

THE GENETIC CHARACTERIZATION OF NORTHEASTERN *QUERCUS* ASSOCIATED *XYLELLA*
FASTIDIOSA POPULATIONS

By

GREGORY BEHRINGER

A dissertation submitted to the Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Graduate Program in Plant Biology

Written under the direction of

Donald Y. Kobayashi

And approved by

New Brunswick, New Jersey

October 2015

ABSTRACT OF THE DISSERTATION

The Genetic Characterization of Northeastern *Quercus* Associated *Xylella fastidiosa* Populations

By GREGORY BEHRINGER

Dissertation Director:

Donald Y. Kobayashi

Xylella fastidiosa is broad-host-range plant pathogen responsible for significant commodity crop damage in much of the Western Hemisphere. Since its description in 1892, focus has centered around disease associated with *Vitis* (grape) hosts. Shade tree host studies of *X. fastidiosa* populations, however, have been both sparse and regionally oriented, making the exploration of infected oak stands an important area for greater understanding of this phytopathogen.

To describe novel genetic profiles of the oak associated pathogen, Northeastern and Mid-Atlantic populations were assessed both phylogenetically and with Principal Component Analyses (PCA) and Minimum Spanning Trees (MSN). Polymerase chain reaction (PCR) based locus recoveries identified previously undescribed genetic diversity and phylogenetically separated oak associated populations based on host geography. Expanded analysis of insertion/deletion regions associated with the oak pathogen was also conducted for fine separation of populations relative to phylogenetic recoveries. Together these provided an efficient means to track the spread of the pathogen at the population level.

To further explore genetic diversity in understudied *X. fastidiosa* oak populations, the genome of a Northeastern *Quercus palustris* associated *X. fastidiosa* isolate, RNB1, was sequenced and analyzed. Existing isolate comparisons described several novel RNB1 genomic regions, including two potential vir genes, and a Gene Ontology procyclic repeat pathogenesis locus. This work provided the first comparative look at an oak associated *X. fastidiosa* genome and described its composition relative to well described isolates.

A final search for novel population specific markers in *X. fastidiosa* colonies targeted prophage segments. Thirteen regions across nineteen genomes were qualitatively described, with phage repressor and terminase suggestive of previously confirmed phylogenetic relatedness at an integrated phage-based locus. This data was then used in several machine learning approaches and proved accurate in predicting taxonomic categories across disparate *X. fastidiosa* populations when trained with matrix transforms of host specific *X. fastidiosa* prophage regions. This final study described evolutionary significance of widely profiled prophage regions and introduced an algorithmic approach for future large-scale genetically themed *X. fastidiosa* based population studies. Overall, the work herein presents previously undescribed genetic aspects of oak associated *X. fastidiosa* populations and posits a novel method for future data synthesis.

ACKNOWLEDGEMENTS

I am truly grateful for the opportunities afforded to me by my main advisor, Dr. Donald Kobayashi. Despite limitations in my scientific background, he remained supportive and trusted my instincts throughout the dissertation process. My committee was also helpful in steering my dissertation to its conclusion. For this, I sincerely thank Dr. Ann Brooks Gould, Dr. Peter Oudemans, Dr. James Polashock, and Dr. Ning Zhang.

A special thank you is also given to my beautiful wife, Laurie, who remained both patient and supportive throughout the process. Thank you believing in me in spite of the crazy trajectories created by a mid-life science degree.

A final thank you is given to all family and friends that helped pull me through the journey as well. Whether it was encouragement, chastisement, or a beer, the parts were vital to the sum.

TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION.....	ii
ACKNOWLEDGEMENTS.....	iv
TABLE OF CONTENTS.....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
CHAPTER 1. Literature Review.....	1
Historical Overview Leading to the Advent of Molecular Techniques.....	1
Early Molecular Techniques and Population Profiling to the first <i>X. fastidiosa</i> genome release.....	3
Early Genetic Characterizations, Multilocus Sequencing Analysis (MLSA), and Continued Population and Speciation Studies.....	6
Expansion of Isolate Sequencing, Large Scale Comparative Genomics, and the Extension into Targeted Locus-Based Research.....	9
Prophage Elements in <i>X. fastidiosa</i> Genomes and Significance in Bacterial Population Distribution and Structure.....	11
References.....	14
CHAPTER 2. “The genetic composition of oak associated <i>X. fastidiosa</i> populations in the Northeastern and Mid-Atlantic United States”	23
Abstract.....	23

Introduction.....	23
Materials and Methods.....	26
Isolate Collection.....	26
DNA Extraction and PCR amplification from environmental samples.....	26
Sequence Alignment, Phylogenetic Analysis, and Tree Imaging.....	27
SNP Analysis, Principal Component Axis (PCA) Analysis, Minimum Spanning tree analysis, and recombination detection.....	28
Results.....	29
Sequence Analysis: Alignment, Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletion (indel) Profiles for Conserved (MLSA) and Variable (MLSA-E) Loci Sets.....	29
Phylogenetic Analysis: gene trees and concatemeric trees for conserved and variable marker sets.....	32
Hierarchical Clustering and Principal Component Analysis (PCA).....	35
Minimum Spanning Tree (MSN).....	37
Recombination Break point Detection.....	38
Discussion.....	39
References.....	45

Chapter 3. “The Intrasubspecies Comparative Genomics of RNB1, a Northeastern <i>Quercus palustris</i> (pin oak) associated <i>Xylella fastidiosa</i> isolate”	74
Abstract.....	74
Introduction.....	74
Materials and Methods.....	77
Isolation, DNA Extraction, and Genomic Sequencing.....	77
Comparative Genomics Pipeline.....	79
Supplementary Locus Amplification from Oak Based Environmental Samples.....	80

Comparative Genomics Data Presentation and Sources.....	80
Results.....	81
Pipeline Based Metrics and Comparative Genomic Analysis.....	81
Environmental Sample Comparison and Multiple Correspondence Analyses (MCA).....	87
Discussion.....	90
References.....	96
 Chapter 4. “The Exploration and Application of Cross-Subspecies <i>Xylella fastidiosa</i> Prophage Regions”.....	118
Abstract.....	118
Introduction.....	119
Materials and Methods.....	121
Genomic Mining, Annotation, and Prophage Region Selection.....	121
Alignment, Matrix Creation, and Machine Learning Analyses.....	122
Results.....	124
Prophage Region Qualitative Analysis.....	124
Dominant Haplotypes and Cross Subspecies Categorization.....	125
Machine Learning Classifier Output.....	126
Discussion.....	128
References.....	134
APPENDIX.....	154

LIST OF TABLES

Table 2.1. Novel New Jersey based and Northeastern/Mid-Atlantic oak associated genotypes considered in this study.....	50
Table 2.2. Southern-based Oak haplotypes described in prior multilocus studies.....	53
Table 2.3. Previously described MLSA and MLSA-E loci used for oak associated population based comparison.....	54
Table 3.1. Polymerase Chain Reaction (PCR) primer list used for verification of <i>X. fastidiosa</i> colonies and exploration of poor scoring RNB1 genomic regions.....	104
Table 3.2. Pfam-A domain hits and Gene Ontology (GO) association for remote BLASTp queries with E-values greater than 1e-10 (E-value > 1.0e-10) and for queries producing no hits (“No hits found”).....	105
Table 3.3. Final categorization of initially identified poor scoring RNB1 protein queries as compared to the total collection of GenBank (NCBI) <i>X. fastidiosa</i> isolates.....	106
Table 3.4. Supplemental Northeastern / Mid-Atlantic environmental samples for variable length tandem repeat (VNTR) analysis of ORF_415 “PELE” and ORF_2081 “POLYL” regions.....	109
Table 4.1. Profile and selection of prophage regions mined from the nineteen selected <i>X. fastidiosa</i> genomes.....	138
Table 4.2. Dominant haplotypes, taxonomy, and domain profiling of mined prophage regions for the nineteen selected <i>X. fastidiosa</i> genomes based on described selection criteria.....	140
Table 4.3. Accuracy percentages and Kappa support for Naïve Bayes, Logistic, and Sequential Minimal Optimization(SMO) classifier results based on both original attribute files with tenfold cross-validation and Synthetic Minority Oversampling Technique (SMOTE) files with tenfold cross-validation.....	149

LIST OF FIGURES

Figure 2.1. Conserved (MLSA) and variable (MLSA-E) SNP profiles for both complete host concatemerized loci and oak associated concatemerized loci.....	55
Figure 2.2. Bayesian phylogenetic tree construction consisting of the concatemerized MLSA-E loci (Parker et al.) and the novel Northeaster and Mid-Atlantic oak associated haplotypes enumerated in Table 2.1.....	59
Figure 2.3. Insertion / Deletion (Indel) profiles and locations for the oak associated gene fragments <i>acvB</i> , <i>copB</i> , and <i>nuoL</i>	61
Figure 2.4. Eigenvalue measure, complete linkage hierarchical clustering, and two axis Principal Component Analysis (PCA) for the concatemeric MLSA and MLSA-E oak associated loci.....	62
Figure 2.5. Minimum Spanning Tree for both the complete MLSA and the complete MLSA-E <i>X.</i> <i>fastidiosa</i> concatemeric haplotypic locus collection.....	69
Supplemental Figure 2.1. Individual Bayesian gene tree files for the seventeen previously described loci used in this study.....	154
Supplemental Figure 2.2. Bayesian phylogenetic tree construction consisting of the concatemerized MLSA loci (Schuenzel et al.) and the novel New Jersey oak associated haplotypes enumerated in Table 1.....	73
Supplemental Figure 2.3. Alignment files for recombination Analysis via Difference of Sums of Squares (DSS) method (McGuire et al. 2000).....	154
Supplemental Figure 2.4. MLSA and MLSA-E concatemerized isolate/environmental sample sequence for all profiled loci.....	154
Supplemental Figure 2.5. RAxML output files for comparative analysis relative to Bayesian methodology / recovered Bayesian topology.....	154
Figure 3.1. Venn Diagram depicting the RNB1 database BLASTp hits resulting from queries derived from putative open reading frames (ORFs) in the <i>X. fastidiosa</i> genomes M12, Griffin, and	

Temecula1.....	100
Figure 3.2. Raw numeric counts of E-value categorizations for reciprocal BLASTp queries derived from putative open reading frames (ORFs) for all RNB1 centered query/database genome combinations.....	102
Figure 3.3. Raw numeric counts of E-value categorizations for remote (NCBI database "nr") BLASTp queries derived from initial poor scoring segments and "No hits found" sequences in reciprocal BLASTp runs.....	103
Figure 3.4. Global genomic snapshot of absolute similarities in M12, Griffin-1, and Temecula1 called open reading frames (ORFs) relative to RNB1 ORFs.....	111
Figure 3.5. Pipeline derived poor scoring RNB1 protein queries relative to multiple technique recovered sequences in the respective genome locations present in M12, Griffin-1, and Temecula1.....	112
Figure 3.6. Multiple Correspondence Analysis (MCA) of categorical data for variable length tandem repeat (VNTR) containing poor scoring RNB1 ORFs: ORF_2081 (Poly L), ORF_415 (PELE), and ORF_1273 (QAQA).....	114
Supplemental 3.1. Associated *.faa, *.fna, *.gbk, and annotation files based on the assembled <i>Quercus palustris</i> associated isolate (RNB1).....	154
Figure 4.1. Comparative confusion matrix results for the three best performing (baseplate894, tailfiber, and terminase) and one worst performing (repressor) prophage region trained classifiers based on kappa support.....	151
Figure 4.2. Comparative cumulative accuracy and error rates among the three classifiers Naïve Bayes, Logistic, and Sequential Minimal Optimization(SMO) for the predicted nominal class "subspecies".....	152

CHAPTER 1. Literature Review

Historical Overview Leading to the Advent of Molecular Techniques

The initial description of *Xylella fastidiosa* symptomology begins with the monograph of Newton B. Pierce in 1892 (Pierce 1892). The tone of Pierce in describing the potential economic impact of this California vine disease that would come to bear his name (Pierce's Disease) rings as true then as it does today when he noted that unlike other commercial crops where neighboring states could continue to supply goods, a grape disease epidemic in California could only see relief via importation from abroad. Although Newton Pierce did not live long enough to witness the establishment of fledgling wineries in other states (Pinney 2005), he was correct in his assessment that California was perhaps the most important seat in global wine making. This has been borne out in the fact that the United States now supports a multibillion dollar travel, tourism, and consumption industry centered around viticulture and wine production (California Agricultural Statistics Service 2005; U.S. Department of Agriculture 2005). Returning to the plight of Pierce and his discovery, it would take many years post first report to understand the nature of the causal agent responsible for this yellowing, leaf drop, and general decline.

Failure to culture the causal agent of this disease led to experimentation surrounding the assumed "viral" agent from the perspective of transmission. A series of experiments were undertaken in which vector transmission was analyzed, and it was found that *Hordnia circellata*, one of the assumed leafhopper vectors, was an efficient agent of transfer (Severin 1949). Continued experimentation refined rates of transfer to *Vitis* genera cuttings and *Medicago sativa* (Severin 1950), and this expanded screening of transmission patterns led to the discovery of spread to alternative hosts which remained largely asymptomatic (Freitag 1951). Simultaneously, a number of questions were being asked about the

relatedness of xylem associated “viruses” described outside of *Vitis* species. The recognition of a possible link between phony peach disease (PDD) and Pierce’s disease (PD) was broached in a study of xylem limited pathogens (Esau 1949). Continued work on vector transmission patterns in phony peach disease hinted at the possibility of a similar agent causing disease in the Southern United States (Turner 1949; Turner and Pollard 1959). Additional observations were made regarding elm stands with a disease phenotype of leaf “scorching” (Wester and Jylkka 1959). Although clues were appearing in both the American landscape and disease reports, a true causal connection among these diseases was lacking. Even more problematic was the fact that an additional decade of research would be needed to reverse the generally held opinion that the potential causal agents were viral in nature.

In the early 1970s, an important stride was made when data supported the finding that the causal agent could be of mycoplasmal origin. This important study (Hopkins and Mortensen 1971) applied exogenous tetracycline hydrochloride and chlorotetracycline in a controlled regiment and noted the amelioration of symptoms in infected vines. The underlying assumption was that response to antibiotic compounds suggested a bacterial or bacterial-like origin. Shortly thereafter, microscopy based studies supported this data when Rickettsia-like organisms were associated with diseased tissue from grape, alfalfa, and plum (Goheen et al. 1973; Kitajima et al. 1975). Despite mounting evidence that these diseases were of bacterial origin, the causal agent remained unculturable by known media. In the meantime, a number of other important diseases were being both described and evaluated in reference to this “agent” currently described in grape, alfalfa, and plum. These diseases included: almond leaf scorch (Mircetich et al. 1976), elm, sycamore, and oak scorch (Hearon et al. 1980), and periwinkle wilt (McCoy et al. 1978).

Continued efforts to achieve a taxonomic designation were then provided by targeted organismal media development, and the confirmation of general biochemical subgroupings. Early media types were developed to culture Pierce’s disease from *Vitis* hosts, optimized for the same disease, and then extended for the isolation of seemingly more fastidious strains (Mission and Flora 1978; Davis et al. 1980; Davis et al. 1981). Modifications in recipes were also made to further target specific pathogen and host populations (Davis et al. 1983; Kostka, et al. 1986). Two such examples of the observed fastidiousness of the organism would be the extended duration for colony appearance in the mulberry specific isolate (18

days) and the periwinkle isolate (10-12 days) relative to the shorter duration for grape associated isolates. The next phase of disease characterization took the form of biochemical and molecular analyses. Another important but limited study was conducted using the strains from plum, peach, grape, and periwinkle, in which hybridizations relative to assumed distant taxa and DNA compositions were analyzed (Kamper et al. 1985). Hard but crude taxonomic breaks were observed, and suggested the presence of a new species, but supplemental strain analysis was needed for complete verification. The main biochemically based categorization of the organism that would come to be labelled *X. fastidiosa* was ultimately carried out with an impressive number of host associated isolates. Host associated population were derived from grape, almond, plum, peach, periwinkle, sycamore, ragweed, elm, mulberry, and oak to provide the most definitive biochemically based taxonomic description at the time. Agreements relative to assumed related Gram-negative classes of bacteria were observed in enzymatic activity, fatty acid distribution and profile, DNA probing, and 16S rRNA composition among other tests (Wells et al. 1987). Results derived therein confirmed the suspected uniqueness of this organism, and *X. fastidiosa* was adopted as a novel genus and species within the (γ) Gammaproteobacterial Class in 1987 (Wells et al. 1987).

Early Molecular Techniques and Population Profiling to the first *X. fastidiosa* genome release

Following successful colony isolations, early attempts at describing the underlying genetic composition of *X. fastidiosa* populations within various hosts came first through restriction fragment length polymorphism (RFLP) comparisons. Digest results reflected strong homogeneity among those strains believed to be causing Pierce's disease, and less homogeneity between strains isolated from other hosts (Chen et al. 1992). Despite being a progenitor method for techniques used in populations genetics, several of the findings reported by using RFLP analyses, especially the similarity among PD isolates, remains true even in the face of the mounting "omics" literature being produced today.

The need for more refined genetic fingerprinting, however, led to the use of SDS-PAGE profiling for isolate separation via the use of protein patterns (Bazzi et al. 1994). This research, however, was more diagnostic in nature, resulting from the need for accurate and rapid pathology based PD identification. Additional

polymerase chain reaction (PCR) work that appeared prior to understanding the genetic subtleties between host associated strains was also conducted across host populations (Minsavage et al. 1994). Simultaneously, enzyme-linked immunosorbent assays (ELISA) were also being explored, but again only provided a diagnostic binary (Hopkins and Adlerz 1988; Sherald, and Lei 1991). The mass use and popularity of the aforementioned PCR methods ushered in an era of amplicon based *X. fastidiosa* analysis for pathogen detection and early population delineation. Direct, base level, DNA analysis set the stage for addressing nuanced questions arising from a need to distinguish *X. fastidiosa* strains colonizing varied plant hosts.

Early attempts at resolving *X. fastidiosa* taxonomy and population differentiation at the molecular level relied upon randomly amplified polymorphic DNA (RAPD) assays (Pooler and Hartung 1995; Pooler and Hartung 1995; Pooler and Hartung 1995; Albibi et al. 1998; Chen et al 1999). One of the most comprehensive studies during this early period utilized seventeen isolates and fourteen RAPDs to characterize and phylogenetically position a number of Southern based strains (Chen et al. 1995). This study considered RAPD banding patterns unique to isolates derived from grape, plum, periwinkle, as well as a large number of Southern based oak associated groups. Despite refined clade formations resulting from RAPD analyses that are currently accepted as incongruous with current *X. fastidiosa* taxonomy, each main clade conformed to suggested host association. In other words, macro level grouping suggested strong host association among *X. fastidiosa* isolates independent of the phylogenetic subtleties currently accepted among a growing number of isolates. While RAPDs provided an early step in unravelling the genetic character of disparate isolates, technological breakthroughs and cost reduction led to targeted amplification, and the production of physically readable and alignable DNA sequence. Further sequencing technology proved especially important in *X. fastidiosa* research as continued reports in the late 1990s described disease phenotypes on several previously unknown hosts like coffee, Sugar Maple, and Sweetgum (Beretta et al. 1996; Hartman et al. 1996), and described disease radiation into unexpected locales like Southern Canada (Goodwin and Zhang 1997). One additional investigation worked on both the descriptive distribution of oak associated *X. fastidiosa* communities in Northeastern United States based vector populations and leveraged existing diagnostic primer sets (Minsavage et al. 1994; Pooler and

Hartung 1995; Henderson et al. 2001; Mehta and Rosato 2001) for pathogen confirmation (Zhang et al. 2011). The discovery and acquisition of DNA regions unique to *X. fastidiosa* populations confirmed the presence of the pathogen and potential disease outbreak, guiding appropriate management strategies. The 16S rRNA gene represented the first prominent genetic locus used for prokaryotic taxonomy and identification. Analysis at this locus provided initial insight into *X. fastidiosa* subspecies diversity in the form of meaningful site specific nucleotide dissimilarity (Chen et al. 2000; Chen et al. 2000). Recognition of additional polymorphic regions within topoisomerase classes (Champoux 2001) in conjunction with computational expansions began to shed further light on genetic differences between host specific strains. In other words, technological advances resulted in more robust phylogenetic analyses that were able to take the form of sophisticated nucleotide substitution modelling, thus discerning potentially unknown strain relationships. The 16S-23S intergenic spacer region (ITS), also used in early genetic analyses, played a prominent role in the differentiation of *X. fastidiosa* strains. This locus was especially important in the early work characterizing the taxonomic positioning of the oleander associated strain (Purcell et al. 1999). Substantive work was also being conducted at the expanded pathogen population level. For instance, previously considered questions regarding alternative hosts were being retested given the appearance of more sensitive modern tools. A major study at the time considered alternative hosts as potential inoculum sources and transmission points, finding that box elder, buckeye, bittersweet, dogwood, and English ivy were all capable of housing bacterial titer high enough to confirm pathogen presence and transmission (McElrone et al. 1999). A means had thus been provided to subject associated populations to comparative analyses with previously unknown rigor.

After approximately one hundred and four years of substantive *X. fastidiosa* research, the start of the second millennium witnessed a landmark event in bacterial phytopathology. In 2000, the genome for the citrus associated *X. fastidiosa* strain was released, providing both the first global perspective of the bacterium, and a rich base for successive genomics based undertakings (Simpson et al. 2000). Given a template for *X. fastidiosa* composition, the Brazilian citrus strain (9a5c) became the working model for understanding the genetic underpinnings of the organism via myriad approaches.

Early Genetic Characterizations, Multilocus Sequencing Analysis (MLSA), and Continued Population and Speciation Studies

The presentation of the first complete *X. fastidiosa* genome allowed mining of the underlying sequence for a wider array of genetic markers and sequence repeats. This allowed for an expansion beyond the limitations of RAPD based assays, and simultaneously built upon the findings of those early assays (Della Coletta-Filho et al. 2001). This quickly led to the genome sequencing of the first North American strain, an isolate associated with California grape and Pierce's Disease (Van Sluys et al. 2003). The counterpoint provided by the sequencing of this strain, (Temecula1), allowed for cross continental examination of two seemingly allopatric subspecies. An additional series of simple sequence repeat (SSR) and variable length tandem repeat(VNTR) assays appeared (Lin et al. 2004; Lin et al. 2005), and population structure and taxonomy among *X. fastidiosa* subspecies was being considered on a more sophisticated scientific level. The first proto assignment of subspecies taxa was carried out in 2004 (Schaad et al. 2004) and based mainly on the use of the 16S-23S (ITS) region for subspecies designation. Regardless, the assignments derived from ITS analysis proved highly consistent with future subsequent findings, (addressed later in the review), and officially proposed names for several of the *X. fastidiosa* subspecies. These taxonomic categories included the grape associated strain (*piercei* / *fastidiosa*), the citrus associated strain (*pauca*), and the broad host range almond, peach, plum and oak associated strain (*multiplex*).

The expanded knowledge provided by *X. fastidiosa* based genomics allowed researchers to co-opt a method initially used for characterization of the bacterium *N. meningitidis*. This method was referred to as multilocus sequence typing (MLST) (Maiden et al. 1998), and broached a previously unknown level of taxonomic scrutiny within *X. fastidiosa* research. Also referred to as multilocus sequencing analysis (MLSA), this technique allowed for more robust analyses of existing organisms, subspecies, and populations by selecting portions of genomes based on desired selectivity pressures (Kimura 1977). From this, one could perform an artificial concatemerization and treat the reduced genomic representation as a proxy for the chromosome at large. Two landmark studies (Schuenzel et al. 2005; Parker et al. 2012), spread seven years apart, then considered representative loci, and proved the repeatability of the early ITS based taxonomic placements (Schaad et al. 2004). The first of these considered nine "housekeeping" genes and

one variable locus used in Type IV pilus assembly (*pilU*) (Schuenzel et al. 2005). In addition, a companion paper by the same group considered clonality and population origin among various *X. fastidiosa* subspecies and arrived at a fourth taxonomic designation for the strain infecting the ornamental shrub oleander (Sally et al. 2005). Careful selections of loci based on defined dN/dS parameters confirmed the aforementioned *piercei/fastidiosa*, *pauca*, and *multiplex* grouping, and added a taxonomic designation for oleander labelled *sandyi*. Although observation about general dissimilarities of oleander based strains had been made in the past (Purcell et al. 1999), this robust approach confirmed prior suspicions. The second of these studies defined a locus set under higher selective pressures and designated it an “environmentally mediated” approach (MLSAS-E) (Parker et al. 2012). Despite adding a number of previously undescribed hosts and geographically segregated populations, the larger *X. fastidiosa* subspecies phylogenetic topology remained intact. This multilocus procedure was additionally repeated for an expanded analysis of a grape and oleander grouping and a coffee and citrus grouping to assess degrees of genetic similarity/dissimilarity persisting in various niches (Almeida et al. 2008; Yuan et al. 2010).

As haplotypic diversity was revealed from the sequencing of numerous *X. fastidiosa* specific loci, unknown population variants, even among similar hosts, drove the continuation of repeat markers as a means of population discrimination and tracking. Additional work along those lines was conducted to determine dissemination of these regions among numerous subspecies populations (Lin et al. 2007; Chen et al. 2010). The effectiveness of this method as a means of population differentiation is apparent at the time of the drafting this review, as a 2015 study using SSR markers in almond associated populations delineated strain representation in nearby and adjacent orchards (Lin et al. 2015). Association of various SSRs and underlying gene functionality may also speak to larger epidemiological issues. For instance, the environmentally mediated study identified several extended repeat regions at the *copB* locus, a protein known to be upregulated in the presence of the bacterioside Cu^{2+} (Rodrigues et al. 2008). Although further experimentation would be needed to validate the effect of this specific residue extension, speculation that it could play a role in tolerating bactericidal levels of copper is plausible.

In addition to the discovery of the previously mentioned oleander (*sandyi*) subspecies, continued

sampling of hosts bearing the *X. fastidiosa* disease phenotype turned up a number of surprising results. In 2007, *X. fastidiosa* was reported in the Southwestern United States ornamental *Chitalpa tashkinensis* and subsequent analysis via the 16S and ITS loci suggested the tentative assignment to a novel subspecies clade (Randall et al. 2007; Randall et al. 2009). Additionally, the supplementation of MLSA with earlier taxonomic methods confirmed the existence of a novel *X. fastidiosa* subspecies derived from the host mulberry (Hernandez-Martinez et al. 2006; Nunney et al. 2014). This subspecies, labelled *morus*, is now included in the standard *X. fastidiosa* taxonomic ranks based on the outcomes of several MLSA studies (Nunney et al. 2014; Hernandez-Martinez et al. 2006). Further, the extent to which bifurcations off the subspecies *multiplex* main clade constitute a single lineage still begs questioning. The *X. fastidiosa* subspecies *multiplex* was so-named because of its host plasticity (Schaad et al. 2004), but as more becomes known about the genomic character of isolates derived from novel hosts, this standard taxonomic thinking may require redefinition. For instance, the recent report of *X. fastidiosa* obtained from the host blueberry (Chang et al. 2009) requires such consideration. Although its lineage is traced to the *multiplex* main clade it does bifurcate away from other *multiplex* associated populations, such as those from the hosts oak, plum, and sycamore (Parker et al. 2012). In addition, a great deal of recent research regarding recombinants in the *multiplex* clade and among broader subspecies represents a novel angle in describing *X. fastidiosa* population genetics (Nunney et al. 2012; Nunney et al. 2013; Nunney et al. 2014a; Nunney et al. 2014b). The general thinking in several of these studies is that *X. fastidiosa* speciation is the result of several non-native introductions with the exception of subspecies *multiplex*. The extent to which non-native introduction, recombination, or an amalgam of the two speaks to a cladogenesis event remains unclear, but the continued unravelling of *X. fastidiosa* genomes promises to move toward greater organismal understanding.

The final integration of *X. fastidiosa* genotyping lies in understanding the movement of this “new world” disease into “old world” niches. Although there are earlier reports regarding the presence of *X. fastidiosa* in Europe based primarily on disease symptomology (Berisha et al. 1998), true confirmation of *X. fastidiosa* as the causal agent of the purported diseases has been marginal. This appears to also be the case in a very recent report using ELISA assays to describe the presence of oleander leaf scorch in

Lebanon (Temsah et al. 2015). Additional reports of *X. fastidiosa* causing disease on both grape and pear have surfaced in Taiwan as well (Leu et al. 1993; Su et al. 2013), however, the recent discovery and investigation into the presence of *X. fastidiosa* in Southern Italy on olive crops has yielded both definitive verification and a provisional genome sequence (Saponari et al. 2013; Cariddi et al. 2014; Giampetruzzi, et al. 2015). These examples support the underlying tenet in pathology of the eventual, global radiation of well-established and persistent pathogens.

Expansion of Isolate Sequencing, Large Scale Comparative Genomics, and the Extension into Targeted Locus-Based Research

The plethora of *X. fastidiosa* genomes that now inhabit GenBank, (eighteen to date) have lent themselves to expanded locus-based studies, but they simultaneously speak to the need for larger, full scale comparative projects. In addition to results stemming from several of the earlier, well annotated genome sequencing projects of the Brazilian citrus strain (9a5c), the Californian Pierce's Disease strain (Temecula1), and the Californian almond leaf scorch strains (M12 and M23) (Simpson et al. 2000; Van Sluys 2003; Chen et al. 2010), the growing number of newly sequenced genomes present an opportunity for further, deeper analyses of *X. fastidiosa* at the subspecies level and beyond. Included among the lesser studied and novel geographical host population genomic releases include GB514 (Pierce's Disease), Griffin-1 (bacterial leaf scorch of oak), Sy-Va (bacterial leaf scorch of sycamore), 6c and 32 (coffee leaf scorch), and the new taxonomic addition to previously existing *X. fastidiosa* phylogenies, Mul-MD (mulberry leaf scorch) (Scheiber et al. 2010; Chen et al. 2013; Guan et al. 2014; Alencar et al. 2014; Guan et al. 2014). Further, the recent availability of the previously mentioned "old world" genomes derived from both Taiwanese pear and Italian olive (Su et al. 2014; Giampetruzzi, et al. 2015) likely contain a wealth of insight as well. Finally, the sequencing of an avirulent type strain derived from an elderberry host, EB92-1 (Zhang et al. 2011), presents a novel counterpoint to its virulent relatives while promising to both unlock organism specific pathogenicity mechanisms and function in a potential biocontrol capacity. As a biocontrol agent, grape inoculation with the EB92-1 strains in various North American locales has

been moderately successful (Appel et al. 2010; Hopkins 2012; Compant et al. 2013), although a complete explanation regarding its non-pathogenic lifestyle has yet to be fully explained (Zhang et al. 2015).

Despite the large number of genomes mentioned herein, there is a surprising lack of full scale comparative genomics work in the body of *X. fastidiosa* literature. While this may be due to the informatics burden placed on researchers when dealing with the storing, sorting, and classifying of large amounts of genomic data, new technologies are arising to deal with such difficulties. One early study used the Brazilian citrus strain (9a5c) genome in isolation, and relied upon putative gene functionality to speculate on modes of pathogenicity within the bacterium (Lambais et al. 2000). A second study investigated comparative structures that incorporated all existing plant pathogenic bacterial genomes up to the year 2002, and used hierarchical categorizations build from such intrinsic qualities as GC content, comparative chromosomal size, and broad gene annotation binnings (Van Sluys et al. 2002). Although insightful, these earlier studies included the *X. fastidiosa* isolates 9a5c and Temecula1 as players in an ensemble cast, assigning them to the broad taxonomic headings of “Class: Gammaproteobacteria” and “Family: Xanthomonadaceae”. The completion of this comparative genomics project led to the first robust comparison between the genomes of *X. fastidiosa* isolates 9a5c and Temecula1 (Van Sluys et al. 2003). Additional studies soon followed, supplementing the existing body of comparisons with a California based almond host *multiplex* strain (Doddapaneni et al. 2006), and an oleander derived *sandyi* strain (Bhattacharyya et al. 2002). Another important study from this time (Moreira et al. 2005) involved the whole genome comparison between the *X. fastidiosa* genomes 9a5c and Temecula1, and the more well studied genomes of *Xanthomonas axonopodis* pv. *citri* and *Xanthomonas campestris* pv. *campestris*. Identification of pathogenicity mechanisms used by *X. fastidiosa* based upon gene inventory counts was suggestive of infective modes and proved an important facet of this study. In addition to these studies, the collective genomes resulted in creation of a searchable database was created where searches could be executed between several of the fully sequenced genomes (9a5c, Temecula1, M12, M23, Ann-1, and Dixon) (Varani et al. 2012). It should be noted that the database has not been actively maintained, and the additional twelve *X. fastidiosa* genomes housed at GenBank are not searchable through this portal. Another recent study compared polymorphic orthologs and speculative paralogs based on both the

recently described *X. fastidiosa* coffee strains (6c and 32), and both Brazilian citrus strain (9a5c), and the California grape strain (Temecula1) (Barbosa et al. 2015). Finally, the most recent comparative project to date considered large scale genomic differences between the elderberry associated biocontrol strain (EB92-1) to identify pathogenicity effectors (Zhang et al. 2015). Despite movement towards more integrated cross subspecies analyses, a great deal of genome to genome analysis remains to be done.

Regardless of the lack of more iterative comparative studies, the presence of a large number of draft sequences nonetheless allows for selective mining and drives general unanswered questions in *X. fastidiosa* biology (Bhattacharyya et al. 2002). Current findings regarding expanded knowledge of *X. fastidiosa* include identification of mechanisms presumed to be associated with host colonization. These include toxin production in which colisin appears to produce an assumed antagonism, attachment mechanisms by way of the role of both fimbrial and afimbrial adhesions, regulatory mechanisms of vir gene products, diffusible signal factors (DSFs) and their role in cell to cell signaling, and recent domain mining for putative pathogenicity functionality deletion assays (Moreira et al. 2005; Cascales et al. 2007; Feil et al. 2003; Lindow et al. 2005; Feil et al. 2007; Newman et al. 2004; Chatterjee et al. 2008; Chatterjee et al. 2010; Cursino et al. 2015).

Another interesting aspect of whole genome analyses is the potential for subspecies specific growth media. Although it has been some ten years since publications have appeared in this area, stymied growth of more fastidious isolates is still a potential obstacle that continues to hinder studies. In one study, post pathway analysis proved that tailored media could support statistically significant growth spikes relative to conventional formulas (de Macedo Lemos et al. 2003). Again, this is of significance due to the early observation of lengthy doubling times that often exceed twenty four hours for some subspecies (Wells et al. 1987; Feil et al. 2001). Despite this impressive body of work spanning numerous aspects of prokaryotic biology, many aspects of *X. fastidiosa* are still poorly understood, and continued large scale variegated host derived analyses are needed.

Prophage Elements in *X. fastidiosa* Genomes and Significance in Bacterial Population Distribution and

Structure

Annotation of the first *X. fastidiosa* genome (9a5c) in 2000, led to the recognition that a small but significant portion of the chromosome contained a multitude of deposited prophage sequence (Simpson et al. 2000). Subsequent genome sequencing projects (VanSluys et al. 2003, Bhattacharyya et al. 2002; Doddapaneni et al. 2006) revealed similar phage deposits, even among populations of varied plant hosts. Earlier microscopy based work had suggested the association of virions with *X. fastidiosa*, but it was then considered tangential to the more important task of characterizing the bacterium itself (Kitajima et al. 1975). Several years after the initial mention of virion association, further microscopy work proposed that the morphological nature of the viral assemblies conformed to those of the *Podoviridae* family (Chen et al. 2008), thus giving the associated virions a taxonomic designation. This was quickly followed by the release of the first plaque forming *X. fastidiosa* associated viral genome, Xfas53 (Summer et al. 2010). Additional studies showed that while several of the prophage elements appeared Podophage specific from the molecular perspective, a degree of hybridization was observed and noted to be a further significant contributor to differentiation among disparate host *X. fastidiosa* isolates (Varani et al. 2008). This hybridization revealed prophage segments with a high degree of similarity to those observed in the viral families *Myoviridae* and *Siphoviridae* as well (Varani et al. 2013; Summer et al. 2010; Chen & Civerolo. 2008; de Mello Varani et al. 2008). These observations provided a direct segue into continued studies of both practicality and significance of these often overlooked segments within the growing collection of *X. fastidiosa* genomic sequences.

The multitude of prophage sequences in the large number of *X. fastidiosa* genomes currently available has yet to be described at an expanded level. A recent study by Varani et al. (2008) profiled phage sequences in the genomes of isolates derived from citrus, grape, oleander, and almond, but this study largely focused its attention on integrase specific regions, and expanded prophage integrase islands constituting significant portions of individual *X. fastidiosa* genomes (Varani et al. 2008). Several interesting companion experiments were also conducted, including a limited analysis of phage related gene upregulation in response to thermal stresses (Varani et al. 2008). Despite the array of data obtained, phylogenetic

positioning was performed on only the aforementioned integrase regions, with only passing descriptive mention of other prophage regions. These regions may further constitute yet unknown subspecies divisions or describe important evolutionary events related to this bacterium. Because of this dearth of literature analyzing prophage regions across *X. fastidiosa* subspecies, continued study regarding these genomic novelties should progress.

The aforementioned observations regarding prophage regions as sectors of genetic diversity between *X. fastidiosa* subspecies effectively states that they and thus likely represent novel coevolutionary relationships that describe hitherto unknown relationships between subspecies inhabiting either different hosts or geographically segregated hosts. Further recent exploration into this area of *X. fastidiosa* research has suggested other viral communities as yet undescribed inhabitants of the bacterium as well (Ahern 2013). This finding, coupled with the recent revelation of vector co-infection and viral based attenuation of *X. fastidiosa* titre (Browmick et al. 2013; Das et al. 2013), beckons further study from a potential control standpoint. In other words, the implied limited infection cycles and limited reinfection cycles suggests this as a burgeoning area of *X. fastidiosa* study where greater understanding has the potential to reduce the spread of this pathogen. This is of increasing economic importance as difficulties in managing this disease relative to high value commodity crops like grape have been well enumerated (Almeida et al. 2005). This could be especially relevant as continued genomic profiling could hint at artificial induction from lysogeny to lytic states, thereby reducing bacterial titre and expanding current limitation in the realm of *X. fastidiosa* biocontrol strategies.

