Ancient, recurrent phage attacks and recombination shaped dynamic sequence-variable mosaics at the root of phytoplasma genome evolution

Wei Wei*, Robert E. Davis*†, Rasa Jomantiene‡, and Yan Zhao*†

*Molecular Plant Pathology Laboratory, U.S. Department of Agriculture-Agricultural Research Service, Beltsville, MD 20705; and ‡Phytovirus Laboratory, Institute of Botany, Vilnius, Lithuania 2021

Contributed by Robert E. Davis, May 30, 2008 (sent for review April 11, 2008)

Mobile genetic elements have impacted biological evolution across all studied organisms, but evidence for a role in evolutionary emergence of an entire phylogenetic clade has not been forthcoming. We suggest that mobile element predation played a formative role in emergence of the phytoplasma clade. Phytoplasmas are cell wall-less bacteria that cause numerous diseases in plants. Phylogenetic analyses indicate that these transkingdom parasites descended from Gram-positive walled bacteria, but events giving rise to the first phytoplasma have remained unknown. Previously we discovered a unique feature of phytoplasmal genome architecture, genes clustered in sequence-variable mosaics (SVMs), and suggested that such structures formed through recurrent, targeted attacks by mobile elements. In the present study, we discovered that cryptic prophage remnants, originating from phages in the order *Caudovirales*, formed SVMs and comprised exceptionally large percentages of the chromosomes of '*Candidatus* Phytoplasma asteris'-related strains OYM and AYWB, occupying nearly all major nonsyntenic sections, and accounting for most of the size difference between the two genomes. The clustered phage remnants formed genomic islands exhibiting distinct DNA physical signatures, such as dinucleotide relative abundance and codon position GC values. Phytoplasma strain-specific genes identified as phage morons were located in hypervariable regions within individual SVMs, indicating that prophage remnants played important roles in generating phytoplasma genetic diversity. Because no SVM-like structures could be identified in genomes of ancestral relatives including *Acholeplasma* spp., we hypothesize that ancient phage attacks leading to SVM formation occurred after divergence of phytoplasmas from acholeplasmas, triggering evolution of the phytoplasma clade.

host-restricted bacteria | mobile gene cassette | pathogenicity island | phytopathogenic bacteria | clade emergence

Phytoplasmas are pleomorphic, cell wall-less bacteria that descended from an acholeplasma-like ancestor and are characterized by small, AT-rich genomes encoding capabilities for transkingdom parasitism and pathogenicity in plants and insects (1, 2). In infected plants, phytoplasmas colonize sieve cells of phloem tissue and typically induce disease symptoms involving impaired amino acid and carbohydrate translocation, disrupted hormonal balance, and rapid senescence (3–7). In their natural insect vectors, phytoplasmas traverse the intestinal wall, circulate in hemolymph, and multiply in tissues including salivary glands, where phytoplasma cells are incorporated into saliva injected into plants during inoculation (8).

Evolutionary adaptation to a broad range of ecological niches has led to emergence of widely divergent phytoplasma lineages (2, 9), and lineages are being discovered at an increasingly rapid pace (10, 11). By contrast, knowledge of their fundamental biology is very limited, largely because of the inability to cultivate phytoplasmas in cell-free media. Genomes of three phytoplasma strains have been completely sequenced (12–14), and genome sequencing of several additional strains is near completion (15, 16), promising insights into phytoplasma genome organization and evolution. Recent work revealed that phytoplasma genomes are highly dynamic structures with unique architecture characterized by the presence of genes repetitively clustered in nonrandomly distributed segments termed sequence-variable mosaics (SVMs), a distinguishing feature of phytoplasma genome architecture (17, 18). It was suggested that the SVMs likely formed through recurrent and targeted mobile element attack and recombination at an early stage in evolution of the phytoplasma clade (18), but the nature and origin of the hypothetical mobile element(s) remained obscure.

Although important features of phytoplasma genome architecture and gene complement were discovered and analyzed (17–20), the occurrence and potential role of prophages in shaping phytoplasma genomes had not been elucidated. In other bacteria, prophages have been identified as major contributors of laterally acquired genes encoding virulence factors (21). Interestingly, alignment of genome sequences from closely related bacterial strains revealed that in some cases all major genome differences can be attributed to prophage sequences (22). Nearly half of the completely sequenced bacterial genomes possess prophage sequences that can constitute a sizable part (10–20%) of a bacterial genome (22, 23). Thus, prophages are major elements of bacterial genomes and a significant driving force for bacterial strain diversification (21). Recent surveys focusing on prophage and other mobile-DNA elements in completely sequenced genomes of walled and wall-less bacteria, including obligately parasitic intracellular bacteria, set the stage for a resurgence of research on microbial mobile elements in host-restricted and other bacteria (24–26).

In our continued work to trace the genesis of phytoplasma genome architecture, we explored the possible role of phages, because repeated phage-related protein genes were prominent features of SVMs (18). In this study, we found that cryptic prophages or prophage genome remnants constituted a significant structural component of phytoplasma genomes. The genomes contained copious amounts of prophage sequences, many in a state of mutational decay, occupying nearly all major nonsyntenic regions of the chromosomes. Here, we report prophage insertions in phytoplasma genomes, describe the gene content of the phytoplasmal prophage elements, and provide evidence to suggest that these phage sequences formed the previously reported SVMs and served as platforms for extensive genetic recombination and capture of laterally transferred genes. These findings, as well as the absence of prophage-based SVMs in the genomes of ancestral relatives *Acho-*

Author contributions: W.W., R.E.D., and Y.Z. designed research; W.W., R.J., and Y.Z. performed research; W.W., R.E.D., R.J., and Y.Z. analyzed data; and W.W., R.E.D., R.J., and Y.Z. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

†To whom correspondence may be addressed. E-mail: robert.davis@ars.usda.gov or yan.zhao@ars.usda.gov.

