Identification and characterization of conserved and variable regions of lime witches' broom phytoplasma genome

Majid Siampour, Keramatollah Izadpanah, Cristina Marzachi and Mohammad Salehi Abarkoohi

Several segments (~20 kbp) of the lime witches' broom (LWB) phytoplasma genome (16SrII group) were sequenced and analysed. A 5.7 kbp segment (LWB-C) included conserved genes whose phylogenetic tree was consistent with that generated using 16S rRNA genes. Another 6.4 kbp LWB phytoplasma genome segment (LWB-NC) was structurally similar to the putative mobile unit or sequence variable mosaic genomic region of phytoplasmas, although it represented a new arrangement of genes or pseudogenes such as phage-related protein genes and tra5 insertion sequences. Sequence- and phylogenetic-based evidence suggested that LWB-NC is a genomic region which includes horizontally transferred genes and could be regarded as a hot region to incorporate more foreign genes into the genome of LWB phytoplasma. The presence of phylogenetically related fragments of retroelements was also verified in the LWB phytoplasma genome. Putative intragenomic retrotransposition or retromoring of these elements might have been determinant in shaping and manipulating the LWB phytoplasma genome. Altogether, the results of this study suggested that the genome of LWB phytoplasma is colonized by a variety of genes that have been acquired through horizontal gene transfer events, which may have further affected the genome through intragenomic mobility and insertion at cognate or incognate sites. Some of these genes are expected to have been involved in the development of features specific to LWB phytoplasma.

INTRODUCTION

Phytoplasmas are wall-less pleomorphic bacteria of the class Mollicutes able to infect a wide variety of plants and cause considerable yield loss in all kinds of crops worldwide. They are poorly culturable on cell-free artificial media (Contaldo et al., 2012), thus being described as ‘Candidatus’ species. In plants, they are mainly restricted to sieve tubes from which they send their signals to other tissues of the plant (Lee et al., 2000; Hogenhout et al., 2008). Phytoplasmas are known to be transmitted by leafhoppers, planthoppers and psyllids in a circulative propagative manner (Lee et al., 2000; Bertaccini & Duduk, 2009). Phylogenetic analyses using the 16S rRNA gene revealed that phytoplasmas make up a single monophyletic clade diverged from Acholeplasma spp. (Lim & Sears, 1992; Gundersen & Lee, 1996; Lee et al., 2000).

The complete genome sequence of several phytoplasmas, including two strains of ‘Ca. Phytoplasma asteris’ [onion yellows (OY; Oshima et al., 2004) and aster yellows-witches’ broom (AY-WB; Bai et al., 2006)], two strains of ‘Ca. Phytoplasma australiense’ (Tran-Nguyen et al., 2008; Andersen et al., 2013) and ‘Ca. Phytoplasma mali’ (Kube et al., 2008), have been determined. In addition, the draft genome of several phytoplasmas, including four phytoplasma strains of the X-disease group (16SrIII; Saccardo et al., 2012), peanut witches’ broom phytoplasma (PnWB, 16SrII; Chung et al., 2013), two strains of the stolbur phytoplasma (16SrXII; Mitrovic et al., 2014) and a strain of...
‘Ca. P. asteris’ (Kakizawa et al., 2014), have been determined. Sequence analyses show that phytoplasmas have gone through reductive evolution by losing some metabolic pathways, making them largely dependent on their hosts for products of these pathways (Oshima et al., 2004; Namba, 2011). With these unique genomic features, phytoplasmas are considered as micro-organisms with the smallest genome of ~500–1200 kbp (Marcone et al., 1999). Special features of phytoplasmas, initially described in the genome of AY-WB phytoplasma, are the putative mobile units (PMUs), which are clusters of gene sequences (Bai et al., 2006) similar to a composite transposon. There are also regions of repeated gene clusters, called sequence variable mosaics (SVMs), in the genome of phytoplasmas (Jomantiene & Davis, 2006; Jomantiene et al., 2007). SVMs and PMUs share similar properties, and may refer to the same genomic regions of phytoplasmas. These elements seem to play a significant role in shaping the genome of phytoplasmas. The AY-WB phytoplasma genome contains four complete and several degenerated PMUs of different sizes, together accounting for 10.2% of its chromosomal genome content (Bai et al., 2006; Hogenhout et al., 2008). SVMs may have resulted from some ancient recurrent phage attacks that have remained in the genome of phytoplasmas as prophage genomes or prophage genome remnants (Wei et al., 2008). These SVMs (phytoplasma prophages) and/or PMUs contain some important genes for phytoplasma biology (Wei et al., 2008; Kube et al., 2012). In addition to phytoplasmas of the 16SrI group, PMUs or SVMs have been found in the genome of phytoplasmas of other 16Sr groups (Kube et al., 2008; Tran- Nguyen et al., 2008; Saccardo et al., 2012; Chung et al., 2013). At present, ~90% of the PnWB phytoplasma genome has been sequenced, providing information on the first representative of 16SrII group phytoplasmas (Chung et al., 2013). Comparative analysis of PnWB with five available phytoplasma genomes uncovered variations in gene content and metabolic capacities amongst phytoplasmas. It was also found that horizontal gene transfer in the PMU region plays a role in shaping the genome of PnWB phytoplasma (Chung et al., 2013). The information from the genome of ‘Ca. P. aurantifolia’ (16SrII), the causal agent of LWB disease, is restricted to a few genes that do not include SVMs or PMUs. Based on the analysis of conserved genes, including 16S rRNA and the immunodominant membrane protein, LWB phytoplasma was shown to be closely related to 16SrII phytoplasmas (Siampour et al., 2013). In the present study, we obtained sequence information from different regions of the genome of this phytoplasma (including SVM regions) and compared them with their counterparts in PnWB or phytoplasmas of other 16Sr groups for which the genome sequence data are available.