REFERENCES

- Ahern, S. J. (2013). Novel Virulent Phages for *Xylella fastidiosa* and Other Members of the Xanthomonadaceae (Doctoral dissertation).
- Albib, R., Chen, J., Lamikanra, O., Banks, D., Jarret, R. L., & Smith, B. J. (1998). RAPD fingerprinting *Xylella fastidiosa* Pierce's disease strains isolated from a vineyard in North Florida. *FEMS microbiology letters*, 165(2), 347-352.
- Alencar, V. C., Barbosa, D., Santos, D. S., Oliveira, A. C. F., de Oliveira, R. C., & Nunes, L. R. (2014). Genomic sequencing of two coffee-infecting strains of *Xylella fastidiosa* isolated from Brazil. *Genome announcements*, 2(1), e01190-13.
- Almeida, R. P., Blua, M. J., Lopes, J. R., & Purcell, A. H. (2005). Vector transmission of *Xylella fastidiosa*: applying fundamental knowledge to generate disease management strategies. *Annals of the Entomological Society of America*, 98(6), 775-786.
- Almeida, R. P., Nascimento, F. E., Chau, J., Prado, S. S., Tsai, C. W., Lopes, S. A., & Lopes, J. R. (2008). Genetic structure and biology of *Xylella fastidiosa* strains causing disease in citrus and coffee in Brazil. *Applied and Environmental Microbiology*, 74(12), 3690-3701.
- Appel, D. N., Black, M., Kamas, J., Hopkins, D. L., Palacios, K., Vineyard, P., & Brenham, T. X. (2010). BIOLOGICAL CONTROL TRIALS WITH EB92-1 IN TEXAS.
- Barbosa, D., Alencar, V. C., Santos, D. S., de Freitas Oliveira, A. C., de Souza, A. A., Colleta-Filho, H. D., ... & Nunes, L. R. (2015). Comparative genomic analysis of coffee-infecting *Xylella fastidiosa* strains isolated from Brazil. *Microbiology, mic-0*.
- Bhattacharyya, A., Stilwagen, S., Ivanova, N., D'Souza, M., Bernal, A., Lykidis, A., ... & Kyrpides, N. C. (2002). Whole-genome comparative analysis of three phytopathogenic *Xylella fastidiosa* strains. *Proceedings of the National Academy of Sciences*, 99(19), 12403-12408.
- Bhattacharyya, A., Stilwagen, S., Reznik, G., Feil, H., Feil, W. S., Anderson, I., ... & Predki, P. F. (2002). Draft sequencing and comparative genomics of *Xylella fastidiosa* strains reveal novel biological insights. *Genome research*, 12(10), 1556-1563.
- Bazzi, C., Stefani, E., & Zaccardelli, M. (1994). SDS-PAGE: a tool to discriminate *Xylella fastidiosa* from other endophytic grapevine bacteria*. *EPPO Bulletin*, 24(1), 121-127.
- Beretta, M. J. G., Harakava, R., Chagas, C. M., Derrick, K. S., Barthe, G. A., Ceccardi, T. L., ... & Ribeiro, I. A. (1996). First report of *Xylella fastidiosa* in coffee. *Plant disease*, 80(7).

- Berisha, B., Chen, Y. D., Zhang, G. Y., Xu, B. Y., & Chen, T. A. (1998). Isolation of Peirce's disease bacteria from grapevines in Europe. *European Journal of Plant Pathology*, 104(5), 427-433.
- Bhowmick, T. S., Das, M., Heinz, K. M., Krauter, P. C., & Gonzalez, C. (2013, June). Transmission of phage by glassy-winged sharpshooter. In *PHYTOPATHOLOGY* (Vol. 103, No. 6, pp. 16-16). 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA: AMER PHYTOPATHOLOGICAL SOC.
- California Agricultural Statistics Service (2005) California Agricultural Statistics (Calif. Agric. Stat. Service, Sacramento).
- Cariddi, C., Saponari, M., Boscia, D., De Stradis, A., Loconsole, G., Nigro, F., ... & Martelli, G. P. (2014). Isolation of a *Xylella fastidiosa* strain infecting olive and oleander in Apulia, Italy. *Journal of Plant Pathology*, 96(3), 1-5.
- Cascales, E., Buchanan, S. K., Duché, D., Kleanthous, C., Lloubes, R., Postle, K., ... & Cavard, D. (2007). Colicin biology. *Microbiology and Molecular Biology Reviews*, 71(1), 158-229.
- Champoux, J. J. (2001). DNA topoisomerases: structure, function, and mechanism. *Annual review of biochemistry*, 70(1), 369-413.
- Chang, C. J., Donaldson, R., Brannen, P., Krewer, G., & Boland, R. (2009). Bacterial leaf scorch, a new blueberry disease caused by *Xylella fastidiosa*. *HortScience*, 44(2), 413-417.
- Chatterjee, S., Killiny, N., Almeida, R. P., & Lindow, S. E. (2010). Role of cyclic di-GMP in *Xylella fastidiosa* biofilm formation, plant virulence, and insect transmission. *Molecular plant-microbe interactions*, 23(10), 1356-1363.
- Chatterjee, S., Wistrom, C., & Lindow, S. E. (2008). A cell-cell signaling sensor is required for virulence and insect transmission of *Xylella fastidiosa*. *Proceedings of the National Academy of Sciences*, 105(7), 2670-2675.
- Chen, J., Chang, C. J., Jarret, R. L., & Gawel, N. (1992). Genetic variation among *Xylella fastidiosa* strains. *Phytopathology*, 82(9), 973-977.
- Chen, J., Lamikanra, O., Chang, C. J., & Hopkins, D. L. (1995). Randomly amplified polymorphic DNA analysis of *Xylella fastidiosa* Pierce's disease and oak leaf scorch pathotypes. *Applied and environmental microbiology*, 61(5), 1688-1690.
- Chen, J., & Albibi, R. (1999). Homogeneity of *Xylella fastidiosa* Pierce's Disease Strains from Bunch Grapes and Muscadines in North Florida. In *PROCEEDINGS-FLORIDA STATE HORTICULTURAL SOCIETY* (Vol. 112, pp. 185-186).
- Chen, J., Banks, D., Jarret, R. L., Chang, C. J., & Smith, B. J. (2000). Use of 16S rDNA sequences as signature characters to identify *Xylella fastidiosa*. *Current microbiology*, 40(1), 29-33.
- Chen, J., Jarret, R. L., Qin, X., Hartung, J. S., Banks, D., Chang, C. J., & Hopkins, D. L. (2000). 16S rDNA sequence analysis of *Xylella fastidiosa* strains. *Systematic and applied microbiology*, 23(3), 349-354.
- Chen, J., & Civerolo, E. L. (2008). Morphological evidence for phages in *Xylella fastidiosa*. *Virology*, 375, 75.
- Chen, J. (2010, June). Evaluation of tandem repeat polymorphisms between two pathogenically similar strains of *Xylella fastidiosa* from almond and grape in California. In *Phytopathology* (Vol. 100, No. 6, pp. S24-S24). 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA: AMER PHYTOPATHOLOGICAL SOC.

- 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(17), 4534-4534.
- 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 announcements*, 1(5), e00756-13.
- Compant, S., Brader, G., Muzammil, S., Sessitsch, A., Lebrihi, A., & Mathieu, F. (2013). Use of beneficial bacteria and their secondary metabolites to control grapevine pathogen diseases. *BioControl*, 58(4), 435-455.
- Cursino, L., Athinuwat, D., Patel, K. R., Galvani, C. D., Zaini, P. A., Li, Y., ... & Mowery, P. (2015). Characterization of the *Xylella fastidiosa* PD1671 Gene Encoding Degenerate c-di-GMP GGDEF/EAL Domains, and Its Role in the Development of Pierce's Disease. *PloS one*, 10(3), e0121851.
- Das, M., Bhowmick, T. S., Ahern, S. J., Young, R. F., & Gonzalez, C. (2013, June). Therapeutic and prophylactic application of phages to control Pierce's disease. In *PHYTOPATHOLOGY* (Vol. 103, No. 6, pp. 34-34). 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA: AMER PHYTOPATHOLOGICAL SOC.
- Davis, M. J., Purcell, A. H., & Thomson, S. V. (1980). Isolation media for the Pierce's disease bacterium. *Phytopathology*, 70(5), 425-429.
- Davis, M. J., French, W. J., & Schaad, N. W. (1981). Axenic culture of the bacteria associated with phony disease of peach and plum leaf scald. *Current Microbiology*, 6(5), 309-314.
- Davis, M. J., Raju, B. C., Bransky, R. H., Lee, R. F., Timmer, L. W., Norris, R. C., & McCoy, R. E. (1983). Periwinkle wilt bacterium: axenic culture, pathogenicity and relationships to other Gram-negative, xylem-inhabiting bacteria. *Phytopathology*, 73(11), 1510-1515.
- Della Coletta-Filho, H., Takita, M. A., de Souza, A. A., Aguilar-Vildoso, C. I., & Machado, M. A. (2001). Differentiation of strains of *Xylella fastidiosa* by a variable number of tandem repeat analysis. *Applied and environmental microbiology*, 67(9), 4091-4095.
- Doddapaneni, H., Yao, J., Lin, H., Walker, M. A., & Civerolo, E. L. (2006). Analysis of the genome-wide variations among multiple strains of the plant pathogenic bacterium *Xylella fastidiosa*. *BMC genomics*, 7(1), 225.
- Esau, K. (1948). Anatomic effects of the viruses of Pierce's disease and phony peach. University of California.
- Feil, H., & Purcell, A. H. (2001). Temperature-dependent growth and survival of *Xylella fastidiosa* in vitro and in potted grapevines. *Plant Disease*, 85(12), 1230-1234.
- Feil, H., Feil, W. S., Detter, J. C., Purcel, A. H., & Lindow, S. E. (2003). Site-directed disruption of the *fimA* and *fimF* fimbrial genes of *Xylella fastidiosa*. *Phytopathology*, 93(6), 675-682.
- 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(3), 318-324.
- Freitag, J. H. (1951). Host range of the Pierce's disease virus of grapes as determined by insect transmission. *Phytopathology*, 41(10), 920-934.

- Giampetruzzi, A., Chiumenti, M., Saponari, M., Donvito, G., Italiano, A., Loconsole, G., ... & Saldarelli, P. (2015). Draft genome sequence of the *Xylella fastidiosa* CoDiRO strain. *Genome announcements*, 3(1), e01538-14.
- Goheen, A. C., Nyland, G., & Lowe, S. K. (1973). Association of a rickettsialike organism with Pierce's disease of grapevines and alfalfa dwarf and heat therapy of the disease in grapevines. *Phytopathology*, 63(3), 341-345.
- Goodwin, P. H., & Zhang, S. (1997). Distribution of *Xylella fastidiosa* in southern Ontario as determined by the polymerase chain reaction. *Canadian Journal of Plant Pathology*, 19(1), 13-18.
- Guan, W., Shao, J., Davis, R. E., Zhao, T., & Huang, Q. (2014). Genome sequence of a *Xylella fastidiosa* strain causing sycamore leaf scorch disease in Virginia. *Genome announcements*, 2(4), e00773-14.
- Guan, W., Shao, J., Zhao, T., & Huang, Q. (2014). Genome sequence of a *Xylella fastidiosa* strain causing mulberry leaf scorch disease in Maryland. *Genome announcements*, 2(2), e00916-13.
- Hartman, J. R., Jarlfors, U. E., Fountain, W. M., & Thomas, R. (1996). First report of bacterial leaf scorch caused by *Xylella fastidiosa* on sugar maple and sweetgum. *Plant disease*.
- Hearon, S. S., Sherald, J. L., & Kostka, S. J. (1980). Association of xylem-limited bacteria with elm, sycamore, and oak leaf scorch. *Canadian Journal of Botany*, 58(18), 1986-1993.
- Hendson, M., Purcell, A. H., Chen, D., Smart, C., Guilhabert, M., & Kirkpatrick, B. (2001). Genetic diversity of Pierce's disease strains and other pathotypes of *Xylella fastidiosa*. *Applied and Environmental Microbiology*, 67(2), 895-903.
- Hernandez-Martinez, R., Pinckard, T. R., Costa, H. S., Cooksey, D. A., & Wong, F. P. (2006). Discovery and characterization of *Xylella fastidiosa* strains in southern California causing mulberry leaf scorch. *Plant disease*, 90(9), 1143-1149.
- Hopkins, D. L., & Mortensen, J. A. (1971). Suppression of Pierce's disease symptoms by tetracycline antibiotics. *Plant Disease Reporter*, 55(7), 610-612.
- Hopkins, D. L., & Adlerz, W. C. (1988). Natural hosts of *Xylella fastidiosa* in Florida. *Plant Disease*, 72(5), 429-431.
- Hopkins, D. L. (2012, July). Long-term control of Pierce's disease in various grape genotypes with a benign strain of *Xylella fastidiosa*. In *PHYTOPATHOLOGY* (Vol. 102, No. 7, pp. 55-55). 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA: AMER PHYTOPATHOLOGICAL SOC.
- Kamper, S. M., French, W. J., & DeKloet, S. R. (1985). Genetic relationships of some fastidious xylem-limited bacteria. *International journal of systematic bacteriology*, 35(2), 185-188.
- Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.
- Kitajima, E. W., Bakarcic, M., & Fernandez-Valiela, M. V. (1975). Association of rickettsia-like bacteria with plum leaf scald disease. *Phytopathology*, 65, 476-479.
- Kostka, S. J., Tattar, T. A., Sherald, J. L., & Hurtt, S. S. (1986). Mulberry leaf scorch, new disease caused by a fastidious, xylem-inhabiting bacterium. *Plant disease*, 70(7), 690-693.

- Lambais, M. R., Goldman, M. H., Camargo, L. E., & Goldman, G. H. (2000). A genomic approach to the understanding of *Xylella fastidiosa* pathogenicity. *Current opinion in microbiology*, 3(5), 459-462.
- de Macedo Lemos, E. G., Alves, L. M. C., & Campanharo, J. C. (2003). Genomics-based design of defined growth media for the plant pathogen *Xylella fastidiosa*. *FEMS microbiology letters*, 219(1), 39-45.
- Leu, L. S., & Su, C. C. (1993). Isolation, cultivation, and pathogenicity of *Xylella fastidiosa*, the causal bacterium of pear leaf scorch disease in Taiwan. *Plant Disease*, 77(6), 642-646.
- Lin, H., Walker, A., & Civerolo, E. (2004, December). Development of SSR markers for genotyping and assessing the genetic diversity of *Xylella fastidiosa* in California. In *Proceedings, Pierce's Disease Research Symposium*, M. Athar Tariq, S. Oswalt, P. Blincoe, R. Spencer, L. Houser, A. Ba, and T. Esser (eds.) (pp. 7-10).
- Lin, H., Civerolo, E. L., Hu, R., Barros, S., Francis, M., & Walker, M. A. (2005). Multilocus simple sequence repeat markers for differentiating strains and evaluating genetic diversity of *Xylella fastidiosa*. *Applied and environmental microbiology*, 71(8), 4888-4892.
- Lin, H., Thimmiraju, S., Walker, A., Stenger, D., Civerolo, E. L., & Groves, R. L. (2007, December). Hierarchical Analysis and Diversity Studies of *Xylella fastidiosa* Populations in California by Multi-locus Simple Sequence Repeat Markers. In *CDFA Pierce's Disease Control Program Research Symposium* (pp. 144-147).
- Lin, H., Islam, M. S., Rosa, J. C. L., Civerolo, E. L., & Groves, R. L. (2015). Population structure of *Xylella fastidiosa* associated with almond leaf scorch disease in the San Joaquin Valley of California. *Phytopathology*, (ja).
- Lindow, S. E., & Feil, H. (2005). Effects of fimbrial (FimA, FimF) and afimbrial (XadA, HxfB) adhesins on the adhesion of *Xylella fastidiosa* to surfaces. In *proceedings of the 2005 Pierce's Disease Research Symposium*. California Department of Food and Agriculture, Sacramento, CA (pp. 173-176).
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., ... & Spratt, B. G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6), 3140-3145.
- McElrone, A. J., Sherald, J. L., & Pooler, M. R. (1999). Identification of alternative hosts of *Xylella fastidiosa* in the Washington, DC, area using nested polymerase chain reaction (PCR). *Journal of Arboriculture*, 25, 258-263.
- McCoy, R. E., Thomas, D. L., Tsai, J. H., & French, W. J. (1978). Periwinkle wilt, a new disease associated with xylem delimited rickettsialike bacteria transmitted by a sharpshooter [*Oncometopia nigricans*, *Catharanthus roseus*, insect vectors]. *Plant Disease Reporter*.
- Mehta, A., & Rosato, Y. B. (2001). Phylogenetic relationships of *Xylella fastidiosa* strains from different hosts, based on 16S rDNA and 16S-23S intergenic spacer sequences. *International journal of systematic and evolutionary microbiology*, 51(2), 311-318.
- Minsavage, G. V., Thompson, C. M., Hopkins, D. L., Leite, R. M. V. B. C., & Stall, R. E. (1994). Development of a polymerase chain reaction protocol for detection of *Xylella fastidiosa* in plant tissue. *Phytopathology*, 84(5), 456-461.
- Mircetich, S. M., Lowe, S. K., Moller, W. J., & Nyland, G. (1976). Etiology of almond leaf scorch disease and transmission of the causal agent. *Phytopathology*, 66(1), 17-24.

- Mission, R. C., & Flora, C. (1978). Pierce's Disease of Grapevines: Isolation of the Causal Bacterium. *Science*, 199, 6.
- Moreira, L. M., De Souza, R. F., Digiampietri, L. A., Da Silva, A. C., & Setubal, J. C. (2005). Comparative analyses of *Xanthomonas* and *Xylella* complete genomes. *Omics: a journal of integrative biology*, 9(1), 43-76.
- 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. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6), 1737-1742.
- Nunney, L., Yuan, X., Bromley, R. E., & Stouthamer, R. (2012). Detecting genetic introgression: high levels of intersubspecific recombination found in *Xylella fastidiosa* in Brazil. *Applied and environmental microbiology*, 78(13), 4702-4714.
- Nunney, L., Vickerman, D. B., Bromley, R. E., Russell, S. A., Hartman, J. R., Morano, L. D., & Stouthamer, R. (2013). Recent evolutionary radiation and host plant specialization in the *Xylella fastidiosa* subspecies native to the United States. *Applied and environmental microbiology*, 79(7), 2189-2200.
- Nunney, L., Hopkins, D. L., Morano, L. D., Russell, S. E., & Stouthamer, R. (2014). Intersubspecific recombination in *Xylella fastidiosa* strains native to the United States: infection of novel hosts associated with an unsuccessful invasion. *Applied and environmental microbiology*, 80(3), 1159-1169.
- Nunney, L., Schuenzel, E. L., Scally, M., Bromley, R. E., & Stouthamer, R. (2014). Large-scale intersubspecific recombination in the plant-pathogenic bacterium *Xylella fastidiosa* is associated with the host shift to mulberry. *Applied and environmental microbiology*, 80(10), 3025-3033.
- Parker, J. K., Havird, J. C., & De La Fuente, L. (2012). Differentiation of *Xylella fastidiosa* strains via multilocus sequence analysis of environmentally mediated genes (MLSA-E). *Applied and environmental microbiology*, 78(5), 1385-1396.
- Pierce, N. B. (1892). The California vine disease: a preliminary report of investigations (No. 2). US Government Printing Office.
- Pinney, T. (2005). A history of wine in America: from prohibition to the present (Vol. 2). Univ of California Press.
- Pooler, M. R., & Hartung, J. S. (1995). RAPDs are Useful for Genetic Analysis of *Xylella fastidiosa* and for Development of Strain-specific PCR Primers. *HortScience*, 30(4), 783-783.
- Pooler, M. R., & Hartung, J. S. (1995). Genetic relationships among strains of *Xylella fastidiosa* from RAPD-PCR data. *Current microbiology*, 31(2), 134-137.
- Pooler, M., & Hartung, J. S. (1995). Genetic relationship among strains of *Xylella fastidiosa* based on RAPD-PCR data. *HortScience*, 30(2), 192-192.
- Purcell, A. H., & Hopkins, D. L. (1996). Fastidious xylem-limited bacterial plant pathogens. *Annual review of phytopathology*, 34(1), 131-151.
- Purcell, A. H., Saunders, S. R., Henderson, M., Grebus, M. E., & Henry, M. J. (1999). Causal role of *Xylella fastidiosa* in oleander leaf scorch disease. *Phytopathology*, 89(1), 53-58.

- Randall, J. J., Radionenko, M., French, J. M., Olsen, M. W., Goldberg, N. P., & Hanson, S. F. (2007). *Xylella fastidiosa* detected in New Mexico in chitalpa, a common landscape ornamental plant. *Plant Disease*, 91(3), 329-329.
- Randall, J. J., Goldberg, N. P., Kemp, J. D., Radionenko, M., French, J. M., Olsen, M. W., & Hanson, S. F. (2009). Genetic analysis of a novel *Xylella fastidiosa* subspecies found in the southwestern United States. *Applied and environmental microbiology*, 75(17), 5631-5638.
- Rodrigues, C. M., Takita, M. A., Coletta-Filho, H. D., Olivato, J. C., Caserta, R., Machado, M. A., & De Souza, A. A. (2008). Copper resistance of biofilm cells of the plant pathogen *Xylella fastidiosa*. *Applied microbiology and biotechnology*, 77(5), 1145-1157.
- Saponari, M., Boscia, D., Nigro, F., & Martelli, G. P. (2013). Identification of DNA sequences related to *Xylella fastidiosa* in oleander, almond and olive trees exhibiting leaf scorch symptoms in Apulia (Southern Italy). *Journal of Plant Pathology*, 95(3).
- Scally, M., Schuenzel, E. L., Stouthamer, R., & Nunney, L. (2005). Multilocus sequence type system for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation to clonal diversity. *Applied and environmental microbiology*, 71(12), 8491-8499.
- Schaad, N. W., Postnikova, E., Lacy, G., Fatmi, M. B., & Chang, C. J. (2004). *Xylella fastidiosa* subspecies: *X. fastidiosa* subsp. *piercei*, subsp. nov., *X. fastidiosa* subsp. *multiplex* subsp. nov., and *X. fastidiosa* subsp. *pauca* subsp. nov. *Systematic and applied microbiology*, 27(3), 290-300.
- Schreiber IV, 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(3), 479-483.
- 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*. *Applied and environmental microbiology*, 71(7), 3832-3839.
- Severin, H. H. (1949). Transmission of the virus of Pierce's disease of grapevines by leafhoppers. *Hilgardia*, 19(6), 190-202.
- SEVERIN, H. (1950). Spittle-insect vectors of Pierce's disease virus. II. Life history and virus transmission. *Hilgardia*, 19(11), 357-376.
- Sherald, J. L., & Lei, J. D. (1991). Evaluation of a rapid ELISA test kit for detection of *Xylella fastidiosa* in landscape trees. *Plant Disease*, 75(2), 200-203.
- Simpson, A. J. G., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., ... & Krieger, J. E. (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406(6792), 151-157.
- Su, C. C., Chang, C. J., Chang, C. M., Shih, H. T., Tzeng, K. C., Jan, F. J., ... & Deng, W. L. (2013). Pierce's disease of grapevines in Taiwan: isolation, cultivation and pathogenicity of *Xylella fastidiosa*. *Journal of Phytopathology*, 161(6), 389-396.
- Su, C. C., Deng, W. L., Jan, F. J., Chang, C. J., Huang, H., & Chen, J. (2014). Draft genome sequence of *Xylella fastidiosa* pear leaf scorch strain in Taiwan. *Genome announcements*, 2(2), e00166-14.
- Summer, E. J., Enderle, C. J., Ahern, S. J., Gill, J. J., Torres, C. P., Appel, D. N., ... & Gonzalez, C. F. (2010). Genomic and biological analysis of phage Xfas53 and related prophages of *Xylella fastidiosa*. *Journal of*

bacteriology, 192(1), 179-190.

Temsah, M., Hanna, L., & Saad, A. (2015). First Report of *Xylella fastidiosa* associated with Oleander Leaf Scorch in Lebanon. *Journal of Crop Protection*, 4(1), 131-137.

Turner, W. F. (1949). Insect vectors of phony peach disease. *Science*, 109(2822), 87-88.

Turner, W. F., & Pollard, H. N. (1959). Insect transmission of phony peach disease (Vol. 1193). US Dept. of Agriculture.

U.S. Department of Agriculture (2005) Agricultural Statistics (Natl. Agric. Stat. Service, Washington, DC).

Van Sluys, M. A., Monteiro-Vitorello, C. B., Camargo, L. E. A., Menck, C. F. M., Da Silva, A. C. R., Ferro, J. A., ... & Simpson, A. J. (2002). Comparative genomic analysis of plant-associated bacteria. *Annual review of phytopathology*, 40(1), 169-189.

Van Sluys, M. A., De Oliveira, M. C., Monteiro-Vitorello, C. B., Miyaki, C. Y., Furlan, L. R., Camargo, L. E. A., ... & Tsukumo, F. (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. *Journal of Bacteriology*, 185(3), 1018-1026.

de Mello Varani, A., Souza, R. C., Nakaya, H. I., De Lima, W. C., Paula de Almeida, L. G., Kitajima, E. W., ... & Van Sluys, M. A. (2008). Origins of the *Xylella fastidiosa* prophage-like regions and their impact in genome differentiation. *PLoS One*, 3(12), e4059.

Varani, A. M., Monteiro-Vitorello, C. B., Nakaya, H. I., & Van Sluys, M. A. (2013). The role of prophage in plant-pathogenic bacteria. *Annual review of phytopathology*, 51, 429-451.

Varani, A. M., Monteiro-Vitorello, C. B., de Almeida, L. G., Souza, R. C., Cunha, O. L., Lima, W. C., ... & Vasconcelos, A. T. (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(1), 149-152.

Wells, J. M., Raju, B. C., Hung, H. Y., Weisburg, W. G., Mandelco-Paul, 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(2), 136-143.

Wester, H. V., & Jylkka, E. W. (1959). Elm Scorch, graft transmissible virus of American elm. *Plant Dis Rep*, 43, 519.

Yuan, X., Morano, L., Bromley, R., Spring-Pearson, S., Stouthamer, R., & Nunney, L. (2010). Multilocus sequence typing of *Xylella fastidiosa* causing Pierce's disease and oleander leaf scorch in the United States. *Phytopathology*, 100(6), 601-611.

Zhang, S., Chakrabarty, P. K., Fleites, L. A., Rayside, P. A., Hopkins, D. L., & Gabriel, D. W. (2015). Three New Pierce's Disease Pathogenicity Effectors Identified Using *Xylella fastidiosa* Biocontrol Strain EB92-1. *PLoS one*, 10(7), e0133796.

Zhang, J., Lashomb, J., Gould, A., & Hamilton, G. (2011). Cicadomorpha insects associated with bacterial leaf scorch infected oak in central New Jersey. *Environmental entomology*, 40(5), 1131-1143.

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(19), 5576-5577.

Zhang, S., Chakrabarty, P. K., Fleites, L. A., Rayside, P. A., Hopkins, D. L., & Gabriel, D. W. (2015). Three New Pierce's Disease Pathogenicity Effectors Identified Using *Xylella fastidiosa* Biocontrol Strain EB92-1. *PloS one*, 10(7), e0133796.

CHAPTER 2. The genetic composition of oak associated *Xylella fastidiosa* populations in the
Northeastern and Mid-Atlantic United States

ABSTRACT

Xylella fastidiosa is an important plant pathogen that continues to undergo radiation events, breaching assumed geographical boundaries. In an effort to understand aspects of understudied *X. fastidiosa* populations and their unique genetic signatures relative to well described isolates, regionally defined, symptomatic *Quercus sp.* (oak) were sampled and subjected to multilocus based phylogenetic analysis and expanded insertion/deletion profiling. Leveraging previously described loci, 37 selected *Quercus* associated bacterial samples were analyzed, producing novel haplotypes and supplementing current limitations regarding the presence of existing genotypes in *X. fastidiosa* oak associated populations. Major findings include a geographically based phylogenetic division between Northeastern/Mid-Atlantic oak derived bacterial populations and Southern United States derived counterparts, and previously unknown insertion/deletion allelic recoveries among the same Northeastern/Mid-Atlantic communities. These results speak to potential spatiotemporal considerations as well as supplement existing information regarding specific genotypic distributions of the oak based pathovar. Information provided herein will likely prove germane in disease tracking, genotypic origin, and better guided disease control strategies.

INTRODUCTION

The vectored phytopathogen, *Xylella fastidiosa* (Wells et al. 1987), is associated with leaf scorch, vascular wilt, and broad range host decline (Purcell et al. 1996). Symptomatic diseases include, but are not limited to: citrus variegated chlorosis (CVC) (Lee et al 1993), Pierce's Disease (PD) (Pierce 1892), almond leaf scorch (ALS) (Davis et al. 1980), oleander leaf scorch (OLS) (Grebus et al. 1996), and bacterial leaf scorch (BLS) of elm, sycamore, and oak (Hearon et al. 1980). Although the genetic composition of this bacterium has been well described in a number of specific hosts (e.g. *Vitis* genera) (Chen et al. 2010; Simpson et al.; VanSluys et al. 2003; Bhattacharyya et al. 2002; Doddapaneni et al. 2006; Barbosa et al. 2015), less is known about the genetic and genomic character of colonies inhabiting less economically impactful plant species. In the case of Northeastern and Mid-Atlantic *Quercus* genera, etiological estimates have shown rampant localized *X. fastidiosa* based infection (Gould et al. 2007), and the need to examine underlying colony genotypes in these infected stands warrants exploration. From the standpoint of disease radiation such concentrated inoculum sources pose the potential for detrimental host to host, and alternative host disease spread.

Assignment of multiple *X. fastidiosa* subspecies taxa using 16S-23S DNA sequencing and phylogenetic analysis (Schaad et al. 2004) confirmed the existence of meaningful genetic diversity among distinct host derived subspecies populations. Succinctly, this study established the general assignment of grape associated isolates to the *fastidiosa* subspecies, citrus associated isolates to the *pauca* subspecies, and shade tree and *Prunus* associated isolates to the *multiplex* subspecies, including BLS infected oak isolates. Since this initial description (Schaad et al. 2004), a number of expanded, multi-isolate, multilocus sequence analyses (MLSA) (Lin et al. 2005; Schuenzel et al. 2005; Yuan et al. 2010; Parker et al. 2012) have been conducted, verifying the positioning of *X. fastidiosa* subspecies in general, and confirming phylogenetic placement of oak isolates within the *multiplex* subspecies clade. This finding was of particular importance because the *multiplex* subspecies designation implies a penchant for host plasticity relative to the stronger host-specific associations observed in other subspecies categories (Hopkins et al. 2002; Schaad et al. 2004). Expanded oak isolate complementation of previous studies is therefore needed to better understand the extent to which oak stands may function as both inoculum reservoirs and nexus

points for possible *X. fastidiosa* disease expansion into yet unidentified hosts and locales.

Several previous studies investigating *X. fastidiosa* subspecies phylogenies (Schuenzel et al. 2005; Parker et al. 2012) included analysis of only a few oak strains specific to locations in the Southern United States. More recent analyses, however, have considered larger scale population studies involving host/subspecies specialization (Nunney et al., 2013). While this latter study greatly expanded knowledge of oak associated genotypes, the results nonetheless omitted the characterization of diversity in Northeastern and Mid-Atlantic United States locales. Additional studies have also expanded geographically limited knowledge of BLS infected oaks, including a thorough survey of infected Kentucky oaks (Mundell 2005), two recent publications of *X. fastidiosa* distribution within the District of Columbia (Washington, D.C.) (Harris et al. 2014, Harris et al. 2015), and the previously mentioned host/species radiation study (Nunney et al., 2013) which presented a wide array of multiplex associated loci, with specific oak sequence haplotypes from Kentucky, the District of Columbia, Georgia, Tennessee, and Florida. Finally, the recent release of the Georgia based Red Oak specific *X. fastidiosa* genome (Griffin-1 from *Quercus rubra*) (Chen et al. 2013) provided the first global genetic perspective of an oak associated strain, but nonetheless suffered from the aforementioned issue of geographical segregation. In support of expanded sequence investigation, recent expansion of known *X. fastidiosa* diversity led to the identification of a novel genotype inhabiting chitalpa trees in the Southwestern United States (Randall et al. 2009), the unexpected genetic composition of the almond strain M23 relative to almond strain M12 despite common host isolation (Chen et al. 2010), and the emergence of a novel mulberry (*morus*) subspecies sister to the well-established *fastidiosa* (grape) and *sandyi* (oleander) clades (Hernandez-Martinez et al. 2006; Nunney et al. 2014).

The objective of this study was to gain a broader perspective of both the phylogenetic relationships and polymorphic signatures of *X. fastidiosa* oak strains residing in the Northeastern and Mid-Atlantic United States. In an effort to explore loci heterogeneity of *X. fastidiosa* oak populations beyond existing regional constraints, the prior results from two MLSA studies (Schuenzel et al. 2005; Parker et al. 2012) were paired with novel oak haplotypes from geographically distinct Northeastern and Mid-Atlantic oak strains. Again, this was done to discern potential genetic and phylogenetic differences in these shade tree

associated populations. Both conserved markers (Schuenzel et al. 2005) and previously defined variable markers (Parker et al. 2012) were used to establish a greater understanding of the bacterial populations inhabiting oak hosts. The thrust of this study forwards the hypothesis that there exists previously undescribed substantive genetic diversity among Northeastern and Mid-Atlantic *X. fastidiosa* oak derived populations. Observed diversity uncovered in this study also suggests general host specific adaptations and may speak to geographical and thermal considerations within this specific pathosystem.

Identification of novel diversity in understudied *X. fastidiosa* populations serves to extend understanding of the pathogen both within oak hosts and the novel geographical space of Northeastern and Mid-Atlantic United States. Results of this study promise to further geographically limited knowledge of oak host populations and expose unknown relationships relative to established *X. fastidiosa* diversity studies.

Materials and Methods:

Environmental Sample / Isolate Collection:

Oak environmental samples used for this study were collected between 2008 and 2011 in various Northeastern and Mid-Atlantic United States locales. Additionally, two environmental samples originating from Missouri were chosen as geographical outliers to compare and contrast the effect of physical distance in underrepresented sampling regions relative to genetic heterogeneity. All environmental samples were collected during the months of August through September and displayed typical scorch disease symptomology. Upon collection, samples were bagged and labeled according to geographical location and *Quercus* species and stored at 4°C until further processing.

DNA Extraction and PCR amplification from environmental samples:

Symptomatic leaves were removed from branches and surface sterilized by 60s immersion in 70% ethanol followed by 60s immersion in a 1% sodium hypochlorite. Samples were then washed three to five times in sterile deionized water and allowed to dry in an aseptic fume hood for 30 minutes. One inch samples (approximately 1/2 inch of leaf petiole and 1/2 inch of leaf midrib) weighing 0.1-0.3 g were excised from leaves. Petiole and midrib tissues were further incised using a sterile razor blade and placed into 2.0 mL conical tubes (Fisher) with 5MM solid glass beads (Fisher) and 0.4 ml AE elution buffer from the QIAamp

DNA Stool Mini Kit (Qiagen). Samples were placed in a bead beater (BIOSPEC PRODUCTS) for 90s using the “Homogenize” setting. Homogenized samples were then used for DNA extraction according to the prescribed QIAamp DNA Stool Mini Kit protocol, and stored at -20°C until further processing.

Conventional polymerase chain reactions (PCR) were carried out using 0.03-0.075 µg of extracted DNA and the following cycling protocol: initial 95°C denaturation for 2 min., subsequent 95°C denaturation for 30 sec., 58°C annealing for 1 min., and 72°C extension for 0.75 -2 min., depending on locus size.

Primer sets (Integrated DNA Technologies) used for MLSA and MLSA-E analyses conformed to those previously described (Schuenzel et al. 2005; Parker et al. 2012) (Table 2.3), excluding the loci *gltT* and *rfbD* due to poor amplification in the processed samples. Poor amplification of the latter in some pathosystems has been previously mentioned in the literature (Alemeida et al. 2008). Loci producing novel insertion/deletion (indel) regions (e.g., *acvB*, *copB*, *xadA*, and *nuoL* loci) were amplified between geographical sampling subsets to both verify and check for consistency across samples, as well as capture the greatest amount of genetic diversity among the oak associated collection.

Thirty-seven of 262 samples, which consistently amplified across the chosen MLSA and MLSA-E loci, were used for further processing and analyses (Table 2.1). All amplified PCR products were purified using the QIAquick PCR purification protocol (Qiagen). Additionally, the QIAquick Gel Extraction Kit protocol (Qiagen) was used for samples requiring gel purification. Both strands of purified PCR products were sequenced (Genewiz South Plainfield, New Jersey) and analyzed using DNASTar software (Lasergene) (Burland 1999). All novel indel and polymorphic regions identified among samples were re-verified by repeat PCR amplification, followed by cloning into the pGEM-T Easy vector system (Promega), and sequencing using universal vector primers. In all cases, chromatograms of sequence profiles appeared to originate from isogenic templates. Because the template DNA was derived from environmental sampling and not from purified culture, the possibility of aberrant base calling resulting from mixed bacterial populations or from PCR error was also considered. To avoid these possibilities, five randomly selected samples from each study were independently re-processed beginning with PCR amplification from environmental samples, followed by re-sequencing amplified products at each respective locus as described for the study. In all cases, results of this random chromatogram verification procedure

confirmed all original sequencing results.

Sequence Alignment, Phylogenetic Analysis, and Tree Imaging:

Consensus sequences for all loci were aligned in MAFFT version 7 (Katoh et al. 2013) with the following specifications: Strategy: FFT-NS-i (Slow; iterative refinement method), and Parameters: 1PAM / k=2, gap opening penalty = 1.53, Offset value = 0.0. Nexus format alignments were exported to BEAUti (Bayesian Evolutionary Analysis Utility Version v1.7.5) (Drummond et al. 2012), and the parameters were set for the resulting XML output as: Substitution Model: GTR (Tavaré 1986) (, Base Frequencies: Empirical, Site Heterogeneity Model: Gamma + Invariant, Model: Lognormal relaxed clock (uncorrelated), Tree Prior: Speciation: Yule Process (Gernhard et al. 2008). The Markov chain Monte Carlo (MCMC) was set in BEAUti according to the parameters for each individual locus: Length of chain: 750,000,000 generations, Echo state to screen every: 1000, Log parameters every: 1000. The Markov chain Monte Carlo (MCMC) was amended for the concatenated loci according to: Length of chain: 1,000,000,000 generations, Echo state to screen every: 1000, Log parameters every: 1000. Resulting XML files were analyzed by BEAST (Bayesian Evolutionary Analysis Sampling Trees Version v1.7.5) (Drummond et al. 2012), and post chain termination, Markov chain convergence was assessed using Tracer (MCMC Trace Analysis Tool Version v1.5.0) (Rambaut et al. 2014). The main criterion for assessing convergence post chain termination was an Effective Sample Size (ESS) statistical posterior greater than 200 for each individual run. Convergent files were then analyzed using TreeAnnotator (Bayesian Evolutionary Analysis Utility Version v1.7.5) (Drummond et al. 2012) with a consensus burnin of 15%, and acquisition of a single maximum credibility tree with corresponding posterior values for contained nodes. Finally, consensus trees were imaged using FigTree (Tree Figure Drawing Tool Version v1.4.2).

A Maximum Likelihood (ML) approach was also run using RAxML 7.3.0 (Stamatakis 2014) to complement the findings of the above Bayes analysis, as the two can sometimes conflict (Beerli 2006). A partitioned dataset was run under the model parameter “GTRGAMMAI”, and rapid bootstrap analysis (Stamatakis et al. 2008) was performed at both 100 and 1000 bootstrap iterations. Consensus trees were also imaged using FigTree.