This article contains supporting information online at www.pnas.org/cgi/content/full/0805237105/DCSupplemental.

leplasma laidlawii (GenBank accession no. CP000896) and *A. palmae* (18), lead us to hypothesize that phages played an important role in shaping the unusual architecture of phytoplasma genomes, fostered adaptations toward a transkingdom parasitic lifestyle, and launched evolution of the phytoplasma clade.

Results and Discussion

We reasoned that, if the nature of the mobile element(s) responsible for phytoplasmal SVM formation could be ascertained, it would be possible to visualize events that triggered evolution of the phytoplasma clade. The presence of repeated phage-related protein-coding genes in the SVMs (18) led us to hypothesize that phages played a significant role in shaping these structures. A global view of the chromosomes of '*Candidatus* Phytoplasma asteris'-related strains OYM and AYWB revealed phage-related genes encoding single-stranded DNA-binding proteins, site-specific DNA methylases, replicative DNA helicases, DNA primases, and viral recombinases, which were scattered but nonrandomly distributed, mainly in SVM regions. Based on these results, we hypothesized that the previously obscure nature of the mobile elements responsible for formation of phytoplasmal genomic SVMs was phage. To test this hypothesis, we analyzed the occurrence, gene content, and organizational features of possible prophage remnants within and external to SVMs.

SVMs Are Composed of Clustered Prophage Remnants. Bacteriophages are one of the most abundant, diverse, and dynamic biological entities in the biosphere (27). The comparatively few phage genomes sequenced are highly varied, and >50% of phage genes encode hypothetical proteins with no current database match (25). Such phage diversity rendered no existing model for easy recognition of unknown prophages that may be present in phytoplasma genomes. Computational analyses of the AYWB and OYM phytoplasma genomes through application of a heuristic program, Prophage Finder (28), revealed prophage loci containing clustered genes encoding phage-related proteins. In addition to those noted above, such genes encoded putative thymidylate kinases, minor tail proteins, and a possible phage tail tape measure protein. Although the program-predicted prophage loci apparently lacked genes that encoded canonical integrases, terminases, or prohead proteases, considered as "corner stone genes" in prophage identification (25, 28–30), the findings warranted further investigation.

Manual inspection of DNA sequences interspersed among the program-predicted prophage loci revealed genes encoding hypothetical proteins having limited database matches and unknown functions. Because these genes, as well as genes encoding assignable phage-related functions, and their intergenic regions were contiguous and not interrupted by operons encoding bacterial housekeeping proteins or ribosomal RNAs, we tentatively considered the regions containing them to be constituent parts of prophage remnants, in accordance with Fouts (25). The repetitive occurrence of the same series of contiguous phage-related genes led us to hypothesize that each series represented no less than one prophage genome remnant. Two or more such presumptive prophage genome remnants, some disrupted by deletions and recombination events, formed tandem arrays or clusters. The gene content and organization of the prophage remnant clusters coincided with that of SVMs (17, 18), supporting the hypothesis that SVMs arose from recurrent and targeted ancestral phage attacks and subsequent genetic recombination.

Prophage Modules and Designation of Phytoplasmal Prophage Genomes. Although each of the presumptive phytoplasmal prophage genome remnants has a highly mosaic structure, close examination revealed modular organizations similar to that of tailed, double-stranded DNA phages in the order *Caudovirales*. This assessment was supported by results from BLAST searches against phage databases, in which predicted proteins encoded by presumptive prophage remnants used as queries yielded significant hits with proteins encoded by sequenced phage genomes, most of which were from nonenveloped, tailed phages having a double-stranded linear DNA genome (supporting information (SI) Table S1 and data not shown). We envision the phytoplasmal prophage modules as follows: a regulation and DNA packaging module (*fliA*, *ssb*, *dam*, *himA*, and *pmp*), a morphogenesis and cell lysis module (tail protein, tail tape measure protein, and other structural component and hypothetical protein-encoding genes), and a replication and recombination module (*yqaJ*, *tmk*, *dnaB*, *dnaG*, *uvrD*, and other replication protein-encoding genes). Note that, in the regulation and DNA packaging module, we designated the previously annotated *hflB* gene as *pmp*, prohead maturation protease gene. Our analysis of the relatively large encoded protein revealed an N-terminal amino acid sequence that we postulate to be a phage capsid domain, an AAA domain near the C-terminal end and a distal protease domain, an architecture consistent with a role as a self-cleaving prohead protease (31). Based on these proposed modular structures, three types of prophage genomes, hereby designated as Ph$\phi\alpha$w, Ph$\phi\alpha$z, and Ph$\phi\alpha$wz, could be distinguished (Fig. 1). Although Ph$\phi\alpha$w is characterized by the presence of a DNA adenine methylase gene (*dam*) and a *uvrD*-type DNA helicase gene, Ph$\phi\alpha$z is characterized by the presence of a thymidylate kinase gene (*tmk*), a viral recombinase gene (*yqaJ*), and a *dnaB*-type DNA helicase gene. The third type of prophage genome, Ph$\phi\alpha$wz, is discernible as a recombinant between Ph$\phi\alpha$w and Ph$\phi\alpha$z.

