AEM Accepted Manuscript Posted Online 22 January 2021
Appl Environ Microbiol doi:10.1128/AEM.02095-20

Allopatric plant pathogen population divergence following disease emergence

Andreina I. Castillo [1], Isabel Bojanini [1], Hongyu Chen[2, 3], Prem P. Kandel[2, 4], Leonardo De La Fuente [2], and Rodrigo P.P. Almeida [1]

[1] Department of Environmental Science, Policy and Management, University of California, Berkeley, CA, USA.

[2] Department of Entomology and Plant Pathology, Auburn University, Auburn, Alabama, USA.

[3] Current address: Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, North Carolina, USA.

[4] Current address: Department of Plant Pathology and Environmental Microbiology, PennState University, University Park, Pennsylvania, USA.

Corresponding author: rodrigoalmeida@berkeley.edu

ABSTRACT

Within the landscape of globally distributed pathogens, populations differentiate via both adaptive and non-adaptive forces. Individual populations are likely to show unique trends of genetic diversity, host-pathogen interaction, and ecological adaptation. In plant pathogens, allopatric divergence may occur particularly rapidly within simplified agricultural monoculture landscapes. As such, the study of plant pathogen populations in monocultures can highlight the distinct evolutionary mechanisms that lead to local genetic differentiation. *Xylella fastidiosa* is a plant pathogen known to infect and damage multiple monocultures worldwide. One subspecies, *Xylella fastidiosa* subsp. *fastidiosa* was first introduced to the USA ~150 years ago, where it was found to infect and cause disease in grapevines (Pierce's disease of grapevines, PD). Here, we studied PD-causing subsp. *fastidiosa* populations, with an emphasis on those found in the USA. Our study shows that following its establishment in the USA, PD-causing strains likely split into populations in the East and West Coast. This diversification has occurred via both changes in gene content (gene gain/loss events) and variations in nucleotide sequence (mutation and recombination). In addition, we reinforce the notion that PD-causing populations within the USA acted as the source for subsequent subsp. *fastidiosa* outbreaks in Europe and Asia.

IMPORTANCE

Compared to natural environments, the reduced diversity of monoculture agricultural landscapes can lead bacterial plant pathogens to quickly adapt to local biological and ecological conditions. Because of this, accidental introductions of microbial pathogens into naïve regions represents a significant economic and environmental threat. *Xylella fastidiosa* is a plant pathogen with an expanding host and geographic range due to multiple intra- and inter-continental introductions.

X. fastidiosa subsp. *fastidiosa*, infects and causes disease in grapevines (Pierce's disease of grapevines; PD). This study focused on PD-causing *X. fastidiosa* populations, particularly those found in the USA but also invasions into Taiwan and Spain. The analysis shows that PD-causing *X. fastidiosa* has diversified via multiple co-occurring evolutionary forces acting at an intra- and inter-population level. This analysis enables a better understating of the mechanisms leading to the local adaptation of *X. fastidiosa*, and how a plant pathogen diverges allopatrically after multiple and sequential introduction events.

INTRODUCTION

The worldwide distribution of microbial plant pathogens is constantly shifting. Global trade and movement of infected plant material enables pathogen introductions from native and endemic aeras to naïve regions (1, 2). Likewise, the intentional introduction of non-native plant species of agronomic and ornamental value to novel environments facilitates the host range expansion of endemic pathogens (3, 4). One crucial factor in the formation of novel plant-pathogen associations is the amount of genetic diversity on which natural selection can act, in other words, the adaptive potential (5). Differences in adaptive potential between host and microbial populations can have a significant role in determining the host and geographic range of a pathogen. For instance, in the case of plant pathogens, higher genetic diversity in effector proteins and virulence genes has a positive effect on host range (6–9). Alternatively, multiple studies have highlighted how reduced genetic diversity in plant hosts can enhance the spread of pathogens within a population (10–12).

Factors that influence genetic diversity, whether via the action of distinct evolutionary mechanisms (13, 14) or as a product of ecological and evolutionary history, affect adaptive potential (15). In plant pathogens, geographical and ecological specialization have been frequently described (16, 17). This is partly explained by plant pathogen differentiation and specialization occurring rapidly within agricultural systems (14, 18, 19). Overall, it is expected that in the absence of gene flow, plant pathogens of agricultural crops will rapidly adapt to local environmental, ecological, and biological conditions (20, 21). Therefore, understanding the mechanisms leading to pathogen adaptation, either to a new crop or environmental condition, has great relevance in developing effective management and control strategies (22). This is particularly pertinent in plant pathogens with a proven capacity to adapt to multiple crops as well as having an expanding geographic and host range. This is the case of the emerging pathogen *Xylella fastidiosa* (23).

