NOVEL INSIGHTS INTO THE ELM YELLOWS PHYTOPLASMA GENOME AND INTO THE METAGENOME OF ELM YELLOWS-INFECTED ELMS

Cristina Rosa, Paolo Margaria, Scott M. Geib, and Erin D. Scully[1]

Abtsract.—In North America, American elms were historically present throughout the northeastern United States and southeastern Canada. The longevity of these trees, their resistance to the harsh urban environment, and their aesthetics led to their wide use in landscaping and streetscaping over several decades. American elms were one of most cultivated plants in the United States until the arrival of Dutch elm disease (DED) and elm yellows disease (EY). EY epidemics have killed large numbers of elm trees in the northeastern United States beginning in the 1940s. Since then, the disease has gradually been spreading to the southern and western regions of the United States while remaining endemic in the Northeast. Today EY, together with DED, is responsible for the death of most of the American species of elm trees, including (*Ulmus americana* (L.), *U. rubra (*Muh.), *U. alata* (Michx.), *U. crassifolia* (Nutt.) *U. serotina* (Sarg)) and of some of their natural hybrids (i.e. *U. pumila* × *rubra*).

We performed next-generation sequencing on EY-infected elm trees to discover EY effector genes involved in plant-phytoplasma interactions and to survey the metagenome of the infected elms. This research is a basic step to understand how EY infection shapes the elm microbial communities and, in the long term, will lead to a better understanding of the pathogenesis of EY infection in elm and the interactions between EY and its leafhopper vectors.

Introduction

Elm yellows (EY) epidemics have killed large numbers of elm trees in the northeastern United States beginning in the 1940s (Carter and Carter 1974, Lanier et al. 1988, Sinclair 1972). EY is an important yet underestimated disease that kills infected trees in 1 to 3 years, depending on the size of the tree (Marcone 2016). Some Asian and European elm species are variably resistant to EY, which led to the hypothesis that the causative agent of EY originated in Europe or Asia. In fact, on those continents, the disease manifests with much milder symptoms, and interestingly, European elms grown in North America have not been found to be naturally infected by EY (Sinclair 1981). Areas afflicted with EY in North America include Canada (Niagara peninsula in Ontario since 1984; Matteoni and Sinclair 1989) and the United States, with presence in approximate latitudes 32 to 46° N and longitudes 71 to 97° W. The states where EY is endemic today include Alabama, Arkansas, Georgia, Iowa, Illinois, Indiana, Kansas, Kentucky, Massachusetts, Minnesota, Mississippi, Missouri, Nebraska, New Jersey, New York, Ohio, Oklahoma, Pennsylvania, Tennessee, West Virginia (CABI 1975) and North Dakota (Stack and Freeman 1988).

The pathogens of EY are wall-less bacteria known as phytoplasmas (Pisi et al. 1981, Wilson et al. 1972), and are vectored by leafhoppers (Baker 1948, 1949). Identified EY vectors include

[1] Assistant Professor of Plant Virology (CR), Pennsylvania State University, 321 Buckhout Labs, University Park, PA 16802; Research Scientist (PM), Braunschweig, Germany; Research Entomologist (SMG), Agricultural Research Service, Hilo, HI; Research Molecular Biologist (EDS), Agricultural Research Service, Manhattan, KS. CR is the corresponding author: to contact, call 814-867-5372 or email at czr2@psu.edu.

Scaphoideus luteolus (Van Duzee) (Barnett 1977); the meadow spittlebug *Philaenus spumarius* (L.); the leafhopper *Allygidius atomarius* (Fabricius); and, more recently, the spittlebug (Cercopidae) *Lepyronia quadrangularis* (Say); and a leafhopper in the genus *Latalus* (Cicadellidae: Deltocephalinae) (Rosa et al. 2014). Adult leafhoppers are widespread geographically and are active from early summer until the first frost in autumn. In temperate regions, leafhoppers overwinter as eggs on elm bark and undergo five instars before molting to adults. Development from first instar to adulthood occurs over a period of about 40 days. Leafhopper nymphs are believed to acquire the EY phytoplasmas in mid-June and begin transmitting it after the incubation period, approximately 3 weeks later from mid-July to September (Sinclair et al. 1976). About 7 weeks post-inoculation, infected trees can serve as reservoirs of new infections. Temperatures below –15 ˚F limit the dissemination of both the vectors and the disease. EY phytoplasmas are vectored exclusively by insects and are obligate pathogens of both their host plants and insects.

Once phytoplasmas are introduced to their host plants, they infect sieve elements in phloem tissues of the elms (Braun and Sinclair 1976). After infection, symptoms typically manifest 3 months post inoculation in young plants and up to 9 months in older trees.[2] Eventually, during the late summer, symptoms such as yellowing of the leaves appear and necrosis of the root system, phloem, and xylem tissues become especially pronounced. Since infected trees are impossible to save, the only solution is to remove them as soon as possible. One extension-type publication claims that EY can be transmitted via root grafting,[3] thus, the root systems of infected trees should also be quickly isolated from the roots of neighboring trees to prevent transmission. Usually the EY population is higher in petioles of brooms of live plants than in dead plants, since the pathogen is an obligate parasite. EY overwinters in the roots of elms, moving into the upper branches in the spring (Braun and Sinclair 1976).

Phytoplasmas are grouped on the basis of their 16S rDNA gene sequence into several ribosomal groups. Strains within the various ribosomal groups are often sub-grouped based on geographical origin and on sequencing of other genes (e.g., elongation factor Tu: *TuF*, variable membrane protein 1: *vmp1*). EY of the reference group *Candidatus* Phytoplasma ulmi [16SrV-A] (Lee et al. 2004) is classified in the 16SrV-A lineage, and, based on host specificity, represents a single species. However, there are three known strains of EY: common, Illinois, and European (OEPP/EPPO 1979). It is possible that other strains exist and that other phytoplasmas, such as aster yellows and clover proliferation, may be causing EY-like symptoms (Jacobs et al. 2003). Phytoplasmas secrete effectors directly into the host cytoplasm of sieve cells via the Sec-dependent protein translocation pathway, and the effectors then unload from the phloem to target other plant cells by symplastic transport (Bai et al. 2009, Hoshi et al. 2009, Sugio et al. 2011b). However, unlike other canonical plant pathogens, genes for the type III and type IV secretion systems and pili are noticeably absent in phytoplasmas, probably because phytoplasmas are introduced into cells directly by their insect vectors during feeding (Kakizawa et al. 2010). Consequently, identification and characterization of phytoplasma effectors are paramount for understanding the processes of host colonization and pathogenicity. Infection with phytoplasmas induces notable changes in plant hormonal balance; specifically, potato purple top phytoplasma causes the reduction of gibberellic acid in tomatoes (Ding et al. 2013) while *Ca*. Phytoplasma mali infection in apple trees stimulates production of plant volatiles that attract insect vectors (Mayer et al. 2008a, 2008b). Furthermore, *Ca*. Phytoplasma asteris effectors interfere with the jasmonic acid (JA) defense pathway (Sugio et al. 2011a) in *Arabidopsis*

[2] Personal communication from Gary W. Moorman, Department of Plant Pathology and Environmental Microbiology, Pennsylvania State University.

