Sustainability in Plant and Crop Protection 12

Chrystel Y. Olivier
Tim J. Dumonceaux
Edel Pérez-López *Editors*

Sustainable Management of Phytoplasma Diseases in Crops Grown in the Tropical Belt

Biology and Detection

Springer

Sustainability in Plant and Crop Protection

Volume 12

Series Editor
Aurelio Ciancio, Sezione di Bari, Consiglio Nazionale delle Ricerche Istituto per la Protezione delle Piante, Bari, Italy

The series describes new perspectives recently developed in pest and disease management, based on innovative tools emerging from basic and applied research. The volumes will aim at interested readers involved in plant protection and crop management, for whom soil biodiversity, crop sustainability and, in general, organic approaches are fundamental. Different cropping systems will be treated by researchers involved in cutting edge studies worldwide. A number of basic issues including sustainability, life-cycle assessment, evolution, plant nutrition and organic cropping technologies will provide a common framework, within which different components of the crop production cycle will be focused on. These will range from roots and endophytes to pest and disease control, through the management of soil microbiome and fertility. These issues will be examined at the field and crop levels, including the effects of invasive species and climate changes on agroecosystems. Recent advancements in massive sequencing will represent the basis of dedicated volumes, dealing with transcriptomics and related approaches. They will illustrate the potentials and benefits of extensive DNA and RNA data analyses and studies, for practical purposes in crop protection, disease management and food production.

Contributions on any of the above cited topics are welcome. Potential Editors proposing a new volume are requested to contact the Series Responsible Editor and provide a short CV (400 words max) listing a selection of their most significant publications. In order to broaden the base of contributors and avoid redundancies, only one volume per Editor is allowed. Exceptionally, in case of many contributed chapters, a two-issues volume can eventually be considered.

More information about this series at http://www.springer.com/series/13031

Chrystel Y. Olivier • Tim J. Dumonceaux
Edel Pérez-López
Editors

Sustainable Management of Phytoplasma Diseases in Crops Grown in the Tropical Belt

Biology and Detection

Springer

Editors
Chrystel Y. Olivier
Agriculture and AgriFood Canada
Saskatoon, SK, Canada

Edel Pérez-López
Department of Biology
University of Saskatchewan
Saskatoon, SK, Canada

Tim J. Dumonceaux
Department of Veterinary Microbiology
University of Saskatchewan
Saskatoon, SK, Canada

Agriculture and Agri-Food Canada,
Saskatoon Research
and Development Centre
Saskatoon, SK, Canada

ISSN 2567-9805 ISSN 2567-9821 (electronic)
Sustainability in Plant and Crop Protection
ISBN 978-3-030-29649-0 ISBN 978-3-030-29650-6 (eBook)
https://doi.org/10.1007/978-3-030-29650-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Phytoplasmas are devastating plant pathogens that are capable of infecting a wide array of plant species, including most crop plants. These are vectored by phloem-feeding insects, mostly leafhoppers, plant hoppers, and psyllids, which transmit the disease during feeding. Phytoplasma-infected plants exhibit symptoms of malformed plants and seeds, sometimes resulting in severe declines in the yield of field crops, fruits, and vegetables worldwide. Economic losses associated with phytoplasma infections are high every year due to their impacts on economically and ecologically important crops. In tropical countries, which commonly features agriculture among the top three income sources, the impact has been especially devastating. The lack of an effective cure means that appropriate management is required to reduce the impact of diseases caused by phytoplasmas. This includes rapid, rigorous, and sensitive methods to detect phytoplasma infections in plants and insects and the ability to accurately differentiate and classify phytoplasmas to facilitate epidemiological investigations of disease incidence and spread. Several books covering different aspects of this important plant pathogen have been published, but to date, this is the first book focusing on the management of phytoplasma diseases affecting plant hosts in selected tropical countries. These countries can be strongly impacted by phytoplasmas due to the importance of agriculture in their economies as well as the continuous growth cycles of insect vectors.

In 12 chapters contributed by experienced scientists worldwide, we have explored the management strategies employed by farmers in Mexico, Cuba, Argentina, Brazil, Colombia, China, and other tropical countries. We are sincerely grateful to all of the contributing authors and to Springer Nature, all of whom were key to the realization of this book. The editors also express special appreciation to Jannet Tam and Devvyn Murphy for their assistance in proofreading this book. We are confident

that this book will be useful for plant pathologists, agronomists, entomologists, extension specialists, and farmers in general interested in the management of phytoplasma diseases.

Saskatoon, Saskatchewan, Canada

Chrystel Y. Olivier
Tim J. Dumonceaux
Edel Pérez-López

Contents

About the Editors

Chrystel Y. Olivier is a Research Scientist at Agriculture and Agri-Food Canada Saskatoon Research and Development Centre (AAFC) since 2001. Her research interests include phytoplasma disease epidemiology and management in field crops. Her lab worked extensively on aster yellow disease epidemiology and control in canola and cereal crops and on leafhopper populations associated with the disease.

Tim J. Dumonceaux has been a Research Scientist at the Agriculture and Agri-Food Canada Saskatoon Research and Development Centre (AAFC) since 2009. His research interests include the molecular diagnostics of plant and animal pathogens, the impacts of microbial communities on a variety of agricultural ecosystems, and the applications of white-rot fungi in the production of biofuels. His lab has developed a suite of tools for identifying and characterizing phytoplasmas based on the universal microbial barcode chaperonin-60 (cpn60), and he maintains an interest in applying these tools to the detection, characterization, and quantification of phytoplasma infections in plant and insect tissues.

Edel Pérez-López is a Postdoctoral Fellow at the University of Saskatchewan, Biology Department, working with clubroot and the soil-borne obligate parasite *Plasmodiophora brassicae*, although he keeps actively collaborating in research projects related to the identification and characterization of phytoplasmas in South America, Saudi Arabia, and Canada. He has been working with phytoplasmas affecting crops in Cuba, Mexico, Peru, Canada, and Saudi Arabia since 2012 and on the development of diagnostic methods to identify and characterize this group of plant pathogenic bacteria.