The bacterial species *X. fastidiosa* has been reported to infect 563 plant species from 82 distinct botanical families (23). However, the host range of *X. fastidiosa* varies among and within described subspecies and phylogenetic clades (24). The geographic distribution of the three main *X. fastidiosa* subspecies is also unique, with most of them having experienced one or several dispersal and establishment events at the continental scale. For this reason, efficient identification and tracking of *X. fastidiosa* subspecies has important implications for the development of adequate disease control and mitigation strategies (25, 26). Three *X. fastidiosa* subspecies have an ancestrally allopatric range that has recently expanded: *X. fastidiosa* subsp. *multiplex* is native to temperate and subtropical North America (27, 28), and has been introduced multiple times into Europe (29); *X. fastidiosa* subsp. *pauca* is native to South America (28) but has been recently reported in the Apulian region in Italy and in Costa Rica (30, 31); finally, *X. fastidiosa* subsp. *fastidiosa* is native to Central America (32, 33), and was introduced to the USA (24, 34), and subsequently to Europe (35) and Taiwan (36). Other non-monophyletic but proposed subspecies include *X. fastidiosa* subsp. *sandyi*, found in Southern regions of the United States (37, 38) and also introduced into Europe (39); and *X. fastidiosa* subsp. *morus*, only found in regions were subsp. *multiplex* and subsp. *fastidiosa* co-occur (24, 40).

The hypothesis that subsp. *fastidiosa* was introduced once to the United States (USA) ~150 years ago leading to the emergence of Pierce’s Disease of grapevines (PD) is well supported (24, 33, 41). PD is a grapevine malady that results in significant economic losses to the wine industry in California (42) and the Southeast USA (43). Current knowledge of the evolution of subsp. *fastidiosa* suggests that the ability to infect grapevines was acquired after its introduction to the USA (33). Furthermore, there is evidence that local adaptation to environmental factors has occurred in grape-infecting isolates across a latitudinal gradient in California (34). Finally, available genomic and MLSA-E data suggest that PD-causing isolates in the West and East Coast of the USA are phylogenetically distinct (34, 44)

These studies are indicative that after its introduction and establishment in the USA, the subsp. *fastidiosa* clade causing disease in grapevines dispersed to different geographic regions and diversified genetically to adapt to a range of biotic and abiotic conditions. To better understand how *X. fastidiosa* evolved with the emergence of a novel plant disease (PD) and diversified in allopatry in different regions of the USA, we studied populations of the pathogen from the USA and abroad. We evaluated the evolutionary relationship between both USA populations and their relationship with recent introduction events derived from them (i.e. introductions to Spain and Taiwan associated with the emergence of PD in those regions). In addition, we identified the evolutionary mechanisms facilitating population diversification by defining intra-population patterns of gene gain/loss, intra-subspecific recombination, and nucleotide diversity.

RESULTS

PD isolates are split into regional clades within the USA, with Europe and Asia introductions originating from these regions. We arbitrarily split grapevine isolates into 3 phylogenetically supported clades, PD-I to -III (Fig.1). These phylogenetically supported clades were also observed in the non-recombinant phylogenetic tree (Fig. S1). PD-I only included isolates from the Southeast USA; PD-II and PD-III were dominated by California isolates, but at the base of those clades there was one isolate from Texas (PD-II) and a sister clade from Georgia

(PD-III). No isolates from California grapevines clustered within the PD-I clade. From the data available alone, it is not possible to infer the dispersal history of the non-California isolates in PD-II and PD-III (i.e. basal sister clades or introductions to California). Isolates from Taiwan were phylogenetically placed within the PD-I clade, while those from Spain were nested in the PD-II clade. These represent two distinct introductions, originating from different regions in the USA. Isolates from the same geographic region tended to cluster together within each major clade. For instance, in the PD-I clade, most Georgia isolates from Site1 (i.e. 14B1, 14B4, 14B6, 16B2, 15B2, 14B3, and 16B4) and Site2 (i.e. 16M5, 16M6, 16M7, 16M8, and 16M9) clustered together. Isolates from each site formed separate subclades within this group (Fig. 2). Other Georgia isolates from Site1 (i.e. 14B2, 14B5, 14B7, 16B1, 16B3, 16B5, and 16B6) were more closely related to those from Florida and North Carolina. In a similar manner, isolates from the West Coast (i.e. California) tended to group geographically. Specifically, isolates obtained from Southern California (i.e. Je81, Je104, Je112, Je110, etc.) were ancestral to those from Northern California (i.e. Hopland, Stag Leap, Conn-Creek, CV17-3, Je65, Je73, etc.) in the PD-III clade.

A total of 141 different haplotypes named using roman numerals (I-CXIV) (Fig. 1a) were found in the PD-causing core genome alignment. Haplotypes were structured by geographic location and largely matched the evolutionary relationships observed in phylogenetic analyses (Fig. 1b,2). Overall, haplotypes were grouped similarly to the phylogenetic analyses. Isolates originating from the West and East Coast were split by 979 mutations. California had the largest number of haplotypes (106) as well as haplotypes with the highest frequency: XXVIII (7), XLIV (6), XXXIV (5), LVIII (4), LXVI (4), LXXIII (3), and XCIV (3). On the other hand, Southeast USA haplotypes (31) were generally found in low frequency (i.e. one or two isolates). In addition, Southeast isolates in PD-III formed a distinct group separated from the California group by 243 mutations. Likewise, GB514 (Texas, PD-II) was closely connected to California isolates, from which it differentiated by 159 mutations. Isolates originating from recent introduction events (i.e. Spain and Taiwan) had unique haplotypes. Spanish associated haplotypes were linked to a haplotype originating from California (PD-II) and were differentiated by 61 mutations. Similarly, the Taiwan haplotypes were closely linked to the haplotype group originating from Southeast USA (PD-I) and differentiated by 13 mutations.