**METHODS**

**PCR, cloning and sequencing of LWB phytoplasma DNA fragments.** An isolate of LWB phytoplasma maintained in periwinkle (Siampour et al., 2013) was used in this study. Total DNA from LWB-infected and healthy periwinkles was extracted using the cetyltrimethylammonium bromide method (Maixner et al., 1995).

The presence of SVM-related genes and the mosaic pattern of these genes were explored using primer pair Fznb/Rb2 originally designed for amplification of SVM regions of 16SrI phytoplasmas (Jomantiene & Davis, 2006; Jomantiene et al., 2007). By using information on PMU-associated DNA primase (duoG) and tra5 genes of PnWB phytoplasma, PrimF (5'-atccatgaatccaacg-3') and TraR (5'-atcggtcattgagc-ag-3') primers were designed for PCR amplification of SVM genes inserted between these genes in the LWB phytoplasma genome. In addition, primers NusF (5'-caacattctgattgatac-3') specific to nusG and RpoR (5'-catgccatcataaacg-3') specific to rpoB genes were designed to amplify a DNA segment of the LWB phytoplasma genome encoding a set of conserved ribosomal proteins. PCR amplification was performed in 50 µL reaction volumes containing 100 ng total nucleic acids, 200 µM dNTPs, 2 µM each primer, 2 mM MgCl2 and 2 U Long-template DNA polymerase (TaKaRa). PCR conditions set up for amplification of long DNA fragments consisted of 5 min initial denaturation at 94 °C followed by 35 cycles of 94 °C for 40 s, 55 °C for 45 s and 68 °C for 5 min. PCR products were extracted from agarose gel and sequenced in three replicates through primer walking.

Two primers IntF1 (5'-ataacacgtggaaagtctg-3') and IntR1 (5'-ttatacagttatgtttgattc-3') were designed to target the terminal regions of the phytoplasma retroelements (also annotated as group II introns) in the genome of LWB phytoplasma and several other phytoplasmas of the 16SrI group. PCR was performed based on the conditions described above. The resultant PCR amplicons of LWB phytoplasma were cloned in pTZ57R/T cloning vector, with a routine protocol. Several clones were selected and the inserted DNA fragments were sequenced in three replicates. Nucleotide sequence data related to different amplicons were assembled using DNAMAN software version 4.02 (Lynnon Biosoft).

**Sequence analysis.** BLAST sequence identity searches (BLASTN or BLASTP) were performed against a custom database containing all phytoplasma sequences (taxid 33926) at the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov) with default settings (in June 2015). Potential ORFs were described using the ORF Finder tool at the NCBI website (http://www.ncbi.nlm.nih.gov/orffinder.cgi). An E value cut-off of <1 × 10⁻10 (with at least 60% amino acid match) was selected as the threshold in BLASTP identity searches. Conserved domains were identified using a BLAST search of the putative protein-coding regions against the Conserved Domain Database at NCBI.

The presence of putative transmembrane helices was predicted using TMHMM (Krogh et al., 2001). For better differentiation of transmembrane helices from signal peptides we used Phobius (Käll et al., 2004) in combination with TMHMM. SignalP (version 3.0; Bendtsen et al., 2004) was further used for prediction of signal peptide domains with default settings defined for Gram-positive bacteria.

**Phylogenetic analysis.** Multiple sequence alignments of the putative protein sequences (translated amino acids of the ORFs) or nucleotide sequences related to the putative ORFs were generated using CLUSTAL Omega (Sievers et al., 2011) via the European Bioinformatics Institute’s web service with default settings (http://www.ebi.ac.uk/Tools/msa/clustalo). The multiple alignments were used to infer the phylogeny for the selected genes according to the maximum-likelihood method by PhyML (Guindon & Gascuel, 2003). Protein- and DNA-based phylogenetic trees were reconstructed under the substitution models of WAG and TN93, respectively. The Γ distribution parameters (with four categories of substitution rates), proportion of invariable sites and transition/transversion ratio were
estimated by the program. The bootstrap support of the phylogenetic
trees was calculated through 100 replicates to evaluate the robustness
of the generated branches.

