

Comparative genomics identifies genes shared by distantly related insect-transmitted plant pathogenic mollicutes

Xiaodong Bai^a, Jianhua Zhang^a, Ian R. Holford^b, Saskia A. Hogenhout^{a,*}

^a Department of Entomology, The Ohio State University, Ohio Agricultural Research and Development Center, 1680 Madison Avenue, Wooster, OH 44691, USA

^b Molecular and Cellular Imaging Center (MCIC), The Ohio State University, Ohio Agricultural Research and Development Center, 1680 Madison Avenue, Wooster, OH 44691, USA

Received 28 January 2004; received in revised form 12 April 2004; accepted 22 April 2004

First published online 10 May 2004

Abstract

Phytoplasmas and spiroplasmas are distantly related insect-transmitted plant pathogens within the class Mollicutes. Genome sequencing projects of phytoplasma strain Aster Yellows-Witches' Broom (AY-WB) and *Spiroplasma kunkelii* are near completion. Complete genome sequences of seven obligate animal and human pathogenic mollicutes (*Mycoplasma* and *Ureaplasma* spp.), and OY phytoplasma have been reported. Putative ORFs predicted from the genome sequences of AY-WB and *S. kunkelii* were compared to those of the completed genomes. This resulted in identification of at least three ORFs present in AY-WB, OY and *S. kunkelii* but not in the obligate animal and human pathogenic mollicutes. Moreover, we identified ORFs that seemed more closely related between AY-WB and *S. kunkelii* than to their mycoplasma counterparts. Phylogenetic analyses using parsimony were employed to study the origin of these genes, resulting in identification of one gene that may have undergone horizontal gene transfer. The possible involvement of these genes in plant pathogenicity is discussed.

© 2004 Federation of European Microbiological Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Comparative genomics; Plant pathogen; Mollicutes; Spiroplasma; Phytoplasma; Mycoplasma

1. Introduction

Mollicutes, characterized by small genomes and no cell wall, are believed to have diverged from a Gram-positive bacterial ancestor in the lactobacillus group [1,2]. Within the class Mollicutes, an early evolutionary split occurred between the AAA (*Asteroleplasma*, *Anaeroplasm*, and *Acholeplasma*) branch and the SEM (*Spiroplasma*, *Entomoplasm*, and *Mycoplasma*) branch, both of which independently underwent genome reductions [2]. Apparently, the conversion of UGA from a stop codon to a tryptophan codon in the SEM branch occurred shortly after the split of the two branches. The SEM branch contains several genera, including *Spi-*

roplasma, *Entomoplasm*, *Mesoplasm*, *Mycoplasma*, and *Ureaplasma*. Spiroplasmas are believed to be evolutionary early mollicutes and did not undergo as many gene loss events as members of other genera [2].

At the start of the genomic era, mollicutes have attracted much attention because of their small genomes and their clinical and agricultural impact. Six mollicutes genomes, five *Mycoplasma* spp., and one *Ureaplasma* sp., have been fully sequenced, representing obligate human and mammal pathogens of the genus *Mycoplasma* of the SEM branch. At the time of preparation of this manuscript, genome sequencing projects of three other mycoplasmas were in progress: the rodent polyarthritid pathogen *Mycoplasma arthritidis*, the contagious caprine pleuropneumonia (CCPP) pathogen *Mycoplasma capricolum*, and the contagious bovine pleuropneumonia (CBPP) pathogen *Mycoplasma mycoides* subsp. *mycoides* SC (small colony). Later, the

* Corresponding author. Tel.: +1-330-263-3730; fax: +1-330-263-3686.

E-mail address: hogenhout.1@osu.edu (S.A. Hogenhout).

complete genome sequences of *M. mycoides* subsp *mycoides* SC and Onion Yellows (OY) phytoplasma were released and published [3,4].

Genome sequencing projects are in progress for *Spiroplasma kunkelii* (<http://www.genome.ou.edu/spiro.html>) and the phytoplasma strain Aster Yellows-Witches' Broom (AY-WB, <http://www.oardc.ohio-state.edu/phytoplasma>). *S. kunkelii* and phytoplasmas are insect-transmitted plant pathogens that replicate in both insect vectors and plant hosts. Interestingly, *S. kunkelii* and phytoplasmas are strikingly similar in their infection patterns of insects and plants. Both are restricted to phloem tissues of plant hosts, from where they are acquired by phloem-feeding insects, and subsequently invade and replicate in the cells of insect gut and other tissues. Interestingly, although *Spiroplasma* species and all phytoplasmas described so far share similar infection patterns and environmental niches, they are distantly related within two branches of the class Mollicutes. Based on phylogenies of 16S rDNA and *tuf* genes, membrane composition, codon usage, and metabolism [2], spiroplasmas were grouped in the SEM branch with *Mycoplasma* and *Ureaplasma* spp., while phytoplasmas were grouped in the AAA branch with *Acholeplasma* spp.

This study was initiated based on the hypothesis that genes shared by evolutionarily divergent insect-transmitted plant pathogens but absent from obligate human and animal pathogens are likely important for insect transmission and/or plant pathogenicity. Using computer-assisted analysis, we have identified at least three open reading frames (ORFs) that were present in *S. kunkelii* and AY-WB but absent from mycoplasmas. We have also identified ORFs that do not match the 16S rDNA and *tuf* phylogenies. The involvement of the ORFs in pathogenicity is discussed.

2. Materials and methods

2.1. Genome sequences

The 16 contigs totaling 695 kb of the estimated 800 kb AY-WB genome were obtained from the phytoplasma genome sequencing project website (<http://www.oardc.ohio-state.edu/phytoplasma>). The 46 contigs totaling 1.5 Mb of the estimated 1.6 Mb *S. kunkelii* CR2-3x genome were obtained from the publicly accessible *S. kunkelii* genome sequencing project website (<http://www.genome.ou.edu/spiro.html>). Complete