Termini of Phytoplasmal Prophages. To define clearly the proposed prophage genomes, it was important to distinguish their termini. Assignment of sequences representing putative termini of ancestral, integrated prophages was first based on criteria related to targeted phage attack. Targeted insertion of phages into phytoplasma chromosomes implies site-specific integration at a specific *attB* site or sites, each integration event giving rise to a pair of direct nucleotide sequence repeats, *attL* and *attR*, at the integrated prophage genome termini. Although such direct repeats may be blurred after a long history of postintegration genetic recombination and mutational decay, we hypothesized that recognizable repeats remain at most if not all prophage genome termini. To test this hypothesis, we designed and applied a Perl script to search for pairs of direct repeats surrounding the presumptive ends of the prophage genomes. A signature sequence, AAGTTAGTCTTTTTTT, characterized by an AT-rich stretch, was found in OYM and AYWB phytoplasma chromosomes adjacent to the first and last ORFs of all prophage genomes except for those that had coalesced, recombined, or severely decayed (Fig. 1, Fig. S1, and Table S2). This signature sequence was interpreted as a common core of primary attachment sites *attB* and corresponding *attP*, and the signature sequence locations were therefore interpreted as putative *attL* and *attR* sites. Interestingly, the second half (GTCTTTTTTT) of the putative *attB*/*attP* core is very similar to the first half (GCTTTTTT) of the well documented 15-bp *attB*/*attP* core of *Escherichia coli* lambda phage, GCTTTTTTATACTAA (32). The high degree of *attL*/*attR* sequence conservation across two phytoplasma genomes is consistent with the possibility that phytoplasma phage(s) remain active.

Next, assignment of sequences representing putative termini of integrated prophages was based on criteria related to horizontal gene transfer involving transduction. Because gene transduction by phages involves covalent linkage of a transduced gene at a prophage excision site (33), we reasoned that a non-phage-related gene in the vicinity of a phage terminus would represent a transduced gene. Because the searches based on site-specific integration had located *attL*/*attR*, we examined distal regions for nonphage sequences, and analyzed whether the putative *attL*/*attR* sequences were also adjacent to putatively transduced genes. BLAST searches indicated that two sets of putative protein coding regions (OYM PAMs 321 and 322; 361 and 362) upstream of *fliA* genes (Fig. S1*A*) were not of phage origin, and therefore, were likely to have been transduced

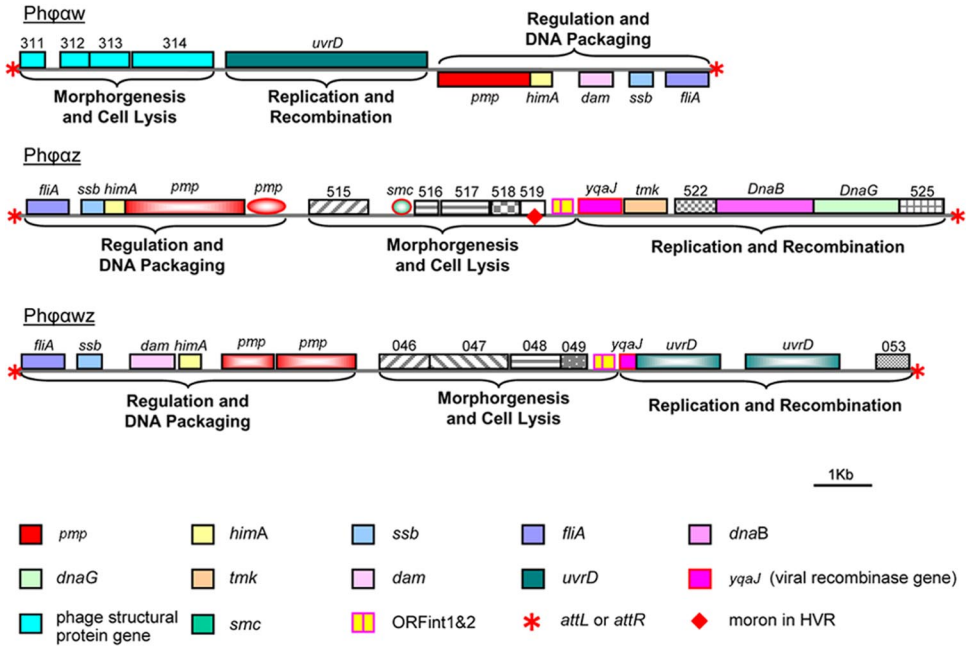

Fig. 1. Diagrammatic representation of three types of prophage genomes, Phφαw, Phφαz, and Phφαwz, integrated in the chromosome of '*Candidatus* Phytoplasma asteris'-related strain onion yellows mild (OYM). Potential protein coding regions (ORFs) are represented by boxes drawn to scale; boxes above and below the horizontal lines represent different directions of transcription. Homologous ORFs share the same color. Truncated ORFs are indicated by hollow shading. ORFs encoding hypothetical proteins are shaded gray; homologous hypothetical protein ORFs have the same pattern. Gene fragment is indicated by an oval. Names of homologous genes are above or below the oval and boxes. Numbers above and below boxes are ORF numbers in the original genome annotation. Horizontal lines represent the DNA sequences. Functional modules are indicated by brackets.

(horizontally transferred) from an unknown host into the recipient phytoplasma chromosome by a transducing phage. A sequence resembling the *attB*/*attP* core sequence was present upstream of the transduced genes. The location of the transduced DNA, adjacent to *fliA* genes, further supports assignment of the *attP* site and the designation of *fliA* as one of the phage genome termini, providing clear definition for designation and delineation of the phytoplasmal prophage genomes.

