Title
Phylogenetics of Historical Host Switches in a Bacterial Plant Pathogen

Journal
Applied and Environmental Microbiology, 88(7)

ISSN
0099-2240

Authors
Kahn, Alexandra Katz
Almeida, Rodrigo PP

Publication Date
2022-04-12

DOI
10.1128/aem.02356-21

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

EVOLUTIONARY AND GENOMIC MICROBIOLOGY

Phylogenetics of Historical Host Switches in a Bacterial Plant Pathogen

Alexandra Katz Kahn,[a] Rodrigo P. P. Almeida[a]

[a]Department of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, California, USA

ABSTRACT *Xylella fastidiosa* is an insect-transmitted bacterial plant pathogen found across the Americas and, more recently, worldwide. *X. fastidiosa* infects plants of at least 563 species belonging to 82 botanical families. While the species *X. fastidiosa* infects many plants, particular strains have increased plant specificity. Understanding the molecular underpinnings of plant host specificity in *X. fastidiosa* is vital for predicting host shifts and epidemics. While there may exist multiple genetic determinants of host range in *X. fastidiosa*, the drivers of the unique relationships between *X. fastidiosa* and its hosts should be elucidated. Our objective with this study was to predict the ancestral plant hosts of this pathogen using phylogenetic and genomic methods based on a large data set of pathogen whole-genome data from agricultural hosts. We used genomic data to construct maximum-likelihood (ML) phylogenetic trees of subsets of the core and pan-genomes. With those trees, we ran ML ancestral state reconstructions of plant host at two taxonomic scales (genus and multiorder clades). Both the core and pan-genomes were informative in terms of predicting ancestral host state, giving new insight into the history of the plant hosts of *X. fastidiosa*. Subsequently, gene gain and loss in the pan-genome were found to be significantly correlated with plant host through genes that had statistically significant associations with particular hosts.

IMPORTANCE *Xylella fastidiosa* is a globally important bacterial plant pathogen with many hosts; however, the underpinnings of host specificity are not known. This paper contains important findings about the usage of phylogenetics to understand the history of host specificity in this bacterial species, as well as convergent evolution in the pan-genome. There are strong signals of historical host range that give us insights into the history of this pathogen and its various invasions. The data from this paper are relevant in making decisions for quarantine and eradication, as they show the historical trends of host switching, which can help us predict likely future host shifts. We also demonstrate that using multilocus sequence type (MLST) genes in this system, which is still a commonly used process for policymaking, does not reconstruct the same phylogenetic topology as whole-genome data.

KEYWORDS *Xylella fastidiosa*, host specificity, quarantine, trade, policymaking, ancestral state reconstruction, genomic diversity, pan-genome, phylogenomics, host-pathogen interactions, phylogenetic analysis, phytopathogens, plant microbiology

Modern plant trade disturbs historical ecological relationships and creates opportunities for the development of novel pathogenic interactions (1, 2), often with correlated genetic changes (3). However, pathogens must be adapted to the environment of the novel host before they meet, or they will not be able to survive and reproduce (4). That does not mean pathogens necessarily preadapted to the exact same host, but they could have adapted to a similar host earlier and retained that adaptation until encountering a novel host. Convergent evolution in diverse pathogen populations can allow for divergent strains to have the ability to infect the same hosts. Three potential

Editor Christopher A. Elkins, Centers for Disease Control and Prevention

Address correspondence to Rodrigo P. P. Almeida, rodrigoalmeida@berkeley.edu.

The authors declare no conflict of interest.

Received 15 December 2021
Accepted 14 February 2022
Published 21 March 2022

mechanisms of genetic change that can accompany host shifts are nucleotide changes leading to different alleles in the core genome of a pathogen (defined as the genes shared by all strains in a set of samples), whole-gene gain and loss in the pan-genome, leading to unique sets of genes in individual strains, or regulatory/epigenetic changes. Due to the recent increase in whole-genome sequencing of plant pathogens, we can now more effectively use phylogenetic analyses to investigate their genetic associations to both novel and historical host plants (5). Understanding the phylogenetic relationships between specific hosts and pathogens should allow the development of preemptive plans to protect natural ecosystems as well as agriculture from the emergence of novel pathogens.

Xylella fastidiosa is an insect-transmitted, xylem-limited bacterial plant pathogen found across the Americas and, as of recently, globally. *X. fastidiosa* is considered to be a generalist pathogen because, as a species, it reportedly infects at least 563 species belonging to 82 botanical families (6). The lack of host specificity that *X. fastidiosa* exhibits as a species contrasts with increased plant host specificity in smaller clades and strains (6–12). It is still debated whether a pathogen like *X. fastidiosa* should be considered a generalist species that "leaps" between phylogenetically distant hosts or, alternatively, a crawler at shallower clades (7, 13). The difference is biological, as there are unique implications for either evolutionary path. *X. fastidiosa* could be repeatedly evolving specialization, or it could have biological and genetic traits as a species that make particular hosts of disparate plant taxa suitable.

From an applied perspective, there have been recent calls from government agencies for increased focus on understanding the host range of *X. fastidiosa*. This is because the pathogen has been deemed likely to spread and to be of extremely high risk to crops of agricultural value (14). *Xylella fastidiosa* causes disease in a range of high-value crops, including Pierce's disease of grapevines (PD), citrus variegated chlorosis disease in sweet oranges, almond leaf scorch, leaf scorch of coffee, and olive quick decline syndrome (OQDS), spanning North and South America, Europe, the Middle East, and Taiwan (7, 15, 16). While there are three distinctive subspecies of *X. fastidiosa* and it would be desirable to be able to use those subspecies for management decisions, so far, the subspecies have not been found to have sufficient resolution to define host range or to infer risk (7). Understanding the molecular basis of plant host specificity in *X. fastidiosa* is vital for predicting and acting upon host shifts, but these are processes yet to be described (7).

Xylella fastidiosa is a member of the group *Xanthomonadaceae* and phylogenetically clusters sister to *Xanthomonas albilineans*, technically within the paraphyletic genus *Xanthomonas*, although *Xylella* is considered a separate genus (17, 18). *Xylella* spp. and *Xanthomonas albilineans* are the only xylem-limited *Xanthomonadaceae* and have convergently reduced genomes compared to the rest of the genus (18). *Xylella* also lacks a type III secretion system (T3SS), a loss compared to its higher-order taxonomic group. As the purpose of the T3SS in phytopathogens is to deliver effectors into living plant cells (19), the loss has been hypothesized to be due to *X. fastidiosa* primarily interacting with nonliving tissue, insect cuticle, and mature xylem vessels (20).

While the molecular basis of host range is not understood, there are consistent patterns in the ability of particular *X. fastidiosa* isolates to infect specific plant hosts regardless of their environmental condition (8, 11). This implies that genetics, as opposed to only environmental conditions, underlie the relationship between isolates and plant hosts that allow for colonization. Recurring pathogen specificity to a particular host can be either explained through phylogenetic signal, where members of a clade have shared traits that allow for pathogenesis in that host, or by pathological convergence, where more distantly related strains have separately acquired mechanisms for virulence. Both processes have underlying genetics, but each shows different phylogenetic patterns (21). Last, we have seen that deletion of *rpfF*, which controls cell-cell signaling via a diffusible signal factor (DSF), can expand the host range of *X. fastidiosa* (22). Other insights into host range have been made in terms of plant immunological studies. For example, removing the O-antigen from the exterior of

X. fastidiosa cells allows the plant to quickly recognize *X. fastidiosa* and initiate immune responses, thus decreasing its likelihood of colonization of the plant (23). O-antigens are highly variable and evolve rapidly and often are shown to have coevolutionary histories between symbiotic organisms, as they are the first exposed part of any bacterium (24). In terms of phylogenetic methods, cophylogenies have shown no cospeciation between plant hosts and *X. fastidiosa* or any other congruence between the evolutionary histories of *X. fastidiosa* and its plant hosts (7). Based on the current data, it is not generally possible to tell if *X. fastidiosa* is undergoing host jumps or range expansions; however, the data available so far suggest that both are occurring given that, in certain situations, we see strains able to infect multiple hosts (8), while in other situations, we see multiple strains coexisting in nature but no cross infections of hosts (11).

Using the influx of whole-genome data generated in the past several years, we searched the genomes of *X. fastidiosa* for correlations with plant host species. The first method we pursued was conducting ancestral state reconstructions. Ancestral state reconstructions use genetic data (phylogenies), with a known phenotype for each taxon, to characterize the most likely state that each ancestral node of the tree would have possessed for the phenotype of interest. This tool has been used to understand host-pathogen interactions via ancestral state reconstructions in fungi and trematodes parasite systems (25, 26). Ideally, we would be able to ask: what was the most likely ancestral host of the ancestor of all *X. fastidiosa*? If we can understand patterns in the past, it can help us better build models to predict future hosts based on the genomic changes associated with historical host shifts. Following the ancestral state reconstructions, we looked further into the pan-genome by calculating correlations between plant host types and the presence/absence of each gene.