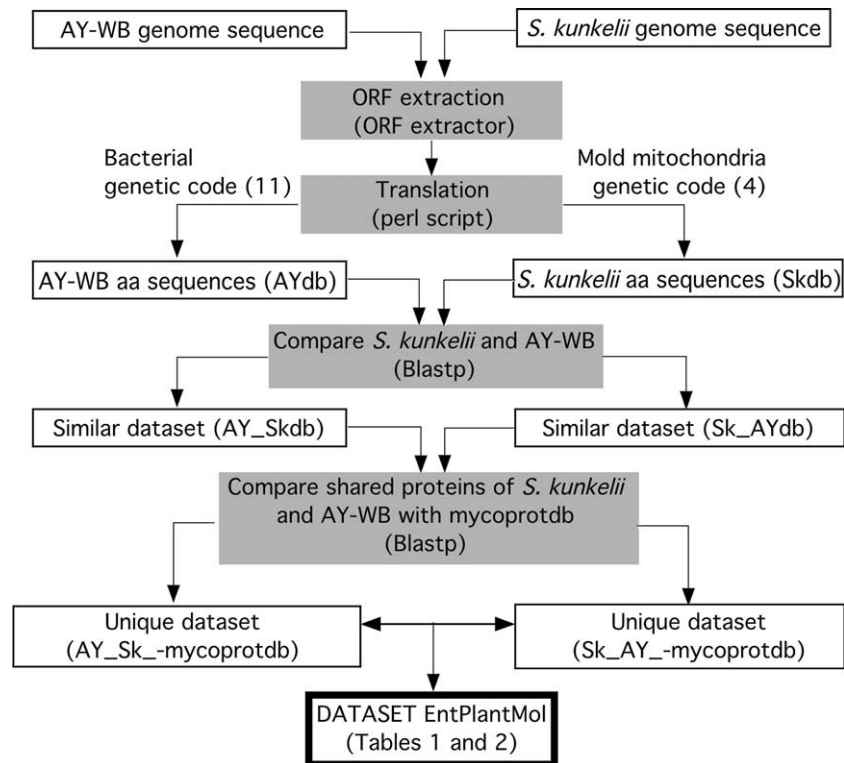


Fig. 1. Algorithms employed to extract proteins that are common between the insect-transmitted plant pathogens AY-WB phytoplasma and *S. kunkelii* but are absent from five *Mycoplasma* spp. and *U. urealyticum*. See Section 2 for details. Similar dataset consists of proteins that are similar in AY-WB and *S. kunkelii*, while unique dataset consists of proteins that are similar between AY-WB and *S. kunkelii* but absent from five *Mycoplasma* spp. and *U. urealyticum*. Shaded text boxes are operations with the programs indicated in parentheses. Open text boxes are datasets either as input or output of the operations. Bacterial and mold mitochondria genetic codes are from NCBI taxonomy databases [11,12].

mycoplasma genome sequences were obtained from GenBank, including *Mycoplasma genitalium* (NC_000908) [5], *Mycoplasma pneumoniae* (NC_000912) [6], *Ureaplasma urealyticum* (NC_002162) [7], *Mycoplasma pulmonis* (NC_002771) [8], *Mycoplasma penetrans* (NC_004432) [9], and *Mycoplasma gallisepticum* (NC_004829) [10].

2.2. Comparative genome analysis

Genome comparisons were conducted as illustrated in Fig. 1. Genome sequences were downloaded onto a Linux workstation and used as input files for the ORF Extractor (http://www.oardc.ohio-state.edu/mcic/bioinformatics/bio_software/bio_software.html). ORFs were defined as starting with ATG and ending with in-frame TAG, TAA, or TGA for AY-WB, or TAG and TAA for *S. kunkelii* and all *Mycoplasma* and *Ureaplasma* spp. [2]. ORFs longer than 90 bp were extracted in FASTA format. Subsequently, only the longest ORF within a set of ORFs having stop codons at the same positions was extracted. ORFs in nucleotide sequences were translated into amino acid (aa) sequences using a Perl translation program, using translation table 11 (bacterial code) for AY-WB and translation table 4 (Mold mitochondria code) for all others [11,12]. This generated datasets AYdb for AY-WB, Skdb for *S. kunkelii*, and mycoprotodb for the five *Mycoplasma* spp. and *U. urealyticum*. Subsequently, AYdb and Skdb were compared using stand-alone BLAST (Basic Local Alignment Search Tool) package [13] with the expectation (*E*)-value threshold of 10^{-8} . Proteins having significant similarity ($E < 10^{-8}$) were extracted from AYdb to generate AY_Skdb and from Skdb to generate Sk_AYdb. Subsequently, AY_Skdb and Sk_AYdb were compared to mycoprotodb and proteins with non-significant hits ($E > 10^{-8}$) or no hits were extracted from AY_Skdb to generate AY_Sk-mycoprotodb and from Sk_AYdb to generate Sk_AY-mycoprotodb. Proteins within AY_Sk-mycoprotodb and Sk_AY-mycoprotodb were annotated based on sequence similarity searches against NCBI non-redundant (nr) database and compared manually to identify common protein sequences. Identified proteins were validated by manual comparison with the annotated genome sequences of mycoplasmas and ureaplasma. After the finish of this study, the genome sequences of OY phytoplasma [4] and *M. mycoides* subsp. *mycoides* SC strain [3] were released. The identified proteins were searched against the annotated proteins of these organisms using the BLAST algorithm [13].

Negative logistic plots of best *E*-values for each query were generated for searches of: (i) AYdb against Skdb and mycoprotodb and (ii) Skdb against AYdb and mycoprotodb. For comparable quantitative assessment, an *E*-value of 0.0 was set to 10^{-200} , and proteins with no significant hits were assigned *E*-values of 1000.

2.3. Phylogenetic analysis

Protein sequences for phylogenetic analysis were extracted from NCBI Entrez database. Sequence alignments were produced using ClustalW [14] and used as inputs for phylogenetic analysis using PAUP (Phylogenetic Analysis Using Parsimony) program [15].

2.4. Accession numbers

AY-WB amino acid sequences identified in this study were deposited in GenBank with the Accession Numbers as follows: AAA type ATPase (AtA), AY533109; cmp-binding factor (CBF), AY533110; cytosine deaminase, AY533111; hypothetical protein, AY533112; cation transport P-ATPase, AY533113; polynucleotide phosphorylase (PNPase), AY533114; ppGpp synthetase, AY533115; YlxR protein, AY533116.