[3] Personal communication from Wayne Sinclair, Cornell University.

plants. These observations suggest that phytoplasmas are adept at manipulating plant-based herbivore defense pathways, allowing insect vectors to feed on host plants for extended periods of time and promoting successful pathogen transmission. Consistent with this hypothesis, it was observed that *Nicotiana attenuata* plants deficient in the JA pathway are more damaged by leafhoppers (Kallenbach et al. 2012). Only four phytoplasma genomes are fully available: two strains belonging to the 16Sr-I group (*Ca.* Phytoplasma asteris; Bai et al. 2006, Oshima et al. 2004); one strain of the 16Sr –X group (*Ca*. P. mali; Kube et al. 2008); and one strain of *Ca.* P. australiense (Tran-Nguyen et al. 2006), related to 16SrXII group. No genome of phytoplasmas belonging to the 16SrV-A lineage have been sequenced yet.

Here, we report the identification and annotation of genome fragments of *Candidatus* Phytoplasma ulmi that include putative bacterial effectors and preliminary observations regarding the composition of the microbial community present on EY infected elms.

Methods

Sample Collection, DNA Extraction and Sequencing, and Metagenomics Analysis

Two elms trees infected with EY were found on the Pennsylvania State University campus (40˚48.408'N, 77˚52.208'W, University Park, PA) and used as sources of plant materials. While the specific genealogy of the trees is not known, one tree resembles an American elm, *Ulmus americana* (L.), and the other a red elm, *U. rubra (*Muh.).

Fifty grams of leaf midribs and phloem from twigs were processed from each of the trees and were used to perform separate total DNA extractions, as in Ahrens and Seemüller (1992), by using CTAB extraction buffer and by adding a partial ultracentrifugation enrichment. The ratio of host DNA to phytoplasma DNA was measured by quantitative real time PCR (qRT-PCR) using the Quantstudio 3D digital PCR System (Applied Biosystems®, Foster City, CA); DNA concentration and quality was quantified by Nanodrop spectrophotometer (Fisher Scientific) and assessed by the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA).

Standard Illumina MiSeq long-insert paired libraries were prepared from the two DNA samples at the Huck Institutes genomics core facility, Pennsylvania State University, University Park, PA. DNA was sequenced on the Illumina MiSeq, generating approximately 17.8 million 300 × 300 nt paired-end reads with fragment lengths of 500 nt (5.3 Gb). Reads were trimmed to remove residual adapter sequences and low quality bases using Trimmomatic (version 0.32) with the following options: SLIDINGWINDOW:4:15 MINLEN:150, and ILLUMINACLIP:TruSeq3-SE (http://www.usadellab.org/cms/?page=trimmomatic). Filtered reads were uploaded to MG-RAST (Meyer et al. 2008) for taxonomic and functional classification. rRNAs were identified using RNAmmer (Lagesen et al. 2007) and taxonomically classified using the Ribosomal Database Project (RDP) classifier tool (Wang et al. 2007) with an 80 percent confidence threshold for taxonomic classifications. Putative coding regions were predicted using Prokka (Seemann 2014) and were functionally classified via blastp (Altschul et al. 1997) comparisons to the COG (Tatusov et al. 2000), SEED (Mitra et al. 2011), and the nonredundant (NR) protein databases with an evalue threshold of 10^{-5}. Putative taxonomies of protein coding reads were predicted by blastp comparisons to the NR protein database and MEGAN's least common ancestor algorithm (Huson et al. 2007). Reads coding for effector proteins were identified via blastp searches using reads taxonomically classified as originating from Tenericutes as queries and a custom database containing other previously identified phytoplasma effectors. Gene ontology terms for reads assigned to phylum Tenericutes were computed using Blast2GO (Conesa et al. 2005).

Table 1.—Quality and annotation metrics from shotgun metagenomic sequencing from DNA collected from EY infected elm trees

Number of paired-end reads sequenced	17,897,952 (10.48 Gb)
Number of reads that passed QC	15,559,395 (9.32 Gb)
Number of reads with predicted proteins	11,611,550 (6.96 Gb)
Number of reads with predicted rRNAs	36,260 (22 Mb)
Number of protein coding reads from bacteria	17,333 (10.58 Mb)
Number of protein coding reads from fungi	21,588 (12.95 Mb)
Number of protein coding reads from viruses	61,806 (37.09 Mb)

Table 2.—Abundance of retroviral sequences found in EY infected trees and their classification based on BlastP, GenBank

Sequenced GenBank Annotation	Abundance (number of reads)
Petunia vein clearing virus	682
Ambrosia asymptomatic virus 2 UKM-2007	30
Pelargonium vein banding virus	19

Retrovirus Identification and Characterization

Since multiple retroviruses were found to be integrated into the elm genome, specific primers were designed to amplify the retrovirus sequences of two of the most highly represented viruses, namely a *Petunia vein clearing* hypothetical *virus* and hypothetical *Ambrosia symptomatic virus*. These primers were used to re-amplify the *in silico* assembled viral sequences from the original trees, and to construct larger viral contigs by genome walking. All PCR generated products were sequenced by Sanger sequencing at the Penn State Genomic Core Facility. Additional DNA samples obtained from eight elm trees grown at the U.S. Forest Service facility, in Delaware, Ohio, were screened for the presence of the two retroviral sequences.

Results

DNA Sequencing and Metagenomics Analysis

Approximately 10.48 Gb paired-end MiSeq reads were sequenced from tissues collected from EY infected elm trees, of which about 9.32 Gb passed all quality filters. Most of these reads (~7 Gb) contained predicted coding regions while rRNA sequences accounted for 22Mb. Although most of the reads originated from the host tree, reads originating from fungi, viruses, and bacteria were also readily identified. Viral proteins accounted for most of the microbial coding regions included in the elm metagenome (39 Mb), while fungal and bacterial proteins accounted for 13 Mb and 11Mb, respectively (See Table 1). Bacterial protein-coding regions were classified to 24 different phyla with Proteobacteria, Tenericutes, and Firmicutes being the most highly represented. Notably, the vast majority of the protein coding reads assigned to the phylum Tenericutes had highest scoring blastp matches to proteins from other phytoplasma species. Fungal protein coding regions were classified to 15 different orders with the Eurotiomycetes, Dothideomycetes, and Leotiomycetes as the most highly represented orders, while viruses were almost exclusively assigned to the family Microviridae (bacteriophages, see Fig. 1). Further analyses of the viral sequences after exclusion of bacteriophages identified pararetroviral sequences with high sequence similarities to *Petunia vein clearing virus*, *Ambrosia asymptomatic virus* 2 UKM-2007 and *Pelargonium vein banding virus* (Table 2). We were able to confirm that sequences belonging to two of these three viruses were present, in variable combinations, not only in the DNA extracted from the original tree tissue used for this analysis,