This study aimed to compare the commonly used genetic data sets available for phylogenetic analyses of *X. fastidiosa* both to compare phylogenetic topologies as well as ancestral host states from each data set. We hypothesized that the pathogen phylogeny would be correlated with host history and that we could observe this trend through ancestral state reconstruction. If there is no relationship between host and the phylogeny, there should not be conclusive ancestral state reconstruction results. We hypothesize that by using either the core genome of *X. fastidiosa*, pan-genome phylogenetic tree, or both, it would be possible to estimate the likelihood of hypothetical plant hosts for ancestral nodes of interest (a node represents a common ancestor of the tips). This would show that the host is largely dependent and predictable based on the phylogeny of bacterial relationships and would lead to further pursuing allelic differences in core genome and/or gene gain/loss in the pan-genome and estimate how either or both are correlated with plant host identity. While not biologically meaningful, since multilocus sequence type (MLST) data are still frequently used in *X. fastidio sa* management, we included that data type in our analysis for comparison as well.

RESULTS

Phylogenetic reconstruction of disparate regions and sizes are topologically similar. The pan-genome of all sequences and the outgroup *Xylella taiwanensis* Wufong1_PLS229 ($n = 349$) contained 17,024 genes (14,564 of which come from the ingroup *X. fastidiosa*). The alignment of MLST genes totaled 4,146 bp in length, while the core genome comprised 1,411 concatenated regions in a total of 354,816 bp. Nonrecombinant regions identified with ClonalFrameML (27) comprised only 32% of the core genome (68% of the alignment showed evidence of recombination), leaving an alignment consisting of only 112,819 bp (Table 1). The alignment contained 130 pairs of sequences that were completely identical to each other, highly reducing the amount of within-subspecies differentiation that is possible with this data set and creating large polytomies of indistinguishable sequences within *X. fastidiosa* subsp. *fastidiosa* (mostly California *Vitis* samples), as well as within *X. fastidiosa* subsp. *pauca* (mostly Italian *Olea* samples). Due to this lack of within-subspecies resolution, the phylogeny with recombinant regions removed is only suitable for between-subspecies comparisons due to the extensive data loss in removing recombinant regions. The strains and

TABLE 1 Summary of alignments and phylogenetic diversity for each of the four alignments and corresponding phylogenetic trees

Phylogeny data source	Total alignment length	Phylogenetic diversity (summed substitutions per site)
Nonrecombinant genome	112,819 bp	3.32
Core genome	354,816 bp	6.65
MLST genes	4,146 bp	8.65
Pan-genome	17,024 genes	59.15

locations in the alignment with recombination can be visualized in Fig. S3 in the supplemental material.

While between-subspecies topologies are similar among the four trees generated, they are not identical. The core genome tree shows consensus of taxonomic division into three subspecies; however, *X. fastidiosa* subsp. *sandyi* and *X. fastidiosa* subsp. *morus* could be either part of subsp. *fastidiosa* or each their own small subspecies without affecting the monophyly of *X. fastidiosa* subsp. *fastidiosa*. (see Fig. 1 and 2 for phylogenetics and Table S1 for strain information). The nonrecombinant tree is similar except that *X. fastidiosa* subsp. *morus* is clustered within *X. fastidiosa* subsp. *fastidiosa*. The pan-genome splits the most basal of the three subspecies, *X. fastidiosa* subsp. subsp. *pauca*, into a paraphyletic cluster; however, it places *X. fastidiosa* subsp. *multiplex*, *X. fastidiosa* subsp. *fastidiosa*, *X. fastidiosa* subsp. *morus*, and *X. fastidiosa* subsp. *sandyi* similar to the core phylogeny (see Fig. 1). The MLST tree shows *X. fastidiosa* subsp. *morus* as the outgroup to *X. fastidiosa* subsp. *fastidiosa*, while *X. fastidiosa* subsp. *sandyi* falls within *X. fastidiosa* subsp. *fastidiosa*. The other difference among the four topologically similar trees is variation in branch length. The phylogenetic diversities were calculated as the summed length of each tree calculated from nodes to root and were core, 6.65; nonrecombinant, 3.32; MLST, 8.65 (in substitutions per site); and pan-genome, 59.15 (in gene gains and losses per site). Since the pan-genome tree was built with gene presence/absence data, it was calculated in gene changes per site. Phylogeny and alignment information are summarized in Table 1. A 16S rRNA gene phylogeny was also built as a comparison (Fig. S5), but the phylogeny provided very poor differentiation among strains (only 40 unique sequences out of the 349 strains).

Within *X. fastidiosa* subsp. *fastidiosa*, the core genome rearticulates the three PD clades that were found in Castillo et al. (28) (Fig. S4). Within the clade defined as PD-III, the sequence similarity in the core has led to extensive polytomies, with many sequences indistinguishable in the core (Fig. 2). The three PD clades are also articulated in the nonrecombinant phylogeny and the pan-genome phylogeny; however, the MLST tree does not differentiate these clades from one another. Not poorly resolved, the MLST does have high bootstrap support for clades that conflict with trees constructed with core- and pan-genome trees, suggesting that using MLST genes have the potential to subvert the analysis of relationships between taxa while showing strong bootstrap support.

Within *X. fastidiosa* subsp. *multiplex*, there have typically been considered two groups, the nonrecombining "non-IHR," as well as the recombining outgroup "IHR" (29). The core genome tree, as well as the MLST tree, both articulate these two groups, the clade non-IHR, as well as the nonmonophyletic recombining group, IHR. The nonrecombinant tree and the pan-genome tree do not recreate these groupings (Fig. S4).

All phylogenies but the pan-genome show a consistent split in *X. fastidiosa* subsp. *pauca* between the strains isolated from the Italian OQDS outbreak and the mixed host strains from Brazil. Within the OQDS strains, as well as several very closely related strains from Costa Rica, there is no clear resolution at this genomic scale. Within the Brazilian clade, strain Hib4 is the outgroup in all phylogenies except the MLST.

