UNIVERSIDADE DE SÃO PAULO

Programa de Pós-Graduação Interunidades em Bioinformática

GUILLERMO UCEDA CAMPOS

Comparative genomics and phylogenomics of *Xylella fastidiosa*

Corrected version of the dissertation defended in 27/09/2019

São Paulo
Date of deposit in SPG: 04/09/2019

UNIVERSIDADE DE SÃO PAULO

Programa de Pós-Graduação Interunidades em Bioinformática

GUILLERMO UCEDA CAMPOS

Comparative genomics and phylogenomics of *Xylella fastidiosa*

Dissertation presented to the Institute of
Mathematics and Statistics of the
University of São Paulo

Advisor: Prof. Aline Maria da Silva
Co-Advisor: Prof. João Carlos Setubal

São Paulo
2019

Ficha Catalográfica

(elaborada pelo autor)

Campos, Guillermo Uceda

Genômica comparativa e filogenômica de *Xylella fastidiosa* (Comparative genomics and phylogenomics of *Xylella fastidiosa*) / Guillermo Uceda Campos – São Paulo, 2019.
85 p.

Dissertação (Mestrado) - Instituto de Matemática e Estatística da Universidade de São Paulo. Programa Interunidades em Bioinformática.
Orientador: da Silva, Aline Maria
Co-Orientador: Setubal, João Carlos

1. X. fastidiosa. 2. Elementos genéticos móveis. 3. prófago, 4. gene de virulência. 5. Anotação genômica. 6. GTACG. I. T. II. da Silva, Aline Maria, orientador

Genômica comparativa e filogenômica de *Xylella fastidiosa*

GUILLERMO UCEDA CAMPOS

Dissertação de Mestrado submetida à Universidade de São Paulo como parte dos requisitos necessários à obtenção do título de “Mestre em Ciências”. Programa: Bioinformática

APROVADO POR:

Profa. Dra. Aline Maria da Silva
(Orientadora e Presidente)

Prof. Dra. Claudia Barros Monteiro Vitorello
ESALQ – USP

Prof. Dr. Julio Cezar Franco de Oliveira
UNIFESP

Prof. Dr. Luciano Antonio Digiampietri
EACH - USP

São Paulo, 27/09/2019

DEDICATION

This Dissertation is dedicated to
my beloved parents Maria Yolanda and Lazaro Guillermo. I thank you
for your endless love, education, pray, sacrifice, patience and everything
you have done since I was born.

ACKNOWLEGEDMENTS

First, I express my deep gratitude to my family, for the education received and for the constant support at each stage of my personal and professional life.

I express my sincere gratitude to Dr. Aline Maria da Silva, my supervisor, for the continuous support to my Master study, research and writing of this thesis, for her patience, motivation, immense knowledge and enthusiasm for science. Besides my supervisor, I would like to thank my co-supervisor, Dr. João Carlos Setubal, for his expertise in the recommendation of the right methodologies throughout my Master study.

I would like to extent my honest gratitude to my research mates Caio Santiago, Oséias Feitosa, Paulo Pierry and Joaquim Martins for the help in providing the necessary information and concepts, formulation of ideas and discussion throughout the development of this thesis. At the same time to my lab friends Fernando Rossi, Deibs Barbosa, Roberta Pereira, Ariosvaldo Pereira, Remigio Rodrigues for the help, advices in the development of this work and for their friendship throughout these two years.

I also thank Alexandre Sanchez and Carlos Morais for the constant technical and professional support during the development of this work.

To Juan Faya and Edgar Llontop, my great friends with whom I shared very pleasant moments since the beginning of my Master study.

Finally, I thank the Bioinformatics Graduate Program, the Mathematics and Statistics Institute and the Chemistry Institute from the University of São Paulo for their support throughout this work.

Funding for this research was provided by grant 3385/2013 from the Coordination for the Improvement of Higher Education Personnel (CAPES). The author was supported by a fellowship from CAPES.

ABSTRACT

CAMPOS, GUILLERMO UCEDA. Comparative genomics and phylogenomics of *Xylella fastidiosa.* 2019. 85 pages. Master dissertation - Bioinformatics Graduate Program. São Paulo University, São Paulo.

The gamma-proteobacterium *Xylella fastidiosa* is an insect-transmitted, xylem-inhabiting pathogen and the causal agent of several plant diseases, most notably Pierce's disease of grapes (PD), citrus variegated chlorosis (CVC) and olive quick decline syndrome (OQDS). The first two complete genomes of *X. fastidiosa* sequenced in the early 2000s were from CVC (9a5c) and PD (Temecula1) strains. Since then, genomes of various isolates of *X. fastidiosa* have been sequenced, which are very similar to each other regardless of the host and/or geographical region from which such strains were isolated. Despite the available genomes in public databases, potential determinants of host adaptation and the heterogenicity among the prophage regions are still unknown in a wide genomic level. In this study, the CDSs of 46 *X. fastidiosa* genomes were compared using the new computational tool GTACG that deals with phylogenetically close organisms. Also, in order to explore the content of Mobile Genetic Elements (MGE) we have analyzed predicted prophages, genomic islands and insertion sequences harbored in *X. fastidiosa* chromosome. A total of 4942 and 1518 orthologs were found in the pan- and core-genome of *X. fastidiosa*, respectively. The phylogenomic trees showed three main clades related with the subspecies *pauca, fastidiosa* and *multiplex*, and subclades most related with the predicted sequence type and geographic region of isolation while the plant host information had less relationship. Most of the virulence and pathogenic-related orthologs were found in the core-genome of *X. fastidiosa*. In one case, the afimbrial adhesin orthologs XadA1 and XadA3 sequence diversities showed a relative congruence with the plant host. The MGE content per genome ranged from 12% to 28% in the 46 strains of *X. fastidiosa* analyzed. The mean of prophages and genomic island regions per genome were 8.6 and 8.9, respectively. Around half (56%) of predicted insertion sequences were located into prophage regions. In summary, the *X. fastidiosa* comparative genomics and phylogenomics analyses showed that the geographic region of isolation is strongly supported at the strains genomic level differently from strain-plant host adaptation. Moreover, a relevant and heterogeneous amount of MGE are harbored in the chromosome of all 46 *X. fastidiosa* strains analyzed.

Keywords: *X. fastidiosa*, mobile genetic elements, prophage, virulence gene, genome annotation, GTACG.

RESUMO

CAMPOS, GUILLERMO UCEDA. Genômica comparativa e filogenômica de *Xylella fastidiosa*. 2019. 85 páginas. Dissertação de Mestrado - Programa de Pós-Graduação Interunidades em Bioinformática. Universidade de São Paulo, São Paulo.

A gamaproteobacteria *Xylella fastidiosa* é transmitida por insetos, restrita ao xilema e agente causal de várias doenças de plantas, principalmente a doença de Pierce de videiras (PD), a clorose variegada dos citros (CVC) e a síndrome do rápido declínio de oliveiras (OQDS). Os dois primeiros genomas completos de *X. fastidiosa* sequenciados no início dos anos 2000 foram das cepas 9a5c e Temecula1, respectivamente causadoras de CVC e PD. Desde então, os genomas de vários isolados de *X. fastidiosa* foram sequenciados, os quais são muito semelhantes entre si, independentemente do hospedeiro e/ou da região geográfica de origem dessas cepas. Apesar da disponibilidade desses genomas em bancos de dados públicos, os potenciais determinantes da adaptação do hospedeiro e a heterogeneidade de regiões do prófagos ainda não foram analisados em nível genômico mais amplo. No presente estudo, as CDSs de 46 genomas de *X. fastidiosa* foram comparadas usando a nova ferramenta computacional GTACG dedicada à análise de genomas de organismos filogeneticamente próximos. Além disso, visando explorar o conteúdo de Elementos Genéticos Móveis (MGE), também foram analisadas as regiões preditas como prófagos, ilhas genômicas e sequências de inserção no cromossomo de *X. fastidiosa*. Foram encontrados 4942 e 1518 ortólogos no *pan*- e *core*-genoma de *X. fastidiosa*, respectivamente. As árvores filogenômicas mostraram os três clados principais relacionados às subespécies *pauca*, *fastidiosa* e *multiplex*, e os subclados foram mais relacionados ao *sequence type* predito e região geográfica de isolamento, enquanto as informações do hospedeiro vegetal tiveram menor relação. A maior parte dos ortólogos relacionados a virulência e patogenicidade foram encontrados no *core*-genoma de *X. fastidiosa*. Em um caso, a diversidade de sequência dos ortólogos das adesinas afimbriais XadA1 e XadA3 apresentou relativa congruência com o hospedeiro vegetal. O conteúdo de MGE por genoma variou de 12% a 28% nas 46 cepas de *X. fastidiosa* analisadas. A média de regiões de prófagos e ilhas genômicas por genoma foi de 8,6 e 8,9, respectivamente. Cerca de metade (56%) das sequências de inserção previstas foram localizadas em regiões de prófagos. Em resumo, as análises de genômica comparativa e de filogenômica de *X. fastidiosa* mostraram forte relação entre a região geográfica de isolamento e os distintos genomas, mas não com a adaptação cepa-hospedeiro vegetal. Além disso, verificamos que todos os 46 genomas de *X. fastidiosa* analisados contém um conteúdo relevante e heterogêneo de MGE.

Palavras-chave: *X. fastidiosa*, elementos genéticos móveis, prófago, gene de virulência, anotação genômica, GTACG

LIST OF FIGURES

Fig. 13. Alignment of the sequences belonging to the LesA-LesB ortholog. In the black boxes are showed the reported catalytic sites. Red and green arrows indicate the lytic LesA and non-lytic LesB in Temecula1, respectively.

Fig. 14. COG functional analysis of the core-genome and accessory-genome. Notably, the category more enriched is X, which is related to the Mobilome.

Fig. 15 Mobile Genetic Elements (MGE) in *X.* fastidiosa genomes. **a** Heatmap showing the density of each MGE analyzed (PPH: prophage, GI: Genetic Island and IS: insertion Sequence). **b** Percentage of the MGE in the genome.

Fig. 16 Total and intact prophages among *X. fastidiosa* genomes.

Fig. 17. Distribution of prophages with lytic potential in *X. fastidiosa* genomes. Each circle represents a prophage, which color green indicates a specific inovirus prophage. In some case more than one homologous prophage was present in the same strain (darker circles). The circle size indicates a proportional size of the prophage according the legend.

Fig. 18. Alignment of filamentous prophage sequences of pph_1 group (Figure 16).

Fig. 19. Total of genomic islands (GIs) (pink) and GI with functional category found (grey) among *X. fastidiosa* genomes.

Fig. 20. Distribution of the Genetic Island (GI) with predicted function in *X. fastidiosa* genomes. RI Resistance island; PAI Pathogenic Island, MI Metabolic Island; SI Symbiotic Island.

Fig. 21 Mapping of the MEG in the *X. fastidiosa* strains with complete genome. The percentage of MGE by genome is showed in parentheses. C: Chromosome, PPH: prophage, GI: genetic island, IS: insertion sequence, I: inovirus.

LIST OF TABLES

LIST OF ABREAVIATIONS

ANI	Average Nucleotide Identity
bp	base pair
BLAST	*Basic Local Alignment Search Tool*
CATG	Center for Advanced Technologies in Genomics
CDS	Protein Coding Sequence
CD-HIT	Cluster Database at High Identity with Tolerance
COG	Cluster of Orthologous Group
CS tree	Consensus tree
GTACG	Gene Tags Assessment by Comparative Genomics
CVC	Citrus Variegated Chlorosis
CWDE	Cell Wall Degrading Enzymes
EPS	Exopolysaccharide
GI	Genomic Island
HQD	High Quality Draft
HTE	Horizontal Transfer Element
IQ-USP	Instituto de Química da Universidade de São Paulo
IS	Insertion Sequence
kbp	kilo (10^3) base pairs
LPS	Lipopolysaccharide
Mbp	mega (10^6) base pairs
MCL	Markov cluster
MGE	Mobile Genetic Element
MI	metabolic island
ML tree	Maximum Likelihood tree
MLST	Multilocus sequence typing
NCBI	National Center for Biotechnology Information
OQDS	Olive Quick Decline Syndrome
P1	subclade 1 of subspecies *pauca*
P2	subclade 2 of subspecies *pauca*
PAI	pathogenicity island
PAMP	Pathogen-associated molecular pattern
PD	Pierce's Disease
PPH	Prophage
RefSeq	NCBI reference sequence database
RI	resistance island
ROARY	Pipeline for rapid large-scale prokaryote pan genome analysis
rRNA	ribosomal RNA
SI	symbiotic island
ST	Sequence Type
TAA	Trimeric Autotransporter Adhesin
tmRNA	transfer-messenger RNA
tRNA	transfer RNA
USA	United States of America

CONTENTS

1. Introduction

The bacterium *Xylella fastidiosa* is a plant endophyte native to the Americas which is transmitted between plants by xylem-feeding insects (Chatterjee *et al.*, 2008a; Janse & Obradovic, 2010; Sicard *et al.*, 2018). Although *X. fastidiosa* has been only isolated and described in the late 1980s, this bacterium was proven to be the causal agent of the Pierce´s disease (PD) of grapes reported in the 1880s in California, USA (Hopkins & Purcell, 2002). A century later (1987), the Citrus Variegated Chlorosis (CVC), a disease caused by *X. fastidiosa*, was reported in the state of Minas Gerais, Brazil (Rossetti *et al.*, 1990). Diseases associated to *X. fastidiosa* infection have been also described in other Latin American countries such as Costa Rica, Argentina, Ecuador and Paraguay (Janse & Obradovic, 2010). In the last few years, this phytopathogen emerged in European (Italia, Portugal, Germany and Spain) and Asian (Taiwan) countries, affecting many crops of economic importance such as olive, grape, cherry and plum (Godefroid *et al.*, 2019).

After the publication of the first *X. fastidiosa* genome sequence in 2000 (Simpson *et al.*, 2000), a number of studies have contributed to a better understanding of the physiology and pathogenesis of this bacterium (Chatterjee *et al.*, 2008a; Lindow, 2019). Over the past 19 years, genome sequences of several *X. fastidiosa* strains/isolates have been released which are being constantly analyzed with different approaches and purposes (Bhattacharyya *et al.*, 2002a; Bhattacharyya *et al.*, 2002b; Van Sluys *et al.*, 2002; Van Sluys *et al.*, 2003; Koide *et al.*, 2004; Monteiro-Vitorello *et al.*, 2005; da Silva *et al.*, 2007; Varani *et al.*, 2008; Varani *et al.*, 2012; Marcelletti & Scortichini, 2016b; Denance *et al.*, 2019; Vanhove *et al.*, 2019).

In this study, we performed a comparative analysis of 46 *X. fastidiosa* genomes which were re-annotated with PROKKA software (Seemann, 2014) and manually curated. For this analysis, we chose the new tool GTACG framework (Santiago *et al.*, 2019) which uses an algorithm developed to deal with phylogenetically close genomes (Santiago *et al.*, 2018). General statistics and analysis of *pan*, *accessory*, *core* and *singleton* genome will be presented aiming to highlight relevant traits found in these genomes. We also constructed phylogenomic and phylogenetic trees as an attempt to infer relationships

between strains and their respective metadata such as geographic region and plant host of isolation as well as strain Sequence Type. Lastly, we performed a comprehensive analysis of the mobile genetic elements harbored in *X. fastidiosa* chromosome, which are known to comprise 9% to 15% of its genome and have been suggested to play an important role in the biology and evolution of *X. fastidiosa* (Monteiro-Vitorello *et al.*, 2005; Varani *et al.*, 2012).

1.1. *Xylella fastidiosa* overview

X. fastidiosa is a phytopathogenic Gram-negative bacterium that belongs to Gamma-proteobacteria class, Xanthomonadaceae family. This bacterium is rod-shaped without flagella and migrates via twitching motility due to its long type IV pilus (Wells *et al.*, 1987; Li *et al.*, 2007; Chatterjee *et al.*, 2008a). Its growth is slow during *in vitro* cultivation, optimal temperature of 26-28°C, closely resembling a facultative anaerobic rather than an obligate aerobic microorganism (Shriner & Andersen, 2014). This bacterium colonizes the xylem vessels of several plant species (Janse & Obradovic, 2010), as well as the buccal apparatus of its xylophagous insect vectors, which belong to the families Cicadellidae and Cercopidae (Backus *et al.*, 2015). In these environments *X. fastidiosa* cells actively aggregate and form a biofilm (Newman *et al.*, 2004; Chatterjee *et al.*, 2008a). The bacterium may also be present in the roots and therefore may be transmitted by contaminated root grafts (Janse & Obradovic, 2010).

X. fastidiosa causes diseases of economical relevance such as Pierce's disease (PD) in grapevines (Hopkins & Purcell, 2002), Citrus Variegated Chlorosis (CVC) (Rossetti *et al.*, 1990; Chang *et al.*, 1993) and the Olive Quick Decline Syndrome (OQDS) (Cariddi *et al.*, 2014; Della Coletta *et al.*, 2016), among others (Janse & Obradovic, 2010; Sicard *et al.*, 2018). On the other hand, this bacterium can be found as a non-pathogenic endophyte in plants such as sagebrush (*Artemisia douglasiana*) and grass species (*Echinochloa cruzgalli*) which might serve as *X. fastidiosa* reservoirs and source of inoculum for vectors (Hopkins & Purcell, 2002; Chatterjee *et al.*, 2008a).

Related plant host species behave differently in response to colonization by *X. fastidiosa*. For instance, *Citrus sinensis* is more susceptible to CVC than other species such as *C. reticulata*, *C. limonia* and others (Rossetti & De Negri, 1990). However, there are reports that the same strain of *X. fastidiosa* can

colonize plants of different species in artificial inoculations (Lopes *et al.*, 2010; Oliver *et al.*, 2015). *X. fastidiosa* lacks type III secretion system (T3SS) (Van Sluys *et al.*, 2002) which is important to host adaptation determination (McCann & Guttman, 2007; Buttner, 2016). It has been suggested that the adaptation of *X. fastidiosa* strains to their respective hosts involves regulation of gene expression (Killiny & Almeida, 2011), but the molecular mechanisms involved remain unknown.

Unlike most phytopathogenic bacteria, *X. fastidiosa* is limited to the xylem (Wells *et al.*, 1987; Newman *et al.*, 2003). Diseases caused by this pathogen are typically associated to leaf scalding and depend on the ability of *X. fastidiosa* to spread away from the point of the insect inoculation, multiply and colonize the xylem of the infected plant. For the colonization of adjacent xylem vessels, *X. fastidiosa* appears to degrade the pit membrane, a primary cell wall barrier that blocks the passage of larger particles such as pathogen cells allowing the movement of xylem sap only (Ellis *et al.*, 2010; Perez-Donoso *et al.*, 2010; Ingel *et al.*, 2019).

Plants infected by *X. fastidiosa* have xylem vessels blocked by bacterial cell aggregates, which are thickened by the presence of exopolysaccharide secreted by the bacteria and other components such as extracellular DNA (Alves *et al.*, 2003; Newman *et al.*, 2004; Roper *et al.*, 2007a; Clifford *et al.*, 2013). Thus, symptoms of *X. fastidiosa*-infected plants appear to be related to water and/or nutritional stresses resulting from xylem occlusion (Purcell & Hopkins, 1996; Hopkins & Purcell, 2002; Chatterjee *et al.*, 2008b). *X. fastidiosa* infection has been shown to cause significant changes in the levels of certain mineral elements in the plant (De La Fuente *et al.*, 2013). Disease symptoms might also derive from the plant response to the *X. fastidiosa* infection. It has been shown that the infected plant release reactive oxygen species (Carazzolle *et al.*, 2011; Zaini *et al.*, 2018) and, as such, *X. fastidiosa* might degrade such toxic compounds to achieve maximal xylem colonization (Wang *et al.*, 2017).

1.2. Pathogenicity and virulence of *X. fastidiosa*

X. fastidiosa produces a wide variety of pathogenicity factors for successful host colonization (Chatterjee *et al.*, 2008a; Janse & Obradovic, 2010). Among these factors fimbrial and afimbrial adhesins as well as

exopolysaccharide (EPS), named fastidian gum, are important for cell adhesion and biofilm formation (Simpson *et al.*, 2000; da Silva *et al.*, 2001; Feil *et al.*, 2007; Chatterjee *et al.*, 2008a; Caserta *et al.*, 2010; Voegel *et al.*, 2010; Zaini *et al.*, 2015). Both in plant and insect hosts, *X. fastidiosa* is exposed to high turbulence in a nutrient-poor environment and subject to host defense responses. Under these conditions, biofilm formation is strategic for its survival and multiplication (de Souza *et al.*, 2003; Guilhabert & Kirkpatrick, 2005; Caserta *et al.*, 2010).

The fimbrial adhesins MrkD and PilY1 are, respectively, components of the short pilus (type I pilus) and long pilus (type IV pilus) which are located at one of the cell poles. While *X. fastidiosa* short pilus is related to cell aggregation, adhesion to surfaces and biofilm formation, the long pilus, in addition to contribute to adhesion, is used by the bacteria for twitching motility (Li *et al.*, 2007; Zaini *et al.*, 2015). Twitching motility allows *X. fastidiosa* to move in the opposite direction of xylem sap flow for systemic colonization of the plant host. This movement also seems to involve chemosensory systems (Cursino *et al.*, 2011).

X. fastidiosa genome encodes different types of afimbrial adhesins. The filamentous hemagglutinins are high molecular weight proteins (>250kDa characterized by having a carbohydrate-dependent agglutination domain, a region containing the RGD motif characteristic of integrin-binding proteins, a heparin-binding domain that mediates cell surface glycolipid binding, and long signal peptides which are cleaved at both ends and various repetitions that vary in length and sequence between members of this protein family (Guilhabert & Kirkpatrick, 2005; Voegel *et al.*, 2010). The genome of Temecula strain (isolated from grapevines) encodes two hemagglutinins, HxfA and HxfB, PD2118 and PD1792, respectively, which are important for cell adhesion, biofilm formation and vector transmission (Guilhabert & Kirkpatrick, 2005; Killiny & Almeida, 2009; Voegel *et al.*, 2010; Killiny & Almeida, 2014).

Another type of afimbrial adhesins found in *X. fastidiosa* are the trimeric autotransporter adhesin (TAA) such as XadA1 and XadA2. The former was detected at all stages of biofilm development, while the latter was mainly detected in the final stage, suggesting that it has a role in biofilm stability (Caserta *et al.*, 2010). The genome of *X. fastidiosa* also encodes classical autotransporter adhesins (AT). These proteins (XatA, XatB and XatC in Temecula1 strain) are

located in the outer membrane and are also important in the aggregation for biofilm formation and virulence (Matsumoto *et al.*, 2012).

The *X. fastidiosa* virulence determinants also include the lipopolysaccharide (LPS). This molecule is composed of a conserved oligosaccharide component, lipid A, and a variable portion of O-antigen. LPS are among the well-known pathogen-associated molecular patterns (PAMPs). They also act as potent selective barriers to the entry of certain substances and are potent elicitors of PAMP-triggered immunity (Kutschera & Ranf, 2019). In *X. fastidiosa,* the O-antigen is important for surface adhesion, cell-cell aggregation, and biofilm maturation (Clifford *et al.*, 2013). Rapicavoli and colleagues have shown that a long-chain O-antigen may delay initial recognition by the plant, effectively preventing elicitation of innate immunity and thus contributing to the establishment of the pathogen in the host (Rapicavoli *et al.*, 2018b). Studies have suggested that among coffee strains there are different LPS profiles that could influence interaction with their respective host and vector (Alencar *et al.*, 2017).

As already mentioned, *X. fastidiosa* lacks type III secretion system (T3SS) (Van Sluys *et al.*, 2002; Killiny & Almeida, 2011) which is commonly used by most phytopathogenic bacteria to secrete protein effectors that enhance virulence and/or suppress plant defenses (McCann & Guttman, 2007; Buttner, 2016). The absence of T3SS in *X. fastidiosa* may reflect the fact that this bacterium does not colonize parenchymal tissue (living cells), but rather xylem, which consists predominantly of dead cells (Chatterjee *et al.*, 2008a). Nevertheless, *X. fastidiosa* could use another secretion systems to release cell wall degrading enzymes (CWDEs), lipases/esterases, proteases and toxins (Ellis *et al.*, 2010; Perez-Donoso *et al.*, 2010; Nascimento *et al.*, 2016; Rapicavoli *et al.*, 2018a; Feitosa-Junior *et al.*, 2019; Ingel *et al.*, 2019) which can contribute for its systemic dissemination and to overcome the host defense response.

The CWDEs encoded by *X. fastidiosa* genome are cellulases, hemicellulases, endoglucanases and polygalacturonases which would possibly be secreted into the extracellular environment by the type II secretion system (Chatterjee *et al.*, 2008a). These enzymes have also been suggested as determinants for host adaptation as the repertoire of active enzymes from different strains may vary as well as the carbohydrate composition in plant host xylem (Roper *et al.*, 2007b; Ingel *et al.*, 2019).

Another important group of enzymes related to virulence and pathogenicity are proteases and lipases/esterases. Several proteases are encoded in *X. fastidiosa* genome, some of which are described as important pathogenicity factors, such as PD0218 and PD0956 in Temecula1 (Zhang *et al.*, 2011; Gouran *et al.*, 2016). Secreted proteases that would play a role in assisting CWDEs in the degradation of vessel communication membranes have been detected in the extracellular proteome of citrus strains (Mendes *et al.*, 2016; Feitosa-Junior *et al.*, 2019). These enzymes also can play a role in pathogen nutrition (Fedatto *et al.*, 2006; Gouran *et al.*, 2016). Among the lipases found in *X. fastidiosa* genomes, LesA was associated with pathogenicity due to its demonstrated hydrolytic activity (Gouran *et al.*, 2016; Nascimento *et al.*, 2016). Moreover, the absence of LesA in the non-pathogenic strain EB92-1 evidenced the role of this enzyme in virulence. (Zhang *et al.*, 2011). Both LesA and LesB have been found associated to outer membrane vesicles in *X. fastidiosa* (Nascimento *et al.*, 2016; Feitosa-Junior *et al.*, 2019).

Hemolysins and microcins are a group of important toxins that *X. fastidiosa* might produce. It has been suggested that the function of hemolysins could be related to pore formation in the host cell membrane (Simpson *et al.*, 2000). In *X. fastidiosa* four RTX class hemolysin coding genes have been reported. Recent studies have shown that these genes are expressed during disease in grapevines and have a complex repetitive structure formed by conserved domains in the C-terminal portion and hypervariable domains in the N-terminal portion. It has been suggested that this organization shows extensive evolutionary restructuring through gene horizontal transfer and heterologous recombination mechanisms in *X. fastidiosa*. The operons where these genes are found have a widely duplicated module that encodes small proteins with homology to the N-terminal region. Surprisingly, some of these small peptides are more similar not to their corresponding full-length RTX, but to the N-terminal portions of the RTXs of other *X. fastidiosa* strains, and even of other remotely related species, suggesting that those modular repeats are new mobile genetic elements (Gambetta *et al.*, 2018). Regarding the microcins, more than a dozen genes for this class of bacteriocins have been annotated in the genome of *X. fastidiosa* strain 9a5c (Rodrigo R. Duarte & Aline M. da Silva, unpublished data) in addition to its three genes encoding colicin V-like microcin (Pashalidis *et al.*, 2005). These microcins may

function as toxins against predators or competitors sharing the same habitat as *X. fastidiosa*. The presence of other bacteria in both the host xylem and the anterior part of the digestive tract of the insect vector has been reported (Lacava *et al.*, 2006; Rogers & Backus, 2014).

The *X. fastidiosa* genome has a set of genes called *rpf* (regulation of pathogenicity factors) which encode proteins responsible for the synthesis and secretion of diffusible signaling factors (DSF) (Chatterjee *et al.*, 2008a; Chatterjee *et al.*, 2008c). DSF comprises a family of related unsaturated fatty acids (Ionescu *et al.*, 2016) that regulates bacterial behavior in a cell density-dependent manner, modulating the transition from less adhesive and motile phenotype to more adhesive sessile cells (Chatterjee *et al.*, 2008a). For example, in the early period of host colonization when DSF concentration is low, there is up-regulation of genes related to long pilus production for bacterial motility and cell wall degradation enzymes. As the bacterial population increases, DSF concentration increases, which signals a phenotypic change, and the cells become more aggregated and able to be acquired by the insect. Large clusters of cells eventually occlude the xylem vessels, which increases the chance that bacterial cells will be acquired by the insect vector and transmitted to other hosts. Due to vessel occlusion, symptoms may appear. Positively regulated genes at this stage include those encoding fimbrial and afimbrial adhesins, as well as increased production and secretion of EPS (Chatterjee *et al.*, 2008a). Accordingly, a mutant blocked in the DSF production has a less adhesive phenotype, is hypervirulent to plant host, and is impaired in insect colonization and transmission (Newman *et al.*, 2004; Ionescu *et al.*, 2013).

1.3. *X. fastidiosa* Genomics

The complete genome of *X. fastidiosa* 9a5c, a strain that causes CVC (Li *et al.*, 1999) was published in 2000 (Simpson *et al.*, 2000). The genome comprises a 2,679,305-base pair (bp) circular chromosome and two plasmids, one of 51,158 bp and one of 1,285 bp. Two years later, the genomes of the Ann-1 and Dixon strains from California, isolated strains of oleander and almond, respectively, were sequenced (Bhattacharyya *et al.*, 2002a; Bhattacharyya *et al.*, 2002b). The chromosome sizes of these strains, 2.78 Mbp for Ann-1 and 2.62 Mbp for Dixon are similar to 9a5c genome, although they have not been closed.