While this manuscript was in preparation, the complete genome sequence of a '*Ca*. Phytoplasma australiense'-related phytoplasma strain appeared in public databases (GenBank accession no. AM422018); our study of this genome revealed prophage sequences corresponding to those present in the OYM and AYWB genomes (data not shown). Although sequence stretches referred to as potential mobile units (PMUs) in the chromosomes of AYWB phytoplasma and the '*Ca*. Phytoplasma australiense'-related phytoplasma (13, 14) contained genes found in SVMs (18) and in phytoplasma prophage genomes elucidated in the present communication, the PMUs lacked criteria for definition and delineation of presumed mobile elements. For example, *tra5* genes, also annotated as phage integrase genes with similarity to phytoplasma transposases and insertion sequences (IS), were designated as components of PMUs (13, 14). However, clear delineation of the phytoplasmal prophages definitively placed the *tra5* genes outside the boundaries of the ancestral prophage genomes (Fig. S1), indicating that the *tra5* genes were separate mobile elements targeted to the prophage islands (SVMs). Therefore, we propose use of the terms prophage genome and prophage genome remnant (prophage remnant) to refer to the complete and partial prophages, respectively, and retention of the term sequence-variable mosaic (SVM) or use of the term prophage island to accurately describe phytoplasma genome regions containing multiple, clustered copies of complete prophage genomes and/or prophage remnants.

Distribution of Prophage Remnants in OYM and AYWB Genomes. A total of 27 and 16 prophage remnants were identified in the genomes of OYM and AYWB, respectively (Fig. S1). The sizes of most prophage remnants ranged from 10.6 kbp to 25.7 kbp. Some of the prophage remnants had coalesced, forming contiguous prophage clusters up to 113.5 kbp (Fig. S1). In the OYM genome, the combined length of prophage clusters reached 264.2 kbp, ≈31.0% of the entire genome, a remarkably high percentage of prophage content (22–24). In the genome of AYWB, prophage-related sequences totaled 160.2 kbp, nearly 22.7% of the circular chromosome. The differential abundance of prophage-related sequences within the OYM and AYWB genomes (104.0 kbp) accounts for the major part (71.0%) of the overall size difference (146.5 kbp) between the two genomes (853.1 kbp vs. 706.6 kbp). Without the prophage sequences, the genome sizes of OYM and AYWB phytoplasmas would be ≈588.9 and 546.4 kbp, respectively, a size comparable to that of the circular chromosome (580.1 kbp) of *Mycoplasma genitalium*, a mollicute thought to possess a minimal gene complement for cellular life (34). Thus, the presence of prophage elements, largely clustered in SVMs (Fig. 2*A*), defines the uniqueness of phytoplasma genomes.

The prophage-rich loci interrupt the synteny between large, normally contiguous OYM and AYWB chromosomal regions encoding essential cellular functions (Fig. 2*B*). The clustering of prophages suggests repeated targeted attacks by these mobile elements, echoing the proposed explanation for SVM formation (17)—the prophage-rich regions representing hot spots for ancient, recurrent attacks by ancestral phages. The presence of hypervariable segments in these regions (18) suggests that they continue to be active as platforms for genetic recombination.

Morons and Hypervariable Regions as Active Platforms for Gene Acquisition. At least 23 AYWB strain-specific genes (AYWB072, AYWB080, AYWB157, AYWB171, AYWB177, AYWB186, AYWB201, AYWB202, AYWB204, AYWB211, AYWB218, AYWB231, AYWB235, AYWB237, AYWB238, AYWB345, AYWB353, AYWB366, AYWB367, AYWB368, AYWB371,

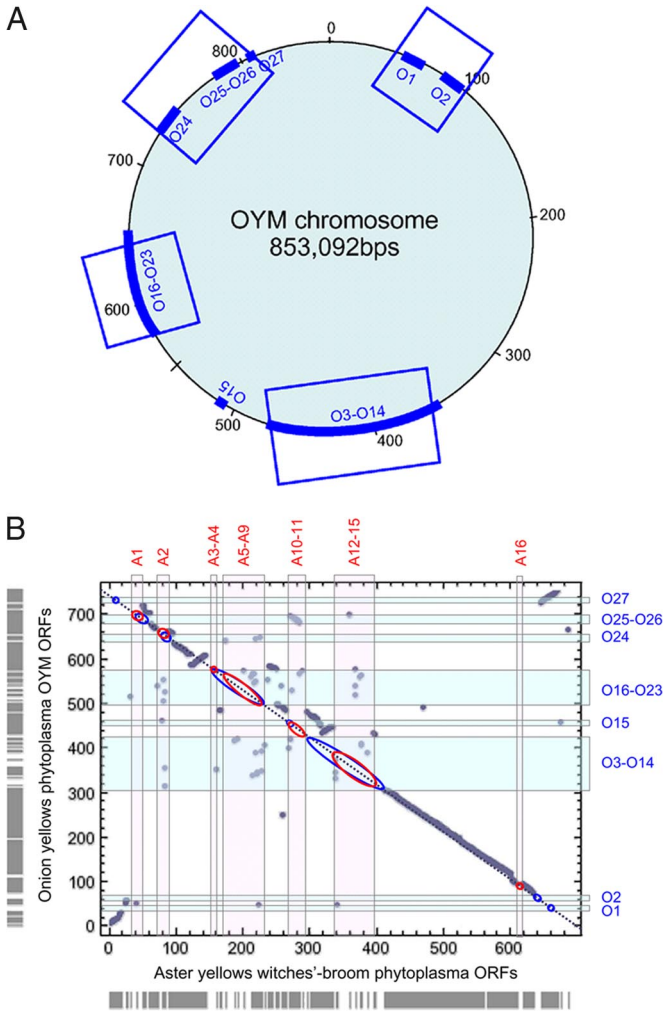

Fig. 2. Distribution of phytoplasmal prophage remnants. (*A*) Diagrammatic representation of the circular chromosome from phytoplasma strain OYM indicating regions (O1 through O27) of prophage integration (thick lines) and regions of sequence variable mosaic (SVM) (boxed). Positions of SVMs are from ref. 17. Numbers represent base positions from the chromosomal origin of replication. (*B*) GenePlot of the chromosomes of '*Ca.* Phytoplasma asteris'-related strains OYM (vertical axis) and AYWB (horizontal axis). Prophage clusters are indicated on the axes by boxes. Blue boxes O1 through O27, OYM prophage clusters. Pink boxes A1 through A16, AYWB prophage clusters. The diagonal dotted line represents the plot expected from complete genome synteny. Blue and red ovals in the OYM and AYWB genomes, respectively, indicate nonsyntenic regions containing prophage genomes and/or genome remnants.