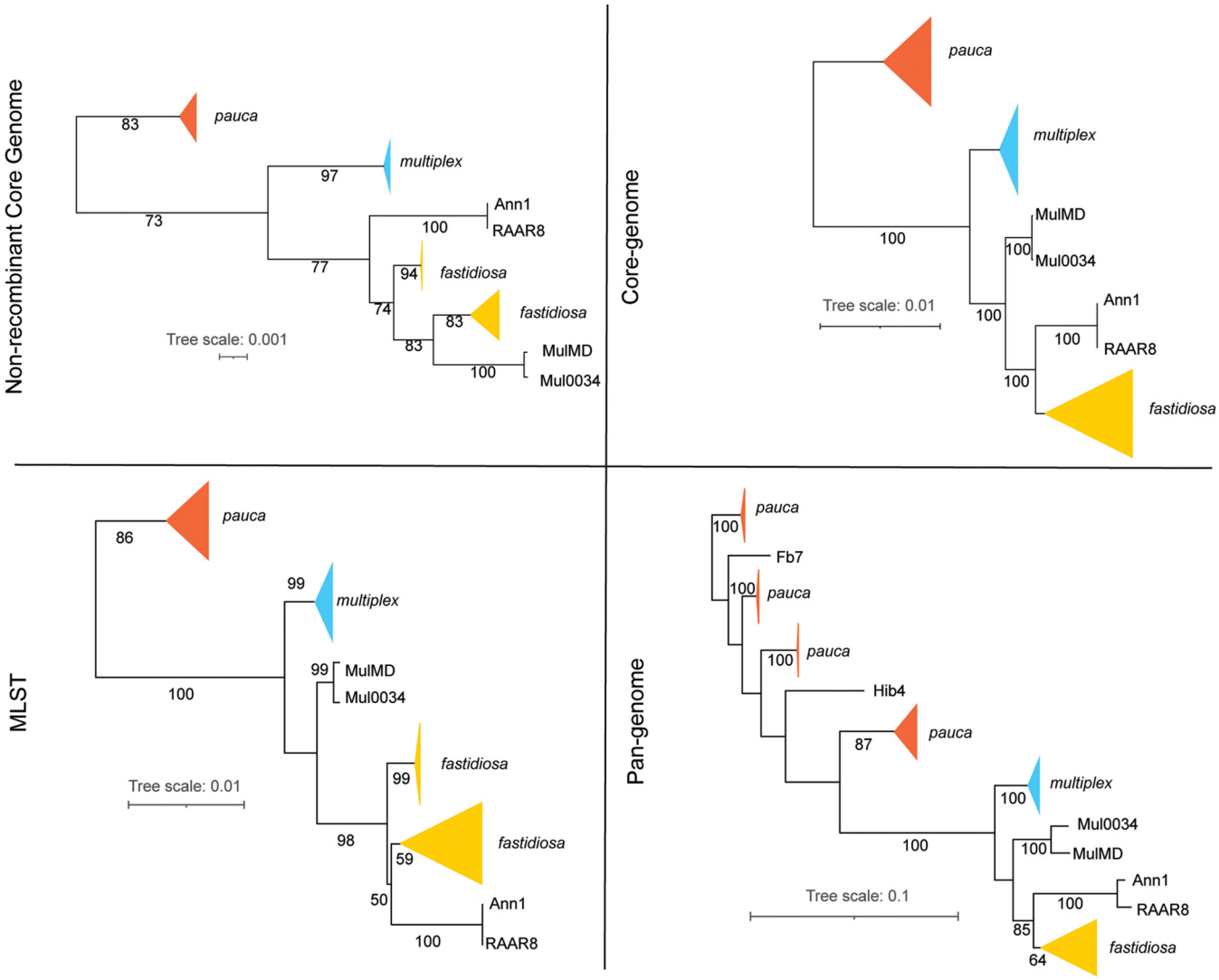

FIG 1 Maximum-likelihood phylogenies for all four genomic subsets with $n = 349$ strains, collapsed to the potential subspecies level. MulMD and Mul0034 are considered examples of the debated *X. fastidiosa* subsp. *morus*, while Ann1 and RAAR8 are examples of the second debated subspecies, *X. fastidiosa* subsp. *sandyi*. *X. fastidiosa* subsp. *fastidiosa*, *X. fastidiosa* subsp. *pauca*, and *X. fastidiosa* subsp. *multiplex* have only been collapsed to clades which do not include those four strains. Each phylogeny has a separate scale of substitutions across its branches. Node support is shown as bootstrap values, with values under 50 not displayed. The outgroup used for all trees is the strain Wufong1, a member of the species *Xylella taiwanensis*, which has been trimmed from these trees for visualization.

The reconstructed ancestral likelihoods suggest ancestral hosts of *X. fastidiosa*. Interrogating the results of the ancestral state reconstruction to the genus level of the core genome phylogeny shows undetermined hosts at the deepest nodes (Fig. 3). However, the ancestral node of *X. fastidiosa* subsp. *fastidiosa* has a significant association with the plant genus *Coffea*, which persists throughout *X. fastidiosa* subsp. *fastidiosa* as the most likely ancestral host for all strains isolated from South and Central America. This changes for the PD clade, where the ancestral host of all nodes except one is *Vitis*, the one exception being an ancestral *Prunus* node. *X. fastidiosa* subsp. *sandyi* and *X. fastidiosa* subsp. *morus* are undetermined in ancestral hosts. *X. fastidiosa* subsp. *multiplex* has a more dynamic history, with *Vaccinium* shown to be the most likely ancestral host for the subspecies, and then within the clade, a switch to a large group of nodes whose most likely host is *Prunus*, as well as two nodes depicting *Platanus* and *Olea*. *X. fastidiosa* subsp. *pauca* does not have a determined ancestral host of the whole subspecies, and internal nodes switch several times between *Citrus* and *Coffea* and once to *Olea*.

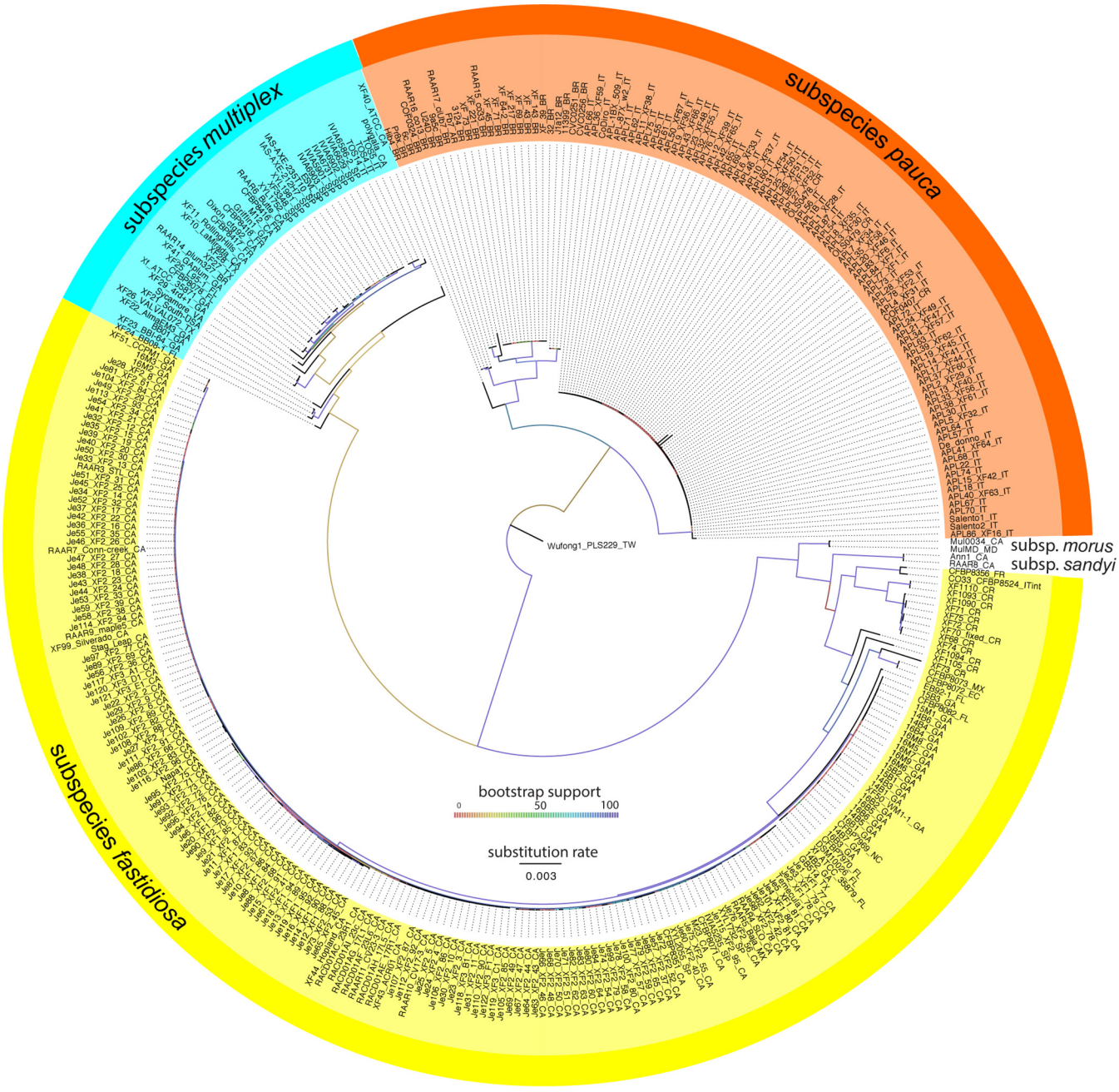

FIG 2 Maximum-likelihood phylogeny of the core genome without any clades collapsed. The branch lengths are quantified by substitution rate, and the bootstrap support is depicted by branch color. Each strain name includes an abbreviated reference to its origin; see Table S1 in the supplemental material for more details on the history of each strain. The branch length leading to the outgroup, Wufong1, has been removed to clarify relationships within the species *X. fastidiosa*. Dashed lines are used to connect tips of the phylogeny to the taxa names.

In terms of the genera across the reconstructions, while the deep nodes (ancestors of a subspecies) are often undetermined, there is more resolution within subspecies (Fig. 4). The node that is consistent across the four reconstructions is that there is a high likelihood of the genus *Coffea* being the ancestral host of the node representing the introduction of *X. fastidiosa* subsp. *fastidiosa* from Central to North America. The genus *Vaccinium* was predicted as the most likely ancestral host of *X. fastidiosa* subsp. *multiplex* in the core genome phylogeny, whereas in the nonrecombinant phylogeny, the ancestor of all but one strain of *X. fastidiosa* subsp. *multiplex* is the genus *Prunus*. All four trees agree upon the ancestor of the internal non-IHR multiplex clade being *Prunus*. In terms of the transition models chosen for each reconstruction, most trees