Chapter 1
The CpnClassiPhyR Facilitates Phytoplasma Classification and Taxonomy Using cpn60 Universal Target Sequences

Kevin Muirhead, Edel Pérez-López, Brian W. Bahder, Janet E. Hill, and Tim J. Dumonceaux

Abstract Phytoplasmas (‘*Candidatus* Phytoplasma’ spp.) are plant pathogenic bacteria that are transmitted by insects and cause developmental changes leading to altered floral morphologies and decreased seed set in a very wide variety of plant species, including most cultivated plants. Using rRNA and protein-encoding genes, 36 ribosomal groups and a similar number of species have been defined within the ‘*Candidatus* Phytoplasma’ genus. The identification of phytoplasma strains infecting plants and insects has been facilitated by the availability of PCR primers that amplify a fragment of the 16S rRNA-encoding gene from any phytoplasma, the sequence of which is subjected to restriction fragment length polymorphism (RFLP) analysis and compared to RFLP patterns from reference sequences. An analogous classification scheme targeting a protein-encoding gene, chaperonin-60 (*cpn*60), has been described. In this work, we present software that automates the determination

K. Muirhead
Department of Microbiology, Immunology and Infectious Diseases, University of Calgary, Calgary, AB, Canada
e-mail: kevin.muirhead@ucalgary.ca

E. Pérez-López
Department of Biology, University of Saskatchewan, Saskatoon, SK, Canada
e-mail: epl733@mail.usask.ca

B. W. Bahder
University of Florida Fort Lauderdale Research and Education Center, Davie, FL, USA
e-mail: bbahder@ufl.edu

J. E. Hill
Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, SK, Canada
e-mail: janet.hill@usask.ca

T. J. Dumonceaux (✉)
Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, SK, Canada

Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre, Saskatoon, SK, Canada
e-mail: tim.dumonceaux@canada.ca

C. Y. Olivier et al. (eds.), *Sustainable Management of Phytoplasma Diseases in Crops Grown in the Tropical Belt*, Sustainability in Plant and Crop Protection 12,
https://doi.org/10.1007/978-3-030-29650-6_1

of *cpn*60-based groups and subgroups. This software, the CpnClassiPhyR (http://cpnclassiphyr.ca), compares input *cpn*60 sequences to a reference database, performs RFLP-based similarity coefficient calculations, and facilitates phylogenetic analysis of the query sequence. We further describe the application of PCR primers that target phytoplasma *cpn*60 genes to generate new sequences from 16S groups II, IV, XI, XIV, and XV. We also used the CpnClassiPhyR to examine *cpn*60 sequences from reported genome sequences of various '*Ca.* Phytoplasma' species, providing results that are consistent with the 16S-based classification of these strains.

Keywords Classification · Phytoplasma · RFLP · cpn60 · Taxonomy

1.1 Introduction

Phytoplasmas ('*Candidatus* Phytoplasma' spp.), which were previously identified as mycoplasma-like organisms (Zreik, Carle, Bove, & Garnier, 1995), are plant pathogenic bacteria that infect a very wide array of plant species, affecting both cultivated and wild plants in all regions of the world (Bertaccini, Duduk, Paltrinieri, & Contaldo, 2014). Phytoplasmas are classified within the taxonomic class Mollicutes, and they are pleomorphic microorganisms that lack a typical bacterial cell wall (Namba, 2011). These phytopathogens propagate well within their plant and insect vector hosts, but they have not been reproducibly cultivated in axenic cultures. For this reason, traditional taxonomic classification criteria cannot be applied to phytoplasmas and the taxonomy of these microorganisms is under the criteria identified for uncultured microorganisms (Firrao et al., 2004). At least 25 species of '*Ca.* Phytoplasma' have been recognized based primarily on DNA sequence analysis of 16S rRNA-encoding genes and other housekeeping genes (Firrao et al., 2004). Phytoplasma taxonomy is further defined by the 16S rRNA gene through the use of restriction fragment length polymorphism (RFLP) analysis of the 16S rRNA F2nR2 fragment with a set of 17 endonucleases (Lee, Davis, & Gundersen-Rindal, 2000; Lee, Gundersen-Rindal, Davis, & Bartoszyk, 1998; Lee et al., 2004). This approach has identified more than 30 groups of phytoplasmas, designated 16SrI-16SrXXXVI, with each group containing subgroups designated by letters (Miyazaki et al., 2017; Naderali et al., 2017). The application of *in silico* RFLP analysis based on DNA sequence data as an alternative to the *in vitro* RFLP, along with the development of the online phytoplasma classification tool *i*PhyClassifier, has increased the accuracy and accessibility of phytoplasma classification based on 16S rRNA gene sequences (Pérez-López, Luna-Rodríguez, Olivier, & Dumonceaux, 2016; Wei, Lee, Davis, Suo, & Zhao, 2007; Zhao et al., 2009). The proliferation of new groups and subgroups of phytoplasma reported in the literature, however, has led to the recognition of a need for a system to formally register and recognize novel groups and subgroups based on specified criteria (Zhao & Davis, 2016).

While phytoplasma classification and taxonomy based on 16S rRNA-encoding gene sequences has proven to be very useful for studying the distribution and epidemiology of phytoplasma infections, there are limitations associated with their exclusive use. In many cases, rRNA-encoding gene sequences share high DNA

sequence identity among closely related taxa, which limits the resolution of closely related but distinct strains. Furthermore, phytoplasma genomes contain two copies of the 16S rRNA-encoding gene (Wei, Davis, Lee, & Zhao, 2007; Zhao et al., 2009), and in some cases these two copies provide distinct RFLP typing results – a phenomenon referred to as 16S rRNA gene heterogeneity (Pérez-López & Dumonceaux, 2016; Zhao et al., 2009). These factors have led to the exploration of other gene sequences for the identification and taxonomic characterization of phytoplasmas. Single-copy, protein-coding genes including *rp* (Martini et al., 2007), *tuf* (Makarova et al., 2012), *secY* (Lee, Zhao, & Bottner, 2006), *rpoB* (Valiunas, Jomantiene, & Davis, 2013), and others have been explored for this use and these gene sequences typically provide a higher resolution of closely related taxa compared to 16S rRNA-encoding genes. However, accessing these genes from a sample infected with a completely unknown phytoplasma type can prove difficult, with complex PCR primer combinations or group-specific primers often required (Martini et al., 2007; Valiunas et al., 2013).

Another gene that has been exploited in phytoplasma detection, classification, and taxonomy is chaperonin-60 (*cpn*60, with synonyms *groEL* and *hsp60*), which encodes a 60 kDa protein with a canonical role in tertiary structure formation for proteins in nearly all bacteria (Hemmingsen et al., 1988), with the exception of certain Mollicutes (Clark & Tillier, 2010; Schwarz, Adato, Horovitz, & Unger, 2018). Phytoplasmas generally contain a single copy of *cpn*60 in their genomes, except for '*Ca.* Phytoplasma' pruni, group 16SrIII (Saccardo et al., 2012). For other groups of phytoplasmas, *cpn*60 has been shown to be a useful marker that provides improved resolution of closely related taxa compared to ribosomal RNA-encoding sequences (Contaldo, Mejia, Paltrinieri, Calari, & Bertaccini, 2012; Mitrović et al., 2011, 2015). A subsequence of *cpn*60 corresponding to nucleotides 274–828 of the *E. coli cpn*60 gene has been identified as a suitable barcode marker for the Domain Bacteria (Links, Dumonceaux, Hemmingsen, & Hill, 2012) and is accessible with a set of universal primers that amplify the *cpn*60 "universal target" (*cpn*60 UT) from bacteria and eukaryotes (Hill, Town, & Hemmingsen, 2006). These universal primers were modified based on available *cpn*60 sequences from phytoplasmas to develop a set of PCR primers and amplification conditions that facilitate the amplification of the *cpn*60 UT from a diverse range of phytoplasmas (Dumonceaux, Green, Hammond, Pérez-López, & Olivier, 2014).