Gene gain and loss events occur following subsp. *fastidiosa* introduction events. Estimated rates of gene gain/loss were highest in branches leading to the introduction of subsp. *fastidiosa* from Central America. Furthermore, a total of 35 core genes were absent in the PD-causing population compared to the Costa Rican isolates, while 49 core genes were present in the PD-causing population but absent in the Costa Rican isolates. In addition, gene gain/loss events also occurred within the USA populations. In California (Fig. S2a) gene gain/loss rates were highest in the branches leading to each cluster than within clusters, but PD-III had higher gene gain/loss rates compared to PD-II. Likewise, two clades were observed within the Southeast USA population (Fig. S2b). The first clade was formed by isolates 16M2, 16M3, XF51_CCPM1 (from Georgia, clustering with PD-III), and GB514 (from Texas, clustering with PD-II); and the second by the remaining Southeast USA isolates (PD-I).

Some unique genes were identified through estimating gene gain/loss rates within each population. We found that, when considering geographical origins of isolates alone, gene presence/absence was similar in PD-II and PD-III isolates regardless of geographical origin (Fig. 3a). In the case of PD-I, PD-II, and PD-III isolates from Southeast USA, three genes were uniquely found in PD-I and nine in PD-III (Table S3). When gene gain/loss was compared between PD-II and PD-III isolates from California and PD-I, three genes coding for hypothetical proteins were found in PD-II and PD-III isolates from California but absent in PD-I. In addition, two genes were absent in isolates from Spain but present in PD-II and PD-III isolates from California. On the other hand, two genes were found in PD-I but absent in PD-II and PD-III isolates from California (Fig. 3b). In addition, five genes were absent in isolates from Taiwan, which was considered as the descendant population of Southeast USA (Table 1).

These unique genes were annotated using eggNOG-mapper and searched in the GenBank and Pfam databases, using both BLAST and interproscan5 (Table 1, Table S4). Two hypothetical proteins and a gene coding for the HTH-type transcriptional regulator (*prtR*) were found for PD-I (Table S3); while nine were hypothetical proteins, the protein coding genes *traC_2* (DNA primase), and *higB_2* (endoribonuclease) were found for the PD-III Southeast USA isolates. Two of the three genes found in PD-II and PD-III isolates from California but absent in PD-I coded for hypothetical proteins and one coded for an alpha/beta fold hydrolase. For the two genes absent in Spain, one of them was listed as glutamate 5-kinase, and another had a conserved LacZ, Beta-galactosidase/beta-glucuronidase domain. For the two genes found in PD-I but absent in PD-II and PD-III isolates from California, one was annotated as a hypothetical protein and the other one as a phage head morphogenesis protein. For the five genes absent in isolates from Taiwan, two were annotated as site-specific DNA-methyltransferase; another two were annotated as peptidoglycan DD-metalloendopeptidase family protein and hypothetical protein, respectively; the last one could be a pseudo gene with unknown function.

Intra-subspecific recombination events are pervasive in both the West and East Coast. Intra-subspecific recombination was pervasive in both populations (Fig. 4 and Fig. S3-4). The r/m estimate (recombination to mutation rates) for the California/Spain core genome alignment was 3.29, while the same estimate for Southeast USA/Taiwan core genome alignment was 5.65. In the Southeast USA (Fig. 4b), recombination events were more frequently observed in isolates from the PD-II/PD-III group (recipient) than in isolates from the PD-I group (donor). Within the PD-II/PD-III group, the Texas isolate GB514 (PD-II) was the most frequent recombinant recipient. Donor sequences for the Texas isolate originated from both PD-I and from an ‘unknown’ donor (representing genetic variability present in the population but not characterized in the original sampling). A total of 188 core genes were entirely contained within recombinant regions in the Southeast USA population; out of this group, 101 genes were classified as hypothetical proteins. The remaining recombinant core genes belonged to a variety of functions (Table S5). These functions were grouped by their COG class resulting in 12 genes belonging to the ‘Cellular Processes and Signaling’ class, 5 genes associated with the ‘Information Storage and Processing’ class, 41 genes from the ‘Metabolism’ class, and 7 genes belonging to two or more functional classes (‘Multiple Categories’). Based on gene annotation, some CDs functions are related to virulence and/or host adaptation. These include vitamin B_{12} import (*btuD*), ferric uptake regulation protein (*fur*), response regulator (*gacA*), virulence protein (PD_1332 in Temecula1 assembly AE009442.1, COG0346), polygalacturonase (*pglA*), export protein (*secB*) and ABC transporter (*uup*).

Likewise, sequence exchange occurred between isolates from the PD-III and the PD-II clusters in California. Recombination events were observed among isolates from the same geographic regions (Fig. 4a). Specifically, recombination was frequent between sequences originating from the Temecula Valley in Southern California (Fig. S3). Sequences in both groups acted as donors and recipients. In addition, Northern California isolates were recipients of recombinant segments from Southern California. This group was also a recipient of ‘unknown’ sequence fragments. A total of 180 core genes were exclusively contained within these recombinant regions (Table S5). Eighty-five genes were described as hypothetical proteins. The remaining genes were classified by their COG as: 'Cellular Processes and Signaling' (19 genes), 'Information Storage and Processing' (6 genes), 'Metabolism' (38 genes), and 'Multiple Categories' (6 genes). From these genes, those with annotated function related to host adaptation/virulence include: biofilm growth-associated repressor (*bigR*), periplasmic serine endoprotease (*degP*) (*htrA* in Temecula1 assembly AE009442), putative TonB-dependent receptor (*phuR* in Temecula1 assembly AE009442, COG1629), virulence protein (PD_1332 in Temecula1 assembly AE009442.1, COG0346), sec-independent translocase protein (*tatA-D*), and PhoH-like protein (*ybeZ*).