RESULTS

Analysis of SVM-related fragments

PCR analysis using primer pair Fzn/Rb2 resulted in amplification of several size-variable specific DNA fragments
(not shown). This finding showed that SVMs were present
in the LWB phytoplasma genome. PCR analysis using
primer pair PrimF/TraR resulted in specific amplification of a DNA segment of ~6400 bp from the LWB phytoplasma
genome. The nucleotide sequence and features of this segment (LWB-NC) were determined and compared
with the related sequences in other phytoplasmas
(Table 1, Fig. 1). Using ORF finding programs and BLAST
searches, 11 putative coding regions (ORF), starting with
\( dnaG \) and terminating with a \( tra5 \) insertion sequence
(\( tra5\)-IS, partial sequence), were identified along this seg-
ment. BLAST search analysis showed that the flanking region (280 nt) upstream of the \( dnaG \) gene had high similarity
to the N-terminal region of the same gene in other phytoplasmas (88 % nucleotide identity to PMU-related
\( dnaG \) in PnWB phytoplasma genome); however, disruption by in-frame stop codons appeared to leave a truncated DNA protein with a predicted translation start codon at nt 280. The putative truncated DNA protein in LWB-NC maintained conserved domains related to DNA proteins and was shown to be highly similar to the PMU-related \( dnaG \) gene in PnWB phytoplasma (Table 1). A predicted pseudogene (an ORF degraded by premature stop codons) was also identified on LWB-NC with the highest similarity to a gene in PnWB phytoplasma annotated as phage variable surface lipoprotein. BLAST searches showed that most features identified in LWB-NC had multiple copies in the genome of some other phytoplasmas within the special genomic clusters called PMUs or prophage-related regions (SVMs; data not shown). In this regard, several ORFs identified in LWB-NC were significantly similar to the repeated prophage-related genes detected in OY phyto-
plasma, strain M (OY-M). This finding confirmed that as with OY-M or AY-WB phytoplasmas, the LWB phyto-
plasma genome contained prophage genome remnants.

One of the features in LWB-NC was the presence, at distant
locations of this fragment, of two nucleotide stretches of
~360 bp, which were 94 % identical to each other
(Table 1, Fig. 1). This ‘direct repeat’ had a stretch of AT
nucleotides at its terminal part. This nucleotide region was partly included in the ORF10 structure in fusion
with a truncated phage-related protein (Table 1). Interest-
ingly, we found that the two stretches outside this direct repeat (including features 1, 2, 14 and 15) showed the high-
est identity to a continuous DNA fragment located on the
PMU region of the PnWB phytoplasma genome (GenBank
accession number AMWZ01000002). In contrast, genes located between the repeats (ORF3–10) were
either absent in the sequenced PnWB genome or closer
to the genes in phytoplasmas of other 16Sr groups (e.g.
16SrI or III; Table 1). Moreover, the gene organization
found between the repeats was not recognized in any other phytoplasma genomes.

In addition to a terminal \( tra5 \), another \( tra5\)-IS composed of three fragmented ORFs (features 5, 6 and 7) was
found in the middle part of LWB-NC (Table 1, Fig. 1).
Analysis of the middle \( tra5 \) gene detected two in-frame
stop codons resulting in fragmentation of this gene into three ORFs. Partial comparisons showed that these two
\( tra5\)-ISs at the middle and terminal part of LWB-NC
were highly variable in their sequences. BLAST search showed that a \( tra5 \) gene (transposase-coding gene) inserted
in prophage remnants of OY-M phytoplasma (PAM_683) was the closest relative of the LWB-NC fragmented \( tra5 \) gene (Table 1). Whilst a conserved sequence region associated
with prophage \( tra5 \) genes was detected upstream of the terminal \( tra5 \) gene, it was absent in the upstream
region of the LWB-NC middle \( tra5 \). BLAST search against the IS Database (https://www.is.biotoul.fr/) showed that both \( tra5\)-ISs identified in LWB-NC belonged to the IS3 family.

Analysis using TmHMM and Phobius predicted that ORF2, 7, 8 and 9 of LWB-NC encompass transmembrane domains
(Table 1). The first three ORFs were absent in the OY-M phytoplasma genome (including prophage-related regions); however, the orthologue of ORF9 was found in the prophage
regions of the same phytoplasma as a ‘moron’, i.e. a phage-unrelated protein. Although the putative transmem-
brane domain detected in ORF7 (20 N-terminal amino acids) showed high amino acid identity to the phytoplasma
SVM signal peptide (80 % identity), it was not detected as a signal sequence using SignalP or Phobius.

Other features in LWB-NC were two palindromic sequences of CGATA–TATCG (located within the direct
repeats) and AAGACCTT–AGGGTCTT (located between
ORF7 and 8) that were found (by BLASTN ) to be conserved
and repetitive in the OY-M phytoplasma genome. Our
BLASTN search analysis showed that these putative repetitive extragenic palindromes (REPs) exist at least in 10 different locations of the OY-M phytoplasma genome.