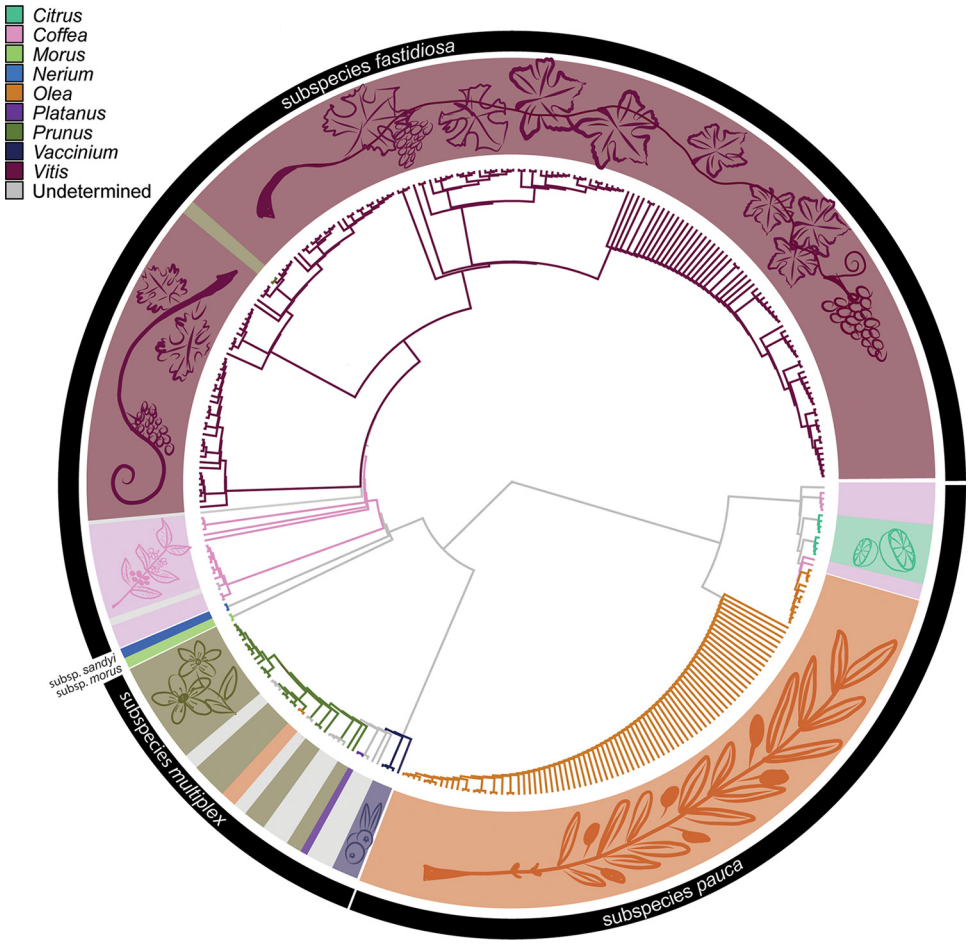

FIG 3 Cladogram of the core genome with the most likely genus of each node mapped onto the branches. Nodes with likelihood of less than 95% for one genus are colored in gray and marked as undetermined. The ancestral state reconstruction was conducted with an equal rates model.

had lower Akaike information criterion (AIC) scores when using the equal rates model with fewer parameters than the symmetrical rates model, the exception being for the pan-genome super order reconstruction having a lower AIC score with the symmetric model than the equal rates (Table S2).

At the node representing the ancestor of the species *X. fastidiosa*, both the nonrecombinant core and pan-genome phylogenies predict that the clade Rosid is the most likely ancestral host (Fig. 5). The core and MLST phylogenies predict Asterid to be the ancestral host but at lower likelihoods of 87 and 78%, respectively, which are visualized, along with all likelihoods under 95%, as undetermined (See Fig. 5 and Table S2). There is enough discordance between reconstructions that a consistent pattern at this host depth is unlikely.

Four plant clades correlated with gene presence and/or absence. Bacteria isolated from the genera *Coffea* and *Vitis*, as well as the superorders Asterid and Rosid, have *X. fastidiosa* genes with which they are significantly correlated, totaling 30 genes (Table 2). Ten of these 30 genes are significantly correlated with both Asterids and Rosids, with paired, opposite relationships (i.e., the same gene is significantly absent for one host, while it is present for another) (Table 2). Some correlations are of significance due to elevated presence of the gene among strains found in a particular host, while most are significant due to an absence of particular genes in the host of interest. Since lineage-specific interdependencies are accounted for with the phylogeny, the correlated genes are representative of convergent processes, either evolutionarily or via lateral gene transfer, not shared ancestry by descent. Genes that are significant

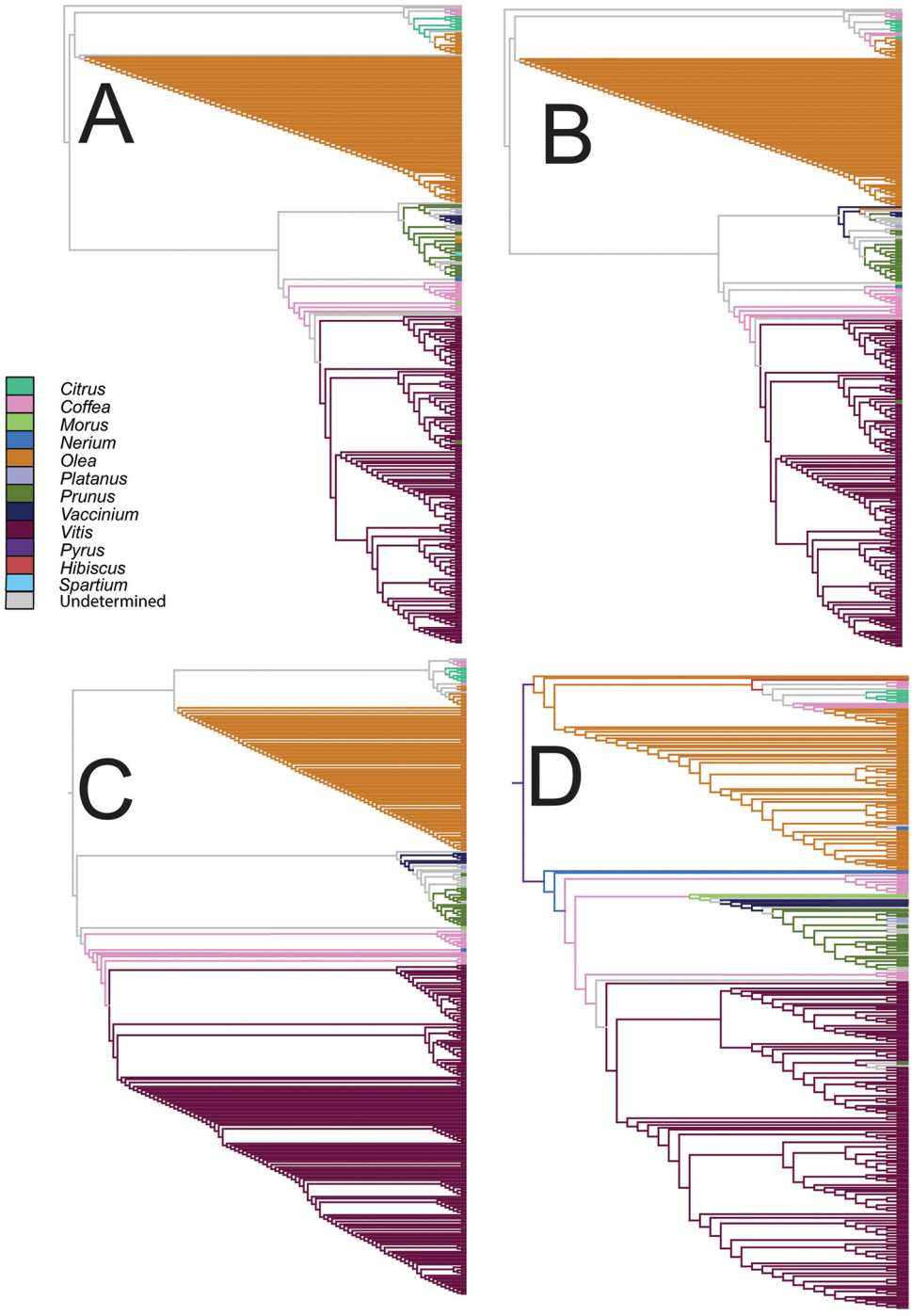

FIG 4 Cladograms of all four genomic regions with the most likely genus of each node mapped onto the branches. Nodes with likelihood of less than 95% for one genus are colored in gray and marked as undetermined. (A) Nonrecombinant core cladogram; (B) core cladogram; (C) MLST cladogram; (D) pan-genome cladogram. All four ancestral state reconstructions were done with an equal rates transition model between hosts.

mark repeated nonvertical descent changes in the pan-genome of strains in convergent patterns specific to the hosts of interest. While most identified genes are hypothetical proteins, genes shown to be correlated with host were *fitB_1* (part of the toxin-antitoxin [TA] system, involved in in-host migration), *vbhT* (part of the TA system, interbacterial effector protein), *socA* (antitoxin to SocB, which inhibits DNA replication), and an HTH-type transcriptional regulator (others known in *X. fastidiosa* to modulate biofilm formation) (Table 2) (30–32).