Following the strategy previously used in the phytoplasma classification scheme based on the 16S rRNA gene, we developed a complementary, coherent system to classify phytoplasmas based on RFLP analysis of *cpn*60 UT sequences with seven endonucleases (Pérez-López, Olivier, Luna-Rodríguez, & Dumonceaux, 2016). This validated classification scheme generally allows a finer differentiation of phytoplasma strains within the same 16S rRNA RFLP subgroup, with the identification of complementary *cpn*60 UT groups and subgroups. Like 16S rRNA-based classification, the *cpn*60 UT-based RFLP analysis relies upon the calculation of a similarity coefficient (F) from the band sizes generated by *in silico* restriction digestion of DNA sequences. To facilitate the broader application of this complementary scheme for phytoplasma classification and taxonomy, we have developed the CpnClassiPhyR, an on-line tool that enables the identification and classification of phytoplasma strains based on the *cpn*60 UT sequence. In this work, we describe this publicly

available resource and its application to the description of groups and subgroups of phytoplasmas for which *cpn*60 sequences have not previously been reported, corresponding to the 16S rRNA groups II, IV, XI, XIV, and XV. We also describe the use of the CpnClassiPhyR to identify novel groups and subgroups of phytoplasmas from reported genome sequences. Finally, we offer to the scientific community a collection of *cpn*60 UT plasmid clones corresponding to the broadest diversity of phytoplasmas that we have been able to obtain.

1.2 Results

1.2.1 Generation of Novel cpn*60 UT Sequences from Previously Unexplored Phytoplasmas*

The phytoplasma *cpn*60 UT primers successfully amplified the target gene from a very wide array of phytoplasma samples, including multiple subgroups within 16S rRNA groups I, II, IV, V, VII, and IX-XV (Fig. 1.1). Many of these groups were

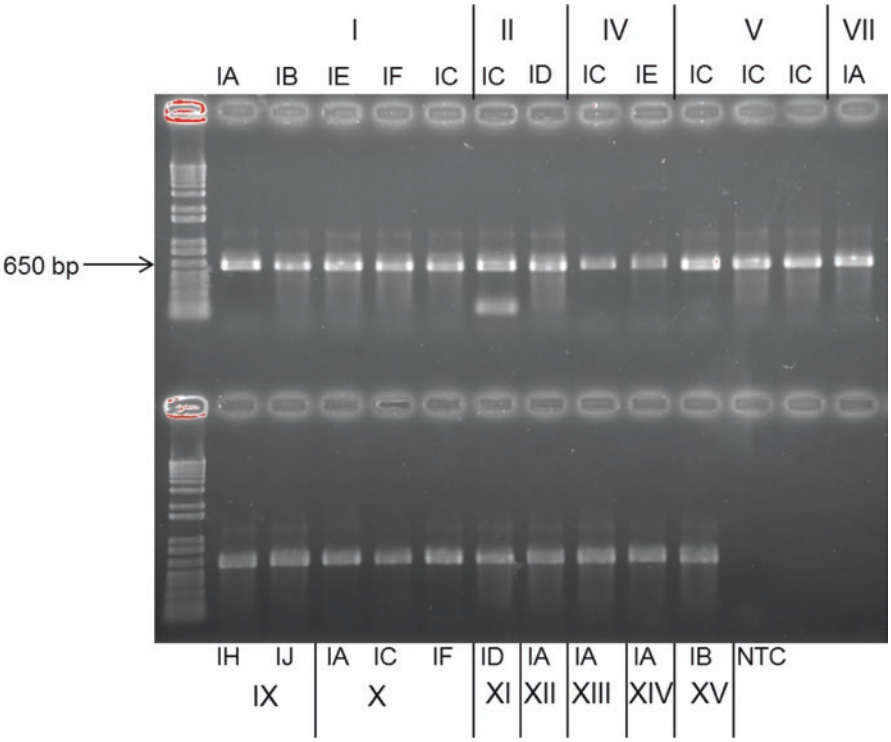

Fig. 1.1 Amplification of phytoplasmas using the *cpn*60 UT primers. Information on the phytoplasma strains used for amplification of each group/subgroup is provided in Table 1.1

previously reported (Dumonceaux et al., 2014), but amplicon from groups IV (‘*Ca.* P.’ palmae, ‘*Ca.* P.’ cocostanzaniae), XI (‘*Ca.* P.’ oryzae) (Zhang et al., 2016), XIV (‘*Ca.* P.’ cynodontis), and XV (‘*Ca.* P.’ brasiliense) had not been previously generated with the phytoplasma *cpn*60 UT PCR primers. Overall, the *cpn*60 genes that were successfully amplified by these primers were highly diverse in sequence, with as low as 61% pairwise sequence identity between strains.

The application of the phytoplasma *cpn*60 UT primers to a wide variety of phytoplasmas collected from around the world has provided us with a collection of cloned *cpn*60 UT fragments from the phytoplasmas identified in Table 1.1 and Fig. 1.2. These are offered to the scientific community for use as plasmid standards, PCR controls, and a means to evaluate novel molecular diagnostics for analytical specificity. The plasmids have been deposited at addgene (www.addgene.org), with accession numbers provided at http://cpnclassiphyr.ca.

1.2.2 The CpnClassiPhyR: A Cpn(60-Based) Classi(Fier for) Phy(Toplasmas Using) R(FLP) Analysis

The CpnClassiPhyR is publicly available at http://cpnclassiphyr.ca. Users are directed to a welcome page containing general information regarding phytoplasmas and the purpose of the software. Along the left-hand side of the welcome page are several tabs, which are described in detail below.

CpnClassiPhyR RFLP analysis of input DNA sequences is performed here. The algorithm used by the software is described in Fig. 1.4. Input sequences are provided by the user and can be in either orientation (protein-coding or reverse-complemented), or can be full-length *cpn*60 genes, and can retain residual primer and cloning vector sequences. The CpnClassiPhyR first checks the length of the sequence and automatically discards those that are <450 bp. Next, the software determines if the sequence is likely derived from a phytoplasma by comparing its sequence using BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) against a non-redundant reference database of *cpn*60 UT nucleotide sequences derived from ‘*Ca.* Phytoplasma’ spp. obtained from cpnDB, the chaperonin database (Hill, Penny, Crowell, Goh, & Hemmingsen, 2004). A cutoff of 70% nucleotide sequence identity has been set for this determination, based on the sequence identity to any phytoplasma sequence returned by the *cpn*60 UT sequence of *A. laidlawii*. Sequences below 70% identity may include non-phytoplasma *cpn*60 sequences, or sequences that do not correspond to *cpn*60 at all. Users are recommended to check public databases to verify the identity of their sequence. BLAST results are returned for phytoplasma *cpn*60 sequences that are less than full-length (552 bp) but these sequences are not subjected to any further steps. Next, retained sequences are trimmed to the universal target and oriented in the protein-coding, reading frame 1 orientation. This trimmed, oriented sequence is made available to the user for download. The trimmed sequence is then subjected to *in silico* RFLP analysis using the previously described scheme (Pérez-López, Olivier, et al., 2016) and a report is provided to the user