Based on the used genome annotations, a total of 13 recombinant genes were shared in both populations. These genes were: *glk_1* and *glk_2* (glucokinases), *glmM_2* (a phosphoglucosamine mutase), *glmS_1* and *glmS_2* (glutamine--fructose-6-phosphate aminotransferases [isomerizing]), *grpE* (a GrpE protein), *grxD* (a glutaredoxin 4), *gshB* (a glutathione synthetase), *gtaB* (a UTP--glucose-1-phosphate uridylyltransferase), *pepQ* (a Xaa-Pro dipeptidase), *petA* (an Ubiquinol-cytochrome c reductase iron-sulfur subunit), *petC* (an ammonia monooxygenase gamma subunit), an unnamed PKHD-type hydroxylase (COG3128), and a unnamed Virulence protein (COG0346).

Grapevine-infecting populations in the East and West USA are largely genetically isolated. Nucleotide diversity (π) varied within and among populations (Table 2). Overall, nucleotide diversity was higher within the Southeast USA (947 SNPs, π=1.36x10e^{-05}) compared to California (458 SNPs, π=3.22x10e^{-06}). When compared to their corresponding source populations, nucleotide diversity was lower within Spain (2 SNPs, π=1.38x10e^{-07}) and Taiwan (6 SNPs, π=4.15x10e^{-07}). When diversity in phylogenetically distinct clusters was evaluated, PD-I

(93 SNPs, π=7.58x10e^{-07}) and PD-II (114 SNPs, π=9.65x10e^{-07}) had lower nucleotide diversity than PD-III (509 SNPs, π=3.25x10e^{-06}).

The frequency of polymorphism present in the population in regard to expectations under neutrality was calculated using a Tajima's D. Briefly, negative Tajima's D values indicate an excess of rare polymorphisms than expected under neutrality, which can be caused by a selective sweep or a recent population expansion. Positive Tajima's D values indicate excess of intermediate frequency polymorphism than expected under neutrality, which could suggest balancing selection or a recent population contraction. Tajima's D in California and the Southeast USA was negative (Table 2); however, the magnitude of the statistic in California was roughly twice that of the Southeast USA (-1.448 and -0.658, respectively). Due to the reduced sample size, it was not possible to estimate Tajima's D in Spain or Taiwan. When populations were divided phylogenetically, PD-I isolates had a lower Tajima's D (-2.060) compared to PD-II (-1.781) and PD-III (-1.743). On the other hand, Watterson's θ estimates the population mutation rate from the observed nucleotide diversity. This estimator decreases with increased sample size or with recombination rate. Watterson's θ estimated a higher mutation rate in the Southeast USA (θ=1.64x10e^{-05}) compared to California (θ=5.75x10e^{-06}). When populations were divided based on phylogeny, mutation rate was higher in PD-III (θ=6.72x10e^{-06}) than PD-I (θ=1.64x10e^{-06}) or PD-II (θ=1.87x10e^{-06}).

In addition, Fst values were used to measure population differentiation across geographic and phylogenetic groups. Briefly, Fst values compare the amount of genetic variability within and between populations, values of 1 indicate complete population structuring while values of 0 indicate complete panmixia. Pairwise Fst values (Table S6) for California vs. Southeast USA (Fst = 0.814) and California vs. Taiwan (Fst = 0.964) were higher than California vs. Spain (Fst

= 0.566). This was also the case for comparisons involving Southeast USA vs. Spain (Fst = 0.847) and Southeast USA vs. Taiwan (Fst = 0.114). Taiwan vs. Spain also showed strong differentiation (Fst = 0.994). Once populations were divided phylogenetically, PD-I was more differentiated from PD-II (Fst=0.987) and PD-III (Fst=0.960), than PD-II and PD-III from each other (Fst=0.541).

An MKT was used to estimate the rate of synonymous and non-synonymous polymorphism vs. the rate of synonymous and non-synonymous fixed differences across geographic populations and phylogenetic groups. Under neutrality, it is expected that both rates will be the same (NI=1). Therefore, departures of neutrality (NI≠1) will indicate either the action of balancing selection (e.g. maintenance of population polymorphisms; NI>1) or the action of positive selection (e.g. accumulation of fixed differences between populations; NI<1). The Neutrality Index (NI) was larger than 1 in all comparisons except for Spain vs. Taiwan. NI was significant only for California vs. Taiwan (p-value=9.87x10e^{-05}) (Table S6). Many polymorphisms were observed in Southeast USA and California, while few were observed within Spain or Taiwan. The largest number of fixed differences were observed for Taiwan vs. California. When populations were divided phylogenetically, the NI values were larger than 1 only in comparison between PD-I with PD-II and PD-III. In this instance, the only significant NI was observed for PD-I vs. PD-III (p-value=6.26x10e^{-05}). The number of polymorphisms was larger in PD-III compared to PD-I and PD-II. The number of fixed differences were similar between PD-I vs. PD-II and PD-III, but smaller in PD-II vs. PD-III.