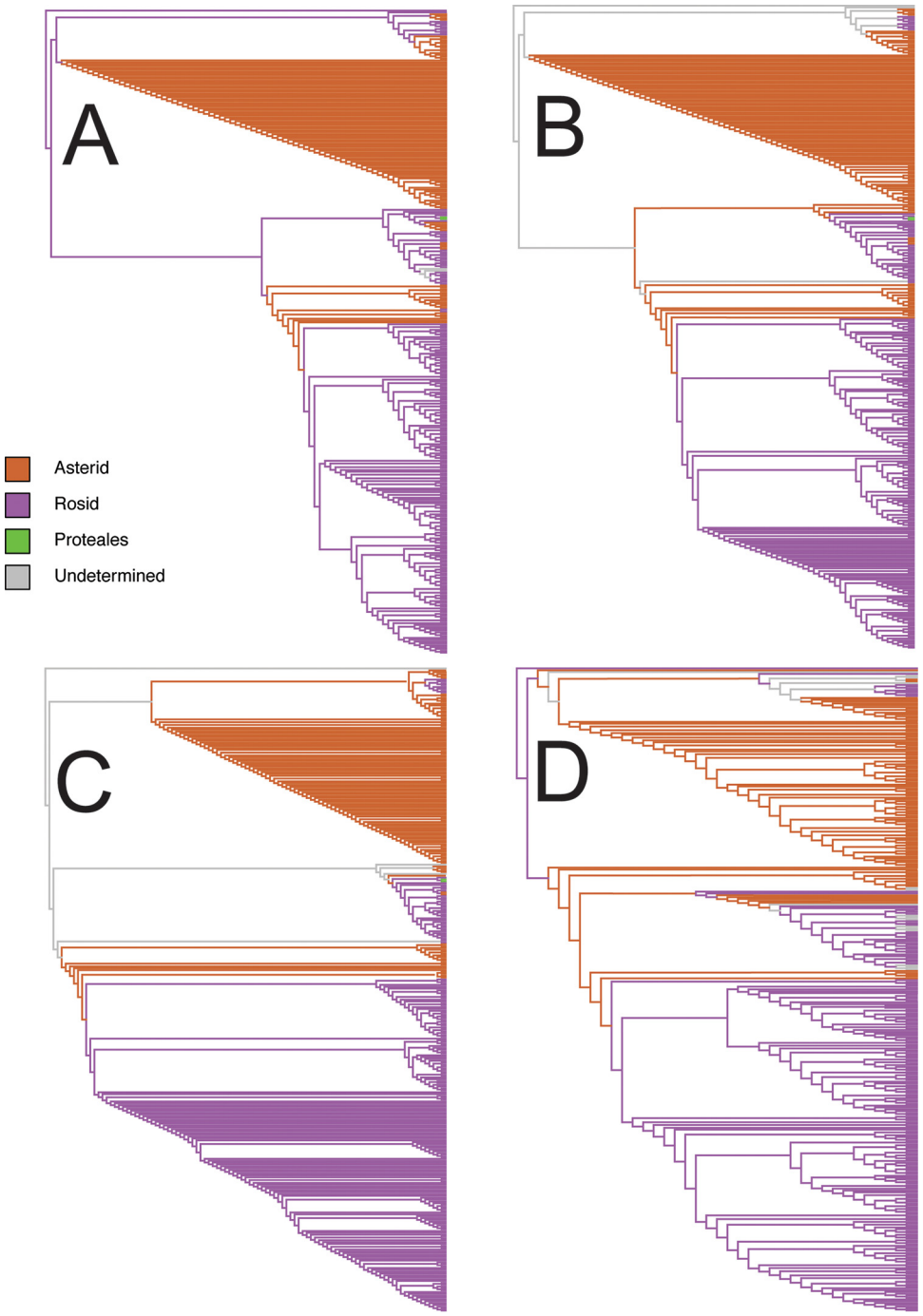

FIG 5 Cladograms of all four genomic regions with the most likely superorder or order of each node mapped onto the branches. Nodes with likelihood of less than 95% for one superorder are colored in gray and marked as undetermined. (A) Nonrecombinant core cladogram; (B) core cladogram; (C) MLST cladogram; (D) pan-genome cladogram. All ancestral state reconstructions were done with an equal rates transition model between hosts except the pan-genome, which performed better with additional parameters of the symmetrical rates model.

DISCUSSION

In this paper, we show that there is a genetic basis to the host range of *X. fastidiosa*. We demonstrate that both the phylogeny and gene gain and loss in the pan-genome are connected to plant host of the diverse species *X. fastidiosa* and that an Asterid of undetermined genus was the most likely ancestral plant host of *X. fastidiosa*. Our

TABLE 2 Significant genes whose presence/absence were correlated with host once phylogenetic history has been corrected for based on common descent

Host taxonomy	Gene	Annotation	No. of hosts in which gene is present	No. of nonhosts in which gene is present	No. of hosts in which gene is not present	No. of nonhost in which gene is not present	Naive *P* value	Tree-corrected *P* value
Asterid	group_943	Hypothetical protein	92	1	36	220	5.19E-53	3.42E-03
	group_949	IS200/IS605 family transposase IS609	13	2	115	219	7.09E-05	7.81E-03
	group_9702	Hypothetical protein	9	0	119	221	9.99E-05	1.56E-02
	group_1963	Hypothetical protein	9	0	119	221	9.99E-05	1.56E-02
	group_1865	Hypothetical protein	0	154	128	67	1.94E-45	3.13E-02
	group_2944	Hypothetical protein	0	41	128	180	2.88E-09	3.13E-02
	fitB_1	Toxin FitB	10	0	118	221	3.50E-05	3.13E-02
	group_2361	Hypothetical protein	10	0	118	221	3.50E-05	3.13E-02
	socA	Antitoxin SocA	30	6	98	215	1.77E-09	3.86E-02
	group_3382	Hypothetical protein	2	23	126	198	1.90E-03	3.91E-02
Rosid	group_943	Hypothetical protein	1	92	218	38	6.13E-52	3.42E-03
	group_949	IS200/IS605 family transposase IS609	2	13	217	117	8.62E-05	7.81E-03
	group_9702	Hypothetical protein	0	9	219	121	1.15E-04	1.56E-02
	group_1963	Hypothetical protein	0	9	219	121	1.15E-04	1.56E-02
	group_1865	Hypothetical protein	154	0	65	130	1.23E-46	3.13E-02
	group_2944	Hypothetical protein	41	0	178	130	1.29E-09	3.13E-02
	fitB_1	Toxin FitB	0	10	219	120	4.11E-05	3.13E-02
	group_2361	Hypothetical protein	0	10	219	120	4.11E-05	3.13E-02
	socA	Antitoxin SocA	6	30	213	100	2.84E-09	3.86E-02
	group_3382	Hypothetical protein	23	2	196	128	1.03E-03	3.91E-02
Vitis	group_3360	Hypothetical protein	15	33	160	141	5.15E-03	1.56E-02
	group_4565	Hypothetical protein	174	3	1	171	3.22E-96	3.13E-02
	group_1893	Hypothetical protein	63	0	112	174	2.30E-22	3.13E-02
	group_4682	Hypothetical protein	58	0	117	174	2.35E-20	3.13E-02
	group_2780	Putative HTH_type transcriptional regulator	0	36	175	138	1.73E-12	3.13E-02
	group_4923	Hypothetical protein	22	0	153	174	2.36E-07	3.13E-02
	group_3389	Hypothetical protein	11	0	164	174	8.30E-04	3.13E-02
	group_3387	Hypothetical protein	9	0	166	174	3.52E-03	3.13E-02
	group_4544	Hypothetical protein	104	50	71	124	8.70E-09	3.91E-02
Coffea	group_84	Hypothetical protein	19	203	1	126	1.49E-03	3.13E-02
	group_4414	Hypothetical protein	5	208	15	121	1.39E-03	3.91E-02
	vbhT	Adenosine monophosphate_protein transferase VbhT	5	207	15	122	1.43E-03	3.91E-02
	group_1250	Hypothetical protein	5	207	15	122	1.43E-03	3.91E-02

results indicate that the evolutionary trajectories of both the core and the pan-genomes allow for a bacterial species with an extensive host range to specialize many times over a broad array of plant hosts. We see this system as an example of one that "leaps," with host genera seemingly changing not via phylogenetic signal to related plant hosts but switching across large regions of plant host phylogenies (13). Prior to this study, we have not been able to trace a pattern of underlying genetic origins of host specificity in *X. fastidiosa*. In this way, our study shows that the phylogeny and gene gain/loss are connected to the adaptations that diversify host specificity in *X. fastidiosa*.

Phylogenies for MLST, pan-genome, core genome, and nonrecombinant core genome data were topologically similar, but not identical. While the subspecies relationships are not important to predicting host range, they are frequently used in management decisions and our ability to converse about outbreaks, so we are including our findings alongside our data on host use (Table S1). In terms of taxonomic subspecies, there are differences between the four trees in whether the two debated subspecies, *X.*

fastidiosa subsp. *morus* and *X. fastidiosa* subsp. *sandyi*, are contained within *X. fastidiosa* subsp. *fastidiosa* or *X. fastidiosa* subsp. *multiplex* or if they should be considered their own subspecies. While there are pairs of strains that are consistently close to each other, like the *X. fastidiosa* subsp. *morus* strains MulMD and Mul0034, the uncertainty in their position from phylogeny to phylogeny likely reflects large gaps in diversity that we have not yet sequenced or horizontal gene transfer more intensely affecting the pan-genome and particular genes used for MLST than the core genes, leading to issues recreating the vertical descent we aim for in a phylogeny (Fig. 1). *X. fastidiosa* subsp. *morus* has been documented to have up to 15.30% of its core genome undergoing intersubspecies homologous recombination, which could account for its uncertain placement in the four phylogenies (33). The two strains that have been described as *X. fastidiosa* subsp. *sandyi*-like, CO33 and CFBP8356, both clustered within *X. fastidiosa* subsp. *fastidiosa*, not with the other potential *X. fastidiosa* subsp. *sandyi* strains Ann1 and RAAR8_XF70, supporting previous work showing that there is not a strong distinction between *X. fastidiosa* subsp. *sandyi* and *X. fastidiosa* subsp. *fastidiosa* (Fig. 2) (34). The core genome tree also has very low bootstrap support for *X. fastidiosa* subsp. *pauca*, which is the most diverse and oldest of the three main subspecies that could be potentially due to conflicting histories between horizontal and vertical descent or, alternatively, reflect that this group is simply not well supported as one subspecies (35). In terms of the poor resolution in the OQDS clade, an analysis has recently been conducted to increase resolution within these strains (36). Given the diversity of *X. fastidiosa* subsp. *pauca*, the Hib4 strain, the outgroup of the subspecies, could be a potentially interesting strain in terms of both function and evolutionary history (33).