AYWB393, and AYWB615) and 17 OYM strain-specific genes (PAM065, PAM069, PAM321, PAM322, PAM342, PAM361, PAM362, PAM382, PAM383, PAM399, PAM404, PAM405, PAM411, PAM412, PAM519, PAM541, and PAM761) were identified in phytoplasmal prophage loci (Fig. S1). These strain-specific genes (except four transduced genes in OYM, PAM321, PAM322, PAM361, and PAM362) are morons by definition, that is, nonphage DNA segments that are inserted between two phage genes in a phage genome (35). Interestingly, one moron was identified as a 2,516-bp group II intron that was located in an OYM prophage at positions 388234.390749 (in GenBank accession no. AP006628) and contained a reverse-transcriptase gene (PAM342). The OYM group II intron was classified into the mitochondrial class based on its A1 type structure (http://www.fp.ucalgary.ca/group2introns/species.htm#Archae). Group II introns are unique genetic elements capable of self-splicing and mobility, including retrohoming and retrotransposition (36). Our recognition of a group II intron as a moron draws increased attention to prophages as platforms for acquisition of foreign genes and to their role in phytoplasma evolution and strain diversity. These data and further evidence indicated that SVMs were composed in major part of compound structures consisting of prophage genomes and morons, some morons having been targeted to moron-specific integration sites.

Previously we proposed that phytoplasmas possessed an unknown mechanism, which shares some similarities with the integron/mobile gene cassette system found in other bacteria, to explain the presence of diverse genes in hypervariable regions (HVRs)—sequence-variable regions each flanked by a conserved ATP-dependent Zn protease gene and a conserved palindromic DNA sequence (18). Based on the phage-derived nature of surrounding sequences, the present study identified, in the genomes OYM and AYWB phytoplasmas, four and two morons, respectively, in HVRs located in the vicinity of a phage-derived viral recombinase. The variable gene content, structural motifs, and presence of the viral recombinase (as well as presence of the group II intron-carried reverse transcriptase, in the case of OYM) support the hypothesis that the HVRs represent potentially active platforms for gene acquisition. Conceivably, morons are targeted for insertion into specific sites in the HVR by mechanisms resembling those of integron/mobile gene cassette systems, which afford targeting of acquired genes to specific sites in bacterial chromosomes (this study; 18, 37). Indeed, it is feasible that morons represent a subclass of mobile gene cassettes, and that morons and mobile gene cassettes have similar origins.

Prophage-Derived Sequences form Distinct Genomic Islands in Phytoplasma Genomes. Evolutionary establishment of phytoplasmal transkingdom parasitism and pathogenicity reasonably must have required the presence of a constellation of genes capable of encoding products that not only served as bona fide virulence factors, but also functioned as fitness factors enhancing the incipient parasite's ability to colonize, multiply in, and shuttle between insect and plant hosts. Although such capabilities may evolve in vertically inherited genes through duplications and mutations, niche expansion may be achieved more rapidly through lateral gene transfer (21). Because numerous genes within the phage-based SVMs encode phytoplasma membrane-targeted proteins (13, 17, 18), the ancient and recurrent attacks by phages, coupled with insertions of morons, must have provided abundant opportunities for phytoplasma progenitors to acquire horizontally transferred genes encoding proteins involved in microbe–host interactions and niche adaptation.

Indeed, massive accumulation of clustered phage-derived sequences in recipient chromosomes of AYWB and OYM, respectively, formed a large pool of horizontally acquired genes in the form of genomic islands exhibiting distinct physical and genetic properties (Tables 1 and 2, Fig. S1, and Table S3). For example, prophage genomic islands of AYWB and OYM phytoplasmas are distinguished, from host genomic segments vertically inherited from acholeplasma-like ancestors, by overall G+C content and codon position G+C content (Table 1). Collectively, phytoplasmal prophage genes have a lower mol% G+C content (25.99% for AYWB and 26.78% for OYM) than do their host genes (29.32% for AYWB and 30.15% for OYM). The difference in mol% G+C content between prophage and host protein coding regions was ≈3.4%, indicating that codon usage differs between phytoplasmal phage genes and host genes. The mol% G+C differences were even more striking in the first (5.27% for AYWB and 5.25% for OYM) and second (4.66% for AYWB and 4.92% for OYM) base positions of the triplet codons. These positions carry more weight in determining amino acid compositions than the third (wobble) codon position, indicating a remarkable difference in overall amino acid compositions between phage and host gene products.

By virtue of their distinct base compositions, the prophage-

Table 1. Codon position G + C contents of phytoplasma host, prophage, transposon, and transduced ORFs in '*Ca*. Phytoplasma asteris'-related strains AYWB and OYM

ORFs	Overall*	Position 1†	Position 2†	Position 3†
AYWB All	28.54 (148,730)	38.58 (67,018)	27.35 (47,507)	19.69 (34,205)
AYWB host	29.32 (118,895)	39.88 (53,913)	28.39 (38,384)	19.68 (26,598)
AYWB phage	25.99 (27,183)	34.61 (12,064)	23.73 (8,272)	19.64 (6,847)
AYWB transposon	24.19 (2,652)	28.48 (1,041)	23.28 (851)	20.79 (760)
OYM All	29.09 (181,747)	39.49 (82,247)	27.59 (57,466)	20.18 (42,034)
OYM host	30.15 (131,069)	41.14 (59,628)	29.10 (42,175)	20.19 (29,266)
OYM phage	26.78 (47,893)	35.89 (21,392)	24.18 (14,411)	20.28 (12,090)
OYM transposon	24.12 (2,147)	29.42 (873)	23.05 (684)	19.89 (590)
OYM transduced	28.43 (638)	47.33 (354)	26.20 (196)	11.76 (88)