Selective sweep signatures were pervasive in both California and Southeast USA (Fig. 5a), thought the magnitude of the sweep was larger in California. Alternatively, CLR peaks were smaller and scattered in Spain and Taiwan. When the populations were split phylogenetically, CLR peaks were more numerous and prominent in PD-III, followed by PD-II, and finally PD-I (Fig. 5b). Regardless if the populations were subdivided geographically or phylogenetically, some CLR peaks co-located across populations, while others were group specific.

DISCUSSION

Our analyses show that after its introduction from Central America (33, 41), PD-causing subsp. *fastidiosa* split into two populations: one in the East Coast (31 haplotypes) and one in the West Coast (106 haplotypes). Apart from PD-II/PD-III isolates from the Southeast USA, each population formed a sister monophyletic clade with long basal branch lengths. This indicates that the populations split shortly after introduction to the USA. Moreover, isolates from the same location clustered together, suggesting stronger sequence similarity within than between locations. With the current information available, it is not possible to know if the clustering of PD-II/PD-III isolates from the Southeast USA with the California clades instead of Southeast USA (PD-I) reflects a recent introduction to California or if there is a higher diversity within Southeast USA isolates than currently represented. Alternatively, it is feasible the East and West Coast populations originated via independent introduction events. Previous studies have pointed out the large genetic diversity of subsp. *fastidiosa* within Central America (33) and the importation of plant material from this region into the USA (67). Our data do not exclude the possibility that additional subsp. *fastidiosa* strains circulate within Central America and could have been introduced to the USA in relatively simultaneous events. This is a hypothesis that should be evaluated as additional whole genomic data from both native and introduced populations of subsp. *fastidiosa* becomes available. However, previously published MLST data (41, 67) and results based on whole genome sequence analysis (monophyly of the PD-causing population, age and diversity of PD-causing clades, and their evolutionary relationship with the native subsp. *fastidiosa* population) are indicative of a single introduction event.

Pathogen introductions into Spain and Taiwan were closely related to isolates from California and the Southeast USA, respectively. Though closely related to their source populations, both Spain and Taiwan had their unique core haplotypes which could be indicative of early local adaptation. Nonetheless, we cannot discard the possibility that differences in unique core haplotypes might also be the result of a founder effect. Small sample size in both populations does not allow to test between these two possibilities; however, this should be addressed once additional genomic data becomes available.

Gene gain/loss events are common between and within populations. Bacterial gene content is in constant flux (68); in bacteria, evolution via gene gain and loss often precedes evolution at the sequence level (i.e. nucleotide substitutions and indels) (69). Therefore variations in gene content can act as a source for adaptive differentiation (70). Gene gain and loss rates were highest following the introduction to the USA (e.g. 35 genes gained and 49 loss vs. 4 genes gained and 5 loss between the East Coast and Taiwan); however, gene content changes were also detected within each geographic population. The higher number of gene gain/loss events observed in basal tree branches can be explained by a founder event. However, they could also be the result of accumulated gene gain/loss events over longer evolutionary time. It is likely that both factors contribute to gene gain/loss between the native and ancestral subsp. *fastidiosa* populations. The highest intra-population gene gain/loss rates were localized in branches following clade splits. Within California, intra-population splits were associated with locations along a latitudinal gradient (PD-II in Southern California vs. PD-III in Southern and Northern California). In other organisms, selection driven gene gain/loss has been described in genes involved in environmental interactions (69, 71, 72). Likewise, previous studies have found evidence of local adaptation to environmental conditions within California (34). Thus, it is possible that changes in gene content might be adaptive to the local environment. This is further supported by PD-III, which encompasses a larger latitudinal gradient, having four times higher gene gain/loss rates compared to PD-II.

Alternatively, while gene gain/loss rates were higher in PD-II/PD-III Southeast samples compared to PD-I, the difference was not as pronounced as that seen in California. Sampling of PD-causing isolates has been more extensive in California; therefore, detecting environmentally linked gene gain/loss might require further sampling in the Southeast USA. Based on the current annotation, it is difficult to interpret the possible benefit or disadvantage of unique genes found in specific *X. fastidiosa* populations. Functional analysis of these genes will be needed to understand their biological role. Still, a small number of genes involved in transcription regulation (*prtR* and *higB_2*) and DNA replication (*traC_2*) were exclusive to PD-I and PD-III Southeast USA isolates. These functions are linked to changes in bacterial transcription and replication in response to environmental cues (73, 74).

However, it should be noted that gene gain/loss events can also be a product of non-adaptive evolution. In bacteria, genetic drift promotes genome reduction and neutral gene losses are favored by small population size (75, 76). In addition, homologous recombination facilitates core genome homogenization but might not affect accessory genes, leading to gene content divergence and pangenome expansion (69). As such, these gene gain/loss events might not be linked to the adaptive potential of each population. Likewise, this could also be the case of more recent introduction events and smaller population sizes (i.e. Spain and Taiwan populations).