It is difficult to know which phylogenies are more accurate than others; however, we assume that the core genome is the most accurate at depicting the descent of this bacterial species, and the topology should be robust to even high levels of recombination (37). While the nonrecombinant core genome might reduce some issues with horizontal gene transfer, the lack of resolution because of too many identical sequences makes it difficult to use. While more data are not intrinsically better, there are known issues with the MLST genes used for *X. fastidiosa* phylogenetics, and having a larger set of unbiased homologous regions should be able to lend data to support nodes that are difficult to differentiate using the smaller MLST data set (29).

Using the core genome phylogeny, the most likely ancestral host was inferred from the phylogeny. These results show us that the phylogenetic history of *X. fastidiosa* is significantly correlated with the agricultural plant host from which the strains were isolated. While the core genome phylogeny depicts mainly vertical descent within this bacterial species, the pan-genome phylogeny likely combines vertical descent with horizontal gene transfer. This is due to the pan-genome's inclusion of the accessory genome, which are genes not shared by all members of the group (38). Based on this, we speculate that there is both adaptation and convergence depicted in these results. Potentially, both convergent horizontal descent via gene gain and loss, as well as vertical descent in the core, leads to our modern distribution of traits. While the ancestral state reconstruction did not show a classic host-parasite story of cospeciating or phylogenetically conserved host specificity, the phylogeny and gene presence/absence are predictive of the hosts from which the strains were isolated, and thus hypothetically, host specificity as well.

While the four ancestral state reconstructions do not show identical histories, they all infer a high likelihood of ancestral hosts at many key branch points of the three subspecies. The pan- and core genome reconstructions predict the genus *Vaccinium* (based on isolates from blueberry) as the most likely ancestral host of the *X. fastidiosa* subsp. *multiplex*, which supports the overall reliability of the reconstruction, as blueberry, like *X. fastidiosa* subsp. *multiplex*, is native to eastern North America (39). *X. fastidiosa* subsp. *pauca*, *X. fastidiosa* subsp. *multiplex*, and *X. fastidiosa* subsp. *fastidiosa* all exhibit host shifts from another genus to *Prunus*, suggesting potential for increased vulnerability in this genus to infection from varied alternative hosts. All four reconstructions also support the genus *Coffea* as the most likely ancestor of the introduced *X. fastidiosa* subsp. *fastidiosa*

strains from Central American to California. This supports a previous hypothesis made by Nunney et al. (40) wherein coffee plants that were imported from Central America to southern California in the mid-1800s might have brought *X. fastidiosa* subsp. *fastidiosa* along with them. Given the potential role of imported *Coffea* in devastating global outbreaks of disease caused by *X. fastidiosa* (California and Italy) (41), it should be much more carefully monitored or restricted in global trade. Given the current policy emphasis on eradication and trade restrictions, it is vital to identify genera such as *Coffea* that are especially relevant to global outbreaks and that should be monitored carefully. The relationship between *X. fastidiosa* and *Coffea* should be further explored as a model host to aid our understanding of the molecular mechanisms of this complex interaction. A potential alternative hypothesis for these nodes could also be that *Coffea* and *Vaccinium* are permissive hosts. From a parsimony perspective, they could be akin to "universal hosts" so that it takes very little change for *X. fastidiosa* strains to switch to *Coffea* or *Vaccinium* from other infected plants. This could be investigated by further interrogating the genes shown to be uniquely absent in *Coffea*-infecting strains. Phylogenetically, this would reflect deep homology in which the underlying genetic framework of the pathogens makes it easy to shift from other plant hosts to *Coffea* or *Vaccinium* (42).

The two plant genera with genes significantly correlated with them, *Vitis* and *Coffea*, had 179 and 20 whole-genome sequences from diverse sampling regions, respectively. The larger clades of Proteales, Asterid, and Rosid were also used to look for convergent gene presence and absence, and again, the two groups with the majority of samples, Asterid ($n = 126$) and Rosid ($n = 194$), had genes correlated with them, while Proteales ($n = 2$) did not. The genes found to be correlated with these host groupings had varied functions. Unfortunately, out of these 23 genes, 20 are hypothetical proteins; the ones with known functions could have very interesting implications for host range. *fitB_1* has been known to be involved in in-host migration and metal binding; similar genes are also frequently gained and lost in other *Xanthomonadaceae* and are hypothesized to affect both gene regulation and resistance mechanisms (43). *vhbT* is an interbacterial effector protein, facilitating bacterial conjugation, another process with potential for large genomic and functional changes (31). Another significant gene (group_2780) contains a helix-turn-helix region, a DNA binding domain that has been found to control metal resistance bacteria generally and biofilm growth in *X. fastidiosa* specifically (32). These genes should be explored further through fitness tests with the presence and absence of these nonessential accessory genes in multiple-host environments to further evaluate if their presence and absence is adaptive or due to drift.

Future research pertaining to host range should focus on both convergent gene gain and loss, as well as the adaptive vertically descended genetics underlying host range. As both genomic assays have identified the pan-genome to be linked to host association, it would be beneficial to our understanding of host specificity to pursue this further. This study has identified a group of candidate genes associated with particular hosts, and they can be tested in the lab to determine if they are significantly linked to fitness in their particular hosts. The study has also identified *Coffea* as an especially relevant host in global plant trade in terms of spreading infection across borders and oceans. Using these data, we can start identifying patterns of likely host shifts that can help make decisions on when eradication and quarantine are necessary based on the historical likelihood of host shifts. However, we should also carry out further whole-genome sequencing of strains outside the classic agricultural settings. To truly understand a biological system, we not only need to understand the relevant biological components but also how they interact both inside and outside agricultural landscapes.

MATERIALS AND METHODS

Whole-genome sequence data set. A total of 349 *Xylella* species genome sequences were used in this study, either downloaded from GenBank (44) or assembled *de novo* in-house from published FASTQ reads (Table S1 in the supplemental material). *De novo* sequences were aligned as described in Castillo et al. (45), and contigs were mapped to complete genomes of each subspecies using Mauve's contig

mover function (46). Three novel genomes are being presented in this study. *X. fastidiosa* subsp. *fastidiosa* scaffolds were reordered using the Temecula1 assembly (GCA_000007245.1), *X. fastidiosa* subsp. *pauca* scaffolds were mapped using the 9a5c assembly (ASM672v1), and *X. fastidiosa* subsp. *multiplex* scaffolds were reordered using the sequence of strain M23. Draft genomes, as well as downloaded sequences, were then annotated using Prokka (47).

Phylogenetics. The first step in creating all phylogenies was building a nucleotide or gene alignment of the genomic regions of interest. Four alignments were created all using the same set of taxa (Table S1 in the supplemental material), a core genome alignment, nonrecombinant core alignment, a multilocus sequence type (MLST) alignment, and a pan-genome alignment.

The core genome was built with Roary (48) to identify nucleotide regions (genes or hypothetical proteins) shared by at least 99% of all taxa. We ran Roary with the parameters -s -ap to cluster paralogs and allow them in the core genome. The nonrecombinant core alignment was based on the core genome, but recombinant sites identified with ClonalFrameML were removed from the alignment using an in-house R script (27). The MLST alignment was based on a nucleotide alignment of the 7 MLST housekeeping genes commonly used for *X. fastidiosa* (*petC*, *nuoL*, *malF*, *leuA*, *holC*, *gltT*, and *cysG*) with reference sequences acquired from the *X. fastidiosa* MLST database (49, 50). We then searched each MLST reference sequence against all whole genomes (Table S1) using the Basic Local Alignment Search Tool (BLAST) at an E value of 10^{-3} in BLAST+, with a database created for each whole genome (51). We concatenated all MLST gene sequences for individual taxa and aligned them to all other taxa using MAFFT v7 (52). The pan-genome alignment was made using Roary's gene presence-absence output by constructing a matrix of all genes as characters with binary presence or absence of that gene in a strain as the character state. As each character represented a known genetic region and there were no gaps in this matrix, no additional alignment algorithm was used. In total, this alignment contained 17,024 characters, representing the 17,024 total genes that make up the pan-genome (every gene present in any strain) of *Xylella* species sequences. The outgroup used for all trees was *Xylella taiwanensis* strain Wufong1 isolated in Taiwan in 2014 from *Pyrus pyrifolia* (53).