Table 1.1 *cpn*60 clone panel. Plasmid DNA containing the cloned *cpn*60 UT sequence from each of the samples described below is available to the phytoplasma research community upon request

clone name	cpn60 UT classification	cpn60 UT GenBank accession	cpnDB ID[1]	16S classification	16S GenBank accession	Phytoplasma species	Host	country of origin	Literature references
O2L[2]	I-IA	KJ939998	b27096	16SrI-A	MH279536	'*Ca. P.* asteris'	Onion	Canada (Saskatchewan)	Dumonceaux et al. (2014)
SF1[3]	I-IB	KJ940013	b27105	16SrI-B	MH279534	'*Ca. P.* asteris'	Flax	Canada (Saskatchewan)	Dumonceaux et al. (2014)
BbSP	I-I(E/AI)AI	KU523402	b27454	16SrI-E/AI[4]	MH279523 (E); MH279522 (AI)	'*Ca. P.* asteris'	Blueberry	Canada (Nova Scotia)	Pérez-López, Olivier, et al. (2016)
CVB	I-IF	KJ939995	b27145	16SrI-F	MH279545	'*Ca.* P. asteris'	Periwinkle	Italy	Dumonceaux et al. (2014)
AY-col	I-IC	KJ939994	b27144	16SrI-R	MH279535	'*Ca.* P. asteris	Periwinkle	Italy	Dumonceaux et al. (2014)
R018	II-IC	MH279495	b31187	16SrII-C	MH266698	'*Ca.* P. aurantifolia'	Lime	Iran	This work
JbWB 190	II-IA	KY704478	b31189	16SrII-D	KY581664	'*Ca.* P. australasiae'	Jojoba	Saudi Arabia	Omar, Pérez-López, Al-Jamhan, and Dumonceaux (2017)
TagTall	IV-IC	MH922781	b32834	16SrIV-C	MH922778	'*Ca.* P. cocostanzaniae'	Coconut	Tanzania	This work
Sab4	IV-IE	MH922782	b32835	16SrIV-D	MH922779	'*Ca.* P. palmae'	Palm	USA (Florida)	This work
Palm A-DR	IV-IE	MH922783	b32836	16SrIV-E	MH922780	'*Ca.* P. palmae'	Coconut	Dominican Republic	This work

FD	V-IC	KJ939992	b27153	16SrV	NA[5]	'*Ca.* P. ulmi'	Grape	France	Dumonceaux et al. (2014)
PGY	V-IC	KJ939991	b27154	16SrV-C	MH279539	'*Ca.* P. ulmi'	Periwinkle	Germany	Dumonceaux et al. (2014)
RS	V-IC	KJ939990	b27155	16SrV-C	MH279544	'*Ca.* P. rubi'	Periwinkle	Germany	Dumonceaux et al. (2014)
AshY	VII-IA	KJ939978	b27156	16SrVII-A	MH279540	'*Ca.* P. fraxini'	Ash	France	Dumonceaux et al. (2014)
Cr	IX-IH	KJ939989	b27158	16SrIX-A	NA[6]	'*Ca.* P.' phoenicium'	Periwinkle	Cuba	Dumonceaux et al. (2014)
ChicBS	IX-IJ	KY986918	b31188	16SrIX-J	KY986922	'*Ca.* P. phoenicium'	Chicory	Saudi Arabia	Pérez-López, Omar, Al-Jamhan, and Dumonceaux (2018)
AP	X-IA	KJ939977	b27159	16SrX-A	MH279542	'*Ca.* P. mali'	Apple	Germany	Dumonceaux et al. (2014)
PYLR	X-IC	KJ940010	b27165	16SrX-C	MH279541	'*Ca.* P. pyri'	Pear	USA	Dumonceaux et al. (2014)
ESFY	X-IF	KJ940007	b27166	16SrX-F	MH279543	'*Ca.* P. prunorum'	Periwinkle	Italy	Dumonceaux et al. (2014)
SCWL	XI-ID	MH279493	b31184	16SrXI-D	KC295286	'*Ca.* P. oryzae'	Sugarcane	China	This work
BN44948	XII-IA	KJ939979	b27168	16SrXII-A	MH279537	'*Ca.* P. solani'	Periwinkle	Germany	Dumonceaux et al. (2014)

(continued)

Table 1.1 (continued)

clone name	cpn60 UT classification	cpn60 UT GenBank accession	cpnDB ID[1]	16S classification	16S GenBank accession	Phytoplasma species	Host	country of origin	Literature references
S31b-GP-MA	XIII-(A/I)I	KT444666	b27178	16SrXIII-A/I[7]	KT444665 (A); KT444664 (I)	'*Ca. P.* hispanicum'	Strawberry	Mexico (Michoacan)	Pérez-López and Dumonceaux (2016)
BGWL	XIV-IA	MH279492	b31185	16SrXIV-A	MH279538	'*Ca.* P. cynodontis'	Berumuda grass	Saudi Arabia	This work
PeruPBT	XV-IB	MH279494	b31186	16SrXV-B	KX810334	'*Ca.* P.' brasiliense	Papaya	Peru	This work and Wei et al. (2017)

In all cases except O2L and SF1, amplicon was generated using the phytoplasma *cpn*60 UT primers described on the "primer sets" tab of the CpnClassiPhyR

[1]chaperonin database ID (http://cpndb.ca)

[2]plasmid contains the 826 bp product of groELF/H280p primers (Dumonceaux et al., 2014). Contains *cpn*60 UT and flanking *cpn*60 sequences

[3]plasmid contains the 1.4 kb product of groEL F/R primers (Mitrović et al., 2011). Contains *cpn*60 UT and flanking *cpn*60 sequences

[4]Sample had F2nR2 sequence heterogeneity (Perez-Lopez et al., manuscript submitted). The sequence MH279522 was suggested as a new subgroup, 16SrI-AI

[5]insufficient template volume remains to generate F2nR2 sequence; 16Sr group information provided by S. Malembic-Maher (personal communication)

[6]NA, not available. The 16S sequence generated from this sample was too short for iPhyclassifier (714 bp), but showed 100% id to GenBank KF941131, which is 16SrIX-A

[7]sample had F2nR2 sequence heterogeneity (Pérez-López & Dumonceaux, 2016)