Unequal recombination frequencies drive inter- and intra-population differentiation. r/m estimates showed that recombination contributes more than mutation to genetic diversity. The r/m values for California/Spain (r/m=3.29) and Southeast USA/Taiwan (r/m=5.65) were higher than previous reports on subsp. *fastidiosa* (r/m = 2.074, (33)). However, both values were lower than reports focused specifically for a California population (r/m=6.797, (34)). Location-specific core genomes analyses can detect nucleotide changes uniquely to a geographic region. Therefore, the high r/m found here is likely due to location-specific SNPs.

The number of genes located within intra-subspecific recombination was similar across functional classes showing that there were no specific gene functions more prone to recombination. These results are like those found in a previous analysis (33). On the other hand, the frequency of recombination varied among phylogenetic clusters. PD-II/PD-III Southeast isolates were recipients to sequence fragments from both PD-I and an ‘unknown’ group. Similarly, recombination occurred among geographically close isolates from the PD-II and PD-III clusters in California. These results show that genetic exchange is actively occurring within the West and East Coast. Variations in recombination frequency across isolates have been reported in native subsp. *fastidiosa* populations (33). Furthermore, recombinant genotypes form distinct phylogenetic groups in subsp. *multiplex* (77). Also, *in vitro* analyses have shown that the natural competency in both subsp. *fastidiosa* and subsp. *multiplex* is strain dependent (78, 79). Taken together this shows that intra- and inter-subspecific recombination does not equally affect all strains and that different gene functions, at least within the core genome, are not differentially prone to recombination.

Recombination events also contribute to the differentiation between the East and West Coast, as well as between PD-I, PD-II, and PD-III. Previous studies have shown allele exchange between co-occurring subsp. *multiplex* and subsp. *fastidiosa* isolates in the Southeast USA, but not in California (40). Therefore, the presence of multiple *X. fastidiosa* subspecies within the same geographic regions can enable divergence of recombinant prone isolates or clades. Moreover, highly recombinant clades also experienced higher gene gain/loss in the East and West Coast. Homologous recombination can aid in maintaining core genome cohesiveness while allowing extensive gene gain/loss in the accessory genome (69) and variations in gene content can enable ecological divergence (80). Therefore, intra-subspecific recombination can act as a source of differentiation in PD-causing isolates, not only by mediating allelic exchange but also by facilitating gene gain/loss.

From the genes found to recombine in the Southeast USA and Californian populations with putative function as host adaptation and/or virulence, most have been already identified as recombinants among *X. fastidiosa* populations (79). Genes with the same annotation found in both studies include *btuD, secB, uup, tatD,* and *ybeZ.* In other cases the identified genes were not exactly the same, but genes with similar functions were found in both studies, including genes related to iron acquisition (*fur* in the current study), biofilm-associated-repressor (*bigR* in the current study) (81–83) and sulfide sensor (84), other members of the *sec* pathway (*tatA-D*) (85, 86), and other serine proteases (*degP* here) (79, 85, 86). Interestingly the vitamin B_{12} transporter BtuD was the single annotated gene with highest recombination inter- and intra-subspecific identified in a previous study (79), and has been described in other bacteria as regulating gene expression, abundance of microorganisms and virulence (87, 88), although no functionality has been attributed yet to *X. fastidiosa*. Genes like *fur* and *gacA* have been identified as transcriptionally regulated by calcium (89), an abundant element inside xylem vessels. Other genes like the putative TonB-dependent receptor (*phuR* in the Temecula1 assembly AE009442, COG1629), are involved in twitching motility and biofilm formation (90); and PhoH-like protein (*ybeZ*), is putatively linked to detection and response to changes of phosphate concentration (91).

West and East Coast populations show unique trends of genetic diversity and mutation rate. At a first glance, isolates originating from the Southeast USA population were more genetically diverse than those originating from California. However, this trend was less clear when populations were assigned phylogenetically. PD-II (California + 1 Texas isolate) had slightly higher than PD-I (exclusively Southeast USA), and PD-III (California + 3 Georgia isolates) had higher genetic diversity than either PD-I or PD-II.

The negative Tajima's D values indicate an excess of rare polymorphisms, which can be caused by a selective sweep or a recent population expansion. In the case of subsp. *fastidiosa*, a population expansion could have occurred following a founder effect. This result, in addition to previously published data (24, 33, 67), supports the hypothesis that subsp. *fastidiosa* was introduced to the USA. Furthermore, they show that limitation on genetic diversity caused by a founder effect can be long lasting. Tajima's D values were markedly reduced in PD-I compared to the geographic Southeast USA population (PD-I+PD-II(Texas)/PD-III(Georgia)). This is indicative that there is more than one phylogenetic cluster circulating in the East Coast. Similarly, Tajima's D was smaller in PD-II and PD-III compared to California, further supporting the idea of ongoing latitudinal distinction within the West Coast.

Watterson's θ estimates were also affected by grouping criteria. In the case of Southeast USA compared to PD-I, the Watterson estimator remained roughly unchanged suggesting that mutation rate in the region is captured by current sampling. Watterson's θ was larger in California compared to either PD-II or PD-III, and lower in PD-III compared to PD-II. The values were comparable to previous reports in California (34). This could be indicative that mutation rate within the West Coast is, to a certain point, location dependent and that mutation itself contributes less to population differentiation than other evolutionary forces.