3. Results

3.1. Extraction of ORFs

As expected, ORF Extractor generated more putative ORFs than currently annotated for the completed mollicute genomes or predicted based on the estimation that ORFs have an average length of 1 kb [16]. However, for this study ORF Extractor was preferred, because, although it generates more false ORFs, it decreases the chance of omitting putative ORFs [17]. Further, since most ORFs starting with alternative start codons have an in-frame ATG elsewhere, the ORF database generated by ORF Extractor included most of the annotated ORFs (complete or partial) of the completed mollicute genomes currently present in GenBank. Of the 4332 mycoplasma and ureaplasma ORFs downloaded from GenBank, 991 ORFs (22.7%) start with an alternative start codon. Of the ORFs starting with an alternative start codon, 984 ORFs (99.3%) contained an in-frame ATG somewhere in the ORF. Thus, only 0.7% of the putative ORFs starting with alternative start codons present in the GenBank database have been excluded from the ORF Extractor database. This is only 0.2% of all the 4332 annotated mycoplasma and ureaplasma (i.e. members of *M. pneumoniae* and *Mycoplasma hominis* groups and *U. urealyticum*) ORFs present in GenBank.

To minimize the number of false-positives produced by the method, only the longest ORF within a set of ORFs having stop codons at the same position was extracted for subsequent analysis. Translation of the ORFs into amino acid sequences generated AYdb for AY-WB, Skdb for *S. kunkelii*, and mycoprotodb for the five *Mycoplasma* species and *U. urealyticum*.

3.2. Identification of four proteins that are present in AY-WB and *S. kunkelii* but absent from mycoplasmas

Amino acid sequence similarity searches were employed to identify proteins shared between AY-WB and *S. kunkelii*. AYdb and Skdb were searched against each other using stand-alone BLAST package [13]. Two hundred and ninety proteins within AYdb had significant similarity (E -value $< 10^{-8}$) to proteins within Skdb, whereas 260 proteins within Skdb had significant similarity to proteins within AYdb. E (expectation)-value in

BLAST search is defined as “the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance”, and it depends on the size of the search database and the scoring system [18]. Thus, it was expected that the number of proteins with significant similarity for the two independent searches would differ because of different database sizes.

To identify shared AY-WB and *S. kunkelii* proteins that are not present in animal and human pathogenic mycoplasmas and ureaplasmas, AY_Skdb and

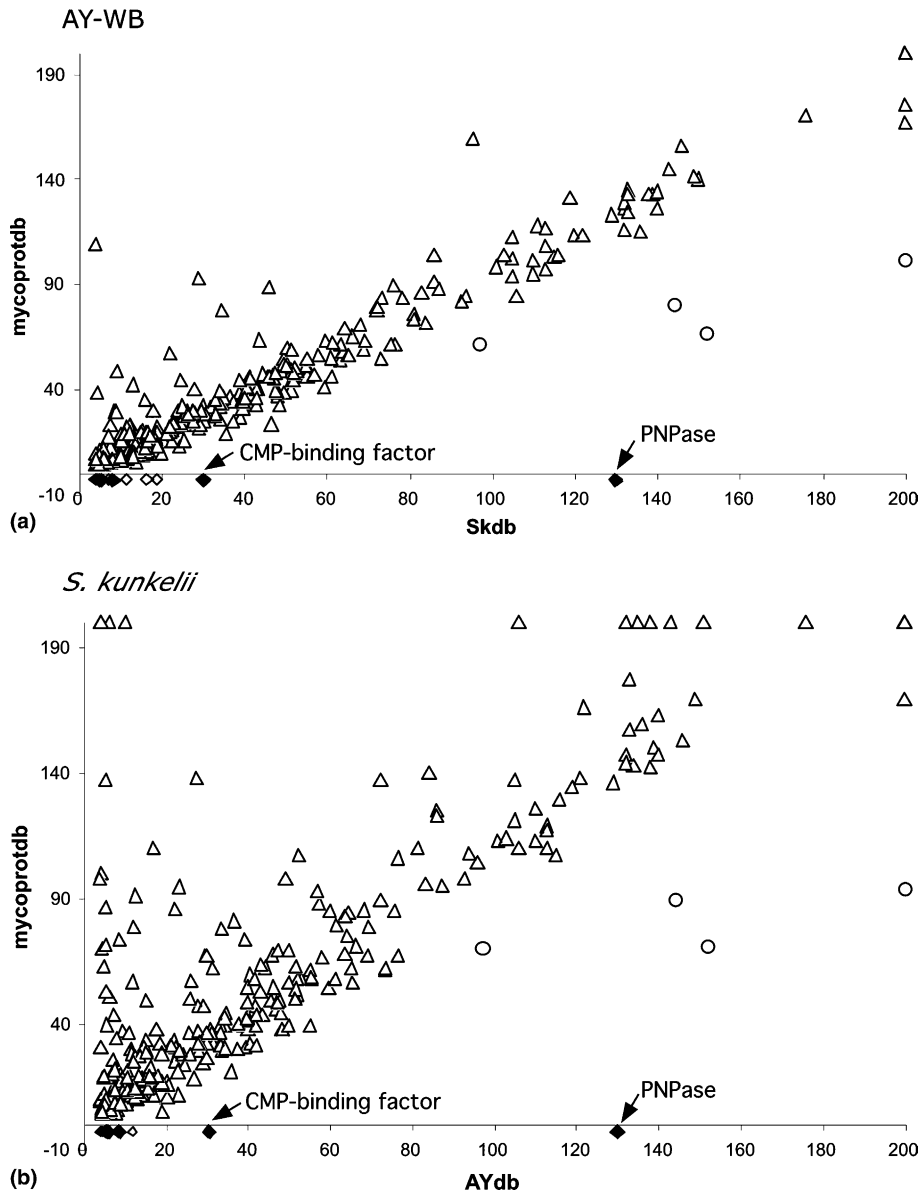


Fig. 2. Graphical representation of comparative analysis results. (a) Negative logistic plots of the top E -values of the BLAST search using AY-Skdb as query and Skdb (x -axis) and mycoprotodb (y -axis) as databases. (b) Negative logistic plots of the top E -values of the BLAST search using Sk_AYdb as query and AYdb (x -axis) and mycoprotodb (y -axis) as databases. Based on criteria described in Section 2, data points in diamonds (◆ or ◇) are proteins shared between AY-WB and *S. kunkelii* but absent from five *Mycoplasma* spp. and *U. urealyticum*, and data points in open triangles (△) are proteins present in AY-WB, *S. kunkelii* and *Mycoplasma* spp. and *U. urealyticum*. Data points in solid diamond (◆) are proteins having similar lengths and annotations, which are detailed in Table 1. Data points in open circles (○) are AY-WB or *S. kunkelii* proteins that are more similar to each other than to counterparts in five *Mycoplasma* spp. and *U. urealyticum*, which are detailed in Table 2.