One of the first studies in *X. fastidiosa* comparative genomics showed that the 9a5c, Ann-1, and Dixon genomes share 82% of the predicted ORF clusters. Analyses of these genomes indicated that fimbriae, adhesins, quorum sensing proteins, and lipid A production should be fundamental in the pathogenicity of this bacterium. Likewise, it was suggested that host adaptation is related to conjugation systems, phage, genomic islands and ribosomal proteins. The presence of gene regions that could have been acquired by horizontal gene transfer, such as the type IV secretion system present in soil bacteria, and genes related to conjugation in enterobacteria were noted (Lambais *et al.*, 2000; Simpson *et al.*, 2000; Bhattacharyya *et al.*, 2002b).

In 2003, the complete genome of the *X. fastidiosa* strain Temecula1, an isolated from grapevine with Pierce's disease (PD) in California (Van Sluys *et al.*, 2003) was published. The chromosome size of this strain is 2.52 Mbp and only a small 1,345 bp plasmid was detected, similar to the one previously reported by for strain 9a5c (Simpson *et al.*, 2000). Compared to strain 9a5c, both strains share 98% of ORFs. Of these orthologous genes, 94.5% have a high percentage of amino acid sequence identity. Genome differences were associated with phage sequences, genomic islands, and deletions that could possibly be related to the adaptation of the strains to their respective hosts (Van Sluys *et al.*, 2003). Few years later, the M12 and M23 strains that caused almond leaf scald (ALSD) in California were sequenced. The chromosome size in these strains was slightly smaller than those mentioned above, with 2.48 Mbp and 2.25 Mbp, respectively. Unlike the M12 strain that has no plasmid, the M23 strain had one with 38,297 bp (Chen *et al.*, 2010).

Subsequently, the genome of strain *X. fastidiosa* GB514 isolated from a grapevine in the state of Texas was sequenced. In this strain a plasmid of 26 Kbp was detected, which shares 98% identity with the mulberry strain plasmid (Schreiber *et al.*, 2010). *X. fastidiosa* EB92-1, a biocontrol strain that is capable of colonizing grapevines but does not cause PD symptoms was also sequenced revealing that its genome may have lost 10 effectors potentially related to pathogenicity (Zhang *et al.*, 2011). In 2013, the genomes of the reference strain *X. fastidiosa* ATCC 35871 and of *X. fastidiosa* Griffin-1 strain, isolated from Quercus rubra in Griffin, Georgia, were sequenced. A comparative analysis of these V with those available to date showed that 5 loci are 100% identical to those

of the M12 strain, which indicated the phylogenetic relationship between them (Chen *et al.*, 2013).

X. fastidiosa also infects coffee trees, for instance the strains 6c and 32. The sequencing of those strains indicated variations in toxin producing genes as well as surface factors, which are likely to be involved with specific host recognition, infectivity and virulence capacity development (Alencar *et al.*, 2014; Barbosa *et al.*, 2015). Over the years, other genomes of diverse origins and hosts have been sequenced and made available on NCBI GenBank (Guan *et al.*, 2014a; Guan *et al.*, 2014b; Jacques *et al.*, 2016; Giampetruzzi *et al.*, 2017a; Van Horn *et al.*, 2017) including the Salento-1 and Salento-2 strains from the Puglia region of southern Italy (Ramazzotti *et al.*, 2018).

Studies on the genomes of *X. fastidiosa* revealed important features such as the existence of plasmids in most strains, considerable number of shared genes among strains, and genomic differences limited to the presence of prophages and indels in the genomes (Bhattacharyya *et al.*, 2002a; Van Sluys *et al.*, 2003). Studies have also described low genomic variation and have suggested that phylogenetic groups that colonize different hosts have similar pathogenicity mechanisms and strong selection possibly due to host adaptation (Schuenzel *et al.*, 2005). Despite the limited differences between strains, other studies have shown the occurrence of phage sequences, plasmids and genomic islands, which are transcriptionally active. It has been proposed that such genomic elements could explain *X. fastidiosa* ability to infect a wide variety of plant species (Nunes *et al.*, 2003). Release of phage particles by *X. fastidiosa* strains has already been demonstrated (Chen & Civerolo, 2008; Ahern *et al.*, 2014). Additionally, it has been described that recombination may occur in relatively high proportions, which may play an important role in the genetic diversity of *X. fastidiosa* (Kung *et al.*, 2013).

The genomic characterization of *X. fastidiosa* allowed the identification of horizontal transfer elements (HTEs). DNA microarray analysis and genome sequencing of coffee-infecting strains in Brazil confirmed the size and importance of HTEs as the main mediators of chromosomal evolution between the *X. fastidiosa* strains. This study also identified differences in gene content primarily in genes that produce toxins and surface-related factors involved in recognizing host-specific factors in different pathogenic bacteria (Barbosa *et al.*, 2015).

Subsequent studies of comparative genomics using strains from Europe associated with the rapid decline syndrome of olive trees confirmed the occurrence of the three subspecies of *X. fastidiosa*, named *fastidiosa*, *multiplex* and *pauca*. The study also showed the occurrence of a clonal genetic complex of four strains within the *pauca* clade that could have evolved in Central America (Marcelletti & Scortichini, 2016a).

Recently new studies used other methods to compare genomes of *X. fastidiosa*, for instance, an specific k-mers identification software that allows to check the synonymy between strains of *X. fastidiosa* from different collections (Denance *et al.*, 2019). In this study, specific SNPs in 16S rRNA sequences could discriminate alleles of the subspecies *fastidiosa*, *multiplex*, *morus*, *sandy* and *pauca*. Nevertheless, when they examined the taxonomy, only three clades strongly supported by k-mers defined the subspecies pauca (i), multiplex (ii) and the combination of *fastidiosa*, *morus* and *sandy*. The genomic diversity and recombination in *X. fastidiosa* have been explored by inter and intra-subspecies analyses, identifying differences in each phylogenetic clade, such as enriched gene ontologies (Denance *et al.*, 2019), as well as distinct recombination rates and events (Vanhove *et al.*, 2019).

1.4. Prophages in *X. fastidiosa*

Bacteriophages, or simply phages, are virus that infect bacteria which are highly abundant in nature and present an extraordinary genomic diversity (Koonin, 2010; Clokie *et al.*, 2011; Hatfull & Hendrix, 2011). Phages depend on their hosts for their replication and propagation. Upon infection of their specific hosts, bacteriophages inject their genetic material into the cell and then follow a lytic or lysogenic cycle. The lytic cycle is characterized by viral replication and subsequent production of new viral particles which released upon host cell lysis. These phages are called lytic phages (Kutter & Sulakvelidze, 2005; Howard-Varona *et al.*, 2017).

Some phages, named temperate phages, although capable of performing a lytic cycle, can establish longer lasting relationship with the host by integrating its genome in the host chromosome. The integrated phage is called prophage, and its genome replicates along with its host where it can remain for infinite generation. Some prophages are active and can be induced by chemical or

physical stresses, particularly those that cause DNA damage, or by spontaneous induction, and as such the lytic cycle is triggered (Howard-Varona *et al.*, 2017). Analysis of prophages sequences have suggested that in several cases, after the being integrated into bacterial genomes, they undergo a complex decay process consisting of inactivating point mutations, genome rearrangements, modular exchanges, invasion by further mobile DNA elements, and massive DNA deletion, turning into defective prophages (Canchaya *et al.*, 2003). Prophages can participate in a number of bacterial cellular processes, including antibiotic resistance, stress response, and virulence and are directly related to the genome diversity of the host cell, contributing positively or not, to bacterial fitness (Touchon *et al.*, 2016; Argov *et al.*, 2017).

About half of the sequenced bacteria are lysogens, representing a tremendous and previously under-explored source of prophages (Touchon *et al.*, 2016). *X. fastidiosa* is among these bacteria. Some studies have investigated at a genomic level the prophages present in *X. fastidiosa* genome. It has been identified that up to 15% of the genome of this pathogen could be formed by phage-like elements and phage remnants (Varani *et al.*, 2008). In this study, the role and diversity of each phage region was determined. It was also revealed the integration mechanism of the integrase sites associated with these regions and the influence on the genome differentiation of *X. fastidiosa* (Varani *et al.*, 2008).

The presumably reason for genomic differences among *X. fastidiosa* strains has been suggested to be the diversity of mobile genetic elements. *X. fastidiosa* genome comparisons have described high number of sequences showing features of prophages and genomic islands, which, together with plasmids, correspond to a flexible gene pool of up 18% of the genomes (Simpson *et al.*, 2000; Bhattacharyya *et al.*, 2002a; Van Sluys *et al.*, 2003; Varani *et al.*, 2008; Barbosa *et al.*, 2015). Prophages can be responsible for moving new important genes between strains through lateral transfer, possibly changing the bacterial host phenotype and may grant advantages in specific niches or environmental conditions. Besides, they may provoke rearrangements and deletions in the genome, significantly contributing to the evolution of the bacterial host (Varani *et al.*, 2008; Barbosa *et al.*, 2015).

Other studies have shown more evidence of the activity of the phage regions of *X. fastidiosa*. Transcriptome analyses of *X. fastidiosa* strains indicated these

elements are transcriptionally active, and suggested that their differential expression could explain the adaptability of *X. fastidiosa* to infect a wide range of plant species (Nunes *et al.*, 2003). Icosahedral viral particles were observed through transmission electron microscopy in two concomitant and independent studies, both *in planta* and *in vitro*, proving that *X. fastidiosa* prophages are indeed active (Chen & Civerolo, 2008; Varani *et al.*, 2008). These phages have been classified in the *Siphoviridae* family (lambda-like phage), although their *in-silico* analyses also found sequences of the *Myoviridae* and *Podoviridae* families (Varani *et al.*, 2008). Accordingly, particles from the *Siphoviridae* (tailed phage particles), *Podoviridae* (~45nm in size), *Microviridae* (~30nm in size) and *Inoviridae* (filamentous particles) having been observed in *X. fastidiosa* cultures (Chen & Civerolo, 2008). Besides the microscopy evidence, plaque formation experiments were also performed in order to confirm the phage activity (Summer *et al.*, 2010; Ahern *et al.*, 2014). The first study isolated the hybrid temperate podophage Xfas53, which was able to form plaques in Temecula1 strain, being the first report of propagation of phage on *X. fastidiosa* (Summer *et al.*, 2010). However, temperate phages may not be used as biocontrol agents, and thus, the same group worked in isolating lytic phages (Ahern *et al.*, 2014). From plant extracts two siphophages and two podophages were isolated and characterized, all them lytic phages that formed plaques in *X. fastidiosa* and *Xanthomonas* hosts. They were suggested to be candidates for phage therapy to control diseases caused by *X. fastidiosa* (Ahern *et al.*, 2014).

1.5. Bioinformatics tools for comparative genomics

Comparing genomes of a group of organisms is a critical step after sequencing. The pangenome, the core genome and other information such as unique genes in a genome group can be identified based on gene content using bioinformatics tools. In some cases, it is of great interest to look for specific genes with important functions, such as virulence genes or drug resistance determinants (Edwards & Holt, 2013). Comparisons made with bioinformatics tools include the identification of orthologous genes, assuming divergence from a common ancestor after a speciation event, and more likely to retain their functions across organisms, unlike the paralog genes that tend to functionally diverge. For this reason, orthologous genes are key elements in genome

annotation and evolutionary studies (Altenhoff *et al.*, 2016), and algorithms for solving bioinformatics problems are fundamental for orthological inference (Contreras-Moreira & Vinuesa, 2013).

Because the evolutionary history of genes is typically unknown, inferring the orthology of a set of genes is not simple. However, several methods of orthology inference as well as efforts to set standards in these methods have already been addressed (Altenhoff *et al.*, 2016). The methods used in orthology inference can be based in trees such as Ensembl Compara[1], PANTHER (Mi *et al.*, 2013) and PhylomeDB (Huerta-Cepas *et al.*, 2013); or based in graphs, constructed from local alignments. Table 1 shows some of the main bioinformatics tools that use graph-based methods. Other methods are based on both tree and graph, for example MetaPhOrs (Tatusov *et al.*, 1997; Altenhoff *et al.*, 2016). However, most bioinformatics tools use graph-based methods for orthology inference. In this clustering method, the sequences of a predicted protein group represent the nodes of the graph (Fig. 1). For this group, all similarities of paired sequences are computed, and the derived values are used to represent the weights of the graph edges (Fig. 1). Some methods apply a clustering step on pair similarities to retrieve groups of orthologs of various species. The sensitivity and specificity of the estimation are highly dependent on the chosen sequence comparison algorithm and the scoring scheme (Setubal *et al.*, 2018). Commonly, BLAST (Altschul *et al.*, 1990) is a popular heuristic method used by various programs to align the sequences (Contreras-Moreira & Vinuesa, 2013; van der Veen *et al.*, 2014; Setubal *et al.*, 2018). The chosen clustering techniques and orthology criteria such as similarity cutoffs or overlap criteria, highly impact the orthology assignment and consequently the pangenome structure (Setubal *et al.*, 2018).

ROARY is one of the newly developed bioinformatics tools for the analysis of genomes of phylogenetically close organisms such as bacterial strains genomes within the same species (Page *et al.*, 2015; Schmid *et al.*, 2018). This tool performs large-scale analyzes of prokaryotic genomes quickly and without compromising the accuracy of the results. This is possible due to runtime segmentation in BLAST comparisons. In the ROARY pipeline, the initial set of sequences is reduced by a pre-clustering step performed by the CD-HIT program.

[1] http://www.ensembl.org/info/genome/compara/homology_method.html

The protein sequences of this substantially reduced set are compared using all-against-all BLASTP and subsequently the sequences are clustered using the Markov Cluster (MCL) algorithm (Enright *et al.*, 2002b; Page *et al.*, 2015).

Another software used in comparative genome is Get_Homologues. This software can cluster homologous gene families using the bidirectional best-hit, COGtriangles, or OrthoMCL clustering algorithms. Clustering stringency can adjust by scanning the domain composition of proteins using the HMMER3 package, by imposing desired pairwise alignment coverage cutoffs, or by selecting only syntenic genes (Contreras-Moreira & Vinuesa, 2013).

Most of the algorithms described above are based on local decisions between neighboring sequences or close groups. However, none of these approaches use of a global metric that considers all sequences in decision making. A graph-based algorithm that uses the clustering coefficient which detects homology in coding sequences obtained from genomes of phylogenetically close organisms was recently been published (Santiago *et al.*, 2018). The clustering coefficient is a topological metric of the graph for which nodes in a graph tend to cluster close together. The density of the components of each graph will be its clustering coefficient, the closest = 1, while components with few edges tend to have their coefficients close to 0. An important feature of the clustering coefficient is that it is not impacted by the isolated components, and if each component has all its vertices fully connected to each other, then the graph clustering coefficient will be = 1 (Watts & Strogatz, 1998). The algorithm that was developed by Santiago and collaborators uses the mean of the clustering coefficient, where the sum of the coefficients of each vertex is divided by the number of vertices. The higher the value of this coefficient, the denser the components of the graph and, therefore, its maximization produces graphs according to the expected characteristics of a homology (for instance, it is expected that the graph formed by the homology relationships is more homogeneous and denser) (Santiago *et al.*, 2018). The recent released tool GTACG framework uses this algorithm (Santiago *et al.*, 2019).

Table 1. Main bioinformatics tools and graph-based algorithms available for comparative genome analysis.

Bioinformatics Tool	Clusterization algorithm to orthology inference	Reference
EDGAR (Efficient Database framework for comparative Genome Analysis)	BLAST score Ratios (BSR)	(Blom *et al.*, 2009; Yu *et al.*, 2017)
PGAP (Pan-Genome Analysis Pipeline)	Multiparanoid (MP) e Gene Family (GF) based in MCL	(Zhao *et al.*, 2018)
GET HOMOLOGUES	OrthoMCL, COG triangle and one version of BHH-BLAST	(Contreras-Moreira & Vinuesa, 2013)
PanFunPro	CD-HIT e HMM	(Lukjancenko *et al.*, 2013)
BPGA (Bacterial Pan Genome Analysis)	USERCH, CD-HIT e OrthoMCL.	(Chaudhari *et al.*, 2016)
Roary	CD-HIT, BLAST all-against-all e MCL	(Page *et al.*, 2015)

[Tabela adaptada de: (Zekic *et al.*, 2018)]

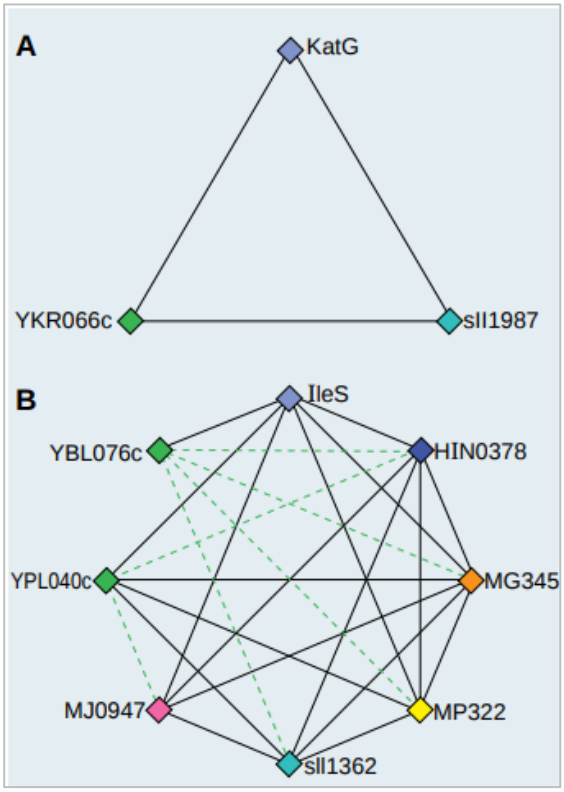

Fig. 1: Orthology inference based on graph. The nodes of the graph represent the predicted protein sequences and the edges Best Hit between them. (A) Graph formed by symmetrical edges (solid lines). (B) Graph formed by asymmetrical edges (broken lines) [Figure from: (Tatusov *et al.*, 1997)].

Mobile genetic elements (MGEs) are an important feature of prokaryote genomes but are rarely well annotated and, consequently, are often underestimated. MGEs include transposons, insertion sequences (ISs), prophages, genomic islands (GIs), integrons, and integrative and conjugative elements (ICEs) (Oliveira Alvarenga *et al.*, 2018). They are involved in genome evolution and promote phenomena such as genomic expansion and rearrangement, emergence of virulence and pathogenicity, and symbiosis. Several programs and databases dedicated to prokaryotic MGE analysis and annotation have been developed as recently reviewed (Oliveira Alvarenga *et al.*, 2018). Below follows a brief description of such programs.

For the prophage prediction, software such as PHASTER, Phispy, Virsorter have been developed. PHASTER, an upgrade of PHAST software identifies and annotates prophages in bacterial and plasmids. This software combines genome-scale ORF prediction and translation (via GLIMMER), protein identification (via BLAST), phage sequence identification (via BLAST matching to a phage-specific sequence database), tRNA identification, attachment site recognition and gene clustering density measurements using density-based spatial clustering of applications with noise (DBSCAN) and sequence annotation text mining (Arndt *et al.*, 2016). PhiSpy uses five distinctive characteristics: protein length, transcription strand directionality, customized AT and GC skew, and the abundance of unique phage DNA sequence words. Among the metrics used, random forest classification algorithm predicts the prophages by ranking genomic regions based on those characteristics (Akhter *et al.*, 2012). Unlike those mentioned, Virsorter detects viral signal in different types of microbial sequence data (such as metagenomic sequencing) in both a reference-dependent and reference-independent manner, leveraging probabilistic models and extensive virome data to maximize detection of novel virus (Roux *et al.*, 2015). Due to revelation of cryptic inoviruses as pervasive in bacteria and archaea across Earth's biomes (Roux et al., 2019), an Inovirus_detector script to detect inovirus sequences have been developed. This script can be used to identify putative inovirus sequences in draft genome assemblies or metagenome assemblies (Roux *et al.*, 2019).

Alien_Hunter, SeqWord Sniffer and GIPSy are some of the software used for the genomic island prediction and analysis. The first one, is an application that

implements Interpolated Variable Order Motifs (IVOMs), an approach that exploits compositional biases using variable order motif distributions and captures more reliably the local composition of a sequence compared to fixed-order methods. Optionally the predictions can be parsed into a 2-state 2nd order Hidden Markov Model (HMM), in a change-point detection framework, to optimize the localization of the boundaries of the predicted regions (Vernikos & Parkhill, 2006). SeqWord Sniffer examines variations in frequencies of oligonucleotides and traces down the distributions of mobile genomic elements across genomes by analyzing the patterns of 4-bases long words (Bezuidt *et al.*, 2009). Finally, GIPSy, developed in Java, is based on the commonly shared features: genomic signature deviation (G+C content and codon usage); presence of transposase genes; factors for virulence, metabolism, antibiotic resistance, or symbiosis; flanking tRNA genes; and absence in other organisms of the same genus or closely related species (Soares *et al.*, 2016).

Other type of MGEs are the Insertion Sequences (ISs) that may occur once in a genome or may consist of a set of almost identical copies. ISs in a particular genome can therefore be classified into two major groups in terms of copy number: multi-copy and single-copy ISs, this is the approach considered by OASIS (Robinson *et al.*, 2012). First, this software groups already-annotated transposase genes that could compose multiple copies of an IS by length and sequence similarity. Those that fit a high similarity threshold are assumed to be in the same group (multi-copy ISs). In the case of isolated transposase in the genome, inverted repeat sequences around the transposases are found and composed single-copy ISs. Multi-copy ISs are also checked for inverted repeats sequences. Once all groups of ISs are identified, hierarchal agglomerative clustering is used to defined unique IS sets. Finally, the classification of IS families is performed using BLASTp against ISFinder database (Robinson *et al.*, 2012). The ISEScan, other IS predicter use a pipeline that consists of the following steps: (i) predicting protein coding genes in the input genome sequence and translating them into protein sequences, (ii) identifying putative transposases by searching the predicted proteins against the library of profile HMMs (based on the transposase sequences in the curated ACLAME dataset), (iii) extending the putative transposase genes into full-length IS elements by locating inverted repeat sequences in the upstream and downstream regions and (iv) refining and

reporting the final set of annotated IS elements in the input genome (Xie & Tang, 2017).

2. Motivation and Objectives

Currently, the RefSeq database of the NCBI provides complete and draft genomes of several *X. fastidiosa* strains from many countries and hosts. In addition, genomes from a few additional *X. fastidiosa* isolates have been sequenced by our research group at IQ-USP and they have not been published. The genomes of the different *X. fastidiosa* isolates present remarkable similarities showing Average Nucleotide Identity (ANI)> 95%, regardless the plant host from which they were isolated. Still, the determinants of *X. fastidiosa* host adaptation remains unclear at the genomic level. Hence the main objective of this study was to investigate the potential determinants of plant host adaptation and other traits of interest, such as the heterogeneity among mobile genetic elements in the available sequenced genomes of *X. fastidiosa*, using comparative genomics and phylogenomics bioinformatics tools.

For this, the following specific objectives were established:

- Retrieve the genomic sequences of 46 strain of *X. fastidiosa* available in the NCBI-*Assembly* database and re-annotate these sequences using PROKKA;
- Compare the 46 genomes of *X. fastidiosa* using a recent developed tool (GTACG framework), a graph-based algorithm that uses the clustering coefficient which detects homology in coding sequences obtained from genomes of phylogenetically close organisms (Santiago *et al.*, 2018; Santiago *et al.*, 2019);
- Perform two phylogenomic trees (Maximum Likelihood Tree and Consensus Tree) based on the information shared by all strains;
- Perform a functional analysis using COG database;
- Verify the association between the phylogenomic analysis and metadata such as geographic region and plant host of isolation as well as strain Sequence Type;
- Predict the sequences of prophages, genomic islands and insertion sequences integrated in the genome of strains of *X. fastidiosa* and hypothesize their potential impact in the genome of these strains.

3. Methodology

The workflow of the analyzes described below is shown in Fig. 2 and the main software used are briefly described in Supplementary information (Table S1).

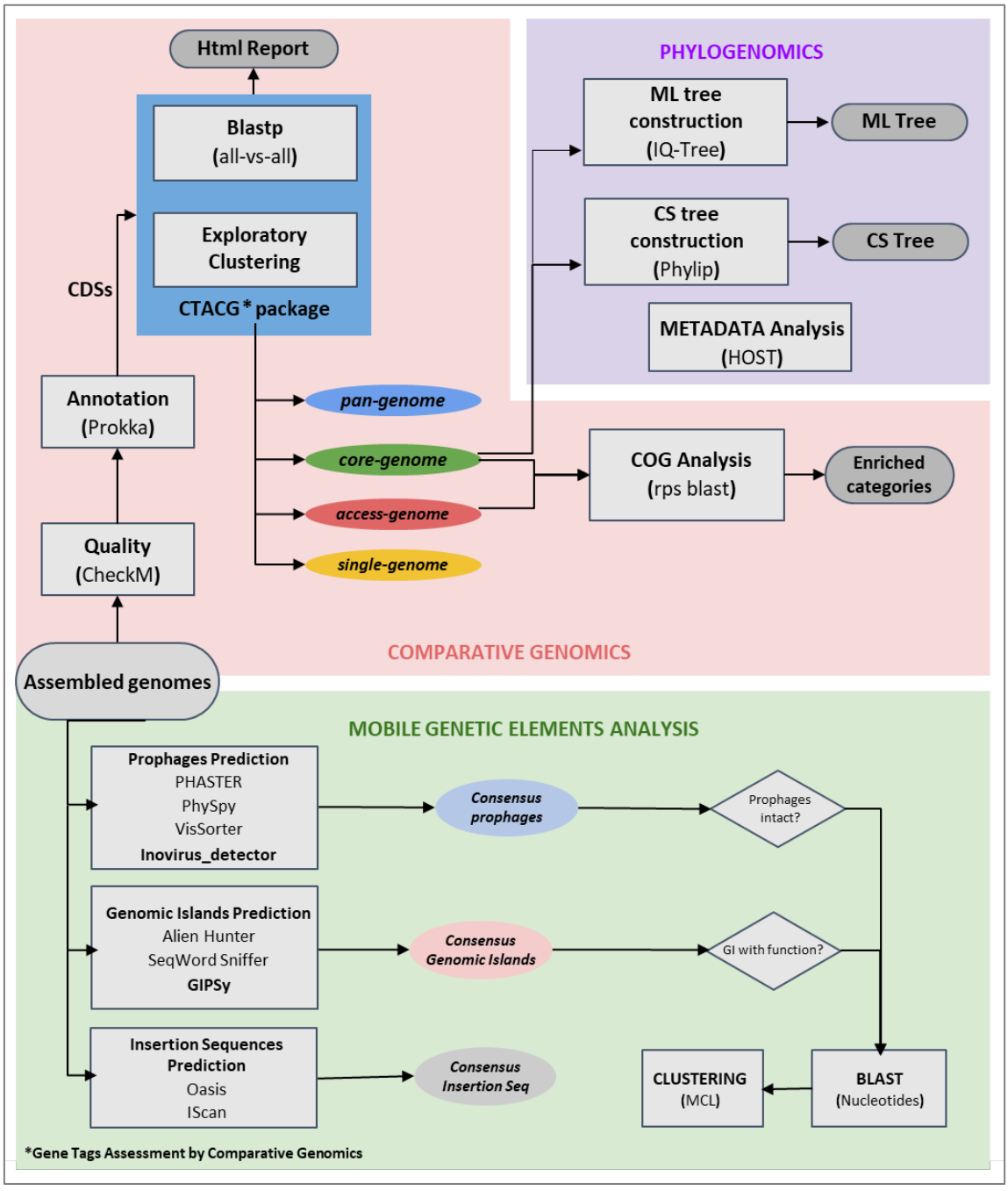

Fig. 2. General pipeline of the methodologies used in this study.

3.1. Quality and Annotation

A total of 46 genomes of *X. fastidiosa* strains were included in this study. Forty-four genomes available until December 2018 were retrieved from the public NCBI RefSeq database[2], while the other two genomes sequenced in our lab were not publicly available by that time. Some features, source and reference of these genomes are presented in the Table 2. For the 16 complete and 30 draft genomes, the contamination and completeness of the genomes were estimated using CheckM package (Parks *et al.*, 2015) with the default parameters. This package assesses the quality of a genome identifying and counting universal single copy genes (SCGs). A list of single copy genes exist is provide by CheckM, and mainly consist of genes encoding for ribosomal proteins and other housekeeping genes. The completeness is the number of unique SCGs present within the genome / the number of unique SCGs in the list. Contamination is estimated by looking at how many SCGs are present in multiple copies, as only one copy should be present of each SCG per genome. This analysis is particularly important for draft genomes.

In order to homogenize the annotation of the all genomes, each genome was re-annotated using the PROKKA v.1.12 software (Seemann, 2014) with the default parameters and a database of the *Xylella* genus which was constructed locally. The PROKKA software uses Prodigal (Hyatt *et al.*, 2010) which is a protein-coding gene prediction software tool for bacterial and archaeal genomes. It provides support for three modes of prediction, normal, anonymous and training mode. To improve the annotation of some protein-coding gene such hemagglutinins (large proteins with repeated motifs), PROKKA was customized with both normal and training mode of Prodigal.

3.1. Comparative genomics

The amino acid and nucleotide coding sequences (CDSs) of each annotated genome were retrieved. To perform the comparative genomics was used the GTACG framework[3] (Santiago *et al.*, 2019) which is based in an algorithm recently published (Santiago *et al.*, 2018). Firstly, a local alignment was

[2] https://www.ncbi.nlm.nih.gov/genome/
[3] https://github.com/caiorns/GTACG-backend

performed using BLAST (Altschul *et al.*, 1990) with an e-value of 1e-10. Secondly, ExploratoryClustering tool was used to cluster and find a threshold that maximizes the cluster coefficient of each cluster. In previous experiments it was realized that with an initial threshold of 45% of the alignment length was enough to produce concise homologous clusters. Later, the result of the clustering of the sequences was used to create a relationship graph among the CDS, to this step ExportGraph tool was used. The next step produced multiple alignments and phylogenies of each homologous families and orthologs using ExportTreeFiles. Finally, the data was exported in html format which was used to visualize and analyze the results for the 46 *X. fastidiosa* genomes. In order to figure out association between strains and specific traits in *X. fastidiosa*, metadata information about host, country of origin and Sequence Type (defined by MLST) was provided to the framework.