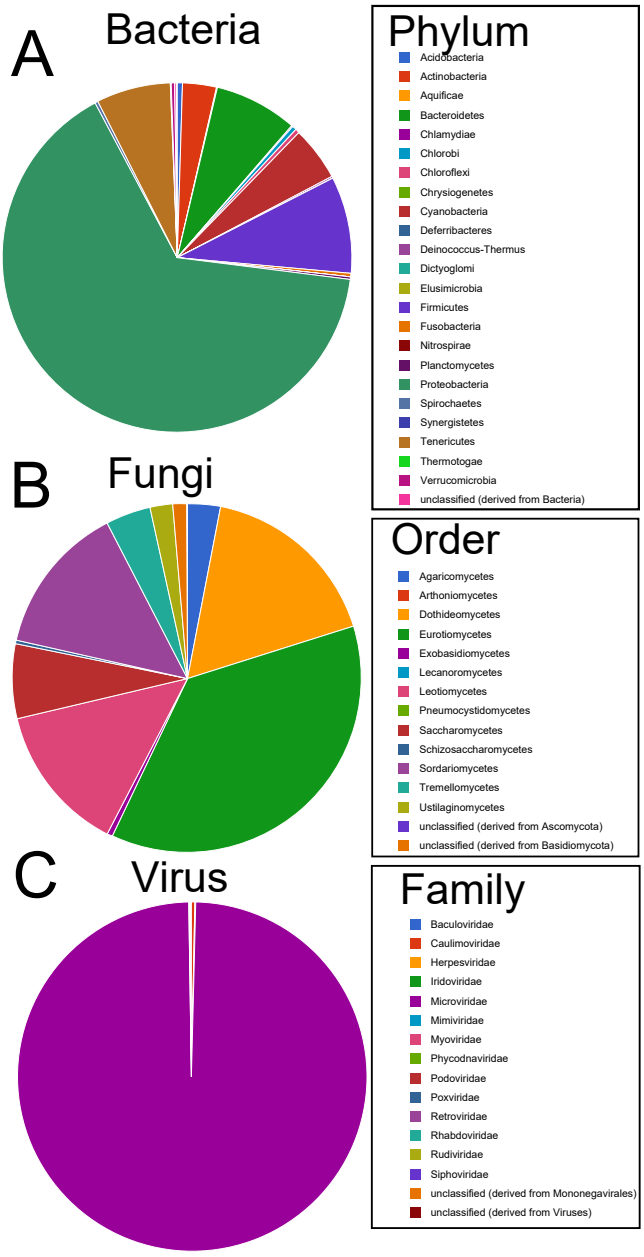

Figure 1.—Taxonomic assignments of predicted protein coding reads from a) bacteria, b) fungi, and c) viruses. Taxonomic assignments of reads predicted to code for proteins were identified via blastp searches to the nonredudant protein database and LCA classification using MEGAN. Bacteria were classified to phylum level, fungi to order level, and viruses to family level. Bacterial reads were classified to 24 different phyla with Proteobacteria, Tenericutes, and Firmicutes being the most highly represented. Fungal reads were classified to 15 different orders with the Eurotiomycetes, Dothideomycetes, and Leotiomycetes as the most represented orders while viruses were almost exclusively assigned to the family Microviridae.

but also in DNA extracted from elm trees collected from another field site in Ohio, suggesting that these viruses are commonly integrated in elm trees, as for other plants.

With regard to rRNA classification, 18 reads containing 16s rRNAs were predicted to originate from Tenericutes while 11 rRNAs were assigned to Burkholderiaceae, and five to Enterobacteriaceae. Other bacterial families detected included Flavobacteriaceae and Cytophagaceae (Fig. 2). Although coding regions containing highest scoring blastp matches to fungi were identified, no fungal rRNAs were identified.

Approximately 850 reads originating from phylum Tenericutes were functionally classified using the SEED database (http://www.theseed.org/). Functional categories including protein metabolism, clustering-based subsystems, carbohydrates, RNA metabolism, and DNA metabolism were highly represented (Fig. 3). The EY key metabolic functions included DNA replication, tRNA aminoacylation for protein translation, nucleobase containing compound

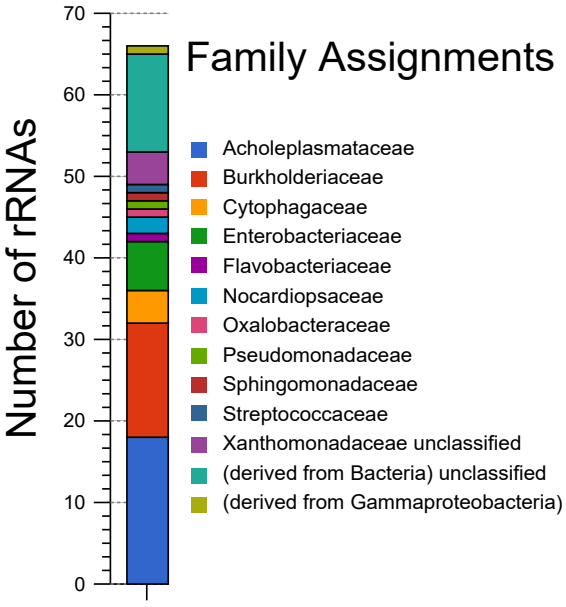

Figure 2.—Class level assignments of bacterial rRNAs detected in EY infected elm tissue. Approximately 65 reads containing bacterial rRNAs were identified using RNAmmer and taxonomically classified to family level using RDP classifier with an 80% confidence threshold. Eighteen rRNAs predicted to originate from Tenericutes, 11 rRNAs predicted to originate from Burkholderiaceae, and five rRNAs predicted to originate from Enterobacteriaceae were identified. Other families detected included Flavobacteriaceae and Cytophagaceae. No rRNAs from fungi or other eukaryotic microbes were identified in this dataset.