Analysis of LWB phytoplasma retroelements

PCR analysis using primer pair IntF1/IntR1 confirmed the presence of several specific and size-variable retroelements
(bacterial group II introns) in the genome of LWB phyto-
plasma (Fig. 2). Amongst these products, we determined
the complete nucleotide sequence of three fragments with
estimated sizes of 1200 (LWBPI1200), 1500 (LWBPI1500)
and 2500 (LWBPI2500) bp, and the partial nucleotide
sequence of a 2200 bp fragment (LWBPI2200). Sequence
analysis showed that these elements were relatively rich in
GC content (36–38 %) and all belonged to the group II
introns with the highest similarity to those identified in
<table>
<thead>
<tr>
<th>Feature*</th>
<th>5’–3’ Boundaries (nt)</th>
<th>GC (%)</th>
<th>Closest relative (% identity over amino acids)†</th>
<th>Closest phytoplasma relative/16Sr group</th>
<th>Predicted function/similar protein/feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: ORF1</td>
<td>278–1330</td>
<td>32</td>
<td>WP_004994632 (94/348)</td>
<td>PnWB/16SrII</td>
<td>Putative DNA primase/DnaG</td>
</tr>
<tr>
<td>2: Pseudogene</td>
<td>1441–2070</td>
<td>30</td>
<td>AMWZ01000002 (97/629)</td>
<td>PnWB/16SrII</td>
<td>Annotated as phase-variable surface lipoprotein in PnWB</td>
</tr>
<tr>
<td>3: DR1</td>
<td>1990–2355</td>
<td>24</td>
<td>AY270153 (82/371)</td>
<td>PnWB/16SrII</td>
<td>ND</td>
</tr>
<tr>
<td>4: ORF2</td>
<td>2344–2395</td>
<td>19</td>
<td>WP_017193131 (43/70)</td>
<td>PnWB/16SrII</td>
<td>Italian clover phyllody/16SrIII</td>
</tr>
<tr>
<td>5: ORF3</td>
<td>3077–2826</td>
<td>30</td>
<td>NP_950935 (54/70)</td>
<td>OY-M/16SrI</td>
<td>Tra5 transposase (truncated)</td>
</tr>
<tr>
<td>6: ORF4</td>
<td>3433–3170</td>
<td>30</td>
<td>NP_950935 (48/84)</td>
<td>OY-M/16SrI</td>
<td>Tra5 transposase (truncated)</td>
</tr>
<tr>
<td>7: ORF5</td>
<td>3790–3437</td>
<td>27</td>
<td>NP_950935 (49/112)</td>
<td>OY-M/16SrI</td>
<td>Tra5 transposase (truncated)</td>
</tr>
<tr>
<td>8: ORF6</td>
<td>4087–4470</td>
<td>36</td>
<td>ABH1165 (75/119)</td>
<td>Maryland aster yellows/16SrII</td>
<td>Hypothetical protein, similar to ATP-dependent Zn protease</td>
</tr>
<tr>
<td>9: ORF7</td>
<td>4506–5027</td>
<td>20</td>
<td>WP_034172311 (89/173)</td>
<td>Chrysanthemum yellows/16SrII</td>
<td>Hypothetical protein, N-terminal similarity to SVM signal peptide of phytoplasmas</td>
</tr>
<tr>
<td>10: ORF8</td>
<td>5147–5365</td>
<td>24</td>
<td>ABU55748 (90/72)</td>
<td>Spiraea stunt/16SrIII</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>11: ORF9</td>
<td>5349–5597</td>
<td>18</td>
<td>YP_002154421 (91/82)</td>
<td>OY-M/16SrI</td>
<td>Hypothetical protein, highly similar to a ‘moron’ in OY-M phytoplasma prophage</td>
</tr>
<tr>
<td>12: ORF10 Ω</td>
<td>5600–5818</td>
<td>34</td>
<td>WP_004994632 (88/50)</td>
<td>PnWB/16SrII</td>
<td>Hypothetical protein, putative phage-related protein Yqf fused with DR2</td>
</tr>
<tr>
<td>13: DR2</td>
<td>5738–6076</td>
<td>24</td>
<td>AMWZ01000002 (89/352)</td>
<td>PnWB/16SrII</td>
<td>ND</td>
</tr>
<tr>
<td>14: Conserved tra5 upstream</td>
<td>6114–6296</td>
<td>24</td>
<td>AY270153 (89/193)</td>
<td>PnWB/16SrII</td>
<td>Putative regulatory sequences upstream of Tra5 transposases</td>
</tr>
<tr>
<td>15: ORF11</td>
<td>6298–6413</td>
<td>23</td>
<td>WP_004994634 (100/38)</td>
<td>PnWB/16SrII</td>
<td>Partial/Tra5 transposase</td>
</tr>
</tbody>
</table>

ND, Not determined.  
*Features identified in LWB-NC are specified by numbers (1–15) and names. Pseudogene, the putative gene degenerated by premature stop codons; ORF2, 7, 8 and 9, putative transmembrane proteins; DR1 and 2, probable ‘direct repeats’ identified based on high identity; Ω, ORF 10 identified as a fusion protein composed of a yqf-related gene (50 aa) fragment fused to part of the DR2 to achieve a 72 aa ORF.  
†GenBank accession number related to the closest relative of each feature amongst the phytoplasma sequence database. The similarity between sequences is shown in parentheses as identity per cent over total matches. The sequence identity analysis for all the ORFs was based on BLASTP, whilst the other features were analysed by BLASTN.
the genome of phytoplasmas. All of these retroelements contained the RT domain (a domain for reverse transcriptase activity), although it was found that the related RT ORF was degenerated by deletions or premature stop codons in all fragments. Compared with a single retroelement detected in OY-M phytoplasma (OYPI1), the LWBPI2500 fragment appeared to be a complete retroelement with a full-size RT region, although the possible expression of full-length RT protein would require frameshifts at two positions. The RT domains (ORFs) in the other three fragments were shown to be degenerated by deletions at different locations. A schematic comparison of these four retroelement fragments within the genomes of LWB phytoplasma and OY-M phytoplasma is shown in Fig. 3.