Based on the number of the orthologs groups found, genomics information of the species *X. fastidiosa* was calculated. This information was: the pan-genome, core-genome, accessory-genome and singleton-genome.

3.2. Phylogenomic analysis

Two phylogenomic trees were built to determine the relatedness of *X. fastidiosa* strains analyzed. These trees, Maximum Likelihood (ML) and Consensus (CS), were based on the orthologs belong to the core-genome finding in the comparative genomics section.

The maximum likelihood method uses standard statistical techniques for inferring probability distribution to assign probabilities to particular phylogenetic trees. To the ML tree, the sequences of nucleotides of the each ortholog were aligned with Clustal Omega 1.2.1 (Sievers *et al.*, 2011) using default parameters. Then the alignments were concatenated and computed using IQ-TREE v.1.5.4 (Nguyen *et al.*, 2015). This software uses a fast and effective stochastic algorithm to reconstruct phylogenetic trees by ML. The model of the ML tree was predicted by the automatic model selection ModelFinder, and an ultrafast bootstrap approximation (UFBoot) of 1,000 replicates was performed to assess branch supports (Minh *et al.*, 2013).

A consensus tree provides an estimate for the level of support for each clade in the final tree. It is built by combining clades which occurred in at least a certain

percentage of the resampled trees. A 100% support threshold result in a strict consensus tree which is a tree where the included clades are those that are present in all the trees of the original set. A 50% threshold result in a Majority rule consensus tree that includes only those clades that are present in the majority of the trees in the original set. A threshold less than 50% gives rise to a Greedy consensus tree [4]. To make the CS tree, the phylogenetic trees of each ortholog were retrieved, and then it was constructed using the consensus tool of Phylip package version 3.69 [4]; the method used was Majority rule consensus.

3.3. Functional analysis

In order to identify whether orthologous clusters of certain functional groups are preferentially found in all strains or if they diverge between the strains, the COG (Cluster of Orthologous Groups) categories (Galperin *et al.*, 2015) were determined for each gene belonging the core and accessory genomes using an e-value of 10^{-6}.

For the case of hypothetical proteins, InterProScan software was used to scan these proteins against InterPro signatures (Mitchell *et al.*, 2018). To verify if some proteins have the potential to be secreted, it was used SecretomeP software (Bendtsen *et al.*, 2005).

3.4. Prediction of prophages, genomic island and insertion sequences

The Mobile Genetic Elements (MGEs) harbored in the chromosomal genome of the 46 *X. fastidiosa* strains, such as prophages, genomic islands and insertion sequences were identified with a pipeline based in a practical guide previously described (Oliveira Alvarenga *et al.*, 2018). This pipeline combines the different prediction software results, which were manually curated, to define consensus regions of prophage, genomic island and insertion sequence in all strains studied.

The individual outputs of each software were merged and compared to define consensus MGE regions. In the case of overlapping, the size of the consensus region was maximized. The presence of prophage in the genome of

[4] http://www0.nih.go.jp/~jun/research/phylip/main.html

X. fastidiosa strains was prioritized. Thus, the consensus regions pointed as prophages were firstly defined using the results of PHASTER (Arndt *et al.*, 2016), VirSorter (Roux *et al.*, 2015) and PhiSpy (Akhter *et al.*, 2012) software. . Inovirus_detector software[5] was used to the classification of consensus prophages as member of the inoviridae family (Roux *et al.*, 2019). The prophages consensus regions were manually evaluated to the presence of at least one integrase ORF.

For the genomic islands, their consensus regions were calculated using the results of Alien_Hunter (Vernikos & Parkhill, 2006) and SeqWord Sniffer (Bezuidt *et al.*, 2009) software. Later these regions without overlapping with prophages consensus were considered as a genomic island consensus. GIPSy software (Soares *et al.*, 2016) was used to assign one or more categories related to the potential function of the genomic island. Finally, to the case of Insertion sequences, the consensus regions were calculated using the software Oasis (Robinson *et al.*, 2012) and ISEScan (Xie & Tang, 2017). For each strain, the consensus regions of the MGE predicted was mapped in the genome to the visualization.

Both the prophages and genomic island consensus were compared to explore homology relationships using Blast all-vs-all with the nucleotide sequences. The hits of Blast results with an identity and coverage alignment higher than 90% and 80%, respectively, were filtered and grouped using the software MCL version 14-137. The algorithm MCL (Markov Cluster) represents a process that captures the concept of random walks in a graph and does it deterministically (Enright *et al.*, 2002a). The weight used was the -log(e-value) of the hits filtered, and the inflation value considered was 10. Finally, the MAUVE software (Darling *et al.*, 2004) was used to visualize the sequences alignment of each group found by MCL. Manual inspection was performed to each group.

[5] https://github.com/simroux/Inovirus

4. Results

4.1. Comparative Genomics Analyses

4.1.1. General features and annotation of *X. fastidiosa* genomes

Table 2 shows the main features of the 46 genomes of *Xylella fastidiosa* strains/isolates included in this study such as genome size, percentage of GC, number of plasmids reported, access to the assembly and reference. Also, metadata information regarding the respective isolate plant host, geographic region of isolation and predicted Sequence Type is presented. Furthermore, genome assembly statistics is provided in Supplementary information (Table S2). Forty-four genomes were retrieved from the public NCBI-assembly database until December 2018. The other two genomes (strains XRB and B111) were sequenced using Illumina MiSeq platform (Ilumina Inc) and assembled using the software CLC Genomics Workbench v.6.5. The assemblies of these genomes yielded incomplete genome with 74 and 77 contigs, respectively (Martins-Jr, J. & Pierry, P.M., unpublished results). However, both genomes have more than 95% of completeness and less than 5% of contamination, which could classify them as being near complete.

The average size of the 46 genomes was 2.6 Mbp ± 0.12 Mbp (standard deviation) (Fig. 3), with the South American strain Hib4 being the largest reported genome so far (2,877,548 bp, chromosome and plasmid) and the North American strain Griffin-1 the smaller with a size of 2,378,314 bp. The percentage of GC varied between 51.34% and 52.91% for the strains DSM10026 and J1a12, respectively. Sixteen strains which have complete assembled genomes harbor at least one plasmid except for the isolates 3124 and M12 isolate from coffee and almond, respectively (Table 2). The plasmids belonging to Hib4 (hibiscus) has 64,251 bp, being the largest *X. fastidiosa* plasmid reported to date. On the other hand, in the draft genomes of CVC0256 and CVC0251 (citrus), COF324 and COF0407 (coffee) and OLS0479 (oleander) up to 4 plasmids were found, with sizes between 1.3 Kbp and 51.2 Kbp. The first three strains (ST 14) were from Brazil and the last two (ST 53) from Costa Rica.

Table 2 General genome features of *X. fastidiosa* strains used in this study.

Strain	Genome Size	%GC	Plasmid	Assembly[b]	Host	Country	MLST	Reference
3124	2.748.594	52,63	0	GCA_001456195.1	Coffea	Brazil	ST_16	not published
9a5c	2.731.750	52,67	2	GCA_000006725.1	Citrus	Brazil	ST_13	(Simpson *et al.*, 2000)
Ann-1	2.780.908	52,1	1	GCA_000698805.1	Nerium	USA	ST_05	(Bhattacharyya *et al.*, 2002a)
De-Donno	2.543.738	51,98	1	GCA_002117875.1	Olea	Italy	ST_53	(Giampetruzzi *et al.*, 2017)
Fb7	2.699.320	52,54	1	GCA_001456335.3	Citrus	Argentina	ST_69	not published
GB514	2.517.383	51,78	1	GCA_000148405.1	Vitis	USA	ST_01	(Schreiber *et al.*, 2010)
Hib4	2.877.548	52,7	1	GCA_001456315.1	Hibiscus	Brazil	ST_70	not published
J1a12	2.867.237	52,91	2	GCA_001456235.1	Citrus	Brazil	ST_11	not published
M12	2.475.130	51,92	0	GCA_000019325.1	Prunus	USA	ST_07	(Chen *et al.*, 2010)
M23	2.573.987	51,76	1	GCA_000019765.1	Prunus	USA	ST_01	(Chen *et al.*, 2010)
MUL0034	2.666.577	51,97	1	GCA_000698825.1	Morus	USA	ST_30	not published
Pr8x	2.705.822	52,63	1	GCA_001456295.1	Prunus	Brazil	ST_14	not published
Salento-1	2.543.366	51,98	1	GCA_002954185.1	Olea	Italy	ST_53	not published
Salento-2	2.543.566	51,98	1	GCA_002954205.1	Olea	Italy	ST_53	not published
Temecula1	2.521.148	51,78	1	GCA_000007245.1	Vitis	USA	ST_01	(Van Sluys *et al.*, 2003)
U24D	2.732.490	52,68	1	GCA_001456275.1	Citrus	Brazil	ST_13	not published
32	2.607.546	52,45	0	GCA_000506405.1	Coffea	Brazil	ST_16	(Alencar *et al.*, 2014)
11399	2.736.060	52,74	1	GCA_001684415.1	Citrus	Brazil	ST_11	not published
6c	2.603.975	52,42	1	GCA_000506905.2	Coffea	Brazil	ST_14	(Alencar *et al.*, 2014)
ATCC35871	2.416.255	51,61	0	GCA_000428665.1	Prunus	USA	ST_41	not published
ATCC35879	2.522.328	51,79	0	GCA_000767565.1	Vitis	USA	ST_02	not published
B111[a]	2.682.276	52,49	2	Not access	Citrus	Brazil	ST_11	not published
BB01	2.511.521	51,76	1	GCA_001886315.1	Vaccinium	USA	ST_42	(Van Horn *et al.*, 2017)
CFBP8072	2.496.662	51,94	0	GCA_001469345.1	Coffea	Ecuador	ST_74	(Jacques *et al.*, 2016)
CFBP8073	2.582.150	51,55	0	GCA_001469395.1	Coffea	Mexico	ST_75	(Jacques *et al.*, 2016)
CFBP8416	2.466.748	51,79	0	GCA_001971475.1	Polygala	France	ST_07	not published
CFBP8417	2.504.981	51,85	0	GCA_001971505.1	Spartium	France	ST_06	not published
CFBP8418	2.513.969	51,88	0	GCA_001971465.1	Spartium	France	ST_06	not published
CO33	2.681.926	51,7	0	GCA_001417925.1	Coffea	Costa Rica	ST_72	(Giampetruzzi *et al.*, 2015b)
CoDiRO	2.542.932	52,03	1	GCA_000811965.1	Olea	Italy	ST_53	(Giampetruzzi *et al.*, 2015a)
COF0324	2.772.556	52,45	4	GCA_001549815.1	Coffea	Brazil	ST_14	not published
COF0407	2.538.474	51,93	4	GCA_001549825.1	Coffea	Costa Rica	ST_53	not published
CVC0251	2.740.246	52,57	4	GCA_001549765.1	Citrus	Brazil	ST_11	not published
CVC0256	2.702.144	52,6	4	GCA_001549745.1	Citrus	Brazil	ST_11	not published
Dixon	2.622.328	52,03	0	GCA_000166835.1	Prunus	USA	ST_06	(Bhattacharyya *et al.*, 2002a)
DSM10026	2.431.652	51,34	0	GCA_900129695.1	Unkonwn	USA	ST_02	not published
EB92-1	2.475.426	51,48	0	GCA_000219235.2	Sambucus	USA	ST_01	(Zhang *et al.*, 2011)
Griffin-1	2.387.314	51,73	0	GCA_000466025.1	Quercus	USA	ST_07	(Chen *et al.*, 2013)
IVIA5235	2.491.574	51,57	1	GCA_003515915.1	Prunus	Spain	ST_01	
Mullberry	2.667.719	51,95	0	IMG[c]	Morus	USA	ST_30	(Guan *et al.*, 2014b)
Mul-MD	2.520.555	51,64	0	GCA_000567985.1	Morus	USA	ST_29	not published
OLS0478	2.555.411	51,95	2	GCA_001549755.1	Nerium	Costa Rica	ST_53	not published
OLS0479	2.539.963	51,96	4	GCA_001549735.1	Nerium	Costa Rica	ST_53	not published
Stags-Leap	2.510.798	51,72	0	GCA_001572105.1	Vitis	USA	ST_01	(Chen *et al.*, 2016)
Sy-VA	2.475.880	51,64	0	GCA_000732705.1	Platanus	USA	ST_08	(Guan *et al.*, 2014a)
XRB[a]	2.714.643	52,27	1	Not access	Citrus	Brazil	ST_11	not published

[a] Strains reporter in this study, [b] NCBI acces, [c] Integrate Microbial Genomes & Microbione, [d] Multi Locus Sequence Typing

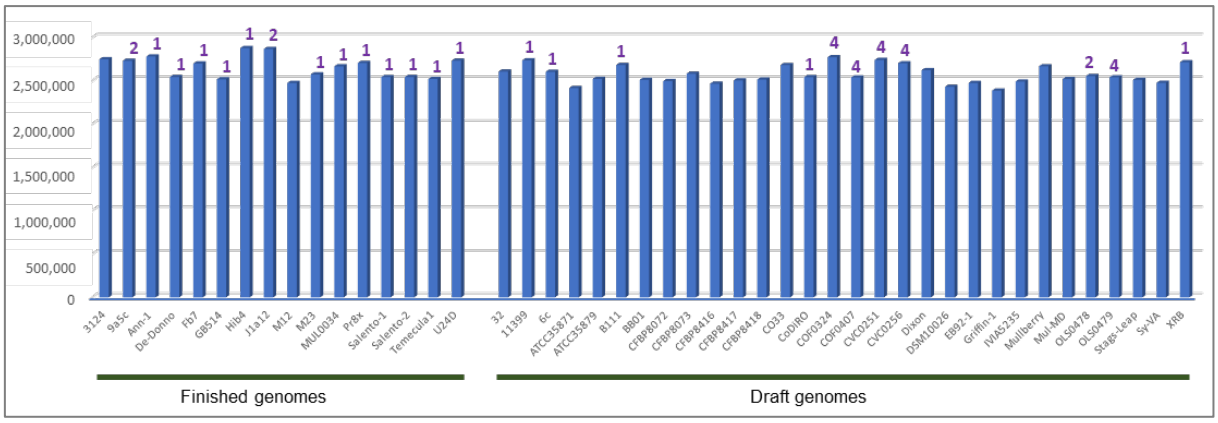

Fig. 3. Total size of *X. fastidiosa* genomes. Above each bar is indicated the number of plasmids by genome.

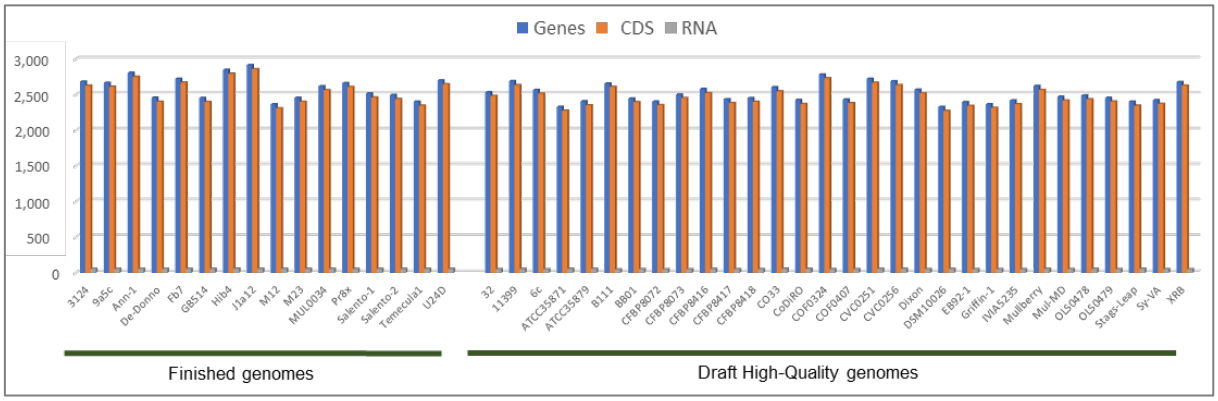

Fig. 4. Re-annotation of *X. fastidiosa* genomes using PROKKA. The CheckM result (Table 3) allowed to classify the contig and scaffold genomes in Draft High-Quality genomes.

Before the re-annotation process, the quality of the 46 genomes was assessed. Table 3 shows the completeness and contamination for the complete and draft genomes according to CheckM software. The draft genomes showed a completeness value greater than 97.42% and a contamination less than 1.09%, thus they were classified as being near complete with low contamination according the CheckM criteria.

The re-annotation with PROKKA software (Table 3 and Fig. 4) provided the localization of gene sequences such as coding sequences (CDS), ribosomal RNA (rRNA), transfer RNA (tRNA) and transfer-messenger RNA (tmRNA). The average of the total of genes in *X. fastidiosa* was 2,544. The CDS varied between 2,857 and 2,271, respectively, for the genomes of J1a12 and DSM10026 strains. The number of the tRNAs was 48 on the average, covering the 20 amino acids. Also, it was observed that all strains with complete (finished) genomes encoded two complete rRNA operons (6 rRNA genes). In the case of draft genomes this number varied from 1 to 8 rRNA genes probably due incorrect genome assemblies.

The standards about minimum information of an assembled genome were defined by Bowers and collaborators (Bowers *et al.*, 2017) as high completeness if >95% and low contamination if <5% as obtained with CheckM. Along with the re-annotation results with PROKKA (all types of rRNA genes, and at least 18 tRNAs), the draft genomes included in this study can be considered as High-Quality Draft (HQD).

Table 3 Re-annotation and quality of *X. fastidiosa* genomes

Strain	CDSs	CDSs per core	tRNAs	rRNA (5S|16S|23S)	Completeness*	Contamin.*	Status
3124	2.626	832	49	6 (2|2|2)	99,64	0	F
9a5c	2.610	827	48	6 (2|2|2)	99,59	0	F
Ann-1	2.750	872	49	6 (2|2|2)	99,64	0	F
De-Donno	2.400	761	49	6 (2|2|2)	99,59	0	F
Fb7	2.668	846	48	6 (2|2|2)	99,28	0	F
GB514	2.397	760	49	6 (2|2|2)	99,18	0	F
Hib4	2.792	885	50	6 (2|2|2)	99,64	1,09	F
J1a12	2.857	906	49	6 (2|2|2)	99,59	0	F
M12	2.308	732	49	6 (2|2|2)	99,64	0	F
M23	2.399	760	49	6 (2|2|2)	99,64	0	F
MUL0034	2.561	812	49	6 (2|2|2)	99,64	0	F
Pr8x	2.606	826	48	6 (2|2|2)	99,59	0	F
Salento-1	2.461	780	49	6 (2|2|2)	99,11	0	F
Salento-2	2.438	773	49	6 (2|2|2)	99,59	0	F
Temecula1	2.346	744	49	6 (2|2|2)	99,64	0	F
U24D	2.645	838	48	6 (2|2|2)	99,59	0	F
32	2.483	787	47	3 (1|1|1)	99,59	0	HQD
11399	2.633	835	49	6 (2|2|2)	99,28	0	HQD
6c	2.515	797	46	3 (1|1|1)	99,64	0	HQD
ATCC35871	2.274	721	47	7 (3|1|3)	99,64	0	HQD
ATCC35879	2.350	745	49	6 (2|2|2)	99,64	0	HQD
B111	2.607	826	44	3 (1|1|1)	99,64	0	HQD
BB01	2.393	759	48	3 (1|1|1)	99,64	0	HQD
CFBP8072	2.354	746	45	3 (1|1|1)	99,64	0	HQD
CFBP8073	2.451	777	46	3 (1|1|1)	99,64	0	HQD
CFBP8416	2.522	799	49	6 (2|2|2)	98,46	0	HQD
CFBP8417	2.383	755	47	3 (1|1|1)	99,28	0	HQD
CFBP8418	2.400	761	47	3 (1|1|1)	99,64	0	HQD
CO33	2.546	807	49	8 (2|2|4)	99,63	0	HQD
CoDiRO	2.369	751	49	6 (2|2|2)	99,64	0	HQD
COF0324	2.730	865	46	5 (1|3|1)	98,31	0,02	HQD
COF0407	2.380	754	47	3 (1|1|1)	99,64	0	HQD
CVC0251	2.666	845	49	6 (2|2|2)	99,63	0	HQD
CVC0256	2.633	835	48	6 (2|2|2)	99,64	0	HQD
Dixon	2.519	799	47	3 (1|1|1)	99,64	0	HQD
DSM10026	2.271	720	49	6 (2|2|2)	99,23	0	HQD
EB92-1	2.342	742	47	4 (1|1|2)	98,91	0	HQD
Griffin-1	2.314	734	47	3 (1|1|1)	97,42	0,18	HQD
IVIA5235	2.367	750	47	3 (1|1|1)	99,64	0,18	HQD
Mullberry	2.565	813	49	6 (2|2|2)	99,64	0	HQD
Mul-MD	2.417	766	47	4 (1|2|1)	99,64	0	HQD
OLS0478	2.433	771	49	6 (2|2|2)	98,91	0	HQD
OLS0479	2.403	762	47	3 (1|1|1)	99,23	0,36	HQD
Stags-Leap	2.345	743	49	6 (2|2|2)	99,23	0	HQD
Sy-VA	2.371	752	47	3 (1|1|1)	99,28	0	HQD
XRB	2.626	832	46	3 (1|1|1)	99,64	0,09	HQD

F: Finisheg genome, HDQ: High Draft Quality, * CheckM results

4.1.2. Comparative genomics of *X. fastidiosa*

As stated in the literature (Zekic *et al.*, 2018), in the present study the pan-genome was defined as the total number of orthologous CDSs (clusters of homologous CDSs) of the 46 genomes analyzed with a clustering algorithm. The

core-genome was defined as the conserved orthologous CDSs present in all genomes. This set of CDSs provides the genomic information to the phylogenomic analysis and would contain the representative orthologs of *X. fastidiosa* species. Regarding the accessory-genome, it was represented as the variable orthologous CDSs among 2 and 45 genomes and could contain CDSs related do supplementary biological process. Finally, the singleton-genome was defined as the CDSs present only in one genome of *X. fastidiosa*.

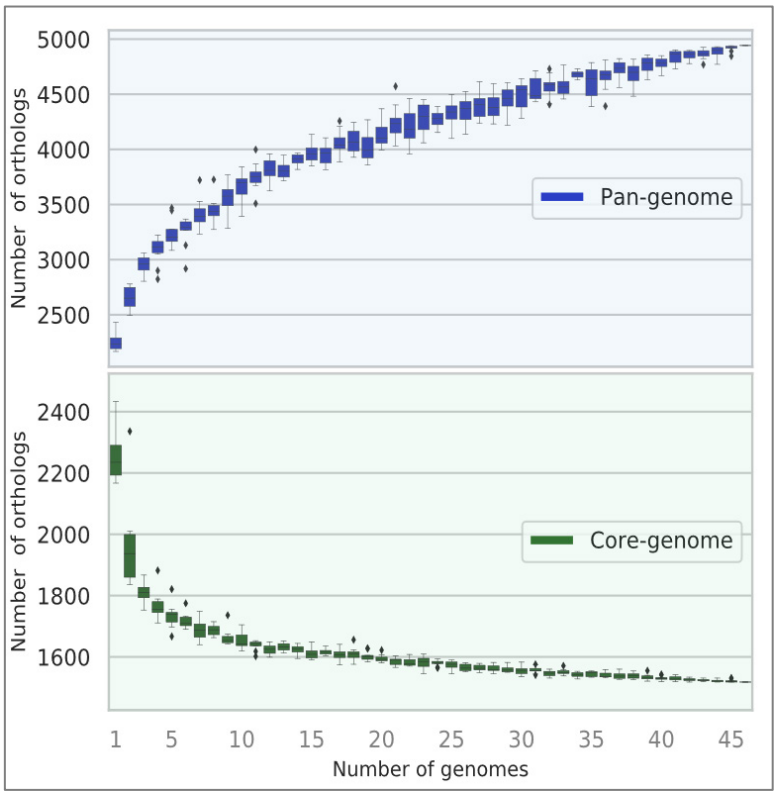

Fig. 5. Pan-genome and core-genome curves. Each single boxplot represents the distribution of the number of orthologs added (pan-genome) or in common (core-genome) with the addition of new genomes.

The pan-genome curve is shown in Fig. 5 with blue boxplots, where it is possible to observe that its size increases with the addition of each new genomes, reaching a total of 4,942 orthologous clusters. However, the curve has not reached a plateau, suggesting that the pan-genome of *X. fastidiosa* species should be considered open. In contrast, the core-genome curve, showed in green box plots (Fig. 5) reached 1,518 orthologous clusters that could represented the total of orthologous clusters sharing among the analyzed *X. fastidiosa* genomes. The percentages of the core-genome and singleton-genome was represented in a plot of frequency of the orthologs within the 46 genomes (Fig. 6), these

percentages were 31.7% (1,518 orthologs clusters) and 23.8% (1,175 orthologs) of the whole CDS pool, respectively. Also, a matrix plot of the presence (green) and absence of the orthologous clusters in the core, accessory and singleton genomes is showed in Fig. 7. It shows the distribution of the CDSs among *X. fastidiosa* strains which the accessory and singleton genome regions (indicated by arrows) constitute the orthologs possibly gained by horizontal gene transfer.

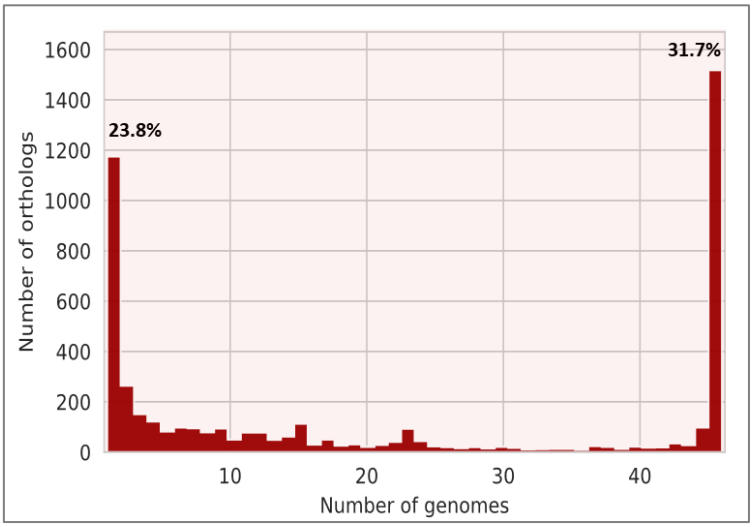

Fig. 6. Number of orthologous CDSs within all genomes. At the extreme left of the x-axis is represented the singleton-genome which is constituted by strain exclusive genes (23.8 % of the pan-genome). At the opposite end, the core-genome is represented (31.7% of the pan-genome).

The CDSs present in the singleton-genome are associated with the unique or exclusive CDS in each strain. Overall, the great majority of exclusive CDS that was found in *X. fastidiosa* has an atypical content of GC. As it was expected, more than half (61.5%) of these exclusive CDSs does not have a defined function, showing the relative great number of unknown genes gained by these strains. Interestingly, an analysis using InterProScan software for these hypothetical proteins showed that almost 300 would have typical signatures for a membrane-associated protein, either to be embedded in the membrane or facing the intra or extracellular milieu. Also, the presence of a signal peptide was predicted in some unique hypothetical proteins, suggesting they may be secreted through the membranes. The exclusive proteins with known functional annotation were related to prophages proteins (integrase, capsid and tail proteins), conjugal transfer protein, integrating conjugative element protein and alpha/beta hydrolases.

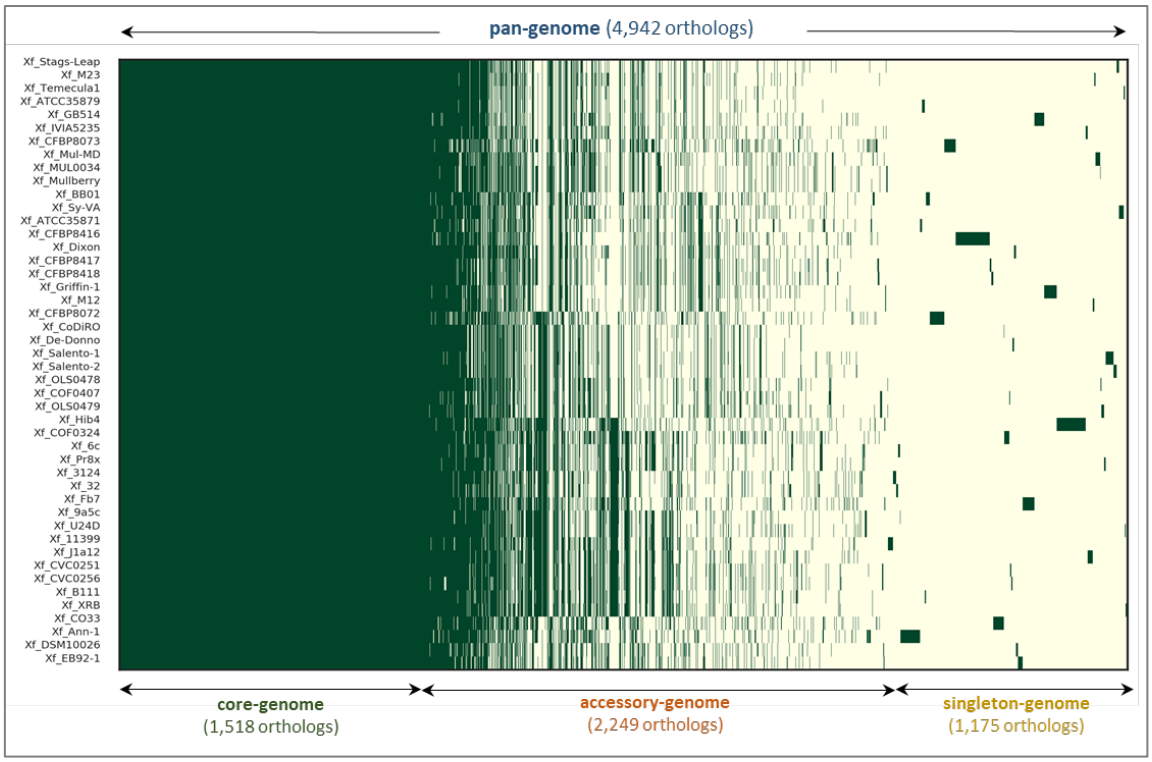

Fig 7. Matrix plot of the CDSs distributed in *X. fastidiosa* genomes. Presence (green bars) and absence of orthologous CDSs in the core-genome, accessory-genome and singleton-genome.