SNP Analysis, Principal Component Axis (PCA) Analysis, Minimum Spanning tree analysis, and

recombination detection:

For SNP and PCA analyses, consensus sequences were aligned in MAFFT version 7 as described above and exported in aligned FASTA format. SNP location charts, complete linkage hierarchical clustering dendrograms, and Principal Component Analysis graphs were created in the R statistical computing /graphics environment version 3.0.2 (2013-09-25) -- "Frisbee Sailing" with the aid of the package "adeigenet" (Jombart et al. 2011).

For minimum spanning tree analysis, Arlequin version 3.5.1.3 (Excoffier et al. 2010) was used to compute the results of individual AMOVAs (Excoffier et al. 1992). AMOVAs were run with 99,999 permutations, and pairwise F_{st} values were computed with 99,999 permutations at a significance level of .05. Derived node and edge data generated from Arlequin version 3.5.1.3 was then imported into Gephi version 0.8.2-beta (Bastian et al. 2009) to build the Minimum Spanning Network (MSN). Visualization of the MSNs were created with both the Force Atlas and Force Atlas 2 algorithm (Jacomy 2009) bearing the following specifications: Edge weight influence 1.0, scaling 10.0, gravity 1.0 Tolerance 0.1 and approximation 1.2. Recombination breakpoint detection was conducted for each individual locus in both the conserved (MLSA) and variable (MLSA-E) gene sets. Detections were carried out using the DSS (Difference of Sums of Squares) method (McGuire et al. 2000) in the TOPALi v2.5 (build 13.04.03) (Milne et al. 2009) analysis suite. The parametric bootstrapping threshold was maintained at 100 runs, under the Jukes Cantor (Jukes and Cantor 1969) nucleotide substitution model. The simplified Jukes Cantor (Jukes and Cantor 1969) model was deemed appropriate based on the empirical evidence that *X. fastidiosa* sequencing project metrics report near 50/50 purine/pyrimidine composition. Recombinant regions were identified by the 95% significance point for the rejection of the null hypothesis (H_0) that the considered regions are not recombinant.

Results:

Sequence Analysis: Alignment, Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletion (indel) Profiles for Conserved (MLSA) and Variable (MLSA-E) Loci Sets.

The degree of genetic similarity between Northeastern and Mid-Atlantic oak loci was first considered at

the SNP level, as summarized for both oak specific and full subspecies concatemers in Figure 1. Following *X. fastidiosa* sequence alignments of MLSA-based loci, SNP analysis of previously described marker loci (Schuenzel et al. 2005) comparing Southern and Northeastern based oak samples revealed no aligned variation to *cysG* (501 sites), *leuA* (577 sites), *nuoL* (500-530 sites), and *petC* (595 sites). The remaining four markers revealed SNP variants in the following forms: *holC* contained a synonymous alanine substitution at residue 43 in New Jersey samples NB26, 1C1, NB21 (GCC to GCT), *lacF* contained a synonymous leucine substitution at residue 6 in New Jersey samples NB27, NB28 (CTG to CTA), *nuoL* contained a non-synonymous substitution from serine to phenylalanine and phenylalanine to leucine at both residue 73 and 78 in New Jersey samples 1C1 and EW10 (TCT to TTT, TTC to TTA), and *nuoN* contained a synonymous phenylalanine substitution at residue 241 in New Jersey samples NB22, and NB23_2. Additionally, the previously reported non-synonymous substitution of valine to alanine at the *pilU* locus (residue 142) in the Southern based sample Oak23 (Schuenzel et al. 2005), was not recovered in sequenced *pilU* fragments within any New Jersey samples. Likewise, no sequence variation was observed among amplicons of Northeastern and Mid-Atlantic based *X. fastidiosa* sample amplicons in seven of nine MLSA-E markers previously described in Southern isolates (Parker et al. 2012). These seven loci were comprised of *acvB* (638-708 sites), *copB* (548-594 sites), *cvaC* (285 sites), *fimA* (506 sites), *pglA* (497 sites), *pilA* (353 sites), and *rpfF* (777 sites). The remaining two genetic loci revealed SNP variants in the following forms: *gaa* contained a non-synonymous substitution from threonine to serine at residue 193 in 4T1_DE, and 5T1_DE (ACT to AGT), *xadA* contained a non-synonymous substitution from asparagine to serine at residue 18 in all Northeastern and Mid-Atlantic isolates except 11T1_DC (AAC to AGC), a non-synonymous substitution from isoleucine to valine at residue 48 in all Northeastern and Mid-Atlantic isolates except 11T1_DC (AUU to GUU), a non-synonymous substitution from glycine to aspartic acid at residue 68 in all Northeastern and Mid-Atlantic isolates except 11T1_DC (GGU to GAU), and two additional non-synonymous substitutions from arginine to glycine and lysine to glutamine at residues 115 and 116 respectively (CGU to GGU, AAG to CAG). These also occurred in all Northeastern and Mid-Atlantic isolates except 11T1_DC. Additionally, the following synonymous SNPs were identified at the *xadA* locus: residues 30, 111, 117, and 120 remained valine, alanine, aspartic acid, and alanine respectively (GUU to

GUG, GCA to GCC, GAC to GAU, GCC to GCG) and appeared in all Northeastern and Mid-Atlantic isolates except 11T1_DC, while residue 79 remained lysine (AAA to AAG) and appeared in all Northeastern and Mid-Atlantic isolates including 11T1_DC. Note, that indel profiles will be considered separate from alignment analysis.

Because BEAST employs a methodology that treats indels and their gap-based alignments as non-contributive or of equal marginal probability for all four nucleotides (Rambaut groups.google.com 2011), phylogenetic trees produced using this algorithm do not capture the effect of indel regions within alignments. For this reason, markers containing indels were secondarily examined to recover the maximum amount of genetic diversity within bacterial populations (Figure 3). Sequencing and alignment of the described partial gene loci *cysG*, *holC*, *lacF*, *leuA*, *nuoN*, *petC*, *pilU* revealed no novel indel patterns for Northeastern and Mid-Atlantic samples relative to the complementary loci used in this study (Schuenzel et al. 2005; Parker et al. 2012). The 530 base pair partial gene sequence for the *nuoL* locus, however, contained an in-frame 30bp deletion for the environmental sample Oak_RO_NB26. Interestingly, this same deletion was also previously observed in several Costa Rican isolate sequences (FJ610211, HM596025, HM243613) from coffee strains of the *X. fastidiosa* subspecies *pauca* (Nunney et al. 2011).

Considering the MLSA-E loci, indel regions were identified in the markers *acvB*, *copB*, and *xadA*. A summary of recovered indels in the profiled loci is provided in Figure 3. For the *acvB* locus, samples 12T2NJ and NB4_NJ each contained indel regions. The former revealed a 13 base pair insertion (TGGTGCCGACGTC) at site 417-429 causing both a gapped alignment relative to the other isolates and a premature stop in the coding sequence. NB4_NJ also contained a deletion segment in sites 361 to 416 again resulting in a premature stop in the coding sequence. Taken together, both isolates showed a truncated protein product for the *acvB* gene. This indel profiling procedure was complemented with the MLSA based sample set. Sample 1C1 revealed an insertion identical to 12T2NJ, and sample NB26 showed a large 172 base pair deletion likely starting at position 226 of the *acvB* gene fragment, but repeat regions contributed to ambiguities in the sequence alignment. In short, the predicted protein products for all four samples were truncated relative to the non-variant haplotypes considered in this study. The non-variant

type presents a 231 amino acid sequence for the partial gene, while the atypical sites resulted in 12T2_NJ: stop - 163 aa in protein, NB4_NJ: stop - 140 aa in protein, 1C1: stop - 163 aa in protein, NB26: stop - 101 aa in protein. All truncated products terminated in the same DAAQQR moiety further underscoring the novel nature of this marker.

For the *copB* locus, indel profiling revealed novel diversity among oak based *multiplex* subspecies considered in this study. Starting at study site 78, two of the previously described southern oak isolates (Oak_95_1 and Oak_92_10) (Parker et al. 2012) shared the distinctive pentapeptide repeat region of "MDHTQ/MDHTQ/MDHTG/MDHAI" with sample isolates 12T1_NH, 13_T1_MD, 14_T2_NJ, and 2T2_VT. Additionally, a longer *X. fastidiosa* subspecies *fastidiosa* associated repeat region represented by the sequence "MDHTQ/MDHTQ/MDHTQ/MDHTQ/MDHTG/MDHAI", was observed in 12_T2_NJ. Exploring the MLSA based grouping for the observed indel profiles, it was found that 1C1 shares the same elongated oak repeat region with 12_T2_NJ, and NB26 contained the repeat region "MDHTQ/MDHTG/MDHTG/MDHTG/MDHAI". This repeat pattern accounted for multiple oak haplotype designations across sampling sets, with the general consensus represented by the moiety MDHTQ(n)/MDHTG/MDHAI.

For the final locus, *xadA*, initial screenings of Northeastern and Mid-Atlantic isolates revealed a rich SNP profile, but no indel regions were observed. This contrasted with prior findings of (Parker et al. 2012) where a 21 base pair in-frame deletion was recovered from several California based *X. fastidiosa* subspecies *fastidiosa* isolates. However, profiling the New Jersey based isolates at the *xadA* locus revealed the same haplotype for this gene region as found in the Southern oak strains NB6 and NB28. Further analysis also uncovered an additional haplotype in the NB26 sample consisting of the following polymorphisms: non-synonymous substitutions were observed at residues 38, 41, 45, and 105, which resulted in shifts from alanine to serine (GCU to UCU), isoleucine to methionine (AUA to AUG), leucine to proline (CUU to CCU), and isoleucine to valine (AUC to GUC). Finally, sample NB26 retained its valine residue at position 48 which was previously noted as characteristic of the predominant oak host derived Northeast/Mid-Atlantic *xadA* locus haplotype.

Phylogenetic Analysis: gene trees and concatemeric trees for conserved and variable marker sets

Considering the topology of the oak groups, the MLSA gene trees (Supplemental 2.1) were largely consistent with the concatemeric tree (Supplemental 2.2) and placed the Southern based oak isolates (Table 2.2) and the Northeastern based isolates (Table 2.1) either within the same clade or phylogenetically sister to a mixture of geographically segregated oak samples as previously dictated by SNP profiling (Schuenzel et al. 2005). In particular, the greatest phylogenetic continuity relative to the concatemeric tree was observed in the largest gene fragments. With the exception of sequence specific differential substitution rates, *lacF* (523 bp), *nuoL* (530 bp), and *nuoN* (751 bp) recreated the topology of the concatemeric tree with the greatest fidelity. The smaller fragments *cysG* (501 bp), *petC* (495 bp), and *pilU* (472 bp) showed moderate divergence from the concatemeric tree regarding several multiplex isolates from the almond host, but preserved the oak based grouping in each instance. The smallest fragment, *holC* (318 bp) produced a polytomy, but nonetheless preserved the same master oak grouping. The MLSA concatemeric tree (Supplemental 2.2) was consistent with the previously recovered topology (Schuenzel et al. 2005). There was little genetic separation between the New Jersey based MLSA multiplex based loci, and the *multiplex* main clade proved consistent with prior taxonomic assignments of *X. fastidiosa multiplex* based subspecies. The phylogenetic profile of conserved loci in New Jersey based oak samples relative to previously reported Southern based conserved loci (Schuenzel et al. 2005) was largely indistinguishable save the minor genetic variations previously mentioned in the *holC*, *lacF*, *nuoN*, and *nuoL* loci.

Considering only the topology of the oak groups once more, the MLSA-E based gene trees (Supplemental 2.1) revealed indistinguishable previously consistent oak clading (Parker et al. 2012) for the markers *acvB*, *copB*, *fimA*, *pglA*, *pilA*, and *rpfF*. The marker *cvaC* revealed an anomalous topology likely due to inconsistencies in phylogenetic recreations when a singleton locus of diminutive size is selected for phylogenetic analysis. Because the *gaa* locus contained the aforementioned non-synonymous substitutions within New Jersey isolates, its gene tree provided some additional resolution relative to the prior observed pattern (Parker et al. 2012) of the *multiplex* main clade in gene reconstructions. The final marker, *xadA*, provided the best resolution and best approximated the concatemeric Bayesian tree with a

high degree of fidelity. Southern and Northeastern / Mid-Atlantic isolates formed distinct, geographical clading patterns not seen in the other eight topologies. This result was not unexpected, as the largest amount of SNP diversity between the geographically distinct oak isolates was be traced to this locus (Figure 2.1).

Visualization of the MLSA-E concatemeric tree (Figure 2.2) revealed consistency between previously described and established clades in subspecies *fastidiosa* and *sandyi* (Parker et al. 2012). Within the *multiplex* main clade, however, the most striking and novel observation was the presence of two unique oak divisions along geographical lines. The haplotypes described by Oak1 and Oak2 (Parker et al. 2012) formed a distinct group with this study's line of Southern demarcation (the Washington D.C. isolate Pin_Oak_11T1_DC). This genetic signature for the *xadA* locus of this sample was more consistent with that previously recovered in other Southern haplotypes (Parker et al. 2012). Seven of the eight concatemerized haplotypes in this study formed a unique grouping phylogenetically sister to the Sycamore loci with further resolution provided by the previously mentioned non-synonymous *gaa* substitutions in isolates RedOak_4T1_DE, and RedOak_5T1_DE (Figure 2.2). These unique topological findings were further supported by calculation of node posterior values for the novel recoveries. In short, the main bifurcation between Southern and Northeastern / Mid-Atlantic oaks supported a posterior probability of 0.94, with the bifurcation between the sycamore loci and the remaining seven Northeastern / Mid-Atlantic oaks revealing a posterior probability of 1.00. Both of these posterior values can be interpreted as proving a high level of support for these recovered Bayesian topologies.

To ensure consistency of the phylogenetic findings, Maximum Likelihood analysis (ML) was provided for comparison to the Bayesian method employed by BEAST. For the RAXML based analysis (Supplemental 2.3), gene trees were excluded and only the concatemerized MLSA and MLSA-E sets were considered. The partitioned MLSA locus set recovered a near identical topology to the MLSA Bayes visualization, as did the partitioned MLSA-E concatemerization. One caveat surrounding ML recovery was the fact that the rapid bootstrapping algorithm produced adequate but low support numbers for many of the nodes (< 50). Considering only the observed oak based bifurcation, however, high support was observed in the main split between the "Southern" and the "Sycamore" and "Northeastern / Mid-Atlantic" groupings at a

bootstrap value of 90, and the node bearing the six of the seven novel haplotypes in this study, was supported at a bootstrap value of 99 (Supplemental 2.3).

Hierarchical Clustering and Principal Component Analysis (PCA):

To further assess the extent to which the indel profiling could place *X. fastidiosa* oak population genotypes relative to one another, underlying genetic sequences were subjected to Principal Component Analysis (PCA) and Hierarchical Clustering (Figure 2.4a-2.4f). Principal Component variation was determined via the binary transform of the underlying polymorphisms into simple dissimilarity matrices derived from multi-sequence alignments. Resulting eigenvalues (the explanatory source of percentages of captured genetic variation within the individual samples) (Jolliffe 2002) were then analyzed, and their cumulative summation determined the number of principal components needed to explain the variation seen in the respective concatemeric loci sets. Per the clustering/PCA methodology, only the highest eigenvalues were retained to explain genetic variation groupings. Specifically, the conserved loci (MLSA) displayed ~98% variation within the first four eigenvalues and ~88% variation within the first two eigenvalues (Figure 2.4a), sufficiently describing the majority of the genetic variation in the sample set. In contrast, the values of the first four eigenvalues within the variable marker set (MLSA-E) achieved a similar total percent variation score of ~98%, but the first two eigenvalues accounted for only ~67% of the underlying sample variation (Figure 2.4c). Here, a 2/3 cut-off was used for PCA visualization, and MLSA-E visualization was limited to two axes.

Complete linkage hierarchical clustering was carried out among the combined oak associated population sets assuming four principal components for each loci set (Figure 2.4b, 2.4e). This clustering was paired with the PCA visualization as a verification of the clustered findings. The conserved MLSA loci produced a deletion driven haplotypic outlier (NB26), and three very tightly clustered groupings (Fig. 2.4b). This layout is indicative of the observed marginal variation within the population that is based on the few novel SNPs uncovered in several of the New Jersey based isolates. Further, the recovered clustering pattern was consistent with the synonymous and non-synonymous substitutions observed in *hoIC* (NB26, 1C1, NB21), *lacF* (NB27, NB28), *nuoL* (1C1 and EW10), and *nuoN* (NB22, and NB23_2) (Fig. 1). A

subsequent analysis of complete linkage hierarchical clustering for the MLSA-E locus set produced a more diffuse clustering pattern indicative of the increased genetic variation (Fig. 2.4e). In this data set two deletion outliers and two salient clusters were formed, including two distinct sub-clusters containing Northeastern and Mid-Atlantic isolates only. The deletion outliers consisted of RedOak_12T2_NJ and PinOak_NB4_NJ, both of which have been previously profiled in Figure 3. Additionally, this method captured the observed genotypic dissimilarity of PinOak_11T1DC by grouping it with the Southern isolate Oak3. Both branching patterns were consistent with their concatenated topologies save for the isolates containing extended indel patterns. Those with the largest indel regions appeared as outliers:

Oak_RO_NB26 for the MLSA locus set and PinOak_NB4 and RedOak_12T2NJ for the MLSA-E locus set. To consider the observed variable MLSA-E grouping further, the effect of the pentapeptide repeat region in *copB* was observed when analyzing the grouping of BlackOak_12T1_NH, PinOak_13T1_MD, PinOak_14T2_NJ, and RedOak_2T2_VT relative to the residual Northeastern/Mid-Atlantic isolates (Figure 2.4b, 2.4e). Indel inclusive results of the hierarchical clustering and PCA methodologies thus serve to provide a more global genetic perspective, complementing the multilocus phylogenies.

An explicit description of the PCA for the conserved loci the dimensionality display was reduced to two axes due to a cumulative eigenvalue percentage of ~88% (Figure 2.4a). This two-axis dimensionality closely paralleled the hierarchical clustering by positioning the haplotypic outlier Oak_NB26 distal from the main clustering (Figure 2.4a, 2.4b). The grouping of Oak_1C1 and Oak_EW10 was accounted for in the *nuoL* based SNP profile. Finally, the genetic invariability of the largest clustered grouping (n=19) confirmed by the near superimposition of the groups in axes one and two.

PCA retained the observed divisions seen in the phylogenetic topology, but provided an additional level of analysis relative to the indel profiles shared within the oak based groupings. The PCA for the variable MLSA-E locus group was considered in the following dimensionality: axes one and two determined ~41% and ~26% of the observed variability, and axes three and four determined ~19% and ~12% of the observed variability (Figure 2.4d). Similar to observations with the MLSA conserved locus group, the first four eigenvalues for the variable loci captured ~98% of the total variability. However, unlike the conserved locus group, more variability was distributed over the first 4 eigenvalues, with only 67%

variability was captured in the first 2 eigenvalues (Figure 2.4a, 2.4d). The plot of axes one and two positioned PinOak_NB4 (group 6) and RedOak_12T2NJ (group 5) as the haplotypic outliers with the latter closer in conserved composition to group one and group two (Figure 2.4f). PCA axis one and PCA axis two placed PinOak_NB4 as such due to the observed *acvB* indel of 55bp and the lack of a pentapeptide repeat region in the *copB* locus (Figure 2.4f). In similar fashion, RedOak_12T2NJ contains both indel regions and is moved away from the main grouping in the two axis display (Figure 2.4f). Switching dimensionality to axis three and four (data not shown), group one confirms the genetic division between the Northeastern/Mid-Atlantic isolates and the Oak3 (Southern) / PinOak_11T1DC. This can be seen in the clustering pattern of the points in the radiating extrema. Additionally, group two shows a similar pattern of extrema confirmed by the known variability observed in that Northeastern/Mid-Atlantic isolate group relative to the previously defined composition of Southern oak isolates Oak1 and Oak2 (Parker et al. 2012). The remaining geographical separation (save PinOak_11T1DC) is due to the previously discussed variability at the *xadA* locus, and the fine sub-grouping is due to the distribution of pentapeptide repeats at the *copB* locus.

Minimum Spanning Tree (MSN):

Because the hierarchical clustering and PCA focused strictly on the oak specific populations, a minimum spanning network (MSN) was chosen to further complement the complete phylogenetic results and place *X. fastidiosa* oak strain diversity within the context of all other analyzed isolates. *Interpretation* of the MSN is based on line distance and line weight (Figure 2.5a, 2.5b), where thicker two-dimensional line width signifies relatedness and line distance (length) signifies genetic difference. Viewing the MLSA spanning tree (Figure 2.5a), sample “s_x”, and “f_x”, representing subspecies *sandyi* (oleander) and *fastidiosa* (grape) respectively, appear in a proximal relationship (Fig. 2.5a) similar to that observed in the recovered basal clade. The subspecies multiplex cluster shows substantial relatedness between nine of the haplotypes based on line weight. The degree of genetic invariance among these haplotypes was visually clear and was further accentuated by the inclusion of the two previously described Southern isolates (Schuenzel et al. 2005) within the m_15 grouping, which centered the cluster. Haplotype representative m_15 contained the largest oak haplotype grouping, composed of eleven New Jersey

based isolates, and the two Southern isolates. The MSN line attenuation in m_16 and m_20 was accounted for by the *nuoL* SNPs (Oak_PO_EW10, Oak_RO_1C1), and the positioning and line weight of m_21 was a reflection of the captured indel in Oak_RO_NB26. The line modularity observed in p_10 conforms to the phylogenetic observations that the subspecies *pauca* (citrus) was the most distant isolate, consistent with this subspecies serving as the outgroup in the Bayesian analysis (Fig. 2.2). The MSN created by the variable loci set (Fig. 5b) supported findings of phylogenetic analysis (Fig. 2.2), and also captured the effects of recovered indel regions present in the *X. fastidiosa* oak isolates represented by haplotypes m_3, m_9-15, m-16, and m_18. The pronounced width of line weighting between haplotypes m_11 and m_12; and also between haplotypes m_13, m_14, and m_16 confirmed strong genetic similarity and correspondence to the segregation of several of the Northeastern/Mid-Atlantic and Southern haplotypes (m_9 and m_11). Because the MSN accounted for indel presence, differences in the previously discussed pentapeptide repeat region within the *copB* gene (Fig. 2.3) was reflected in similarity of haplotypes group m_9 and group m_10 to each other, and their separation from the oak haplotypes. The more distally located outlier MSN groupings of oak haplotypes m_3 and m_18 could be explained by the *acvB* indel in the former and the *acvB* deletion and extended *copB* pentapeptide repeat region and in the latter (Figure 2.3).

Recombination Break point Detection:

Although this study was concerned with the description of genetic diversity in geographically novel oak associated *X. fastidiosa* populations, a recombination detection analysis was run. The results of a recombination event can cause differential evolutionary processes to appear on the same locus and may partially invalidate certain types of analyses (Schierup and Hein 2000). Potential recombination breakpoints were sought within loci using the DSS method (McGuire et al. 2000) (Supplemental 2.4). While breakpoints were detected by the DSS method at every conserved and variable locus, only breakpoints within the variable loci genes *copB* and *gaa* loci were found to be statistically significant ($p < 0.05$). Residual relaxation of the p value above $p = 0.05$ resulted in recombination detection at the 84%-70% significance point for *acvB*, *cvaC*, *fimA*, *pglA*, and *xadA*. The remaining markers *pilA* and *rpff*, as well as the conserved locus set showed evidence of recombination only below this relaxed threshold and were

thus considered non-significant beyond the logic of this modified recombination detection analysis (McGuire et al. 2000; Milne et al. 2009). Because an additional haplotype was uncovered at the *xadA* locus between sampling sets, a second DSS locus specific recombination analysis was performed using the three haplotypes uncovered in this study and the previously reported Southern haplotype (Parker et al. 2012). The aforementioned relaxed p-value method detected a breakpoint at a 70% significance point among these *xadA* haplotypes.

Discussion:

This study uncovered and described meaningful genetic differences in oak associated *X. fastidiosa* strains obtained from previously uncharacterized geographic locales that includes the northeastern and mid-Atlantic United States. Further, this study supplemented existing genotypic descriptions of oak derived *X. fastidiosa* haplotypes by expanding several phylogenetic interpretations previously described (Schuenzel et al. 2005; Parker et al., 2012). A small number of novel oak haplotypes were described using the conserved MLSA loci, however, these genetic variations lacked the SNP richness to provide substantive differentiation between sample isolates within the larger oak genotype pool. Examining the loci profiled in the MLSA set, it would be reasonable to expect that those genes described as “housekeeping” would be neutrally variant. Without considering the large amount of recent work to explain *X. fastidiosa* introgressions (Nunney et al. 2012, 2014), this was generally found to be the case within the considered oak populations. Regardless of the absence of these point mutations, a high degree of similarity was preserved among the remaining bases suggesting conservation among oak isolates. It remains the case, however, that a *pilU* variant appeared in the Southern oak population relative to the New Jersey supplementation (Schuenzel et al 2005), and more work should be done to consider the usefulness of this marker for geographical and host specific origins.

The recovered topology generated from the variable MLSA-E locus set (Fig. 2.2) revealed a surprising amount of genetic similarity between Northern and Southern oak populations, with the exception of the *xadA* locus. Putting this locus aside, this observation would seem to strengthen the generally accepted genetic relationships of *X. fastidiosa* host/pathogen associations, despite the ultimate branching patterns

(Figure 2.2) recovered by the phylogram. Given the unexpected lack of genetic diversity in the MLSA-E analysis among oak associated populations, it would be tempting to dismiss the clading as anomalous. This would be erroneous, however, because the underlying importance of these novel haplotypes in the Northeastern and Mid-Atlantic populations may lie in the functionality of the gene itself. The *xadA* gene is known to encode for an afimbrial adhesion protein associated with *X. fastidiosa* virulence (Feil et al. 2007). Disruption of the *xadA* locus showed attenuated glass surface attachment, a reduction in the ability to produce mature biofilms, and reduced disease expression in grape hosts (Feil et al. 2007). Additionally, this attachment gene was shown to be significantly upregulated in the xylem fluid of grape hosts, leading to speculation that increased transcript abundance could be linked to both enhanced vessel attachment and biofilm formation (Shi et al. 2010). While the role of *xadA* in virulence has not been investigated in other *X. fastidiosa* host systems, it remains a strong candidate among strains for genetic diversity resulting from host specific selectivity. Analysis of the *xadA* locus influenced findings of oak populations findings with respect to geographical radiation Literature describing *X. fastidiosa* intrasubspecific homologous recombination (IHR) (Nunney et al. 2013, 2014), and the presence of clonal complexes (CCs) (Sally et al. 2005) have been critical of phylogenetic findings which may superficially appear to be cladogenesis events. The populations chosen for MLSA-E phylogenetic visualization resulted in geographical division (Figure 2.2), and reanalysis of the New Jersey-based collection for indel presence within the *xadA* locus recovered both the Southern haplotype and a novel intermediate haplotype (Table 1). This observation challenges the extent to which the Northeastern and Mid-Atlantic *xadA* haplotype genetically diverged from the southern haplotype, however, the two individual isolates (NB26 and NB28) housing this haplotype (Table 1) were recovered from young, transplanted trees within urban settings. This haplotype was not recovered in any other instances in the Northeastern sampling as periodic annual increments (PAIs) evaluations (Chapman 1921) (data not presented) predicted the majority of oak hosts to be greater than 60 years in age. These observations suggested that the newly discovered “Southern” haplotype identified among Northeastern *X. fastidiosa* populations may have been recent transplants. For completeness, the *xadA* locus from the aforementioned *X. fastidiosa* Red Oak associated genome (Griffin-1) was also analyzed. Although several novel SNPs relative to both Northeastern/Mid-Atlantic and

Southern haplotypes were observed, the prevailing nature of the locus appeared consistent with Southern haplotype recoveries. The extent to which these polymorphisms could represent either yet undescribed haplotype diversity or artifacts of the sequencing project would, therefore, require further analysis.

X. fastidiosa indel profiling of *acvB* revealed three observed variants within the New Jersey and Northeastern and Mid-Atlantic populations that result in a translational truncation. The *acvB* gene consists of a 900 bp ORF in *X. fastidiosa* subsp. *fastidiosa* strain Temecula1 (PD_1902) (VanSluys et al. 2003). Since this indel occurs near the C-terminus in the predicted gene product in both isolate RedOak_12T2NJ and PinOak_NB4NJ modulation must also be considered. Mutation of this gene is known to produce a hypervirulent phenotype in grape (Hernandez-Martinez et al. 2006). Similar to *xadA*, the *acvB* gene represents another vir-associated gene that potentially warrants further study for better understanding of oak-based *X. fastidiosa* subsp. *multiplex* populations, and thus the epidemiology of BLS in the Northeastern United States.

The previously reported *copB* indel region found among *X. fastidiosa* strains (Lin et al. 2005; Parker et al. 2012) were similarly observed within Northeastern/Mid-Atlantic isolates (Figure 2.3). Recovered genotypes revealed similar pentapeptide repeats as those observed in their Southern Oak haplotypes (Parker et al. 2012), but this study also revealed the presence of two novel variants in the New Jersey locale: the elongated repeat regions (MDHTQ/MDHTQ/MDHTQ/MDHTQ/MDHTG/MDHAI) identified in both RedOak_12T2NJ and Oak_RO_1C1 (Table 2.1, Figure 2.3), and (MDHTQ/MDHTG/MDHTG/MDHTG/MDHAI) in Oak_RO_NB26. In addition to potentially enhanced enzymatic activity, it has been speculated that these repeat regions may play a role in conformational structure, such as elongation of either alpha helices or beta sheets (Bateman et al. 1998; Gemayel et al. 2010). From a control standpoint, isolates containing such regions may have enhanced catalytic sites and thus greater sequestration properties.

The previously reported 21 base pair in frame deletion within the *xadA* locus of several grape strains (Parker et al. 2012) was not recovered in either the Northeastern/Mid-Atlantic oak grouping, or the expanded testing of the New Jersey based oak grouping looking for indels across samplings. The revealed

absence of this indel in the profiled oak populations lent credence to the observation that this could represent a host based modification (Parker et al. 2012). This is especially important because the aforementioned *xadA* locus is associated with prokaryotic adhesion, and locus disruption has been directly linked to a hypovirulent phenotype in grape (Feil et al. 2007).

The only MLSA detected indel was uncovered in the *nuoL* gene encoding the NADH-ubiquinone oxidoreductase, NQO12 subunit, a key component of aerobic respiration. As stated prior, this deletion was previously observed (Nunney et al. 2011) in several *X. fastidiosa* coffee isolates from the *pauca* subspecies. Since this study remains focused on the differences observed between regionally segregated *X. fastidiosa* oak isolates, the extent to which this deletion could serve as an indicative marker of origin or niche remains unclear.

Another facet of this study was to also visually capture the indel rich nature of *X. fastidiosa* in the profiled oak associated samples. Presented eigenvalues can be considered a good approximation of sample variability, where n-component displays (where n is the number of eigenvalues selected) can describe the complete dataset in both a macro and micro sense (Jolliffe 2002). For the conserved MLSA locus set, the magnitude of the first two eigenvalues account for nearly all the variation observed within the underlying dataset (Figure 2.4a). Considering the hierarchical clustering for the conserved loci, a four-component representation was chosen to highlight minor variation between several of the New Jersey based samples in relation to their Southern counterparts (Figure 2.4b). The corresponding PCA, however, adequately described the genetic composition using only two axes (Figure 2.4c). In contrast, the eigenvalues for the variable MLSA-E locus set showed that hierarchical clustering necessitated a four-component display (Figure 2.4d). The two axis PCA for the variable locus set could likely benefit from the inclusion of a third and fourth axis display, but the two axis display account for almost 70% of the underlying variation, and its visualization successfully captured an a priori distribution of the genetic composition of the oak derived populations (Figure 2.4f). Clustering and PCA complementation resulted in the segregation of subpopulations ...by how be specific? . This study uncovered several *acvB* haplotypes tied to truncated protein products. Prior research linking this locus to a hypervirulent phenotype suggests that knowing details of the distribution of this genotype could be important in both predicting disease outbreaks and

crafting control strategies. In the same vein, recognition of extended *copB* regions in certain niches could be used to assuage the spread of problematic subpopulations.

The minimum spanning trees for both loci sets were highly consistent with the findings of the concatemeric phylogenetic trees (Figure 2.2, Figure 2.5a, 2.5b). The conserved loci showed predictable groupings and line weights according to the low amount of genetic variation and indel poor recoveries observed in the profiled oak samples. Those samples showing variation were assigned a line weight suggestive of variation (m_16, m_20 for example) and moved away from the central grouping consisting of 13 oak haplotypes (m_15) (Figure 2.5a). A similar but more diffuse pattern was observed in the spanning tree for the variable genetic loci consistent with the increased overall genetic diversity in the examined samples. The groupings bear a similarity to that recovered in the variable phylogenetic topology, although some ambiguous relationships were observed due to weight of indels (Figure 2.5b). One caveat to this approach, however, is that while the output builds a single spanning tree, various paths to points, here represented as haplotypes, may exist (Sailpante et al. 2011). Nonetheless, taking the observed genetic variation in total, from a qualitative perspective, the spanning tree may present a more complete picture of genetic variance in this case. The variation uncovered in this study for oak isolates validates the supplemental use of a minimum spanning tree, and the relationships described herein conform well to the accompanying phylogenetic and PCA based data.

Without considering the ramifications of IHRs, the novel establishment of a clade sister to the Southern oak haplotypes is used in descriptive fashion, suggesting previously unknown genetic diversity in oak populations of *X. fastidiosa*. Although the topology results in the formation of a novel clade, no claim to cladogenesis is made (Figure 2.2). Given the extent of the work examining IHRs (Nunney et al.) it may be true that the resulting *xadA* haplotypes are the result of recombination events, but recombination breakpoints were only included in this work as a tangential hint at potential origin. It may also be the case that the DSS method is inadequate to describe all of the recombination breakpoints present in the considered loci, but this study stops short of addressing the significance of recombination breakpoints within the studied oak populations and instead seeks to further the current understanding of oak host

associated genetic composition existing within the parameters of this sampling schema.

This study posits the existence of several novel oak associated haplotypes and indel profiles within Northeastern and Mid-Atlantic geographical locales. In addition to Bayesian phylogenetic results, populations were subcategorized with complementary indel composition. Because the understudied oak associated *X. fastidiosa* populations in Northeastern and Mid-Atlantic stands are typically juxtaposed with *Prunus*, *Vaccinium*, and *Vitis* genera (Frecon et al. 2001; Telfer 2002; Strik 2004), the findings presented here should function as a prelude to continued *X. fastidiosa* population profiling. Despite the observance of close host/pathogen associations in *X. fastidiosa* phylogenies, and limited evidence regarding host shift, many inoculum reservoirs may sit at the nexus of future vector habitat expansion, and thus disease spread. The intuition that predicted Northeastern temperature elevation will likely be impactful to forest pathology also presents an additional fulcrum for the possibility of *X. fastidiosa* radiation. In short, the extent to which temperature zones remain effective allopatric boundaries may fade in the coming decades, making a more global understanding of this disease critical. The work presented here has, therefore, aided in an extended description of current oak associated *X. fastidiosa* genotypes, and furthered understanding of subspecies taxonomy, geographical origins, epidemiological ramifications, and potential starting points for strategic control.

REFERENCES

- Almeida, R. P., Nascimento, F. E., Chau, J., Prado, S. S., Tsai, C. W., Lopes, S. A., & Lopes, J. R. (2008). Genetic structure and biology of *Xylella fastidiosa* strains causing disease in citrus and coffee in Brazil. *Applied and Environmental Microbiology*, 74(12), 3690-3701.
- Barbosa, D., Alencar, V. C., Santos, D. S., de Freitas Oliveira, A. C., de Souza, A. A., Colleta-Filho, H. D., ... & Nunes, L. R. (2015). Comparative genomic analysis of coffee-infecting *Xylella fastidiosa* strains isolated from Brazil. *Microbiology*, mic-0.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8, 361-362.
- Bateman, A., Murzin, A. G., & Teichmann, S. A. (1998). Structure and distribution of pentapeptide repeats in bacteria. *Protein science*, 7(6), 1477-1480.
- Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22(3), 341-345.
- Bhattacharyya, A., Stilwagen, S., Ivanova, N., D'Souza, M., Bernal, A., Lykidis, A., ... & Kyrpides, N. C. (2002). Whole-genome comparative analysis of three phytopathogenic *Xylella fastidiosa* strains. *Proceedings of the National Academy of Sciences*, 99(19), 12403-12408.
- Burland, T. G. (1999). DNASTAR's Lasergene sequence analysis software. In *Bioinformatics Methods and Protocols* (pp. 71-91). Humana Press.
- Chapman, H. H. (1921). *Forest mensuration*. John Wiley & sons, Incorporated.
- 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(17), 4534-4534.
- 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 announcements*, 1(5), e00756-13.
- Davis, M. J., Thomson, S. V., & Purcell, A. H. (1980). Etiological role of a xylem-limited bacterium causing Pierce's disease in almond leaf scorch. *Phytopathology*, 70(472), 5.
- Doddapaneni, H., Yao, J., Lin, H., Walker, M. A., & Civerolo, E. L. (2006). Analysis of the genome-wide variations among multiple strains of the plant pathogenic bacterium *Xylella fastidiosa*. *BMC genomics*, 7(1), 225.
- Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, 29(8), 1969-1973.

Dukes, J. S., Pontius, J., Orwig, D., Garnas, J. R., Rodgers, V. L., Brazee, N., ... & Ayres, M. (2009). Responses of insect pests, pathogens, and invasive plant species to climate change in the forests of northeastern North America: What can we predict? This article is one of a selection of papers from NE Forests 2100: A Synthesis of Climate Change Impacts on Forests of the Northeastern US and Eastern Canada. *Canadian Journal of Forest Research*, 39(2), 231-248.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479-491.

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, 10(3), 564-567.

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(3), 318-324.

Frecon, J., Belding, R., & Lokaj, G. (2001, July). Evaluation of white-fleshed peach and nectarine varieties in New Jersey. In V International Peach Symposium 592 (pp. 467-477).

Gemayel, R., Vincens, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, 44, 445-477.

Gernhard, T., Hartmann, K., & Steel, M. (2008). Stochastic properties of generalised Yule models, with biodiversity applications. *Journal of mathematical biology*, 57(5), 713-735.

Gould, A. B., & Lashomb, J. H. (2007). Bacterial leaf scorch (BLS) of shade trees. The Plant Health Instructor.

Grebus, M. E., Henry, J. M., Hartin, J. E., & Wilen, C. A. (1996). Bacterial leaf scorch of oleander: a new disease in southern California. *Phytopathology*, 86, 110.

Harris, J. L., Di Bello, P. L., Lear, M., & Balci, Y. (2014). Bacterial Leaf Scorch in the District of Columbia: Distribution, Host Range, and Presence of *Xylella fastidiosa* Among Urban Trees. *Plant Disease*, 98(12), 1611-1618.

Harris, J. L., & Balci, Y. (2015). Population Structure of the Bacterial Pathogen *Xylella fastidiosa* among Street Trees in Washington DC. *PloS one*, 10(3).

Hearon, S. S., Sherald, J. L., & Kostka, S. J. (1980). Association of xylem-limited bacteria with elm, sycamore, and oak leaf scorch. *Canadian Journal of Botany*, 58(18), 1986-1993.