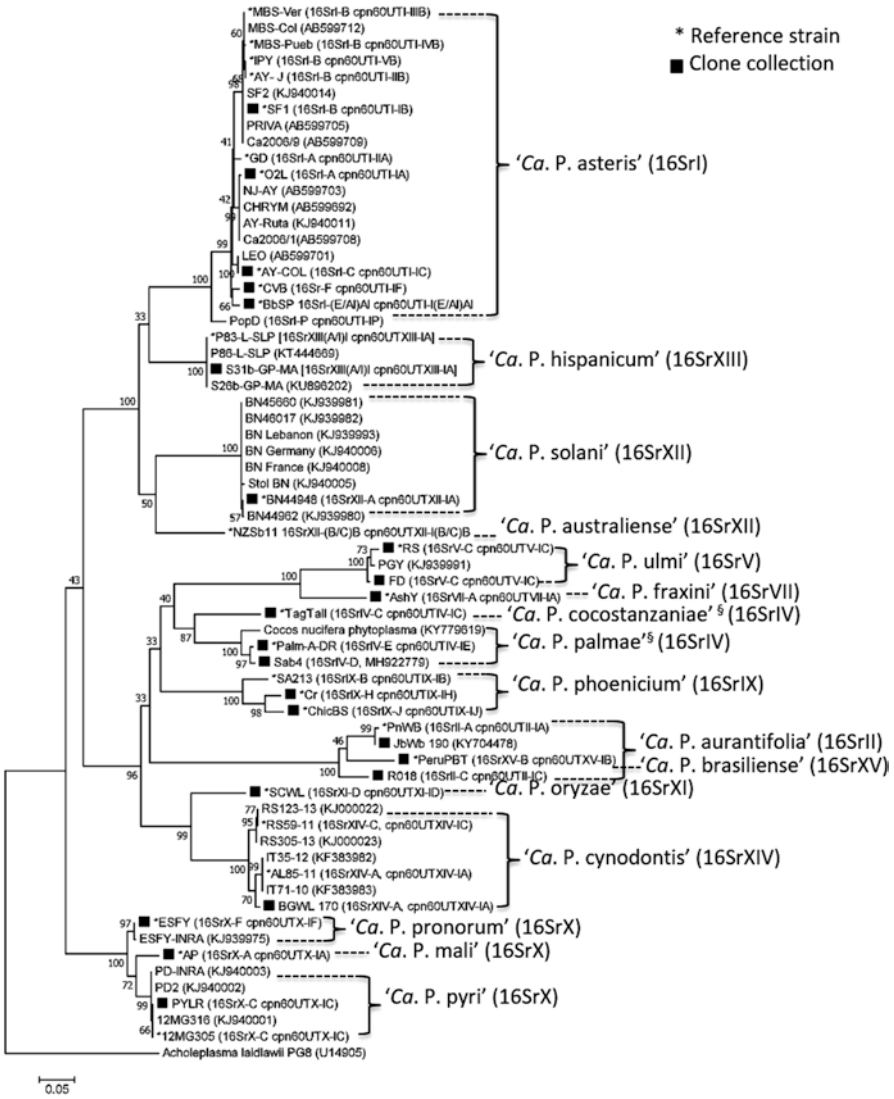

Fig. 1.2 The known diversity of phytoplasmas based on *cpn*60 UT sequences. This phylogenetic tree was constructed using the maximum likelihood method and incorporates *cpn*60 UT sequences generated by our group since 2014 and sequences retrieved from Genbank. Strains for which cloned *cpn*60 gene fragments are available to the scientific community (Table 1.1) are indicated (■). Reference strains for each group/subgroup (Table S1) are indicated (∗). The phylogenetic tree was bootstrapped 1000 times. Bar, 5 substitutions in 100 positions

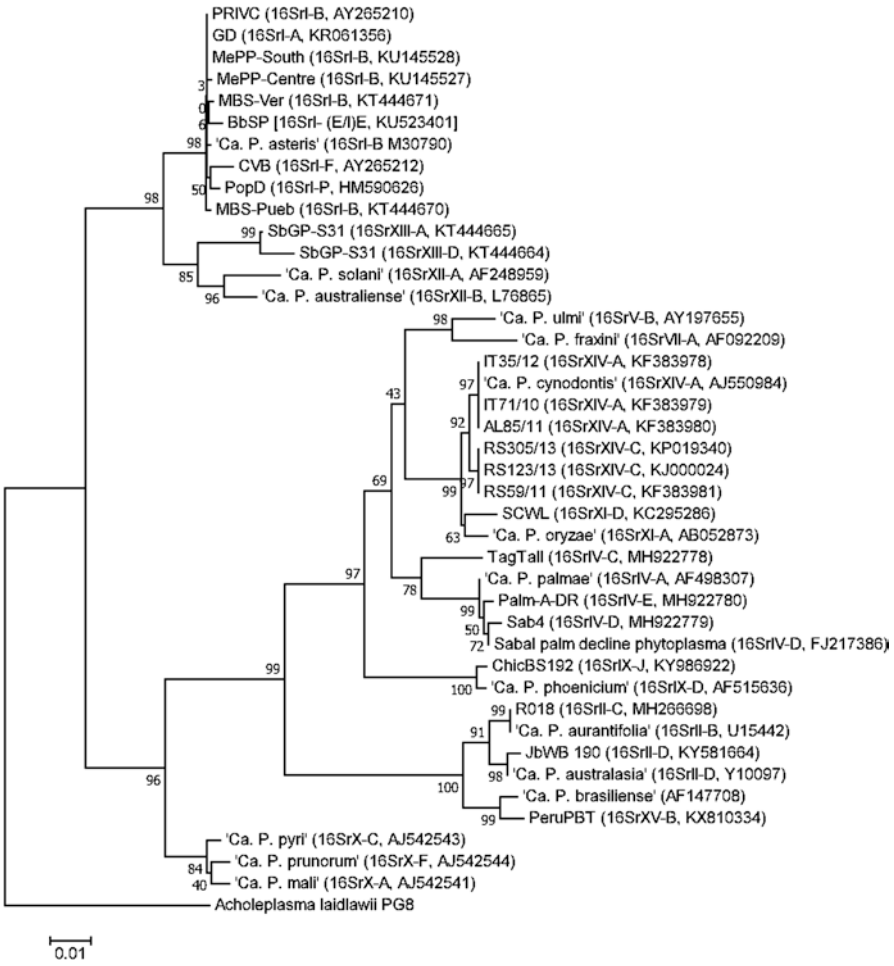

Fig. 1.3 Phylogenetic tree reconstructed through the maximum likelihood method of F2nR2 sequences from the phytoplasma strains examined in this study. The phylogenetic tree was bootstrapped 1000 times. Bar, 1 substitution in 100 positions

depending on the F-value determined for the nearest match. This is achieved by comparison to a reference database of *cpn*60 UT sequences from phytoplasmas. This table is dynamic and will be updated as new validated *cpn*60 UT sequences accrue. Samples that match a previously described group and subgroup ($F \geq 0.97$) are reported to the user. New subgroups are identified by samples with $F \geq 0.59$ and $F < 0.97$ compared to their closest match; similarly, new groups are identified by samples showing $F < 0.59$ to any previously reported group. When these criteria for group/subgroup inclusion are not met, users are provided the opportunity to register their sequence as a new group or subgroup (described below), and new, validated