PCR analysis using primer pair IntF1/IntR1 showed that different fragments of retroelements are also present in the genome of several other 16SrII phytoplasmas. As shown in Fig. 2, the number and size of these retroelement DNA fragments are specific to some of these phytoplasma isolates.

Analysis of a conserved DNA segment of LWB phytoplasma

A DNA fragment of ~5700 bp (LWB-C) was identified following PCR amplification of LWB phytoplasma in infected plants using primer pair NusF/RpoR. The nucleotide sequence of this fragment was fully determined. Analysis revealed that LWB-C consisted of four complete and two partial ORFs. BLAST searches against NCBI amino acid and conserved domain databases determined that the predicted ORFs included the nusG gene (partial) followed by four conserved phytoplasma 50S ribosomal protein genes (including rp1k, rp1A, rp1J and rp1L) and ~820 amino acids related to the N terminus part of the rpoB gene, coding for DNA-directed RNA polymerase β protein (Table 2). A single copy of these genes with similar organization was also found in the genome of other phytoplasmas. All these genes in LWB phytoplasma showed the highest similarity to the corresponding genes in PnWB phytoplasma. The identity scores between LWB-C ORFs and the counterpart genes in other phytoplasmas are shown in Table 2.

Phylogenetic study

A maximum-likelihood phylogenetic tree for 10 phytoplasmas was reconstructed using a concatenated amino acid sequence alignment (1792 aligned positions) related to...
six conserved genes found in the LWB-C fragment (Table 2, Fig. 4). The topology of this phylogenetic tree was in agreement with the phylogenetic tree reconstructed using the 16S rRNA gene (Fig. 4a, b). In this regard, LWB and PnWB phytoplasmas of the 16SrII group clustered together and were more closely related to phytoplasma phylogenetic groups 16SrIII and X than to those of 16SrI and XII. This phylogenetic relationship was used as a reference for evaluation of phylogenetic trees reconstructed using genes or elements in the LWB-NC fragment.

Fig. 3. Sequence map and phylogenetic position of different retroelements identified in LWB phytoplasma. (a) Schematic representation of the nucleotide alignment of four retroelement fragments identified in LWB phytoplasma (LWBPI2500, LWBPI2200, LWBPI1500 and LWBPI1200) compared with a 2516 bp complete retroelement characterized in OY-M phytoplasma (OYP1). The gaps between the thick lines in LWBPI1500 and LWBPI1200 elements represent the deleted parts of these elements compared with OYP1. The nucleotide sequence of the inner region in the LWBPI2200 retroelement (dotted line) was not determined. Numbers above the OY-M element correspond to the nucleotide end points of the OYP1 retroelement in the OY-M phytoplasma complete genome (GenBank accession number AP006628). The thick red arrow corresponds to the RT ORF gene of the OYP1 and the dotted arrows above it show its related domains. RT, reverse transcriptase; X, maturase activity domain; En, DNA endonuclease. (b) Maximum-likelihood phylogenetic tree reconstructed based on 1293 aligned nucleotide sequence positions (corresponding to the LWBPI1200 element) of different retroelements inserted in the genome of phytoplasmas belonging to 16Sr groups I, II, III and XII. ●, OY phytoplasma; ■, PnWB phytoplasma; ▼, Italian clover phyllody phytoplasma; □, strawberry lethal yellowing phytoplasma. The phylogenetic positions of four retroelement sequences identified in LWB phytoplasma are shown by arrowheads. The GenBank accession numbers by which the sequence for each retroelement can be retrieved are shown on the tree. Different retroelements identified in the genome of strawberry lethal yellowing phytoplasma are specified by their locus tag in the brackets. Bootstrap values >50 % are shown next to nodes.

Fig. 4(c) shows a representative maximum-likelihood phylogenetic gene tree based on the PMU (or SVM)-related dnaG gene of phytoplasmas in which the LWB-NC dnaG was clustered with its homologue in PnWB (16SrII). Contrary to the phylogenetic relationship of conserved genes (Fig. 4b), however, the dnaG genes in LWB and PnWB phytoplasmas were closer to some dnaG genes in OY-M phytoplasma (GenBank accession numbers NP_950896 and NP_950824) rather than to that in ‘Ca. P. mali’ phytoplasma. In this regard, it was also found that PMU-related
four retroelements identified in LWB phytoplasma from other phytoplasmas. The nucleotide information related to common regions of tra5 gene in ‘Ca. P. mali’ (16SrII; GenBank accession number AKIM00000000.1) and OY-M (16Sr-I; GenBank accession number AP006628) phytoplasmas.

