimark, H., Ragozzino, A., Lauer, U., Seemüller, E. (1999) Chromosome sizes of phytoplasmas composing major phylogenetic groups and subgroups. Phytopathology 89, 805–810.

[51] Mitrović, J., Siewert, C., Duduk, B., Hecht, J., Mölling, K., Broecker, F., Beyerlein, P., Büttner, C., Bertaccini, A., Kube, M. (2014) Generation and analysis of draft sequences of "stolbur" phytoplasma from multiple displacement amplification templates. J. Mol. Microbiol. Biotechnol. 24, 1–11.

[52] Miura, C., Komatsu, K., Maejima, K., Nijo, T., Kitazawa, Y., Tomomitsu, T., Yusa, A., Himeno, M., Oshima, K., Namba, S. (2015) Functional characterization of the principal sigma factor RpoD of phytoplasmas via an in vitro transcription assay. Sci. Rep. 5, 11893.

[53] Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W192–W195.

[54] Murolo, S., Romanazzi, G. (2015) In-vineyard population structure of "*Candidatus* Phytoplasma solani" using multilocus sequence typing analysis. Infect. Genet. Evol. 31, 221–230.

[55] Orlovskis, Z., Canale, M.C., Haryono, M., Lopes, J.R.S., Kuo, C.H., 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.

[56] Oshima, K., Kakizawa, S., Arashida, R., Ishii, Y., Hoshi, A., Hayashi, Y., Kagiwada, S., Namba, S. (2007) Presence of two glycolytic gene clusters in a severe pathogenic line of *Candidatus* Phytoplasma asteris. Mol. Plant Pathol. 8, 481–489.

[57] Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.Y., Wei, W., Suzuki, S., Arashida, R., Nakata, D., Miyata, S.I., Ugaki, M., Namba, S. (2004) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nat. Genet. 36, 27–29.

[58] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055.

[59] Plavec, J., Križanac, I., Budinšćak, Ž., Škorić, D., Musić, M.Š. (2015) A case study of FD and BN phytoplasma variability in Croatia: multigene sequence analysis approach. Eur. J. Plant Pathol. 142, 591–601.

[60] Quaglino, F., Kube, M., Jawhari, M., Abou-Jawdah, Y., Siewert, C., Choueiri, E., Sobh, H., Casati, P., Tedeschi, R., Lova, M.M., Alma, A., Bianco, P.A. (2015) "*Candidatus* Phytoplasma phoenicium" associated with almond witches'-broom disease: from draft genome to genetic diversity among strain populations. BMC Microbiol. 15, 148.

[61] Quaglino, F., Zhao, Y., Casati, P., Bulgari, D., Bianco, P.A., Wei, W., Davis, R.E. (2013) "*Candidatus* Phytoplasma solani", a novel taxon associated with stolbur- and bois noir-related diseases of plants. Int. J. Syst. Evol. Microbiol. 63, 2879–2894.

[62] Razin, S., Yogev, D., Naot, Y. (1998) Molecular biology and pathogenicity of mycoplasmas. Microbiol. Mol. Biol. Rev. 62, 1094–1156.

[63] Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P. (2011) Integrative genomics viewer. Nat. Biotechnol., 24–26.

[64] Saccardo, F., Martini, M., Palmano, S., Ermacora, P., Scortichini, M., Loi, N., Firrao, G. (2012) Genome drafts of four phytoplasma strains of the ribosomal group 16SrIII. Microbiology (United Kingdom) 158, 2805–2814.

[65] Šeruga, M., Škorić, D., Botti, S., Paltrinieri, S., Juretić, N., Bertaccini, A.F. (2003) Molecular characterization of a phytoplasma from the aster yellows (16SrI) group naturally infecting *Populus nigra* L. "Italica" trees in Croatia. For. Pathol. 33, 113–125.

[66] Sugio, A., Kingdom, H.N., MacLean, A.M., Grieve, V.M., Hogenhout, S.A. (2011) Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating plant development and defense hormone biosynthesis. Proc. Natl. Acad. Sci. 108, E1254–E1263.

[67] Tamura, K., Dudley, J., Nei, M., Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

[68] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Smirnov, S., Nikolskaya, A.N., Rao, B.S., Mekhedov, S.L., Sverlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform. 4, 41.

[69] Thompson, C.C., Vieira, N.M., Vicente, A.C.P., Thompson, F.L. (2011) Towards a genome based taxonomy of *Mycoplasmas*. Infect. Genet. Evol. 11 (7), 1798–1804.

[70] Toruño, T.Y., Seruga Musić, M., Simi, S., Nicolaisen, M., Hogenhout, S.A. (2010) Phytoplasma PMU1 exists as linear chromosomal and circular extrachromosomal elements and has enhanced expression in insect vectors compared with plant hosts. Mol. Microbiol. 77 (6), 1406–1415.

[71] Tran-Nguyen, L.T.T., Kube, M., Schneider, B., Reinhardt, R., Gibb, K.S. (2008) Comparative genome analysis of "*Candidatus* Phytoplasma australiense" (subgroup tuf-Australia I; rp-A) and "*Ca.* phytoplasma asteris" strains OY-M and AY-WB. J. Bacteriol. 190 (11), 3979–3991.

[72] Tsai, Y.-M., Chang, A., Kuo, C.-H., Sloan, D. (2018) Horizontal gene acquisitions contributed to genome expansion in insect-symbiotic *Spiroplasma clarkii*. Genome Biol. Evol. 10 (6), 1526–1532.

[73] Wang, Y., Coleman-Derr, D., Chen, G., Gu, Y.Q. (2015) OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 43 (W1), W78–W84.

[74] Zamorano, A., Fiore, N. (2016) Draft genome sequence of 16SrIII-J phytoplasma, a plant pathogenic bacterium with a broad spectrum of hosts. Genome Announc. 4, e00602-16.

[75] Zerbino, D.R., Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

[76] Zhao, Y., 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.

[77] Zhao, Y., Davis, R.E., Wei, W., Lee, I.M. (2015) Should '*Candidatus* Phytoplasma' be retained within the order Acholeplasmatales? Int. J. Syst. Evol. Microbiol. 65, 1075–1082.