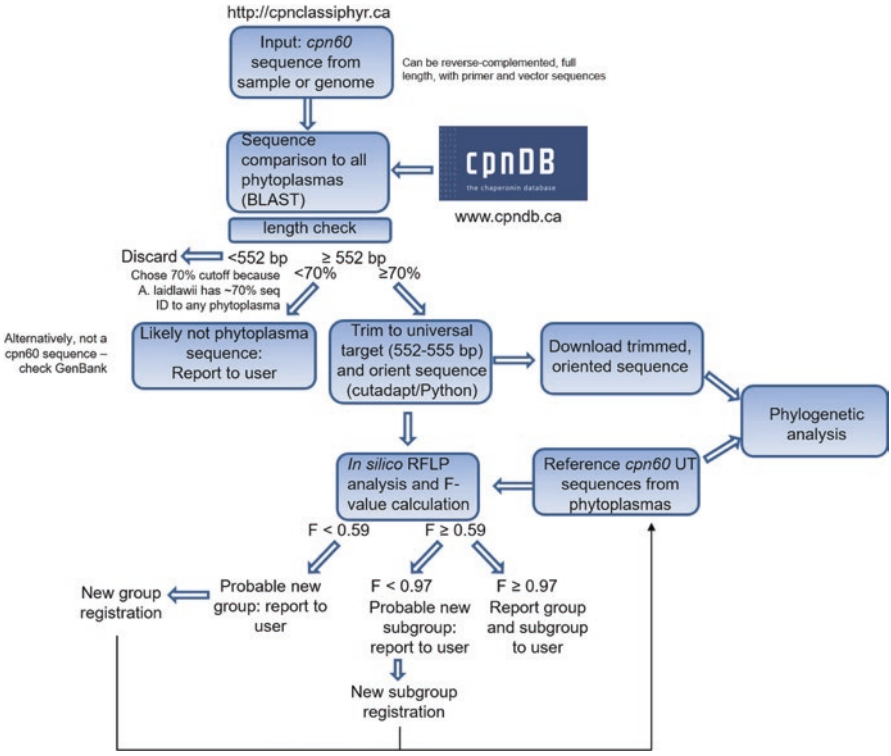

Fig. 1.4 Schematic diagram of the algorithm used by the CpnClassiPhyR for assigning group/subgroups to *cpn*60 UT sequences from phytoplasmas

sequences are added to the reference database to facilitate the identification of future phytoplasma query sequences.

Subgroups Detailed information is provided for each of the reference strains under the "subgroups" tab. The Subgroup Overview presents the known characteristics of all the reference strains in the database, including the *cpn*60 UT and 16S-based classifications, cpnDB identification number, the Genbank accession number, strain name and description including the '*Ca*. Phytoplasma' species, host information, sequence length and the virtual RFLP gel. When available, a link to the corresponding 16S rRNA-encoding sequence is also provided so that users can verify the 16S-based group/subgroup of all strains. The *cpn*60 UT sequences of all the reference strains can be downloaded from this tab. The second tab, "Number of Bands", shows the band sizes generated after the *in silico* digestion of the *cpn*60 UT sequences belonging to the reference strains with the seven endonucleases that form the *cpn*60 UT classification scheme. The third tab in this window is the Similarity Coefficient Matrix, which shows the F values determined by pairwise comparisons

of all the *cpn*60 UT reference strains. The last tab is the Subgroups Tree, which allows the user to view the phylogenetic relationship of all the *cpn*60 UT reference sequences – this is described in more detail below.

Primer Sets The primer sets used to access the *cpn*60 UT from any phytoplasma have been previously described (Dumonceaux et al., 2014) and are also presented here along with detailed information on PCR setup and conditions. It is expected that these primers may need to be modified as new *cpn*60 sequences are described and this will be updated accordingly.

Publications A list of relevant publications that provide context for the CpnClassiPhyR in the scientific literature is provided.

Downloads The code that comprises the CpnClassiPhyR is provided as a Github link (https://github.com/kevmu/CpnClassiPhyR). Here the code that comprises the CpnClassiPhyR is made openly available to the scientific community.

cpnDB We provide users with a direct link to cpnDB, the publicly available chaperonin database (http://cpndb.ca) (Hill et al., 2004).

1.2.3 New Group/Subgroup Registration Service

The "dashboard" tab opens a login portal where registered users can manually enter information required to define a novel *cpn*60 UT-based group or subgroup of phytoplasma. Information is collected from users on the phytoplasma species, strain, and proposed *cpn*60 group/subgroup. In addition, users are required to include the 16S-based classification of the strain, along with the GenBank accession numbers of both the *cpn*60 UT sequence and the 16S F2nR2 sequence. Users are also asked to provide data on the host, the geographic origin of the sample, and if possible the literature reference for the sequence being submitted. Upon submission, the new sequence is automatically downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) and subjected to analysis using the CpnClassiPhyR – a detailed report is sent by email to the user. If the CpnClassiPhyR results determine that the new sequence is a classified as a probable new group or subgroup, the new sequence and its associated information are added dynamically to the *cpn*60 reference dataset ("subgroups") used to calculate F values and phylogenetic trees to facilitate the identification of phytoplasma strains by all users.

1.2.4 CpnClassiPhyR Output

The analysis of a query sequence with CpnClassiPhyR results in an output window with five tabs. To illustrate the functionality of the software, we used the *cpn*60 sequence (full length) from the genome of the New Jersey Aster Yellows strain (NJ-AY; GenBank AB599703) (Sparks, Bottner-Parker, Gundersen-Rindal, & Lee, 2018). The first tab is the CpnClassiPhyR Overview, and it provides the major characteristics of the sequence such as its best match in cpnDB, the alignment length, the strandedness (+ or -) of the query sequence and the *E*-value with the best match. Users are also provided the option of downloading the trimmed, oriented *cpn*60 UT sequence in FASTA format (552 bp in this case). The second tab, Best Subgroup Matches, provides the key output of the software: the *cpn*60 UT-based RFLP classification of the best subgroup match of the query sequence, the similarity coefficient calculation, the F value, the *cpn*60 UT classification, and the RFLP patterns of the query sequence and the best match presented as virtual gels, which are downloadable. The results of the analysis can also be downloaded as a csv file. The third tab in this window is the "Number of Bands", where the software shows in a table the number of bands generated with each of the seven endonucleases (*Alu*I, *Bfa*I, *Hinf*I, *Hpa*I, *Mse*I, *Rsa*I, and *Taq*I) after analyzing the query sequence. The fourth tab in this window is the Similarity Coefficient Matrix, where the software shows the F values resulting from the comparison of the query sequence to all reference strains. In the present example, strain NJ-AY was identified as *cpn*60 UT I-IA, with an F-value of 1.00 compared to the reference strain for this subgroup, O2L. This is consistent with the identification of NJ-AY as a 16SrI-A strain (Sparks et al., 2018). The complete matrix of F-values generated in this tab can be downloaded by users as a csv file. The last tab, CpnClassiPhyR tree, shows the results of automated phylogenetic tree calculation of the input sequence in the context of the entire set of reference *cpn*60 UT sequences from phytoplasmas, with *Acholeplasma laidlawii* as an outgroup. Users can view an interactive tree (Fig. 1.5) with a link to CpnClassiPhyR_PhyD3_tree.php, a visualization tool powered by PhyD3 (https://phyd3.bits.vib.be/). Various options are provided for downloading phyloxml tree XML, SVG and PNG files associated with the phylogenetic tree. The input FASTA, MEGA options, summary, and CLUSTALW alignment files are also made available for download, allowing users to reproduce results.