The mean number of exclusive CDSs in each genome was 25.6 with a standard deviation of 34.6. This large deviation from the mean indicates a high degree of genome diversity in the *X. fastidiosa* species. The genomes of the strains CFBP8416 and Hib4 contained the two greatest numbers of exclusive CDSs, 167 and 133, respectively. Both strains have ornamental plants as hosts from the genus *Poligala* and *Hibiscus*, respectively (Table 2). In the case of Hib4, 50 exclusive orthologs belong to a plasmid of 64.2 kbp, representing 75% of its CDSs. A recent study has pointed for this exclusiveness of this plasmid in *X. fastidiosa* strains (Denance *et al.*, 2019).

In order to explore the exclusivity of some orthologous groups and the metadata, it was considered the categories: plant host, country of origin and Sequence Type (ST, defined by MLST analysis). Each category is represented by groups, which are integrated by at least three strains (Table 4). As well as in the exclusive orthologs of each strain, the majority of orthologous found in these groups has unknown function. In the host category, only in *Citrus*, *Olea* and *Morus* groups we have found exclusive orthologs. In the first group, two exclusive orthologs are located in a plasmid; one of them is an ortholog with an alpha/beta

hydrolase fold and predicted to be secreted. The other exclusive protein is a hypothetical protein located downstream of an XRE family transcriptional regulator. The second group of host category (*Olea*) showed two exclusive orthologous genes, both coding for proteins with hypothetical function. In the case of the *Morus* group, an integrase, a topoisomerase and a DNA cytosine methyltransferase were found in all three strains isolated from this host, as well as other 14 proteins with hypothetical function.

Table 4 Analysis of the orthologous genes based in the metadata

Metadata*	Num. of orthologs	Annotation
Host category**		
Citrus (9)	2	Alpha/beta hydrolase and hypothetical protein.
Coffea (8)	0	
Prunus (5)	0	
Olea (4)	2	Hypothetical proteins
Vitis (4)	0	
Morus (3)	17	Integrase, topoisomerase, DNA cytosine methyltransferase and 14 HPs
Nerium (3)	0	
Country of origin category**		
USA (17)	0	
Brazil (14)	2	Integrating conjugative element relaxase and hypothetical protein
France (5)	0	
Italy (5)	0	
Costa Rica (3)	7	CopG family transcriptional regulator, type II toxin-antitoxin system RelE/ParE family toxin (plasmid) and hypothetical proteins
ST category**		
ST53 (7)	19	ATP-binding protein, phage antirepressor, filamentous phage Cf1c related protein, M48 family peptidase, lipase chaperone and 14 hypothetical proteins
ST01 (6)	0	
ST11 (6)	4	Acyl-ACP-UDP-N-acetylglucosamine O-acyltransferase, phosphoenolpyruvate-protein phosphotransferase, restriction endonuclease - SacI family and a hypothetical protein.
ST14 (3)	16	Pyridoxal phosphate-dependent aminotransferase, SMP-30/gluconolactonase/LRE family protein and 14 hypothetical proteins
ST07 (3)	1	Head completion protein
ST06 (3)	9	Hypothetical proteins

* We considered the groups with at least 3 strains
**Number of genomes is indicate in parentheses.

Regarding the country category, in the groups of Brazil and Costa Rica were found 2 and 7 exclusive orthologs, respectively. The orthologs are an integrating conjugative element relaxase and a hypothetical protein for Brazil group. A CopG family transcriptional regulator and the RelE/ParE family toxin (type II toxin-antitoxin system located in the plasmid) were detected in the Costa Rica group.

The ST categorization yield the greatest number of exclusive orthologs per group, except for the ST_01. Some orthologs found belong to phage-related proteins, such the ones found in ST_53 and ST_07 groups. Transferases were also found in two categories (ST_11 and ST_14). Several proteins with hypothetical function were found in all categories, except for ST_01 and ST_07.

4.1.3. Phylogenomics of *X. fastidiosa*

Both the Maximum Likelihood (ML) Tree and the Consensus (CS) Tree, allowed to group the 46 strains in three main clades. These clades correspond to three subspecies already reported: *fastidiosa*, *pauca*, and *multiplex* (Marcelletti & Scortichini, 2016a). Despite clustering in three main clades, the topology of the clades within each main clade showed a different pattern in the trees, especially in subsp. *pauca* and *fastidiosa*.

The ML tree (Fig. 8) showed a topology where the strain CFBP8072 (from coffee trees of Ecuador) was located in the base of the *pauca* clade. Two derived clades were found, dividing strains from Italy and Costa Rica (P2) from other South American strains from Brazil and Argentina (P1), with the *Hibiscus* strain in the base of this. In the CS tree (Fig. 9), it was also found the same clear division in the *pauca* clade (displayed as P1 and P2) with the same component strains, although the Ecuador strain was not in the base of the entire *pauca* clade, but instead was in the base of the P2 clade only, while the *Hibiscus* strain was found again in the base of P1 clade. Regarding the subsp. *fastidiosa*, one subclade, formed by the strains Mul-MD, MUL0034 and Mullberry, was observed in both ML and CS trees. Another subclade of subsp. *fastidiosa* grouped the strains Ann-1 and CO33 only in our CS tree. The last subspecies *multiplex* was shown as a monophyletic group in both trees.

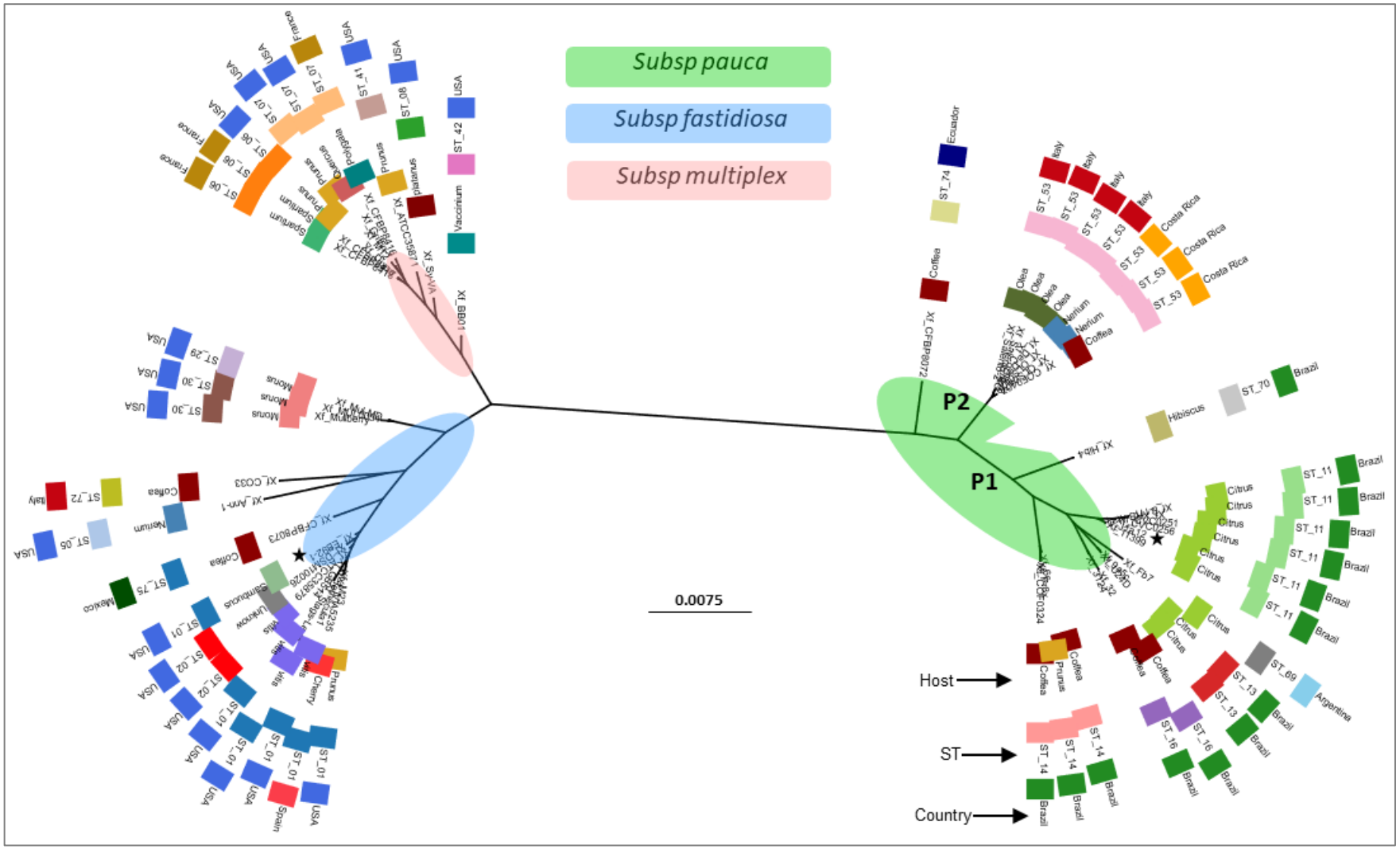

Fig. 8. Maximum Likelihood tree of *X. fastidiosa* strains. The 46 strains included in this study were grouped in three main clades. Each clade represents the known subspecies *pauca*, *fastidiosa* and *multiplex*. P1 indicate a subclade formed only by south American strains and P2 formed principally by Italian and Costa Rica strains. The strain CFBP8072 from Ecuador showed an early divergence in the subsp. *pauca*. In asterisk: Non-pathogenic strains.

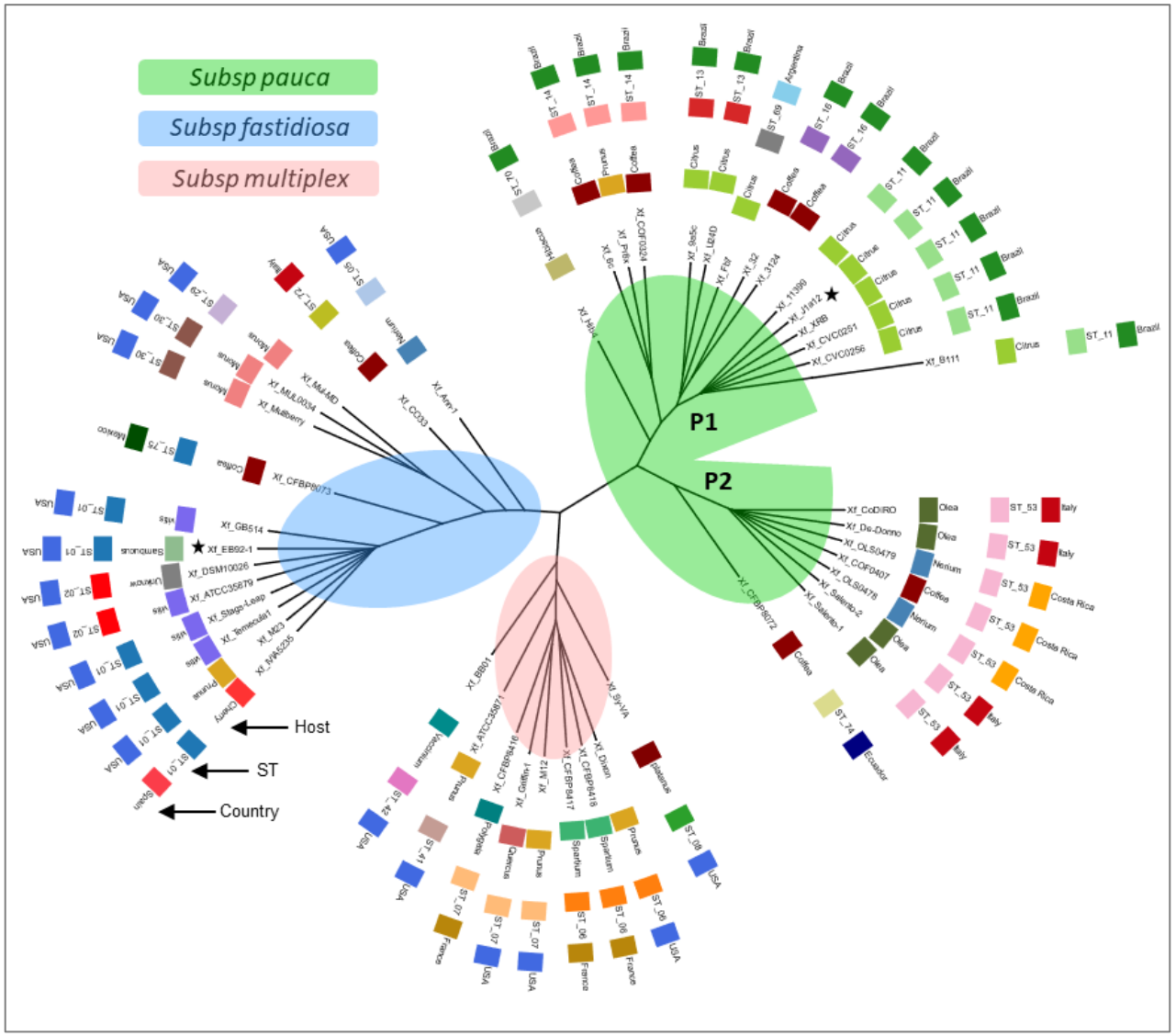

Fig. 9. Consensus of the core-genome orthologous trees. As in Fig. 8, three main clades related to the subspecies *pauca*, *fastidiosa* and *multiplex*, are showed. P1 indicate a subclade formed only by south American strains and P2 formed principally by Italian and Costa Rica strains. The strains CFBP8072 and Hib4 were in the bases of the subclades P1 and P2, respectively. In asterisk: Non-pathogenic strains.

The metadata information (Table 4) were added in each branch of the phylogenomic trees (Fig. 8 and Fig. 9). Most of the clades in both trees has some relationship with the ST and country of origin categories. The ST_01 (*fastidiosa*), ST_11 (*pauca,* P1) and ST_53 (*pauca,* P2) were the most frequently. Regarding to the country of origin, P1 of both trees, grouped mainly Brazilian strains. Another example was observed in the subsp. *fastidiosa* where there are subclades formed mostly by USA strains. Although the host metadata showed less relationship with the clades, most strains isolated from *Citrus* were grouped in one clade of P1. The four strains from *Olea* was grouped in one clade of P2. In the subsp. *fastidiosa*, both strains from *Vitis* and *Morus* were grouped in different clades. However, the last one was shown to have a monophyletic clade. In contrast, the strains from *Coffea* are present in more than one clade, that is the *fastidiosa* and *pauca* subspecies clades.

4.1.4 Virulence orthologs in genomic context

PROKKA re-annotation found orthologs related with the virulence and pathogenesis of *X. fastidiosa*, such as the fimbrial and afimbrial adhesins essential for biofilm formation (Chatterjee *et al.*, 2008b; Caserta *et al.*, 2010), plant cell wall degrading enzymes (Kubicek *et al.*, 2014), biosynthesis of exopolysaccharides (EPS) (da Silva *et al.*, 2001), among others. Notably, various of these orthologs were detected in the core genome (Table S3). At least 50 orthologs were related to pilus and fimbrial proteins, which 29 were found in all genomes. One of these orthologs is related to a prepilin peptidase and could be also related with protease activity (Paetzel, 2019). Regarding the trimeric autotransporter adhesins (TAAs) reported in Temecula1 by three paralogs (XadA1, XadA2 and XadA3) (Caserta *et al.*, 2010; Zaini *et al.*, 2015), PROKKA re-annotation found one ortholog for each of the three TAAs in all genomes. Phylogenetic trees of fimbrial adhesins (Fig. 11) showed clades with sequences that share more similarities than TAA orthologs (Fig. 10). Remarkably, the XadA1 and XadA3 showed clades more related with the host *citrus*, *olive*, *vitis* and *morus* (Fig. 10). In the case of the orthologs related to hemagglutinins, only two of eleven were presented in the core-genome. Besides, in one ortholog was detected up to 13 domains distributed among their sequences.

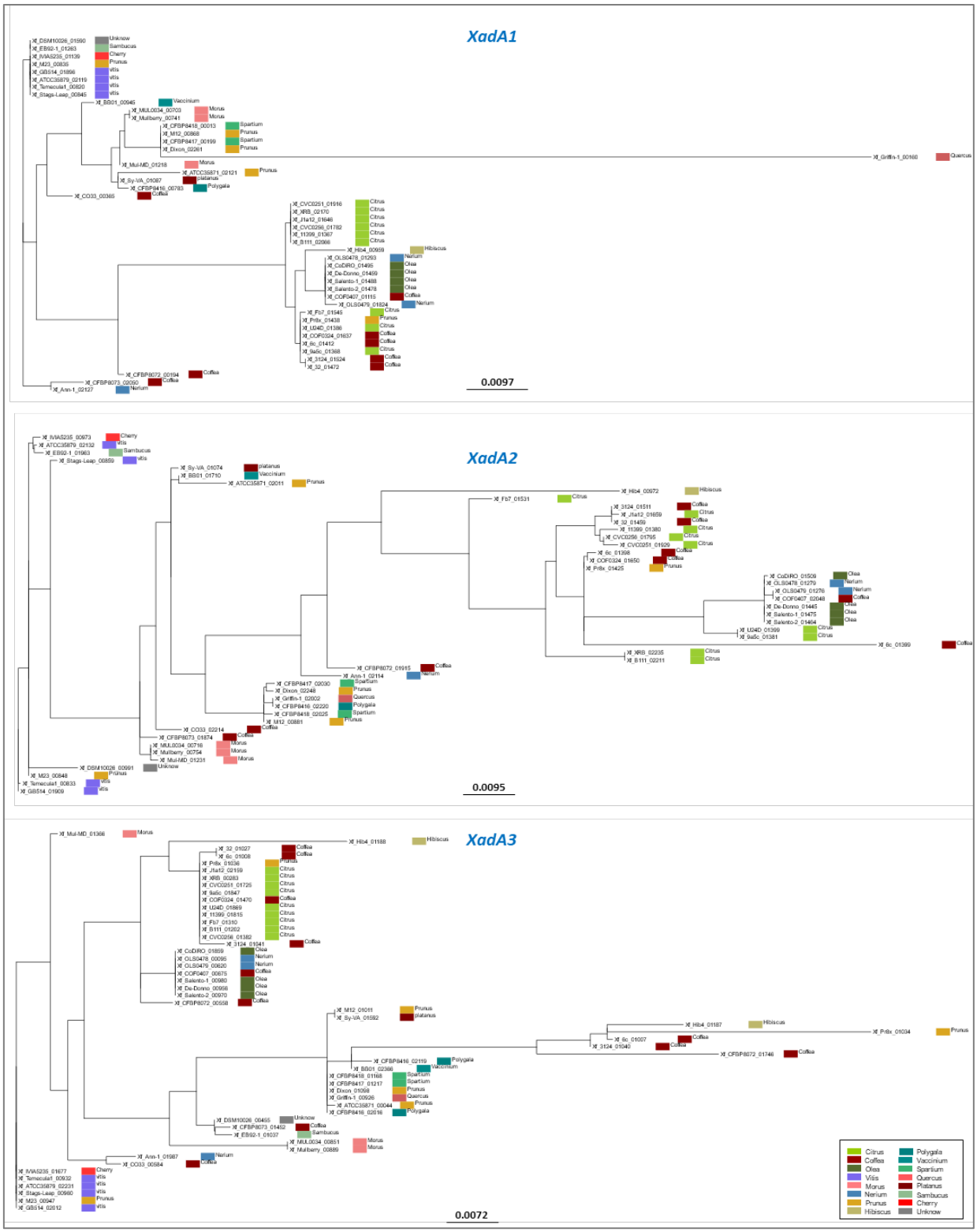

Fig. 10. Phylogenetic trees belonging to afimbrial adhesins orthologs XadA1, XadA2 and XhadA3 present in all strains of *X. fastidiosa*. The host is shown in each branch.

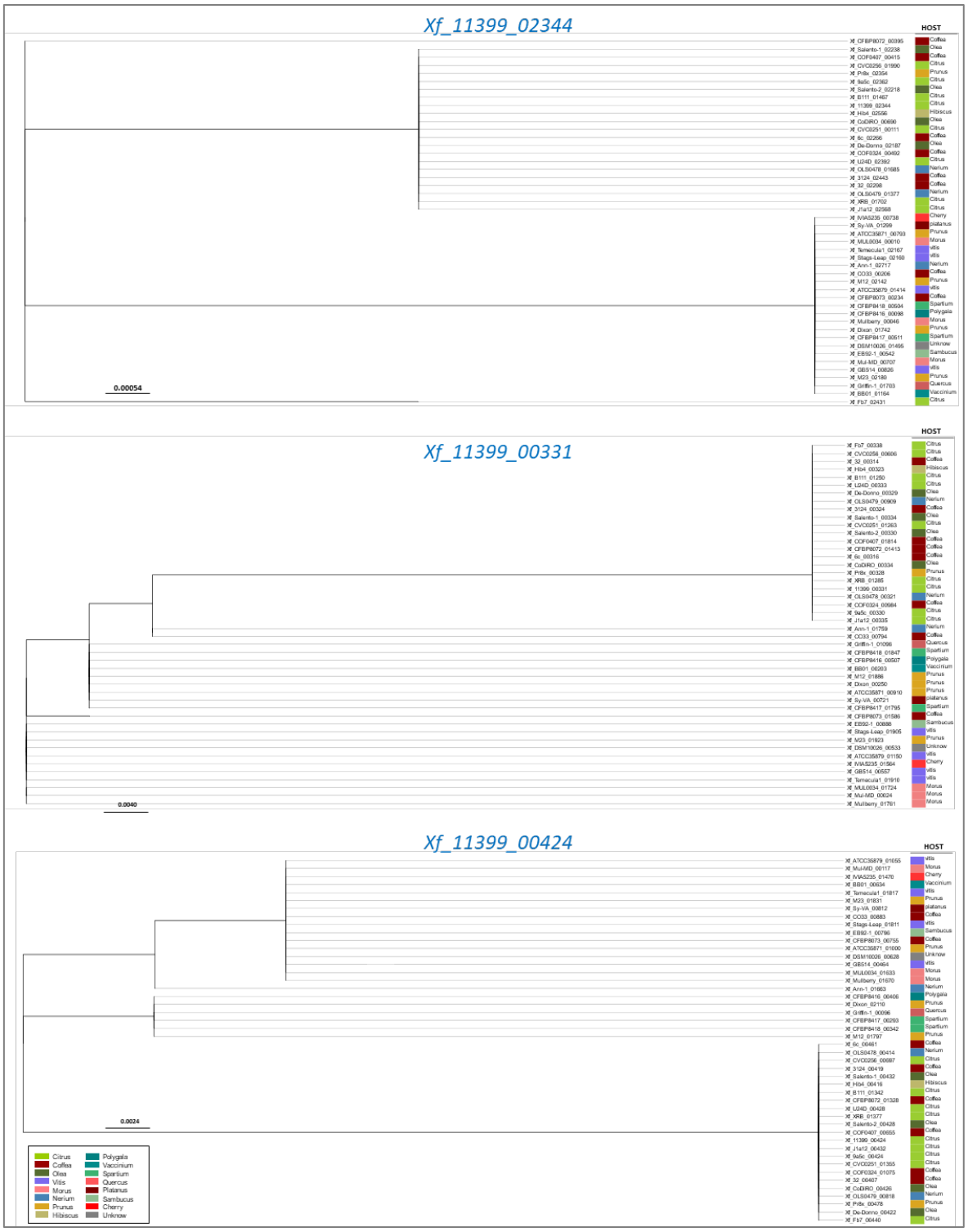

Fig. 11. Phylogenetic trees of orthologous fimbrial adhesins present in all genomes.

The cell wall degrading enzymes (CWDEs), such as polygalacturonase, cellulase and pectinases are enzymes required by phytopathogens for the hydrolysis of the plant cell wall (Kubicek *et al.*, 2014). The CWDEs detected in genomes of *X. fastidiosa* showed the potential to be secreted according to

SecretomeP results (Table 5). One polygalacturonase ortholog was present in all genomes, however an alignment of the sequences showed the truncation of this protein in 11 genomes due to a frameshift mutation (an insertion of a single adenine nucleotide at the position 498 in 9a5c strain) (Fig. 12a). It was detected this frameshift in strains isolated from citrus (9a5c, U24D, Fb7, J1a12, CVC0251, CVC0256, 11399, B111 and XRB) and coffee (32 and 3124), all belonging to the subsp. *pauca*. Nonetheless, the strains Pr8x, 6c, Hib4, COF0324 and CFBP8072, also originated in South America, did not show this mutation and, thus, possess a complete sequence of polygalacturonase. They share this feature with all strains of subsp. *multiplex* and *fastidiosa* (Fig. 12b), pointing to a fully functional enzyme.

Table 5 Number of CWDE, lipases and proteases in the analyzed genomes

CWDEs	Total	In the core-genome
Polygalacturonases	1	1
Cellulases*	5	2
Galactosidases	1	1
Glucosidases	3	2
Lipases	11	3
Proteases	24	13

* Include cellobiosidases and endoglucanases

Others secreted enzymes such as lipases and proteases were also detected (Table 5). Unlike CWDEs, these enzymes have more orthologs distributed among strains. Regarding the lipases, one case was found. Three lipases originated by duplication (LesA, LesB and LesC) have been reported in Temecula1 (Nascimento *et al.*, 2016). Only LesA have hydrolytic activity due to two catalytic sites (Gouran *et al.*, 2016). In these analyses, LesA and LesB were grouped together in one unique ortholog. The genomes from South America strains have LesA, with the exception of Hib4, which has only LesB. A similar case was detected in the subsp. *fastidiosa* where LesA is absent in EB92.1, CFBP8073 and DSM100026 strains. Here it was reported others strains that could have lost one of these lipases, important to bacterial pathogenesis. Interesting, the sequences of LesB in the genomes of Hib4 and EB92.1 have high similarity with the non-hydrolytic LesB of Temecula1. However, only the LesB of Hib4 has one of the catalytic sites of the hydrolytic LesA (Fig. 13).

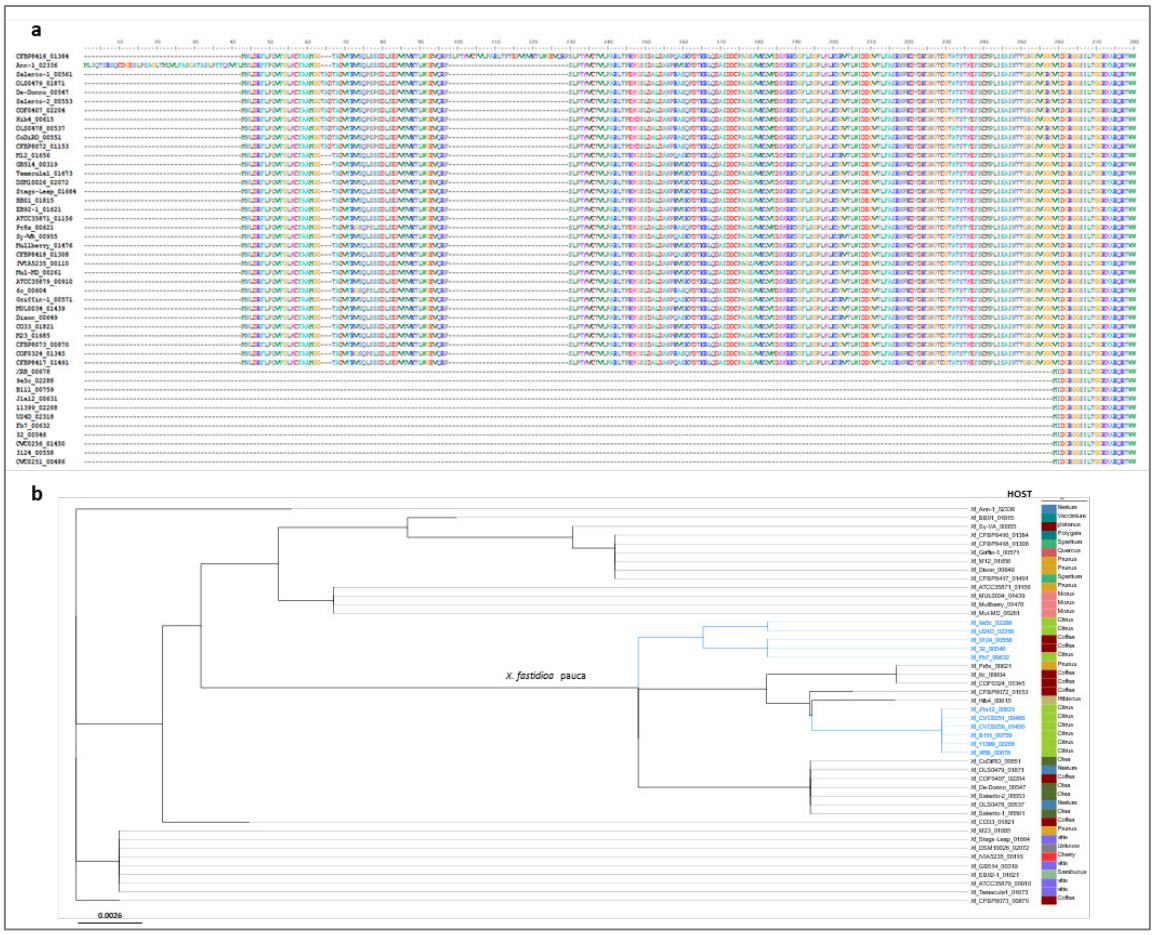

Fig. 12. Ortholog of polygalacturonase belonging to the core-genome of *X. fastidiosa*. **a** Alignment of the amino acid sequences showing the loss of a region in the last eleven ones. **b** Phylogenetic trees showing the strains of the subsp. *pauca* with frameshift mutation (light blue).