Hernandez-Martinez, R., Dumenyo, K. C., & Cooksey, D. A. (2006). Site-directed mutagenesis of *acvB* gene in a Pierce's disease strain of *Xylella fastidiosa*. *Phytopathology*, 96(6).

Hernandez-Martinez, R., Pinckard, T. R., Costa, H. S., Cooksey, D. A., & Wong, F. P. (2006). Discovery and characterization of *Xylella fastidiosa* strains in southern California causing mulberry leaf scorch. *Plant disease*, 90(9), 1143-1149.

Hopkins, D. L., & Purcell, A. H. (2002). *Xylella fastidiosa*: cause of Pierce's disease of grapevine and other emergent diseases. *Plant disease*, 86(10), 1056-1066.

Jacomy, M. (2009). Force-atlas graph layout algorithm. URL: <http://gephi.org/2011/forceatlas2-the-new->

version-of-our-home-brew-layout.

Jolliffe, I. (2002). *Principal component analysis*. John Wiley & Sons, Ltd.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21), 3070-3071.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 3, 21-132.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), 772-780.

Lee, R. F., Beretta, M. J. G., Hartung, J. H., Hooker, M. E., & Derrick, K. S. (1993). Citrus variegated chlorosis: confirmation of a *Xylella fastidiosa* as the causal agent. *Summa Phytopathologica*, 19(2), 123-125.

Lin, H., Civerolo, E. L., Hu, R., Barros, S., Francis, M., & Walker, M. A. (2005). Multilocus simple sequence repeat markers for differentiating strains and evaluating genetic diversity of *Xylella fastidiosa*. *Applied and environmental microbiology*, 71(8), 4888-4892.

McGuire, G., & Wright, F. (2000). TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, 16(2), 130-134.

Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F., & Wright, F. (2009). TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics*, 25(1), 126-127.

Mundell, J. N. (2005). Phylogenetic analysis of Kentucky strains of *Xylella fastidiosa*.

Nunney, L. (2011). Homologous recombination and the invasion of a new plant host by the pathogenic bacterium, *Xylella fastidiosa*. *Phytopathology*, 101, S130.

Nunney, L., Vickerman, D. B., Bromley, R. E., Russell, S. A., Hartman, J. R., Morano, L. D., & Stouthamer, R. (2013). Recent evolutionary radiation and host plant specialization in the *Xylella fastidiosa* subspecies native to the United States. *Applied and environmental microbiology*, 79(7), 2189-2200.

Nunney, L., Hopkins, D. L., Morano, L. D., Russell, S. E., & Stouthamer, R. (2014). Intersubspecific recombination in *Xylella fastidiosa* strains native to the United States: infection of novel hosts associated with an unsuccessful invasion. *Applied and environmental microbiology*, 80(3), 1159-1169.

Nunney, L., Ortiz, B., Russell, S. A., Sánchez, R. R., & Stouthamer, R. (2014). The Complex Biogeography of the Plant Pathogen *Xylella fastidiosa*: Genetic Evidence of Introductions and Subspecific Introgression in Central America. *PloS one*, 9(11), e112463.

Nunney, L., Schuenzel, E. L., Scally, M., Bromley, R. E., & Stouthamer, R. (2014). Large-scale intersubspecific recombination in the plant-pathogenic bacterium *Xylella fastidiosa* is associated with the host shift to mulberry. *Applied and environmental microbiology*, 80(10), 3025-3033.

Parker, J. K., Havird, J. C., & De La Fuente, L. (2012). Differentiation of *Xylella fastidiosa* strains via multilocus sequence analysis of environmentally mediated genes (MLSA-E). *Applied and environmental microbiology*, 78(5), 1385-1396.

Pierce, N. B. (1892). The California vine disease: a preliminary report of investigations (No. 2). US Government Printing Office.

Purcell, A. H., & Hopkins, D. L. (1996). Fastidious xylem-limited bacterial plant pathogens. Annual review of phytopathology, 34(1), 131-151.

Purcell, A. H., Saunders, S. R., Hendson, M., Grebus, M. E., & Henry, M. J. (1999). Causal role of *Xylella fastidiosa* in oleander leaf scorch disease. Phytopathology, 89(1), 53-58.

Rambaut A, Suchard MA, Xie D & Drummond AJ (2014) Tracer v1.6, Available from <http://beast.bio.ed.ac.uk/Tracer>

Randall, J. J., Goldberg, N. P., Kemp, J. D., Radionenko, M., French, J. M., Olsen, M. W., & Hanson, S. F. (2009). Genetic analysis of a novel *Xylella fastidiosa* subspecies found in the southwestern United States. Applied and environmental microbiology, 75(17), 5631-5638.

Salipante, S. J., & Hall, B. G. (2011). Inadequacies of minimum spanning trees in molecular epidemiology. Journal of clinical microbiology, 49(10), 3568-3575.

Scally, M., Schuenzel, E. L., Stouthamer, R., & Nunney, L. (2005). Multilocus sequence type system for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation to clonal diversity. Applied and environmental microbiology, 71(12), 8491-8499.

Schaad, N. W., Postnikova, E., Lacy, G., Fatmi, M. B., & Chang, C. J. (2004). *Xylella fastidiosa* subspecies: *X. fastidiosa* subsp. *piercei*, subsp. nov., *X. fastidiosa* subsp. *multiplex* subsp. nov., and *X. fastidiosa* subsp. *paucapauca* subsp. nov. Systematic and applied microbiology, 27(3), 290-300.

Schierup, M. H., & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2), 879-891.

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*. Applied and environmental microbiology, 71(7), 3832-3839.

Shi, X., Bi, J., Morse, J. G., Toscano, N. C., & Cooksey, D. A. (2010). Differential expression of genes of *Xylella fastidiosa* in xylem fluid of citrus and grapevine. FEMS microbiology letters, 304(1), 82-88.

Simpson, A. J. G., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., ... & Krieger, J. E. (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature, 406(6792), 151-157.

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. Systematic biology, 57(5), 758-771.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312-1313.

Strik, B. (2004, May). Blueberry production and research trends in North America. In VIII International Symposium on Vaccinium Culture 715 (pp. 173-184).

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences, 17, 57-86.

Team, R. C. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Telfer, D. J. (2002). e Northeast Wine Route: wine tourism in Ontario. Wine tourism around the world: Development, management and markets, 253.

Van Sluys, M. A., De Oliveira, M. C., Monteiro-Vitorello, C. B., Miyaki, C. Y., Furlan, L. R., Camargo, L. E. A., ... & Truffi, D. (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. *Journal of Bacteriology*, 185(3), 1018-1026.

Wells, J. M., Raju, B. C., Hung, H. Y., Weisburg, W. G., Mandelco-Paul, 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(2), 136-143.

Yuan, X., Morano, L., Bromley, R., Spring-Pearson, S., Stouthamer, R., & Nunney, L. (2010). Multilocus sequence typing of *Xylella fastidiosa* causing Pierce's disease and oleander leaf scorch in the United States. *Phytopathology*, 100(6), 601-611.

Table 2.1. Novel New Jersey and Northeastern/Mid-Atlantic oak associated genotypes considered in this study

Study novel oak isolates	host	sampling date	location	Loci set	amplicon indels	SNP locus locations	Haplotype Grouping	Haplotype representative in Tree
Oak_RO_NB27	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA	no	<i>lacF</i> , <i>xadA</i> ***	1_MLSA	Yes
Oak_RO_NB28	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA MLSA	no	<i>lacF</i> ***	1_MLSA	No
RedOak_12T2_NJ	<i>Q. rubra</i>	2011	Cranbury, NJ	-E	<i>copB</i> , <i>acvB</i>	<i>xadA</i>	1_MLSA-E	Yes
Oak_RO_NB22	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA	no	<i>nuoN</i> , <i>xadA</i> ***	3_MLSA	Yes
Oak_RO_NB232	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA MLSA	no	<i>nuoN</i> , <i>xadA</i> ***	3_MLSA	No
BlackOak_12T1_NH	<i>Q. velutina</i>	2008	Salem, NH	-E MLSA	<i>copB</i>	<i>xadA</i>	3_MLSA-E	Yes
PinOak_13T1_MD	<i>Q. palustris</i>	2008	Salisbury, MD	-E MLSA	<i>copB</i>	<i>xadA</i>	3_MLSA-E	No
PinOak_14T2_NJ	<i>Q. palustris</i>	2011	Cranbury, NJ	-E MLSA	<i>copB</i>	<i>xadA</i>	3_MLSA-E	No
RedOak_2T2_VT	<i>Q. rubra</i>	2008	Vermont	-E	<i>copB</i>	<i>xadA</i>	3_MLSA-E	No
Oak_PO_NB4	<i>Q. palustris</i>	2011	New Brunswick, NJ**	MLSA	no	<i>xadA</i>	4_MLSA	Yes
Oak_RO_NB5	<i>Q. rubra</i>	2011	New Brunswick, NJ**	MLSA	no	<i>xadA</i>	4_MLSA	No
Oak_PO_NB6	<i>Q. palustris</i>	2011	New Brunswick, NJ	MLSA	no	<i>none</i> ***	4_MLSA	No
Oak_PO_PK6	<i>Q. palustris</i>	2011	Cranbury, NJ	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_PO_2C4	<i>Q. palustris</i>	2011	Cranbury, NJ**	MLSA	no	<i>xadA</i>	4_MLSA	No

Oak_RO_2C1	<i>Q. rubra</i>	2011	Cranbury, NJ	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_RO_NB23	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_RO_NB24	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_PO_5OUT	<i>Q. palustris</i>	2008	Missouri*	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_EBO_0923	<i>Q. velutina</i>	2011	Woodbine, NJ	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_EBO_6923	<i>Q. velutina</i>	2011	Woodbine, NJ	MLSA	no	<i>xadA</i> ***	4_MLSA	No
Oak_PO_EW10	<i>Q. palustris</i>	2011	East Windsor, NJ	MLSA	no	<i>xadA</i> ***	5_MLSA	Yes
PinOak_11T1_DC	<i>Q. palustris</i>	2008	Washington, D.C.	MLSA	no	<i>xadA</i>	5_MLSA-E	Yes
Oak_RO_NB21	<i>Q. rubra</i>	2011	New Brunswick, NJ	-E	no	<i>holC,</i> <i>xadA</i> ***	6_MLSA	Yes
PinOak_2C4_NJ	<i>Q. palustris</i>	2011	Cranbury, NJ**	MLSA	no	<i>xadA</i>	6_MLSA-E	Yes
Oak_RO_1C1	<i>Q. rubra</i>	2011	Cranbury, NJ	-E	<i>copB</i>	<i>holC,</i> <i>xadA</i> ***	7_MLSA	Yes
PinOak_15T1_MD	<i>Q. palustris</i>	2008	Del Mar, MD	MLSA	no	<i>xadA</i>	7_MLSA-E	Yes
PinOak_16T2_NJ	<i>Q. palustris</i>	2011	Cranbury, NJ	MLSA	no	<i>xadA</i>	7_MLSA-E	No
RedOak_1T1_PA	<i>Q. rubra</i>	2008	Chatham, PA	MLSA	no	<i>xadA</i>	7_MLSA-E	No
PinOak_2T1_PA	<i>Q. palustris</i>	2008	Philadelphia, PA	MLSA	no	<i>xadA</i>	7_MLSA-E	No
RedOak_3T2_WV	<i>Q. rubra</i>	2008	Shepherdstown, WV	MLSA	no	<i>xadA</i>	7_MLSA-E	No
RedOak_4T2_WV	<i>Q. rubra</i>	2008	Keyser, WV	MLSA	no	<i>xadA</i>	7_MLSA-E	No
BurOak_8T2_MO_ou tlier	<i>Q. macrocarpa</i>	2008	Missouri*	MLSA	no	<i>xadA</i>	7_MLSA-E	No
RedOak_NB5_NJ	<i>Q. rubra</i>	2011	New Brunswick, NJ**	MLSA	no	<i>xadA</i>	7_MLSA-E	No

Oak_RO_NB26	<i>Q. rubra</i>	2011	New Brunswick, NJ	MLSA MLSA	<i>copB, acvB,</i> <i>nuoL</i>	<i>holC,</i> <i>xadA***</i>	8_MLSA	Yes
RedOak_4T1_DE	<i>Q. rubra</i>	2008	Wilmington, DE	-E MLSA	no	<i>gaa, xadA</i>	8_MLSA-E	Yes†
RedOak_5T1_DE	<i>Q. rubra</i>	2008	Wilmington, DE	-E MLSA	no	<i>gaa, xadA</i>	8_MLSA-E	Yes†
PinOak_NB4_NJ	<i>Q. palustris</i>	2011	New Brunswick, NJ**	-E MLSA	<i>acvB</i>	<i>xadA</i>	9_MLSA-E	Yes

* = Oak_PO_5OUT and BurOak_8T2_MO_outlier represent distal geographical samples and provide additional data to test whether conservation/variation extends beyond the geographical scope of this study

**= Oak_PO_NB4/PinOak_NB4_NJ, Oak_RO_NB5/RedOak_NB5_NJ, and Oak_PO_2C4/PinOak_2C4_NJ represent identical samples in the MLSA (conserved) and MLSA-E (variable) studies respectively

***= *xadA* amplicons were obtained from the MLSA isolates to determine the extent of cross-study genotypic variation at that locus

†=Although RedOak_4T1_DE and RedOak_5T1_DE represent a haplotypic redundancy the samples were included to emphasis the geographically distinct SNP present in both individuals

Table 2.2. Southern-based Oak haplotypes profiles for those described in previous multilocus studies

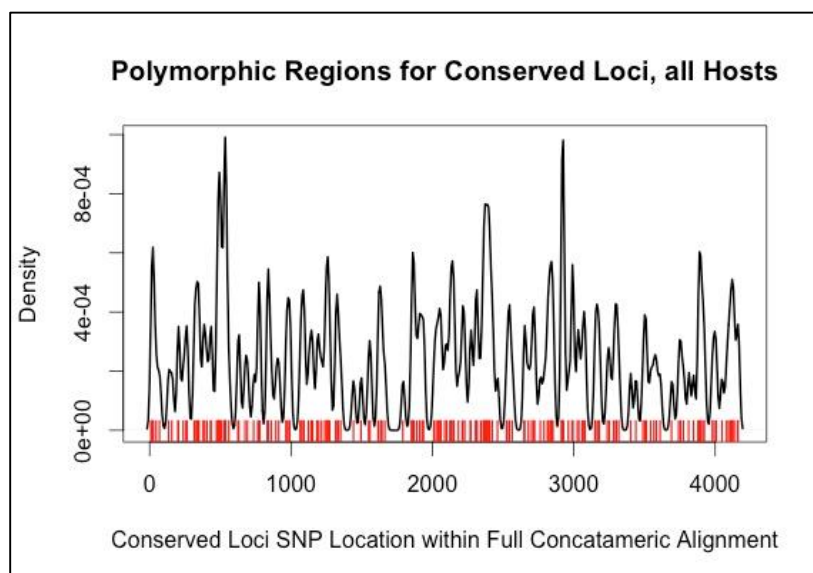
Host	Location	Original Name	Current Study Designation	Locus Set	Origin
<i>Quercus</i> sp.	Georgia	Oak17	Oak17	MLSA	Hendson et al.
<i>Quercus</i> sp.	Georgia	Oak24	Oak24	MLSA	Hendson et al.
<i>Quercus</i> sp.	Florida	Oak23	Oak23	MLSA	Hendson et al.
<i>Quercus cerris</i>	Lake County, Florida	Oak 95-1	Oak2	MLSA-E	Donald Hopkins
<i>Quercus nigra</i>	Lake County, Florida	Oak 92-10	Oak1	MLSA-E	Donald Hopkins
<i>Quercus</i> sp.	Palm Beach County, Florida	Oak 92-6	Oak3	MLSA-E	Donald Hopkins

Table 2.3. Previously described loci used for oak population based comparison

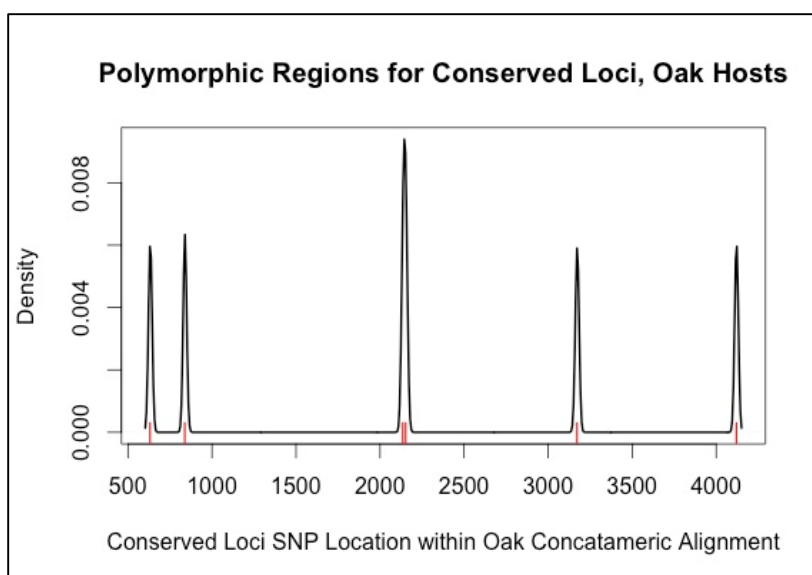
Gene Name/Gene ID	study designation	study amplicon size	primer set source
Virulence protein (<i>acvB</i>)	MLSA-E variable group	708	Parker et al.
copper resistance protein B precursor (<i>copB</i>)	MLSA-E variable group	789	Parker et al.
colicin V precursor (<i>cvaC</i>)	MLSA-E variable group	285	Parker et al.
fimbrial subunit precursor (<i>fimA</i>)	MLSA-E variable group	506	Parker et al.
glutaryl-7-ACA acylase precursor (<i>gaa</i>)	MLSA-E variable group	1064	Parker et al.
polygalacturonase precursor (<i>plgA</i>)	MLSA-E variable group	497	Parker et al.
pilin subunit (<i>pilA</i>)	MLSA-E variable group	353	Parker et al.
regulator of pathogenicity factors (<i>rpfF</i>)	MLSA-E variable group	777	Parker et al.
outer membrane afimbrial adhesin protein (<i>xadA</i>)	MLSA-E variable group	1060	Parker et al.
seroheme synthase (<i>cysG</i>)	MLSA housekeeping / <i>pilU</i> group*	501	Schuenzel et al.
DNA polymerase III holoenzyme chi subunit (<i>hoIC</i>)	MLSA housekeeping / <i>pilU</i> group*	318	Schuenzel et al.
ABC transporter sugar permease (<i>lacF</i>)	MLSA housekeeping / <i>pilU</i> group*	523	Schuenzel et al.
2-isopropylmalate synthase (<i>leuA</i>)	MLSA housekeeping / <i>pilU</i> group*	577	Schuenzel et al.
NADH quinone dehydrogenase (<i>nuoL</i>)	MLSA housekeeping / <i>pilU</i> group*	530	Schuenzel et al.
NADH-ubiquinone oxireductase subunit N (<i>nuoN</i>)	MLSA housekeeping / <i>pilU</i> group*	751	Schuenzel et al.
ubiquinol cytochrome C oxidoreductase cytochrome C1 subunit (<i>petC</i>)	MLSA housekeeping / <i>pilU</i> group*	495	Schuenzel et al.
twitching motility protein (<i>pilU</i>)	MLSA housekeeping / <i>pilU</i> group*	472	Schuenzel et al.

Figure 2.1. Conserved (MLSA) and Variable (MLSA-E) SNP Profiles for both Complete Isolate/Environmental Sample Concatemerized Loci and Oak only Concatemerized Loci.

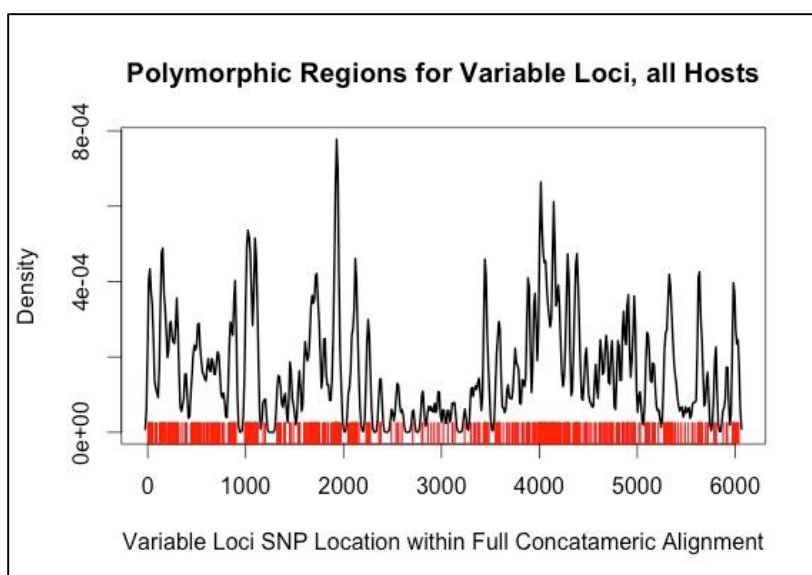
The horizontal axis is the relative position for non-indel based polymorphisms in both complete isolate/environmental sample and oak only concatemerizations and the vertical axis represents a variability density measuring the dissimilarity of the considered region relative to respective genetic profile. Both pan-host SNP profiles confirm the high level of diversity known to exist in *X. fastidiosa* subspecies loci. Considering the conserved oak profile, there are several discrete polymorphic regions spread over several loci (*holC*, *lacF*, *nuoL*, *nuoN*, and *pilU*). The variable oak loci show one discrete region (*gaa*) and a highly variable region containing many polymorphisms (*xadA*). The exact positions of the concatemerized loci are provided adjacent to the figures as a location aid.



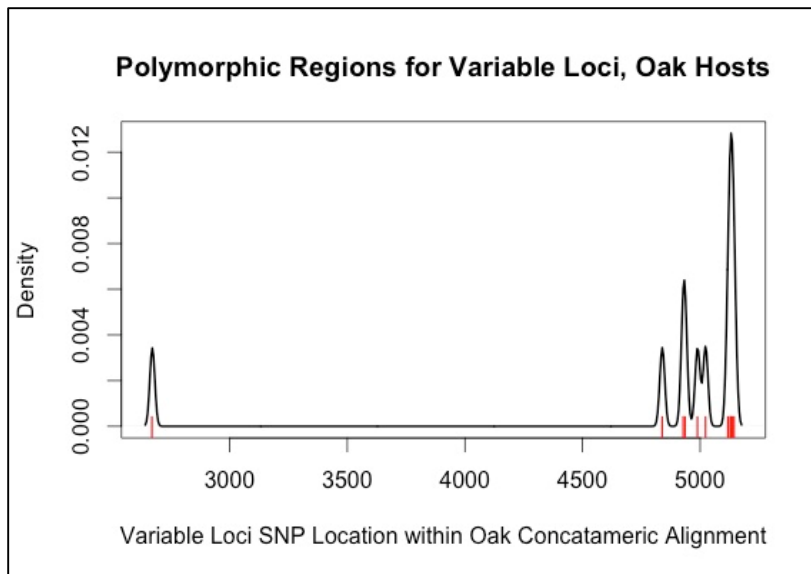
Gene	Concatemic position
<i>cysG</i>	1-501
<i>holC</i>	502-819
<i>lacF</i>	820-1342
<i>leuA</i>	1343-1919
<i>nuoL</i>	1920-2449
<i>nuoN</i>	2450-3200
<i>petC</i>	3201-3695
<i>pilU</i>	3696-4167



Gene	Concatemeric position
<i>cysG</i>	1-501
<i>hoIC</i>	502-819
<i>lacF</i>	820-1342
<i>leuA</i>	1343-1919
<i>nuoL</i>	1920-2449
<i>nuoN</i>	2450-3200
<i>petC</i>	3201-3695
<i>pilU</i>	3696-4167



Gene	Concatemeric position
<i>acvB</i>	1-708
<i>copB</i>	709-1542
<i>cvaC</i>	1543-1827
<i>fimA</i>	1828-2333
<i>gaa</i>	2334-3397
<i>pglA</i>	3398-3894
<i>pilA</i>	3895-4247
<i>rpfF</i>	4248-5024
<i>xadA</i>	5025-6084



Gene	Concatemeric position
<i>acvB</i>	1-708
<i>copB</i>	709-1542
<i>cvaC</i>	1543-1827
<i>fimA</i>	1828-2333
<i>gaa</i>	2334-3397
<i>pglA</i>	3398-3894
<i>pilA</i>	3895-4247
<i>rpfF</i>	4248-5024
<i>xadA</i>	5025-6084

Figure 2.2. Bayesian phylogeny consisting of the concatenated MLSA-E loci for the leveraged MLSA-E data and the novel Northeastern and Mid-Atlantic oak haplotypes categorized in Table 1.

Bayesian reconstruction via BEAST v 1.7.5 of a 1,000,000,000 generation Markov chain Monte Carlo (MCMC) for the concatenated MLSA-E loci. A burnin of 15% was used to obtain a single maximum credibility tree. The resulting tree was rooted using the subspecies *pauca* as an out group (Orange). Corresponding posterior probability values greater than 0.50 were placed at the appropriate nodes as a measure of construction confidence. The substitution rate of 0.006 is reported with a corresponding length to assess evolutionary changes corresponding to branch lengths. The oak specific clading within the multiplex main clade is set off from the master tree via differential shading, and oak haplotype groupings are labelled. The additional subspecies (*fastidiosa*, *sandyi*, and *pauca*) are labelled and mark previously described phylogenetic segregation within *X. fastidiosa* subspecies taxonomy.

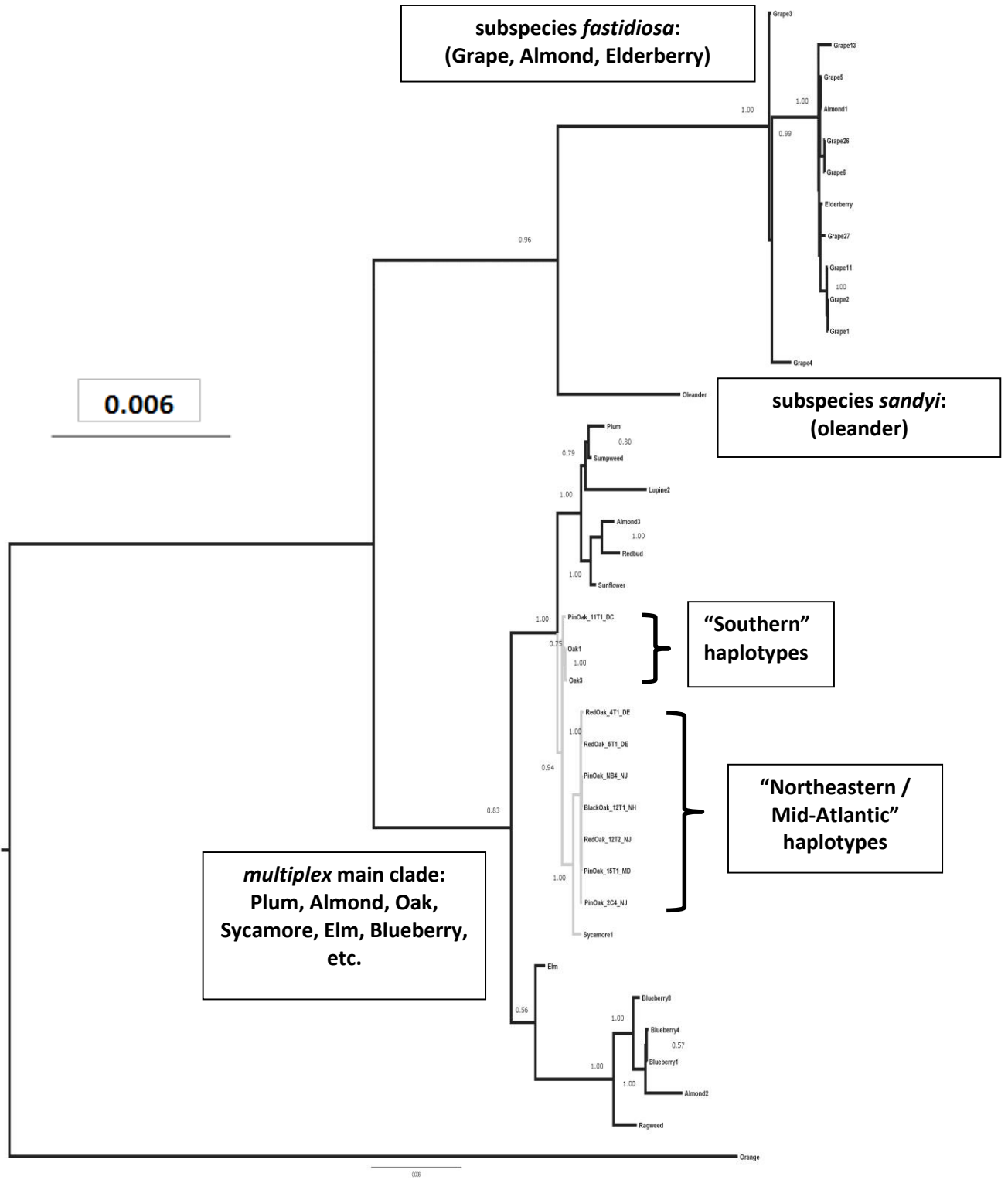


Figure 2.3. Insertion / Deletion (Indel) Profiles and Locations for the gene fragments *acvB*, *copB*, and *nuoL*.

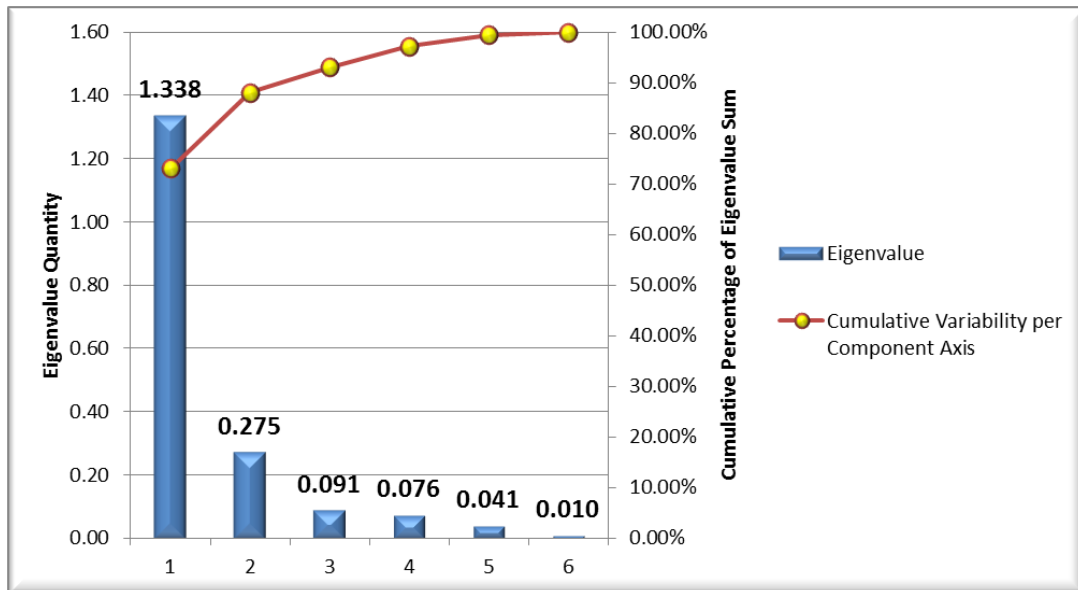
Previously described MLSA-E (*acvB*, *copB*) and MLSA (*nuoL*) oak based genotypes are compared to those containing the indel profiles from this study. Pertinent residue positions are provided to further assess indel character. Indel patterns causing truncated protein products are marked "STOP" at their terminus, while those with uninterrupted translation are marked "Translational Continuity".

*PinOak_NB4_NJ/Oak_PO_NB4 and RedOak_NB5_NJ are serving as consensus amino acid sequences in the population

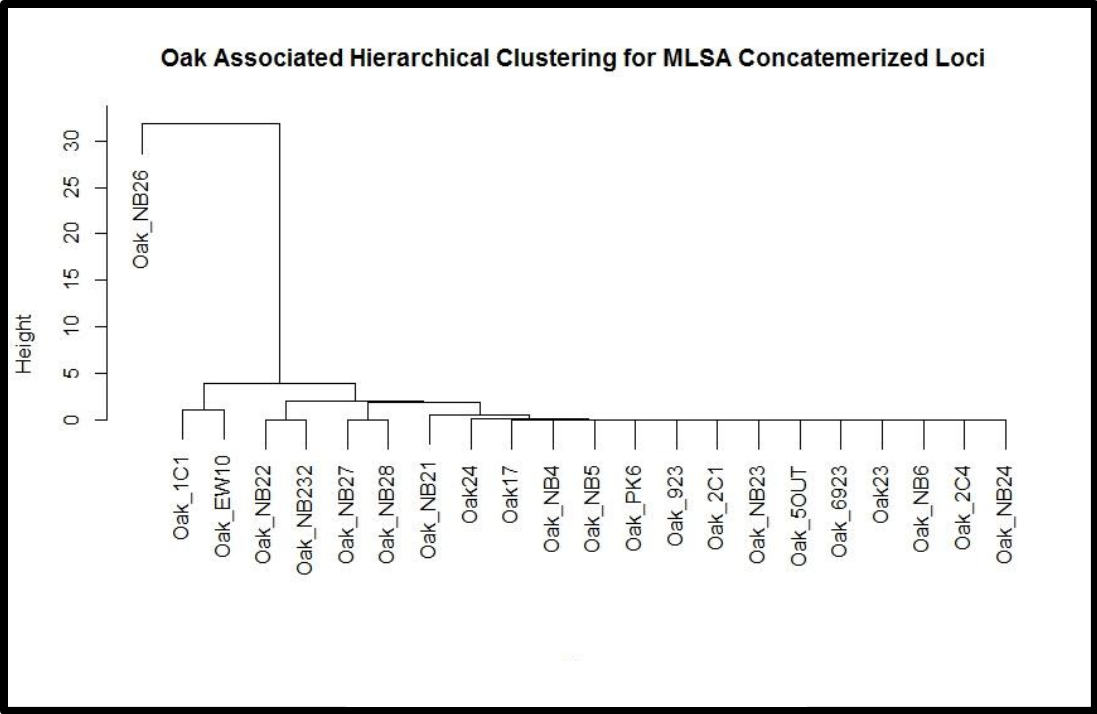
<i>acvB</i>	
Oak2	D L S R V I A D Y Q R R W G A H Q V I L I G Y S F G A D V - - - - -
Oak1	D L S R V I A D Y Q R R W G A H Q V I L I G Y S F G A D V - - - - -
Oak3	D L S R V I A D Y Q R R W G A H Q V I L I G Y S F G A D V - - - - -
RedOak_12T2_NJ	D L S R V I A D Y Q R R W G A H Q V I L I G Y S F G A D V - 13 BP INSERTION
PinOak_NB4_NJ	D L S R V I - - - - 55 BP DELETION - - - - -
Oak_RO_NB26	- - - - - 172 BP DELETION - - - - -
RedOak_NB5_NJ*	D L S R V I A D Y Q R R W G A H Q V I L I G Y S F G A D V - - - - -
Residue	1 1 5 1 6 3
<i>copB</i>	
Oak2	L S E H T Q M D H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
Oak1	L S E H T Q M D H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
Oak3	L S E H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
BlackOak_12T1_NH	L S E H T Q M D H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
PinOak_13T1_MD	L S E H T Q M D H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
PinOak_14T2_NJ	L S E H T Q M D H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
RedOak_12T2_NJ	L S E H T Q M D H T Q M D H T Q M D H T Q M D H T Q M D H T G M D H A I H G A T T R Translational Continuity
RedOak_2T2_VT	L S E H T Q M D H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
Oak_RO_1C1	L S E H T Q M D H T Q M D H T Q M D H T Q M D H T Q M D H T G M D H A I H G A T T R Translational Continuity
Oak_RO_NB26	L S E H T Q M D H T Q - - - - M D H T G M D H T G M D H T G M D H A I H G A T T R Translational Continuity
PinOak_NB4_NJ*	L S E H T Q M D H T Q - - - - - M D H T G M D H A I H G A T T R Translational Continuity
Residue	2 0 6 1
<i>nuoL</i>	
Oak17	G H D A D D H V N T H T S N D D H A H G V H Translational Continuity
Oak24	G H D A D D H V N T H T S N D D H A H G V H Translational Continuity
Oak23	G H D A D D H V N T H T S N D D H A H G V H Translational Continuity
Oak_RO_NB26	G H D A D D - - - - - H A H G V H Translational Continuity
Oak_PO_NB4*	G H D A D D H V N T H T S N D D H A H G V H Translational Continuity
Residue	3 2 5 3

Figure 2.4. Eigenvalue measure, Complete Linkage Hierarchical Clustering, and Two Axis Principal Component Analysis (PCA) for Concatemeric MLSA and MLSA-E Oak Associated Loci.

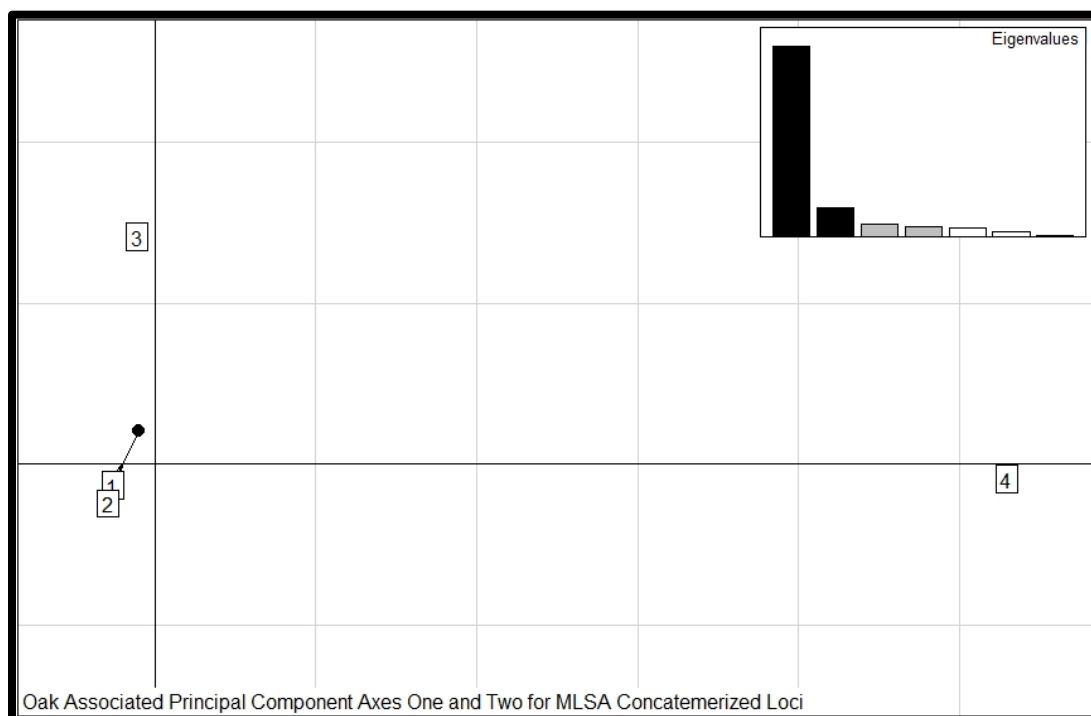
Panel "A" and panel "D" are measures of eigenvalue magnitude within the concatemeric MLSA and MLSA-E loci. The first vertical axis represents the magnitude of each eigenvalue component, the second vertical axis represents the cumulative percentage of the eigenvalue sum, and the horizontal axis is a count of eigenvalue measure derived from the underlying datasets. The eigenvalue analysis serves as a measure of assurance for groupings in the respective complete linkage hierarchical clustering visualizations (panel "B" and "E"), and the two axis PCA displays (panel "C" and "F"). Panels "B" and "E" show complete linkage hierarchical clustering diagrams where the vertical axis represents a measure of difference in the respective oak associated concatemeric datasets. The complete linkage hierarchical clustering is built from the cumulative eigenvalue sums and represents four groupings for the MLSA oak dataset and six groupings for the MLSA-E oak dataset. Finally, panels "C" and "F" show a two axis PCA display of the oak associated groupings. Enumeration of membership is provided adjacent to the PCA to further define oak associated *X. fastidiosa* relationships.