*Overall G + C content of all three positions in mol percentage. Numbers in parentheses are the total number of G + C residues.
†Cumulative G + C content of nucleotide positions in triplet codons.

derived genomic islands represent low G+C isochors (macroisochors) in the phytoplasma genomes. Such low G+C isochors are believed to promote genic and segmental duplications and subsequent genetic recombination in resident genomes (38, 39). Notably in this regard, a 21-kbp large segmental duplication (LSD) is present in a prophage-derived genomic island in the OYM phytoplasma chromosome (Fig. S1). Spanning genome locations 357003.382562 and 398072.423447, the LSD encompasses four clustered phage remnants and contains 56 ORFs. LSDs frequently occur in eukaryotic genomes and play an important role in genome evolution (40, 41).

The AYWB and OYM prophage genomic islands were further distinguished from their respective host DNAs on the basis of differential dinucleotide relative abundance (DRA) (Table 2 and Table S3), a unique signature of an organism's DNA that has been exploited in studies of genomic heterogeneity and lateral gene transfer (42, 43). Our results showed that the DRA distance (DRAD) of phytoplasmal prophage islands vs. host DNA is 116.51 for AYWB and 127.47 for OYM; both are much higher than those of bacterial pathogenicity islands (PAIs) identified in plant pathogens including those causing crown gall, citrus canker, Pierce's disease, and black rot in plants, as well as in bacteria causing diseases of humans and animals (http://www.pathogenomics.sfu.ca/islandpath/current/IPindex.pl) (44, 45). Thus, several lines of evidence indicate that SVMs are prophage-derived phytoplasmal PAIs (This communication). Because they contain numerous genes encoding proteins with an N terminus signal peptide and/or transmembrane segment(s) (data not shown), functional studies are warranted to determine whether these secreted and membrane-targeted proteins are involved in phytoplasma–host interactions—parasitism and pathogenicity, as suggested earlier (13, 18).

Table 2. Pairwise dinucleotide relative abundance distance (DRAD) between phytoplasma host, prophage, transposon, and transduced ORFs of AYWB and OYM

ORFs	Total DRAD	Average DRAD × 1,000
AYWB		
Phage vs. host DRAD [δ(p, h)]	1.86419	116.51
Transposon vs. host DRAD [δ(tn, h)]	2.09038	130.65
Phage vs. transposon DRAD [δ(p, tn)]	1.71796	107.37
OYM		
Phage vs. host DRAD [δ(p, h)]	2.03956	127.47
Transposon vs. host DRAD [δ(tn, h)]	2.56746	160.47
Transduced vs. host DRAD [δ(td, h)]	2.14788	134.24
Phage vs. transposon DRAD [δ(p, tn)]	1.62702	101.69
Phage vs. transduced DRAD [δ(p, td)]	3.34990	209.37
Transposon vs. transduced DRAD [δ(tn, td)]	4.07758	254.85

Within the phage-derived genomic islands are targeted insertions of transposable elements (*tra5*), whose DNA physical signatures indicate that their origins differ from those of the prophage genomes (Tables 1 and 2 and Table S3). Independent origins are also consistent with the occurrence of *tra5* sequences outside of the SVMs.

Clues to the Origin of Phytoplasma Phages. The origin(s) of the phage(s) giving rise to these unique features of phytoplasmal genome architecture remained unresolved. To resolve the origin(s) of the prophages, we queried proteins encoded by the prophage remnants against phage protein databases. Proteins yielding the most significant hits (matches) in the similarity-based BLAST searches were those of phages from low G+C Gram-positive walled bacteria and from Gram-negative Gamma *Proteobacteria* (Table S1). These results yielded no clear indication of a particular bacterial group as the origin or major source of phytoplasmal prophage genes, undoubtedly because available phage genome sequence information is still relatively limited compared with the planet's estimated 10^{31} phages (46).

We reasoned that successful adaptation of phages to bacterial host niches involves coevolution with their bacterial hosts, during which opportunities arise for gene exchanges between host, phage, and genomes of organisms occupying the same niche. We hypothesized that traces or footprints from such hypothetical ancient exchanges remain and point to bacterial host origin(s) of the phytoplasmal prophages. To test this hypothesis, we selected three phytoplasmal prophage gene products (HimA, DnaG, and TMK) and their phytoplasma host counterparts, and used the sequences as protein BLAST queries against all microbial genome sequences available. Preliminary phylogenetic analyses indicated that these proteins were distinct from their respective host counterparts and clustered with orthologues from class *Mollicutes* or from *Buchnera aphidicola*, a symbiont of aphids (data not shown); however, the analysis of very divergent proteins from diverse bacteria could have caused very divergent lineages to cluster together. Nevertheless, phytoplasmal prophage TMK proteins, for example, shared greater amino acid sequence similarity with TMKs of *Buchnera* spp. than with phytoplasmal host TMKs (Table S4). These findings point to insect-inhabiting microflora as a possible source of phytoplasma phage genes. Like aphids that harbor symbiotic bacteria, leafhoppers and psyllids, the principal insect vectors of phytoplasmas, similarly host symbiotic bacteria (47). Conceivably, horizontal gene exchange occurred among insect genomes, insect-inhabiting microflora, including symbionts, and prophages or their progenitors. This concept implicates insects as an important niche contributing to the origins of phages involved in formation of phytoplasmal pathogenicity islands, SVMs.