Table 1

Four AY-WB and *S. kunkelii* homologues that were absent from mycoprotodb consisting of the whole genome sequences of *M. genitalium*, *M. pneumoniae*, *U. urealyticum*, *M. pulmonis*, *M. penetrans*, and *M. gallisepticum*

ID	AY-WB and <i>Spiroplasma kunkelii</i> homologues absent from mycoprotodb			Best hit against NCBI nr database			Cellular location ^c
	Source	ORF ID ^a	Length ^b	Accession #, Homology	Organism	E-value	
1	AY-WB	246_1F	716	29377522, PNPase	<i>Enterococcus faecalis</i>	1e – 180	Cytoplasm
	<i>Spiroplasma kunkelii</i>	100_74F	719	15902560, PNPase	<i>Streptococcus pneumoniae</i>	0	Cytoplasm
2	AY-WB	247_200F	321	27468441, cmp-binding factor 1	<i>Staphylococcus aureus</i>	1e – 39	Cytoplasm
	<i>Spiroplasma kunkelii</i>	109_633F	313	16078057, cmp-binding factor 1	<i>Bacillus subtilis</i>	3e – 59	Cytoplasm
3	AY-WB	247_187F	161	20806575, cytosine/adenosine deaminases	<i>Thermoanaerobacter tengcongensis</i>	4e – 26	Cytoplasm
	<i>Spiroplasma kunkelii</i>	98_127R	159	20806575, cytosine/adenosine deaminases	<i>Thermoanaerobacter tengcongensis</i>	1e – 16	Cytoplasm
4	AY-WB	247_205R	85	541414, conserved hypothetical protein YlxR	<i>Bacillus subtilis</i>	2e – 14	Cytoplasm
	<i>Spiroplasma kunkelii</i>	107_113R	91	541414, conserved hypothetical protein YlxR	<i>Bacillus subtilis</i>	2e – 06	Cytoplasm

^a ORF ID identified by ORF Extractor.

^b Length of deduced amino acid sequence.

^c Cellular location was determined by pSORT [37].

Table 2

Identities of AY-WB and *S. kunkelii* proteins that are more similar to each other than to proteins in mycoprotodb

ID	Proteins shared between AY-WB and <i>Spiroplasma kunkelii</i>			Best hit against NCBI nr database			Cellular location ^c
	Organism	ORF ID ^a	Length ^b	Accession #, Homology	Organism	E-value	
1	AY-WB	235_4R	414	15613820, BH1257 unknown conserved	<i>Bacillus halodurans</i>	1e – 94	Cytoplasm
	<i>Spiroplasma kunkelii</i>	94_78R	414	15613820, BH1257 unknown conserved	<i>Bacillus halodurans</i>	2e – 87	Cytoplasm
2	AY-WB	248_157F	889	15673239, cation-transporting P-ATPase (EC 3.6.3.2)	<i>Lactococcus lactis</i>	2e – 179	Membrane
	<i>Spiroplasma kunkelii</i>	77_20F	910	30022224, Mg ²⁺ transport ATPase, P type (EC 3.6.3.2)	<i>Bacillus cereus</i>	0	Membrane
3	AY-WB	247_48R	745	10443847, ppGpp synthetase	<i>Geobacillus stearothermophilus</i>	e – 154	Cytoplasm
	<i>Spiroplasma kunkelii</i>	106_196R	749	6647842, ppGpp synthetase	<i>Spiroplasma citri</i>	0	Cytoplasm
4	AY-WB	246_186F	528	28378886, Hypothetical exported protein/HAD hydrolase	<i>Lactobacillus plantarum</i>	1e – 116	Membrane or outside
	<i>Spiroplasma kunkelii</i>	96_41R	509	401696, Hypothetical exported protein/HAD hydrolase	<i>Mycoplasma mycoides</i>	1e – 92	Membrane or outside

^a ORF ID identified by ORF Extractor.

^b Length of deduced amino acid sequence.

^c Cellular location was determined by pSORT [37].