<table>
<thead>
<tr>
<th>ORF name/product</th>
<th>Size [aa (GC %)]</th>
<th>Amino acid sequence identity*</th>
<th>PnWB</th>
<th>‘Ca. P. mali’</th>
<th>Italian clover phylldy</th>
<th>OY-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>nusG/transcription anttermination protein (partial)</td>
<td>162 (25)</td>
<td>92/157</td>
<td>51/161</td>
<td>66/161</td>
<td>55/152</td>
<td></td>
</tr>
<tr>
<td>rplK/50S ribosomal protein L11</td>
<td>141 (26)</td>
<td>94/141</td>
<td>74/141</td>
<td>74/141</td>
<td>67/140</td>
<td></td>
</tr>
<tr>
<td>rplA/ribosomal protein L1</td>
<td>229 (29)</td>
<td>89/229</td>
<td>49/223</td>
<td>57/224</td>
<td>51/228</td>
<td></td>
</tr>
<tr>
<td>rplJ/50S ribosomal protein L10</td>
<td>173 (23)</td>
<td>84/167</td>
<td>35/161</td>
<td>44/165</td>
<td>33/168</td>
<td></td>
</tr>
<tr>
<td>rplL/50S ribosomal protein L7/L12</td>
<td>128 (26)</td>
<td>84/128</td>
<td>36/121</td>
<td>35/122</td>
<td>30/122</td>
<td></td>
</tr>
<tr>
<td>rpoB/DNA-directed RNA polymerase subunit β (partial)</td>
<td>822 (27)</td>
<td>94/824</td>
<td>60/860</td>
<td>69/804</td>
<td>55/856</td>
<td></td>
</tr>
</tbody>
</table>

*Amino acid identity (%) / total number of amino acids in BLASTP comparisons between LWB-C ORFs and the homologous genes in PnWB (16Sr-II; GenBank accession number AMWZ00000000), ‘Ca. P. mali’ (16Sr-X; GenBank accession number CU469464), Italian clover phylldy (16Sr-III; GenBank accession number AKIM00000000.1) and OY-M (16Sr-I; GenBank accession number AP006628) phytoplasmas.

*DNA* in ‘Ca. P. mali’ (16SrX) represented the most divergent sequence compared with all counterparts in other phytoplasmas. Altogether, the *dnaG* phylogeny corroborated the hypothesis that horizontal transfer of PMU-related *dnaG*, supposedly from 16SrI to 16SrII and III phytoplasma groups, developed after divergence of the 16SrII and III phytoplasma groups from the 16SrX group.

Based on the maximum-likelihood phylogenetic tree reconstructed from the nucleotide alignment of *tra5* genes (21 genes from six phytoplasma isolates of four 16Sr phylogenetic groups), seven phylogenetic clusters could be distinguished (Fig. 4d). As shown, the *tra5* genes in a single phytoplasma genome were not necessarily the closest relatives to each other. For example, the *tra5*-ISs identified in OY-M phytoplasma were distributed in phylogenetic groups 1, 4 and 5 clustering with their orthologues in PnWB, AY-WB and LWB phytoplasmas. Moreover, in some cases *tra5* genes from phytoplasmas belonging to the same 16Sr group were clearly separated from each other. As a clear example, the LWB-NC middle *tra5* gene (nt 2826–3790) was phylogenetically closer to one of those identified in OY-M phytoplasma (PAM_683) than to the *tra5* gene in PnWB phytoplasma. Partial sequence (38 aa) of another *tra5*-IS at the terminal part of LWB-NC was largely different to its middle counterpart, although 100% identical to the *tra5* gene of PnWB. Thus, it can be concluded that at least two divergent paralogues of *tra5*-ISs are present in the LWB phytoplasma genome. Based on the same phylogenetic tree, different *tra5* genes of ‘Ca. P. australiensis’ could also be differentiated into two diverged paralogues.

Similar to the PMU-related *dnaG* phylogenetic tree (Fig. 4c), the *tra5* phylogenetic tree revealed that the *tra5* gene in ‘Ca. P. mali’ from 16SrX was distantly related to *tra5* from other phytoplasmas.

The nucleotide information related to common regions of four retroelements identified in LWB phytoplasma (corresponding to the LWBPI12000, Fig. 3) was used to infer the phylogenetic relationship amongst these elements in different phytoplasmas. The reconstructed maximum-likelihood phylogenetic tree revealed that all of the analysed retroelements identified in the LWB genome clearly clustered together (Fig. 3). Similar behaviour was also observed in the case of various retroelements identified in ‘Ca. P. australiensis’. The single ‘ORF included’ retroelement identified in OY-M phytoplasma (16SrI) clustered with a counterpart in Italian clover phylldy phytoplasma (16SrIII) as supported by a high bootstrap value. The closest relatives of LWB phytoplasma retroelements were those identified in PnWB phytoplasma. Contrary to *tra5* and *dnaG* genes, different copies of retroelements inserted in a given phytoplasma often tended to cluster together and hence were more related to each other than to any retroelement from other phytoplasmas. An exception was a retroelement in PnWB phytoplasma (GenBank accession number EF193164) that was closer to the LWB phytoplasma retroelements than to the other elements in the same phytoplasma genome (Fig. 3).