We constructed four maximum-likelihood phylogenies using RAxML v8.2.11 (54) under a generalized time-reversible model. Node support was measured with 1,000 bootstrap replicates (35). Trees were visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and the Interactive Tree of Life (55). Phylogenetic diversity was calculated as the summation of total branch lengths for each phylogeny (not including the outgroup) using the R package adephylo (56–58).

Ancestral state reconstructions. To conduct ancestral state reconstructions, we used an extant distribution of characters (heritable traits of interest), in this case, the genera of plants from which we isolated the bacteria. Using that distribution, we constructed the most likely history of hosts across the phylogeny at all internal (ancestral) nodes. We are, in essence, seeking parameter values that maximize the probability of the data (the observed character states) given the hypothesis (a model of character evolution and a phylogeny relating the observed sequences or taxa). Based on available data, the identity of the host plant from which each strain was isolated in the field is identifiable to at least the genus level. This value is used as a point proxy for the true state of interest, potential host range. Since the host range must be experimentally determined, in this study, we use the host from which each strain was isolated as a point representative of an unknown range of susceptibility. Due to this, any subsequent results cannot infer specificity to a given host but imply the ability to infect said host. Because sampling is heavily biased toward symptomatic agricultural crops in the case of *X. fastidiosa*, we interpret each ancestral state as the most likely agriculturally relevant host that the pathogen would have been isolated from.

All taxa were coded based on plant host genus and super order/order (the deepest clade grouping that combined our genera into more than one group). This included 2 superorders (Asterid and Rosid), 1 order (*Proteales*), and 26 genera that were potential hosts for *X. fastidiosa*'s hypothetical ancestors at each internal node of the phylogenies. The marginal ancestral state likelihood estimates of each host for all internal nodes of the ML phylogenetic trees were calculated using the rerooting method of Yang et al. (59) in the R package phytools (60) and mapped using the package APE (61). This method uses the phylogeny of extant taxa to reconstruct ancestral traits of extinct ancestors by analyzing phylogenetic parameters (topology and branch length), along with a model of nucleotide substitution, to build posterior probabilities of character states at each interior node by randomly rerooting the tree at each internal node and calculating the probability of observing the extant distribution of traits over all possibilities of that internal node character identity. The ML estimates at each internal node were calculated based on both the equal rates transition model (i.e., fixed rate of change between any two hosts) and the symmetrical rates transition model (i.e., fixed rates of host change symmetrically pairwise between hosts, but not between all hosts). The fit of the two models to the data was compared using the Akaike information criterion and can be seen in Table S2 (62).

Correlation between host and gene presence/absence. Information on plant host taxonomy was gathered on NCBI's Taxonomy Browser (63). Scoary (64) was used to test if the pan-genome was correlated with hosts at either the superorder scales or the genus scale by conducting a Fisher's exact test (FET) (65). FET measures the association of each gene in the pan-genome to a trait of interest, which, in this case, is plant host. While FET requires no association between data points, Scoary uses a phylogeny in order to remove lineage-specific interdependencies and corrects the *P* value based on those interdependencies. Significance was evaluated by the "worst pairwise comparison *P*" for the phylogenetic corrections, not the naive *P* values from FET. Individual analyses were conducted to test for correlation of gene presence and absence with each of the 29 coded host groups (26 genera, Asterid, Rosid, and Proteales).

Data availability. Genomic data are available in GenBank Nucleotide Database with BioSample accession numbers in Table S1.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 7.7 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.03 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.3 MB.

ACKNOWLEDGMENTS

We acknowledge and thank Andreina Castillo Siri, Mathieu Vanhove, Adam Zeilinger, Michael Voeltz, Dylan Beal, Anne Sicard, Aidee Guzman, and Kirk Pearson for helpful discussions, ideas, edits, and support. We also thank Monica Donegan, Elizabeth Clark, and Isabel Bojanini who have helped with final revisions.

This work was supported by the California Department of Food and Agriculture PD/GWSS Research Program and the European Union's Horizon 2020 research program (XF-ACTORS). A.K. received support from the UC Berkeley William Carroll Smith Plant Pathology Fellowship.

REFERENCES

1. Hulme PE. 2009. Trade, transport and trouble: managing invasive species pathways in an era of globalization. J Appl Ecol 46:10–18. https://doi.org/10.1111/j.1365-2664.2008.01600.x.
2. Lockwood JL, Cassey P, Blackburn T. 2005. The role of propagule pressure in explaining species invasions. Trends Ecol Evol 20:223–228. https://doi.org/10.1016/j.tree.2005.02.004.
3. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F, Rangel Alvarez B, Lundberg DS, Lu TY, Lebeis S, Jin Z, McDonald M, Klein AP, Feltcher ME, Rio TG, Grant SR, Doty SL, Ley RE, Zhao B, Venturi V, Pelletier DA, Vorholt JA, Tringe SG, Woyke T, Dangl JL. 2017. Genomic features of bacterial adaptation to plants. Nat Genet 50:138–150. https://doi.org/10.1038/s41588-017-0012-9.
4. Woolhouse MEJ, Haydon DT, Antia R. 2005. Emerging pathogens: the epidemiology and evolution of species jumps. Trends Ecol Evol 20:238–244. https://doi.org/10.1016/j.tree.2005.02.009.
5. Huang W, Reyes-Caldas P, Mann M, Seifbarghi S, Kahn A, Almeida RPP, Béven L, Heck M, Hogenhout SA, Coaker G. 2020. Bacterial vector-borne plant diseases: unanswered questions and future directions. Mol Plant 13:1379–1393. https://doi.org/10.1016/j.molp.2020.08.010.
6. European Food Safety Authority (EFSA). 2018. Update of the Xylella spp. host plant database. EFSA J 16:e05408. https://doi.org/10.2903/j.efsa.2018.5408.
7. Sicard A, Zeilinger AR, Vanhove M, Schartel TE, Beal DJ, Daugherty MP, Almeida RPP. 2018. Xylella fastidiosa: insights into an emerging plant pathogen. Annu Rev Phytopathol 56:181–202. https://doi.org/10.1146/annurev-phyto-080417-045849.
8. Nunney L, Azad H, Stouthamer R. 2019. An experimental test of the host-plant range of nonrecombinant strains of North American Xylella fastidiosa subsp. multiplex. Phytopathology 109:294–300. https://doi.org/10.1094/PHYTO-07-18-0252-FI.
9. Sanderlin RS. 2017. Host specificity of pecan strains of *Xylella fastidiosa* subsp. *multiplex*. Plant Dis 101:744–750. https://doi.org/10.1094/PDIS-07-16-1005-RE.
10. Barrett LG, Kniskern JM, Bodenhausen N, Zhang W, Bergelson J. 2009. Continua of specificity and virulence in plant host-pathogen interactions: causes and consequences. New Phytol 183:513–529. https://doi.org/10.1111/j.1469-8137.2009.02927.x.
11. Almeida RPP, Nascimento FE, Chau J, Prado SS, Tsai CW, Lopes SA, Lopes JRS. 2008. Genetic structure and biology of Xylella fastidiosa strains causing disease in citrus and coffee in Brazil. Appl Environ Microbiol 74:3690–3701. https://doi.org/10.1128/AEM.02388-07.
12. Purcell AH, Almeida RPP, Purcell AH. 2003. Biological traits of Xylella fastidiosa strains from grapes and almonds. Appl Environ Microbiol 69:7447–7452. https://doi.org/10.1128/AEM.69.12.7447-7452.2003.
13. Park AW, Farrell MJ, Schmidt JP, Huang S, Dallas TA, Pappalardo P, Drake JM, Stephens PR, Poulin R, Nunn CL, Davies TJ. 2018. Characterizing the phylogenetic specialism-generalism spectrum of mammal parasites. Proc R Soc Lond B Biol Sci 285:20172613. https://doi.org/10.1098/rspb.2017.2613.
14. EFSA Panel on Plant Health. 2015. Scientific Opinion on the risks to plant health posed by Xylella fastidiosa in the EU territory, with the identification and evaluation of risk reduction options. EFSA J 13:3989. https://doi.org/10.2903/j.efsa.2015.3989.
15. Amanifar N, Babaei G, Mohammadi AH. 2019. Xylella fastidiosa causes leaf scorch of pistachio (Pistacia vera) in Iran. Phytopathol Mediterr 58:369–378.
16. Della Coletta-Filho H, Castillo AI, Laranjeira FF, de Andrade EC, Silva NT, de Souza AA, Bossi ME, Almeida RPP, Lopes JRS. 2020. Citrus variegated chlorosis: an overview of 30 years of research and disease management. Trop Plant Pathol 45:175–191. https://doi.org/10.1007/s40858-020-00358-5.
17. Rodriguez-R LM, Grajales A, Restrepo S, Bernal A, Salazar C, Arrieta-Ortiz M. 2012. Genomes-based phylogeny of the genus Xanthomonas. BMC Microbiol 12:43. https://doi.org/10.1186/1471-2180-12-43.
18. Szurek B, Couloux A, Poussier S, Koebnik R, Pieretti I, Carrere S, Rott P, Mangenot S, Arlat M, Gouzy J, Darrasse A, Verdier V, Cociancich S, Jacques M-A, Lauber E, Segurens B, Royer M, Barbe V, Manceau C. 2009. The complete genome sequence of Xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. BMC Genomics 10:616. https://doi.org/10.1186/1471-2164-10-616.
19. Galán J, Collmer A. 1999. Type III secretion machines: bacterial devices for protein delivery into host cells. Science 284:1322–1329. https://doi.org/10.1126/science.284.5418.1322.
20. Chatterjee S, Almeida RPP, Lindow S. 2008. Living in two worlds: the plant and insect lifestyles of *Xylella fastidiosa*. Annu Rev Phytopathol 46:243–271. https://doi.org/10.1146/annurev.phyto.45.062806.094342.
21. Hajri A, Brin C, Hunault G, Lardeux F, Lemaire C, Manceau C, Boureau T, Poussier S. 2009. A «repertoire for repertoire» hypothesis: repertoires of type three effectors are candidate determinants of host specificity in Xanthomonas. PLoS One 4:e6632. https://doi.org/10.1371/journal.pone.0006632.
22. Killiny N, Almeida RPP. 2011. Gene regulation mediates host specificity of a bacterial pathogen. Environ Microbiol Rep 3:791–797. https://doi.org/10.1111/j.1758-2229.2011.00288.x.
23. Rapicavoli JN, Blanco-Ulate B, Muszyński A, Figueroa-Balderas R, Morales-Cruz A, Azadi P, Dobruchowska JM, Castro C, Cantu D, Roper MC. 2018. Lipopolysaccharide O-antigen delays plant innate immune recognition of Xylella fastidiosa. Nat Commun 9:390. https://doi.org/10.1038/s41467-018-02861-5.
24. Zipfel C, Felix G. 2005. Plants and animals: a different taste for microbes? Curr Opin Plant Biol 8:353–360. https://doi.org/10.1016/j.pbi.2005.05.004.
25. Navaud O, Barbacci A, Taylor A, Clarkson JP, Raffaele S. 2018. Shifts in diversification rates and host jump frequencies shaped the diversity of host range among Sclerotiniaceae fungal plant pathogens. Mol Ecol 27:1309–1323. https://doi.org/10.1111/mec.14523.
26. Razo-Mendivil U, Pérez-Ponce de León G. 2011. Testing the evolutionary and biogeographical history of Glypthelmins (Digenea: plagiorchiida), a parasite of anurans, through a simultaneous analysis of molecular and morphological data. Mol Phylogenet Evol 59:331–341. https://doi.org/10.1016/j.ympev.2011.02.018.
27. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041.
28. Castillo AI, Bojanini I, Chen H, Kandel PP, De La Fuente L, Almeida RPP. 2021. Allopatric plant pathogen population divergence following disease