A

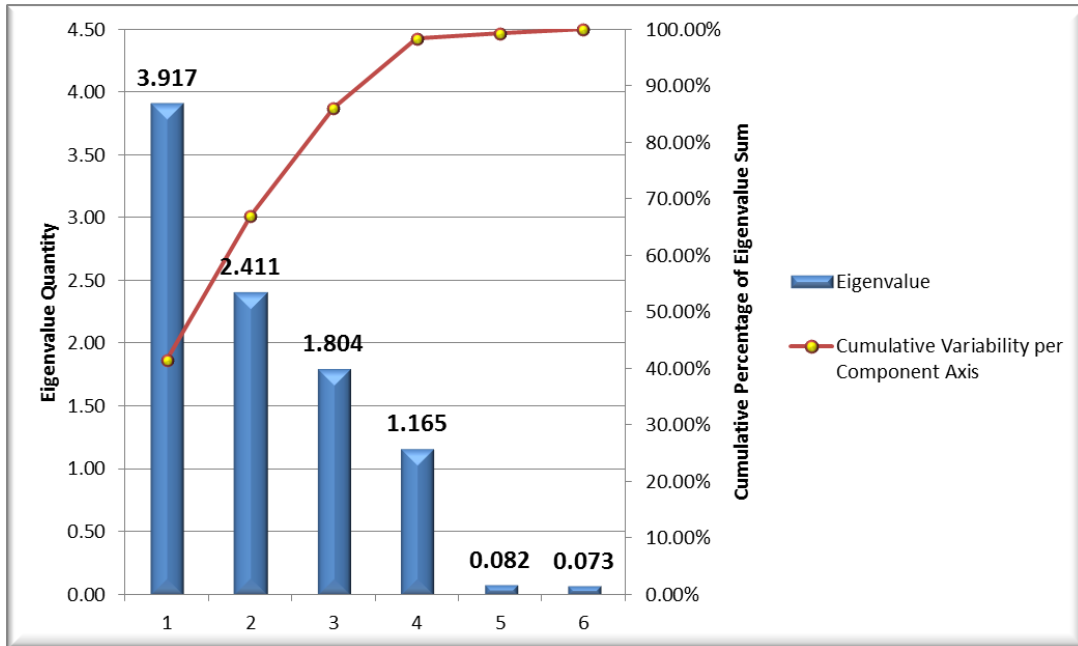


B

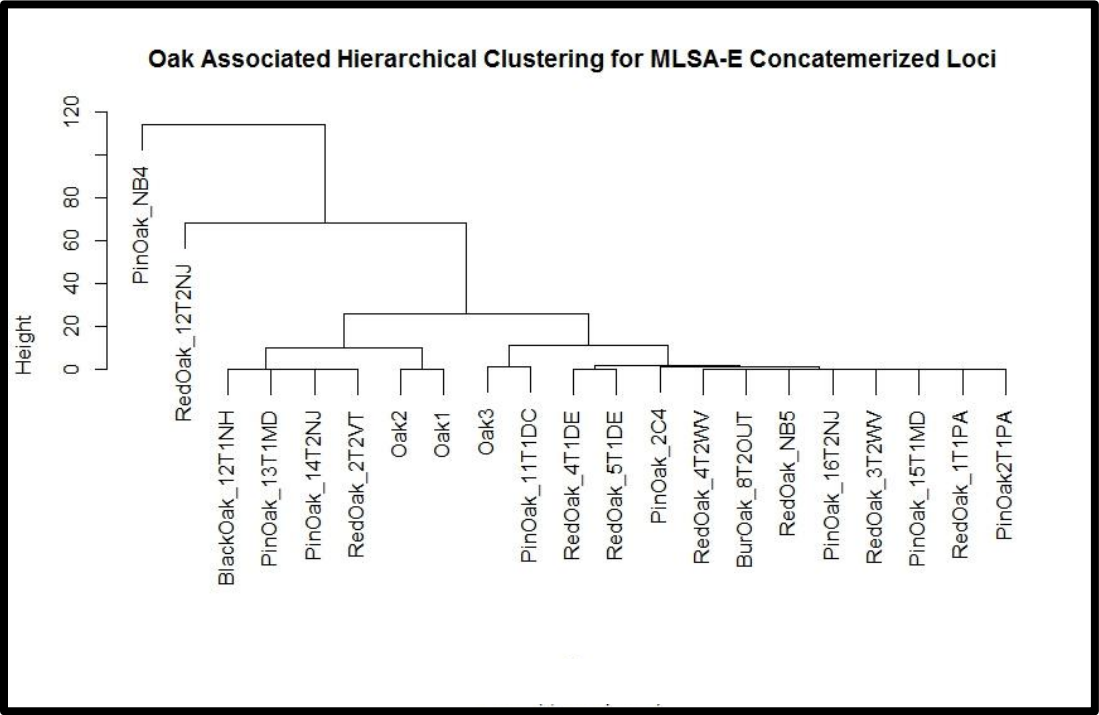


C

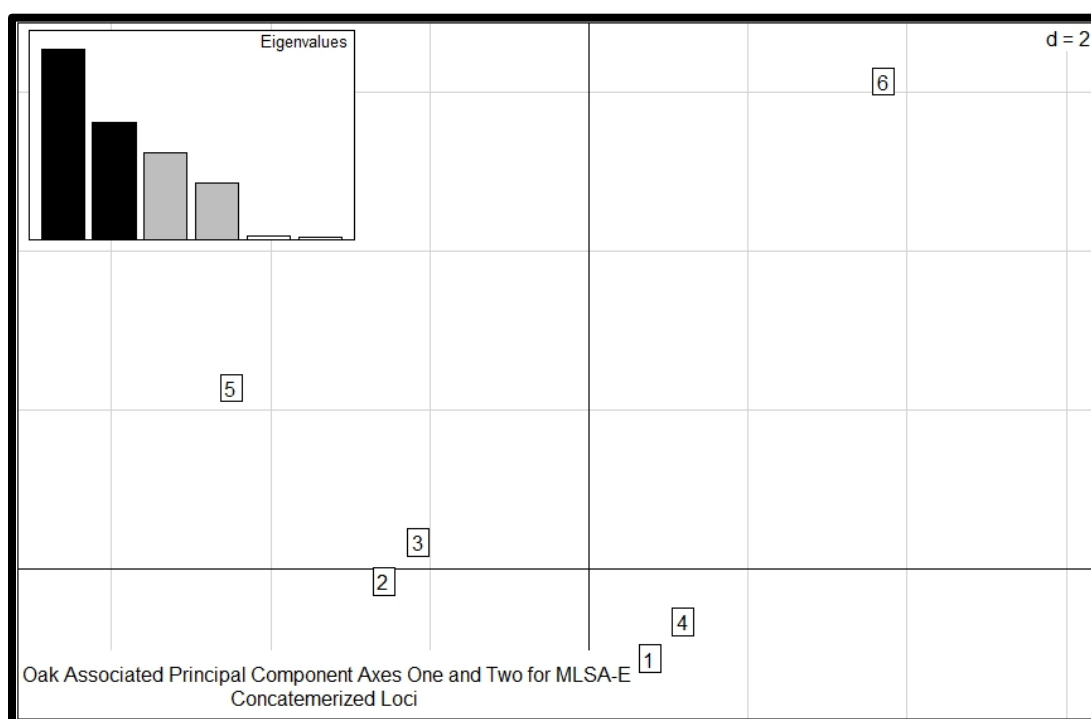
Isolate Name	Category Membership
Oak_EBO_0923	1
Oak_EBO_6923	1
Oak_PO_2C4	1
Oak_PO_5OUT	1
Oak_PO_NB4	1
Oak_PO_NB6	1
Oak_PO_PK6	1
Oak_RO_2C1	1
Oak_RO_NB21	1
Oak_RO_NB23	1
Oak_RO_NB24	1
Oak_RO_NB27	1
Oak_RO_NB28	1
Oak_RO_NB5	1
Oak17	1
Oak23	1
Oak24	1
Oak_RO_NB22	2
Oak_RO_NB232	2
Oak_PO_EW10	3
Oak_RO_1C1	3
Oak_RO_NB26	4



D



E



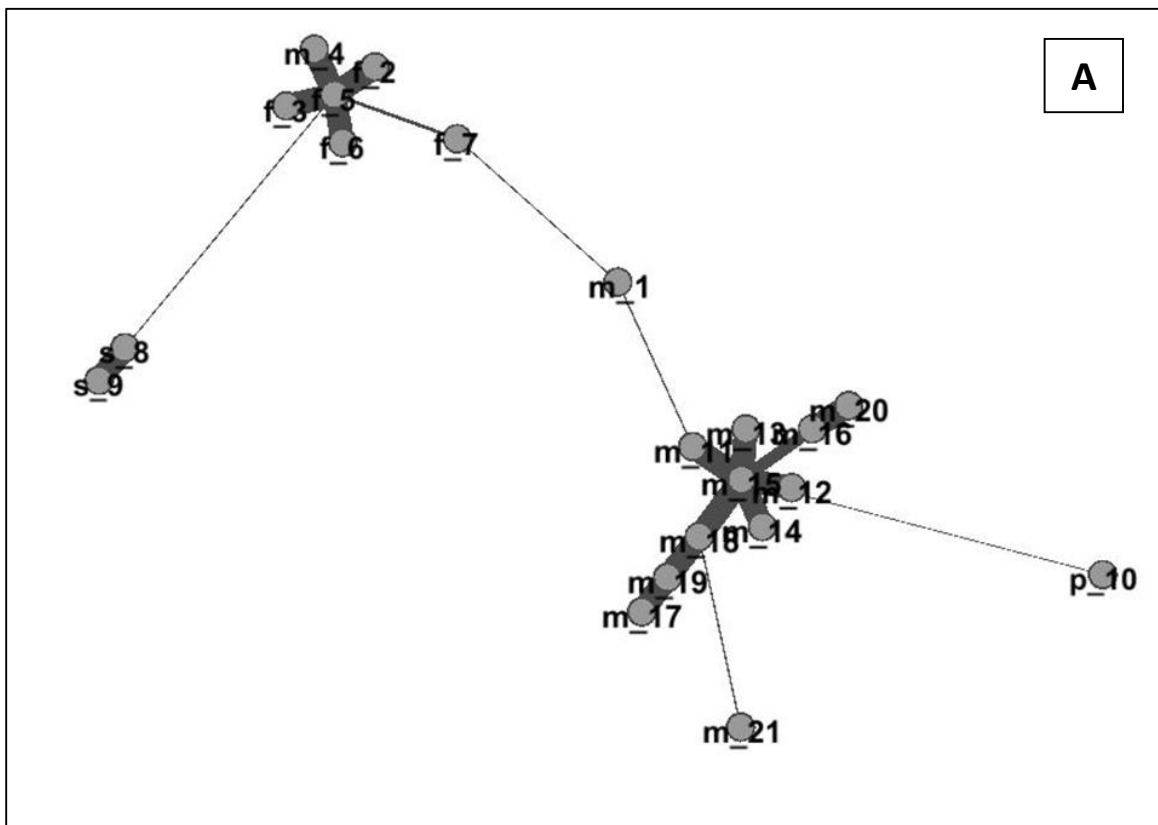
F

Isolate Name	Category Membership
Oak3	1
PinOak_11T1_DC	1
Oak1	2
Oak2	2
BlackOak_12T1_NH	3
PinOak_13T1_MD	3
PinOak_14T2_NJ	3
RedOak_2T2_VT	3
BurOak_8T2_MO_outlier	4
PinOak_15T1_MD	4
PinOak_16T2_NJ	4
PinOak_2C4_NJ	4
PinOak_2T1_PA	4
RedOak_1T1_PA	4
RedOak_3T2_WV	4
RedOak_4T1_DE	4
RedOak_4T2_WV	4
RedOak_5T1_DE	4
RedOak_NB5_NJ	4
RedOak_12T2_NJ	5
PinOak_NB4_NJ	6

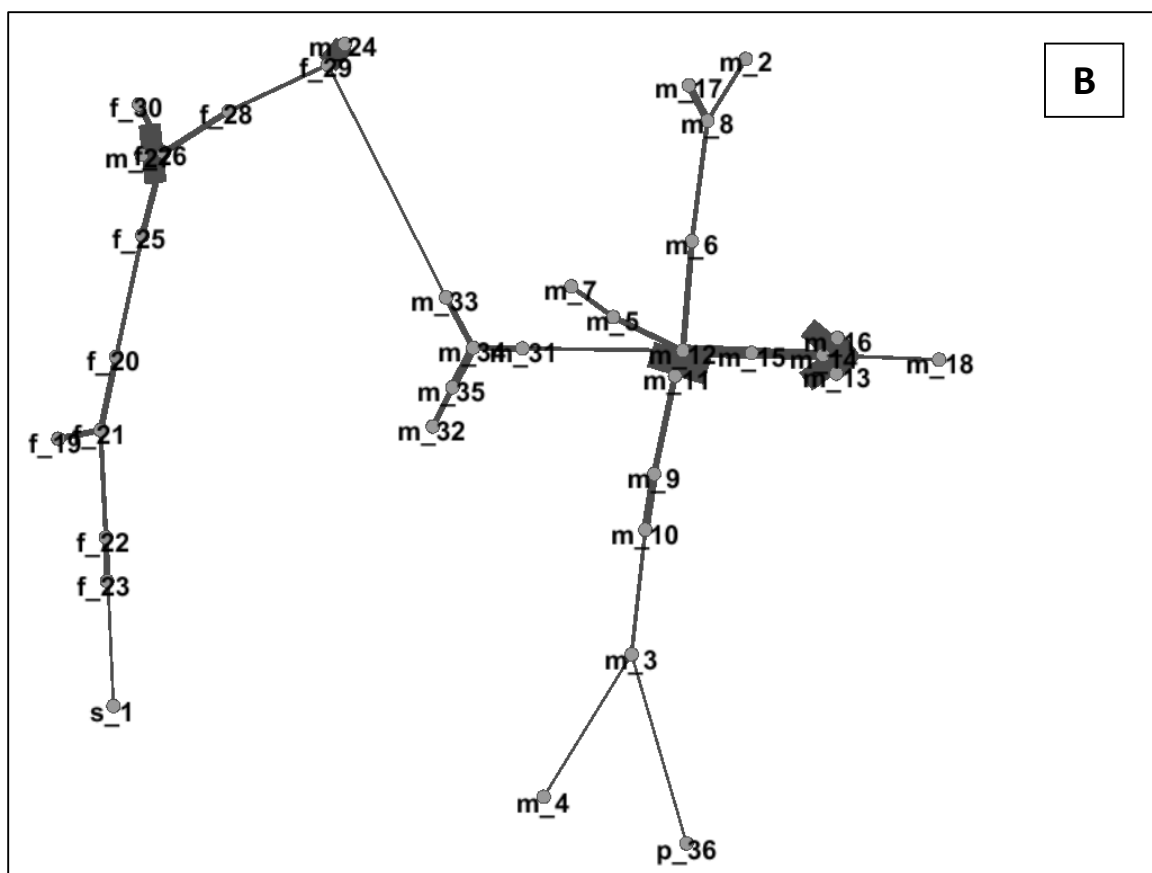
Figure 2.5. Minimum Spanning Tree for both the complete MLSA and complete MLSA-E haplotype *X. fastidiosa* Concatemeric Locus Collection

Points along both the MLSA ("A") and MLSA-E ("B") trees correspond to host associated haplotypes. Euclidean distance between points can be considered a measure of dissimilarity and line weighting (width) can be considered a measure of similarity between closely associated haplotypes.

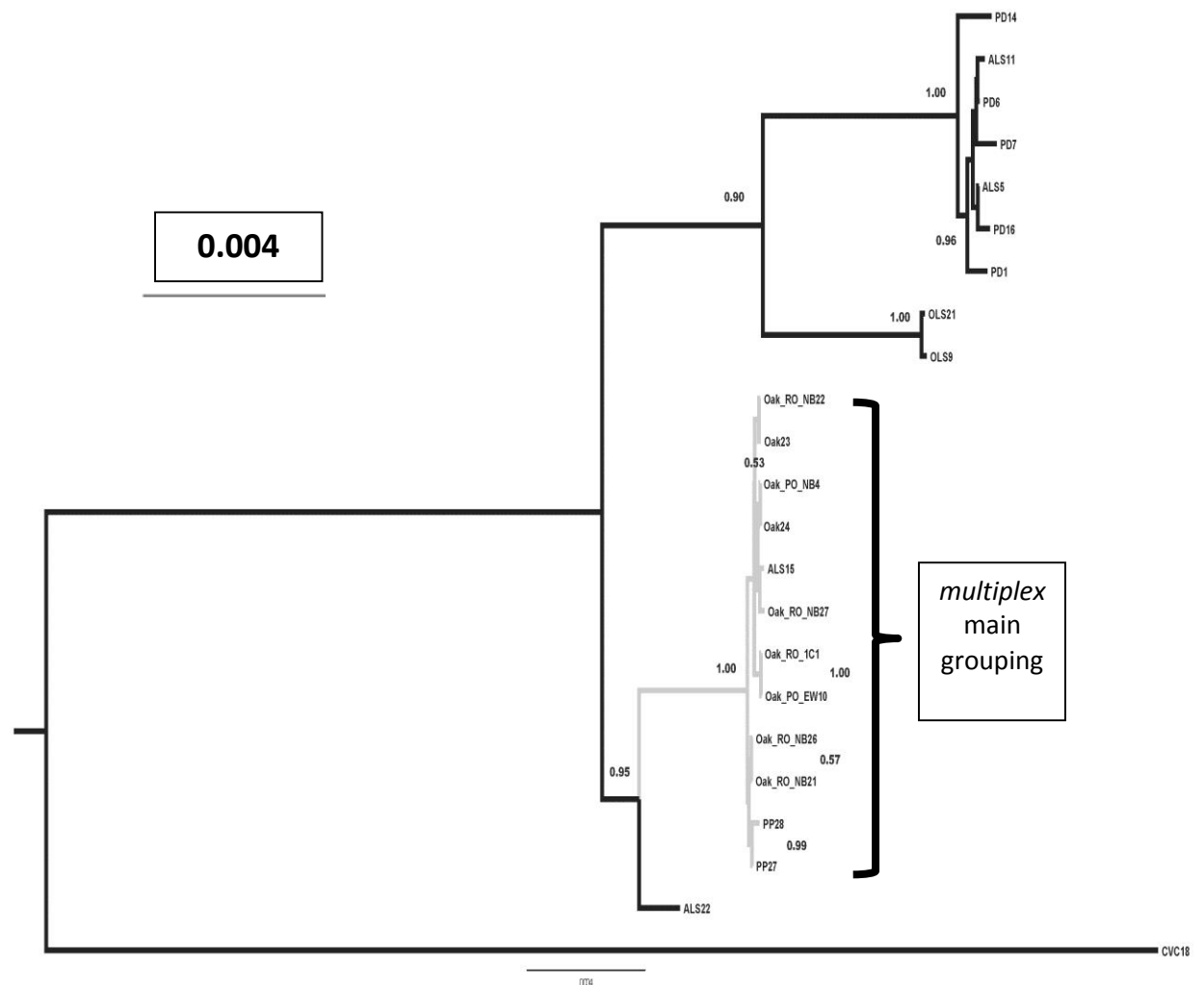
*Naïve category assignment was chosen to distinguish haplotypes, therefore, the previously established taxonomy of the considered elderberry and almond associated haplotypes was ignored. The minimum spanning tree correctly represented both as more subspecies *fastidiosa*-like despite bearing a multiplex moniker.



MLSA Haplotype Designation	Total Individuals	Host	Grouping and Representative
f_2	1	Grape	PD1
f_3	1	Grape	PD16
f_5	4	Grape	ALS11
f_6	1	Grape	PD7
f_7	1	Grape	PD14
m_1	2	Almond	ALS22
m_11	3	Almond	ALS15
m_12	2	Oak (Northern)	Oak_RO_NB27
m_13	1	Oak (Southern)	Oak23
m_14	2	Oak (Northern)	Oak_RO_NB22
m_15	13	Oak (Northern/Southern)	Oak_PO_NB4
m_16	1	Oak (Northern)	Oak_PO_EW10
m_17	1	Plum	PP28
m_18	1	Oak (Northern)	Oak_RO_NB21
m_19	1	Plum	PP27
m_20	1	Oak (Northern)	Oak_RO_1C1
m_21	1	Oak (Northern)	Oak_RO_NB26
m_4	1	Almond	ALS5
p_10	1	Orange	CVC18
s_8	5	Oleander	OLS8
s_9	1	Oleander	OLS21



Haplotype Designation	Total Individuals	Host	Grouping and Representative
f_19	1	Grape	Grape (<i>fastidiosa</i>) grouping
f_20	1	Grape	Grape (<i>fastidiosa</i>) grouping
f_21	5	Grape	Grape (<i>fastidiosa</i>) grouping
f_22	1	Grape	Grape (<i>fastidiosa</i>) grouping
f_23	1	Grape	Grape (<i>fastidiosa</i>) grouping
f_25	1	Grape	Grape (<i>fastidiosa</i>) grouping
f_26	2	Grape	Grape (<i>fastidiosa</i>) grouping
f_28	3	Grape	Grape (<i>fastidiosa</i>) grouping
f_29	11	Grape	Grape (<i>fastidiosa</i>) grouping
f_30	1	Grape	Grape (<i>fastidiosa</i>) grouping
m_10	4	Oak (Northeastern/Mid-Atlantic)	BlackOak_12T1_NH
m_11	1	Oak (Southern)	Oak3
m_12	1	Oak (Northeastern/Mid-Atlantic)	11T1_DC
m_13	1	Oak (Northeastern/Mid-Atlantic)	2C4NJ
m_14	8	Oak (Northeastern/Mid-Atlantic)	15T1_MD
m_15	2	Sycamore	Sycamore1
m_16	2	Oak (Northeastern/Mid-Atlantic)	4T1_DE / 5T1_DE
m_17	1	Almond	Almond3
m_18	1	Oak (Northeastern/Mid-Atlantic)	RedOak_12T2
m_2	1	Lupine	Lupine2
m_24*	2	Elderberry	Elderberry
m_27*	1	Almond	Almond1
m_3	1	Oak (Northeastern/Mid-Atlantic)	PinOak_NB4
m_31	1	Almond	Almond2
m_32	1	Ragweed	Ragweed
m_33	3	Blueberry	Blueberry4
m_34	3	Blueberry	Blueberry1
m_35	2	Blueberry	Blueberry8
m_4	1	Elm	Elm
m_5	1	Plum	Plum
m_6	1	Redbud	Redbud
m_7	1	Sumpweed	Sumpweed
m_8	1	Sunflower	Sunflower
m_9	2	Oak (Southern)	Oak1
p_36	1	Orange	Orange
s_1	1	Oleander	Oleander



CHAPTER 3. The Intrasubspecies Comparative Genomics of RNB1, a Northeastern *Quercus palustris* (Pin Oak) associated *Xylella fastidiosa* isolate"

ABSTRACT

Xylella fastidiosa is a Gram negative plant pathogen that has been well described in many economically important crops. While much is known about the populations infecting various host species like *Vitis* (grape), little is known about the genetic composition of *X. fastidiosa* in hardwood stands. In particular, *Quercus* species make up a large portion of Northeastern forest expanse and present concentrated inoculum sources that can be either contiguous to or within reasonable proximity of commercial growing interests. In an effort to expand the findings in previous locus oriented *X. fastidiosa* studies, a Northeastern *Quercus palustris* (Pin Oak) isolate was sequenced (Pacific Bio) at 91x coverage and designated RNB1. While it is true that NCBI currently houses a Southern United States *Quercus rubra* (red oak) derived isolate designated Griffin-1, locus based studies have also shown that meaningful genetic differences exists among regionally separated oak populations. The approach to this study was twofold. First, RNB1 based comparative analyses were conducted among two well described isolates in grape (Temecula1) and almond (M12). Additionally, the red oak derived isolate, Griffin-1, was used as geographical counterpoint for the comparative analysis. Supplemental analysis with fourteen additional NCBI derived isolates was also performed. Second, this study forwards a simple, modular, open source driven pipeline for prokaryotic annotation. This work, therefore, expands current knowledge of the global genetic character of Northeastern oak associated *X. fastidiosa* isolates relative to other host and geographically specific genomes. Enhanced understanding of this pathogen may lend insight into vector acquisition, pathogen transmission, and the potential amelioration of symptomology.

The plant pathogen *Xylella fastidiosa* (Wells et al. 1987) has been responsible for substantial loss in select commodity crops for much of the twentieth century (Hopkins et al. 2002). Within North America, those genera most affected include " *Vitis* (Pierce's disease), *Prunus* (Almond Leaf Scorch/Plum Scald/Phony Peach Disease), *Vaccinium* (Blueberry Scorch), and *Quercus* (Bacterial Leaf scorch of oak) (Pierce 1892; Davis et al. 1980; Hearon et al. 1980; Davis et al. 1981; Grebus et al. 1996). Amidst the backdrop of the recent "omics" revolution technologies have ushered in methods to obtain more global understandings of this pathogen's genetic character. Since the completion of the first South American *X. fastidiosa* sequencing project in 2000 (Simpson et al. 2000), the North American strains Temecula1 (Grape host), M12 /M23 (Almond/Grape host), and Ann-1 (Oleander host) were sequenced shortly thereafter (Van Sluys et al. 2003; Bhattacharyya et al. 2002; Chen et al. 2010). These watershed projects provided complementation to the existing body of RFLP, RAPD, and short nucleotide composition based publications focused on pathogen detection and taxonomic assignment (Chen et al. 1992; Chen et al. 1995; Henderson et al. 2001; Rodrigues et al. 2003). At the time of drafting this article, no less than eighteen *X. fastidiosa* genomes now reside in the NCBI database with more likely on their way. Despite such a proliferation, much of the comparative genomics work has been focused away from the inclusion of the oak strain. Complementing the findings of recent multilocus sequence analyses (Schuenzel et al. 2005; Parker et al. 2012), previously undescribed genetic diversity has been detected in oak associated *X. fastidiosa* populations in the

Northeastern and Mid-Atlantic regions of the United States (Behringer et al. pending submission). Especially important was the description of oak derived polymorphic loci that were found to be unique relative to similarly profiled Southern *Quercus* pathovars. For this reason, there is an underlying need for a more global genetic perspective within populations tied to this host. In 2013, a substantive step was taken with the publication of the first oak associated *X. fastidiosa* genome (Chen et al. 2013) isolated in Griffin, Georgia. Despite the appearance of an oak associated genome, the aforementioned geographically unique polymorphic findings justified the need for a Northeastern based genome. To date, no comparative studies have been conducted to unveil aspects of the oak population's genetic character relative to other important *X. fastidiosa* strains. While it is true that oak presence has declined in the North American forest landscape in the past forty years (Thompson et al. 2013), it nonetheless remains a dominant fixture in many Northeastern woodland stands. Such potential inoculum concentrations are often contiguous to areas of significant agricultural development, and need to be understood in the context of all *X. fastidiosa* subspecies prior to etiological calamities like non-native vector induced disease spread (Purcell and Saunders 1999).

From a taxonomic standpoint, the existing body of *X. fastidiosa* subspecies assignment largely associates host and strain (Schaad et al. 2004; Schuenzel et al. 2005; Hernandez-Martinez et al. 2006; Nunney et al. 2014). While this is generally true for subspecies assignments of Pierce's Disease (subspecies *fastidiosa*), Oleander Leaf Scorch (subspecies *sandyi*), Citrus Variegated Chlorosis (subspecies *pauca*), and the recently designated Mulberry Leaf scorch (subspecies *morus*), the fifth dominant taxonomic category (subspecies *multiplex*), contains a large grouping of scorch diseases associated with shade, nut, and stone fruit trees, alternative hosts like Sunflower, Sumpweed, and Lupine, and Blueberry (Parker et al. 2012). The oak isolate's membership in the *multiplex* main clade, therefore, makes further study of this protean

subspecies imperative. In conjunction with the above general pathology concerns, phylogenetic based topologies add evidence that multiplex based host jumping may be a tangible concern. Beyond filling knowledge gaps of oak population genomics relative to other *X. fastidiosa* subspecies, analyzing increased sample sizes of genomes for this important bacterium will continue to shed light on the global nature of its genetic underpinnings, and in doing, may lead toward a path of continued subspecies positioning and novel loci based diagnostic recognition. The objectives of this study are twofold. First, this study provides a comparative genomic analysis of a Northeastern *Quercus palustris* (Pin Oak) derived bacterial isolate relative to several existing NCBI *X. fastidiosa* genomes. Several loci are profiled across an expanded list of *X. fastidiosa* genomes relative to the Pin Oak derived isolate as well. This oak associated *X. fastidiosa* strain was isolated at Rutgers, the state university of New Jersey and, post genomic sequencing, has been designated RNB1. The chosen genomes for reciprocal comparison were Temecula1 (grape) (Van Sluys et al. 2003), M12 (almond) (Chen et al. 2010), and Griffin-1 (Red Oak) (Chen et al. 2013). The two former are among the best annotated and economically important genomes within the North American isolate collection. The latter most was chosen as both geographical and host counterpoint since its origination is Griffin, Georgia and its host is *Quercus rubra* (Red Oak). Second, from the vast array of genome-oriented analytical algorithms, this study presents a simple, scalable, and modular annotation based pipeline for both *X. fastidiosa* and general prokaryotic genomes. The results of the RNB1 sequencing project (Supplemental 1), therefore, expand the current genetic knowledge of oak associated *X. fastidiosa* populations relative to several major *X. fastidiosa* subspecies.

Materials and Methods:

Isolation, DNA Extraction, and Genomic Sequencing:

Foliated oak branches bearing the greatest amount of “scorch” symptomology were removed using either a telescoping pole trimmer or hand shears. Leaves displaying scorch symptoms were then removed from branches and the tissue was surface sterilized using a 60 second 70% ethanol immersion followed by a 60 second 1% sodium hypochlorite immersion. Specimens were then washed three to five times in sterile deionized water and allowed to dry in an aseptic fume hood for 30 minutes. Approximately 1/2 inch of leaf petiole and 1/2 inch of leaf midrib were excised bearing an approximate weight of 0.100g - 0.300g. The excised portions of leaf tissue were placed in 1.5 mL microcentrifuge tubes filled with 1 mL of sterile molecular biology grade water (Fischer Scientific), agitated for five minutes, and left immersed for an additional 30 minutes. 200µL of the resulting solution was then plated on Periwinkle Wilt Gel media (PWG) (Davis et al. 1981), and the resulting petri dishes were placed in a 28°C incubator. Putative colonies bearing *X. fastidiosa* morphology were selected, and single colony quadrant streaked onto additional aseptic PWG plates. Colony Polymerase Chain Reactions (PCR) was then performed utilizing previously described primer sets (Barry et al. 1991; Schuenzel et al. 2005; Parker et al. 2012) (Table 3.1). Colony PCR amplicons were excised from 1% agarose gels and purified using the QIAquick Gel Extraction Kit protocol (Qiagen). Final verification of *X. fastidiosa* assignment was done via Sanger sequencing (Genewiz, South Plainfield, New Jersey). *X. fastidiosa* genomic DNA extraction was performed according to Wizard Genomic DNA Purification Kit Gram negative protocols (Promega). Sequencing was performed per standard protocols (Yale Center for Genome Analysis - YCGA). In short, libraries were prepared following Pacific Biosciences guidelines and sequenced on SMRT cells using Pacific Biosciences RS sequencing technology. The fragment library preparation was performed using the Pacific Biosciences DNA Template Prep Kit 2.0 (3kb to 10 kb range), and size selection and library purification was performed using 0.45X AMPure beads (Beckmann-Coulter Genomics). The

library was bound to C2 DNA polymerase, mag-bead loaded into two SMRT cells and observed.

N50 was confirmed to be 9,408 bps.

Assembly was performed using the RS_HGAP_Assembly.3 (Hierarchical Genome Assembly Process) protocol incorporated into SMRT Analysis v2.3 (Pacific Biosciences). The 64,426 post-filter reads were assembled into 22 overall contigs with total contig length 2,878,121 bps. GC content was confirmed at 53.16% with AT content being 46.84%, and average coverage was 91x for the project with mapping concordance of the assembly being 99.92%. Initial genome annotation was performed by The RAST Server (Rapid Annotations using Subsystems Technology) (Aziz et al. 2008) and GLIMMER (Gene Locator and Interpolated Markov ModelER) (Delcher et al. 1999). The unitig was later analyzed via GeneMark (Lukashin et al. 1998) for a more conservative estimate of putative open reading frame (ORF) calling.

Comparative Genomics Pipeline:

To ensure continuity between assembly and putative ORF calling between RNB1 and the other compared genomes, the more conservative GeneMark (Borodovsky and McIninch 1993) non-Hidden Markov Model (HMM) methodology was employed for the genomes Temecula1 (Van Sluys et al. 2003), M12(Chen et al. 2010), Griffin-1 (Chen et al. 2013), and RNB1. The non-HMM GeneMark algorithm predicts fewer putative ORFs and was chosen for its conservative gene calling methodology (Lukashin and Borodovsky 1998). This also resulted in a reduced number of putative ORFs relative to the initial GLIMMER (Delcher et al. 1999) based calls for RNB1.

Customized databases composed of *X. fastidiosa* genomic sequence were then built to effectively target query hits. For each of the genomes post GeneMark (Lukashin et al. 1998) ORF calling, BLASTp (Altschul et al. 1990) analysis was run in reciprocal fashion for Temecula1, M12, Griffin1, and RNB1. This totaled three queries against a the customized database composed of

only RNB1 sequence, and an additional three queries against a database composed first of M12 sequences, second of Temecula1 sequences, and finally Griffin-1 sequences. The reciprocal process ensures that post BLASTp runs, the database is then selected as a query with the initial query set then serving as the database. This serves as confirmation that non-hitting queries (“No hits found”) are captured for each genome- to- genome pair. The resulting hits were then parsed using SEQIO BioPerl (Stajich et al. 2002), and the top scoring hit by E-value for each reciprocal pair was captured. All results were then binned according to designated E-value categories or non-hitting returns. Those poor scoring parsed hits with E-values greater than “1e-10” were then subjected to a second BLASTp analysis against the remote Non-redundant protein sequences collection (nr) (NCBI). This was done to account for low point coverage in any of the considered sequencing projects where incorrect base calling or scaffold omissions could result in erroneously high E-value recovery. Resulting hits from the secondary BLASTp screening were then placed in secondary bins per the previously described categorizations and subjected to Pfam 27.0 (Finn et al. 2013) domain calling against the Pfam-A database using “hmmScan” (HMMER 3.1b1). Because (identity/similarity) domain motifs can be indicative of function, the Pfam based E-values were relaxed to include everything less than a Pfam-A full sequence E-value of 1.0. Subsequent discussion of any domains will be accompanied by the Pfam-A full sequence E-value as an added measure of confidence. Post Pfam-A annotation, results were then collated and run through “hmm2go” (HMMER2GO version 0.11) for Gene Ontology (GO) (Ashburner et al. 2000) categorization.

Supplementary Locus Amplification from Oak Based Environmental Samples:

Environmental samples for locus comparison were derived from infected oak stands and processed as previously described (Behringer et al. pending submission) (Table 3.4). Primer sets

for the two considered ORFs (ORF_415 and ORF_2081) are listed in Table 1 as well.

Comparative Genomics Data Presentation and Sources:

Venn diagram construction was accomplished via Venny 2.0 (Oliveros 2007), and supplementary “Sparklines” (Tuft 2004) were used to visualize the E-value binning of successive BLAST runs (Figure 1, Figure 3.2a, 3.2b). Ring rendering of both the full Pin Oak (RNB1) genome relative to the genomes of subjects M12, Griffin-1, and Temecula1 as well as the comparison of low scoring Pin Oak (RNB1) to the same respective genomes was accomplished via BLAST Ring Image Generator Version (BRIG Version 0.95)(Alikhan et al. 2011) (Figure 3.4, Figure 3.5). Finally, Multiple Correspondence Analysis (MCA) analysis figures were created in the R statistical computing /graphics environment version 3.0.2 (2013-09-25) -- "Frisbee Sailing"(Team 2012) with the aid of the dependency FactoMineR (Lê et al. 2008) and ggplot2 (Wickham 2011) (Figure 3.6a, 3.6c, 3.6d).

All genomes used for RNB1 comparison are housed at NCBI under the following RefSeq/INSDC links: Temecula1 (NC_004556.1 , AE009442.1), M12 (NC_010513.1 , CP000941.1), Griffin-1 (NZ_AVGA000000000.1 , AVGA000000000.1), 6c (NZ_AXBS000000000.1 , AXBS000000000.1), 9a5c (NC_002488.3 , AE003849.1), Ann-1 (NZ_CP006696.1 , CP006696.1), 32 (NZ_AWYH000000000.1 , AWYH000000000.1), ATCC 35879(NZ_JQAP000000000.1 , JQAP000000000.1), Dixon (NZ_AAAL000000000.2 , AAAL000000000.2), GB514 (NC_017562.1 , CP002165.1), M23 (NC_010577.1 , CP001011.1), MUL0034 (NZ_CP006740.1 , CP006740.1), Mul-MD (NZ_AXDP000000000.1 , AXDP000000000.1),sycamore Sy-VA (NZ_JMHP000000000.1 , JMHP000000000.1), ATCC 35871 (NZ_AUAJ000000000.1 , AUAJ000000000.1), Ann-1 (NZ_AAAM000000000.4 , AAAM000000000.4), EB92.1 (NZ_AFDJ000000000.1 , AFDJ000000000.1) (NCBI).

Results:

Pipeline based metrics and comparative genomic analysis:

The Pin Oak associated *multiplex* strain (RNB1) genome was run through GeneMark (non-HMM) and 2,487 ORFs were recovered. To maintain analytical continuity, the same process was repeated for the California almond strain (M12), the Georgia Red Oak strain (Griffin-1), and California grape strain (Temecula1). These respective runs yielded the following number of non-HMM derived ORFs: 1,913, 1,940, and 1,927.