Strain differentiation using the CpnClassiPhyR was also demonstrated using *cpn*60 sequences amplified from a symptomatic blueberry (*Vaccinium corymbosum*) plant in Nova Scotia, Canada, as well as from an insect collected in the same field. The insect was identified by CO1 barcoding analysis (Ratnasingham & Hebert, 2007) as *Penthimia americana* (CO1 sequence deposited to GenBank with accession number MF958490). This insect is not known to transmit phytoplasma infections in blueberry. Consistent with this, the *cpn*60 sequences amplified from the insect and the symptomatic plant in the same field were distinct, with discernable differences in digestion patterns produced by *Mse*I and *Rsa*I (Fig. 1.6). The

Fig. 1.5 Interactive phylogenetic analysis of the New Jersey Aster Yellows strain (NJ-AY; GenBank AB599703)

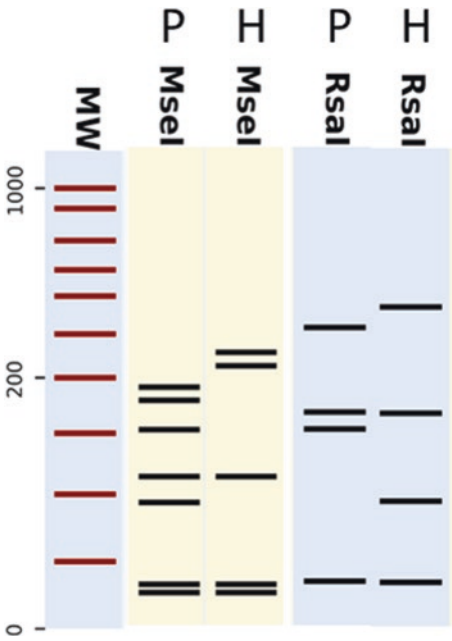

Fig. 1.6 Distinct restriction patterns with *Mse*I and *Rsa*I from *cpn*60 sequences amplified from infected plants (P) and a leafhopper (H) collected in the same field

Table 1.2 CpnClassiPhyR output resulting from analysis of *cpn*60 sequences amplified from an infected plant and an insect identified as *Penthimia americana* collected in the same field in Nova Scotia, Canada, in 2016

Strain	Best match	Similarity coefficient calculation	F value	Classification
Blueberry-cpn60	I-I(E/AI) AI	F = (2 ∗ 25) / (25 + 25)	1.00	Exact match to Cpn60 UT I-I(E/AI)AI
Insect-cpn60	I-IC	F = (2 ∗ 23) / (23 + 23)	1.00	Exact match to Cpn60 UT I-IC

Table 1.3 Similarity coefficient matrix generated by the CpnClassiPhyR of the insect-derived *cpn*60 sequence compared to other group I phytoplasma sequences

cpn60 UT	I-I(E/AI)AI	I-IA	I-IB	I-IC	I-IF	I-IIA	I-IIB	I-IIIB	I-IP	I-IVB	I-VB
Blueberry-cpn60	1.00	0.79	0.76	0.79	0.82	0.61	0.61	0.64	0.85	0.76	0.82
Insect-cpn60	0.79	0.74	0.75	1.00	0.81	0.72	0.64	0.67	0.71	0.79	0.72

CpnClassiPhyR placed these sequences in distinct *cpn*60 subgroups, with the plant-derived sequence matching the known sequence of blueberry stunt phytoplasma (*cpn*60 I-I(E/AI)AI) and the insect sequence matching *cpn*60 I-IC (AY-col) (Table 1.2). Moreover, the similarity coefficient matrix generated by the CpnClassiPhyR showed that the insect-derived sequence had an F value that placed it in the same group as all other group I phytoplasma sequences while the sequence was clearly identified as *cpn*60 UT I-IC (Tables 1.2 and 1.3).

1.2.5 Application of the CpnClassiPhyR to the Identification of New Groups and Subgroups

Using the phytoplasma *cpn*60 universal primers, we generated and analyzed *cpn*60 UT sequences from five phytoplasma strains belonging to four 16S rRNA based groups. The strains we have added to the *cpn*60 reference database for this study are Peru-PBT (16SrXV-B) described by Wei et al. (2017); strain SCWL (16SrXI-D) from China (Zhang et al., 2016), strain BGWL (16SrXIV-A) from Saudi Arabia (Omar, 2016), and strain R018 (16SrII-C) from Iran. In addition, the primers were applied to the phytoplasma strains TagTall (16SrIV-C), obtained from a Tagnanan Tall coconut variety in Tanzania, Sab4 (16SrIV-D), and Palm-A-DR (16SrIV-E), which was derived from a coconut palm in the Dominican Republic. No *cpn*60 sequences have previously been reported for 16S groups XI, XIV, and XV, and only other subgroups within groups 16SrIV (GenBank KY779619), 16SrXIV (Mitrović et al., 2015) and 16SrII (Chung, Chen, Lo, Lin, & Kuo, 2013) have been identified. All available information regarding the strains used in this study is provided in Table 1.1. For each of the strains included in this work, the 16S rRNA-encoding gene and *cpn*60 UT were amplified, cloned, and sequenced.

```
                TGRTTITCICCAAAACCAGGIGCATT    - H280p
                TGRTKITCICCAAAACYWGGIGCWTC    - D0318
845TGATTATCTCCGAATCCTGGAGCTTC820 - 16SrIV
                  *   *
```

Fig. 1.7 Alignment of the downstream primer landing site of the 16SrIV 'Cocos nucifera' phytoplasma *cpn*60 sequence (GenBank KY779619) with the phytoplasma *cpn*60 UT primers. The sites of mismatch are indicated (∗)

The results of phylogenetic analysis of phytoplasma strains based on *cpn*60 UT sequences was consistent with the 16S rRNA gene-based classification of the strains (Figs. 1.2 and 1.3). The BGWL strain belonging to the 16SrXIV group branched with the other strains within this group, and the new sequence from R018 clustered with Peanut Witches' Broom (Group 16SrII). The *cpn*60 UT sequence amplified from the strain PeruPBT, a member of the 16SrXV group ('*Ca*. P.' brasiliense), branched with the strains within the 16SrII group, as previously reported (Harrison et al., 2014). Similarly, the new *cpn*60 UT sequences from the strains in group 16SrIV reported in this study were closely related to strains belonging to the group 16SrV and 16SrVI, as previously reported (Harrison et al., 2014). While amplicon was successfully generated with the phytoplasma *cpn*60 primers using genomic DNA from 16SrIV-infected samples, examination of the amplified sequences and of a full-length *cpn*60 gene from 'Cocos nucifera' phytoplasma (GenBank KY779619) revealed two mismatches in the downstream primer landing site for 16SrIV phytoplasmas (Fig. 1.7). This resulted in the CpnClassiPhyR initially failing to trim correctly the full-length *cpn*60 gene to the UT; however the script has been adjusted to accommodate this observation.

In this study we have expanded the number of described *cpn*60 UT groups from nine to twelve (Fig. 1.2), which represents a third of the phytoplasma 16Sr groups that have been reported to date (Naderali et al., 2017). The known *cpn*60-based diversity of phytoplasmas was similarly expanded to include sixteen '*Ca.* Phytoplasma' species (Fig. 1.2). The phylogenetic tree obtained using the full diversity of known phytoplasma *cpn*60 sequences (Fig. 1.2) showed a clear differentiation of the three major phytoplasma subclades previously described (Chung et al., 2013; Hogenhout et al., 2008; Zhao, Davis, Wei, Shao, & Jomantiene, 2014; Zhao, Wei, Davis, & Lee, 2010), and corresponded well with the 16S rRNA-based phylogenetic tree (Fig. 1.2, S4).