emergence. Appl Environ Microbiol 87:1–19. https://doi.org/10.1128/AEM.02095-20.
29. Landa BB, Castillo AI, Giampetruzzi A, Kahn A, Román-Écija M, Velasco-Amo MP, Navas-Cortés JA, Marco-Noales E, Barbé S, Moralejo E, Coletta-Filho HD, Saldarelli P, Saponari M, Almeida RPP. 2020. Emergence of a plant pathogen in Europe associated with multiple intercontinental introductions. Appl Environ Microbiol 86:1–15. https://doi.org/10.1128/AEM.01521-19.
30. Aakre CD, Phung TN, Huang D, Laub MT. 2013. A bacterial toxin inhibits DNA replication elongation through a direct interaction with the β sliding clamp. Mol Cell 52:617–628. https://doi.org/10.1016/j.molcel.2013.10.014.
31. Harms A, Liesch M, Körner J, Québatte M, Engel P, Dehio C. 2017. A bacterial toxin-antitoxin module is the origin of inter-bacterial and inter-kingdom effectors of Bartonella. PLoS Genet 13:e1007077. https://doi.org/10.1371/journal.pgen.1007077.
32. Barbosa RL, Benedetti CE. 2007. BigR, a transcriptional repressor from plant-associated bacteria, regulates an operon implicated in biofilm growth. J Bacteriol 189:6185–6194. https://doi.org/10.1128/JB.00331-07.
33. Vanhove M, Retchless AC, Sicard A, Rieux A, Coletta-Filho HD, Fuente LD, La Stenger DC, Almeida PP, De La Fuente L, Stenger DC, Almeida RPP, La Fuente LD, Stenger DC, Almeida PP. 2019. Genomic diversity and recombination among Xylella fastidiosa subspecies. Appl Environ Microbiol 85:1–17. https://doi.org/10.1128/AEM.02972-18.
34. Denancé N, Briand M, Gaborieau R, Gaillard S, Jacques MA. 2019. Identification of genetic relationships and subspecies signatures in Xylella fastidiosa. BMC Genomics 20:1–21. https://doi.org/10.1186/s12864-019-5565-9.
35. Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791. https://doi.org/10.2307/2408678.
36. Sicard A, Saponari M, Vanhove M, Castillo AI, Giampetruzzi A, Loconsole G, Saldarelli P, Boscia D, Neema C, Almeida RPP. 2021. Introduction and adaptation of an emerging pathogen to olive trees in Italy. Microb Genom 7:000735. https://doi.org/10.1099/mgen.0.000735.
37. Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. mBio 5:e02158-14. https://doi.org/10.1128/mBio.02158-14.
38. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482. https://doi.org/10.1038/nrg3962.
39. Wood GW. 2008. The wild blueberry industry—past. Small Fruits Rev 3:11–18. https://doi.org/10.1300/J301v03n01_03.
40. Nunney L, Yuan X, Bromley R, Hartung J, Montero-Astúa M, Moreira L, Ortiz B, Stouthamer R. 2010. Population genomic analysis of a bacterial plant pathogen: novel insight into the origin of pierce's disease of grapevine in the U.S. PLoS One 5:e15488. https://doi.org/10.1371/journal.pone.0015488.
41. Marcelletti S, Scortichini M. 2016. Xylella fastidiosa CoDIRO strain associated with the olive quick decline syndrome in southern Italy belongs to a clonal complex of the subspecies pauca that evolved in Central America. Microbiology (Reading) 162:2087–2098. https://doi.org/10.1099/mic.0.000388.
42. Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. Nature 457:818–823. https://doi.org/10.1038/nature07891.
43. Martins PMM, Machado MA, Silva NV, Takita MA, De Souza AA. 2016. Type II toxin-antitoxin distribution and adaptive aspects on Xanthomonas genomes: focus on Xanthomonas citri. Front Microbiol 7:652. https://doi.org/10.3389/fmicb.2016.00652.
44. Benson D. a, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2010. GenBank. Nucleic Acids Res 38:D46–D51. https://doi.org/10.1093/nar/gkp1024.
45. Castillo AI, Chacón-Díaz C, Rodríguez-Murillo N, Coletta HD, Almeida RPP, Rica C, Diaz CC, Colleta-Filho H, Almeida RPP. 2020. Impacts of local population history and ecology on the evolution of a globally dispersed pathogen. BMC Genom 21:369. https://doi.org/10.1186/s12864-020-06778-6.
46. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. 2009. Reordering contigs of draft genomes using the Mauve Aligner. Bioinformatics 25:2071–2073. https://doi.org/10.1093/bioinformatics/btp356.
47. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
48. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.
49. Jolley KA, Chan MS, Maiden MCJ. 2004. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics 5:86. https://doi.org/10.1186/1471-2105-5-86.
50. Scally M, Schuenzel EL, 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. Appl Environ Microbiol 71:8491–8499. https://doi.org/10.1128/AEM.71.12.8491-8499.2005.
51. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:1–9. https://doi.org/10.1186/1471-2105-10-421.
52. Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518. https://doi.org/10.1093/nar/gki198.
53. 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 Announc 2:2–3. https://doi.org/10.1128/genomeA.00166-14.
54. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
55. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
56. Jombart T, Balloux F, Dray S. 2010. Adephylo: exploratory analyses for the phylogenetic comparative method. Bioinformatics 26:1907–1909. https://doi.org/10.1093/bioinformatics/btq292.
57. Faith DP. 1992. Conservation evaluation and phylogenetic diversity. Biol Conserv 61:1–10. https://doi.org/10.1016/0006-3207(92)91201-3.
58. R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
59. Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650. https://doi.org/10.1093/genetics/141.4.1641.
60. Revell LJ. 2016. Package 'phytools.' https://cran.r-project.org/web/packages/phytools/phytools.pdf.
61. Paradis E, Schliep K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528. https://doi.org/10.1093/bioinformatics/bty633.
62. Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723. https://doi.org/10.1109/TAC.1974.1100705.
63. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, McVeigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020:1–21. https://doi.org/10.1093/database/baaa062.
64. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 17:1–9. https://doi.org/10.1186/s13059-016-1108-8.
65. Fisher RA. 1934. Statistical methods for research workers (5th ed). Oliver and Boyd: Edinburgh.