PD-causing strains have differentiated phylogenetically and geographically. The Fst values for different groups of PD-causing isolates were higher than those reported for other global bacterial pathogens(92). It is possible that these values reflect rapid differentiation of PD-causing populations. Pairwise Fst values between PD-I (Southeast only) vs. PD-II (California + 1 Texas) and PD-III (California + 3 Georgia) were higher than between Southeast USA and California. These results further support the phylogenetic and geographic separation of the East and West Coast, and the more recent differentiation within California. How much this differentiation can be linked to the Southeast USA PD-II/PD-III group, needs to be further analyzed. Our Fst analyses indicate a complex phylogeographic history between USA populations, yet, the effects of sample size in these calculations should not be ignored. For example, recently introduced populations (e.g. Spain and Taiwan) showed even higher population differentiation than comparisons involving their source populations. Whether this suggests higher differentiation as a product of a founder effect remains to be determined.

In general, genetic diversity has a high impact on adaptive potential (93); however, some genetic variants might be considered neutral and can be estimated based on the number of synonymous polymorphisms (94). Variables associated with local adaptation are linked to non-synonymous polymorphisms. There were more non-synonymous than synonymous polymorphism in both the East and West Coast. This suggests that, though the number of polymorphisms might be limited due to a recent introduction event, each population maintains a certain level of genetic variation (as evidenced by NI > 1) which could be a source for local adaptation (95).

When populations were divided according to their phylogenetic relationships, a significant NI > 1 was only observed between PD-I and PD-III. Polymorphism largely accumulated in PD-III compared to PD-I. However, the number of fixed differences was comparable between PD-I vs. PD-II and PD-III. This shows that a significant number of intra-clade polymorphisms in PD-III have not yet been fixed. Instead, fixed differences seem to mostly reside between the PD-I compared to PD-II and PD-III. This further supports the idea that East and West Coast populations split early following introduction to the USA, with local population differentiation within a latitudinal gradient in the West Coast.

Selective sweeps have occurred following the introduction of *X. fastidiosa* to the USA. Many CLR peaks were co-localized in the same region while others were group exclusive. The localized nature of CLR peaks in the core genome alignment suggests that selective sweeps can only be detected on certain genes. The location and intensity of selective sweeps are the product of evolutionary and ecological variables. A founder effect can result in reduced selection strength, but it might not affect recombination potential, particularly in *X. fastidiosa* (96). Therefore, the CLR patterns observed here likely reflect genes undergoing strong selection, either following subsp. *fastidiosa* introduction from Central America (co-localized CLR peaks) or via selective pressures associated to a specific environment (group specific CLR peaks). Strong CLR signals in both PD-II and PD-III are indicative that selective sweeps have been more prevalent within the West Coast. Some genes located in CLR peak include: outer membrane protein assembly factors (*BamA-B*), a beta-barrel assembly-enhancing protease (*bepA_4*), a ubiquinol cytochrome C oxidoreductase (*fbcH*), a glycine cleavage system transcriptional repressor (*gcvR*), a glutamine--fructose-6-phosphate aminotransferase (*glmS_2*), a proton/glutamate-aspartate symporter (*gltP*), and a sensor histidine kinase (*rcsC*). Branch-site analyzes aimed to detect signals of positive selection should be performed to further evaluate these results.

CONCLUSIONS

We identified a series of evolutionary mechanisms that led to the diversification of PD-causing subsp. *fastidiosa* populations. Diversification has occurred in core genome sequences via mutation and recombination, and in gene content via gain/loss events. These differences have the potential of facilitating local adaptation to environmental conditions, and in the absence of gene flow, lead to pathogen specialization. The host range and geographic distribution of *X. fastidiosa* is expanding and each new introduction can result in significant economic and ecological damage. Understanding the mechanisms and speed of local adaptation in *X. fastidiosa* is important to manage emerging *X. fastidiosa* diseases and hopefully limit the number of novel epidemics.

MATERIALS AND METHODS

Sampling, culturing, and isolation. The following study encompasses 175 *X. fastidiosa* subsp. *fastidiosa* isolates obtained from infected PD-symptomatic grapevines from diverse geographic regions. The number of isolates from each region were: California (N=140), Southeast USA (N=31), Spain (N=2), and Taiwan (N=2). In addition, three non-grapevine-infecting *X. fastidiosa* subsp. *fastidiosa* isolates from Costa Rica were used as an outgroup (33). New subsp. *fastidiosa* isolates were obtained from infected grapevines in the Southeast USA during 2014-2016; these isolates were cultured from symptomatic leaves as previously described (44). Colonies growing after ~1-2 weeks under 28°C incubation were re-streaked, cloned, and had identity confirmed with *X. fastidiosa* specific PCR primer sets (45). Isolates were obtained from different grapevine varieties. Specifically, the varieties found in Site1 were: Merlot (N=5, years 2014-2016), Mourvedre (N=1, year 2014), Cabernet Sauvignon (N=1, year 2014), Chardonnay (N=5, years 2014-2016), Viognier (N=2, years 2014-2015), Sangiovese (N=1, year 2014), and Touriga (N=1, year 2014). The varieties found in Site2 were: Montaluce (N=1, year 2015), Merlot (N=3, year 2016), Pinot grigio (N=3, year 2016), and Vidal (N=3, year 2016). Except for Site1 (N=16) and Site2 (N=8), all data included in the following study has been previously made publicly available. Detailed metadata on each assembly has been compiled in Table S1; assembly statistics for new whole genome sequences are provided in Table S2.