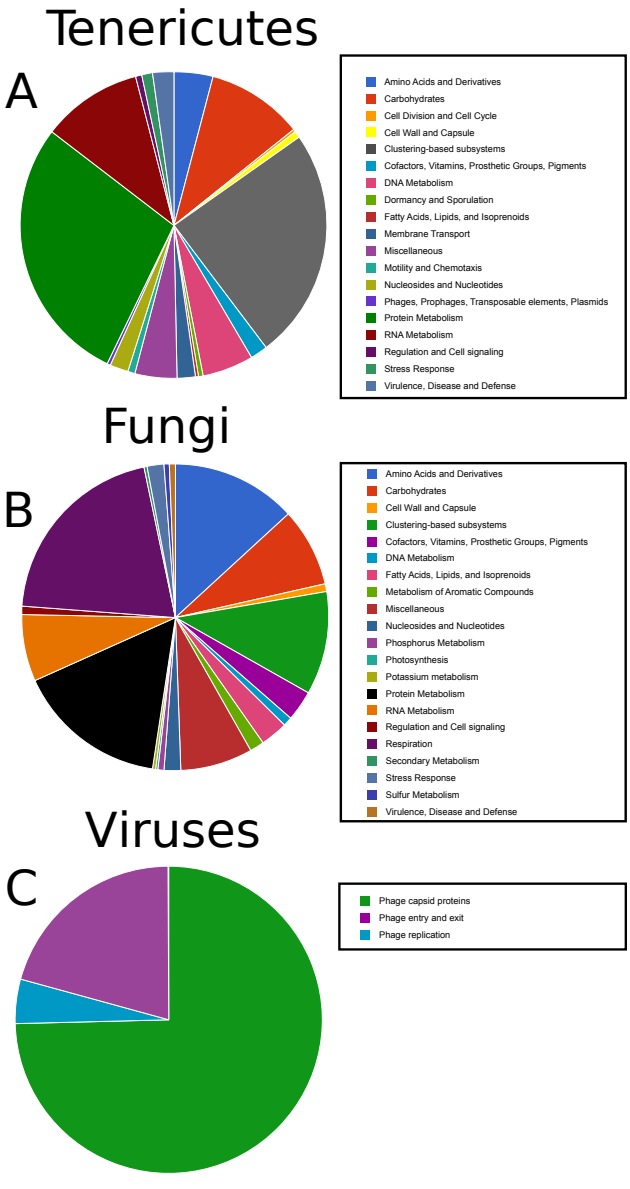

Figure 3.—SEED subsystems assignments for reads assigned to phylum Tenericutes, as Fungi, and as Viruses The main functional categories associated with A) Tenericutes were protein metabolism, clustering-based subsystems, carbohydrates, RNA metabolism, and DNA metabolism. Also, there were a little over 700 reads that were classified as Tenericutes and had similarities to proteins in SEED. For fungi, B), there were 340 reads classified as fungi that had similarity to proteins in SEED.

Again, protein metabolism, clustering based subsystems, and carbohydrates were three of the most prominent categories. In addition, the categories amino acids and derivatives, respiration, cofactors, vitamins, prosthetic groups, pigments, and acids, lipids, and isoprenoids were also well represented.

For viruses, C), there were 28,230 reads classified as virus that had similarity to proteins in seed. The majority of these were capsid proteins, with small numbers of entry and exit and replication proteins identified.

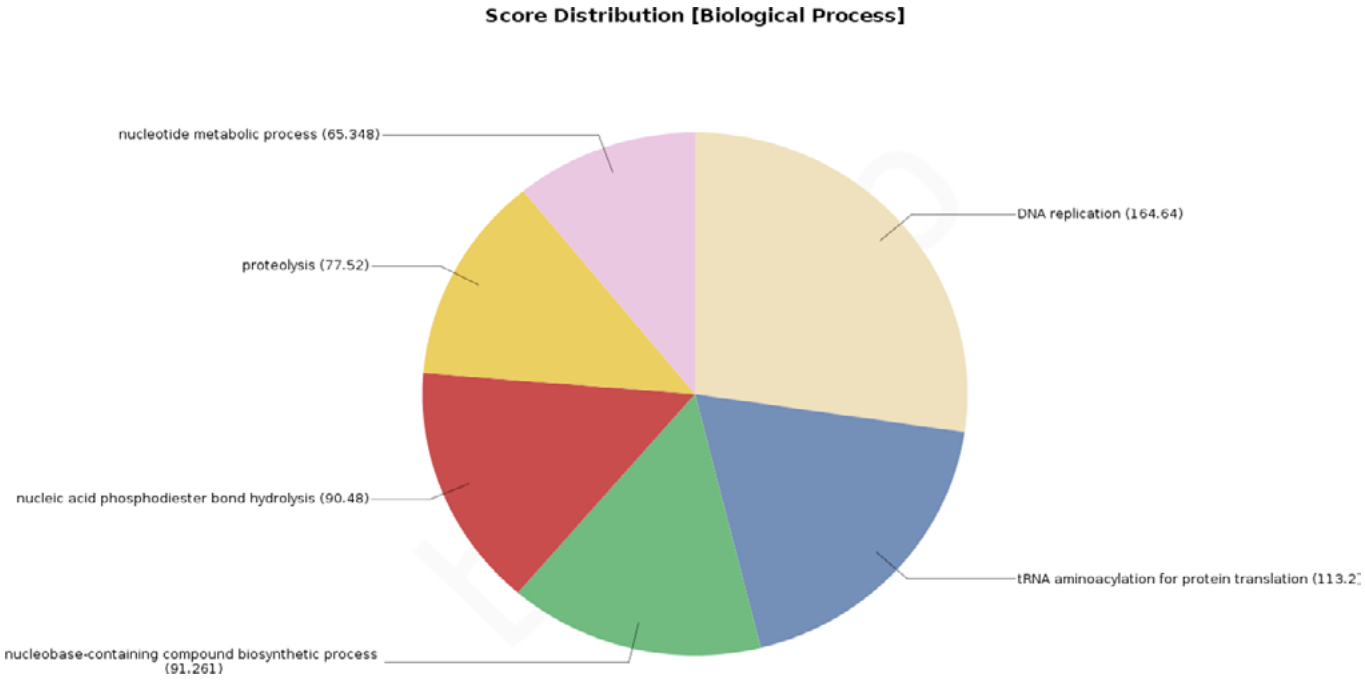

Figure 4:—Score Distribution of EY Biological Processes. The main GO biological process identified for EY were DNA replication, tRNA aminoacylation for protein translation, nucleobase containing compound biosynthetic process, nucleic acid phosphodiester bond hydrolysis, proteolysis and nucleotide metabolic process.

biosynthetic process, nucleic acid phosphodiester bond hydrolysis, proteolysis and nucleotide metabolic process (Fig. 4). EY molecular functions were: ATP binding, metal ion binding, nuclease activity, RNA binding, nucleotidyltransferase activity, DNA binding, ligase activity, and nucleoside-triphosphatase activity (Fig. 5). In addition, Table 3 contains the reads derived from phytoplasmas with their highest scoring blast match, and with the number of reads for each annotation. Using this information, a number of reads coding for putative effectors were positively identified including: endo-beta-1,4-glucanase (break down plant cell walls), protein hupB (siderophore), endopeptidase la, hemolysin channel proteins, hemolysin, ABC maltose transport system, ABC sugar transporters, spermidine/putrescine, ABC transporter permease, transcriptional inducers, and repressors of HrcA heat shock proteins. Two components of the Sec transport system were also readily identified, SecY and SecA.

In fungi, functional catagories corresponding to protein metabolism, clustering based subsystems, and carbohydrates were also three of the most prominent categories while other categories such as amino acids and derivatives, respiration, cofactors, vitamins, prosthetic groups, pigments, fatty acids, lipids, and isoprenoids were also well represented. Viral proteins were mainly capsid proteins, while comparatively smaller numbers of entry and exit and replication proteins were also identified (Fig. 3).