4.1.5. Functional analysis of the core and accessory genomes

To identify whether other orthologs of specific functional groups are found in all strains (core-genome) or if they diverge between strains (accessory-genome), COG (Cluster of Orthologous Groups) (Galperin *et al.*, 2015) functional categories for the core and accessory genes of all 46 strains were explored (Fig. 14). As expected, the COG functional categories of some highly conserved biological processes, such as “Translational, ribosomal structure, and biogenesis” (Class J) and “Cell wall/membrane/envelope biogenesis” (Class M), were more enriched among the core genes than accessory genes. In contrast, the category X of “Mobilome: prophages, transposons” was highly enriched in the accessory genome reaching almost 20%. Other categories more enriched in the accessory-genome were “Replication, recombination and repair” (Class L) and

“Defense mechanisms” (Class V). The high percentage of Class L could be explained by the great proportion of transposases (Class X), or by the presence of homologous DNA replication enzymes found also in phages, which are found in most of the strains. The abundance of genes categorized in Class V could reflect the capacity of the strains to survive under different stress conditions in different environments. As well as, host´s defense system against phages, such as genes of restriction enzymes or CRISPR/Cas system. Finally, genes of unknown function (Class S) were more abundant in the accessory-genome, as well as genes not assigned to any COG class.

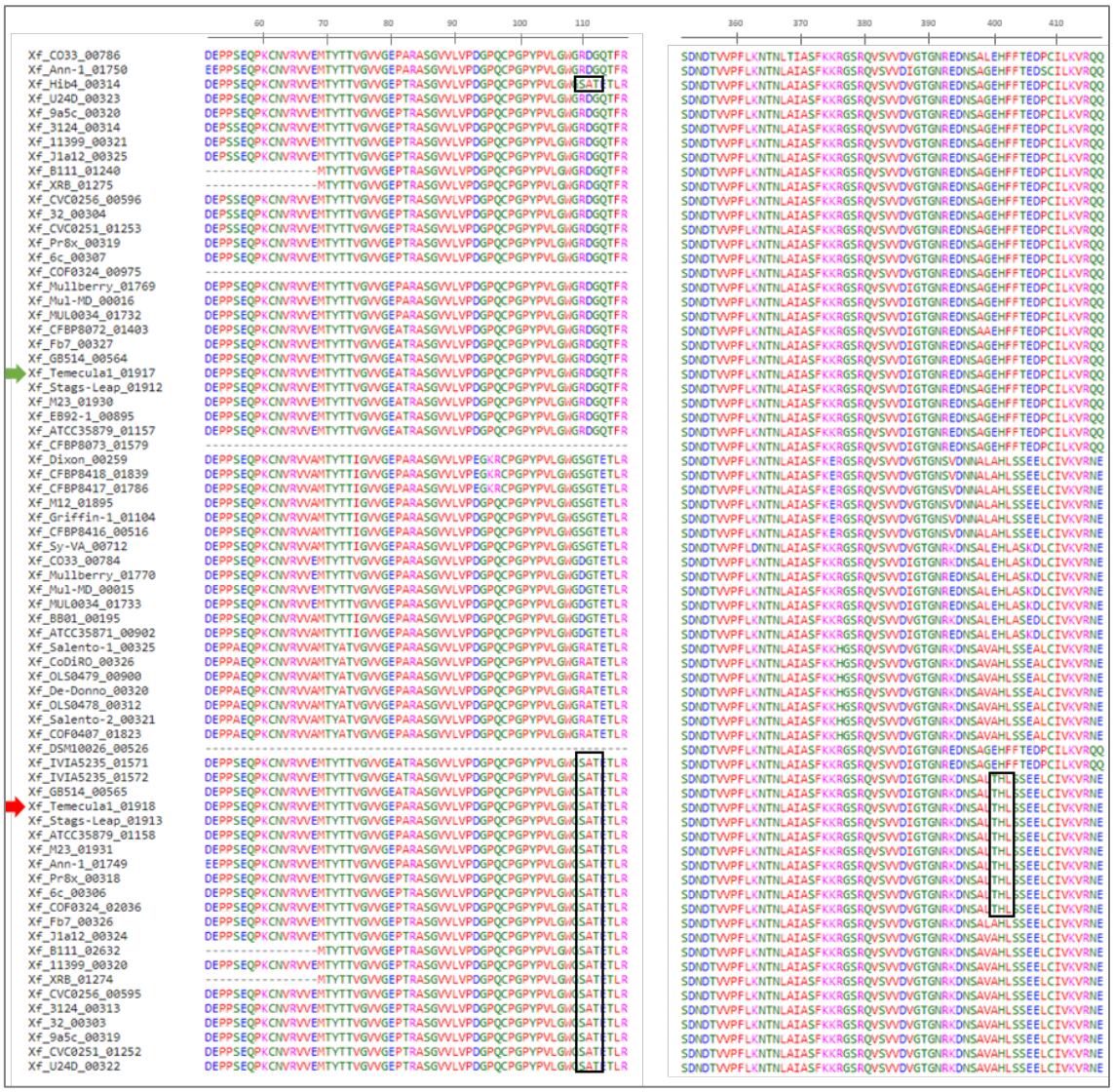

Fig. 13. Alignment of the sequences belonging to the LesA-LesB ortholog. In the black boxes are showed the reported catalytic sites. Red and green arrows indicate the lytic LesA and non-lytic LesB in Temecula1, respectively.

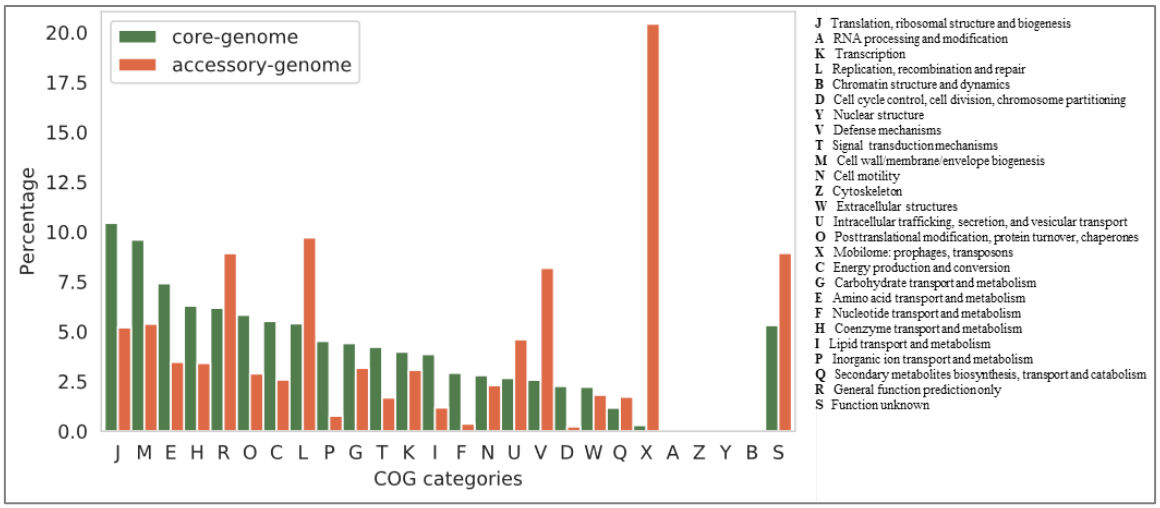

Fig. 14. COG Functional analysis of the core-genome and accessory-genome. Notably, the category more enriched is X, which is related to the Mobilome.

4.2. Prophages, genomic islands and insertion sequences

4.2.1. Mobile elements of the chromosome of *X fastidiosa* strains

As indicated in the previous section the analyses of COG categories in the *X. fastidiosa* genomes (Fig. 14) revealed that around 20% of the accessory-genome comprise mobilome-related genes (Category X). Furthermore, in the analysis of the singleton-genome could be observed exclusive proteins related to phages, highlighting the horizontal gene transfer (HGT) events in *X. fastidiosa*. The genes acquired by HGT are introduced via mobile genetic elements (MGEs) and incorporated into the chromosome by homologous or illegitimate recombination (Llop, 2015). Taking in consideration that results in *X. fastidiosa* genomes, and the significance of the MGEs, it was explored, compared and visualized the prophages, genomic islands and insertion sequences harbored in the chromosomal genome of the 46 *X. fastidiosa* strains described before.

Recently, a practical guide to the analysis of MGEs has suggested the combination of available software and database tailored to small scale genomes analysis (Oliveira Alvarenga *et al.*, 2018). Thus for prediction of the MGEs in *X. fastidiosa* genomes we built a pipeline based in the previously reported protocol (Oliveira Alvarenga *et al.*, 2018) with some modifications (Fig. 2). The results of the prophage predictors (Table 6) show that VirSorter predicted more prophage regions than PHASTER and PhiSpy. For Insertion sequences, similar situation was observed in the ISEScan results comparing with Oasis (Table 8). For

genomic island (Table 7), that trend was not observed with results of Alien_hunter and SeqWord Sniffer. To combine the results, a consensus region for each MGEs in *X. fastidiosa* genomes was calculated (see in the methodology). In the case of prophages consensus regions, the number of those classified as member of *Inoviridae* family (by Inovirus_detector software) is indicated in parentheses in Table 6. For the genomic island consensus, the parentheses indicate the number of genomic islands with category assigned by GIPSy software.

Table 6 Results of prophages predictor software

strain	PHASTER	VirSorter	PhiSpy	Consensus*
9a5c	7	6	7	9 (1)
3124	7	7	7	11 (1)
Ann-1	13	9	9	12 (2)
De-Donno	7	4	4	7 (2)
Fb7	9	8	7	10 (2)
GB514	8	5	5	10 (1)
Hib4	10	7	8	11 (2)
J1a12	7	8	8	9 (2)
M12	7	4	5	6 (1)
M23	10	5	4	10 (2)
MUL0034	9	7	8	10 (1)
Pr8x	7	7	9	10 (2)
Salento-1	7	4	5	9 (3)
Salento-2	6	4	4	8 (3)
Temecula1	9	5	4	10 (1)
U24D	7	6	7	9 (1)
32	8	14	7	9 (1)
11399	8	11	7	11 (1)
6c	8	7	7	9 (2)
ATCC35871	2	7	2	6 (2)
ATCC35879	9	9	5	11 (1)
B111	2	15	6	7 (2)
BB01	5	9	5	9 (1)
CFBP8072	2	6	3	3 (1)
CFBP8073	4	7	6	8 (1)
CFBP8416	4	3	7	7 (1)
CFBP8417	1	3	3	4 (1)
CFBP8418	1	4	2	3 (1)
CO33	9	12	2	12 (3)
CoDiRO	6	4	4	5 (2)
COF0324	4	12	8	13 (3)
COF0407	3	9	5	10 (1)
CVC0251	2	12	8	11 (2)
CVC0256	3	10	5	10 (2)
Dixon	9	10	6	8 (1)
DSM10026	4	6	4	7 (1)
EB92-1	4	8	8	13 (2)
Griffin-1	5	5	2	3 (1)
IVIA5235	4	10	6	8 (2)
Mullberry	10	7	8	10 (1)
Mul-MD	5	7	3	5 (2)
OLS0478	4	9	6	7 (2)
OLS0479	3	12	4	7 (1)
Stags-Leap	8	8	4	9 (2)
Sy-VA	3	12	4	8 (2)
XRB	4	12	7	12 (2)

*In parentheses: number of inovirus

Table 7 Results of genomic island predictor software

strain	Alien_Hunter	SeqWord Sniffer	Consensus*
9a5c	9	14	11 (3)
3124	8	8	8 (1)
Ann-1	10	6	8 (1)
De-Donno	10	5	9 (3)
Fb7	5	8	7 (1)
GB514	8	7	6 (1)
Hib4	7	13	9 (2)
J1a12	8	9	10 (2)
M12	11	6	8 (1)
M23	9	8	7 (1)
MUL0034	7	7	8 (2)
Pr8x	7	8	7 (1)
Salento-1	10	7	8 (2)
Salento-2	10	5	9 (3)
Temecula1	11	11	10 (2)
U24D	10	15	11 (3)
32	4	6	6 (1)
11399	7	13	11 (3)
6c	6	6	6 (2)
ATCC35871	13	6	11 (0)
ATCC35879	8	7	6 (1)
B111	3	9	7 (0)
BB01	8	7	10 (0)
CFBP8072	7	8	9 (1)
CFBP8073	9	8	9 (0)
CFBP8416	10	10	13 (0)
CFBP8417	15	8	14 (0)
CFBP8418	9	8	12 (0)
CO33	8	7	10 (1)
CoDiRO	4	9	9 (1)
COF0324	5	6	8 (1)
COF0407	7	7	9 (0)
CVC0251	5	4	4 (0)
CVC0256	6	5	6 (0)
Dixon	15	17	17 (1)
DSM10026	7	8	10 (0)
EB92-1	6	9	8 (1)
Griffin-1	5	16	13 (1)
IVIA5235	5	12	11 (0)
Mullberry	9	7	9 (1)
Mul-MD	12	12	14 (1)
OLS0478	6	5	5 (1)
OLS0479	2	8	7 (0)
Stags-Leap	7	7	8 (0)
Sy-VA	6	6	7 (0)
XRB	4	9	4 (0)

* In parentheses the number of GI with category assigned

Fig. 15 shows the percentage of total MGE in the chromosomal genome of *X. fastidiosa* strains. The content of MGE found ranged from 12% to 28% with a mean value of 19.68% being probably one of the species with more content of MGE by genome. The distribution of the percentages had an association with the size chromosomal genome; the larger chromosomal genome, the greater percentage of MGEs. At the same time, complete genomes tended to have higher percentage of MGE except in a few cases, such as the scaffold genome of Dixon

strain, which was shown to have the highest percentage of MGE (28.67%). Other strains with high content of MGEs are Ann-1, Hib4 and U24D (more than 25%). The strains with lower content of MGE were Sy-VA, CFBP8418 with 12%.

Table 8 Results of insertion sequence predictor software

strain	Oasis	IScan	Consensus
9a5c	1	2	**2**
3124	9	2	**11**
Ann-1	2	1	**3**
De-Donno	2	2	**2**
Fb7	6	2	**6**
GB514	4	1	**5**
Hib4	6	2	**8**
J1a12	7	2	**7**
M12	8	4	**8**
M23	5	1	**5**
MUL0034	11	1	**12**
Pr8x	5	2	**6**
Salento-1	2	2	**2**
Salento-2	2	2	**2**
Temecula1	5	1	**5**
U24D	0	2	**2**
32	8	0	**8**
11399	0	2	**2**
6c	0	2	**2**
ATCC35871	0	0	**0**
ATCC35879	7	1	**7**
B111	2	2	**4**
BB01	0	2	**1**
CFBP8072	0	1	**1**
CFBP8073	0	2	**1**
CFBP8416	0	0	**0**
CFBP8417	2	0	**2**
CFBP8418	0	0	**0**
CO33	0	0	**0**
CoDiRO	2	2	**2**
COF0324	4	2	**6**
COF0407	0	1	**1**
CVC0251	0	2	**1**
CVC0256	0	2	**1**
Dixon	11	5	**14**
DSM10026	0	5	**1**
EB92-1	0	3	**2**
Griffin-1	0	1	**1**
IVIA5235	0	1	**1**
Mullberry	11	1	**12**
Mul-MD	0	1	**1**
OLS0478	2	2	**4**
OLS0479	0	1	**1**
Stags-Leap	5	1	**6**
Sy-VA	0	1	**1**
XRB	4	2	**6**

In summary, the proportion of MGE represented in each genome of *X. fastidiosa*, the number of prophages, genomic islands and insertion sequences varies considerably. Nevertheless, it was possible to differentiate five groups

based in their enrichment (Fig. 15). Both the groups (i) and (ii) are enriched by all MGE and they are conformed principally by strains belong to subsp. *pauca* from South America, subsp. *fastidiosa* and only M12 and Dixon from subsp. *multiplex*. In the case of the group (iii), it was more enriched by genomic island and the strains that conformed this group are principally from subsp. *multiplex*. Finally, the group (v) it was enriched only by prophages and genomic island. This group is also distributed among the subsp. *pauca* (from South America and Europe) and subsp. *fastidiosa*, but none of subsp. *multiplex*.

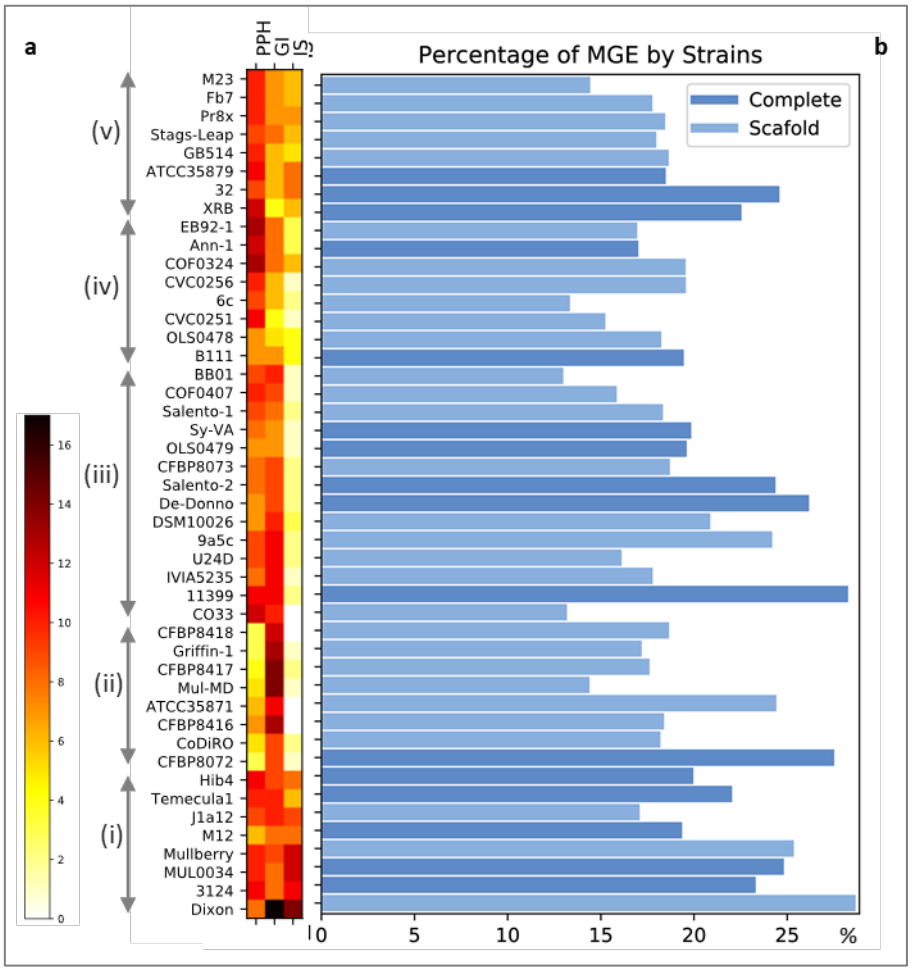

Fig. 15 Mobile Genetic Elements (MGE) in *X.* fastidiosa genomes. **a** Heatmap showing the density of each MGE analyzed (PPH: prophage, GI: Genetic Island and IS: insertion Sequence). **b** Percentage of the MGE in the genome.

It is important to highlight the usually uniform MGE organization into the chromosomal genome (with complete status) among the *X. fastidiosa* strains (see Fig. 21 below). The prophages are slightly concentrated in the central region of the genome, where the replication process is ended in bacterial cells. Genomic islands are usually at the terminal regions of genomes and most IS are into or

flanking a respective prophage, although few IS are isolated entities with no other near MGE.

4.2.2. Prophages

The prophages are important to the evolution, virulence and fitness of the bacterial pathogens (Fortier & Sekulovic, 2013). Three prophage predictor software were used for *X. fastidiosa*: PHASTER, VirSorter and PhiSpy. The results of the prophages identified in all strain by each predictor were detailed in the Table 6. The pipeline constructed to calculate prophages consensus regions was able to identify 396 regions distributed among the 46 strains analyzed. That prophage regions have a mean of 33.2 Kbp, the smallest (3.2 Kbp) found in the strains CVC0251 and the largest (143.1 Kbp) in J1a12. Regarding the number of prophages by strains, it was found an average of 8.6. The strains COF03424 and EB92-1, isolate from coffee and elderberry respectively, showed to have the greatest number of prophages (13) harbored in their chromosomal genome, and the strains with the least prophages (3) were CFBP8072, CFBP8418 and Griffin-1.

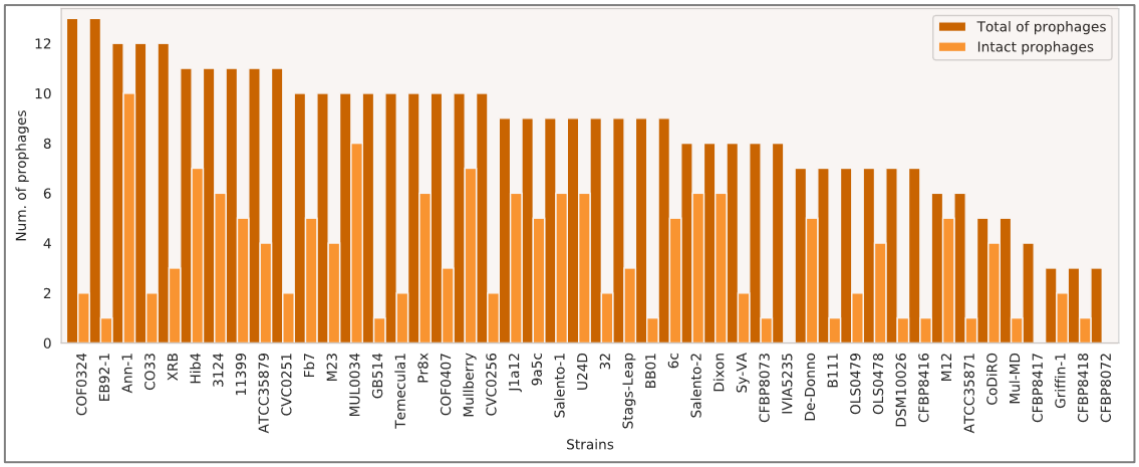

Fig. 16 Total and intact prophages among *X. fastidiosa* genomes

The software PHASTER, identified intact prophages which is related with the potential viability (Arndt *et al.*, 2016). Considering this classification, 39.65 % (157) of the total prophage predicted were identified as intact prophages. Within this proportion, the size mean of intact prophages was 38.9 Kbp ranging from 4.6 to 143.1 Kbp in ATTCC35879 and J1a12 strains, respectively. The strains that showed to have a higher proportion of intact/total prophages were Ann-1, M12,

MUL0034 and CoDiRO with more the 80% (Fig. 16). Moreover, it was found strains without intact prophages such as IVIA 5235, CFBP8417 and CFBP8072. The detection of inovirus was performed using the software inovirus_detector and at least one inovirus was identified in the genomes analyzed. The strains that harbored up to 3 inovirus were CO33 belonging to the subsp. *fastidiosa*, COF0324, Salento-1 and Salento-2 belonging to the subsp. *pauca*.

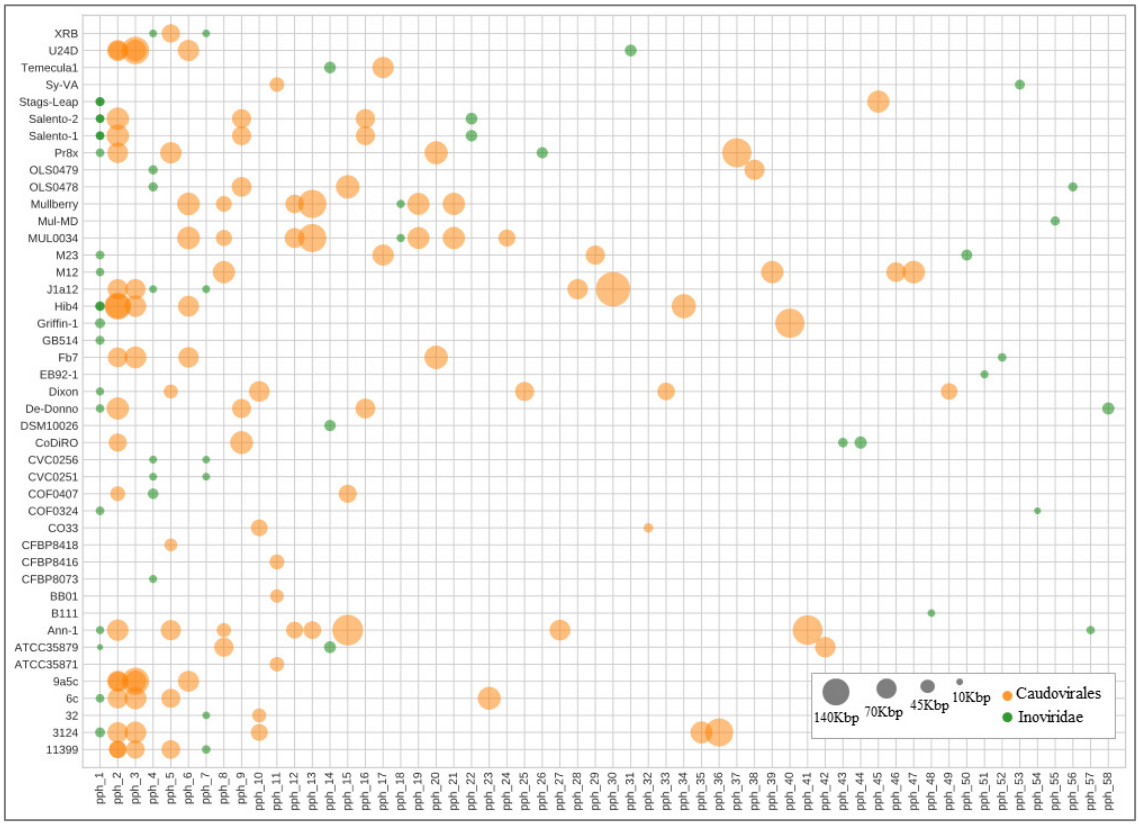

Fig. 17. Distribution of prophages with lytic potential in *X. fastidiosa* genomes. Each circle represents a prophage, which color green indicates a specific inovirus prophage. In some case more than one homologous prophage was present in the same strain (darker circles). The circle size indicates proportional prophage length in Kbp according to the legend.

In order to explored homology relationships of potential lytic prophages in *X. fastidiosa* strains, the 157 intact prophages were compared using Blast. The -log(e-value) of the hits with an identity and coverage greater than 90% and 80% respectively, was used for grouping with MCL software. After manual inspection of each group, 58 groups of intact prophages were found (Fig. 17). The group with more prophages (pph_1) contains 20 inovirus prophages. In the case of the strains Salento-1, Salento-2 and Hib4, two inovirus with related sequence were found in the same chromosomal genome. The second and third groups are also

integrated by strains with more than one prophage in their genome. Thirty-five unique prophages sequences (from pph_22 to pph58 in Fig. 17) were found distributed among 26 strains. Within these prophages, 13 were inovirus. Fig. 18 shows the alignment of the prophage sequences in the group pph_1.