Recognition and delineation of the prophage-derived genomic islands presents a model for understanding phytoplasma genome

EVOLUTION

evolution. Thus, emergence and adaptive evolution of the phytoplasma clade can be envisioned as a process enabled by ancient and recurrent phage attacks, site-specific chromosomal integration of prophage genomes, and subsequent genetic recombination and duplication events in an acholeplasma-like ancestor. Presence of phage-derived SVMs in phylogenetically diverse phytoplasma lineages and their absence in *Acholeplasma* genomes supports the concept that SVMs formed at an early stage in phytoplasma evolution, after evolutionary divergence of acholeplasmas and phytoplasmas, and that prophage integration is at the root of evolutionary emergence of the phytoplasma clade. We hypothesize that phage-mediated gene exchange enabled phytoplasmal transkingdom parasitism and pathogenicity and triggered events that launched evolution of the phytoplasma clade.

Methods

Initial identification of prophage elements in the genomes of two '*Candidatus* Phytoplasma asteris'-related strains OYM (GenBank accession no. AP006628) and AYWB (GenBank accession no. CP000061) was performed by using prophage loci prediction tool devised by Bose and Baber (28). Prophage elements identified by Prophage_Finder were subjected to multiple bioinformatics analyses (See *SI Methods*).

Genetic synteny between AYWB and OYM was assessed by using GenPlot (http://www.ncbi.nlm.nih.gov/sites/entrez?db = genomeprj), which performs pairwise comparisons on two sets of predicted proteins encoded sequentially along the genome sequences, and plots the symmetrical best matches according to the indices of the corresponding gene in the two genomes as the X and Y coordinates.

Variation in overall and codon position G+C content was analyzed by a Perl program developed in the present study. The program accepts multiple protein-encoding sequences (ORFs) in FASTA format, breaks the ORFs into triplet codons, counts the occurrence of guanidine and cytosine bases in each position, and calculates mol percentage G+C content in each codon position for the input ORFs.

The dinucleotide relative abundance (DRA) values (ρ_{XY}) were calculated according to the indices defined by Burge *et al.* (48): $\rho_{XY} = f_{XY}/f_X \times f_Y$, where f_{XY} is the observed frequency of the dinucleotide XY, and f_X and f_Y are the expected frequencies of bases X and Y, respectively. DRA distance (DRAD) values between the prophages and their respective phytoplasma host [δ(p,h)] were determined by using the formula of Karlin and Mrázek (49): $\delta(p,h) = (1/16)\,\Sigma\,|\,\rho_{XY}(p) - \rho_{XY}(h)\,|$. Final DRAD values were expressed as $\delta(p,h) \times 1{,}000$.

Putative phage integration sites (*attL* and *attR*) were identified by application of a Perl program developed in the present study. The program searches through a pair of input sequences and identifies perfect and imperfect pattern matches (direct repeats) with a user-definable pattern length and allowable mismatches (16 bp and 6 bp, respectively, in the present study).

Minimal evolution analysis was conducted with software MEGA4 (50) by using the close neighbor interchange (CNI) algorithm. The initial tree for the CNI search was obtained by the neighbor-joining method. The reliability of analysis was subjected to a bootstrap test with 5,000 replicates. The consensus tree from each individual analysis was used to infer phylogenetic relationships.

ACKNOWLEDGMENTS. We thank Xiaobing Suo for writing Perl programs to calculate codon position G+C content, DRA and DRAD values; for direct repeats search; and for batch retrieval of protein sequence data for phylogenetic analyses. This work was funded by the U.S. Department of Agriculture-Agricultural Research Service.