The initial stage pipeline reciprocal BLASTp procedure was then executed to examine RNB1. RNB1 was first queried three times against the respective customized databases consisting of the M12, Griffin-1, and Temecula1 genomes, and was then, itself, made the database for queries derived from M12, Griffin-1, and Temecula1 respectively. This initial query procedure, representing half the reciprocal BLAST process, is presented as a Venn diagram in Figure 1. In this presentation, each query from either M12, Griffin-1, or Temecula1 was associated with the specific hit against the RNB1 database and pooled for subspecies overlap. The large number of database hits and their underlying low E-values confirms that a high degree of overall similarity is shared between the three *X. fastidiosa* subspecies. Those hits unique to each of the respective subspecies largely represent poor scoring and thus high E-value sequences. For example, 222 genes are unique to Temecula1 relative to RNB1. Existing phylogenetic work suggests accord with the Venn recovery given the larger number of Temecula1 specific hits (222) relative to the smaller number of M12 specific hits (73). The existing body of current *X. fastidiosa* taxonomic research consistently confirms that Temecula1 is recovered in the *fastidiosa* subspecies clade and M12 is recovered in the *multiplex* clade (Schuenzel et al. 2005; Parker et al. 2012; Nunney et al. 2014). The higher number of Griffin-1 non-overlapping hits,

113, may be an artifact of sequencing as analyzing identical host genera at the subspecies level would suggest “oak to oak” BLASTp runs with near identical results. Data from previously described population studies, however, suggests that this finding may also be indicative of yet uncovered genetic diversity in disparate *X. fastidiosa* communities (Behringer et al. pending submission).

The detailed results of the initial RNB1 reciprocal process are summarized in Figure 3.2a with the aid of a trend depicting sparkline. This aforementioned table/figure hybrid provides a data trend as well as the raw numeric representations. As expected, the initial *X. fastidiosa* subspecies comparisons mirrored a decreasing linear function, with the greatest number of query hits at or near an E-value of zero. This is consistent with overall organismal similarity. Despite a general trend toward subspecies similarity, post hoc analysis conveyed a significant number of reciprocal query hits exceeding the E-value threshold of “ $> 1e-10$ ”. To account for the possibility of gapped regions post artificial contig assembly, a second, BLASTp was run against the previously mentioned “nr” collection for those poor scoring (“ $> 1e-10$ ”) and “No hits found” results. The artificial concatemerization of contigs for gene calling gives the false impression of genomic continuity and could mistakenly predict genes as poor scoring when they may in fact be truncated due to problematic assemblies. After employing this schema, many of the poor scoring and “No hits found” segments were pushed into bins (E-values of zero, zero < E-values < $1.0e-100$, $1.0e-100 < \text{E-values} < 1.0e-10$) reflective of prior observed similarities between *X. fastidiosa* subspecies strains. The results for this process are summarized in Figure 3.2b with the aid of an additional sparkline chart for trend visualization. The data for this pipeline iteration suggests that the vast majority of the initial poor scoring segments actually belong to the “ $1.0e-100 < \text{E-values} < 1.0e-10$ ” category. Because this study is concerned with

the absolute genomic differences between RNB1 relative to M12, Griffin-1, and Temecula, those poor scoring segments that were pushed into bins suggestive of greater sequence similarity reduced the number of poor scoring RNB1 sequences in the final comparative analysis. It is likely, however, that reviewing this data could yield phylogenetically relevant loci for evolutionarily driven assays. Here, the focus is limited to absolute genetic difference of the oak isolate RNB1 relative to the considered genomes.

The resulting poor scoring and “No hits found” sequences identified through the two iterations of BLASTp runs were then subjected to Pfam-A and Gene Ontology (GO) database calling. This process, designed to add descriptors to the sequences in question, was initially performed on the following: RNB1 query/M12 database (50 sequences), RNB1 query/Griffin-1 database (53 sequences), RNB1 query/Temecula1 database (51 sequences), M12 query/RNB1 database (18 sequences), Griffin-1 query/RNB1 database (21 sequences), and Temecula1 query/RNB1 database (16 sequences). Considering the full length proteins, unique Pfam-A E-values scoring < 1.0 averaged five total hits in the six iterations of the reciprocal pipeline process. Among the same six iterations, total Pfam-A names domains, irrespective of redundancy and E-value, averaged 12 annotations. Those sequences returning no annotation averaged 30 null assignments across the same dataset. A full summary of all Pfam-A domain hits and Gene Ontology association for Remote BLASTp Queries with E-values greater than 1e-10 (E-value > 1.0e-10) and for queries producing no hits (“No hits found”) is listed in Table 3.2. Several of the best scoring recoveries hit the proteobacterial associated domains of unknown function (DUFs) DUF3470 and DUF3011 thus providing no further resolution. Two of the best scorings category returns, however, contained a Pfam-A annotation of “Trypan_PARP”, and produced a non-trivial E-value full domain score (Table 3.1). This named domain is associated with *Trypanosoma* species which also fall into a class of vectored pathogen, be it mammalian. This was not

observed in the recovered ORF from Temecula1 one because of a unique multi-valine residue substitution relative to the “Trypan_PARP” regions observed in RNB1, M12, and Griffin-1. A similar observation was made when GO assignments were tied to the results of the residual poor and “No hits found” scoring pipeline sequences. Again, five of the six poor scoring analyses returned a GO assignment of either “GO:0016020 Procyclic acidic repetitive protein (PARP)”, “GO:0007157 Cytadhesin P30/P32”, or both for the highest or second highest non-trivial E-value full Pfam-A domain score (Table 3.2). DUFs failed to call a GO annotation because of uncertainty surrounding putative functionality.

Again, the absolute ORF differences contained within RNB1 relative to M12, Griffin-1, and Temecula1 are presented in Table 3.1. As a final check regarding the validity of these ORFs two additional tasks were undertaken. First, unitig files were run through the GeneMark ORF calling algorithm for both the conservative and HMM calling procedure. This was supplemented with either recovery or non-recovery of the RNB1 poor scoring ORFs within the additional fourteen *X. fastidiosa* genomes housed at NCBI. Second, one final BLASTp with the base poor scoring RNB1 ORF was run to capture those hits that upon full alignment would be either homologous or near homologous sequence matches. The latter was performed subsequent to the prior two BLASTp pipeline iterations due to the observations that gapped amino acid deposits in NCBI caused inconsistent retrieval behavior in the BLAST algorithm. The most significant RNB1 based findings relative to complete analysis presented in Table 3.3 were the variable number tandem repeat (VNTR) patterns observed in ORF_415, ORF_1273, and ORF_2081, and the subspecies multiplex subspecies specific orfs: ORF_2081, ORF_727, and ORF_1357 (Table 3.3). The uniqueness of both the former and latter groups has potential utilization as both diagnostic and phylogenetic subspecies aides. The possibility of these markers as population trackers exists as well. This

would be especially important if novel hosts were found to express the disease phenotype, or if a survey on asymptomatic hosts (McElrone et al. 1999; Winstrom et al. 2005) was being considered.

Post pipeline implementation and secondary checking of recovered poor scoring and “No hits found” RNB1 sequences, several additional analyses were performed to punctuate the genetic character of RNB1 relative to M12, Griffin-1, and Temecula1. To first present a global view of the RNB1 isolate relative to the M12, Griffin-1, and Temecula1, a BLAST Ring Image Generator (BRIG) rendering was created Figure 3.4. Using RNB1 as the reference database, a BLASTn of the individual genomes comprised of GeneMark (non-HMM) ORFs was performed, and a global rendering was imaged where differential shading patterns correlated to allelic similarity. The greater the transparencies observed in base colors representing individual genomes relative to RNB1 corresponded with, the greater the dissimilarity at the specified location. Conversely, those regions with the darkest shading or least transparency of the base color conveyed the greatest similarity at the specified region. Succinctly, the degree of transparency as prescribed by the legend threshold insert conveys a succinct semblance of spot genome to genome similarity. Examining the BRIG results directly, the nucleotide to nucleotide similarity agrees best between those sharing the multiplex subspecies designations Griffin-1 and M12. The more phylogenetically distant subspecies *fastidiosa*, represented by Temecula1, has more numerous ring breaks and shadings. Note, this study limited the categorized transparency to three distinct classifications: 100%, 70%, and 50%. These differentially shaded markers thus present polymorphic regions, genetic gaps, or confirm homology at the nucleotide level.

Following the nucleotide presentation of the genomes, the specific, identified poor scoring and “No hits found” RNB1 ORFs are also presented in similar ring format in Figure 3.5. In this

instance, the resulting genes have been artificially concatemerized for ease of display, but follow the same logic as the initial BRIG rendering. The outermost black arc represents the full concatemerized suite of poor scoring and “No hit found” RNB1 genes relative to the inner concentric arcs which are once more representative of gene presence, absence, or variation in M12, Griffin-1, and Temecula1.

These enumerated findings show that relative to the concatemerized collection of poor scoring and “No hits found” in RNB1, both the conserved and HMM GeneMark ORF calling methods predict absences in M12 (ORF_948, ORF_1273), Griffin-1 (ORF_848, ORF_948, ORF_1273), and Temecula1 (ORF_948, ORF_2081, ORF_727, ORF_1357). Importantly, Figure 3.5 reasserts the finding that ORF_2081, ORF_727, and ORF_1357 may be *multiplex* specific due to the observed absence in Temecula1, a member of the *fastidiosa* subspecies. Re-referencing Table 3.3, the expansion of this ORF finding methodology show that the addition of fourteen more *X. fastidiosa* genomes provides stringer confirmation that ORF_2081, ORF_727, and ORF_1357 may again be *multiplex* specific.

Environmental Sample Supplementation and Multiple Correspondence Analysis (MCA):

The presence of VNTR regions in several low scoring RNB1 ORFs provided potentially informative allelic results, and the previously described primer sets specific to VTNR regions ORF_415 and ORF_2081 were used to analyze eight additional previously described environmental samples (Table 3.4). Amplicons were obtained from previously described Northeastern and Mid-Atlantic environmental samplings (Behringer et al. pending submission) which targeted shade trees expressing the associated scorch-like disease phenotype. Results herein were paired with the existing data obtained from the M12, Griffin-1, Temecula1, and fourteen additional NCBI *X. fastidiosa* genomes for overall comparison. Summarizing the data

contained within Table 3.4, the largest repeat in ORF_415 (“PELE”), at 119 amino acids, was recovered in sample 7024_NH, a Black Oak sample from Salem, New Hampshire. Extending these results to the additional NCBI *X. fastidiosa* genomes, this was the third largest repeat with isolates 9a5c (citrus host) and MUL0034 (mulberry host) superseding that number at 133 amino acids and 131 amino acids respectively. The smallest repeat lengths were observed in M12 (almond host) and Griffin-1 (Red Oak host) at 79 amino acids. The largest repeat in ORF_2081 (“POLYL”), at 69 amino acids, was recovered in sample 6654_DE, a Red Oak sample from Wilmington, Delaware. This was also the largest repeat region recovered post comparison to the extended NCBI *X. fastidiosa* collection. Within this larger group, it was found that the next largest repeat regions in ORF_2081 were also derived from oak associated populations. The samples 7024_NH, the aforementioned Black Oak sample from Salem, New Hampshire, and 1C1_NJ, a Red oak sample from Cranbury New Jersey, each contained repeats of 47 and 59 residues respectively. Supplementation with the chosen Northeastern and Mid-Atlantic environmental samples showed this putative ORF to be unique to the *multiplex* subspecies *in silico*.

Because much of the descriptive allelic data generated in this study was categorical in nature, and non-numeric, MCA was a logical choice to cluster and associate aspects of the respective datasets (Le Roux, and Rouanet 2004). The MCA rendered a polarized depiction of the initial findings that the “poly L” repeat resides only in the *multiplex* subspecies (Table 3.3, Figure 3.6a). The only ambiguous clustering categorical was “Southern_United_States”, which occurred because several subspecies lacking the ORF overlapped in geographical membership with Southern based *multiplex* isolates containing the Leucine repeat. The *pauca* subspecies appeared as an outlier based on the non-overlapping categoricals in “Subspecies” and “Geographical_Locale”. The subspecies *fastidiosa*, *morus*, and *sandyi* all clustered together

based on subspecies designation and the lack of the aforementioned poly L containing ORF.

Based on MCA support, this ORF appears to be an excellent candidate for confirmation of multiplex subspecies designation, and perhaps lineage identification.

As further verification of the *in silico* prediction that this ORF was subspecies multiplex specific, each genome was mined at the nucleotide level (BLASTn) for the presence of the characteristic leucine repeat. The *in silico* prediction proved valid as the variant ORFs contained both upstream and downstream polymorphisms, and a repeat consisting of “LP” residues (Figure 3.6b). The detected variant was consistent across all subspecies predicted to be absent the poly leucine region save *pauca*. A summary of the findings in the subspecies *fastidiosa*, *sandyi*, and *moris* are summarized in Figure 3.6b relative to the described patterns observed in the *multiplex* subspecies.

The MCA correctly predicted the genetic ambiguity of the "PELE" repeat marked in the observed similarity between geometric distances for the categorical pairs "Over_Mean" and "*multiplex*" and "Under_Mean" and "*multiplex*" (Figure 3.6c). In general the subspecies *fastidiosa* and *sandyi* clustered in the "Under_Mean" region, but several multiplex isolates were also designated as either "Over_Mean" or "Under_Mean" across hosts. This fact essentially nullifies its use for diagnostic and phylogenetic purposes if the repeat is considered in isolation. Interestingly, the repeat region is considered here in the absolute size sense only, and actually reveals polymorphic variation that could be considered informative as a strain-specific marker. Finally, the non-environmentally supplemented “QAQA” MCA correctly clustered the categoricals "*fastidiosa*", "*sandyi*", and "Under_Mean" which reflected the smaller than average "QA" repeat region for the respective subspecies (Figure 3.6D). Mixed data regarding the presence and relative length of the repeat region in the *multiplex* subspecies led to less defined

clustering, but the MCA did place categorical attributes of those isolates with the largest "QA" repeat regions closest to the "Over_Mean" designation. Those containing the largest "QA" repeats were RNB1 (36 amino acids), 6c (33 amino acids), 9a5c (29 amino acids), 32 (29 amino acids), and ATCC_35871_Plum (28 amino acids). Plum was clustered slightly away from the "Over_Mean" categorical because of overlapping memberships with isolates containing different composite characters. Although this "QA" containing ORF appears promising as a marker for either diagnostics or nucleic sequence analysis population characterization, its *in silico* absence in the closely related Red Oak isolate (Griffin-1) should be confirmed via traditional experimentation.

Discussion:

The recent appearance of *X. fastidiosa* in Southern Italy (Saponari et al. 2013; Cariddi et al. 2014) underscores the tenet that host range plasticity may be more the rule than the exception for this plant pathogen. As new epidemics akin to this recent “new world” to “old world” radiation emerge, they cannot simply be dismissed as singleton events. For this reason, this study seeks to forward the thought that bacterial leaf scorch of oak is an understudied and abundant pathogen that may threaten a greater number of hosts than just namesake shade trees. Even at the municipality level, the frequency of disease and incorrect association with abiotic stress (Gould et al. 2007) may lead to systemic titre that further exacerbates vector acquisition and transfer beyond the confines of local forests.

The suggested host promiscuity within the *multiplex* main clade may have important ramifications in the Northeastern and Mid-Atlantic United States where natural oak range is in close proximity to commercial *Vaccinium*, *Prunus*, and burgeoning *Vitis* operations

(www.nj.gov/agriculture; Atanassov et al. 2002; McCormick 1979). The idea of host shift and testing for reciprocal infectivity is not a new concept in *X. fastidiosa* research, with early work addressing the limitations of elm and sycamore strain reciprocity (Sherald et al. 1993), and several later assays looking at transmission in both mulberry and blueberry (Hernandez-Martinez et al 2006; Oliver et al. 2015). Work also exists showing the use of the Solanaceous host *Nicotiana tabacum* cv. SR1 as a model for inoculation, infectivity, and *in planta* study (Francis et al. 2008). To generalize, continuity in cross-host infection has been difficult to predict. Although the work herein does not test the possibility of expanding host infectivity of the analyzed oak strain RNB1, it seeks to show that, absent that knowledge, understanding the genetic underpinnings remains of understudied strains remains important given the growing knowledge of this problematic pathogen.

A means to further knowledge of the *X. fastidiosa* isolate RNB1 was provided by a simple, scalable, freeware-driven pipeline. This pipeline makes no claims to uniqueness, and instead provides a map to leverage existing tools and synthesize their respective output. One aspect to consider in this pipelines is the ORF calling. In general, non-HMM driven algorithms call far fewer purported genes than their HMM driven counterparts (Borodovsky and McIninch 1993). On average the HMM method in this particular pipeline called an average of 17.5% more putative gene calls. It is, therefore, plausible that the conservative approach taken in this study may have resulted in an underreporting of the true genetic dissimilarity among *X. fastidiosa* subspecies. As a precaution, part of the final analysis (Table 3.3) was to run the HMM algorithm as counterpoint to the initial non-HMM method. This was done as an added guard before declaring a putative ORF absent from the considered genomes post pipeline analysis. Additionally, all called genes would have to be subjected to actual existence via laboratory techniques. The thrust of this study, however, remains the absolute observed differences

between RNB1 and the selected strains, supplementary strains, and oak derived environmental samples on an *in silico* basis. Again it should be noted that gene absence or dissimilarity could be an artifact of the underlying sequencing assay used to produce the genome in question. For example, Griffin-1 consists of 85 contigs that are artificially concatenated to form a unitig in order to execute the initial ORF calling procedure. Unitig assembly in this case assumes that the contig gap can simply be bridged when, in reality, underlying sequence may be missing. While such a concern is valid, supplementation with genomic loci from an additional three multiplex isolates (ATCC_35871, Dixon, and Sycamore Sy-VA) represents a reasonable approach to addressing this potential deficiency.

Several poor scoring RNB1 sequences were noteworthy for their presence of VNTRs. To start, ORF_2081 ("POLYL") was called via the GeneMark and GeneMarkHMM method for only members of the multiplex subspecies. Namely, it was identified in ATCC_35871_plum, Dixon, Griffin-1, M12, sycamore Sy-VA in addition to its aforementioned discovery in RNB1. Correlation between hosts, however, was less successful as oak host bacterial amplicons ranged from a low Poly-L region of size 25 amino acids in the New Hampshire environmental sample 7043_NH to a high Poly-L region of size 69 amino acids in the Delaware environmental sample 6654_DE. The largest recovered lengths of the Poly-L regions in almond, sycamore, and plum were 42, 49, and 47 amino acid residues respectively. It is likely that larger sample size analysis among hosts may reveal ambiguities in host/size trends when expanded population genetics approaches are taken. Although literature on the topic of Poly-L repeats is sparse, it is interesting to note that one study working with human surfactant protein engineered synthetic Poly-L repeats and discovered that such repeats lowered overall surface tension and enhanced surface spreading at air-water interfaces (Takei et al. 1996). It is not unreasonable to speculate that such a molecule

could have utility in the xylem environment. Adding to this, a signal peptide region was uncovered via SMART analysis (Schultz et al. 1998) confirming secretion of the protein. Of final note is the non-HMM recovery of a variant amino acid sequence in the *fastidiosa*, *morus*, and *sandyi* subspecies of the form “VIAAL(Poly-LP)NDTHIM”. The cyclic nature of proline implies that it would dramatically effect adopted secondary confirmation relative to poly leucine, but that is based on the assumption that this called ORF represents a true protein product. Further experimentation would be needed to verify the *in silico* prediction herein. Suffice it to say, this predicted ORF is derived from genomic sequence and nonetheless still functions as a useful marker to distinguish subspecies.

Next, ORF_415 (“PELE”) was called via the GeneMark and GeneMarkHMM algorithm for all subspecies. Although several polymorphic variants existed in multiple subspecies, the appearance of a “Valine” residue in the main body of the repeat marked membership in the subspecies *fastidiosa*, *sandyi*, and *morus*. Sequencing errors present in the genomes of EB92.1, Ann-1_NZ, and ATCC_35871 precluded the use of those sequences for comparison, but the valine variation was explicitly detected in Ann-1_CP, ATCC_35879, GB514, M23, MUL0034, Mul-MD, and Temecula1. The multiplex repeat variant within this gene matched a Pfam-A annotation specific to the procyclic acidic repetitive proteins (PARP) of *Trypanosoma brucei*, another vectored organism. While it may be a coincidence that this recovery was made, the Pfam-A full length domain E-value of 1.50E-07 suggests that similarity is present. This “EP-PARP” protein present in *T. brucei* is a surface protein expressed in the procyclic phase of the organism, and is believed to play a role during its persistence in the tsetse fly mid-gut (Mehlert et al. 1998). While this concurrence between vectored pathogens is noteworthy, a complication is the confirmed presence of *X. fastidiosa* in the foregut of xylophagous vectors (Purcell and

Finlay 1979; Hill et al. 1995). The association of this protein with Tsetse fly mid-gut persistence may speak to a different role in the bacterium. This observation is meaningful because it has been shown to have direct consequences on bacterial transmission and loss of pathogenicity post molt (Hill et al. 1995; Almeida et al. 2003; Almeida et al. 2005). Thus, the role of this protein remains uncertain in the absence of supplementary assays. This cross-kingdom domain similarity is nonetheless worthy of further consideration.

The final ORF under consideration is ORF_1273 (“QAQA”). Before discussing the potential utilities of this repeat, it must be reiterated that the additional environmental oak samples are not present in the overall analysis, therefore, conclusions may be considered more speculative than those for either ORF_2081 (“POLYL”) or ORF_415 (“PELE”). The longest observed “QAQA” repeat was present in RNB1 (36 amino acid residues), and the shortest non-zero repeat was present in Dixon (14 amino acid residues). Subspecies association for this repeat is problematic, however, because the recovered ORF list shows this gene absent from Griffin-1, the Red Oak isolate. While it is a possibility that this finding could accurately represent the true genomic composition of this geographically oak associated isolate, more investigation is likely necessary to prove that a sequencing artifact is not present. Additional complications arise due to the absence of the “QAQA” repeat in the closely related sycamore (sycamore Sy-VA) strain as well. Any more definitive conclusion surrounding this ORF clearly requires additional samplings in order to more accurately assess the allelic distribution in question. Putative functionality is also unclear. A search of the perfect amino acid repeat (PAAR) portal revealed this repeat revealed both mixed kingdom association and greater Eukarya representation (Kumar et al. 2015). The largest “QAQA” repeat segments corresponding in length to those segments recovered from *X. fastidiosa* isolates were derived from genes of annotated eukaryotic function only (RNA polymerase II transcription subunit 15, Transcription elongation regulator 1, and Zinc finger

protein 384).

For the final summarization, the categorically driven MCA provides a succinct, multivariate method for handling qualitative data. Beyond its power to reduce the dimensionality of datasets to broader but often meaningful associations, it secondarily writes a localized, terse narrative that provides an intuitive interpretation of the dataset. In other words, by reading off the clustered terms and noting their position relative to other visualized groupings, associations are revealed. For example, Figure 6d shows that ORF_2081 is characteristic of the multiplex subspecies via the clustered terms “Northeast_Mid_Atlantic”, “Under_mean”, “Over_Mean”, “multiplex”, and “Yes”. In contrast, the observation that the Poly-L repeat is absent in other taxa is made apparent by the clustering of “morus”, “fastidiosa”, “sandyi”, “Western_United_States”, “Zero”, and “No”. The intermediate presentation of “Southern_United_States” captures the fact that this geographical designation houses members of multiple subspecies where the ORF in question is both present and absent depending on the host in question.

REFERENCES

- Alikhan, N. F., Petty, N. K., Zakour, N. L. B., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics*, 12(1), 402.
- Almeida, R. P., & Purcell, A. H. (2003). Transmission of *Xylella fastidiosa* to grapevines by *Homalodisca coagulata* (Hemiptera: Cicadellidae). *Journal of economic entomology*, 96(2), 264-271.
- Almeida, R. P., Blua, M. J., Lopes, J. R., & Purcell, A. H. (2005). Vector transmission of *Xylella fastidiosa*: applying fundamental knowledge to generate disease management strategies. *Annals of the Entomological Society of America*, 98(6), 775-786.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25-29.
- Atanassov, A., Shearer, P. W., Hamilton, G., & Polk, D. (2002). Development and implementation of a reduced risk peach arthropod management program in New Jersey. *Journal of economic entomology*, 95(4), 803-812.
- Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, 9(1), 75.
- Barry, T., Colleran, G., Glennon, M., Dunican, L. K., & Gannon, F. (1991). The 16s/23s ribosomal spacer region as a target for DNA probes to identify eubacteria. *Genome Research*, 1(1), 51-56.
- Bhattacharyya, A., Stilwagen, S., Reznik, G., Feil, H., Feil, W. S., Anderson, I., ... & Predki, P. F. (2002). Draft sequencing and comparative genomics of *Xylella fastidiosa* strains reveal novel biological insights. *Genome research*, 12(10), 1556-1563.
- Borodovsky, M., & McIninch, J. (1993). GENMARK: parallel gene recognition for both DNA strands. *Computers & chemistry*, 17(2), 123-133.

- Cariddi, C., Saponari, M., Boscia, D., De Stradis, A., Loconsole, G., Nigro, F., ... & Martelli, G. P. (2014). Isolation of a *Xylella fastidiosa* strain infecting olive and oleander in Apulia, Italy. *Journal of Plant Pathology*, 96(3), 1-5.
- 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(17), 4534-4534.
- Chen, J., Lamikanra, O., Chang, C. J., & Hopkins, D. L. (1995). Randomly amplified polymorphic DNA analysis of *Xylella fastidiosa* Pierce's disease and oak leaf scorch pathotypes. *Applied and environmental microbiology*, 61(5), 1688-1690.
- Chen, J., Chang, C. J., Jarret, R. L., & Gawel, N. (1992). Genetic variation among *Xylella fastidiosa* strains. *Phytopathology*, 82(9), 973-977.
- 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 announcements*, 1(5), e00756-13.
- Davis, M. J., Thomson, S. V., & Purcell, A. H. (1980). Etiological role of a xylem-limited bacterium causing Pierce's disease in almond leaf scorch. *Phytopathology*, 70(472), 5.
- Davis, M. J., French, W. J., & Schaad, N. W. (1981). Axenic culture of the bacteria associated with phony disease of peach and plum leaf scald. *Current Microbiology*, 6(5), 309-314.
- Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic acids research*, 27(23), 4636-4641.
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., ... & Punta, M. (2013). Pfam: the protein families database. *Nucleic acids research*, gkt1223.
- Francis, M., Civerolo, E. L., & Bruening, G. (2008). Improved bioassay of *Xylella fastidiosa* using *Nicotiana tabacum* cultivar SR1. *Plant Disease*, 92(1), 14-20.
- Giampetruzzi, A., Chiumenti, M., Saponari, M., Donvito, G., Italiano, A., Loconsole, G., ... & Saldarelli, P. (2015). Draft genome sequence of the *Xylella fastidiosa* CoDiRO strain. *Genome announcements*, 3(1), e01538-14.
- Gould, A. B., & Lashomb, J. H. (2007). Bacterial leaf scorch (BLS) of shade trees. *The Plant Health Instructor*.
- Grebus, M. E., Henry, J. M., Hartin, J. E., & Wilen, C. A. (1996). Bacterial leaf scorch of oleander: a new disease in southern California. *Phytopathology*, 86, 110.
- Hearon, S. S., Sherald, J. L., & Kostka, S. J. (1980). Association of xylem-limited bacteria with elm, sycamore, and oak leaf scorch. *Canadian Journal of Botany*, 58(18), 1986-1993.
- Hendson, M., Purcell, A. H., Chen, D., Smart, C., Guilhabert, M., & Kirkpatrick, B. (2001). Genetic diversity of Pierce's disease strains and other pathotypes of *Xylella fastidiosa*. *Applied and Environmental Microbiology*, 67(2), 895-903.
- Hernandez-Martinez, R., Pinckard, T. R., Costa, H. S., Cooksey, D. A., & Wong, F. P. (2006). Discovery and characterization of *Xylella fastidiosa* strains in southern California causing mulberry leaf scorch. *Plant*

disease, 90(9), 1143-1149.

Hill, B. L., & Purcell, A. H. (1995). Acquisition and retention of *Xylella fastidiosa* by an efficient vector, *Graphocephala atropunctata*. *Phytopathology*, 85(2), 209-212.

HMMER 3.1b1 (May 2013); <http://hmmer.org/>
 hmmscan :: search sequence(s) against a profile database
 Copyright (C) 2013 Howard Hughes Medical Institute.
 Freely distributed under the GNU General Public License (GPLv3).

Hopkins, D. L., & Purcell, A. H. (2002). *Xylella fastidiosa*: cause of Pierce's disease of grapevine and other emergent diseases. *Plant disease*, 86(10), 1056-1066.

Kumar, H., Srivastava, S., & Varadwaj, P. K. (2015). DPAAR: a Database of Perfect Amino Acid Repeat. *International Journal for Computational Biology (IJC)*, 4(1), 62-66.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software*, 25(1), 1-18.

Le Roux, B., & Rouanet, H. (2004). *Geometric data analysis: from correspondence analysis to structured data analysis*. Springer Science & Business Media.

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark. hmm: new solutions for gene finding. *Nucleic acids research*, 26(4), 1107-1115.

McCormick, J. (1979). The vegetation of the New Jersey pine barrens. *Pine Barrens: ecosystem and landscape*, 229-243.

McElrone, A. J., Sherald, J. L., & Pooler, M. R. (1999). Identification of alternative hosts of *Xylella fastidiosa* in the Washington, DC, area using nested polymerase chain reaction (PCR). *Journal of Arboriculture*, 25, 258-263.

Mehlert, A., Zitzmann, N., Richardson, J. M., Treumann, A., & Ferguson, M. A. (1998). The glycosylation of the variant surface glycoproteins and procyclic acidic repetitive proteins of *Trypanosoma brucei*. *Molecular and biochemical parasitology*, 91(1), 145-152.

Nunney, L., Schuenzel, E. L., Scally, M., Bromley, R. E., & Stouthamer, R. (2014). Large-scale intersubspecific recombination in the plant-pathogenic bacterium *Xylella fastidiosa* is associated with the host shift to mulberry. *Applied and environmental microbiology*, 80(10), 3025-3033.

Oliver, J., Cobine, P., & 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*, (ja).

Oliveros, J. C. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams. <http://bioinfo.gp.cnb.csic.es/tools/venny/index.html>.

Parker, J. K., Havird, J. C., & De La Fuente, L. (2012). Differentiation of *Xylella fastidiosa* strains via multilocus sequence analysis of environmentally mediated genes (MLSA-E). *Applied and environmental microbiology*, 78(5), 1385-1396.

Pierce, N. B. (1892). The California vine disease: a preliminary report of investigations (No. 2). US Government Printing Office.

Purcell, A. H., & Finlay, A. H. (1979). Evidence for noncirculative transmission of Pierce's disease bacterium by sharpshooter leafhoppers. *Phytopathology*, 69(4), 393-395.

Purcell, A., & Saunders, S. (1999). Glassy-winged sharpshooters expected to increase plant disease. *California Agriculture*, 53(2), 26-27.

Rodrigues, C. M., Takita, M. A., Coletta-Filho, H. D., Olivato, J. C., Caserta, R., Machado, M. A., & De Souza, A. A. (2008). Copper resistance of biofilm cells of the plant pathogen *Xylella fastidiosa*. *Applied microbiology and biotechnology*, 77(5), 1145-1157.

Rodrigues, J. L., Silva-Stenico, M. E., Gomes, J. E., Lopes, J. R. S., & Tsai, S. M. (2003). Detection and diversity assessment of *Xylella fastidiosa* in field-collected plant and insect samples by using 16S rRNA and *gyrB* sequences. *Applied and environmental microbiology*, 69(7), 4249-4255.

Saponari, M., Boscia, D., Nigro, F., & Martelli, G. P. (2013). Identification of DNA sequences related to *Xylella fastidiosa* in oleander, almond and olive trees exhibiting leaf scorch symptoms in Apulia (Southern Italy). *Journal of Plant Pathology*, 95(3).

Schaad, N. W., Postnikova, E., Lacy, G., Fatmi, M. B., & Chang, C. J. (2004). *Xylella fastidiosa* subspecies: *X. fastidiosa* subsp. *piercei*, subsp. nov., *X. fastidiosa* subsp. *multiplex* subsp. nov., and *X. fastidiosa* subsp. *pauca* subsp. nov. *Systematic and applied microbiology*, 27(3), 290-300.

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*. *Applied and environmental microbiology*, 71(7), 3832-3839.

Schultz, J., Milpetz, F., Bork, P., & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11), 5857-5864.

Sherald, J. L. (1993). Pathogenicity of *Xylella fastidiosa* in American elm and failure of reciprocal transmission between strains from elm and sycamore. *Plant Disease*, 77(2), 190-193.

Simpson, A. J. G., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., ... & Krieger, J. E. (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, 406(6792), 151-157.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., ... & Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, 12(10), 1611-1618.

Takei, T., Hashimoto, Y., Ohtsubo, E., Sakai, K., & Ohkawa, H. (1996). Characterization of poly-leucine substituted analogues of the human surfactant protein SP-C. *Biological & pharmaceutical bulletin*, 19(12), 1550-1555.

Team, R. C. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Thompson, J. R., Carpenter, D. N., Cogbill, C. V., & Foster, D. R. (2013). Four centuries of change in northeastern United States forests. *PloS one*, 8(9), e72540.

Tufte, E. (2004). Sparklines: Intense, simple, word-sized graphics. *Beautiful Evidence*, 1, 46-63.

Van Sluys, M. A., De Oliveira, M. C., Monteiro-Vitorello, C. B., Miyaki, C. Y., Furlan, L. R., Camargo, L. E. A.,

... & Truffi, D. (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. *Journal of Bacteriology*, 185(3), 1018-1026.

WEB: www.nj.gov/agriculture

Wells, J. M., Raju, B. C., Hung, H. Y., Weisburg, W. G., Mandelco-Paul, 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(2), 136-143.

Wickham, H. (2011). *ggplot2*. Wiley Interdisciplinary Reviews: Computational Statistics, 3(2), 180-185.

Wistrom, C., & Purcell, A. H. (2005). The fate of *Xylella fastidiosa* in vineyard weeds and other alternate hosts in California. *Plant Disease*, 89(9), 994-999.

Figure 3.1. Venn Diagram depicting RNB1 database BLASTp hits resulting from queries derived from putative open reading frames (ORFs) in the *X. fastidiosa* genomes M12, Griffin, and Temecula1.

The Venn categorized overlapping hits herein consist of matching database hits in common between the respective query sequences from M12, Griffin-1, and Temecula1 in each of the three unique runs. A large number of overlapping genes (1643) are observed between the subspecies, and the unique resultants correspond to previously observed taxonomic differences in host associated *X. fastidiosa* populations.

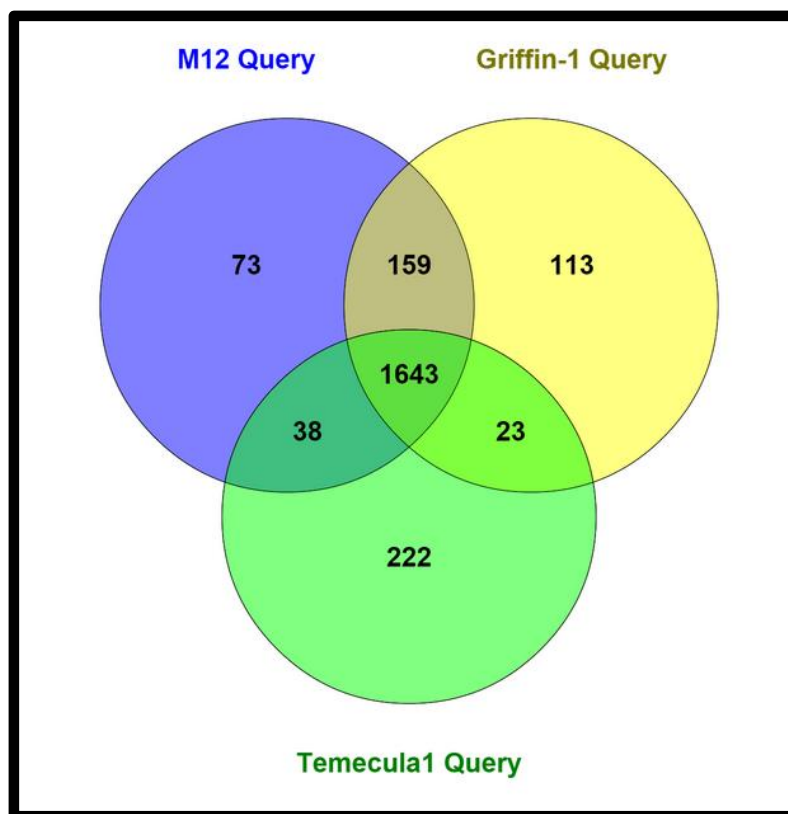


Figure 2. Raw numerical counts of E-Value categorizations for reciprocal BLASTp queries derived from putative open reading frames (ORFs) for all RNB1 centered query/database genome combinations. A Sparkline chart is supplied to quickly assess the E-value "binning" trends between these reciprocal BLASTp runs.

Note that a Sparkline chart is a simple multipoint trend line where the magnitude of each point corresponds proportionally to a vertical height. The chart is drawn without axes or coordinates and presents data trends pictorially.

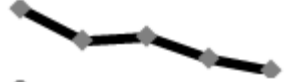
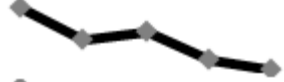
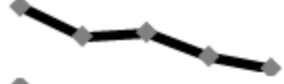



BLASTp Queries	E-values of zero	0 < E-values < 1.0e-100	1.0e-100 < E-values < 1.0e-10	E-values > 1.0e-10	"No hits found"	Sparkline Depiction
RNB1 Query Versus M12 Database	1122	534	603	226	2	
RNB1 Query Versus Temecula1 Database	1083	530	672	197	5	
RNB1 Query Versus Griffin1 Database	1083	548	627	229	0	
M12 Query Versus RNB1 Database	1046	465	364	38	0	
Temecula1 Query Versus RNB1 Database	1008	456	376	86	1	
Griffin1 Query Versus RNB1 Database	1002	485	408	43	2	

Figure 3. Raw numerical counts of E-Value categorizations for remote (NCBI database "nr") BLASTp queries derived from initial poor scoring segments and "No hits found" sequences in reciprocal BLASTp runs.

A spark chart is supplied to quickly assess the E-value "binning" trends between these remote (NCBI) BLASTp runs. Again, note that a Sparkline chart is a simple multipoint trend line where the magnitude of each point corresponds proportionally to a vertical height. The chart is drawn without axes or coordinates and presents data trends pictorially.