**DISCUSSION**

In this study, different fragments of the LWB phytoplasma genome with a total size of ~20 kbp were characterized. One of these fragments, LWB-NC, had several features including a long sequence repeat, *tra5*-ISs, phage-related protein genes and REPs resembling SVMs (Jomantiene & Davis, 2006; Jomantiene et al., 2007; Wei et al., 2008) or PMUs (Bai et al., 2006) of the phytoplasma genomes. OY-M and AY-WB phytoplasma SVMs are mainly composed of phage-related protein genes, thereby being considered as prophage genomes or prophage genome remnants (Wei et al., 2008). Interestingly, some of the genes or elements identified in LWB-NC were significantly similar to OY-M and AY-WB prophage-related protein genes, and hence probably originated from phages of the
Fig. 4. Phylogenetic analysis showing phylogeny of 16S rRNA gene and genes in the LWB-C fragment compared with the phylogeny of some genes in the LWB-NC fragment. (a) Maximum-likelihood phylogenetic tree of 10 phytoplasmas (belonging to 16Sr groups I, II, III, X and XII) as inferred by analysis of their 16S rRNA gene. *Acholeplasma laidlawii* (GenBank accession number NC_010163) was used as outgroup. (b) Phytoplasmal maximum-likelihood phylogenetic tree reconstructed using a concatenate alignment of amino acids (1792 aligned amino acid positions) related to conserved genes identified in the LWB-C fragment and their homologues in nine other phytoplasmas of 16Sr groups I, II, X and XII. *A. laidlawii* was used as outgroup. (c) Maximum-likelihood phylogenetic tree of PMU (or SVM)-related dnaG genes identified in the LWB-NC fragment and their homologues in nine other phytoplasmas of 16Sr groups I, II, III, X and XII. *A. laidlawii* was used as outgroup. Elements marked with asterisks are those composed of separated ORFs. (d) Maximum-likelihood phylogenetic tree of tra5 genes (1282 aligned nucleotide positions) reconstructed by the nucleotide alignment of the tripartite tra5 gene identified in the inner part of LWB-NC and those identified in the genome of phytoplasmas belonging to 16Sr groups I, II, X and XII. The tra5-Is of different phytoplasmas are specified based on their locus tag in the respective phytoplasma genome. The tra5 gene in *Erysipelothrix rhusiopathiae* (GenBank accession number AP012027; a member of the IS3 family) was used as outgroup. Elements marked with asterisks are those composed of separated ORFs. The tra5 phylogenetic groups 1–7 are shown. In all phylogenetic trees, numbers next to each node show the bootstrap support as percentage of 100 replicates (only values >50 are shown). ▼, Milkwed yellow phytoplasma (16SrI; GenBank accession number AKIL00000000); ▽, Italian clover phytoflyd phytopylasma (16SrII; GenBank accession number AKIMO00000000); ▲, poinsettia branch-inducing phytoplasma (16SrII; GenBank accession number AKIK00000000); ◊, strawberry lethal yellowing phytoplasma (16SrII; GenBank accession number CP002548); ◊, ‘Ca. P. australiensis’ (16SrII; GenBank accession number AM422018); ■, PnWB phytoplasma (16SrII; GenBank accession number AMWZ00000000); ◵, ‘Ca. P. malii’ (16SrX; GenBank accession number CU469464); ◆, AY-WB phytoplasma (16SrI; GenBank accession number CP000061); ●, OY phytoplasma (16SrI).
order Caudovirales. This finding suggests that phage attacks on phytoplasmas played a role in shaping and reorganizing the LWB phytoplasma genome.

The LWB-NC fragment included two tra5-Is: one in the middle and the other (partially sequenced) in its terminal part. These elements have the potential to affect their host genome by mobility and insertion at different locations. Although tra5 genes are known as the essential, core part of PMUs, it was shown that they are not the constitutive components of phytoplasma prophylogenome genomes (Wei et al., 2008; Kube et al., 2012). The high sequence variability between the terminal and the inner tra5-Is, and lack of conserved flanking sequences upstream of the latter, suggest that these two elements are not derived from each other. Analysis showed that a −1 frameshift (between ORFs 5 and 4) and read-through (between ORF4 and 3) strategies are needed for complete expression of the inner tra5 gene. In most members of the bacterial IS3 family, the frameshift strategy is a common mechanism to encode the active transposases by fusing of two separate ORFs, A and B (Mahillon & Chandler, 1998). Although UGA serves as a stop codon in phytoplasmas, this in-frame codon in the LWB-NC inner tra5 can be suppressed through the mistranslation or read-through strategies as reported in several bacterial systems (Engelberg-Kulka, 1981; Matsugi et al., 1998; González et al., 2003). Experimental evidence is needed to verify whether LWB phytoplasma employs such strategies for expression of this tra5 element.

A new family of REPs called PhREP has been found in the genome of OY-M phytoplasma (Jomantiene & Davis, 2006). Our analysis also proposes that at least two such elements in OY-M phytoplasma are also maintained in the LWB-NC extragenic region. Although the role of these elements in phytoplasmas is still unknown, the stem–loop structure of these elements suggests their involvement in transcription termination, genome stability or recombination.

Contrary to phytoplasma conserved genes in the LWB-C fragment, the phylogenetic trees reconstructed from LWB-NC fragments were not similar to that generated using the 16S rRNA gene. The PMU-related dnaG phylogeny of phytoplasmas revealed that LWB-NC dnaG was clustered with its homologue in PnWB phytoplasma. This cluster, in contrast to the evolutionary history of phytoplasmas, was more closely related to dnaG in OY-M phytoplasma (16SrI) than to that identified in the 16SrX phytoplasma group. Similarly, Chung et al. (2013) showed that PMU-related dnaG in PnWB phytoplasma is closer to PMU-related dnaG genes in the 16SrI phytoplasma group. Hence, it can be concluded that transduction of dnaG to the genome of phytoplasmas has been followed by its horizontal transfer from phytoplasmas of the 16SrI group to 16SrII and III phytoplasma groups. As shown in the phylogenetic tree, copies (paralogues) of dnaG in a given phytoplasma (e.g. 16SrI) do not form a distinct cluster; rather, they are distributed throughout the tree. Conceivably, this is due to multiple introduction of dnaG into a phytoplasma genome, which is also consistent with recurrent phage attacks to the genome of phytoplasmas (Wei et al., 2008). It could also be argued that the genomes of some phytoplasmas (e.g. OY-M) that possess a vast variety of dnaG or other phage-related genes have served as the origin of the counterpart genes identified in other phytoplasmas (e.g. LWB).