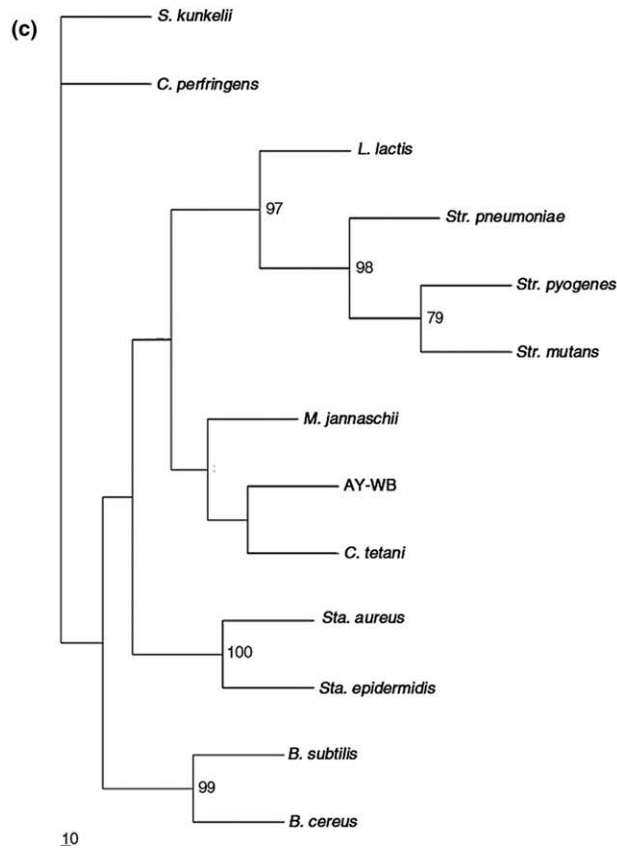
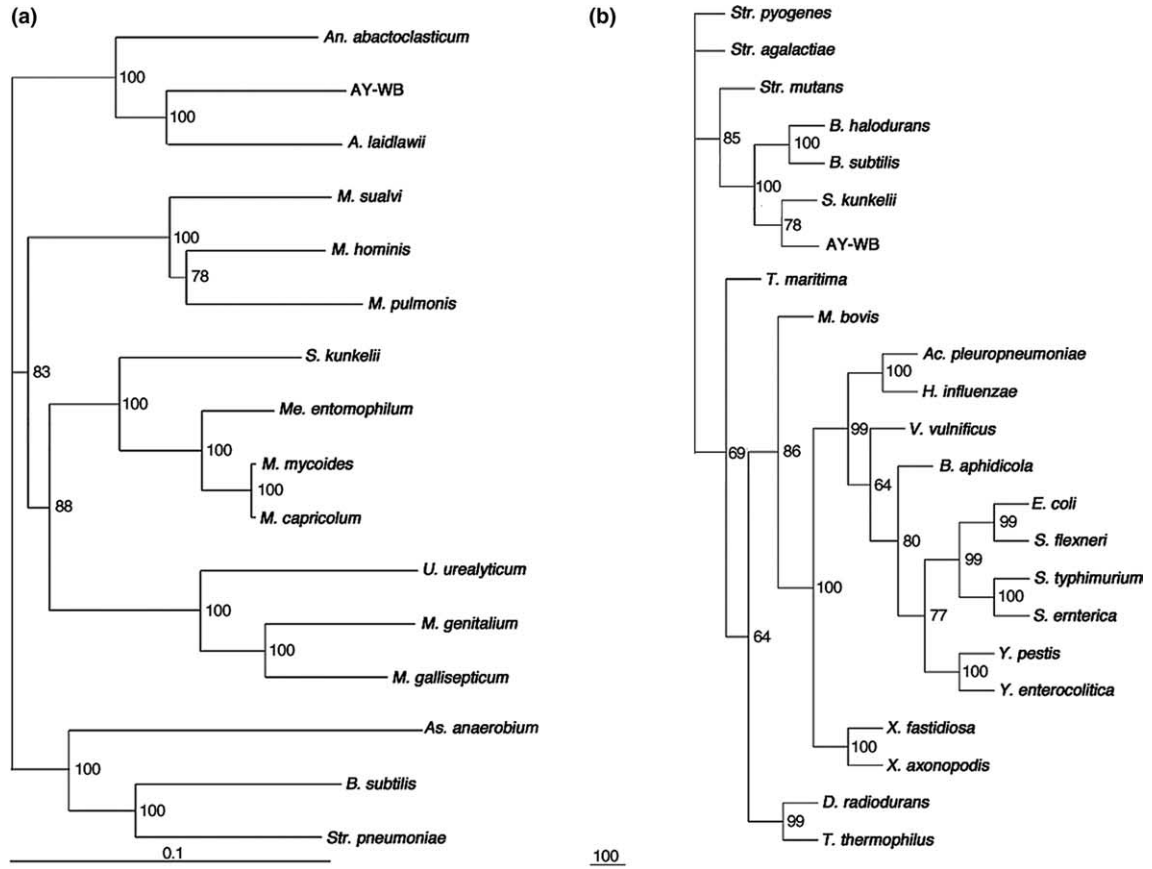


Fig. 3. Phylogenetic analyses of proteins that are present in insect-transmitted plant pathogenic AY-WB and *S. kunkelii* but absent from animal and human pathogenic mycoplasmas. Phylogenetic trees were generated following the procedure described in Materials and methods. Bars under the trees represent evolutionary distances. (a) Phylogenetic tree derived from 16S rDNA sequences. (b) Phylogenetic tree derived from polynucleotide phosphorylase (PNPase). (c) Phylogenetic tree derived from *cmp*-binding factor (CBF). Protein sequences were obtained from GenBank and aligned with ClustalW [14]. The alignments were used for parsimony analysis in PAUP version 4.0 [15]. Trees were bootstrapped 1000 times and the bootstrap values above 50% are indicated as a percentage at the branches. Accession Numbers for protein sequences were as follows: (a) *Acholeplasma laidlawii*, M23932; *Anaeroplasmabactoclasticum*, M25050; *Asteroleplasma anaerobium*, M22351; *Bacillus subtilis*, AB042061; *Mesoplasma entomophilum*, AF305693; *Mycoplasma capricolum*, U26048; *M. gallisepticum*, M22441; *M. genitalium*, X77334; *M. hominis*, AJ002268; *M. mycoides*, U26050; *M. pulmonis*, AF125582; *Mycoplasma svalvi*, AF412988; *Streptococcus pneumoniae*, AY281083; *U. urealyticum*, U06098. (b) *Actinobacillus pleuropneumoniae*, ZP_00134571; *Bacillus halodurans*, NP_243273; *B. subtilis*, NP_389551; *Buchnera aphidicola*, NP_777952; *Deinococcus radiodurans*, NP_295786; *Escherichia coli*, NP_312072; *Haemophilus influenzae*, NP_438401; *Mycobacterium bovis*, CAD94991; *S. enterica*, NP_806878; *S. typhimurium*, AAL22154; *Shigella flexneri*, NP_708965; *Streptococcus agalactiae*, CAD45842; *Streptococcus mutans*, NP_720625; *Streptococcus pyogenes*, BAC64773; *Thermotoga maritime*, NP_229146; *Thermus thermophilus*, CAB06341; *Vibrio vulnificus*, NP_935490; *Xylella fastidiosa*, NP_778440; *Xanthomonas axonopodis*, NP_642994; *Yersinia enterocolitica*, CAA71697; *Yersinia pestis*, NP_668031. (c) *B. subtilis*, CAB12833; *Bacillus cereus*, NP_830807; *Clostridium perfringens*, NP_560939; *Clostridium tetani*, NP_783025; *Lactococcus lactis*, NP_268079; *Methanococcus jannaschii*, NP_247831; *S. aureus*, NP_374949; *Sta. Epidermidis*, NP_765078; *Str. mutans*, NP_720807; *Str. pneumoniae*, NP_359386; *Str. pyogenes*, NP_268621.

Sk_AYdb were searched against mycoprotodb using the blastp algorithm. Sequences that had non-significant similarity (E -value $> 10^{-8}$) or no similarities were extracted from Skdb and AYdb. This resulted in two datasets of AY_Sk_-mycoprotodb with 14 entries and Sk_AY_-mycoprotodb with seven entries.