BLASTp Queries	E-values of zero	0 < E-values < 1.0e-100	1.0e-100 < E-values < 1.0e-10	E-values > 1.0e-10	"No hits found"	Sparkline Depiction
RNB1 Query Versus M12 Database	26	26	126	20	30	
RNB1 Query Versus Temecula1 Database	28	28	95	26	25	
RNB1 Query Versus Griffin1 Database	27	27	122	25	28	
M12 Query Versus RNB1 Database	0	1	19	9	9	
Temecula1 Query Versus RNB1 Database	10	11	50	8	8	
Griffin1 Query Versus RNB1 Database	0	5	19	10	11	

Table 3.1. Polymerase Chain Reaction (PCR) primer list used for verification of putative *X. fastidiosa* colonies and exploration of poor scoring RNB1 genomic regions

Genomic region / Locus	Putative functionality	Forward sequence (5'->3')	Reverse Sequence (5'->3')	Source
16s/23s ribosomal spacer region (ITS)	Non coding	AAC AAG GTA GCC GTA TCG GAA GGT	GTG TGC GCT TAT TCG CTT GAC CAT	Modified from Barry et al. 1991
<i>copB</i>	copper sequestration / divalent copper tolerance	ATG AAC ACC CGT ACC TGG TTC GTA	ATT TAG TCT CCA CCA TGA GCC GCA	Rodrigues et al. 2008
<i>holC</i>	DNA replication	GAT TTC CAA ACC GCG CTT TC	TCA TGT GCA GGC CGC GTC TCT	Schuentzel et al. 2005
<i>lacF</i>	sugar transport	TTG CTG GTC CTG CGG TGT TG	CCT CGG GTC ATC ACA TAA GGC	Schuentzel et al. 2005
putative ORF "Poly L" (ORF_2081)	Unknown	GAG ACA CGA GCA CAG CAC ATA G	CCT TAG CGG CAT ACT TTC AGA G	This Study
putative ORF "PELE" (ORF_415)	Unknown	GTC GCA CTC CAT AGG GTC TG	CAG GAT GCA GGG ATA GGT TTA	This Study

Table 3.2. Pfam-A domain hits and Gene Ontology association for Remote BLASTp Queries with E-values greater than 1e-10 (E-value > 1.0e-10) for queries producing no hits ("No hits found")

Reciprocal Run Category	Unique Pfam-A complete E- value Proteins scoring < 1.0	total named domains Pfam-A database	"No Pfam-A matches"	Best Scoring Pfam-A full sequence E-value	Gene Ontology (GO) Associations
RNB1 Query Versus M12 Database	6	19	44	Trypan_PARP PF05887.6 / 1.1e-05	GO:0016020 / Procylic acidic repetitive protein (PARP)
RNB1 Query Versus Temecula1 Database	9	23	42	DUF3470 PF11953.3 / 3.5e- 09	None / DUF Association
RNB1 Query Versus Griffin1 Database	7	21	46	Trypan_PARP PF05887.6 / 1.1e-05	GO:0016020 / Procylic acidic repetitive protein (PARP)
M12 Query Versus RNB1 Database	2	3	16	DUF3011 PF11218.3 / 1.3E- 36*	None / DUF Association**
Temecula1 Query Versus RNB1 Database	2	2	14	DUF3011 PF11218.3 / 1.3E- 36*	None / DUF Association**
Griffin1 Query Versus RNB1 Database	4	6	17	DUF3011 PF11218.3 / 1.3E- 36*	None / DUF Association**

* Second best scoring Pfam-A full sequence domain was

Trypan_PARP PF05887.6 / 1.50E-07

**Trypan_PARP PF05887.6 GO

Associations:

GO:0016020 Procylic acidic repetitive
protein (PARP)

GO:0007157 Cytadhesin P30/P32

GO:0009405 Cytadhesin P30/P32

				M0000000 0.4, Ann- 1_NZ_CP00 6696.1, EB92.1, M23, Mul- MD, Temecula1		
VSSGCKGEFELGAEDGSGFVDMPWRLICESKKR QRMSCGTSVQHEVSVFLQLSTTPCEKDRNWG WDADRI WVDGGCRAEFLVY	ORF_120	None†	32, ATCC_35871_plum, ATCC_35879_FL_grape, 6c, 9a5c, Ann- 1_NZ_AAAM00000000.4, Ann-1_NZ_CP006696.1, GB514, M23, MUL_0034, Mul-MD	Dixon, EB92.1, Griffin-1, M12, sycamore Sy-VA, Temecula1	N/A	
VIAVLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL LPNNVSTALKRDDDTRIM	ORF_2081	None	ATCC_35871_plum, Dixon, Griffin-1, sycamore Sy- VA††	M12††	N/A	
VARSWRTYLASLASSETRQTLLQLQQQMSALQ AALDSAHSAPPSGTASSHRLAQHGRHIS	ORF_727	None	ATCC_35871_plum, Dixon, Griffin-1, M12, sycamore Sy-VA††	No Additional	N/A	
VQRNANPPHEDTTMPAPKRASTDVHREPFH DVSEALFMENFSAHGKKPEDSLLASCYDIRSNA VQQCM DLVNSVRRVYANPTLNSVQQDIEGRQAAEAPH DRAHQRLRCFAPARAGAKG	ORF_1357	Yes	ATCC_35871_plum, Dixon, Griffin-1, M12, sycamore Sy-VA††	No Additional	N/A	
VADIEEWKRRKIAAEAQREELHLA	ORF_2414	None†††	All	No Additional	N/A	

*RNB1 partial residue sequence of conjugal transfer protein *traL*

**Repeat region noted as highly variable across all subspecies

*** unusable sequence present in ATCC_35871_plum, Ann-1_NZ_AAAM00000000.4, and EB92.1

†Partial of full DUF3011

††ORF and variants appear indicative of multiplex subspecies designation

†††Partial of full Phage_Nu1

Table 3.4. Supplemental Northeastern / Mid-Atlantic environmental samples for Variable Length Tandem Repeat (VNTR) analysis of ORF_415 "PELE" / ORF_2081 "POLYL"

Environmental Sample Name	Collection Location	Host	Collection Date	Recovered Sequence (ORF_415 "PELE" / ORF_2081 "POLYL")	Repeat Length *
7043_NH	Salem, New Hampshire	<i>Quercus velutina</i> (Black Oak)	2008	VILEVIGFLVGSEAPGLKPEPGLKPEPELEPGLEPGLEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPEL EPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELGSGVRLISG	99 amino acids
7024_NH	Salem, New Hampshire	<i>Quercus velutina</i> (Black Oak)	2008	VIAVLNNTALKRDDDTRIM VILEVIGFLVGSEAPGLKPEPGLKPEPELEPGLEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPEL ELEPELGS GVRLISG	25 amino acids
NB4_NJ	New Brunswick, New Jersey	<i>Quercus palustris</i> (Pin Oak)	2010	VIAVLNNTALKRDDDTRIM VILEVIGFLVGSEAPGLKPEPGLKPEPELEPGLEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPEL ELEPELGS GVRLISG	119 amino acids
2C1_NJ	Cranbury, New Jersey	<i>Quercus rubra</i> (Red Oak)	2010	VIAVLNNTALKRDDDTRIM VILEVIGFLVGSEAPGLKPEPGLKPEPELEPGLEPGLEPELEPELEPELEPELEPELEPELEPELEPELEPELEPELEPEL ELEPELGS GVRLISG	47 amino acids
				VIAVLNNTALKRDDDTRIM	97 amino acids
				VIAVLNNTALKRDDDTRIM	29 amino acids
				VIAVLNNTALKRDDDTRIM	91 amino acids
				VIAVLNNTALKRDDDTRIM	33 amino acids

[illegible]

Figure 4. Global genomic snapshot of absolute similarities in M12, Griffin-1, and Temecula1 called open reading frames (ORFs) relative to RNB1 ORFs.

Each concentric ring represents a corresponding genome and the degree of transparency at any individual point in that genome depicts overall similarity of the respective genome to RNB1.

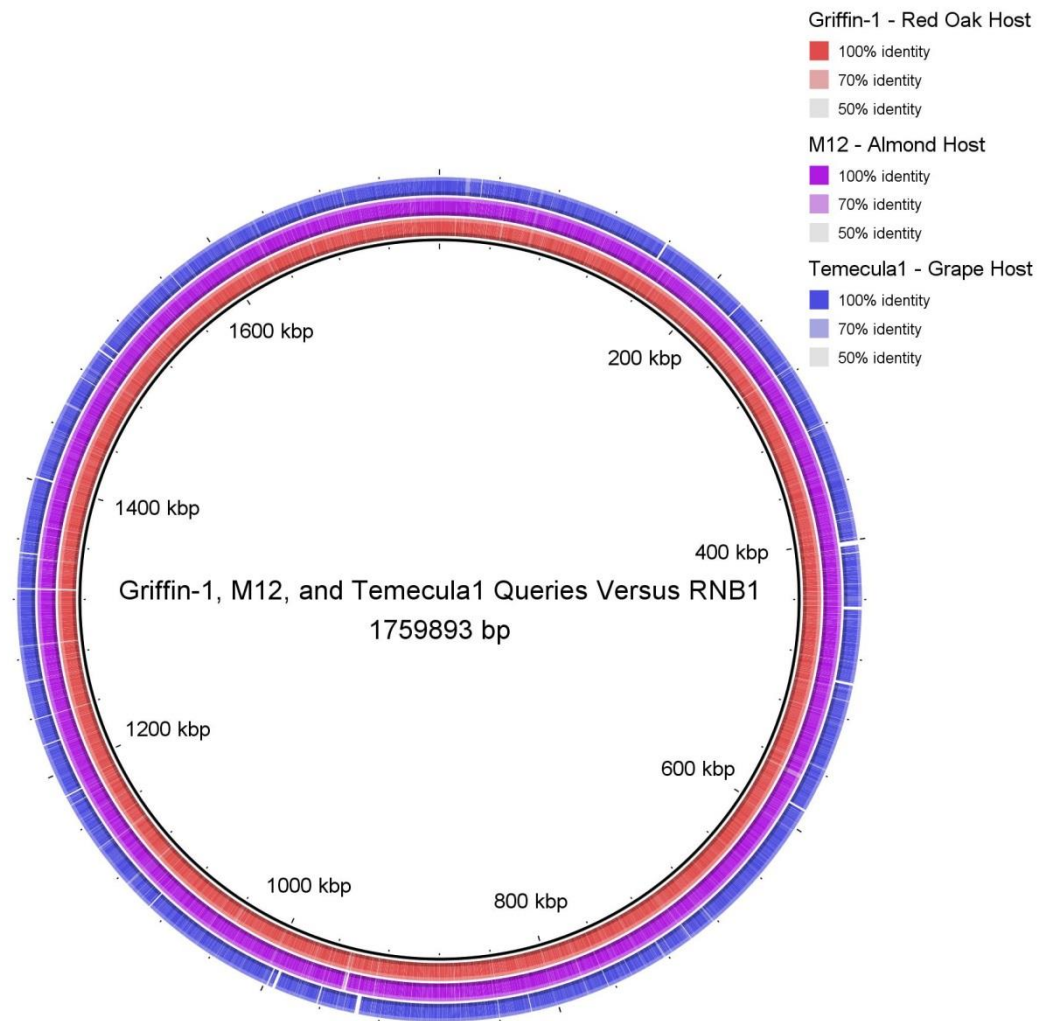
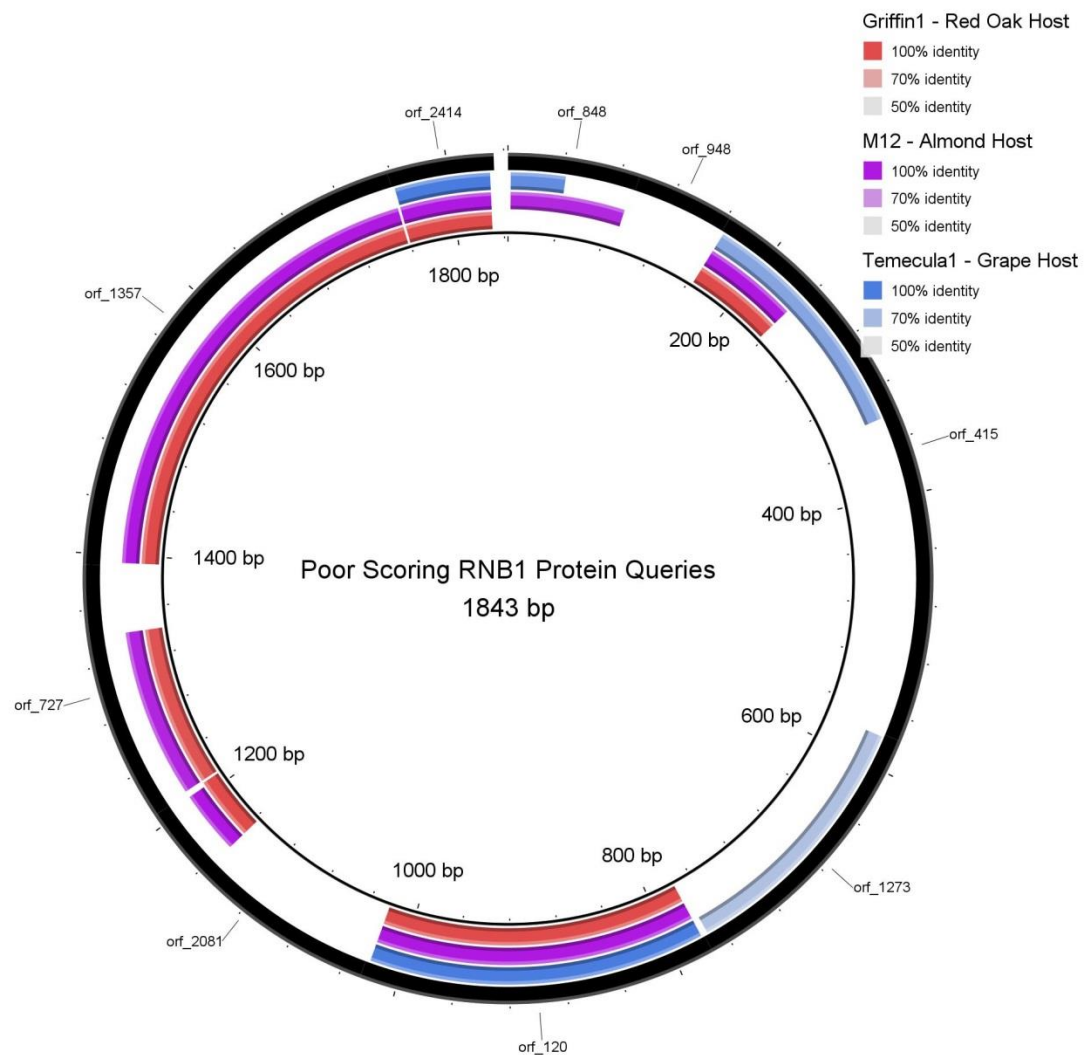


Figure 5. Pipeline derived poor scoring RNB1 protein queries relative to multiple technique recovered sequences in the respective genome locations present in M12, Griffin-1, and Temecula1. The outer black arc represents those sequences present in RNB1 relative to those recovered by various methods in M12, Griffin-1, and Temecula1. Note that these poor scoring RNB1 putative genes represent recoveries from disparate parts of the entire *X. fastidiosa* genome and have been concatemerized for the sake of arc continuity. The degree of transparency depicts genetic overlaps of identified poor scoring queries in RNB1 relative to that respective genome. Blank or white segments relative to the outer black arc depicts those segments deemed "missing" per this method of genome comparison. Exact coordinates of this concatemerized representation are provided adjacent to Figure 5.



Position	Presence / Absence	Presence / Absence	Presence / Absence
RNB1 Positioning	M12	Griffin-1	Temecula1
ORF_848: 1-93	ORF_848	ABSENT	ORF_848
ORF_948: 94-162	ABSENT	ABSENT	ABSENT
ORF_415: 163-576	ORF_415	ORF_415	ORF_415
ORF_1273: 577-774	ABSENT	ABSENT	ORF_1273
ORF_120: 775-1026	ORF_120	ORF_120	ORF_120
ORF_2081: 1027-1209	ORF_2081	ORF_2081	ABSENT
ORF_727: 1210-1392	ORF_727	ORF_727	ABSENT
ORF_1357: 1393-1758	ORF_1357	ORF_1357	ABSENT
ORF_2414: 1759-1833	ORF_2414	ORF_2414	ORF_2414

Figure 6. Multiple Correspondence Analysis (MCA) of categorical data for variable length tandem repeat (VNTR) containing poor scoring RNB1 ORFs: ORF_2081 (Poly L), ORF_415 (PELE), and ORF_1273 (QAQA).

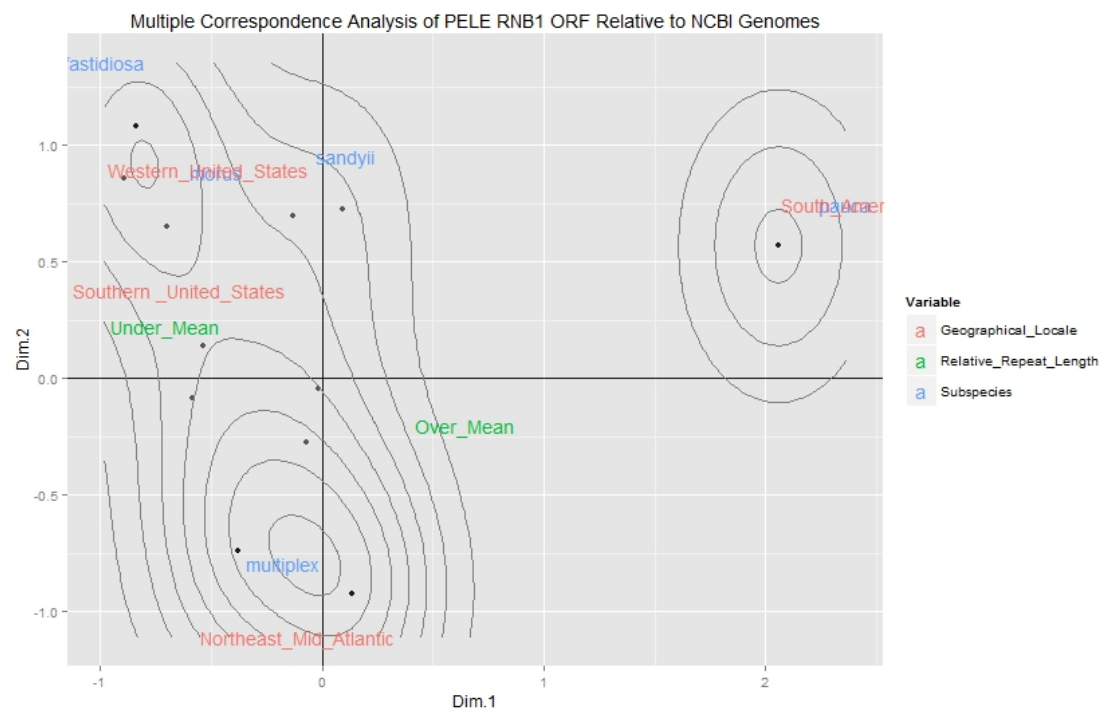
The clustering pattern of the various categorical responses under the provided "Variable" heading, namely "Geographical_Locale", "Relative_Repeat_Length", "Repeat_Presence", and "Subspecies" provide a composite qualitative description of each sample's similarity to another. Categories containing categorical responses are listed to the right of each figure under the "variable" heading. A table is provided listing the categorical responses assigned after surveying each subspecies member. Dimensionalized, clustered nominals signify similarity. Additionally, an alignment of the ORF_2081 variant versus two representative oak associated populations (B). The variant is profiled in representative isolates from the subspecies *fastidiosa*, *morus*, and *sandyi*.

Variable	Responses
Geographical_Locale	Western_United_States, Southern_United_States, Northeast_Mid_Atlantic
Relative_Repeat_Length	Zero, Under_Mean, Over_Mean
Repeat_Presence	Yes, No
Subspecies	<i>fastidiosa</i> , <i>sandyi</i> , <i>morus</i> , <i>multiplex</i> , <i>pauca</i>

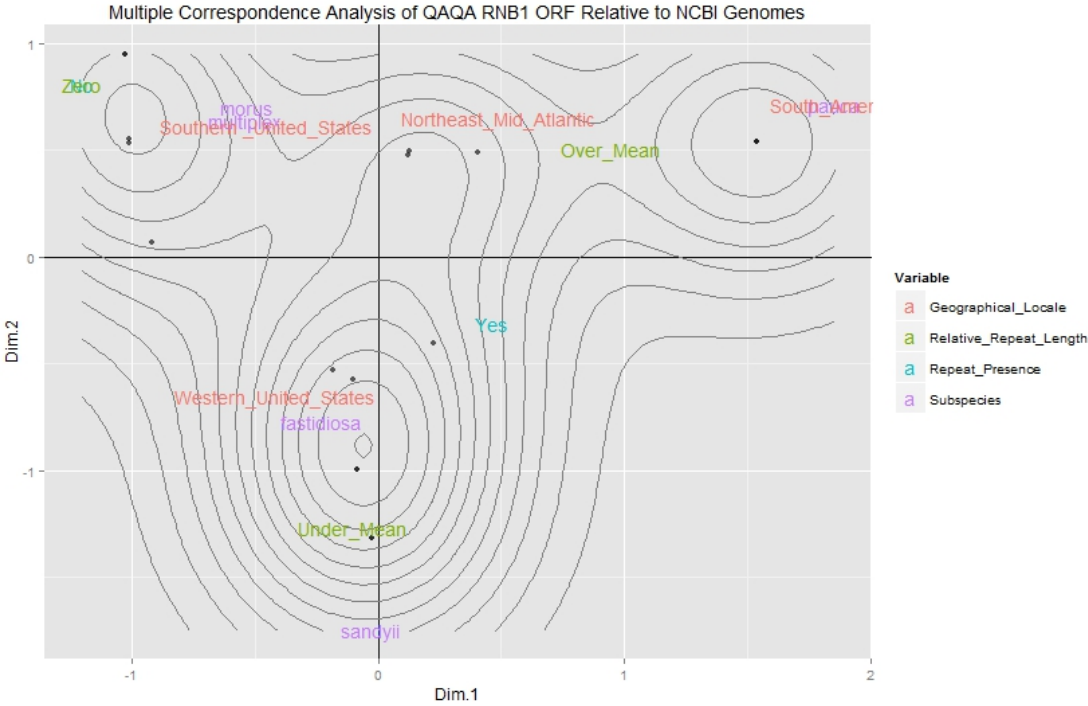
A

[illegible]

B



C



D

CHAPTER 4. The Exploration and Application of Cross-Subspecies *Xylella fastidiosa* Prophage

Regions

ABSTRACT

Since the completion of the initial *Xylella fastidiosa* genome projects in early part of the millennium, a key observation has been the contribution of prophage regions as significant drivers of intrasubspecific genetic diversity within this pathogen. Although several smaller scale studies exist regarding descriptive aspects of prophage regions within *X. fastidiosa* genomes, this study supplements and expands prior findings by profiling unique prophage regions across all currently available, NCBI based *X. fastidiosa* genomes. In this analysis, meaningful diversity among various prophage regions was uncovered in both isolate based haplotype profiling and at cross host phage loci. This was especially apparent when the recently described European based olive isolate (CoDiRo) was considered relative to those derived from Western hemisphere based hosts. In addition to shedding light on the prophage deposits across the existing body of *X. fastidiosa* isolates, this project also used these variable regions as data for the application of machine learning (ml). Methods described herein supplemented qualitative aspects of prophage regions and posited a method for the classification and categorization of prokaryotic genomic data through the lens of several predictive machine learning classifiers: Naïve Bayes, Logistic Regression, and Sequential Minimal Optimization (SMO). Results suggested both Logistic Regression and SMO as highly effective taxonomic predictors when trained with *X. fastidiosa* prophage data. As the "omics" age continues to generate large datasets, the ml specific approaches described in this study promise to synthesize large amounts of data and provide useful categorizations in the field of phytopathology and beyond.

Since the advent of the second millennium, *Xylella fastidiosa* (Wells et al. 1987) has become one of the most sequenced bacterial organisms in the field of phytopathology. This prolific number of sequencing projects is wholly justified given the import of both the threat posed by infection and the recently increased spectrum of previously defined host associated ranges. Specifically, GenBank (NCBI) currently houses eighteen genomes from both disparate hosts and physical locales. Beginning with the first fully sequenced citrus isolate in 2000 (9a5c) (Simpson et al. 2000), successive annotated genomic projects quickly established that a major point of subspecies variation could be found in the prophage regions unique to individual host associated populations (Bhattacharyya et al. 2002). This explicit observation first became apparent after the sequencing of both the first grape isolate (Temecula1) and the first oleander (Ann-1) and almond (Dixon) isolates (Van Sluys et al. 2003; Bhattacharyya et al. 2002). Subsequent analyses of these prophage regions after the appearance of additional *X. fastidiosa* strains confirmed this to be a prevailing theme within *X. fastidiosa* genomes (de Mello Varani et al. 2008). In addition to these lysogenic remnants, a genomic sequence for a specific plaque forming phage (Xfas53) has also been sequenced (Summer et al. 2010). Initial cross-host population analyses have confirmed *Podoviridae* (Xfas53) association within integrase and polymerase prophage segments (de Mello Varani et al. 2008), but analysis across all existing

genomes also revealed that many prophage regions are associated with several other viral families. Expanded genomic analysis verified both previously reported and unreported taxonomic representation from both the *Myoviridae* and *Siphoviridae* families (Summer et al. 2010; Chen and Civerello 2008; de Mello Varani et al. 2008) as well as undescribed association with *Podoviridae*.

The presence of suggested phage diversity within *X. fastidiosa* genomes lends itself to the exploration of the complexity of lysogenic phage existing in the ever expanding array of sequenced *X. fastidiosa* isolates. Although the argument could be made that these foreign insertions can promote general genomic decay and the possibility of greater than expected level of recombination, it is plausible to state that the opposite could also be true. Consistent recoveries of prophage regions that are congruent with the body of phylogenetic work regarding *X. fastidiosa* (Nunney et al. 2014; Parker et al 2012; Hernandez-Martinez et al. 2006; Schaad et al. 2004) may represent near permanent lysogenic regions capable of describing important aspects of *X. fastidiosa* biological phenomena.

Here we explore nine distinct prophage regions across eighteen *X. fastidiosa* genomes and one additional *Quercus palustris* (Pin Oak) genome to describe genetic continuity between bacterial hosts and viral parasites given the current state of *X. fastidiosa* genomic knowledge. Post qualitative description, machine learning approaches were applied to the data. Because machine learning takes dataset patterns and attempts to find substantive associations in the datasets (Langley and Simon 1995), it proved a logical choice for *X. fastidiosa* based prophage region analysis. Machine learning, therefore, possesses utility when analysis of large datasets is required. In short, the objective of this paper is to mine all currently existing *X. fastidiosa* genomes for prophage regions and determine the extent to which they may collectively shed continued light on subspecies similarities, dissimilarities, and population radiation. This has

become an increasing vital undertaking given the recent Old World emergence of *X. fastidiosa* in Italy on olive hosts (Cariddi et al. 2014; Saponari et al. 2013), and its appearance on pear and grape hosts in Taiwan (Leu et al. 1993; Su et al. 2014). This study supplements existing bodies of knowledge regarding earlier *X. fastidiosa* prophage descriptions and incorporates several novel machine learning based approaches to describe relationships between these genomic observations. An approach such as this expands current *X. fastidiosa* knowledge and provides additional tools from the epidemiological standpoint to track aspects of this pathogen and potential protean host movement. It may also serve to forward a greater understanding of the parasitism within *X. fastidiosa* populations hinting at future virion driven biocontrol approaches.

Material and Methods:

Genomic mining, annotation, and prophage region selection:

All genomic eighteen genomic sequences comprising this study were imported from GenBank (NCBI) in contig form and artificially concatemerized (<http://www.ncbi.nlm.nih.gov/genome/?term=xylella>). To ensure continuity in gene calling across sequencing projects, the Genemark HMM algorithm (Lukashin et al. 1998) was chosen for each individual isolate. Selection of the Genemark HMM calling procedure was done to procure a larger putative gene set relative to more conservative gene calling algorithms. One additional genome sequence representing a *Quercus palustris* (Pin Oak) associated isolate (RNB1) was also included and subjected to the same regiment.

Three annotated genomes (Temecula1, 9a5c, and M12) comprising three of the five major subspecies divisions were chosen as a base point for prophage annotation. Whole genomes of the closest lineage to one of the three annotated archetypes were selected as BLASTn (Altschul et al. 1990) queries against the known annotated genome, and all resulting hits comprising E values of $<1e-05$ were procured for inclusion. Omissions among subspecies of consistently

recovered prophage segments were secondarily checked at the protein level via recovered prophage segments as queries against protein databases comprised of individual non-reporting isolates.

A large number of prophage regions were recovered from each of the examined genomes, but were further filtered with additional criteria. First, sequences were selected for representation across all eighteen GenBank plus RNB1genomes to insure a maximum degree of continuity between analyses. Contigs for all sequencing projects used in this study are available at the URL: “<http://www.ncbi.nlm.nih.gov/genome/?term=xylella>”. Several exceptions were made where major subspecies appeared to be lacking a largely consistent prophage region or contained remnant regions too abbreviated for adequate domain calling. This was the case among the oleander sequencing projects analyzed in this study for the repressor protein region in the Ann1AAA/Ann1CP (INSDC: AAAM00000000.4; CP006696.1). Due to the epidemiological importance of a transoceanic radiation event, one additional exception was made in order to profile the recently described olive associated pathogen (CoDiRO). A second criterion was employed to limit the downstream effect of marginally alignable and abbreviated viral sequence so as not to unduly influence subsequent machine learning analyses. A summary for inclusion and exclusion counts is contained in Figure 1. This approach resulted in two additional subdivisions within the integrated baseplate and tailprotein segments. Explicitly, baseplate regions with prevailing approximate sizes 324, 588, and 894 bps, and tailprotein regions with prevailing approximate sizes of 246, 570, and 1083 bps were observed. Finally, haplotype determination for additional downstream prophage categorization was performed in GenAIEx 6.501 (Peakall and Smouse, 2012).

Alignment, matrix creation, and machine learning analyses:

Selected prophage regions were aligned via MAFFT version 7 (Katoh et al. 2013) with specifications: Strategy: FFT-NS-i (Slow; iterative refinement method), and Parameters: 1PAM / k=2, gap opening penalty = 1.53, Offset value = 0.0. All resulting polymorphic sites were extracted and converted to a simple dissimilarity matrix without the benefit of transition/transversion weighting. A differential gap weighting was employed. Both discriminative (Logistic regression)(Le Cessie, S & Van Houwelingen 1992) and Generative (Naïve Bayes)(John et al. 1995) machine learning classifiers were employed as well as a Sequential Minimum Optimization classifier (Platt 1999) via the Waikato Environment for Knowledge Analysis (WEKA) Version 3.6.12 (Hall et al. 2009). The respective classifiers were run with tenfold cross validation for prediction of two nominal classes: subspecies and geography. Cross validation was chosen over the percentage split method to not only reduce variance measures but to also aid in training with underrepresented populations. Because the sample size was limited to the collection of isolates currently housed at GenBank, the rebalancing tool Synthetic Minority Over-sampling Technique (SMOTE)(Chawla et al. 2002) was additionally employed to determine if simulated sample corrections could more accurately predict either of the two nominal categories under consideration. Because computational demands for logistic fold building caused excessive delays in obtaining results, a subset of representative attributes was chosen via the Weka provided “Select attributes” tab: FilterAttributeEval / Ranker. All non-zero rank attributes were then included for each of the two classifiers. Finally, a kappa statistic was computed for each accompanying confusion matrix output (Fleiss et al. 1969; Cohen 1960). The first nominal class being considered for prediction, “subspecies”, conforms to existing *X. fastidiosa* subspecies taxonomic designation where host/subspecies association conforms with few exceptions to the following: citrus/coffee/olive = pauca; oleander = sandyi;

grape/elderberry = fastidiosa; oak/almond/plum/sycamore = multiplex; mulberry = morus (Nunney et al. 2014; Hernandez-Martinez 2006 et al. 2006; Schaad et al. 2004). The second nominal class being considered for prediction, “location”, conforms to an arbitrary geographical division derived from the isolate locales of both the GenBank specific genomes and RNB1. They are: “SouthAmerica”, “WesternUS”, “SouthwesternUS”, “SouthernUS”, “NortheastUS”, and “OldWorld”. The nominal attribute “host” was considered but was dismissed due to the large number of singletons and perceived redundancy with the nominal attribute “subspecies”. The classifier “subspecies” was therefore considered and excellent approximation for “host” data pending the ability to analyze a wider array of genomic data.

RESULTS:

Prophage region selection for qualitative analysis:

Post mining, standardized gene calling (GeneMark), and annotating, integrated regions were selected as described in the materials and methods. Again, the greatest representation of prophage regions both within the selected *X. fastidiosa* genome projects and among the subspecies designations with consistently alignable regions were chosen for the study. Exceptions made for the olive isolate (CoDiRO) were due to the novel nature of its recent disease appearance. One additional criteria potentially affecting secondary analysis was found in the called repressor region. A lack of redundant recoveries in the genomes led to an overinclusion of segments with more divergence than was allowed in other included prophage regions. A summary of the selections made is listed in table 4.1. Table 4.1 shows nine annotated prophage regions that occur across all nineteen genomes considered in this study (baseplate, integrase, lysozyme, polymerase, portal, repressor, tailfiber, tail protein, and terminase). Significant similarity was observed in three unique subcategories within both the

baseplate and tailprotein annotated regions across *X. fastidiosa* subspecies. Additionally, the former and the latter had the second and third highest selection percentage among the nine distinct regions next to the repressor range, indicating a high degree of observed similarity between the subdivided regions in question. Divergence in the integrase region led to the lowest selection percentage at ~37% of recovered annotated sequence. The integrase region among a smaller cohort of *X. fastidiosa* subspecies has been previously well described and categorized (de Mello Varani et al. 2008), however, the methodology of consistent non-divergent recovery across all represented genomes reduced average inclusion relative to the other selected prophage regions.

Dominant haplotypes and cross subspecies categorization:

The top three haplotypes were chosen for profiling among the included regions across subspecies. This methodology was employed due to the high degree of divergence observed and singleton categorization among mined prophage regions. Although several of the observations that have been made herein conform to previously investigated prophage regions, namely redundant isolate *Podoviridae*, *Myoviridae*, and *Siphoviridae* detection (Varani et al. 2013; Summer et al. 2010; Chen & Civerolo. 2008; de Mello Varani et al. 2008), the expanded profiling provided by this study extends dominant haplotypes into previously undescribed *X. fastidiosa* isolates. The description of these haplotype groupings, domain associations, and taxonomic rankings is provided in Table 4.2. Specifically, RNB1 was found to be among the most representative haplotypes in seven of the thirteen prophage regions, and the understudied isolates EB92.1 and Mul-MD contained unique genetic signatures at the repressor and terminase prophage regions respectively. Unlike the conserved viral taxonomy observed in the other dominant haplotypes, the repressor sequence showed significant E-Value annotation

from two distinct viral families within its respective dominant haplotype list. In addition to the taxonomic designations for the repressor haplotypes displayed in Table 4.2., a third singleton haplotype from the coffee derived isolate³², produced a non-trivial E-Value for Podoviridae assignment. Additionally, the European olive isolate was found to be unique to its haplotypic designation relative to all considered prophage regions. Of greatest interest was the terminase assignment. While the first returned BLASTp hit produced a metagenomic type return, the second near identical E-Value hit recovered a *Wolbachia* associated phage, WO (3e-77). This was highly distinct from any of the other returned dominant haplotype considerations which produced only γ proteobacterial assignments. The presence of an α proteobacterial related prophage region could be of significance in tracking post radiation events from the surprising discovery of *X. fastidiosa* in Italian olive hosts.

Result of machine learning classifier runs:

Post creation of dissimilarity matrices for polymorphic prophage regions, three classifiers were run to determine the continuity both between prophage regions between existing *X. fastidiosa* subspecies and among their geographic distribution. In other words, to what extent could the thirteen selected prophage regions (Table 4.1) show consistency via machine learning classifier metrics both between subspecies geographical location and known taxonomic designations.

To reiterate, predictive metrics were generated with three classifiers in the Weka environment described in the materials and methods section: Naïve Bayes, Logistic, and Sequential Minimal Optimization (SMO). Both Naïve Bayes and Logistic can be thought of as probabilistic approaches to data recognition, while SMO can be thought of as a means to solve a non-probabilistic approach to data classification. The overall performance of each classifier, post

training, is presented in Table 3. In short, Naïve Bayes proved to be the worst cumulative predictor of the “subspecies” nominal class given individual dissimilarity matrices for each prophage region. For the same nominal prediction, Logistic and SMO proved highly effective as evidenced by the supplied Kappa statistics for all prophage regions save the repressor regions (Table 4.3). The ability to correctly predict the subspecies nominal for a varied sized range of prophage insertions (Table 4.1) at near 85% or greater shows these classifiers to be of potential use (Table 4.3). The large scale failure for all three classifiers to make any meaningful, kappa supported prediction with regard to the nominal classifier “location”, however, is not surprising given the ability to correctly predict the nominal “subspecies” class. For instance, the nineteen genomes for all subspecies are drawn from disparate geographies across subspecies, and says more about the nature of prophage regions contained within specific subspecies independent of geographical location. The individual confusion matrix results of the two best classifiers for three regions (baseplate 894, tailfiber, and terminase) juxtaposed with the results from the naïve Bayes classifier are presented in figure 4.1. The confusion matrices contain the same row and column headers and are read along their respective diagonals for interpretation. Reading across the rows gives the actual count for the category in question, and reading down the columns gives the predictive results relative to the actual category count. Therefore, a valid interpretation is the higher the count along the respective diagonals, the more accurate the classifier in predicting the class in question. Taking the baseplate894 prophage region under the logistic classifier as an example, it is apparent that a high number of correct “subspecies’ predictions were made relative to two incorrect predictions where the multiplex subspecies was misidentified as pauca and the fastidiosa subspecies was also misidentified as pauca. The confusion matrices for the worst performing region, repressor, is also presented for naïve Bayes, logistic, and SMO in the same table. In general, the prevailing larger values along the confusion

matrix diagonals for both the logistic and SMO classifiers relative to naïve Bayes shows them to be superior predictors for this type of dataset. Again the repressor region is shown as counterpoint juxtaposed with the large number of successful predictions for the logistic and SMO classifiers. To account for random error, kappa statistic also accompany the raw percentage of correct predictions. A conservative approach to the kappa measure was taken and only those $>.75$ were considered as valid findings independent of descriptive percentages (Fleiss et al. 1969; Cohen 1960). Kappa should be considered a measure from 0.0 - 1.0 where larger numbers indicate predictive agreement. A kappa value in the range of 0.80-0.75 is generally considered a supported outcome (Landis and Koch 1977; Fleiss et al. 1969). Because confusion matrices are presented for only select prophage regions, and additional table highlighting the cumulative predictive power for all positive and negative “subspecies” responses for the three classifiers is also presented with an adjoining spark chart (Figure 4.2). Note that “location” was not supported at a meaningful Kappa level for any of the prophage regions across classifiers. For the cumulative positive predictions, the adjoining Sparkchart shows a downward pointing chevron. This indicates reduced accuracy for positive predictions given that the classifiers logistic and SMO are located at the extrema. Cumulative negative predictions show an upward pointing chevron in the adjoining Sparkchart, indicating that both logistic and SMO are again more accurate and suppress error rate better than naïve Bayes given the same extrema location.