Bacterial inhabitants of insect vectors of phytoplasmas have been considered as the possible origin for prophage (or PMU)-related genes in phytoplasmas (Wei et al., 2008). In this regard, the distinct phylogenetic position of ‘Ca. P. mali’ in the dnaG-based phylogenetic tree may be explained by its distinct evolution in a different vector type (i.e. psyllid) and in association with its special bacterial inhabitants. In other words, the origin of horizontally transferred genes in phytoplasmas may be linked to their lifestyle.

It was also shown that the topology of the tra5 phylogenetic tree is largely different from that reconstructed using conserved genes. Interestingly, the fragmented tra5 gene in the inner LWB-NC was phylogenetically closer to a tra5 in OY-M (PAM_683) than to that in PnWB phytoplasma. This suggests the horizontal transfer of tra5 elements between the genome of phytoplasmas related to diverged 16S rRNA groups. Moreover, the distribution of the tra5 genes related to a given phytoplasma (e.g. OY-M) throughout the phylogenetic tree suggests multiple introductions of these elements into the genome of a phytoplasma lineage. However, considerable sequence variation between the tra5 genes in the inner and the terminal parts of LWB-NC suggests that they are not derived from each other through gene duplication.

Sequence identity analysis showed that the closest sequence to terminal parts of LWB-NC (including direct repeats and external flanking regions) was a single continuous fragment in PnWB phytoplasma. By contrast, genes included between these terminal sequences were absent in PnWB phytoplasma or were closer to their counterparts in other phytoplasmas. This finding was also corroborated by phylogenetic analysis where selected genes from these two parts (dnaG and tra5) showed different behaviour in grouping with their counterparts in PnWB. Moreover, a PCR analysis also failed to detect the inner region of LWB-NC in carrot witches’ broom phytoplasma belonging to the 16SrI group (data not shown). These findings may raise the hypothesis that the linkage between terminal parts of LWB-NC is interrupted by the insertion of the inner part between the direct repeats. The closest counterparts of the protein genes in the inner region of LWB-NC, however, were identified in distant phytoplasmas belonging to the 16SrI and III groups. It may also be hypothesized that genes in the inner region of LWB-NC were not introduced from a single phytoplasma at the same time. These findings indicate the prophage (or PMU)-related regions of LWB phytoplasma as the favourable platform for acquisition of foreign genes.

In this study we also analysed several size-variable, albeit homologous, fragments of LWB phytoplasma retroelements...
that all included a degraded RT ORF domain. As a rare case in the bacterial world, an ORF-less group II intron has been recently identified in the OY-M phytoplasma genome (OYPI2; Simon et al., 2008). The RT of a coexistent ORF-including retroelement in the OY-M phytoplasma chromosome (OYPI1) has been postulated to aid the OYPI2 intron in trans (Simon et al., 2008). Similarly, it can be suggested that retrotransposition of ORF-degraded retroelements in LWB phytoplasma is dependent on transaction of a functional (yet unidentified) RT protein from another retroelement. Whilst genomes of LWB phytoplasma and some other phytoplasmas including PnWB and members of the 16SrXII group contained several copies of retroelement fragments, these elements were absent in the AY-WB phytoplasma (16SrI) and ‘Ca P. mali’ (16SrX) genomes (Cimerman et al., 2006; Tran-Nguyen et al., 2008; Chung et al., 2013). This information clearly shows that there are barriers to the insertion and spread of these elements into the genome of some phytoplasmas. These barriers may include different horizontal transfer systems amongst phytoplasmas or differential adaptation of retroelements to phytoplasma genomes. Evidence of horizontal transfer of retroelements between phytoplasmas of distant phylogenetic groups (e.g., between phylogenetic groups 16SrI and III) proposes the intergenomic motion of these elements in phytoplasmas. Phylogenetic analysis of this study also showed that different copies of retroelements in the LWB phytoplasma genome are clustered strictly together in close relationship with retroelements of PnWB phytoplasma. This finding suggests that multiple copies of retroelements in LWB and PnWB phytoplasmas belonging to the 16SrII group may have been derived from a primary single element (common ancestor) followed by its intragenomic retrotransposition or retrohoming (mobility and retention of these elements into non-cognate or cognate sites).

Altogether, the results of this study suggest that the putative intragenomic and/or intergenomic motion of retroelements, phage-related proteins and tras5-1s may have been involved in shaping the LWB phytoplasma genome at different stages in the course of its evolution.

ACKNOWLEDGEMENTS

The authors thank the Iranian Council of Centers of Excellence and National Foundation of Elites for their support.

REFERENCES


Edited by: G. Preston