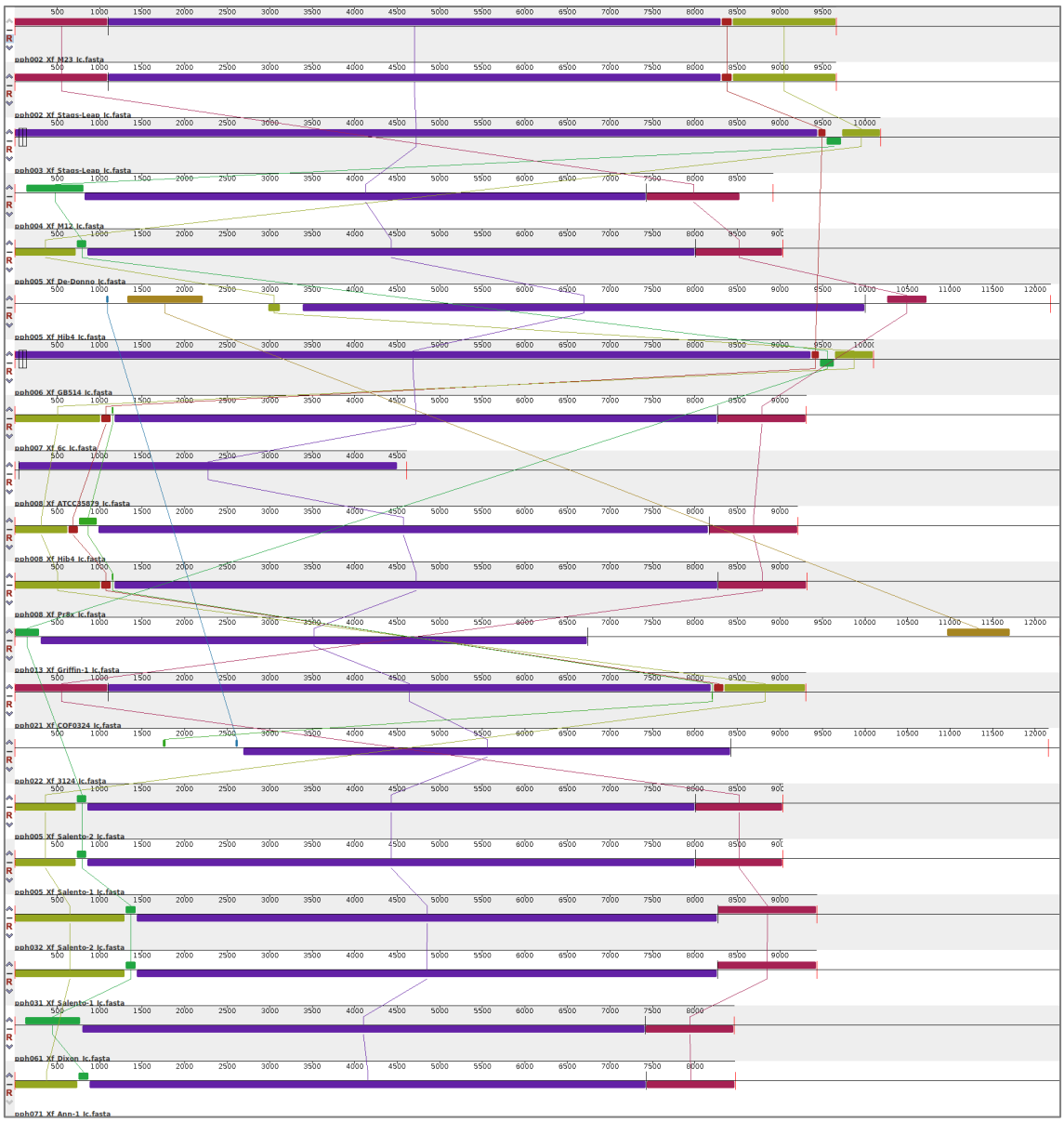

Fig. 18. Alignment of filamentous prophage sequences of pph_1 group (Figure 16).

4.2.3. Genomic islands

The genomic islands are also relevant to the bacterial fitness, eventually carries virulence factors and other effectors, influencing on how the bacterium handles external world. Alien_Hunter and SeqWord Sniffer were the software

used to predict genomic islands in the chromosomal genome of *X. fastidiosa* strains. A total of 409 genomic islands consensus were calculated among the 46 strains analyzed (Table 7). The mean of the size between these sequences was 25.41 Kbp, the smallest was found in Ann-1 and the largest in IVIA5235 with a size of 12.1 and 84.1 Kbp, respectively. Related to the number of genomic islands by genome, the mean was 8.9. Dixon strain belonging to the subsp. *multiplex* showed the highest number of genomic islands (17). Other strains of the same subspecies, such as CFBP8416, CFBP8417, CFBP8418 also showed to have a greater number of genomic islands. On the other hand, the genome with less genomic islands detected was of the citrus strain XRB.

The GIPSy software (Soares *et al.*, 2016) used to assign the category to each genomic island consensus was able to identify specific genomic island such as pathogenicity island (PAIs), metabolic islands (MIs), antibiotic resistance islands (RIs) and symbiotic island (SIs) (Fig. 19). For the case of genomic island consensus in the genomes of *X. fastidiosa*, 46 genomic islands were categorized among 30 strains (Table 7, showed in parentheses). Within these 46 specific genomic islands, 37 genomic islands were classified as RIs and only 3 were classified as PAI. Six genomic islands were classified with more than one category (Fig. 20). Among these latter, RI was the most frequently. And only in three islands, the type of SI was found in CFBP8072, Fb7, and Mul-MD. The comparison of the sequences of the 46 specific genomic islands using MCL allowed to group in 15 groups (Fig. 20). The first, third and fourth groups are formed only by RI. The second group also contains RI and one with two possible function RI-PAI (Hib4). From GI_5 to GI 15 correspond to unique genomic island, where some ones have mixed function as is shown in the Fig. 20.

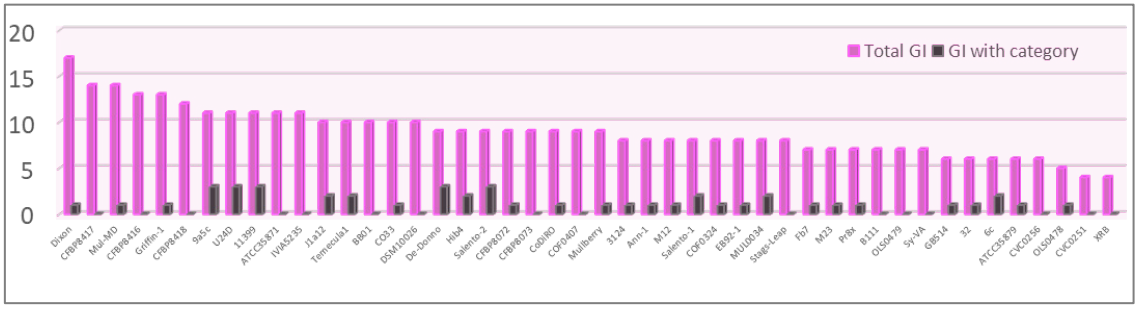

Fig. 19. Total of genomic islands (GIs) (pink) and GI with functional category found (grey) among *X. fastidiosa* genomes.

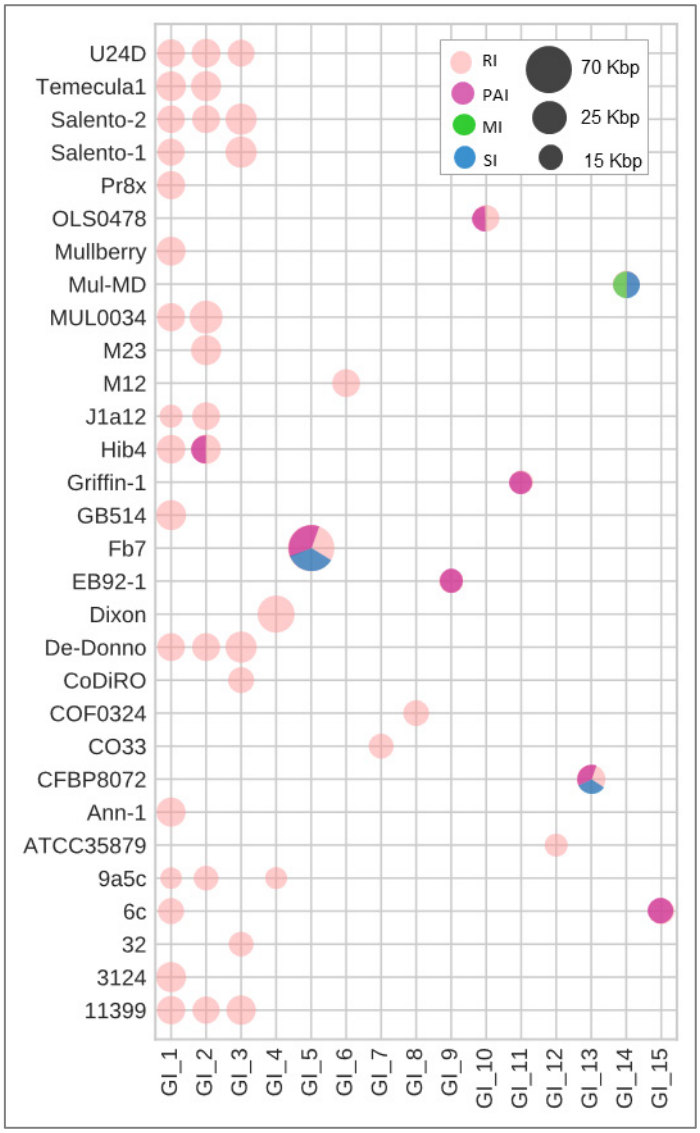

Fig. 20. Distribution of the Genetic Island (GI) with predicted function in *X. fastidiosa* genomes. RI Resistance island; PAI Pathogenic Island, MI Metabolic Island; SI Symbiotic Island.

4.2.4. Insertion Sequences

Insertion sequences are a driving force towards reallocations and transversions of chromosome regions, which eventually causes diminish reduction of functionality for certain genes. 175 insertion sequences consensus were detected among 42 strains of *X. fastidiosa*. Unlike prophages and genomic island, the amount of insertion sequences by genomes varies widely among the 42 strains. As it shown in the Table 8, the strains Dixon, Mulberry, MULL0034 and 3124 were the strains with a greater number of insertion sequences by

genome. In some strains it was not detected, such the strains CFBP8416, CFBP8418, ATCC3587 and CO33. Of the 175 insertion sequences, 56.0% is located inside of the prophages, 42.3% is in non-MGEs region but in some cases these insertion sequences are flanking the prophages regions. Only 1,7% were found into a genomic island (Fig. 21 and Figure S4).

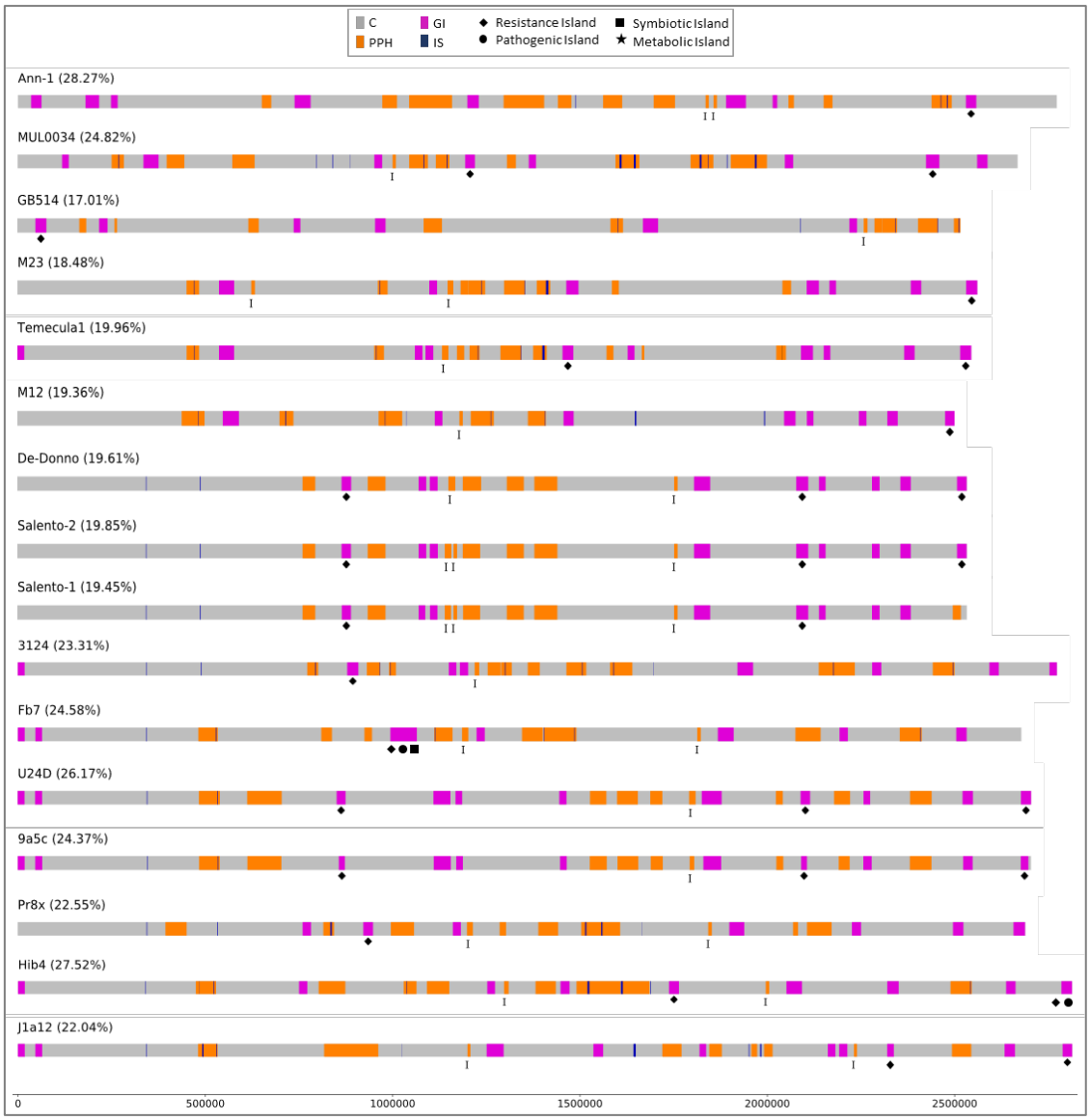

Fig. 21 Mapping of the MEG in the *X. fastidiosa* strains with complete genome. The percentage of MGE by genome is showed in parentheses. C: Chromosome, PPH: prophage, GI: genetic island, IS: insertion sequence, I: inovirus.

5. Discussion

5.1. Comparative genomic and phylogenomic analysis

In the present study, we have analyzed 46 genomes of *Xylella fastidiosa* from different sequencing projects and annotation pipelines. Therefore, it was decided to standardize the annotation process, using PROKKA. Forty-four genomes were retrieved from the NCBI database (genomes available until December 2018), while the other two genomes, sequenced in our lab, were not publicly available by that time. Among all strains, only 16 had a finished genome assembly, while the others had an incomplete or draft genome. However, the results of CheckM program and the evaluations of other parameters proposed by (Bowers *et al.*, 2017), have shown that these draft genomes have a high quality status and completeness. Therefore, our results comprise almost the complete genomic information about *X. fastidiosa* species.

We have applied a new approach, the GTACG framework, (Santiago *et al.*, 2018; Santiago *et al.*, 2019) to compare *X. fastidiosa* genomes, which is different from comparative analysis recently published (Giampetruzzi *et al.*, 2017b; Denance *et al.*, 2019; Vanhove *et al.*, 2019). This new approach uses a global metrics that considers all sequences in decision making. Its algorithm is based on the clustering coefficient to find and maximize the density of orthologous groups in genomes from closely related strains, as is the case of *X. fastidiosa*. Relevant biological criteria, such as the differences in sequences evolution, are considered by the algorithm to the definition of orthologous groups (Santiago et al., 2018). Among the advantages of this new approach, it was possible to highlight important traits related to biology and pathogenicity of *X. fastidiosa*.

The pan, core, and unique genome calculated by GTACG framework were compared with the other bioinformatics tools such as ROARY and Get_homologues using their default parameters. These last two software found more orthologs in the pan-genome of *X. fastidiosa* with an observable variation (Figure S1). Regarding to the number of orthologs in the core-genome, the variation between the tree software was minor. These results were expected because each one uses their own algorithm. However, some aspects must be considered. For instance, according the ROARY results, only the 12% would be

share among the *X. fastidiosa* strains. Although there is not defined proportion pan/core genome in bacterial, that percentage could not be entirely real. This due to strict clustering parameters in the formation of orthologs in *X. fastidiosa*.

In the genomic comparison analysis, it was detected the pan-genome of *X. fastidiosa* species is still open (Fig. 5), which means by adding new genome sequences, the curve will continue to grow. Our results showed a core-genome representing 31.5% of the pan-genome. A similar proportion of 29.3% have been reported in *Xanthomonas citri* (Bansal *et al.*, 2017), a phytopathogen phylogenetically related to *X. fastidiosa*. Organisms that are closely related to each other, such as the strains of *X. fastidiosa* analyzed here, are characterized to share more orthologs than unrelated species. Therefore, the set of *X. fastidiosa* strains used tended to have a relatively small pan-genome and a large core-genome. One interesting aspect to highlight is the presence of various orthologs related to pathogenicity in the core-genome. The phylogenetic trees constructed for each ortholog of the core-genome, as expected, showed the sequences are quite similar. However, the ortholog sequences related to afimbrial proteins such as XadA, showed more differences between them (Fig. 10). Moreover, it was observed a congruency related to the host metadata in XadA1 and XadA3 orthologous trees.

As it has been mentioned, the CDWEs are important for the *X. fastidiosa* pathogenesis (Chatterjee *et al.*, 2008a). Recent studies have found these enzymes as abundant components in the secretome of Temecula1 strain (Mendes *et al.*, 2016; Nascimento *et al.*, 2016; Feitosa-Junior *et al.*, 2019). Here we showed some CDWEs not reported before with the potential to be secreted and contribute to the *X. fastidiosa* pathogenicity. Considering the traits discussed and the results showed, apparently *X. fastidiosa* have the necessary machinery to infect plants. However, more detailed analyses for each gene needed to be performed. For instance, despite the gene for a polygalacturonase being an ortholog present in all strains, some South American strains have showed an insertion of a single adenine nucleotide that create a premature translation termination codon (Fig. 12), causing a frameshift mutation and consequently the depletion of its function. This trait has been reported only for the strains 9a5c (Van Sluys *et al.*, 2003) and 32 (Barbosa *et al.*, 2015), being correlated with the

differences in their infective capacity (Roper *et al.*, 2007b). Nonetheless, the possibility that another enzyme could replacing that function is still considered.

An interesting aspect was found in the lipases LesA and LesB, secreted enzymes also reported in *X. fastidiosa* secretome (Nascimento *et al.*, 2016). Nonetheless, only LesA have hydrolytic activity due to the presence of two catalytic sites (Gouran *et al.*, 2016). The absence of LesA in the genome of strain EB92.1 was previously reported, suggesting to contribute to its non-virulent phenotype (Zhang *et al.*, 2011). Here we showed other strains that lack LesA, for instance, Hib4 strain (Fig. 13); however, Hib4 has a LesB sequence with one of the catalytic sites of LesA, showing an important aspect of its evolution. Overall, the evidence for the secretion of enzymes such as CWDEs, lipases and proteases showed in our results reflects their importance for *X. fastidiosa* biology, contributing not only to the pathogenicity but also to nutrient acquisition, as proposed for Temecula1 strain (Nascimento *et al.*, 2016). It was also shown that other strains, like Pr8x, that have the two catalytic sites in LesA, suggesting the same activity reported in Temecula1. Other orthologs related to virulence feature are shown in Table S3.

The information of the core-genome (1,518 orthologs) was used to construct two phylogenomic trees (ML tree and CS tree). As shown in Fig. 8 and Fig. 9, both trees reflected a consistent grouping of the strains in three principal clades, which were related to subspecies *pauca*, *fastidiosa* and *multiplex*. These three clades also have been found and reported using ANIb, k-mers, MLSA (Denacé et al 2019) and other methods (Vanhove et al., 2019). However, the taxonomic definition of other subspecies such as *morus* and *sandyi* vary depending the phylogenomic data used and is still debatable (Almeida & Nunney, 2015; Giampetruzzi *et al.*, 2017b; Denance *et al.*, 2019; Godefroid *et al.*, 2019; Vanhove *et al.*, 2019). The CS tree (Fig. 9) showed a division of the subsp. *pauca* in the subclades P1 and P2, with the strains Hib4 and COF8072 located at the base of each subclade, respectively. This topology suggests an early divergency of that strains in its respectively subclades. Both phylogenomic trees allowed to show the relationship with the host, country of origin and ST, of which the last one was the most related with the subclades found. Our phylogenomic analysis also showed the closed relationship between the recently European strains and American strains. For instance, in one subclade of the subsp. p*auca*, the Italian

and Costa Rica strains are phylogenetically related. Genetic analyses have suggested an accidentally introduction in Italy from Costa Rica (Martelli *et al.*, 2016). Another example was observed in one subclade of the subsp. *fastidiosa* were the strain IVIA5235 from Spain was grouped with USA strains.

Both accessory-genome and singleton-genome of *X. fastidiosa*, represent more than half of the pan-genome (69.3%). The COG analysis showed a high proportion (20% of the accessory-genome) of genes related to the Mobile Genetic Elements category, indicating a high lateral gene transfer among the genomes of *X. fastidiosa*, in which the prophages and transposons have an important role. These genes could be moving between strains of *X. fastidiosa* and forming new trait combinations, possibly reflecting their adaptation to diverse environments (Segerman, 2012; Andam *et al.*, 2017) and possibly contributing to the pathogenesis (Nakamura, 2018). The inspection of the singleton-genome revealed a high amount of genes coding for hypothetical proteins, which indicates that a relative great number of these acquired genes are still unknown. It might also be implicated in the adaptation process of *X. fastidiosa* strains to their respective plant hosts. The prediction of domains in these proteins showed that they could be mainly related with membrane proteins and signaling peptides. Additional functional genomic studies (e.g. whole transcriptomic/proteomic profiling) will be important to understand the function of this large set of unique genes with hypothetical functions. We also observed the presence of prophages related genes in the singleton-genome. The transfer MGE such as plasmids and integrated prophages have been reported as the major player in *X. fastidiosa* evolution, and could partially answer the adaptation to particular life styles or niches, as well as host-association and virulence (Alcaraz, 2014). Interestingly, exclusive orthologs found in the ST_53 group (Table 2) are similar to proteins of the filamentous phage CfIc. Considering the high proportion of the Mobilome category in COG analysis, and the importance of the MGE, we have analyzed the prophages, genetic islands and insertion sequences in the genomes of *X. fastidiosa*.

5.2. Prophages, genetic islands and insertion sequences

Our analyses showed that the mean percentage of MGEs in the genomes of *X. fastidiosa* was 19.69%. In some strains, the percentage was more than 25%

such as Dixon, Ann-1, Hib4, U24D and Mullbery. A previously work in a pathogenic *Escherichia coli* has identified that the MGEs represent overall 8.7–19.8% of the total genome size (Delannoy *et al.*, 2017). Moreover, it has been suggested that prophages could constitute as much as 10-20% of a bacterial genome and are the major contributors to differences between individual within species (Casjens, 2003). *X. fastidiosa* strains have a high percentage of MGEs per genome, and for the purpose of this study we have considered prophages, genomic island and insertion sequences harbored on its chromosomal genome.

The mobilome is one of the responsible agents for the bacterial evolution, adaptation to niche, symbiosis, virulence, and increasing antibiotic resistance (Tettelin *et al.*, 2008; Aminov, 2011; Bozcal, 2019). Nevertheless, they are often underestimated due to inaccurate genome annotation. The heterogeneity of the results may be due to non-uniform criteria established by each software such as has been reported by Canchaya and collaborators (Canchaya *et al.*, 2003). Thus, construction of pipelines using more than one prediction software is need to improve the MGE identification (Oliveira Alvarenga *et al.*, 2018; Partridge *et al.*, 2018). In this work, we have determined consensus regions for each MGEs considered, in which at least two software were used, and a manual inspection was performed to confirm. A previous study about the prophages in complete genomes of 9a5c and Temecula1 had reported 12 and 10 regions, respectively (Varani *et al.*, 2008). The number of consensus prophages we found was 9 for 9a5c and 10 for Temecula1 (Table 6) and they are consistent with the prophages already published. For the case of 9a5c, the nine prophages consensus were corresponding with ten of the prophages reported. And for the other strain Temecula1, nine prophages consensus matched eight prophages reported. Also, in 9a5c, one genomic island consensus was corresponding with a region reported as prophage. For the first time, we show a comprehensive MGE analysis for 46 *X. fastidiosa* genomes including not only prophages, but also genomic island and insertion sequences.

According with our results, the presence of prophages consensus was observed in the 46 strains studied, with a mean of 8.6 prophages by genome. It suggests the present of poly-lysogeny in the species *X. fastidiosa* which might be important in the ecology of the bacteria to improve its competitiveness *in vivo* (Burns *et al.*, 2015). However, it is important to highlight that not necessarily those

with more consensus prophages in total, have more intact prophages (for instance COF0324, EB92-1, CO33) (Fig. 16). Even in the case of IVIA 5235, CFBP8417 and CFBP8072 no intact prophage was found. The mean (8.9) of genomic island consensus was similar to the prophage consensus, however, only 46 of them (11.2%), distributed among 30 strains (Fig. 19), were classified such a specific island (RI, PAI, MI or SI). The most of them was resistance island. It is thought that genomic islands may be transferred *in bloc* and possibly evolved from bacteriophages that have lost genes required for replication (Sundin, 2007). The results found suggest that possibly the ancestral *X. fastidiosa* were infected for other bacteriophages and degraded along the time, originating that genomic islands.

Despite the simplicity of the MCL algorithm similarity criteria (negative logarithm of the e-value) used to grouping prophages and genomic island, the algorithm may still detect homology relationships as it has been previously reported (Costa *et al.*, 2014). On the other hand, the complicated genomic structure of the prophages requires of new approaches to figure out homology between this type of MGEs. The prophages can be present in many different forms ranging from inducible prophages to prophages showing deletions, insertions, and rearrangements to prophage remnants that have lost most of the phage genome (Canchaya *et al.*, 2003). The intact prophages we have found could be clustered in 58 groups based on their homology and may have the potential to enter in a lytic cycle.

Some groups of intact prophages (pph_1, pph_2 and pph_3) are present in two copies in the same strain which could suggest superinfection events such was observed for inovirus (Roux *et al.*, 2019). One interesting aspect of the clustering was the presence of at least 5 groups related to inovirus family showing the importance of these inovirus in *X. fastidiosa*. Recent studies have shown the importance of the filamentous phages in the structure of the biofilm in others strains (Secor *et al.*, 2017), considering that *X. fastidiosa* is a bacterial that form biofilm, this kind filamentous phages could be giving some advantage to *X. fastidiosa*. It was important to mention that the incomplete prophages could also play a role in the biological of *X. fastidiosa*. It is known that many defective prophage remain functional and may carry a pool of beneficial genes to the host such have been reported in others strains (Casjens, 2003; Zhou *et al.*, 2011;

Bertelli *et al.*, 2018). More studies are necessary to understand the contribution of the content of prophages to *X. fastidiosa* species.

6. Conclusion

The results of comparative genomics and phylogenomics analyses of 46 *Xylella fastidiosa* strains, described in this work, revealed that 4942 and 1518 orthologs, respectively, in the pan- and core-genome of this bacterium. These analyses strongly support the strains geographic region of isolation at the genomic level while host-adaptation is weakly supported. All analyzed strains have the potential to be pathogens due to the presence of virulence and pathogenic orthologs in the core-genome. The analyses performed until now indicate that the afimbrial adhesin orthologs XadA1 and XadA3 sequence diversity has a relative congruence with the plant host.

A relevant and heterogeneous amount (12% to 28%) of Mobile Genetic Elements are harbored in the chromosome of the all 46 *X. fastidiosa* genomes. The mean number of prophages and genomic island regions per genome is similar (8.6 and 8.9, respectively). And among the predicted prophages, at least one inovirus was found highlighting the importance of the filamentous phages in *X. fastidiosa*.

7. References

1. **Ahern, S. J., Das, M., Bhowmick, T. S., Young, R. & Gonzalez, C. F.** 2014. Characterization of novel virulent broad-host-range phages of Xylella fastidiosa and Xanthomonas. J Bacteriol 196**:**459-71.

2. **Akhter, S., Aziz, R. K. & Edwards, R. A.** 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res 40**:**e126.

3. **Alcaraz, L. D.** 2014. Pan-genomics: Unmasking the gene diversity hidden in the bacteria species. :, PeerJ PrePrints, vol. 2. PeerJ.

4. **Alencar, V. C., Barbosa, D., Santos, D. S., Oliveira, A. C., de Oliveira, R. C. & Nunes, L. R.** 2014. Genomic Sequencing of Two Coffee-Infecting Strains of Xylella fastidiosa Isolated from Brazil. Genome Announc 2.

5. **Alencar, V. C., Jabes, D. L., Menegidio, F. B., Sassaki, G. L., de Souza, L. R., Puzer, L., Meneghetti, M. C. Z., Lima, M. A., Tersario, I. L. D., de Oliveira, R. C. & Nunes, L. R.** 2017. Functional and Evolutionary Characterization of a UDP-Xylose Synthase Gene from the Plant Pathogen Xylella fastidiosa, Involved in the Synthesis of Bacterial Lipopolysaccharide. Biochemistry 56**:**779-792.

6. **Almeida, R. P. P. & Nunney, L.** 2015. How Do Plant Diseases Caused by Xylella fastidiosa Emerge? Plant Disease 99**:**1457-1467.

7. **Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T. *et al.*** 2016. Standardized benchmarking in the quest for orthologs. Nature Methods 13**:**425.

8. **Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J.** 1990. Basic local alignment search tool. J Mol Biol 215**:**403-10.

9. **Alves, E., Kitajima, E. W. & Leite, B.** 2003. Interaction of *Xylella fastidiosa* with different cultivars of *Nicotiana tabacum*: a comparison of colonization patterns. Journal of Phytopathology-Phytopathologische Zeitschrift 151**:**500-506.

10. **Aminov, R. I.** 2011. Horizontal gene exchange in environmental microbiota. Front Microbiol 2**:**158.

11. **Andam, C. P., Challagundla, L., Azarian, T., Hanage, W. P. & Robinson, D. A.** 2017. Population Structure of Pathogenic Bacteria, p. 51-70. *In* M. Tibayrenc (ed.), Genetics and Evolution of Infectious Diseases (Second Edition). Elsevier, London.

12. **Argov, T., Azulay, G., Pasechnek, A., Stadnyuk, O., Ran-Sapir, S., Borovok, I., Sigal, N. & Herskovits, A. A.** 2017. Temperate bacteriophages as regulators of host behavior. Curr Opin Microbiol 38**:**81-87.

13. **Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. & Wishart, D. S.** 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44**:**W16-21.