DISCUSSION:

The continued profiling of prophage regions within *X. fastidiosa* genomes remains important for several reasons. First, there is the aforementioned contribution to general genomic variation (Van Sluys et al. 2003; Bhattacharyya et al. 2002), and as shown in this study, rapid, selective subspecies differentiation given a wider array of analytic approaches. As multi-attribute

repositories grow, efficient means of data sensemaking are needed. Given the limited sample size currently available to *X. fastidiosa* researchers, the results reported here suggest that expanded machine learning approaches using *X. fastidiosa* datasets are good candidate technologies. Expansion beyond prophage regions into data collections housing larger genomic collections may benefit from this approach, but the selective use of prophage regions for accurate categorization has been shown here. Next, the potential for lysogenic conversion and potential virulence in bacterial systems is yet another reason to study such regions. This phenomenon has been observed in numerous (γ) Gammaproteobacteria (Neely and Friedman 1998; Waldor and Mekalanos 1996; Hayashi et al. 1990). This has direct bearing on pathogen control as reports of pathogen movement from New World to Old World locales continue to emerge. A very recent and alarming publication has also raised the possibility of the presence of *X. fastidiosa* in Lebanon (Temsah et al 2015). Finally, recent work aimed at potential biocontrol measures has considered the uptake and retention of viruses in known *X. fastidiosa* vectors (Bhowmick et al. 2013; Das et al. 2013).

Considering the selection process and the choice to exclude prophage segments due to both high variability from common domain overlap, and insisted presence across all represented hosts, the appearance of bias along applications in machine learning is a possibility. Despite this semblance of manipulation, it should be argued that there exists biological reasons for the inclusion and exclusion of genomic regions in this way. First, from the diagnostic standpoint, highly alignable sequences lend themselves to universal or near universal primer sets for desired loci amplification. This in turn leads to reduced experimentation and rapid turnaround for amplicon production. An even more relevant reason is the necessity of region overlap for all existing hosts with a prophage region of interest. While this study in no way discounts the continued need to explore *X. fastidiosa* genomes, continuity for the sake of comparison is highly

relevant for the types of short term, applicability based questions plant pathologists are likely to ask. Reiterating the recent radiation epidemics caused by of *X. fastidiosa* in locales such as Italy, Taiwan, and Lebanon, no assumption can be made regarding underlying subspecies origins. To summarize, it would not be readily known whether the appearance of the pathogen was a result of recent careless or illegal transport, or had was a much more established denizen whose expression a confluence of factors.

Inclusive mining of current *X. fastidiosa* genomic deposits produced results consistent with prior mining assays, updating the patterns of *Podoviridae*, *Myoviridae*, and *Siphoviridae* distribution for more recent *X. fastidiosa* genome releases. Several noteworthy findings were also present. The repressor region , unlike other prophage sequences considered, had two representative viral families present in its dominant haplotype listing (Table 4.2). Further exploration of the mined repressor regions revealed a third viral family present based on E-value cutoffs of $<1e-10$. For this reason, it seems likely to conclude that this may have been an underlying reason for poor performance of this region across all classifiers in the machine learning aspects of this study. The variation observed in the repressor region across subspecies isolates showed that it lacked consistent genetic signatures to accurately classify it from the currently accepted taxonomic standpoint (Schuenzel et al. 2005; Parker et al. 2012; Nunney et al. 2014).

Another genomic novelty was found in the olive associated phage integrase region (Table 4.2). Subsequent BLASTp analysis called a GB3 synthase domain, which is largely associated with glycosphingolipids moieties within higher eukaryotic genomes (Keusch et al. 2000). Although the domain association could be considered marginal given its accompanying E-Value($8.7e-03$), its uniqueness among the recovered integrase regions suggests potential utility as a novel genetic signature. This is of even more importance given its association with the olive genome and the substantiated movement of *X. fastidiosa* into Europe. Regardless of its origin, it could

be viewed as a another potential tool to track clonal spread from the base olive tree isolation region. A final note regarding the olive isolate is its hybrid terminase region. The observed pattern of γ proteobacterial association with recovered prophage sequences was violated by this region in the olive genome as it associated with an α proteobacterial phage typical of *Wolbachia* genera (3e-77). A prophage region associated with both a novel geographical locale and a known invertebrate endosymbiont (Werren et al. 2008) suggest several possibilities. The genetic reorganization resulting from Old World phage distribution and parasitism renews concerns surrounding lysogenic conversion. It also suggests horizontal transfer via insects housing virions yet unknown in traditional *X. fastidiosa* vector studies. Either of these considerations, however, is speculative and supplemental studies would be needed for verification.

A comparative analysis of the classifiers starts with the both the limitations of the prophage selection region and the limitations of currently available genomes housed at GenBank (NCBI). First, it is important to understand the nature of what a confusion matrix says indicates about the performance of a classifier in relation to the underlying data. For the data set considered herein, the general workflow consists of the training data, the machine learning algorithm, and the construction of a classifier based on the chosen machine leaning algorithm. Test data is then taken and sent into the classifiers for predictive metrics. In this assay, a confusion matrix and kappa statistics describe the performance of the chosen classifier. As referenced in the material and methods section, tenfold cross validation was chosen for its ability to consider every data point for testing. Current limitations are ascribed to what will likely come to be known as a paltry number of *X. fastidiosa* genomes in coming years. Continued expression of Moore's law (Hayden 2014) in conjunction with machine learning classifiers promises the potentially more predictive power across disciplines. For instance, as the number of sequenced

X. fastidiosa genomes grow, such a genomic repository could serve as a master training set with user defined test sets submitted for immediate results. One example could be the use of the master repository to predict a subspecies nominal for hundreds of amplicons derived from an epidemiological study. In addition, many nominal classes could accompany coded polymorphic data to evaluate any number of attributes pertinent to the study at hand.

Analyzing the tendencies of the performances of the classifiers Naïve Bayes, Logistic, and SMO, the general approaches of each relative to underlying dissimilarity matrices suggests why the latter two may appeared to have performed better in predictive accuracy than Naïve Bayes, which . The generative classifier (Naïve Bayes) takes the approach of first considering the joint probability $p(x,y)$ and then fitting it into the schema $p(y|x)$ (Jordan 2002). To place it within a more colloquial context, this preliminary step leads to feature correlation that may bias and result in incorrect classification. Bringing this explanation back to the realm of biology, a prophage region is often characteristic of a remnant states, perhaps producing an alignment containing both domain overlap but sufficient gaps as well. Where gapped data is equally weighted, it appears that dissimilarity matrices constructed as they were in this paper can cause the Naïve Bayes classifier to incorrectly categorize subspecies. Logistic attempts to categorize $p(y|x)$ immediately (Jordan 2002), thereby, functioning better in assessing accurate “subspecies” calls for data of this sort. SMO attempts to segregate data by a series of optimization steps so as to create a border between classes by means of a hyperplane (Platt 1999). To summarize, a 2-D representation of data would be separated by a hyperplane which would take the form of a line in a 2-D space. This line would serve as a boundary between the groupings of points. Subsequent test data would then reside on one side of the line or the other. In this simplified binary it would be then classified according to that schema. A longer exposition of sequential minimal optimization is beyond the scope of this paper, but suffice it to

say that the segregation of data achieved by the SMO algorithm makes it an intriguing choice for polymorphic regions of this sort.

While the use of machine learning techniques is not novel in broad sense of scientific literature, it is seemingly underutilized in the realm of plant pathology. Though studies do exist (Villordon et al 2010; Perez-Ariza et al. 2012) the handling of large amounts of data in an efficient manner has become a rate limiting step in many biological subfields. For a pathologist, epidemiological considerations are often time critical and require rapid responses and recommendations. The findings in this study indicate that as genomic data proliferates across all kingdoms of life, rapid coalescence and analysis are imperative for the timely dissemination of new scientific findings. Here, *X. fastidiosa* prophage regions are characterized relative to the most inclusive set of *X. fastidiosa* genomes housed at GenBank, and used to illustrate machine learning applications both in plant pathology and beyond.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.
- Bhattacharyya, A., Stilwagen, S., Reznik, G., Feil, H., Feil, W. S., Anderson, I., ... & Predki, P. F. (2002). Draft sequencing and comparative genomics of *Xylella fastidiosa* strains reveal novel biological insights. *Genome research*, 12(10), 1556-1563.
- Bhattacharyya, A., Stilwagen, S., Ivanova, N., D'Souza, M., Bernal, A., Lykidis, A., ... & Kyripides, N. C. (2002). Whole-genome comparative analysis of three phytopathogenic *Xylella fastidiosa* strains. *Proceedings of the National Academy of Sciences*, 99(19), 12403-12408.
- Bhowmick, T. S., Das, M., Heinz, K. M., Krauter, P. C., & Gonzalez, C. (2013, June). Transmission of phage by glassy-winged sharpshooter. In *PHYTOPATHOLOGY* (Vol. 103, No. 6, pp. 16-16). 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA: AMER PHYTOPATHOLOGICAL SOC.
- Cariddi, C., Saponari, M., Boscia, D., De Stradis, A., Loconsole, G., Nigro, F., ... & Martelli, G. P. (2014). Isolation of a *Xylella fastidiosa* strain infecting olive and oleander in Apulia, Italy. *Journal of Plant Pathology*, 96(3), 1-5.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321-357.
- Chen, J., & Civerolo, E. L. (2008). Morphological evidence for phages in *Xylella fastidiosa*. *Virology*, 5, 75.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.
- Das, M., Bhowmick, T. S., Ahern, S. J., Young, R. F., & Gonzalez, C. (2013, June). Therapeutic and prophylactic application of phages to control Pierce's disease. In *PHYTOPATHOLOGY* (Vol. 103, No. 6, pp. 34-34). 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA: AMER PHYTOPATHOLOGICAL SOC.
- Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and

weighted kappa. *Psychological Bulletin*, 72(5), 323.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18.

Hayashi, T., Baba, T., Matsumoto, H., & Terawaki, Y. (1990). Phage-conversion of cytotoxin production in *Pseudomonas aeruginosa*. *Molecular microbiology*, 4(10), 1703-1709.

Hayden, E. C. (2014). The \$1,000 genome. *Nature*, 507(7492), 294-295.

Hernandez-Martinez, R., Pinckard, T. R., Costa, H. S., Cooksey, D. A., & Wong, F. P. (2006). Discovery and characterization of *Xylella fastidiosa* strains in southern California causing mulberry leaf scorch. *Plant disease*, 90(9), 1143-1149.

John, G. H., & Langley, P. (1995, August). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Morgan Kaufmann Publishers Inc..

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 841.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14), 3059-3066.

Keusch, J. J., Manzella, S. M., Nyame, K. A., Cummings, R. D., & Baenziger, J. U. (2000). Cloning of GB3 synthase, the key enzyme in globo-series glycosphingolipid synthesis, predicts a family of α 1, 4-glycosyltransferases conserved in plants, insects, and mammals. *Journal of Biological Chemistry*, 275(33), 25315-25321.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11), 54-64.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied statistics*, 191-201.

Leu, L. S., & Su, C. C. (1993). Isolation, cultivation, and pathogenicity of *Xylella fastidiosa*, the causal bacterium of pear leaf scorch disease in Taiwan. *Plant Disease*, 77(6), 642-646.

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark. hmm: new solutions for gene finding. *Nucleic acids research*, 26(4), 1107-1115.

Neely, M. N., & Friedman, D. I. (1998). Arrangement and functional identification of genes in the regulatory region of lambdoid phage H-19B, a carrier of a Shiga-like toxin. *Gene*, 223(1), 105-113.

Nunney, L., Schuenzel, E. L., Scally, M., Bromley, R. E., & Stouthamer, R. (2014). Large-scale

intersubspecific recombination in the plant-pathogenic bacterium *Xylella fastidiosa* is associated with the host shift to mulberry. *Applied and environmental microbiology*, 80(10), 3025-3033.

Parker, J. K., Havird, J. C., & De La Fuente, L. (2012). Differentiation of *Xylella fastidiosa* strains via multilocus sequence analysis of environmentally mediated genes (MLSA-E). *Applied and environmental microbiology*, 78(5), 1385-1396.

Peakall, R., & Smouse, P. E. (2012). GenAEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, 28(19), 2537-2539.

Perez-Ariza, C. B., Nicholson, A. E., & Flores, M. J. (2012). Prediction of coffee rust disease using Bayesian networks. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models* (pp. 259-266).

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods—support vector learning*, 3.

Schaad, N. W., Postnikova, E., Lacy, G., Fatmi, M. B., & Chang, C. J. (2004). *Xylella fastidiosa* subspecies: *X. fastidiosa* subsp. *piercei*, subsp. nov., *X. fastidiosa* subsp. *multiplex* subsp. nov., and *X. fastidiosa* subsp. *pauca* subsp. nov. *Systematic and applied microbiology*, 27(3), 290-300.

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*. *Applied and environmental microbiology*, 71(7), 3832-3839.

Saponari, M., Boscia, D., Nigro, F., & Martelli, G. P. (2013). Identification of DNA sequences related to *Xylella fastidiosa* in oleander, almond and olive trees exhibiting leaf scorch symptoms in Apulia (Southern Italy). *Journal of Plant Pathology*, 95(3).

Su, C. C., Chang, C. J., Chang, C. M., Shih, H. T., Tzeng, K. C., Jan, F. J., ... & Deng, W. L. (2013). Pierce's disease of grapevines in Taiwan: isolation, cultivation and pathogenicity of *Xylella fastidiosa*. *Journal of Phytopathology*, 161(6), 389-396.

Summer, E. J., Enderle, C. J., Ahern, S. J., Gill, J. J., Torres, C. P., Appel, D. N., ... & Gonzalez, C. F. (2010). Genomic and biological analysis of phage Xfas53 and related prophages of *Xylella fastidiosa*. *Journal of bacteriology*, 192(1), 179-190.

Temsah, M., Hanna, L., & Saad, A. (2015). First Report of *Xylella fastidiosa* associated with Oleander Leaf Scorch in Lebanon. *Journal of Crop Protection*, 4(1), 131-137.

Van Sluys, M. A., De Oliveira, M. C., Monteiro-Vitorello, C. B., Miyaki, C. Y., Furlan, L. R., Camargo, L. E. A., ... & Tsukumo, F. (2003). Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. *Journal of Bacteriology*, 185(3), 1018-1026.

de Mello Varani, A., Souza, R. C., Nakaya, H. I., De Lima, W. C., Paula de Almeida, L. G., Kitajima, E. W., ... & Van Sluys, M. A. (2008). Origins of the *Xylella fastidiosa* prophage-like regions and their impact in genome differentiation. *PLoS One*, 3(12), e4059.

Varani, A. M., Monteiro-Vitorello, C. B., Nakaya, H. I., & Van Sluys, M. A. (2013). The role of prophage in plant-pathogenic bacteria. *Annual review of phytopathology*, 51, 429-451.

Villordon, A., Clark, C., Smith, T., Ferrin, D., & LaBonte, D. (2010). Combining linear regression and machine learning approaches to identify consensus variables related to optimum sweetpotato transplanting date. *HortScience*, 45(4), 684-686.

Waldor, M. K., & Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, 272(5270), 1910-1914.

Wells, J. M., Raju, B. C., Hung, H. Y., Weisburg, W. G., Mandelco-Paul, 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(2), 136-143.

Werren, J. H., Baldo, L., & Clark, M. E. (2008). *Wolbachia*: master manipulators of invertebrate biology. *Nature Reviews Microbiology*, 6(10), 741-751.

Table 4.1. Profile and selection of prophage regions mined from the nineteen *X. fastidiosa* genomes

Annotated Region	HMM Candidate Sequences / Fragments	HMM Chosen Sequences / Fragments	Percentage Used	Full <i>X. fastidiosa</i> subspecies representation?	Length (MIN/MAX)
baseplate (combined)	133	96*	~72%	NO**	273 / 324; 588 / 588; 558 / 894
integrase	186	67	~36%	YES	108 / 1020
lysozyme	143	85	~59%	YES	108 / 534
polymerase	134	54	~40%	YES	336 / 2181
portal	53	29	~55%	NO**	312 / 1491
repressor	22	18	~82%	NO**†	648 / 780
tailfiber	49	32	~65%	YES	249 / 1245
tailprotein (Combined)	134	91††	~68%	NO**	219 / 219; 246 / 570; 1083 / 1083
terminase	60	30	50%	YES	171 / 1953

**X. fastidiosa* phage baseplate recoveries were split into three unique subcategories as unique similarity was observed between sequences of approximately 324bps, 588bps, and 894bps

**Post contig concatemerization, these phage elements were not detected via the GeneMark HMM calling algorithm in the recently described *X. fastidiosa* olive isolate

†Called phage repressor regions were not detected at a significant E-Value ($< 1e-05$) in the two oleander associated *X. fastidiosa* subspecies (*sandyi*)

††*X. fastidiosa* phage tail protein recoveries were split into three unique subcategories as unique similarity was observed between sequences of approximately 219bps, 570bps, and 1083bps

Table 4.2. Dominant haplotypes, taxonomy, and domain profiling of mined prophage regions for nineteen *X. fastidiosa* genomes based on described selection criteria

Prophage HMM called genes	Haplotype		BLAS Tp E- valu e	Bacterial Class	Viral Family rank	Xfas53 annotati on	Specific Domain Hit	Superfamily hit
haplotype representatives	Count / Breakdown	BLASTp Returns		Rank				
<u>Baseplate 324</u>								
Ann1CP_618 (oleander)	4 (<i>sandyi</i> , 4)	putative phage tail protein Escherichia phage vB_EcoM_ECO1230-10	1E- 43	γ proteob acteria	Myovirid ae	No	COG3628	GPW_gp25
RNB1_2653 (oak)	4 (<i>multiplex</i> , 3; <i>morus</i> , 1)	putative phage tail protein Escherichia phage vB_EcoM_ECO1230-10	4E- 46	γ proteob acteria	Myovirid ae	No	COG3628	GPW_gp25
Temecula1_396 (grape)	6 (<i>fastidiosa</i> , 6)	putative phage tail protein Escherichia phage vB_EcoM_ECO1230-10	3E- 43	γ proteob acteria	Myovirid ae	No	COG3628	GPW_gp25
<u>Baseplate588</u>								
M12_405 (almond)	3 (<i>multiplex</i> , 3)	putative baseplate assembly protein V Pseudomonas phage PPpW-3	6E- 37	γ proteob acteria	Myovirid ae	No	gpV / COG4540	Phage_base_V
Temecula1_1223 (grape)	3 (<i>fastidiosa</i> , 3)	putative baseplate assembly protein V Pseudomonas phage PPpW-3	1E- 38	γ proteob acteria	Myovirid ae	No	gpV / COG4540	Phage_base_V
Temecula1_401 (grape)	3 (<i>fastidiosa</i> , 3)	putative baseplate assembly protein V Pseudomonas phage PPpW-3	2E- 35	γ proteob acteria	Myovirid ae	No	gpV / COG4540	Phage_base_V

Baseplate 894

Ann1CP_1050 (oleander)	3 (<i>sandyi</i> , 3)	baseplate assembly protein Stenotrophomonas phage Smp131	6E-89	γ proteobacteria	P2-like Myoviridae	No	COG3948	Baseplate_J
M12_399 (almond)	3 (<i>multiplex</i> , 3)	baseplate assembly protein Stenotrophomonas phage Smp131	9E-90	γ proteobacteria	P2-like Myoviridae	No	none	Baseplate_J
Temecula1_395 (grape)	6 (<i>fastidiosa</i> , 6)	baseplate assembly protein Stenotrophomonas phage Smp131	3E-88	γ proteobacteria	P2-like Myoviridae	No	none	Baseplate_J

integrase

RNB1_335 (oak)	3 (<i>multiplex</i> , 3)*	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA_BRE_C
Temecula1_1132 (grape)	4 (<i>fastidiosa</i> , 4)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA_BRE_C
Temecula1_836 (grape)	4 (<i>fastidiosa</i> , 4)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA_BRE_C

lysozyme

RNB1_1204 (oak)	8 (7 <i>multiplex</i> ; 1 <i>morus</i>)	putative endolysin/autolysin Acinetobacter bacteriophage AP22	7E-35	γ proteobacteria	Myovirid ae	No	endolysin_ autolysin	lysozyme_like
RNB1_905 (oak)	4 (<i>multiplex</i> , 4)	putative endolysin/autolysin Acinetobacter bacteriophage AP22	5E-38	γ proteobacteria	Myovirid ae	No	endolysin_ autolysin	lysozyme_like
Temecula1_1102 (grape)	4 (<i>fastidiosa</i> , 4)	bacteriophage lysis protein; endolysin; lysozyme Phage Gifsy-2 (Salmonella)	2E-36	γ proteobacteria	Myovirid ae	No	endolysin_ autolysin	lysozyme_like

polymerase

GB514_600 (grape)	2 (<i>fastidiosa</i> , 2)**	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podovirid ae	Yes - (0.0)	None	DNA pol A
Temecula1_1262 (grape)	4 (<i>fastidiosa</i> , 4)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podovirid ae	Yes - (0.0)	None	DNA pol A
Temecula1_1486 (grape)	4 (<i>fastidiosa</i> , 4)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podovirid ae	Yes - (0.0)	None	DNA pol A

Portal***

M12_411 (almond)	3 (<i>multiplex</i> , 3)	capsid protein Escherichia phage vB_EcoM-ep3	2E-157	γ proteobacteria	Myovirid ae	No	portal_lam bda	Phage_portal
------------------	---------------------------	---	--------	----------------------------	----------------	----	-------------------	--------------

RNB1_3 (oak)	4 (<i>multiplex</i> , 4)	capsid protein Escherichia phage vB_EcoM-ep3	4E-159	γ proteobacteria	Myoviridae	No	portal_lambda	Phage_portal
--------------	---------------------------	--	--------	-------------------------	------------	----	---------------	--------------

Repressor

EB921_572 (eldeberry)	2 (<i>fastidiosa</i> , 2)	prophage repressor Enterobacteria phage mEp043 c-1	5.00 E-13	γ proteobacteria	Siphoviridae	No	S24_LexA-like	Peptidase_S24_S26
M12_1149 (almond)	2 (<i>multiplex</i> , 2)	putative transcriptional regulator Cronobacter phage ENT47670	9.00 E-16	γ proteobacteria	Myoviridae	No	S24_LexA-like	Peptidase_S24_S26
teme_1115 (grape)	3 (<i>fastidiosa</i> , 3)	prophage repressor Enterobacteria phage mEp043 c-1	3E-13	γ proteobacteria	Siphoviridae	No	S24_LexA-like	Peptidase_S24_S26

tailfiber†

Ann1CP_1383 (oleander)	2 (<i>sandyi</i> , 2)	putative tail-fiber protein Pectobacterium phage ZF40	7E-31	γ proteobacteria	Myoviridae	No	Collar	Collar superfamily
------------------------	------------------------	---	-------	-------------------------	------------	----	--------	--------------------

RNB1_1664 (oak)	2 (<i>multiplex</i> , 2)	putative tail-fiber protein Pectobacterium phage ZF40	6E-31	γ proteobacteria	Myoviridae	No	Collar	Collar superfamily
Temecula1_1310 (grape)	2 (<i>fastidiosa</i> , 2)	putative tail-fiber protein Pectobacterium phage ZF40	7E-31	γ proteobacteria	Myoviridae	No	Collar	Collar superfamily
<hr/> tail protein ~200++ <hr/>								
ann1CP2XXgene_608 (oleander)	3 (<i>sandyi</i> , 3)	tail protein Escherichia phage vB_EcoM-ep3	9E-17	γ proteobacteria	Myoviridae	No	None	Phage Tail X
rnb1_2XXgene_768 (oak)	3 (<i>multiplex</i> , 3)	tail protein Escherichia phage vB_EcoM-ep3	1E-16	γ proteobacteria	Myoviridae	No	None	Phage Tail X
Temecula_1211 (grape)	7 (<i>fastidiosa</i> , 5; <i>sandyi</i> , 2)	tail protein Escherichia phage vB_EcoM-ep3	1E-18	γ proteobacteria	Myoviridae	No	None	Phage Tail X
<hr/> tail protein ~500 <hr/>								
Ann1CPgene_616_5XX (oleander)	3 (<i>sandyi</i> , 3)	baseplate assembly protein Stenotrophomonas phage Smp131	2E-63	γ proteobacteria	P2-like Myoviridae	No	gpl	Tail P2 I
M12gene_398_5XX (almond)	3 (<i>multiplex</i> , 3)	baseplate assembly protein Stenotrophomonas phage Smp131	2E-62	γ proteobacteria	P2-like Myoviridae	No	gpl	Tail P2 I
Teme1gene_1218_5XX (grape)	8 (<i>fastidiosa</i> , 8)	baseplate assembly protein Stenotrophomonas phage Smp131	2E-63	γ proteobacteria	P2-like Myoviridae	No	gpl	Tail P2 I

tail protein
~1000+++

Ann1CPgene_607_10XX (oleander)	3 (<i>sandyi</i> , 3)	transcriptional regulator Escherichia phage vB_EcoM-ep3	6E-119	γ proteobacteria	Myovirid ae	No	gpD	Phage GPD
M12gene_390_10XX (almond)	2 (<i>multiplex</i> , 2)	transcriptional regulator Escherichia phage vB_EcoM-ep3	8E-119	γ proteobacteria	Myovirid ae	No	gpD	Phage GPD
rnb1gene_2146_10XX (oak)	2 (<i>multiplex</i> , 2)	transcriptional regulator Escherichia phage vB_EcoM-ep3	3E-118	γ proteobacteria	Myovirid ae	No	gpD	Phage GPD

terminase‡

M12gene_413_term (almond)	3 (<i>multiplex</i> , 3)	putative large terminase subunit Escherichia phage vB_EcoM_ECO1230-10	0.00	γ proteobacteria	Myovirid ae	No	Terminase GpA	Terminase GpA
MULMD_2166_term (mulberry)	2 (<i>morus</i> , 2)	putative large terminase subunit Escherichia phage vB_EcoM_ECO1230-10	0.00	γ proteobacteria	Myovirid ae	No	Terminase GpA	Terminase GpA
teme1gene_1231_term (grape)	4 (<i>fastidiosa</i> , 4)	putative large terminase subunit Escherichia phage vB_EcoM_ECO1230-10	0.00	γ proteobacteria	Myovirid ae	No	Terminase GpA	Terminase GpA

OLIVE
SUPPLEMENTAL‡
‡

Integrase

olive_1554	1 (<i>pauca</i> ,1)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA BRE C
olive_1632	1 (<i>pauca</i> ,1)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA BRE C
olive_1767	1 (<i>pauca</i> ,1)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA BRE C / Gb3 Synth

lysozyme

olive_1564	2 (<i>pauca</i> , 2)	putative endolysin/autolysin [Acinetobacter bacteriophage AP22]	2.00 E-36	γ proteobacteria	Myoviridae	No	endolysin _autolysin	lysozyme_like
olive_2194	1 (<i>pauca</i> ,1)	hypothetical protein ORF033 [Pseudomonas phage PA11]	2.00 E-32	γ proteobacteria	Podoviridae	No	endolysin _autolysin	lysozyme_like
olive_2195	1 (<i>pauca</i> ,1)	Insufficient residue count for identification	> 1.0 e-10	NA	NA	NA	NA	NA

polymerase

olive_1184	1 (<i>pauca,1</i>)	integrase Xylella phage Xfas53	1.00 E-128	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA Q like Exo
olive_2176	1 (<i>pauca,1</i>)	integrase Xylella phage Xfas53	0.00	γ proteobacteria	Podoviridae	Yes - (0.0)	None	DNA pol A

repressor

olive_1198	1 (<i>pauca,1</i>)	C2 Salmonella enterica bacteriophage SE1	1.00 E-26	γ proteobacteria	Myoviridae	No	HTH XRE / S24_LexA	HTH XRE / Peptidase_S24_S26
------------	----------------------	--	--------------	----------------------------	------------	----	-----------------------	--------------------------------

tailfiber

olive_1698	1 (<i>pauca,1</i>)	putative tail-fiber protein Pectobacterium phage ZF40	2.00 E-25	γ proteobacteria	Myoviridae	No	Collar	Collar superfamily
------------	----------------------	---	--------------	----------------------------	------------	----	--------	--------------------

terminase

olive_1958	1 (<i>pauca,1</i>)	terminase large subunit uncultured Mediterranean phage uvMED†††	4.00 E-79	NA	NA	No	YbcX	Terminase GpA
------------	----------------------	---	--------------	----	----	----	------	---------------

*three members haplotype groups included multiple *fastidiosa* and *multiplex* members,
multiplex chosen as representative
 **two, two member *fastidiosa*
 haplotypes present
 ***singletons omitted for portal regions after the first two haplotype groupings
 †eight categories of two haplotypes
 ††one additional *multiplex* specific
 three member haplotype
 †††one additional *multiplex* specific
 two member haplotype
 ‡one additional *pauca* specific two
 member haplotype
 ‡‡putative *pauca* assignment based
 on whole genome alignment
 ‡‡‡next viral associate hit to gp15 Wolbachia phage WO / E-Value 3e-77 / α
 proteobacteria taxonomic rank

Table 4.3. Accuracy percentages and Kappa support for Naïve Bayes, Logistic, and Sequential Minimal Optimization(SMO) classifier results based on both original attribute files and tenfold cross-validation and Synthetic Minority Oversampling Technique (SMOTE) tenfold cross-validation rebalanced files. The nominal categories under predictive consideration are "subspecies" and "location" and are labelled as such.

Proph age Regio n	Naïve Bayes predictive attribute "subspecies" Correctly Classified Instances percentage / Kappa Statistic	Logistic predictive attribute "subspecies" Correctly Classified Instances percentage / Kappa Statistic	SMO predictive attribute "subspecies" Correctly Classified Instances percentage / Kappa Statistic	Naïve Bayes predictive attribute "location" Correctly Classified Instances percentage / Kappa Statistic	Logistic predictive attribute "location" Correctly Classified Instances percentage / Kappa Statistic	SMO predictive attribute "location" Correctly Classified Instances percentage / Kappa Statistic
basep late3 24	89.1892% (SMOTE) / .8597	89.1892% (SMOTE) / .8584	91.8919% (SMOTE) / .8947	Kappa fails to support	Kappa fails to support	Kappa fails to support
basep late5 88	84.375% / .7911	88.5714%(SMOTE) / .8537	88.5714% (SMOTE) / .8536	Kappa fails to support	Kappa fails to support	Kappa fails to support
basep late8 94	Kappa fails to support	93.9394%(SMOTE) / .9224	96.9697 (SMOTE) / .9609	Kappa fails to support	Kappa fails to support	Kappa fails to support
integr ase	Kappa fails to support	86.5672% / .8221	89.5522% / .8625	Kappa fails to support	Kappa fails to support	Kappa fails to support
lysozy me	Kappa fails to support	82.4176% (SMOTE) / .7636	88.2353 % / .8375	Kappa fails to support	Kappa fails to support	Kappa fails to support
polym erase	Kappa fails to support	88.5246% (SMOTE) / .8547	90.1639 % (SMOTE) / .8753	Kappa fails to support	Kappa fails to support	Kappa fails to support
portal	Kappa fails to support	90.3226% (SMOTE) / .8569	83.871 % (SMOTE) / .752	Kappa fails to support	Kappa fails to support	Kappa fails to support

repre	Kappa fails to support	Kappa fails to support	Kappa fails to support	Kappa fails to support	Kappa fails to support	Kappa fails to support
ssor			94.4444 % (SMOTE) /			
tailfib	84.375 / .7949	96.875% / .9593	.9287	Kappa fails to support	Kappa fails to support	Kappa fails to support
er						
tailpr		83.871% (SMOTE) /	83.871 % (SMOTE) /			
otein	86.2069 / .8083	.7885	.7885	Kappa fails to support	Kappa fails to support	Kappa fails to support
2XX						
tailpr	83.3333 (SMOTE) /	83.3333% (SMOTE) /	83.3333 % (SMOTE) /			
otein	.7692	.7741	.7791	Kappa fails to support	Kappa fails to support	Kappa fails to support
5XX						
tailpr		80.6452% (SMOTE) /	87.0968% (SMOTE) /			
otein	Kappa fails to support	.7483	.8256	Kappa fails to support	Kappa fails to support	Kappa fails to support
10XX		90.625% (SMOTE) /	90.625 % (SMOTE) /			
termi		.8657	.8611			
nase	Kappa fails to support			Kappa fails to support	Kappa fails to support	Kappa fails to support

Figure 1. Comparative confusion matrix results for the three best performing (baseplate894, tailfiber, and terminase) and worst performing (repressor) classifiers based on kappa support. The nominal under prediction is "subspecies" with identical labelling for both rows and columns. The rows contain the actual "subspecies" counts and the columns contain the predicted "subspecies" counts.

Accurate predictions for the "subspecies" nominal are read along individual matrix diagonals. Row populations outside of the diagonals represent erroneous "subspecies" nominal calls. Conditional formatting is provided to highlight the magnitude of non-zero cells. A differential color scale is provided to the right of the figure.

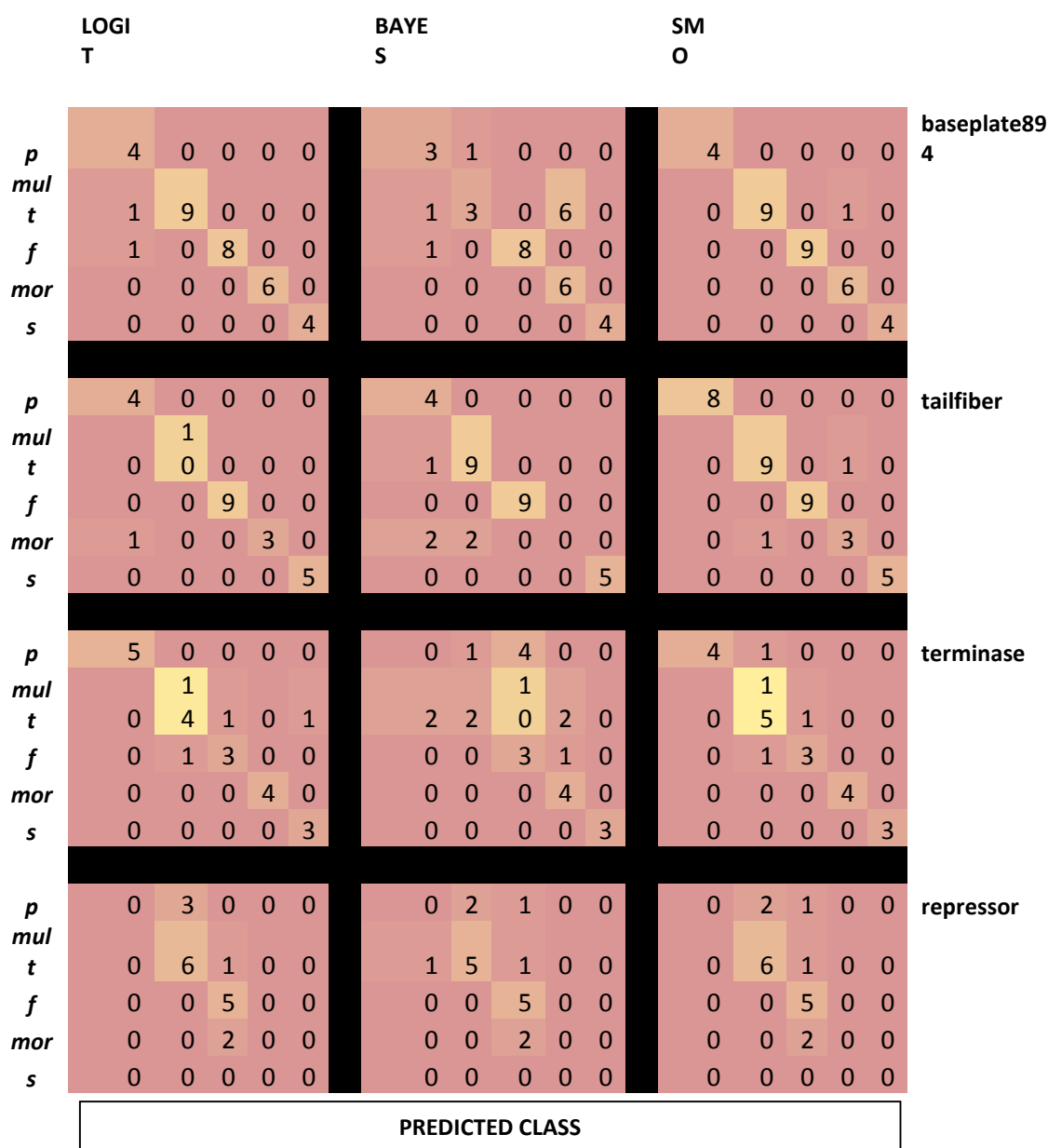




Figure 2.

Comparative cumulative accuracy rates among the three classifiers Naïve Bayes, Logistic, and Sequential Minimal Optimization(SMO) for the predicted nominal "subspecies".

"A" is representative of the cumulative error rate for the predicted nominal "subspecies" across all thirteen defined prophage regions, and "B" represents the accuracy for the "subspecies" nominal prediction across the same thirteen defined prophage regions. Sparklines follow each count to elucidate trends.

"A" shows an upward pointing chevron indicating that both logistic and Sequential Minimal Optimization(SMO) classifiers have an overall lower kappa supported error rate for "subspecies" prediction.

"B" shows a downward pointing chevron indicating that both logistic and Sequential Minimal Optimization(SMO) classifiers have overall higher kappa supported accuracy for "subspecies" prediction.

	LOGIT	BAYES	SMO	
<i>p</i>	0.161	0.383	0.129	
<i>mult</i>	0.128	0.436	0.101	
<i>f</i>	0.088	0.310	0.071	
<i>mor</i>	0.263	0.366	0.229	
<i>s</i>	0.066	0.390	0.105	
	LOG	BAYES	SMO	
<i>p</i>	0.840	0.617	0.871	
<i>mult</i>	0.872	0.564	0.899	
<i>f</i>	0.912	0.690	0.929	
<i>mor</i>	0.737	0.634	0.771	

s 0.934 0.610 0.895

APPENDIX

Supplemental Figure 2.1. Individual Bayesian gene tree files for the seventeen previously described loci used in this study

(hosted <http://www.eden.rutgers.edu/~gregbehr/>)

Supplemental Figure 2.3. Alignment files for recombination Analysis via Difference of Sums of Squares (DSS) method (McGuire et al. 2000)

(hosted <http://www.eden.rutgers.edu/~gregbehr/>)

Supplemental Figure 2.4. MLSA and MLSA-E concatemerized isolate/environmental sample sequence for all profiled loci

(hosted <http://www.eden.rutgers.edu/~gregbehr/>)

Supplemental Figure 2.5. RAxML output files for comparative analysis relative to Bayesian methodology / recovered Bayesian topology

(hosted <http://www.eden.rutgers.edu/~gregbehr/>)

Supplemental 3.1. Associated *.faa, *.fna, *.gbk, and annotation files based on the assembled *Quercus palustris* associated isolate (RNB1)

(hosted <http://www.eden.rutgers.edu/~gregbehr/>)