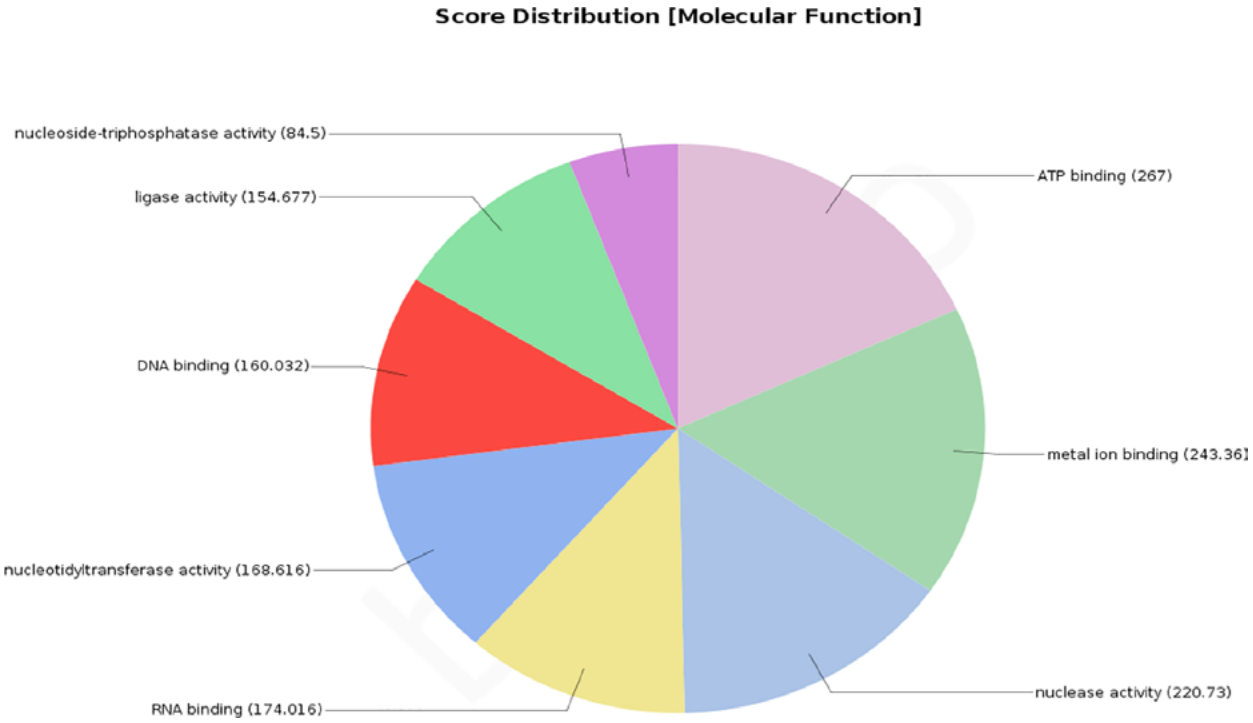

Figure 5.—EY score distribution of molecular functions. EY GO molecular functions were ATP binding, metal ion binding, nuclease activity, RNA binding, nucleotidyltransferase activity, DNA binding, ligase activity and nucleoside-triphosphatase activity.

Table 3.—Phytoplasma reads, their highest scoring blast match, and number of reads for each annotation

Number of reads	Annotation
138	hypothetical protein
30	DNA primase
25	replicative DNA helicase
25	DNA polymerase III subunit alpha
17	DNA double-strand break repair Rad50 ATPase
16	AAA+ ATPase
14	Thymidylate kinase
14	PolC-type DNA polymerase
14	phage-Associated protein
12	DNA helicase
12	AAA+ ATPase, partial
11	exonuclease VII large subunit
11	DNA polymerase III, alpha subunit, Gram-positive
10	conserved hypothetical protein
9	MULTISPECIES: endonuclease
9	endonuclease
9	conserved hypothetical protein, partial sequence,
9	Cell division protein ftsH-like protein

continued

Number of reads	Annotation
9	ATP-dependent DNA helicase
8	zinc ABC transporter permease
8	protein hupB
8	cell division protein FtsH
7	phage-associated protein
7	hypothetical protein S284_00240
7	DNA primase, partial CDS, partial
6	MULTISPECIES: DNA primase
6	hypothetical protein, partial sequence, partial
6	DNA-directed RNA polymerase specialized sigma
6	dihydrolipoyl dehydrogenase
6	ATPase
6	50S ribosomal protein L9
5	valine—tRNA ligase
5	tRNA uridine-5-carboxymethylaminomethyl(34) synthesis
5	sugar ABC transporter permease
5	leucine—tRNA ligase
5	Holliday junction resolvase RecU
5	endopeptidase La
5	DNA gyrase subunit A
5	DNA-directed RNA polymerase subunit beta
5	cysteine—tRNA ligase
4	tRNA (adenosine(37)-N6)-threonylcarbamoyltransferase
4	MULTISPECIES: hypothetical protein
4	methionine—tRNA ligase
4	MBL fold metallo-hydrolase
4	DNA replication protein
4	DNA polymerase III alpha subunit, partial CDS, partial
4	conserved hypothetical protein, partial CDS, partial
4	asparagine—tRNA ligase
4	alanine—tRNA ligase
3	tRNA(Ile)-lysidine synthetase
3	translation initiation factor IF-2
3	spermidine/putrescine ABC transporter permease
3	rRNA (cytidine-2′-O-)-methyltransferase
3	ribosomal RNA small subunit methyltransferase
3	ribonuclease Y
3	peptide chain release factor 2
3	lipoprotein ABC transporter ATP-binding protein
3	histidine—tRNA ligase
3	glutamine—tRNA ligase
3	dTMP kinase

continued

Number of reads	Annotation
3	DNA-directed RNA polymerase subunit beta'
3	DNA-binding protein
3	deoxyribonuclease IV
3	chromosomal replication initiator protein DnaA
3	channel protein, hemolysin III family
3	cation transport ATPase
3	ABC-type maltose transport system, permease protein
3	50S ribosomal protein L3
3	23S rRNA (guanosine(2251)-2'-O)-methyltransferase
2	tRNA (guanosine(37)-N1)-methyltransferase TrmD
2	transcription elongation factor GreA
2	threonine—tRNA ligase
2	sugar ABC transporter substrate-binding protein
2	spermidine/putrescine ABC transporter ATP-binding
2	signal recognition particle protein
2	ribonuclease J
2	pyruvate dehydrogenase (acetyl-transferring) E1
2	primosomal protein N'
2	preprotein translocase subunit SecE
2	PolC-type DNA polymerase III
2	phosphohydrolase
2	peptide chain release factor 1
2	MULTISPECIES: ribosome-recycling factor
2	multidrug ABC transporter permease
2	metallo-beta-lactamase superfamily protein, partial
2	isoleucine—tRNA ligase
2	hypothetical protein S284_01810
2	hypothetical protein S284_01080
2	GTP-binding protein YchF
2	glucose inhibited division protein A, partial sequence,
2	glucose-6-phosphate isomerase
2	exopolyphosphatase
2	excinuclease ABC subunit A
2	elongation factor Ts
2	DNA topoisomerase I
2	DNA polymerase III subunit beta
2	DNA polymerase III, delta prime subunit
2	DNA polymerase III alpha subunit, partial sequence,
2	DNA ligase (NAD(+)) LigA
2	DNA-formamidopyrimidine glycosylase
2	DNA-directed RNA polymerase subunit delta
2	DNA-directed RNA polymerase sigma-70 factor