Including these new sequences, the reference *cpn*60 UT database for the CpnClassiPhyR now includes 30 reference strains, with 11 subgroups within the *cpn*60 UT-I group, 2 subgroups within the *cpn*60 UT-II group, 2 subgroups within the *cpn*60 UT-IV, 1 subgroup within group *cpn*60 UT-V, 1 subgroup within group *cpn*60 UT-VII, 3 subgroups within *cpn*60 UT-IX, 3 subgroups within *cpn*60 UT-X, 1 subgroup within *cpn*60 UT-XI, 2 subgroups within *cpn*60 UT-XII, 1 subgroup within *cpn*60 UT-XIII, 2 subgroups within *cpn*60 UT-XIV, and 1 subgroup within group *cpn*60 UT-XV. This reference dataset is readily expandable as new *cpn*60 UT

sequences accumulate, either from genomic sequencing efforts or through the application of existing or adapted *cpn*60 phytoplasma-specific primers.

1.2.6 Discrepancies

The *cpn*60 UT RFLP patterns generated by strains 16SrIV-D (Sab4) and 16SrIV-E (Palm A-DR) were identical, so the RFLP scheme as currently described cannot discern these subgroups. Despite this, the *cpn*60 UT sequences from these strains were slightly more divergent than the corresponding 16S sequences – the strains shared 99.3% sequence identity at F2nR2 compared to 99.0% sequence identity at *cpn*60. Conversely, RFLP using F2nR2 sequences was unable to discern 16SrXIV-A (AL85/11) from 16SrXIV-C (RS59/11), while these strains were readily differentiated into distinct subgroups using *cpn*60 UT RFLP.

1.2.7 Application of the CpnClassiPhyR to cpn*60 Sequences Extracted from Genomes*

To determine the *cpn*60 UT group and subgroup classification for publicly available phytoplasma genomes, *cpn*60 and, where available, 16S rRNA-encoding gene sequences were extracted from 17 phytoplasma genomes, including a variety of 16S-based groups and subgroups (Table 1.4). Wherever possible we analyzed one or both 16S rRNA-encoding genes (depending on their availability within the genome sequence) using iPhyClassifier along with the full-length *cpn*60 gene using CpnClassiPhyR. In several cases, published genomes did not contain 16S rRNA-encoding genes, or the genes were of insufficient length for classification using the *i*PhyClassifier (e.g. *'Ca.* P.' solani strains 284/09 and 231/09; *'Ca.* P.' phoenicium strains SA213 and CHiP; *'Ca.* P.' oryzae strain Mbita1; Table 1.4). However, *cpn*60 genes were found in all analyzed genomes (Table 1.4). In nearly all cases in which 16S rRNA-encoding genes were available, the classification provided by the CpnClassiPhyR was consistent with the 16S rRNA group/subgroup classification. In the case of '*Ca.* P.' phoenicium strain CHiP, which lacks 16S rRNA-encoding gene sequence data, the *cpn*60 UT sequence extracted from the genome suggested that the strain may belong to a new *cpn*60 subgroup, with F = 0.93 to group IX-IJ (Table 1.4). In the absence of any known 16S rRNA-encoding gene sequences from this strain, we have not added it to the phytoplasma *cpn*60 UT RFLP reference database.

Several of the genome sequences also displayed clear evidence of 16S rRNA-encoding gene heterogeneity, including two strains of *'Ca.* P.' australiense (NZSb11 and NC_010544) with identical *cpn*60 sequences that placed them in a clade with *'Ca.* P.' solani', also a group 16SrXII phytoplasma (Fig. 1.2). Due to this clear evi-

Table 1.4 CpnClassiPhyR results using *cpn*60 sequences from published phytoplasma genomes

Species	strain	16S group-subgroup (F)	*cpn*60 UT group-subgroup (F)	accession no.
'*Ca.* P.' solani	284/09	ND[a]	XII-IA (1.00)	GCA_000970375.1
'*Ca.* P.' solani	231/09	ND[a]	XII-IA (1.00)	GCA_000970395.1
'*Ca.* P.' phoenicium	SA213	16SrIX[b]	IX-IB (1.00)	GCA_001189415.1
'*Ca.* P.' phoenicium	CHiP	ND[a]	IX-IJ (0.93)	GCA_002968365.1
'*Ca.* P.' australiense	NZSb11	16SrXII-B (0.98); 16SrXII-C (1.00)	XII-I(B/C)B (1.00)	CP002548.1
'*Ca.* P.' australiense	NC_010544	16SrXII-B (1.00); 16SrXII-C (1.00)	XII-I(B/C)B (1.00)	AM422018.1
Echinacea purpurea witches'-broom phytoplasma	NCHU2014	II-A (1.00)	II-IA (1.00)	NZ_LKAC00000000
'*Ca.* P.' oryzae	Mbita1	16SrXI[c]	X-IF (0.5)	NZ_LTBM00000000.1
'*Ca.* P.' asteris	New Jersey Aster yellows	16SrI-A	I-IA (1.00)	MAPF00000000.1
'*Ca.* P.' asteris	Onion yellows OY-M	16SrI-B (1.00)	I-IB (1.00)	AP006628.2
'*Ca.* P.' asteris	Maize bushy stunt phytoplasma	16SrI-B (1.00)	I-IIIB (1.00)	NZ_CP015149.1
'*Ca.* P.' asteris	TW1	16SrI-A (0.97); 16SrI-B (1.00)	I-IB (1.00)	QGKT00000000
'*Ca.* P.' mali		16SrX-A (1.00)	X-IA (1.00)	NC_011047.1
'*Ca.* P.' asteris	Wheat blue dwarf phytoplasma	16SrI-R (1.00); 16SrI-S (1.00)	I-IC (1.00)	NZ_AVAO00000000.1
'*Ca.* P.' asteris	Rice orange leaf phytoplasma	16SrI-B (1.00)	I-IIIB (1.00)	NZ_MWKN00000000.1
'*Ca.* P.' aurantifolia		ND[a]	II-IC (1.00)	NZ_MWKN00000000.1
Peanut witches' broom phytoplasma		16SrII-A (1.00)	II-IA (1.00)	NZ_AMWZ00000000

[a]ND, not determined. No 16S rRNA-encoding genes were found in the genomic sequence
[b]Subgroup undetermined. A single 16S rRNA-encoding gene was found in the genomic sequence, which was of insufficient length for classification using the iPhyClassifier. An accession number is provided for a 16S rRNA-encoding gene sequence for SA213 (KM275491), but this sequence types as 16SrIX-D while the strain is identified as 16SrIX-B (Quaglino et al., 2015)
[c]Subgroup undetermined. No 16S rRNA-encoding genes were found in the genomic sequence and the strain was identified only to the group level using 16S amplicon analysis (Fischer et al., 2016)