14. **Backus, E. A., Shugart, H. J., Rogers, E. E., Morgan, J. K. & Shatters, R.** 2015. Direct Evidence of Egestion and Salivation of Xylella fastidiosa Suggests Sharpshooters Can Be "Flying Syringes". Phytopathology 105**:**608-20.

15. **Bansal, K., Midha, S., Kumar, S. & Patil, P. B.** 2017. Ecological and Evolutionary Insights into Xanthomonas citri Pathovar Diversity. Appl Environ Microbiol 83.

16. **Barbosa, D., Alencar, V. C., Santos, D. S., Oliveira, A. C. D., de Souza, A. A., Coletta, H. D., de Oliveira, R. C. & Nunes, L. R.** 2015. Comparative genomic analysis of coffee-infecting Xylella fastidiosa strains isolated from Brazil. Microbiology-Sgm 161**:**1018-1033.

17. **Bendtsen, J. D., Kiemer, L., Fausboll, A. & Brunak, S.** 2005. Non-classical protein secretion in bacteria. BMC Microbiol 5**:**58.

18. **Bertelli, C., Tilley, K. E. & Brinkman, F. S. L.** 2018. Microbial genomic island discovery, visualization and analysis. Briefings in Bioinformatics.

19. **Bezuidt, O., Lima-Mendez, G. & Reva, O. N.** 2009. SeqWord Gene Island Sniffer: a Program to Study the Lateral Genetic Exchange among Bacteria. International Journal of Computer and Information Engineering 3**:**2399-2404.

20. **Bhattacharyya, A., Stilwagen, S., Ivanova, N., D'Souza, M., Bernal, A. *et al.*** 2002a. Whole-genome comparative analysis of three phytopathogenic *Xylella fastidiosa* strains. Proc Natl Acad Sci U S A 99**:**12403-8.

21. **Bhattacharyya, A., Stilwagen, S., Reznik, G., Feil, H., Feil, W. S. *et al.*** 2002b. Draft sequencing and comparative genomics of *Xylella fastidiosa* strains reveal novel biological insights. Genome Res 12**:**1556-63.

22. **Blom, J., Albaum, S. P., Doppmeier, D., Puhler, A., Vorholter, F. J., Zakrzewski, M. & Goesmann, A.** 2009. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10**:**154.

23. **Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D. *et al.*** 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35**:**725-731.

24. **Bozcal, E.** 2019. Insight into the mobilome of Escherichia coli, p. 14p. *In* M. S. Erjavec (ed.), The Universe of Escherichia coli [Working Title], vol. open-access. Intechopen, London, UK.

25. **Burns, N., James, C. E. & Harrison, E.** 2015. Polylysogeny magnifies competitiveness of a bacterial pathogen in vivo. Evolutionary applications 8**:**346-351.

26. **Buttner, D.** 2016. Behind the lines-actions of bacterial type III effector proteins in plant cells. Fems Microbiology Reviews 40**:**894-937.

27. **Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H.** 2003. Prophage genomics. Microbiol Mol Biol Rev 67**:**238-76, table of contents.

28. **Carazzolle, M. F., Rabello, F. R., Martins, N. F., de Souza, A. A., do Amaral, A. M., Freitas-Astua, J., Pereira, G. A. G., Machado, M. A. & Mehta, A.** 2011. Identification of defence-related genes expressed in coffee and citrus during infection by Xylella fastidiosa. European Journal of Plant Pathology 130**:**529-540.

29. **Cariddi, C., Saponari, M., Boscia, D., De Stradis, A., Loconsole, G., Nigro, F., Porcelli, F., Potere, O. & Martelli, G. P.** 2014. Isolation of a *Xylella fastidiosa*

strain infecting olive and oleander in Apulia, Italy. Journal of Plant Pathology 96:425-429.

30. **Caserta, R., Takita, M. A., Targon, M. L., Rosselli-Murai, L. K., de Souza, A. P., Peroni, L., Stach-Machado, D. R., Andrade, A., Labate, C. A., Kitajima, E. W., Machado, M. A. & de Souza, A. A.** 2010. Expression of *Xylella fastidiosa* fimbrial and afimbrial proteins during biofilm formation. Appl Environ Microbiol 76:4250-9.

31. **Casjens, S.** 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49:277-300.

32. **Chang, C. J., Garnier, M., Zreik, L., Rossetti, V. & Bove, J. M.** 1993. Culture and serological detection of the xylem-limited bacterium causing citrus variegated chlorosis and its identification as a strain ofXylella fastidiosa. Curr Microbiol 27:137-42.

33. **Chatterjee, S., Almeida, R. P. P. & Lindow, S.** 2008a. Living in two worlds: The plant and insect lifestyles of Xylella fastidiosa. Annual Review of Phytopathology 46:243-271.

34. **Chatterjee, S., Newman, K. L. & Lindow, S. E.** 2008b. Cell-to-cell signaling in *Xylella fastidiosa* suppresses movement and xylem vessel colonization in grape. Molecular Plant-Microbe Interactions 21:1309-1315.

35. **Chatterjee, S., Wistrom, C. & Lindow, S. E.** 2008c. A cell-cell signaling sensor is required for virulence and insect transmission of *Xylella fastidiosa*. Proc Natl Acad Sci U S A 105:2670-5.

36. **Chaudhari, N. M., Gupta, V. K. & Dutta, C.** 2016. BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep 6:24373.

37. **Chen, J., Huang, H., Chang, C. J. & Stenger, D. C.** 2013. Draft Genome Sequence of Xylella fastidiosa subsp. multiplex Strain Griffin-1 from Quercus rubra in Georgia. Genome Announc 1.

38. **Chen, J., Xie, G., Han, S., Chertkov, O., Sims, D. & Civerolo, E. L.** 2010. Whole Genome Sequences of Two Xylella fastidiosa Strains (M12 and M23) Causing Almond Leaf Scorch Disease in California. Journal of Bacteriology 192:4534-4534.

39. **Chen, J. C. & Civerolo, E. L.** 2008. Morphological evidence for phages in Xylella fastidiosa. Virology Journal 5.

40. **Clifford, J. C., Rapicavoli, J. N. & Roper, M. C.** 2013. A Rhamnose-Rich O-Antigen Mediates Adhesion, Virulence, and Host Colonization for the Xylem-Limited Phytopathogen Xylella fastidiosa. Molecular Plant-Microbe Interactions 26:676-685.

41. **Clokie, M. R. J., Millard, A. D., Letarov, A. V. & Heaphy, S.** 2011. Phages in nature. Bacteriophage 1:31-45.

42. **Contreras-Moreira, B. & Vinuesa, P.** 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79:7696-701.

43. **Costa, G. G. L., Digiampietri, L. A., Ostroski, E. H. & Setubal, J. C.** 2014. Evaluation of graph based protein clustering methods. UNICAMP.

44. **Cursino, L., Galvani, C. D., Athinuwat, D., Zaini, P. A., Li, Y., De La Fuente, L., Hoch, H. C., Burr, T. J. & Mowery, P.** 2011. Identification of an operon, Pil-Chp, that controls twitching motility and virulence in Xylella fastidiosa. Molecular plant-microbe interactions : MPMI 24**:**1198-206.

45. **da Silva, F. R., Vettore, A. L., Kemper, E. L., Leite, A. & Arruda, P.** 2001. Fastidian gum: the *Xylella fastidiosa* exopolysaccharide possibly involved in bacterial pathogenicity. Fems Microbiology Letters 203**:**165-171.

46. **da Silva, V. S., Shida, C. S., Rodrigues, F. B., Ribeiro, D. C., de Souza, A. A., Coletta-Filho, H. D., Machado, M. A., Nunes, L. R. & de Oliveira, R. C.** 2007. Comparative genomic characterization of citrus-associated *Xylella fastidiosa* strains. BMC Genomics 8**:**474.

47. **Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T.** 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14**:**1394-403.

48. **De La Fuente, L., Parker, J. K., Oliver, J. E., Granger, S., Brannen, P. M., van Santen, E. & Cobine, P. A.** 2013. The Bacterial Pathogen Xylella fastidiosa Affects the Leaf Ionome of Plant Hosts during Infection. Plos One 8**:**9.

49. **de Souza, A. A., Takita, M. A., Coletta-Filho, H. D., Caldana, C., Goldman, G. H., Yanai, G. M., Muto, N. H., de Oliveira, R. C., Nunes, L. R. & Machado, M. A.** 2003. Analysis of gene expression in two growth states of *Xylella fastidiosa* and its relationship with pathogenicity. Mol Plant Microbe Interact 16**:**867-75.

50. **Delannoy, S., Mariani-Kurkdjian, P., Webb, H. E., Bonacorsi, S. & Fach, P.** 2017. The Mobilome; A Major Contributor to Escherichia coli stx2-Positive O26:H11 Strains Intra-Serotype Diversity. Frontiers in Microbiology 8.

51. **Della Coletta, H., Francisco, C. S., Lopes, J. R. S., De Oliveira, A. F. & Da Silva, L. F. D.** 2016. First report of olive leaf scorch in Brazil, associated with Xylella fastidiosa subsp pauca. Phytopathologia Mediterranea 55**:**130-135.

52. **Denance, N., Briand, M., Gaborieau, R., Gaillard, S. & Jacques, M. A.** 2019. Identification of genetic relationships and subspecies signatures in Xylella fastidiosa. Bmc Genomics 20**:**21.

53. **Edwards, D. J. & Holt, K. E.** 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp 3**:**2.

54. **Ellis, E. A., McEachern, G. R., Clark, S. & Cobb, B. G.** 2010. Ultrastructure of pit membrane dissolution and movement of Xylella fastidiosa through pit membranes in petioles of Vitis vinifera. Botany-Botanique 88**:**596-600.

55. **Enright, A. J., Van Dongen, S. & Ouzounis, C. A.** 2002a. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30**:**1575-84.

56. **Enright, A. J., Van Dongen, S. & Ouzounis, C. A.** 2002b. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30**:**1575-1584.

57. **Fedatto, L. M., Silva-Stenico, M. E., Etchegaray, A., Pacheco, F. T. H., Rodrigues, J. L. M. & Tsai, S. M.** 2006. Detection and characterization of protease secreted by the plant pathogen Xylella fastidiosa. Microbiological Research 161**:**263-272.

58. **Feil, H., Feil, W. S. & Lindow, S. E.** 2007. Contribution of fimbrial and afimbrial adhesins of Xylella fastidiosa to attachment to surfaces and virulence to grape. Phytopathology 97**:**318-324.

59. **Feitosa-Junior, O. R., Stefanello, E., Zaini, P. A., Nascimento, R., Pierry, P. M., Dandekar, A. M., Lindow, S. E. & da Silva, A. M.** 2019. Proteomic and Metabolomic Analyses of Xylella fastidiosa OMV-Enriched Fractions Reveal Association with Virulence Factors and Signaling Molecules of the DSF Family. Phytopathology 109**:**1344-1353.

60. **Fortier, L. C. & Sekulovic, O.** 2013. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence 4**:**354-65.

61. **Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V.** 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43**:**D261-9.

62. **Gambetta, G. A., Matthews, M. A. & Syvanen, M.** 2018. The Xylella fastidosa RTX operons: evidence for the evolution of protein mosaics through novel genetic exchanges. Bmc Genomics 19**:**11.

63. **Giampetruzzi, A., Saponari, M., Almeida, R. P. P., Essakhi, S., Boscia, D., Loconsole, G. & Saldarelli, P.** 2017a. Complete Genome Sequence of the Olive-Infecting Strain Xylella fastidiosa subsp. pauca De Donno. Microbiology Resource Announcements 5**:**2.

64. **Giampetruzzi, A., Saponari, M., Loconsole, G., Boscia, D., Savino, V. N., Almeida, R. P. P., Zicca, S., Landa, B. B., Chacon-Diaz, C. & Saldarelli, P.** 2017b. Genome-Wide Analysis Provides Evidence on the Genetic Relatedness of the Emergent Xylella fastidiosa Genotype in Italy to Isolates from Central America. Phytopathology 107**:**816-827.

65. **Godefroid, M., Cruaud, A., Streito, J. C., Rasplus, J. Y. & Rossi, J. P.** 2019. Xylella fastidiosa: climate suitability of European continent. Sci Rep 9**:**8844.

66. **Gouran, H., Gillespie, H., Nascimento, R., Chakraborty, S., Zaini, P. A., Jacobson, A., Phinney, B. S., Dolan, D., Durbin-Johnson, B. P., Antonova, E. S., Lindow, S. E., Mellema, M. S., Goulart, L. R. & Dandekar, A. M.** 2016. The Secreted Protease PrtA Controls Cell Growth, Biofilm Formation and Pathogenicity in Xylella fastidiosa. Sci Rep 6**:**31098.

67. **Guan, W., Shao, J., Davis, R. E., Zhao, T. & Huang, Q.** 2014a. Genome Sequence of a Xylella fastidiosa Strain Causing Sycamore Leaf Scorch Disease in Virginia. Genome Announc 2.

68. **Guan, W., Shao, J., Zhao, T. & Huang, Q.** 2014b. Genome Sequence of a Xylella fastidiosa Strain Causing Mulberry Leaf Scorch Disease in Maryland. Genome Announc 2.

69. **Guilhabert, M. R. & Kirkpatrick, B. C.** 2005. Identification of Xylella fastidiosa antivirulence genes: Hemagglutinin adhesins contribute to X-fastidiosa biofilm maturation and colonization and attenuate virulence. Molecular Plant-Microbe Interactions 18**:**856-868.

70. **Hatfull, G. F. & Hendrix, R. W.** 2011. Bacteriophages and their genomes. Current Opinion in Virology 1**:**298-303.

71. **Hopkins, D. L. & Purcell, A. H.** 2002. Xylella fastidiosa: Cause of Pierce's disease of grapevine and other emergent diseases. Plant Disease 86**:**1056-1066.

72. **Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B.** 2017. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. Isme Journal 11**:**1511-1520.

73. **Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón, T.** 2013. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Research 42**:**D897-D902.

74. **Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11**:**119.

75. **Ingel, B., Jeske, D. R., Sun, Q., Grosskopf, J. & Roper, M. C.** 2019. Xylella fastidiosa Endoglucanases Mediate the Rate of Pierce's Disease Development in Vitis vinifera in a Cultivar-Dependent Manner. Molecular Plant-Microbe Interactions 32**:**1402-1414.

76. **Ionescu, M., Baccari, C., Da Silva, A. M., Garcia, A., Yokota, K. & Lindow, S. E.** 2013. Diffusible signal factor (DSF) synthase RpfF of Xylella fastidiosa is a multifunction protein also required for response to DSF. J Bacteriol 195**:**5273-84.

77. **Ionescu, M., Yokota, K., Antonova, E., Garcia, A., Beaulieu, E., Hayes, T., Iavarone, A. T. & Lindow, S. E.** 2016. Promiscuous Diffusible Signal Factor Production and Responsiveness of the Xylella fastidiosa Rpf System. MBio 7.

78. **Jacques, M. A., Denance, N., Legendre, B., Morel, E., Briand, M., Mississipi, S., Durand, K., Olivier, V., Portier, P., Poliakoff, F. & Crouzillat, D.** 2016. New Coffee Plant-Infecting Xylella fastidiosa Variants Derived via Homologous Recombination. Applied and Environmental Microbiology 82**:**1556-1568.

79. **Janse, J. D. & Obradovic, A.** 2010. Xylella Fastidiosa: Its Biology, Diagnosis, Control and Risks. Journal of Plant Pathology 92**:**S35-S48.

80. **Killiny, N. & Almeida, R. P.** 2009. Xylella fastidiosa afimbrial adhesins mediate cell transmission to plants by leafhopper vectors. Appl Environ Microbiol 75**:**521-8.

81. **Killiny, N. & Almeida, R. P.** 2011. Gene regulation mediates host specificity of a bacterial pathogen. Environ Microbiol Rep 3**:**791-7.

82. **Killiny, N. & Almeida, R. P. P.** 2014. Factors Affecting the Initial Adhesion and Retention of the Plant Pathogen Xylella fastidiosa in the Foregut of an Insect Vector. Applied and Environmental Microbiology 80**:**420-426.

83. **Koide, T., Zaini, P. A., Moreira, L. M., Vencio, R. Z. N., Matsukuma, A. Y., Durham, A. M., Teixeira, D. C., El-Dorry, H., Monteiro, P. B., da Silva, A. C. R., Verjovski-Almeida, S., da Silva, A. M. & Gomes, S. L.** 2004. DNA microarray-based genome comparison of a pathogenic and a nonpathogenic strain of Xylella fastidiosa delineates genes important for bacterial virulence. Journal of Bacteriology 186**:**5442-5449.

84. **Koonin, E. V.** 2010. The wonder world of microbial viruses. Expert Rev Anti Infect Ther 8**:**1097-9.

85. **Kubicek, C. P., Starr, T. L. & Glass, N. L.** 2014. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. Annu Rev Phytopathol 52**:**427-51.

86. **Kung, S. H., Retchless, A. C., Kwan, J. Y. & Almeida, R. P. P.** 2013. Effects of DNA Size on Transformation and Recombination Efficiencies in Xylella fastidiosa. Applied and Environmental Microbiology 79**:**1712-1717.

87. **Kutschera, A. & Ranf, S.** 2019. The multifaceted functions of lipopolysaccharide in plant-bacteria interactions. Biochimie 159**:**93-98.

88. **Kutter, E. & Sulakvelidze, A.** 2005. Bacteriophages: Biology and Applications. CRC Press, Boca Raton, FL.

89. **Lacava, P. T., Andreote, F. D., Araujo, W. L. & Azevedo, J. L.** 2006. Characterization of the endophytic bacterial community from citrus by isolation, specific PCR and DGGE. Pesquisa Agropecuaria Brasileira 41**:**637-642.

90. **Lambais, M. R., Goldman, M. H., Camargo, L. E. & Goldman, G. H.** 2000. A genomic approach to the understanding of *Xylella fastidiosa* pathogenicity. Curr Opin Microbiol 3**:**459-62.

91. **Li, W. B., Zreik, L., Fernandes, N. G., Miranda, V. S., Teixeira, D. C., Ayres, A. J., Garnier, M. & Bove, J. M.** 1999. A triply cloned strain of Xylella fastidiosa multiplies and induces symptoms of citrus variegated chlorosis in sweet orange. Current Microbiology 39**:**106-108.

92. **Li, Y. X., Hao, G. X., Galvani, C. D., Meng, Y. Z., De la Fuente, L., Hoch, H. C. & Burr, T. J.** 2007. Type I and type IV pili of Xylella fastidiosa affect twitching motility, biofilm formation and cell-cell aggregation. Microbiology-Sgm 153**:**719-726.

93. **Lindow, S.** 2019. Money Matters: Fueling Rapid Recent Insight Into Xylella fastidiosa-An Important and Expanding Global Pathogen. Phytopathology 109**:**210-212.

94. **Llop, P.** 2015. Genetic islands in pome fruit pathogenic and non-pathogenic Erwinia species and related plasmids. Front Microbiol 6**:**874.

95. **Lopes, J. R. S., Daugherty, M. P. & Almeida, R. P. P.** 2010. Strain origin drives virulence and persistence of Xylella fastidiosa in alfalfa. Plant Pathology 59**:**963-971.

96. **Lukjancenko, O., Thomsen, M. C., Voldby Larsen, M. & Ussery, D. W.** 2013. PanFunPro: PAN-genome analysis based on FUNctional PROfiles [version 1; peer review: 3 approved with reservations]. . F1000Research 2**:**16.

97. **Marcelletti, S. & Scortichini, M.** 2016a. Genome-wide comparison and taxonomic relatedness of multiple Xylella fastidiosa strains reveal the occurrence of three subspecies and a new Xylella species. Arch Microbiol 198**:**803-12.

98. **Marcelletti, S. & Scortichini, M.** 2016b. Xylella fastidiosa CoDiRO strain associated with the olive quick decline syndrome in southern Italy belongs to a clonal complex of the subspecies pauca that evolved in Central America. Microbiology 162**:**2087-2098.

99. **Martelli, G. P., Boscia, D., Porcelli, F. & Saponari, M.** 2016. The olive quick decline syndrome in south-east Italy: a threatening phytosanitary emergency. European Journal of Plant Pathology 144**:**235-243.

100. **Matsumoto, A., Huston, S. L., Killiny, N. & Igo, M. M.** 2012. XatA, an AT-1 autotransporter important for the virulence of Xylella fastidiosa Temecula1. MicrobiologyOpen 1**:**33-45.

101. **McCann, H. C. & Guttman, D. S.** 2007. Evolution of the type III secretion system and its effectors in plant-microbe interactions. New Phytologist 177**:**33-47.

102. **Mendes, J. S., Santiago, A. S., Toledo, M. A. S., Horta, M. A. C., de Souza, A. A., Tasic, L. & de Souza, A. P.** 2016. In vitro Determination of Extracellular Proteins from Xylella fastidiosa. Frontiers in Microbiology 7**:**15.

103. **Mi, H., Muruganujan, A. & Thomas, P. D.** 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41**:**D377-86.

104. **Minh, B. Q., Nguyen, M. A. & von Haeseler, A.** 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30**:**1188-95.

105. **Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P. *et al.*** 2018. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research 47**:**D351-D360.

106. **Monteiro-Vitorello, C. B., De Oliveira, M. C., Zerillo, M. M., Varani, A. M., Civerolo, E. & Van Sluys, M. A.** 2005. Xylella and Xanthomonas Mobil'omics. Omics-a Journal of Integrative Biology 9**:**146-159.

107. **Nakamura, Y.** 2018. Prediction of Horizontally and Widely Transferred Genes in Prokaryotes. Evol Bioinform Online 14**:**1176934318810785.

108. **Nascimento, R., Gouran, H., Chakraborty, S., Gillespie, H. W., Almeida-Souza, H. O., Tu, A., Rao, B. J., Feldstein, P. A., Bruening, G., Goulart, L. R. & Dandekar, A. M.** 2016. The Type II Secreted Lipase/Esterase LesA is a Key Virulence Factor Required for Xylella fastidiosa Pathogenesis in Grapevines. Scientific Reports 6**:**18598.

109. **Newman, K. L., Almeida, R. P., Purcell, A. H. & Lindow, S. E.** 2003. Use of a green fluorescent strain for analysis of *Xylella fastidiosa* colonization of *Vitis vinifera*. Appl Environ Microbiol 69**:**7319-27.

110. **Newman, K. L., Almeida, R. P., Purcell, A. H. & Lindow, S. E.** 2004. Cell-cell signaling controls *Xylella fastidiosa* interactions with both insects and plants. Proc Natl Acad Sci U S A 101**:**1737-42.

111. **Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q.** 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32**:**268-74.

112. **Nunes, L. R., Rosato, Y. B., Muto, N. H., Yanai, G. M., da Silva, V. S., Leite, D. B., Goncalves, E. R., de Souza, A. A., Coletta-Filho, H. D., Machado, M. A., Lopes, S. A. & de Oliveira, R. C.** 2003. Microarray analyses of *Xylella fastidiosa* provide evidence of coordinated transcription control of laterally transferred elements. Genome Res 13**:**570-8.

113. **Oliveira Alvarenga, D., Moreira, L. M., Chandler, M. & Varani, A. M.** 2018. A Practical Guide for Comparative Genomics of Mobile Genetic Elements in Prokaryotic Genomes, p. 213-242. *In* J. C. Setubal, J. Stoye & P. F. Stadler (ed.), Comparative Genomics: Methods and Protocols. Springer New York, New York, NY.

114. **Oliver, J. E., Cobine, P. A. & De La Fuente, L.** 2015. Xylella fastidiosa Isolates from Both subsp. multiplex and fastidiosa Cause Disease on Southern Highbush Blueberry (Vaccinium sp.) Under Greenhouse Conditions. Phytopathology 105**:**855-62.

115. **Paetzel, M.** 2019. Bacterial Signal Peptidases, p. 187-219. *In* A. Kuhn (ed.), Bacterial Cell Walls and Membranes. Springer International Publishing, Cham.

116. **Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A. & Parkhill, J.** 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31**:**3691-3.

117. **Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.** 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25**:**1043-55.

118. **Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O.** 2018. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clinical Microbiology Reviews 31**:**e00088-17.

119. **Pashalidis, S., Moreira, L. M., Zaini, P. A., Campanharo, J. C., Alves, L. M. C., Ciapina, L. P., Vencio, R. Z. N., Lemos, E. G. M., Da Silva, A. M. & da Silva, A. C. R.** 2005. Whole-genome expression profiling of Xylella fastidiosa in response to growth on glucose. Omics-a Journal of Integrative Biology 9**:**77-90.

120. **Perez-Donoso, A. G., Sun, Q., Roper, M. C., Greve, L. C., Kirkpatrick, B. & Labavitch, J. M.** 2010. Cell Wall-Degrading Enzymes Enlarge the Pore Size of Intervessel Pit Membranes in Healthy and Xylella fastidiosa-Infected Grapevines. Plant Physiology 152**:**1748-1759.

121. **Purcell, A. H. & Hopkins, D. L.** 1996. Fastidious xylem-limited bacterial plant pathogens. Annual Review of Phytopathology 34**:**131-151.

122. **Ramazzotti, M., Cimaglia, F., Gallo, A., Ranaldi, F., Surico, G., Mita, G., Bleve, G. & Marchi, G.** 2018. Insights on a founder effect: the case of Xylella fastidiosa in the Salento area of Apulia, Italy. Phytopathologia Mediterranea 57**:**8-25.

123. **Rapicavoli, J., Ingel, B., Blanco-Ulate, B., Cantu, D. & Roper, C.** 2018a. Xylella fastidiosa: an examination of a re-emerging plant pathogen. Molecular Plant Pathology 19**:**786-800.

124. **Rapicavoli, J. N., Blanco-Ulate, B., Muszynski, A., Figueroa-Balderas, R., Morales-Cruz, A., Azadi, P., Dobruchowska, J. M., Castro, C., Cantu, D. & Roper, M. C.** 2018b. Lipopolysaccharide O-antigen delays plant innate immune recognition of Xylella fastidiosa. Nature Communications 9**:**12.

125. **Robinson, D. G., Lee, M. C. & Marx, C. J.** 2012. OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. Nucleic Acids Res 40**:**e174.

126. **Rogers, E. E. & Backus, E. A.** 2014. Anterior Foregut Microbiota of the Glassy-Winged Sharpshooter Explored Using Deep 16S rRNA Gene Sequencing from Individual Insects. Plos One 9**:**9.

127. **Roper, M. C., Greve, L. C., Labavitch, J. M. & Kirkpatrick, B. C.** 2007a. Detection and visualization of an exopolysaccharide produced by *Xylella fastidiosa* in vitro and in planta. Appl Environ Microbiol 73**:**7252-8.

128. **Roper, M. C., Greve, L. C., Warren, J. G., Labavitch, J. M. & Kirkpatrick, B. C.** 2007b. *Xylella fastidiosa* requires polygalacturonase for colonization and pathogenicity in *Vitis vinifera* grapevines. Molecular Plant-Microbe Interactions 20**:**411-419.

129. **Rossetti, V. & De Negri, J. D.** 1990. Clorose variegada dos citros: revisão. Laranja 11**:**1-14.

130. **Rossetti, V., Garnier, M., Bove, J. M., Beretta, M. J. G., Teixeira, A. R. R., Quaggio, J. A. & Denegri, J. D.** 1990. Occurrence of xylem-restricted bacteria in sweet orange trees affected by chlorotic variegation, a new citrus disease in Brazil. Comptes Rendus de L Academie Des Sciences 310**:**345-349.

131. **Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B.** 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3**:**e985.

132. **Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P. B., Cheng, J. F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C. & Eloe-Fadrosh, E. A.** 2019. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat Microbiol.

133. **Santiago, C., Pereira, V. & Digiampietri, L.** 2018. Homology Detection Using Multilayer Maximum Clustering Coefficient. J Comput Biol.

134. **Santiago, C. R. N., Assis, R. A. B., Moreira, L. M. & Digiampietri, L. A.** 2019. Gene Tags Assessment by Comparative Genomics (GTACG): A User-Friendly Framework for Bacterial Comparative Genomics. Frontiers in Genetics 10.

135. **Schmid, M., Muri, J., Melidis, D., Varadarajan, A. R., Somerville, V., Wicki, A., Moser, A., Bourqui, M., Wenzel, C., Eugster-Meier, E., Frey, J. E., Irmler, S. & Ahrens, C. H.** 2018. Comparative Genomics of Completely Sequenced Lactobacillus helveticus Genomes Provides Insights into Strain-Specific Genes and Resolves Metagenomics Data Down to the Strain Level. Frontiers in Microbiology 9.

136. **Schreiber, H. L., Koirala, M., Lara, A., Ojeda, M., Dowd, S. E., Bextine, B. & Morano, L.** 2010. Unraveling the First Xylella fastidiosa subsp fastidiosa Genome from Texas. Southwestern Entomologist 35**:**479-483.

137. **Schuenzel, E. L., Scally, M., Stouthamer, R. & Nunney, L.** 2005. A multigene phylogenetic study of clonal diversity and divergence in North American strains of the plant pathogen *Xylella fastidiosa*. Appl Environ Microbiol 71**:**3832-9.

138. **Secor, P. R., Michaels, L. A., Smigiel, K. S., Rohani, M. G., Jennings, L. K., Hisert, K. B., Arrigoni, A., Braun, K. R., Birkland, T. P., Lai, Y., Hallstrand, T. S., Bollyky, P. L., Singh, P. K. & Parks, W. C.** 2017. Filamentous Bacteriophage Produced by Pseudomonas aeruginosa Alters the Inflammatory Response and Promotes Noninvasive Infection In Vivo. Infect Immun 85.

139. **Seemann, T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30**:**2068-2069.

140. **Segerman, B.** 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Frontiers in cellular and infection microbiology 2**:**116-116.