continued

Number of reads	Annotation
2	diadenosine tetraphosphate hydrolase
2	CTP synthetase
2	copy number control protein (plasmid)
2	conserved hypothetical protein, phage-associated protein
2	class 1b ribonucleoside-diphosphate reductase
2	ATP-dependent zinc metalloprotease FtsH
2	Arginyl-tRNA synthetase, partial sequence, partial
2	adenylate kinase
2	acetate kinase
2	ABC transporter ATP-binding protein
2	50S ribosomal protein L4
2	50S ribosomal protein L35
2	50S ribosomal protein L28
2	50S ribosomal protein L20P, partial CDS, partial
2	50S ribosomal protein L16
2	2,3-bisphosphoglycerate-independent phosphoglycerate
1	YihA family ribosome biogenesis GTP-binding protein
1	Xaa-Pro aminopeptidase, partial sequence, partial
1	Valyl-tRNA synthetase
1	uracil-DNA glycosylase
1	tyrosine—tRNA ligase
1	type I methionyl aminopeptidase
1	type I glyceraldehyde-3-phosphate dehydrogenase
1	tryptophan--tRNA ligase
1	tRNA pseudouridine synthase B
1	tRNA (adenosine(37)-N6)-dimethylallyltransferase
1	triose-phosphate isomerase
1	trigger factor (FKBP-type peptidyl-prolyl cis-trans
1	trigger factor
1	transcription elongation factor NusA, partial sequence,
1	thymidylate synthase
1	sugar permease
1	sugar ABC transporter ATP-binding protein
1	Spermidine/putrescine-binding periplasmic protein
1	sodium transporter
1	site-specific integrase
1	serine protease
1	segregation protein B
1	SAM-dependent methyltransferase
1	rRNA maturation RNase YbeY
1	RNA polymerase sigma factor RpoD
1	ribosome biogenesis GTPase RsgA

continued

Number of reads	Annotation
1	ribosome biogenesis GTPase Der
1	ribosome-binding factor A
1	ribonuclease P protein component
1	Ribonuclease III
1	ribonuclease HIII
1	pyruvate kinase, partial sequence, partial
1	Pyruvate kinase
1	putative endo-1,4-beta-glucanase, partial sequence,
1	pseudouridylate synthase
1	protein translocase component YidC
1	Protein translocase
1	proteasome-activating nucleotidase
1	Prolyl-tRNA synthetase
1	preprotein translocase subunit SecY
1	preprotein translocase subunit SecA
1	Predicted HAD-superfamily hydrolase
1	predicted ATPase AAA-type, contains CbxX/CfqX motif
1	predicted AAA+ ATPAse
1	predicted AAA+ ATPase
1	(p)ppGpp synthetase
1	Phosphoglyceromutase
1	Phosphoglycerate kinase
1	phosphatidylserine decarboxylase
1	phosphatidate cytidylyltransferase
1	phenylalanine--tRNA ligase subunit alpha
1	Phage-Associated Protein
1	peptidyl-tRNA hydrolase
1	peptide transporter
1	peptide ABC transporter substrate-binding protein
1	peptide ABC transporter permease
1	O-methyltransferase
1	nucleotide exchange factor GrpE
1	NAD+ synthetase
1	Na+-driven multidrug efflux pump
1	NADH oxidase
1	MULTISPECIES: transcription elongation factor
1	MULTISPECIES: SsrA-binding protein
1	MULTISPECIES: ribosomal RNA small subunit methyltransferase
1	MULTISPECIES: manganese ABC transporter ATP-binding
1	MULTISPECIES: elongation factor Ts
1	MULTISPECIES: elongation factor 4
1	MULTISPECIES: dTMP kinase

continued

Number of reads	Annotation
1	MULTISPECIES: DNA ligase (NAD(+)) LigA
1	MULTISPECIES: 50S ribosomal protein L33
1	MULTISPECIES: 50S ribosomal protein L28
1	MULTISPECIES: 30S ribosomal protein S2
1	MULTISPECIES: 30S ribosomal protein S15
1	MULTISPECIES: 16S rRNA maturation RNase YbeY
1	Multidrug resistance ABC transporter ATP-binding and
1	multidrug ABC transporter ATP-binding protein
1	molecular chaperone DnaK
1	methionine adenosyltransferase
1	Malate/Na+ symporter
1	malate:citrate symporter
1	lysine—tRNA ligase
1	lipoate—protein ligase
1	kinase
1	inorganic pyrophosphatase
1	hypothetical protein, YrdC-like domain protein
1	hypothetical protein S284_01820
1	hypothetical protein, partial CDS, partial
1	HrcA family transcriptional regulator
1	Holliday junction DNA helicase RuvB
1	Holliday junction DNA helicase RuvA
1	hemolysin
1	Heat-inducible transcription repressor HrcA
1	haloacid dehalogenase
1	guanosine polyphosphate pyrophosphohydrolase
1	GTP pyrophosphokinase
1	GTPase ObgE
1	glycine—tRNA ligase
1	glycerol-3-phosphate acyltransferase
1	Glyceraldehyde-3-phosphate dehydrogenase
1	glutamate—tRNA ligase
1	Glucose-inhibited division protein A, partial sequence,
1	fructose-1,6-bisphosphate aldolase, class II
1	Formamidopyrimidine-DNA glycosylase
1	fatty acid-binding protein DegV
1	Excinuclease ATPase subunit A, partial sequence, partial
1	energy-coupling factor transporter ATP-binding
1	energy-coupling factor transporter ATPase
1	elongation factor Tu
1	elongation factor P
1	elongation factor 4

continued

Number of reads	Annotation
1	DNA-directed RNA polymerase subunit alpha
1	DNA-directed RNA polymerase beta chain, partial sequence,
1	dipeptide transport ATP-binding protein DppF
1	dipeptide/oligopeptide/nickel ABC transporter
1	dimethyladenosine transferase
1	DEAD/DEAH box helicase family protein, SrmB-like
1	cytidine(C)-cytidine(C)-adenosine (A)-adding
1	cobalt ABC transporter ATP-binding protein
1	CMP-binding protein
1	Chromosomal replication initiator protein DnaA, partial
1	CDP-diacylglycerol--serine O-phosphatidyltransferase
1	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
1	cation uptake P-type ATPase
1	Cation transport ATPase, partial sequence, partial
1	Calcium-translocating P-type ATPase A
1	cadmium-transporting ATPase
1	ATP-dependent Zn protease
1	ATP-dependent zinc metalloprotease FtsH 3
1	ATP-dependent zinc metalloprotease FtsH 2
1	aspartyl-tRNA synthetase, partial sequence, partial
1	aspartate--tRNA ligase
1	arginine--tRNA ligase
1	acyl carrier protein
1	ABC transporter substrate-binding protein
1	AAA+ ATpase
1	6-phosphofructokinase
1	5′-3′ exonuclease
1	50S ribosomal protein L7/L12
1	50S ribosomal protein L6
1	50S ribosomal protein L24
1	50S ribosomal protein L21
1	30S ribosomal protein S8
1	30S ribosomal protein S6
1	30S ribosomal protein S16
1	30S ribosomal protein S15
1	1-acyl-sn-glycerol-3-phosphate acyltransferase
1	16S rRNA pseudouridylate synthase
1	16S rRNA (adenine(1518)-N(6)/adenine(1519)-N(6))-dimethyltransferase

Conclusions

Phytoplasmas can infect about 1,000 different plant species (McCoy et al. 1989). However, despite their importance as plant pathogens, only a handful of phytoplasma genomes have been sequenced. This lack of sequence availability is due to the intrinsic properties of these bacteria that make them particularly challenging to work with. Though phytoplasmas are evolutionarily derived from gram positive ancestors, they lack a cell wall and cannot be cultured in axenic conditions (Firrao et al. 2004). Furthermore, their AT-rich genomes are significantly reduced in size relatively to other bacterial plant pathogens (Marcone and Seemuller 2001, Marcone et al. 1999, Neimark and Kirkpatrick 1993), ranging in size from 300 to 700 Mb. The high richness further impedes genome sequencing efforts as designing specific primers for PCR-based sequencing is very difficult. Adding further to these complications are the presence of large numbers of mobile genetic elements (Kube et al. 2008) and potential mobile units (PMUs) within the genomic DNA that have the potential to reshuffle gene orders (Bai et al. 2006). PMUs are suggested to be mobile elements involved in phytoplasma host switching (Toruno et al. 2010). In addition, many phytoplasmas contain plasmids (Kube et al. 2008, Tran-Nguyen 2006); however, not much is known about their function.