Plotting the negative logs of the blastp E -values showed that the majority of the predicted protein sequences shared by AY-WB and *S. kunkelii* had homologs in the five *Mycoplasma* spp. and *U. urealyticum* (Fig. 2). However, 9 AY-WB and 8 *S. kunkelii* proteins did not have significant similarity to proteins in mycoprotodb. Among these, four proteins present in both the AY_Sk_-mycoprotodb and Sk_AY_-mycoprotodb datasets were analyzed because they were similar in length and had significant similarity to proteins in NCBI nr database (closed diamonds, Fig. 2). The four proteins were identified as polynucleotide phosphorylase (PNPase), *cmp*-binding factor (CBF), cytosine deaminase, and YlxR protein (Table 1). The PNPase protein sequences of AY-WB and *S. kunkelii* were 62% (452/719) similar, the CBFs 59% (138/231), cytosine deaminases 60% (86/141), and YlxR proteins 61% (46/74). To ensure that the sequences are not present in the genomes of mycoplasmas and ureaplasmas, sequences in common between AY-WB and *S. kunkelii* were searched against the mycoplasma and ureaplasma GenBank databases. Further, the annotated protein databases of *Mycoplasma* spp. and *U. urealyticum* were searched by keywords. Both analyses showed that no proteins for these organisms were annotated as PNPase, CBF, cytosine deaminase, or YlxR protein. Thus, these data suggested that these four genes are present in AY-WB and *S. kunkelii* but absent from *Mycoplasma* spp. and *U. urealyticum* genomes. All these four proteins have homologs in OY phytoplasma genome. However, all but PNPase have homologs in *M. mycoides* subsp. *mycoides* SC strain.

3.3. Identification of proteins more closely related between AY-WB and *S. kunkelii* than to other mollicutes

Four proteins were identified from the negative logistic plots that were more similar between AY-WB and *S. kunkelii* than to mycoplasmas (open circles, Fig. 2). These proteins were identified as ppGpp synthetase, HAD hydrolase, AtA (AAA type ATPase), and P-type Mg^{2+} transport ATPase (Table 2). Amino acid sequence similarities between AY-WB proteins and *S. kunkelii* proteins were ppGpp synthetase, 59% (305/503); HAD hydrolase, 59% (449/750); AtA, 88% (362/407); and P-type Mg^{2+} transport ATPase, 56% (512/902). All proteins have homologs in the genomes of OY phytoplasma and *M. mycoides* subsp. *mycoides* SC strain, except for the AtA sequence that is lacking from OY phytoplasma.

3.4. Phylogenetic analysis of proteins present in AY-WB and *S. kunkelii* but absent from mycoplasmas

Phylogenetic analyses were performed to investigate the origin of the proteins identified in this study. The PNPases from AY-WB and *S. kunkelii* clustered with those from the Gram-positive *Bacillus* and *Streptococcus* spp. and were clearly distinct from those of Gram-negative bacteria (Fig. 3(b)). Thus, the PNPase phylogenetic trees are consistent with the proposed evolutionary status of mollicutes as descendants of Gram-positive bacterial ancestors [1,19]. Phylogenetic analysis of CBFs (Fig. 3(c)) resulted in a tree different from the phylogenetic tree based on 16S rDNA sequences (Fig. 3(a)) with the CBF sequences of AY-WB and *S. kunkelii* separated by CBF sequences of Gram-positive bacteria. Phylogenetic analyses of cytosine deaminases and YlxR proteins resulted in trees with most branches having low bootstrap values (data not shown).

3.5. Phylogenetic analysis of proteins more closely related between AY-WB and *S. kunkelii* to other mollicutes

Phylogenetic analysis was employed to analyze the possible origins of the four proteins that were more closely related between AY-WB and *S. kunkelii* than to other mollicutes. Most branches of the phylogenetic trees generated using ppGpp synthetase, HAD hydrolase, and P-type Mg²⁺ transport ATPase had bootstrap values lower than 50% (data not shown). However, bootstrap values of the phylogenetic tree based on AtA sequences were statistically significant. Interestingly, in the AtA phylogeny, the phytoplasma AtA sequence clustered together with the AtA sequence of *S. kunkelii* in a cluster of AtA sequences of mycoplasmas belonging

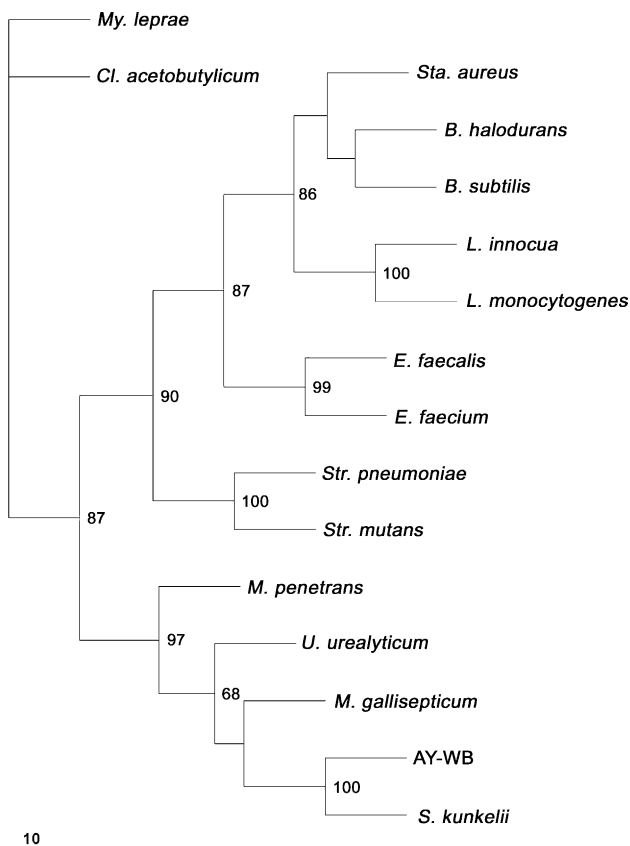


Fig. 4. Phylogenetic analysis for AtA (AAA type ATPase). Phylogenetic trees were generated following the procedure described in Material and Methods. Bars under the trees stand for evolutionary distances. Protein sequences were obtained from GenBank and aligned with ClustalW [14]. The alignments were used for parsimony analysis in PAUP version 4.0. Trees were bootstrapped 1000 times and the bootstrap values above 50% are indicated as a percentage at the branches. Accession Numbers for protein sequences follows. *B. halodurans*, NP_242123; *B. subtilis*, NP_390631; *Clostridium acetobutylicum*, NP_348297; *Enterococcus faecalis*, NP_815655; *Enterococcus faecium*, ZP_00036045; *Listeria innocua*, NP_470885; *Listeria monocytogenes*, NP_465039; *Mycobacterium leprae*, CAA19102; *M. gallisepticum*, NP_853308; *M. penetrans*, NP_757529; *S. aureus*, NP_646394; *Str. mutans*, NP_722348; *Str. pneumoniae*, NP_346223; *Ureaplasma urealyticum*, NP_078028.

to the SEM branch (Fig. 4). Thus the AtA phylogeny is different from the 16S rDNA phylogeny (Fig. 3(a)). The AtA homolog is present in *M. mycoides* subsp. *mycoides* SC, which is also a member of the SEM branch of mollicutes, but it is absent from the OY phytoplasma genome.