141. **Setubal, J. C., Almeida, N. F. & Wattam, A. R.** 2018. Comparative Genomics for Prokaryotes, p. 55-78. *In* J. C. Setubal, J. Stoye & P. F. Stadler (ed.), Comparative Genomics: Methods and Protocols. Springer New York, New York, NY.

142. **Shriner, A. D. & Andersen, P. C.** 2014. Effect of oxygen on the growth and biofilm formation of Xylella fastidiosa in liquid media. Curr Microbiol 69**:**866-73.

143. **Sicard, A., Zeilinger, A. R., Vanhove, M., Schartel, T. E., Beal, D. J., Daugherty, M. P. & Almeida, R. P. P.** 2018. Xylella fastidiosa: Insights into an Emerging Plant Pathogen. Annu Rev Phytopathol.

144. **Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D. & Higgins, D. G.** 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7**:**539.

145. **Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M. *et al.*** 2000. The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406**:**151-7.

146. **Soares, S. C., Geyik, H., Ramos, R. T., de Sa, P. H., Barbosa, E. G., Baumbach, J., Figueiredo, H. C., Miyoshi, A., Tauch, A., Silva, A. & Azevedo, V.** 2016. GIPSy: Genomic island prediction software. J Biotechnol 232**:**2-11.

147. **Summer, E. J., Enderle, C. J., Ahern, S. J., Gill, J. J., Torres, C. P., Appel, D. N., Black, M. C., Young, R. & Gonzalez, C. F.** 2010. Genomic and Biological Analysis of Phage Xfas53 and Related Prophages of Xylella fastidiosa. Journal of Bacteriology 192**:**179-190.

148. **Sundin, G. W.** 2007. Genomic insights into the contribution of phytopathogenic bacterial plasmids to the evolutionary history of their hosts. Annu Rev Phytopathol 45**:**129-51.

149. **Tatusov, R. L., Koonin, E. V. & Lipman, D. J.** 1997. A genomic perspective on protein families. Science 278**:**631-7.

150. **Tettelin, H., Riley, D., Cattuto, C. & Medini, D.** 2008. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11**:**472-7.

151. **Touchon, M., Bernheim, A. & Rocha, E. P.** 2016. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J 10**:**2744-2754.

152. **van der Veen, B. E., Harris, H. M., O'Toole, P. W. & Claesson, M. J.** 2014. Metaphor: finding bi-directional best hit homology relationships in (meta)genomic datasets. Genomics 104**:**459-63.

153. **Van Horn, C., Chang, C. J. & Chen, J. C.** 2017. De Novo Whole-Genome Sequence of Xylella fastidiosa subsp. multiplex Strain BB01 Isolated from a Blueberry in Georgia, USA. Microbiology Resource Announcements 5**:**2.

154. **Van Sluys, M. A., de Oliveira, M. C., Monteiro-Vitorello, C. B., Miyaki, C. Y., Furlan, L. R. *et al.*** 2003. Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J Bacteriol 185**:**1018-26.

155. **Van Sluys, M. A., Monteiro-Vitorello, C. B., Camargo, L. E., Menck, C. F., Da Silva, A. C., Ferro, J. A., Oliveira, M. C., Setubal, J. C., Kitajima, J. P. & Simpson, A. J.** 2002. Comparative genomic analysis of plant-associated bacteria. Annu Rev Phytopathol 40**:**169-89.

156. **Vanhove, M., Retchless, A. C., Sicard, A., Rieux, A., Coletta-Filho, H. D., De La Fuente, L., Stenger, D. C. & Almeida, R. P. P.** 2019. Genomic Diversity and Recombination among Xylella fastidiosa Subspecies. Appl Environ Microbiol 85.

157. **Varani, A. M., Monteiro-Vitorello, C. B., de Almeida, L. G. P., Souza, R. C., Cunha, O. L., Lima, W. C., Civerolo, E., Van Sluys, M. A. & Vasconcelos, A. T. R.** 2012. Xylella fastidiosa comparative genomic database is an information resource to explore the annotation, genomic features, and biology of different strains. Genetics and Molecular Biology 35**:**149-152.

158. **Varani, A. M., Souza, R. C., Nakaya, H. I., de Lima, W. C., Paula de Almeida, L. G., Kitajima, E. W., Chen, J., Civerolo, E., Vasconcelos, A. T. & Van Sluys, M. A.** 2008. Origins of the *Xylella fastidiosa* prophage-like regions and their impact in genome differentiation. PLoS ONE 3**:**e4059.

159. **Vernikos, G. S. & Parkhill, J.** 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22**:**2196-203.

160. **Voegel, T. M., Warren, J. G., Matsumoto, A., Igo, M. M. & Kirkpatrick, B. C.** 2010. Localization and characterization of Xylella fastidiosa haemagglutinin adhesins. Microbiology-Sgm 156**:**2172-2179.

161. **Wang, P., Lee, Y., Igo, M. M. & Roper, M. C.** 2017. Tolerance to oxidative stress is required for maximal xylem colonization by the xylem-limited bacterial phytopathogen, Xylella fastidiosa. Molecular Plant Pathology 18**:**990-1000.

162. **Watts, D. J. & Strogatz, S. H.** 1998. Collective dynamics of 'small-world' networks. Nature 393**:**440-442.

163. **Wells, J. M., Raju, B. C., Hung, H. Y., Weisburg, W. G., Mandelcopaul, L. & Brenner, D. J.** 1987. Xylella-Fastidiosa Gen-Nov, Sp-Nov - Gram-Negative, Xylem-Limited, Fastidious Plant Bacteria Related to Xanthomonas-Spp. International Journal of Systematic Bacteriology 37**:**136-143.

164. **Xie, Z. & Tang, H.** 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics 33**:**3340-3347.

165. **Yu, J., Blom, J., Glaeser, S. P., Jaenicke, S., Juhre, T., Rupp, O., Schwengers, O., Spanig, S. & Goesmann, A.** 2017. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. J Biotechnol 261**:**2-9.

166. **Zaini, P. A., Burdman, S., Igo, M. M., Parker, J. K. & De La Fuente, L.** 2015. Fimbrial and Afimbrial Adhesins Involved in Bacterial Attachment to Surfaces, p. 492. *In* N. Wang, J. B. Jones, G. W. Sundin, F. F. White, S. A. Hogenhout, C. Roper, L. De La Fuente & J. H. Ham (ed.), Virulence mechanisms of plant-

pathogenic bacteria. American Phytopathological Society Press, St. Paul, Minnesota, EUA.

167. **Zaini, P. A., Nascimento, R., Gouran, H., Cantu, D., Chakraborty, S., Phu, M., Goulart, L. R. & Dandekar, A. M.** 2018. Molecular Profiling of Pierce's Disease Outlines the Response Circuitry of Vitis vinifera to Xylella fastidiosa Infection. Frontiers in Plant Science 9**:**16.

168. **Zekic, T., Holley, G. & Stoye, J.** 2018. Pan-Genome Storage and Analysis Techniques, p. 29-53. *In* J. C. Setubal, J. Stoye & P. F. Stadler (ed.), Comparative Genomics: Methods and Protocols. Springer New York, New York, NY.

169. **Zhang, S., Flores-Cruz, Z., Kumar, D., Chakrabarty, P., Hopkins, D. L. & Gabriel, D. W.** 2011. The Xylella fastidiosa biocontrol strain EB92-1 genome is very similar and syntenic to Pierce's disease strains. Journal of Bacteriology 193**:**5576-7.

170. **Zhao, Y., Sun, C., Zhao, D., Zhang, Y., You, Y., Jia, X., Yang, J., Wang, L., Wang, J., Fu, H., Kang, Y., Chen, F., Yu, J., Wu, J. & Xiao, J.** 2018. PGAP-X: extension on pan-genome analysis pipeline. BMC Genomics 19**:**36.

171. **Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S.** 2011. PHAST: a fast phage search tool. Nucleic Acids Res 39**:**W347-52.

8. Supplementary information

Table S1: Brief description of the Bioinformatics software used in this work

Bioinformatics software	Description	Reference
CheckM	Provides a set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes. CheckM assesses the quality of genomes recovered from isolates, single cells, or metagenomes estimating the completeness and contamination by using collocated sets of genes that are ubiquitous and single-copy within a phylogenetic lineage.	Parks et al., 2015
Prokka	Command line software tool to annotate bacterial genome in two stages. First, prokka identifies the coordinates of candiate gene (Prodigal in the case of CDSs). Then, each gene is comparing with a database in hierarchal manner, starting with a smaller trustworthy database, moving to medium-sized but domain-specific databases, and finally to curated models of protein families.	Seemann, 2014
GTACG	Tool for identifying in the subgroups of bacterial genomes whose microorganisms have common phenotypic characteristic, to find data that differentiates them from other associated genomes in a simple and fast way. The GTACG analysis is bases on the formation of homologous CDS clusters from local alignments.	Santiago et al., 2019
IQ-Tree	IQ-TREE is a time and search efficient ML-tree reconstruction program that uses the combination of hill-climbing approaches and a stochastic perturbation method to reduce the computation time efficiently.	Nguyen et al., 2015
PHASTER	An upgrade of PHAST which identify and annotate prophages in bacterial and plasmids. PHAST uses GLIMMER to the identification of ORF and BLAST for the protein identification (by homology). Therefore, BLAST is used to phage sequence identification (phage-specific sequence database), tRNA identification, attachment site recognition and gene clustering density measurements using density-based spatial clustering of applications with noise (DBSCAN) and sequence annotation text mining.	Arndt et al., 2016
PhiSpy	It was developed for identifying prophages using five distinctive similarity-agnostic characteristics: protein length, transcription strand directionality, customized AT and GC skew, and the abundance of unique phage DNA sequence words. Among the metrics used, random forest classification algorithm predicts the prophages by ranking genomic regions based on those characteristics.	Akhter et al., 2012
VirSorter	Tool designed to detect viral signal in different types of microbial sequence data (such as metagenomic sequencing) in both a reference-dependent and reference-independent manner, leveraging probabilistic models and extensive virome data to maximize detection of novel virus.	Roux et al., 2015
Inovirus_detector	It is a set of script can be used to identify putative inovirus sequences in draft genome assemblies or metagenome assemblies. Frist, de-novo identification of putative inovirus sequences is performed. To this, the script performs a custom and simple annotation for 30kbp windows around the putative pI-like protein(s). Then, the prediction is refine using random forest classifier, and detecting canonical and non-canonical attachment (att) site.	Roux et al., 2019
Alien Hunter	Tool for the prediction of putative Horizontal Gene Tranfer (HGT) events with the implementation of Interpolated Variable Order Motifs (IVOM) which is a method that exploits compositional biases at various levels (e.g. codon, dinucleotide and aminoacid bias, structural constraints) by implementing variable order motif distributions.	Vernikos & Parkhill, 2006
SeqWord Sniffer	A tool that examines variations in frequencies of oligonucleotides and traces down the distributions of mobile genomic elements across genomes by analyzing the patterns of 4-bases long words.	Bezuidt et al., 2009
GIPSy	A genomic island prediction software, a standalone and user-friendly software for the prediction of GEIs, built on our previously developed pathogenicity island prediction software (PIPS)	Soares et al., 2016
Oasis	This software groups already-annotated transposase genes that could compose multiple copies of an IS by length and sequence similarity. Those that fit a high similarity threshold are assumed to be in the same group (multi-copy ISs). In the case of isolated transposase in the genome, inverted repeat sequences around the transposases are found and composed single-copy ISs. Multi-copy ISs are also checked for inverted repeats sequences. Once all groups of ISs are identified, hierarchal agglomerative clustering is used to defined unique IS sets. Finally, the classification of IS families is performed using BLASTp against ISFinder database.	Robinson et al., 2012
ISEScan	The ISEScan pipeline consists of the following steps: (i) predicting protein coding genes in the input genome sequence and translating them into protein sequences, (ii) identifying putative transposases by searching the predicted proteins against the library of profile HMMs (based on the transposase sequences in the curated ACLAME dataset), (iii) extending the putative transposase genes into full-length IS elements by locating inverted repeat sequences in the upstream and downstream regions and (iv) refining and reporting the final set of annotated IS elements in the input genome	Xie & Tang, 2017

Table S2: Genome assembly statistics

Strain	Genome Size	# contigs	Largest contig	N50	N75	L50	L75
3124	2.748.594	1	2.748.594	2.748.594	2.748.594	1	1
9a5c	2.731.750	3	2.679.306	2.679.306	2.679.306	1	1
Ann-1	2.780.908	2	2.750.603	2.750.603	2.750.603	1	1
De-Donno	2.543.738	2	2.508.465	2.508.465	2.508.465	1	1
Fb7	2.699.320	2	2.659.912	2.659.912	2.659.912	1	1
GB514	2.517.383	2	2.491.203	2.491.203	2.491.203	1	1
Hib4	2.877.548	2	2.813.297	2.813.297	2.813.297	1	1
J1a12	2.867.237	3	2.788.789	2.788.789	2.788.789	1	1
M12	2.475.130	1	2.475.130	2.475.130	2.475.130	1	1
M23	2.573.987	2	2.535.690	2.535.690	2.535.690	1	1
MUL0034	2.666.577	2	2.642.186	2.642.186	2.642.186	1	1
Pr8x	2.705.822	2	2.666.242	2.666.242	2.666.242	1	1
Salento-1	2.543.366	2	2.508.097	2.508.097	2.508.097	1	1
Salento-2	2.543.566	2	2.508.296	2.508.296	2.508.296	1	1
Temecula1	2.521.148	2	2.519.802	2.519.802	2.519.802	1	1
U24D	2.732.490	2	2.681.334	2.681.334	2.681.334	1	1
32	2.607.546	56	258.764	104.769	79.083	9	16
11399	2.736.060	35	331.284	183.019	86.822	6	12
6c	2.603.975	19	801.503	264.629	117.733	3	7
ATCC35871	2.416.255	58	252.088	106.692	73.902	7	13
ATCC35879	2.522.328	16	770.940	500.648	372.305	2	4
B111	2.682.276	77	254.258	130.109	55.367	8	15
BB01	2.511.521	83	197.455	96.868	74.967	10	17
CFBP8072	2.496.662	184	224.836	87.332	46.649	11	21
CFBP8073	2.582.150	212	247.073	67.596	39.060	13	24
CFBP8416	2.466.748	89	304.552	113.891	70.159	6	13
CFBP8417	2.504.981	141	229.226	118.419	57.355	8	16
CFBP8418	2.513.969	149	229.246	96.035	56.940	9	17
CO33	2.681.926	68	406.234	176.135	98.074	6	13
CoDiRO	2.542.932	12	678.618	375.307	224.376	3	5
COF0324	2.772.556	135	267.945	100.091	50.933	9	18
COF0407	2.538.474	133	176.368	56.951	32.728	15	29
CVC0251	2.740.246	128	255.920	135.650	54.575	7	15
CVC0256	2.702.144	122	340.191	116.891	57.551	7	15
Dixon	2.622.328	32	703.569	201.792	84.131	4	9
DSM10026	2.431.652	63	260.199	103.747	57.986	9	16
EB92-1	2.475.426	124	149.098	79.650	49.093	12	21
Griffin-1	2.387.314	84	142.581	74.017	48.205	12	21
IVIA5235	2.491.574	90	399.902	103.359	57.832	7	14
Mullberry	2.667.719	3	2.642.186	2.642.186	2.642.186	1	1
Mul-MD	2.520.555	101	395.385	134.146	57.693	6	14
OLS0478	2.555.411	45	391.174	246.999	97.240	4	8
OLS0479	2.539.963	130	218.420	68.456	35.530	11	23
Stags-Leap	2.510.798	15	731.756	559.515	393.232	2	4
Sy-VA	2.475.880	75	342.206	119.240	87.370	7	13
XRB	2.714.643	74	263.947	138.935	90.512	8	14

Software used: Quast (http://quast.bioinf.spbau.ru/)

Figure S1: Comparative genomics using other software

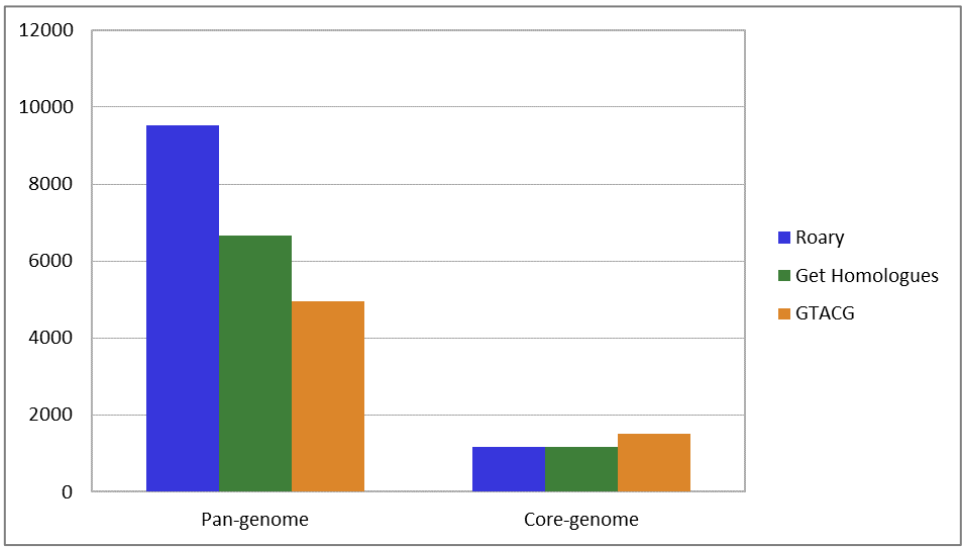

Table S3: Orthologs related to the virulence found in the core-genome

Prokka annotation locus	NCBI annotation locus	Annotation	Num. Sequences per genome
CDS related to pilus and fimbrial adhesins			
Xf_9a5c_00424	XF_RS01965	fimbrial assembly protein/fimbrial protein	1.00
Xf_9a5c_00330	XF_RS01555	fimbrial assembly protein/fimbrial protein	1.00
Xf_9a5c_00329	XF_RS01550	fimbrial assembly protein/pilus assembly protein PilM	1.00
Xf_9a5c_00071	XF_RS00335	fimbrial biogenesis outer membrane usher protein	1.00
Xf_9a5c_00425	XF_RS01970	fimbrial protein	1.00
Xf_9a5c_00332	XF_RS01565	fimbrial protein	1.00
Xf_9a5c_00331	XF_RS01560	fimbrial protein	1.00
Xf_9a5c_00068	XF_RS00325	fimbrial protein/type 1 fimbrial protein	2.09
Xf_9a5c_00027	XF_RS00125	hypothetical protein/pilus assembly protein PilW	1.00
Xf_9a5c_00876	not in ncbi	hypothetical protein/pilus assembly protein/prepilin-type cleavage/methylation domain-containing protein/type 4 fimbrial biogenesis protein	1.00
Xf_9a5c_00421	XF_RS01950	hypothetical protein/pre-pilin like leader sequence/prepilin-type cleavage/methylation domain-containing protein	1.00
Xf_9a5c_00025	XF_RS00115	hypothetical protein/prepilin	1.00
Xf_9a5c_00072	XF_RS00340	molecular chaperone/molecular chaperone, partial/pilus assembly protein PapD	1.00
Xf_9a5c_00030	XF_RS00140	PilE protein/hypothetical protein/pilus assembly protein PilE	1.00
Xf_9a5c_02364	XF_RS11055	pilin/prepilin-type cleavage/methylation domain-containing protein/prepilin-type cleavage/methylation domain-containing protein, partial	2.00
Xf_9a5c_01477	XF_RS06915	PilT/PilU family type 4a pilus ATPase/twitching motility protein PilT/type IV pili twitching motility protein PilT	1.02
Xf_9a5c_00423	XF_RS01960	pilus assembly protein PilW	1.00
Xf_9a5c_00028	XF_RS00130	PilX protein	1.00
Xf_9a5c_01362	XF_RS06380	PilZ domain-containing protein	1.00
Xf_9a5c_02362	XF_RS11045	prepilin peptidase	1.00
Xf_9a5c_00073	XF_RS00345	type 1 fimbrial protein	1.00
Xf_9a5c_01478	XF_RS06920	type IV pili twitching motility protein PilT	1.02
Xf_9a5c_00426	XF_RS01975	type IV pilin protein	1.00
Xf_9a5c_00608	XF_RS02820	type IV pilus assembly PilZ	1.00
Xf_9a5c_00411	XF_RS01905	type IV pilus biogenesis/stability protein PilW	1.00
Xf_9a5c_00026	XF_RS00120	type IV pilus modification protein PilV	1.00
Xf_9a5c_00422	XF_RS01955	type IV pilus modification protein PilV	1.00
Xf_9a5c_00333	XF_RS01570	type IV pilus secretin PilQ	1.17
Xf_9a5c_02369	XF_RS11075	type IV-A pilus assembly ATPase PilB	1.00
CDS related to afimbrial adhesins			
Xf_9a5c_01143	not in ncbi	autotransporter/autotransporter domain-containing protein/hypothetical protein/outer membrane autotransporter barrel	1.15
Xf_9a5c_02187	XF_RS10195	autotransporter/autotransporter domain-containing protein	1.04
Xf_9a5c_00710	XF_RS03320	autotransporter domain-containing esterase/autotransporter domain-containing protein	1.00
Xf_9a5c_01714	XF_RS08015	autotransporter domain-containing protein/autotransporter domain-containing protein, partial/serine protease/serine protease, partial	3.11
Xf_9a5c_00812	XF_RS03825	hemagglutinin/hemagglutinin, partial/hemagglutinin-like protein, partial/hemagglutinin-related protein/hypothetical protein	15.91
Xf_9a5c_02057	XF_RS09565	hemagglutinin	1.00
Xf_9a5c_01381	XF_RS06465	adhesin/adhesin (XadA2)/cell surface protein/surface protein/surface protein (XadA2)	1.02
Xf_9a5c_01368	XF_RS06405	hypothetical protein/membrane protein/membrane protein (XadA1)	1.00
Xf_9a5c_01847	XF_RS08585	adhesin/autotransporter adhesin/autotransporter adhesin (XadA3)	1.13

Table S3: Orthologs related to the virulence found in the core-genome (*continuation*)

CDS related to hemolysins			
Xf_9a5c_02563	XF_RS11960	hemolysin/peptidase M23	1.41
Xf_9a5c_00273	XF_RS01295	hemolysin D	1.43
Xf_9a5c_02375	XF_RS11105	ShlB/FhaC/HecB family hemolysin secretion/activation protein	1.00
Xf_9a5c_01100	XF_RS05165	hemolysin D	1.00
Xf_9a5c_00159	XF_RS00750	HylII/hemolysin D/hemolysin III	1.00
CDS related to CWDEs with evidence to be secreted			
Xf_9a5c_02510	XF_RS11735	cellulase	1.00
Xf_9a5c_00734	XF_RS03440	endoglucanase	1.00
Xf_9a5c_00759	XF_RS03550	beta-galactosidase	1.00
Xf_9a5c_00763	XF_RS03570	beta-glucosidase/glycoside hydrolase family 3	1.00
Xf_9a5c_00392	XF_RS01820	beta-glucosidase/beta-glucosidase BglX	1.00
Xf_9a5c_02288	not in ncbi	polygalacturonase	1.00
CDS related to gum production			
Xf_9a5c_02202	XF_RS10270	polysaccharide biosynthesis protein GumD/undecaprenyl-phosphate glucose phosphotransferase	1.02
Xf_9a5c_02200	XF_RS10260	GumF protein/acyltransferase/acyltransferase 3/hypothetical protein/polysaccharide biosynthesis protein GumF	1.00
Xf_9a5c_02201	XF_RS10265	polysaccharide biosynthesis protein GumE	1.00
Xf_9a5c_02203	XF_RS10275	GumC protein/LPS biosynthesis protein/exopolysaccharide biosynthesis protein/polysaccharide biosynthesis protein GumC	1.00
Xf_9a5c_02204	not in ncbi	polysaccharide biosynthesis protein GumB/polysaccharide transporter	1.00
CDS related to lipases with evidence to be secreted			
Xf_9a5c_02013	XF_RS09325	alpha/beta hydrolase/lipase	1.00
Xf_9a5c_00319	XF_RS01510	alpha/beta hydrolase/alpha/beta hydrolase, partial/lipase	1.54
Xf_9a5c_00986	XF_RS04630	cardiolipin synthase/phospholipase D family protein	1.00
CDS related to proteases with evidence to be secreted			
Xf_9a5c_01021	XF_RS04785	metalloprotease PmbA	1.00
Xf_9a5c_01067	XF_RS05000	ATP-binding protein/ATP-dependent protease	1.00
Xf_9a5c_01075	XF_RS05040	ATP-dependent Clp endopeptidase, proteolytic subunit ClpP/ATP-dependent Clp protease proteolytic subunit	1.00
Xf_9a5c_01688	XF_RS07895	peptidase S41/tail-specific protease	1.00
Xf_9a5c_02095	XF_RS09755	DegQ family serine endoprotease/PDZ domain-containing protein/peptidase S1	1.00
Xf_9a5c_00740	XF_RS03465	peptidase M20/peptidase M28	1.00
Xf_9a5c_02046	XF_RS09455	rhomboid family intramembrane serine protease	1.00
Xf_9a5c_01716	XF_RS08020	carboxyl-terminal processing protease/peptidase S41/peptidase S41, partial/protease	1.91
Xf_9a5c_00081	XF_RS00385	ATP-dependent metallopeptidase FtsH/Yme1/Tma family protein/ATP-dependent zinc metalloprotease FtsH	1.00
Xf_9a5c_00006	not in ncbi	CPBP family intramembrane metalloprotease/hypothetical protein/membrane protein	1.04
Xf_9a5c_00405	XF_RS01875	protease modulator HflC	1.00
Xf_9a5c_00404	XF_RS01870	FtsH protease activity modulator HflK	1.02

Figure S2: Mapping of the MEG in the *X. fastidiosa* strains with draft genome

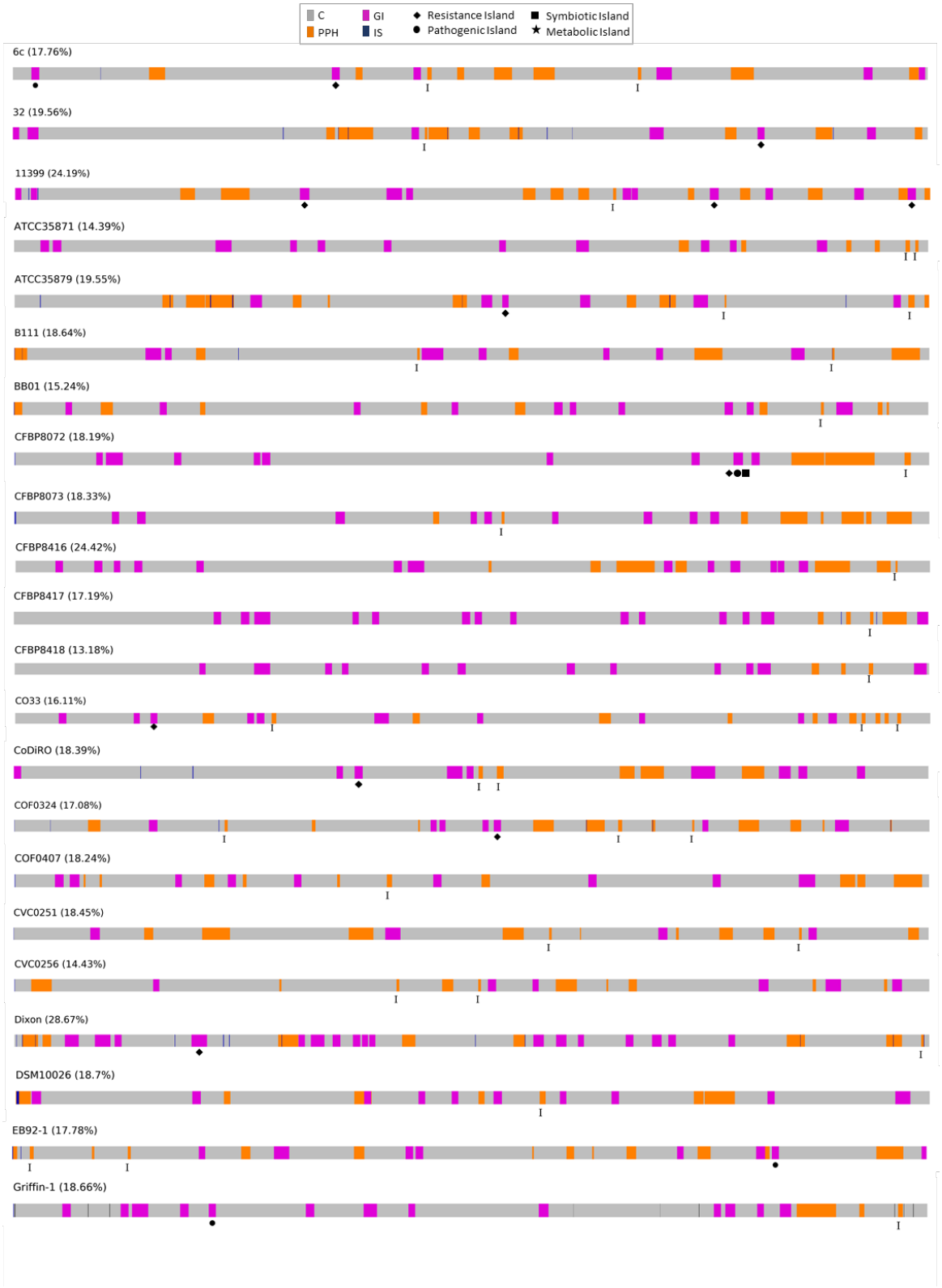

Figure S2: Mapping of the MEG in the *X. fastidiosa* strains with draft genome (*continuation*)