After the arrival of EY at Pennsylvania State University campus (University Park, PA), researchers developed EY detection techniques via a highly specific real time RT-PCR assay (Herath et al. 2010), monitored the EY incidence on the PSU campus during the last 3 years, and determined the seasonal distribution pattern of the phytoplasma in infected trees. Researchers tested more than 1000 elm samples from 471 trees (Herath et al. 2010), and identified two new insect species as EY vectors (Rosa et al. 2014). The next step in our research at Penn State is to offer novel information on EY phytoplasma genome, and especially to identify putative phytoplasma effectors (SAP). Effectors are molecules secreted by the bacteria into the cells of the hosts. SAPs can change flower development and leaf shape and can modify plant-insect interactions, increasing phytoplamsa fitness. For instance, the aster yellows (AY) phytoplasma strain witches' broom (AY-WB) SAP11 is localized in the cell nuclei, deregulates jasmonic acid production, and produces symptoms (witches' broom phenotype). AY-WB has more than 50 effectors (Sugio et al. 2011a).

Based on our preliminary analyses, EY infected trees contain a metagenomics core that includes many bacteria and fungi. The bacteria found belong to families containing plant pathogenic bacteria as well as bacteria associated with plant, soil, and insects.

Data obtained in this study did not allow us to classify the fungi below the order level, but several of the coding regions have highest scoring blastp matches to Dutch elm disease associated fungi. Studying the identity of these fungi could bring some knowledge as to their use as biocontrols against DED.

The elm genome contains many pararetroviral sequences. We don't know if these integrated viruses generate episomal infections, but our tests suggest that the presence and number of pararetroviral sequences could be used as elm phylogenetic tool. Elm phylogeny is complicated and will eventually rely on classification based on key elm genes, but the use of retroviral sequences for classification could be an easier way that could be used until the elm phylogeny is not completely resolved. Many bacteriophage sequences were also found in the EY microbial reads, probably integrated in the plant genome as well as in the genomes of the plant associated microbes. In conclusion, this study represents the first step in the study of EY genome and of the metagenomics community associated with EY infected trees.

Acknowledgments

Funding for this research was provided by the Pennsylvania State University, Office of Physical Plant. We thank Elizabeth McCarthy for technical support. Blast searches were performed using computing resources available at Agricultural Research Service's Daniel K. Inouye Pacific Basin Agricultural Research Center (Moana cluster; Hilo, HI). Mention of commercial products and organizations in this manuscript is solely to provide specific information. It does not constitute endorsement by USDA Agricultural Research Service over other products and organizations not mentioned.

Literature Cited

Ahrens, U.; Seemüller, E. 1992. **Detection of DNA of plant pathogenic mycoplasmalike organisms by a polymerase chain reaction that amplifies a sequence of the 16 S rRNA gene.** Phytopathology. 82(8): 828-832.

Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. 1997. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** Nucleic Acids Research. 25(17): 3389-3402.

Bai, X.; Zhang, J.; Ewing, A.; Miller, S.A.; Radek, A.J. [et al.]. 2006. **Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts.** Journal of Bacteriology. 188(10): 3682-3696.

Bai, X.; Correa, V.R.; Toruño, T.Y.; Ammar, E.D.; Kamoun, S.; Hogenhout, S.A. 2009. **AY-WB phytoplasma secretes a protein that targets plant cell nuclei.** Molecular Plant-Microbe Interactions. 22(1): 18-30.

Baker, W.L. 1948. **Transmission by leaf hoppers of the virus causing phloem necrosis of American elm.** Science. 108(2803): 307-308.

Baker, W.L. 1949. **Studies on the transmission of the virus causing phloem necrosis of American elm, with notes on the biology of its insect vector.** Journal of Economic Entomology. 42(5): 729-732.

Barnett, D.E. 1976. **A revision of the Nearctic species of the genus Scaphoideus (Homoptera: Cicadellidae).** Transactions of the American Entomological Society. 102(4): 485-593.

Braun, E.J.; Sinclair, W.A. 1976. **Histopathology of phloem necrosis in *Ulmus americana*.** Phytopathology. 66: 598-607.

CABI. 1975. **Elm phloem necrosis virus, 2nd edition [distribution map]. Map 107.** In: Distribution maps of plant diseases. Wallingford, UK: CAB International. http://www.cabi.org/dmpd/ (June 6, 2017).

Carter, J.E.; Carter, L.R. 1974. **An urban epiphytotic of phloem necrosis and Dutch elm disease, 1944-1972.** Illinois Natural History Survey Bulletin. 31(4): 112-143.

Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. 2005. **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** Bioinformatics. 21(18): 3674-3676.

Ding, Y.; Wei, W.; Wu, W.; Davis, R.E.; Jiang, Y. [et al.]. 2013. **Role of gibberellic acid in tomato defence against potato purple top phytoplasma infection.** Annals of Applied Biology. 162(2): 191-199.

Firrao, G.; Andersen, M.; Bertaccini, A.; Boudon, E.; Bove, J.M. [et al.] 2004. **Candidatus phytoplasma', a taxon for the wall-less, non-helical prokaryotes that colonize plant phloem and insects.** International Journal of Systematic and Evolutionary Microbiology. 54(4): 1243-1255.

Herath, P.; Hoover, G.A.; Angelini, E.; Moorman, G.W. 2010. **Detection of elm yellows phytoplasma in elms and insects using real-time PCR.** Plant Disease. 94(11): 1355-1360.

Hoshi, A.; Oshima, K.; Kakizawa, S.; Ishii, Y.; Ozeki, J. [et al.]. 2009. **A unique virulence factor for proliferation and dwarfism in plants identified from a phytopathogenic bacterium.** Proceedings of the National Academy of Sciences. 106(15): 6416-6421.

Huson, D.H.; Auch, A.F.; Qi, J.; Schuster, S.C. 2007. **MEGAN analysis of metagenomic data.** Genome Research. 17(3):377-386.

Jacobs, K.A.; Lee, I.M.; Griffiths, H.M.; Miller Jr, F.D.; Bottner, K.D. 2003. **A new member of the clover proliferation phytoplasma group (16SrVI) associated with elm yellows in Illinois.** Plant Disease. 87(3): 241-246.

Kakizawa, S.; Oshima, K.; Namba, S. 2010. **Functional genomics of phytoplasmas.** In: Weintraub, P.G.; Jones, P., eds. Phytoplasmas: genomes, plant hosts and vectors. Oxford, UK: CAB International: 37-50.