4. Discussion

In this study, we have identified several proteins that appear to be present in AY-WB and *S. kunkelii* but absent from *Mycoplasma* spp. and *U. urealyticum*. These proteins are PNPase, CBF, cytosine deaminase, and YlxR. These proteins are also present in the genome of OY phytoplasma, another insect-transmitted plant pathogenic mollicute closely related to AY-WB.

PNPase is an exoribonuclease belonging to the PDX family that also includes RNase PH [20]. Most prokaryotes have PNPase homologs, however, thus far, none have been sequenced from mycoplasmas and *U. urealyticum*. PNPase genes are also present in the genomes of plants [21] and *Drosophila* [22]. PNPases are highly conserved proteins that are involved in mRNA degradation and regulation of gene expression [23]. PNPase has been shown to be a global regulator of virulence factors of *Salmonella enterica*, because a single point mutation of the PNPase gene resulted in a significant decrease in efficiency of invasion and intracellular replication of this bacterium [24]. Both AY-WB and *S. kunkelii* invade and replicate cells of insects and plants [25] and, consequently, have to adjust their gene expression patterns continuously to different environments. In contrast, the *Mycoplasma* and *Ureaplasma* spp. are restricted to animal hosts in which they are able to attach to and mostly invade epithelial cell layers [2]. Thus, PNPases in plant pathogenic bacteria, AY-WB, *S. kunkelii*, and OY phytoplasma, could be important for gene expression regulation allowing adaptation to multiple environmental niches, including insect gut lumen, insect cells, plant phloem, and plant cells. However, the involvement of PNPase in regulation of virulence of plant pathogenic mollicutes, AY-WB and *S. kunkelii*, remains to be investigated. At this time, spiroplasmas are more suitable candidates for such an investigation, because, unlike phytoplasmas, they can be cultured [26] and transformed [27].

CBF is a protein identified in *Staphylococcus aureus*. It binds to the *cmp* sequence, a replication enhancer identified in the pT181 plasmid of *S. aureus*, to stimulate plasmid replication [28]. Spiroplasmas and phytoplasmas have plasmids [2], whereas plasmids have not been reported in members of *M. pneumoniae* and *M. hominis* groups and *U. urealyticum* that do not have CBF. Interestingly, a CBF homolog is present within the re-

cently released complete genome of *M. mycoides* subsp. *mycoides* SC strain [3]. Although plasmids have not been reported in the SC type strain, plasmids are common in *M. mycoides* spp. *mycoides* [29,30]. It is possible that CBF is required for regulation of plasmid replication in spiroplasmas and phytoplasmas. Interestingly, spiroplasma and phytoplasma plasmid appear to harbor virulence factors [31,32].

Cytosine deaminase is an enzyme involved in nucleotide metabolism and can affect protein synthesis if transiently expressed in human cells [33]. Thus, apparently, *S. kunkelii*, AY-WB, and OY have an additional housekeeping gene that is absent from other mollicutes sequenced so far. YlxR protein is expressed from the nusA/infB operon in bacteria and proposed to be an RNA-binding protein [34].

We also identified four AY-WB and *S. kunkelii* ORFs that appear to be more closely related to each other than their mycoplasma counterparts. Of these, the AtA sequence is most interesting, because the phylogenetic tree suggests that phytoplasmas might have obtained the AtA sequence from spiroplasmas, possibly *S. kunkelii*, by horizontal gene transfer. This hypothesis is supported by additional data. First, AtA is absent from the OY phytoplasma genome [4]. OY phytoplasmas is a plant pathogen in Japan where there is no occurrence of *S. kunkelii*. But, in the American continent, *S. kunkelii* and AY-WB co-occur and occasionally share similar insect and plant host ranges [35]. Secondly, AtA sequences of both AY-WB and *S. kunkelii* are flanked by insertion sequences that often part of mobile elements [36]. AY-WB AtA is flanked by a truncated transposase gene at its 5' end and an intact transposase gene at its 3' end, and *S. kunkelii* AtA is located in an IS (insertion sequence) element-rich region.

In summary, the comparative genomics study presented herein successfully identified proteins that are common among insect-transmitted plant pathogenic mollicutes. Further studies of these proteins may elucidate their roles in insect transmission and plant pathogenicity.

This research was supported by OSU-OARDC Research Enhancement Competitive Grants Program and MCIC.

Acknowledgements

The authors thank Dr. Sophien Kamoun in the Department of Plant Pathology, OSU-OARDC, for constructive advice; Dr. Tea Meulia for setup of the Linux workstation and design of ORF Extractor; and B.A. Roe, S.P. Lin, H.G. Jia, H.M. Wu, D. Kupfer, and R.E. Davis and the *S. kunkelii* Genome Sequencing Project funded by US Department of Agriculture, Agricultural Research Service Project Number: 1275-22000-144-02 for the *S. kunkelii* genome sequences.