1. Gundersen DE, Lee IM, Rehner SA, Davis RE, Kingsbury DT (1994) Phylogeny of mycoplasmalike organisms (phytoplasmas): A basis for their classification. *J Bacteriol* 176:5244–5254.
2. Lee I-M, Davis RE, Gundersen-Rindal DE (2000) Phytoplasma: Phytopathogenic mollicutes. *Annu Rev Microbiol* 54:221–255.
3. Chang C-J (1998) Pathogenicity of aster yellows phytoplasma and *Spiroplasma citri* on periwinkle. *Phytopathology* 88:1347–1350.
4. Lepka P, Stitt M, Moll E, Seemüller E (1999) Effect of phytoplasmal infection on concentration and translocation of carbohydrates and amino acids in periwinkle and tobacco. *Physiol Mol Plant Pathol* 55:59–68.
5. Bertamini M, Grando MS, Muthuchelian K, Nedunchezhian N (2002) Effect of phytoplasmal infection on photosystem II efficiency and thylakoid membrane protein changes in field grown apple (*Malus pumila*) leaves. *Physiol Mol Plant Pathol* 61:349–356.
6. Bertamini M, Nedunchezhian N, Tomasi F, Grando MS (2002b) Phytoplasma [Stolbur-subgroup (Bois Noir-BN)] infection inhibits photosynthetic pigments, ribulose-1,5-bisphosphate carboxylase and photosynthetic activities in field grown grapevine (*Vitis vinifera* L. cv. Chardonnay) leaves. *Physiol Mol Plant Pathol* 61:357–366.
7. Curković-Perica M, Lepedus H, Seruga Musić M (2007) Effect of indole-3-butyric acid on phytoplasmas in infected *Catharanthus roseus* shoots grown *in vitro*. *FEMS Microbiol Lett* 268:171–177.
8. Seemüller E, Garnier M, Schneider B (2002) Mycoplasmas of plants and insects. *Molecular Biology and Pathology of Mycoplasmas*, eds Razin S, Herrmann R (Kluwer Academic/Plenum Publishers, London), pp 91–116.
9. Davis RE, Jomantiene R, Zhao Y (2005) Lineage-specific decay of folate biosynthesis genes suggests ongoing host adaptation in phytoplasmas. *DNA Cell Biol* 24:832–840.
10. Lee I-M, Zhao Y, Davis RE, Wei W, Martini M (2007) Prospects of DNA-based systems for differentiation and classification of phytoplasmas. *Bull Insectol* 60:239–244.
11. Wei W, Davis RE, Lee I-M, Zhao Y (2007) Computer-simulated RFLP analysis of 16S rRNA genes: Identification of ten new phytoplasma groups. *Int J Syst Evol Microbiol* 57:1855–1867.
12. Oshima K, *et al.* (2004) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat Genet* 36:27–29.
13. Bai X, *et al.* (2006) Living with genome instability: The adaptation of phytoplasmas to diverse environments of their insect and plant hosts. *J Bacteriol* 188:3682–3696.
14. Tran-Nguyen LT, Kube M, Schneider B, Reinhardt R, Gibb KS (2008) Comparative genome analysis of '*Candidatus* Phytoplasma australiense' (subgroup *tuf*– Australia I; *rp*-A) and '*Ca.* P asteris' strains OY-M and AY-WB. *J Bacteriol* 190:3979–3991.
15. Cimerman A, Arnaud G, Foissac X (2006) Stolbur phytoplasma genome survey achieved using a suppression subtractive hybridization approach with high specificity. *Appl Environ Microbiol* 72:3274–3283.
16. Kube M, Schneider B, Reinhardt R, Seemüller E (2007) First look into the genome sequence of '*Candidatus* Phytoplasma mali' compared with '*Candidatus* Phytoplasma asteris' strains OY-M and AY-WB. *Bull Insectol* 60:113–114.
17. Jomantiene R, Davis RE (2006) Clusters of diverse genes existing as multiple, sequence-variable mosaics in a phytoplasma genome. *FEMS Microbiol Lett* 255:59–65.
18. Jomantiene R, Zhao Y, Davis RE (2007) Sequence-variable mosaics: Composites of recurrent transposition characterizing the genomes of phylogenetically diverse phytoplasmas. *DNA Cell Biol* 26:557–564.
19. Kakizawa S, Oshima K, Namba S (2006) Diversity and functional importance of phytoplasma membrane proteins. *Trends Microbiol* 14:254–256.
20. Arashida R, *et al.* (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.
21. Brüssow H, Canchaya C, Hardt WD (2004) Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* 68:560–602.
22. Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H (2003) Prophage genomics. *Microbiol Mol Biol Rev* 67:238–276.
23. Hayashi T, *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8:11–22.
24. Bordenstein SR, Reznikoff WS (2005) Mobile DNA in obligate intracellular bacteria. *Nat Rev Microbiol* 3:688–699.
25. Fouts DE (2006) Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 34:5839–5851.
26. Srividhya KV, *et al.* (2007) Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS ONE* 2:e1193.
27. Brüssow H, Hendrix RW (2002) Phage genomics: Small is beautiful. *Cell* 108:13–16.
28. Bose M, Barber RD (2006) Prophage Finder: A prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* 6:223–227.
29. Rao GV, Mehta P, Srividhya KV, Krishnaswamy S (2005) A protein similarity approach for detecting prophage regions in bacterial genomes. *Genome Biol* 6:11.
30. Srividhya KV, *et al.* (2006) Database and comparative identification of prophages. *LNCIS* 344:863–868.
31. Casjens SR, *et al.* (2005) The generalized transducing *Salmonella* bacteriophage ES18: Complete genome sequence and DNA packaging strategy. *J Bacteriol* 187:1091–1104.
32. Landy A, Ross W (1977) Viral integration and excision: Structure of the lambda *att* sites. *Science* 197:1147–1160.
33. Miller RV (2004) Bacteriophage-mediated transduction: An engine for change and evolution. *Microbial Evolution: Gene Establishment, Survival, and Exchange*, eds Miller RV, Day MJ (ASM Press, Washington DC), pp 144–157.
34. Fraser CM, *et al.* (1995) The minimal gene complement of Mycoplasma genitalium. *Science* 270:397–403.
35. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. *Trends Microbiol* 8:504–508.
36. Dai L, Zimmerly S (2002) Compilation and analysis of group II intron insertions in bacterial genomes: Evidence for retroelement behavior. *Nucleic Acids Res* 30:1091–1102.
37. Recchia GD, Hall RM (1997) Origins of the mobile gene cassettes found in integrons. *Trends Microbiol* 5:389–394.
38. Forsdyke DR (2004) Regions of relative GC% uniformity are recombinational isolators. *J Biol Sys* 12:261–271.
39. Lee S-J, Mortimer JR, Forsdyke DR (2004) Genomic conflict settled in favour of the species rather than the gene at extreme GC percentage values. *Appl Bioinformatics* 3:219–228.
40. Zhang LQ, Henry HS, Lu WY, Chung JY, Li WH (2005) Patterns of segmental duplication in the human genome. *Mol Biol Evol* 22:135–141.
41. Koszul R, Dujon B, Fischer G (2006) Stability of large segmental duplications in the yeast genome. *Genetics* 172:2211–2222.
42. Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet* 11:283–290.
43. Karlin S (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* 1:598–610.
44. Karlin S (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* 9:335–343.
45. Hsiao W, Wan I, Jones SJ, Brinkman FS (2003) IslandPath: Aiding detection of genomic islands in prokaryotes. *Bioinformatics* 19:418–420.
46. Wommack KE, Colwell RR (2000) Virioplankton: Viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* 64:69–114.
47. Bordenstein SR, Wernegreen JJ (2004) Bacteriophage flux in endosymbionts (*Wolbachia*): Infection frequency, lateral transfer, and recombination rates. *Mol Biol Evol* 21:198.
48. Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* 89:1358–1362.
49. Karlin S, Mrázek J (1997) Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci USA* 94:10227–10232.
50. 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.