Kallenbach, M., Bonaventure, G., Gilardoni, P.A., Wissgott, A. and Baldwin, I.T. 2012. **Empoasca leafhoppers attack wild tobacco plants in a jasmonate-dependent manner and identify jasmonate mutants in natural populations.** Proceedings of the National Academy of Sciences. 109(24): E1548-E1557.

Kube, M.; Schneider, B.; Kuhl, H.; Dandekar, T.; Heitmann, K.; Migdoll, A.M.; Reinhardt, R.; Seemüller, E. 2008. **The linear chromosome of the plant-pathogenic mycoplasma 'Candidatus Phytoplasma mali'.** BMC Genomics. 9(1): 306.

Lagesen, K.; Hallin, P.; Rødland, E.A.; Stærfeldt, H.H.; Rognes, T.; Ussery, D.W. 2007. **RNAmmer: consistent and rapid annotation of ribosomal RNA genes.** Nucleic Acids Research. 35(9): 3100-3108.

Lanier, G.N.; Schubert, D.C.; Manion, P.D. 1988. **Dutch elm disease and elm yellows in central New York: out of the frying pan into the fire.** Plant Disease. 72(3): 189-194.

Lee, I.-M.; Martini, M.; Marcone, C.; Zhu, S.F. 2004. **Classification of phytoplasma strains in the elm yellows group (16SrV) and proposal of 'Candidatus Phytoplasma ulmi' for the phytoplasma associated with elm yellows.** International Journal of Systemic and Evolutionary Microbiology. 54: 337-347.

Marcone, C. 2016. **Elm yellows: A phytoplasma disease of concern in forest and landscape ecosystems.** Forest Pathology. 47(1): e12324.

Marcone, C.; Neimark, H.; Ragozzino, A.; Lauer, U.; Seemüller, E. 1999. **Chromosome sizes of phytoplasmas composing major phylogenetic groups and subgroups.** Phytopathology. 89(9): 805-810.

Marcone, C.; Seemüller, E. 2001. **A chromosome map of the European stone fruit yellows phytoplasma.** Microbiology. 147(5): 1213-1221.

Matteoni, J.A.; Sinclair, W.A. 1989. **A note on the presence of elm yellows in the Niagara Peninsula.** Phytoprotection. 70(3): 137-139.

Mayer, C.J.; Vilcinskas, A.; Gross, J. 2008a. **Pathogen-induced release of plant allomone manipulates vector insect behavior.** Journal of Chemical Ecology. 34(12):1518-1522.

Mayer, C.J.; Vilcinskas, A.; Gross, J. 2008b. **Phytopathogen lures its insect vector by altering host plant odor.** Journal of Chemical Ecology. 34(8): 1045-1049.

McCoy, R.E.; Caudwell, A.; Chang, C.J. [et al.]. 1989. **Plant diseases associated with mycoplasma-like organisms.** In: Whitcomb, R.F.; Tully, J.G., eds. The mycoplasmas, Vol. 5: Spiroplasmas, Acholeplasmas, and Mycoplasmas of plants and arthropods. New York, NY: Academic Press: 546-640.

Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.M. [et al.]. 2008. **The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes.** BMC Bioinformatics. 9(1): 386.

Mitra, S.; Rupek, P.; Richter, D.C.; Urich, T.; Gilbert, J.A. [et al.]. 2011. **Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics.** 12(1): S21.

Neimark, H.; Kirkpatrick, B.C. 1993. **Isolation and characterization of full-length chromosomes from non-culturable plant-pathogenic Mycoplasma-like organisms.** Molecular Microbiology. 7(1): 21-28.

OEPP/EPPO [European and Mediterranean Plant Protection Organization]. 1979. **Data sheet on quarantine organisms no. 26. Elm phloem necrosis (mycoplasm).** OEPP/EPPO Bulletin. 9 (2).

OEPP/EPPO [European and Mediterranean Plant Protection Organization]. 1982. **Data sheets on quarantine organisms no. 63. *Ceratocystis ulmi*.** OEPP/EPPO Bulletin. 12 (1).

Oshima, K.; Kakizawa, S.; Nishigawa, H.; Jung, H.Y.; Wei, W. [et al.]. 2004. **Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma.** Nature Genetics. 36(1): 27-29.

Pisi, A.; Marani, F.; Bertaccini, A. 1981. **Mycoplasma-like organisms associated with elm witches' broom symptoms.** Phytopathologie Mediterranea. 20(2/3): 189-191.

Rosa, C.; McCarthy, E.; Duong, K.; Hoover, G.; Moorman, G. 2014. **First report of the spittlebug Lepyronia quadrangularis and the leafhopper *Latalus* sp. as vectors of the elm yellows associated phytoplasma, Candidatus phytoplasma ulmi in North America.** Plant Disease. 98(1): 154.

Seemann, T. 2014. **Prokka: rapid prokaryotic genome annotation.** Bioinformatics. 30(14): 2068-99.

Sinclair, W.A. 1972. **Phloem necrosis of American and slippery elms in New York.** Plant Disease Reporter. 56(2): 159-161.

Sinclair, W.A.; Braun, E.J.; Larsen, A.O. 1976. **Update on phloem necrosis of Elms.** Journal of Arboriculture. 2(6): 106-113.

Sinclair, W.A. 1981. **Elm yellows.** In: Stipes, R.J.; Campana, R.J., ed. Compendium of elm diseases. St. Paul, MN: American Phytopathological Society: 25-31.

Stack, R.W.; Freeman, T.P. 1988. **First report of elm yellows in North Dakota.** Plant Disease. 72(10): 912.

Sugio, A.; MacLean, A.M.; Grieve, V.M.; Hogenhout, S.A. 2011a. **Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating plant development and defense hormone biosynthesis.** Proceedings of the National Academy of Sciences. 108(48): E1254-E1263.

Sugio, A.; MacLean, A.M.; Kingdom, H.N.; Grieve, V.M.; Manimekalai, R.; Hogenhout, S.A. 2011b. **Diverse targets of phytoplasma effectors: from plant development to defense against insects.** Annual Review of Phytopathology. 49: 175-195.

Tatusov, R.L.; Galperin, M.Y.; Natale, D.A.; Koonin, E.V. 2000. **The COG database: a tool for genome-scale analysis of protein functions and evolution.** Nucleic Acids Research. 28(1): 33-36.

Toruño, T.Y.; Seruga Musić, M.; Simi, S.; Nicolaisen, M.; Hogenhout, S.A. 2010. **Phytoplasma PMU1 exists as linear chromosomal and circular extrachromosomal elements and has enhanced expression in insect vectors compared with plant hosts.** Molecular Microbiology. 77(6): 1406-1415.

Tran-Nguyen, L.T.T.; Gibb, K.S. 2006. **Extrachromosomal DNA isolated from tomato big bud and Candidatus Phytoplasma australiense phytoplasma strains.** Plasmid. 56(3): 153-166.

Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. 2007. **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.** Applied and Environmental Microbiology. 73(16): 5261-5267.

Wilson, C.L.; Seliskar, C.E.; Krause, C.R. 1972. **Mycoplasmalike bodies associated with elm phloem necrosis.** Phytopathology. 62: 140-143.