References

- [1] Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* 51 (2), 221–271.
- [2] Razin, S., Yogeve, D. and Naot, Y. (1998) Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.* 62 (4), 1094–1156.
- [3] Westberg, J., Persson, A., Holmberg, A., Goesmann, A., Lundeberg, J., Johansson, K.E., Pettersson, B. and Uhlen, M. (2004) The genome sequence of *Mycoplasma mycoides* subsp. *mycoides* SC type strain PG1T, the causative agent of contagious bovine pleuropneumonia (CBPP). *Genome Res.* 14 (2), 221–227.
- [4] Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.-Y., Wei, W., Suzuki, S., Arashida, R., Nakata, D., Miyata, S., Ugaki, M. and Namba, S. (2004) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat. Genet.* 36 (1), 27–29.
- [5] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G.G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J.L., Nguyen, D.T., Utterback, T., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O. and Venter, J.C. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403.
- [6] Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C. and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24 (22), 4420–4449.
- [7] Glass, J.I., Lefkowitz, E.J., Glass, J.S., Heiner, C.R., Chen, E.Y. and Cassell, G.H. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* 407, 757–762.
- [8] Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., Moszer, I., Dybvig, K., Wroblewski, H., Viari, A., Rocha, E.P.C. and Blanchard, A. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* 29 (10), 2145–2153.
- [9] Sasaki, Y., Ishikawa, J., Yamashita, A., Oshima, K., Kenri, T., Furuya, K., Yoshino, C., Horino, A., Shiba, T., Sasaki, T. and Hattori, M. (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. *Nucleic Acids Res.* 30 (23), 5293–5300.
- [10] Papazisi, L., Gorton, T.S., Kutish, G., Markham, P.F., Browning, G.F., Nguyen, D.K., Swartzell, S., Madan, A., Mahairas, G. and Geary, S.J. (2003) The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain R(low). *Microbiology (Reading, Engl.)* 149, 2307–2316.
- [11] Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L. (2003) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 31 (1), 28–33.
- [12] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acids Res.* 31 (1), 23–27.
- [13] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- [14] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.
- [15] Swofford, D. (2001) PAUP* 4.0. Sinauer Associates.
- [16] Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* 32, 339–377.

- [17] Dandekar, T., Huynen, M., Regula, J.T., Ueberle, B., Zimmermann, C.U., Andrade, M.A., Doerks, T., Sanchez-Pulido, L., Snel, B., Suyama, M., Yuan, Y.P., Herrmann, R. and Bork, P. (2000) Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res.* 28 (17), 3278–3288.
- [18] Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* 6, 119–129.
- [19] Weisburg, W.G., Tully, J.G., Rose, D.L., Petzel, J.P., Oyaizu, H., Yang, D., Mandelco, L., Sechrest, J., Lawrence, T.G., Van Etten, J., Maniloff, J. and Woese, C.R. (1989) A phylogenetic analysis of the mycoplasmas: basis for their classification. *J. Bacteriol.* 171, 6455–6467.
- [20] Zuo, Y. and Deutscher, M.P. (2001) Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. *Nucleic Acids Res.* 29 (5), 1017–1026.
- [21] Li, Q.S., Gupta, J.D. and Hunt, A.G. (1998) Polynucleotide phosphorylase is a component of a novel plant poly(A) polymerase. *J. Biol. Chem.* 273 (28), 17539–17543.
- [22] Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferreira, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Siden-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M. and Venter, J.C. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.
- [23] Carpousis, A.J. (2002) The *Escherichia coli* RNA degradosome: structure, function and relationship in other ribonucleolytic multienzyme complexes. *Biochem. Soc. Trans.* 30 (2), 150–155.
- [24] Clements, M.O., Eriksson, S., Thompson, A., Lucchini, S., Hinton, J.C., Normark, S. and Rhen, M. (2002) Polynucleotide phosphorylase is a global regulator of virulence and persistency in *Salmonella enterica*. *Proc. Natl. Acad. Sci. USA* 99 (13), 8784–8789.
- [25] Ozbek, E., Miller, S.A., Meulia, T. and Hogenhout, S.A. (2003) Infection and replication sites of *Spiroplasma kunkelii* (Class: Mollicutes) in midgut and Malpighian tubules of the leafhopper *Dalbulus maidis*. *J. Invertebr. Pathol.* 82 (3), 167–175.
- [26] Sgaglio, P., L'hospital, M., Lafleche, D., Dupont, G., Bove, J.M. and Freundt, E.A. (1973) *Spiroplasma citri* gen. and sp. n.: a mycoplasma-like organism associated with 'stubborn' disease of citrus. *Int. J. Syst. Bacteriol.* 23, 191–204.
- [27] Lartigue, C., Duret, S., Garnier, M. and Renaudin, J. (2002) New plasmid vectors for specific gene targeting in *Spiroplasma citri*. *Plasmid* 48, 149–159.
- [28] Zhang, Q., Soares de Oliveira, S., Colangeli, R. and Gennaro, M.L. (1997) Binding of a novel host factor to the pT181 replication enhancer. *J. Bacteriol.* 179 (3), 684–688.
- [29] King, K.W. and Dybvig, K. (1994) Mycoplasma cloning vectors derived from plasmid pKMK1. *Plasmid* 31 (1), 49–59.
- [30] Djordjevic, S.R., Forbes, W.A., Forbes-Faulkner, J., Kuhnert, P., Hum, S., Hornitzky, M.A., Vilei, E.M. and Frey, J. (2001) Genetic diversity among Mycoplasma species bovine group 7: clonal isolates from an outbreak of polyarthritis, mastitis, and abortion in dairy cattle. *Electrophoresis* 22 (16), 3551–3561.
- [31] Melcher, U., Sha, Y., Ye, F. and Fletcher, J. (1999) Mechanisms of spiroplasma genome variation associated with SpV1-like viral DNA inferred from sequence comparisons. *Microb. Comp. Genomics* 4 (1), 29–46.
- [32] Oshima, K., Miyata, S., Sawayanagi, T., Kakizawa, S., Nishigawa, H., Jung, H.-Y., Furuki, K., Yanazaki, M., Suzuki, S., Wei, W., Kuboyama, T., Ugaki, M. and Namba, S. (2002) Minimal set of metabolic pathways suggested from the genome of Onion Yellowings phytoplasma. *J. Gen. Plant Pathol.* 68 (3), 225–236.
- [33] Kreuzer, J., Denger, S., Reifers, F., Beisel, C., Haack, K., Gebert, J. and Kubler, W. (1996) Adenovirus-assisted lipofection: efficient in vitro gene transfer of luciferase and cytosine deaminase to human smooth muscle cells. *Atherosclerosis* 124, 49–60.
- [34] Osipiuk, J., Gornicki, P., Maj, L., Dementieva, I., Laskowski, R. and Joachimiak, A. (2001) *Streptococcus pneumoniae* YlxR at 1.35 Å shows a putative new fold. *Acta Crystallogr. D Biol. Crystallogr.* 57, 1747–1751.
- [35] Nault, L.R. (1990) Evolution of an insect pest: maize and the corn leaf hopper, a case study. *Maydica* 35, 165–175.
- [36] Mahillon, J. and Chandler, M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.* 62 (3), 725–774.
- [37] Nakai, K. and Horton, P. (1999) pSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24 (1), 34–36.