Sequencing, assembly, and annotation of *X. fastidiosa* subsp. *fastidiosa* isolates. All isolates were sequenced using Illumina HiSeq2000. Samples were sequenced at the University of California, Berkeley Vincent J. Coates Genomics Sequencing Laboratory (California Institute for Quantitative Biosciences; QB3), and the Center for Genomic Sciences, Allegheny Singer Research Institute, Pittsburgh, PA. All raw reads and information regarding each newly sequenced strain can be accessed under the NCBI BioProject accession PRJNA655351. The quality of raw paired FASTQ reads was evaluated using FastQC (46) and visualized using MultiQC (47). Low quality reads and adapter sequences were removed from all paired raw reads using seqtk v1.2 (https://github.com/lh3/seqtk) and cutadapt v1.14 (48) with default parameters. After pre-processing, isolates were assembled *de novo* with SPAdes v3.13 (49, 50) using the -*careful* parameter and -k of 21, 33, 55, and 77. Assembled contigs were reordered using Mauve's contig mover function (51) with the complete publicly available Temecula1 assembly (GCA_000007245.1) used as reference. Assembled and reordered genomes were then individually annotated using the Prokka pipeline (52). In addition, published genome sequences were also re-annotated with Prokka.

Core genome alignments, construction of Maximum Likelihood trees, and haplotype network. Roary v3.11.2 (53) was used to calculate the number of genes in the core (genes shared between 99-100% strains), soft-core (genes shared between 95-99% strains), shell (genes shared between 15-95% strains), and cloud (genes shared between 0-15% strains) genomes of PD-causing isolates (N=175).A core genome alignment of PD-causing isolates plus the three Costa Rica isolates (non-PD) was created using the -e (codon aware multisequence alignment of core genes) and -n (fast nucleotide alignment) flags in Roary. This core genome alignment was used to build a Maximum Likelihood (ML) tree with RAxML (54). The GTRCAT substitution model was used on tree construction, while tree topology and branch support were assessed with 1000 bootstrap replicates. In addition, a non-recombinant tree was constructed by removing detected recombinant segments from the core genome alignment (see later methods). The ML non-recombinant tree was constructed using the same parameters as the recombinant tree. Finally, a haplotype network for PD-causing isolates was built following removal of the outgroup sequences (non-PD). Core genome haplotypes were calculated based on the number of mutations among the analyzed strains, and the haplotype network was built using the HaploNet function in the R package ‘pegas’ (55). Haplotypes were then color-coded by geographic location.

Estimation of recombinant segments and gene gain/loss rates within populations. Isolates were divided based on their geographical origin: California, Southeast USA, Spain, and Taiwan. California and the Southeast USA were the source population for Spain and Taiwan, respectively. Source and descendant relationships between populations were phylogenetically determined (see results). A core genome alignment was created for California/Spain (N=142), and Southeast USA/Taiwan (N=33). The alignment was used to estimate the frequency and location of recombinant events. FastGEAR (56) was used with default parameters to identify lineage-specific recombinant segments (ancestral) and strain-specific recombinant segments (recent). The size and location of recombinant segments across the length of the core genome alignment were mapped within California/Spain and Southeast USA/Taiwan using the R package 'circlize' (57). Donor/recipient recombinant regions were visualized using fastGEAR's plotRecombinations script. In addition, the number of substitutions introduced by recombination vs. random point mutation (r/m) (58) was estimated for the California/Spain and Southeast USA/Taiwan core genome alignments using ClonalFrameML (59). It should be noted that fastGEAR was designed to test recombination in individual gene alignments instead of core genome alignments; a previous study found that fastGEAR was more conservative than other more appropriate recombination detection methods such as ClonalFrameML (34). Future research should perform an empirical comparison of recombination detection methods for *X. fastidiosa.*

Additionally, the stochastic probability of gene gain/loss per tree branch was estimated with GLOOME using default parameters (60). Briefly, RAxML was used to build a ML phylogenetic tree for the California/Spain and Southeast USA/Taiwan core genome alignments. The parameters used were the same that for the PD-causing ML tree. Roary v3.11.2 was used to calculate a binary gene presence (1)/absence (0) matrix within the California/Spain and the Southeast USA/Taiwan populations. A binary accessory genome matrix was created by removing core genome genes from the dataset. Subsequently, the binary accessory presence/absence matrix was transposed and converted into FASTA format. The binary accessory genome and the ML trees were used as inputs to the GLOOME analysis. Unique genes were identified through estimating gene gain/loss rates within each population. These genes were annotated by eggNOG-mapper v1.0.3 (https://github.com/eggnogdb/eggnog-mapper) and searched in the Genebank and Pfam databases using BLAST and interproscan v5.47 (https://github.com/ebi-pf-team/interproscan).

Population genomics analyses. Global measures of genetic diversity, population differentiation, and selective sweeps were estimated for the PD-causing dataset using the R package ‘PopGenome’ (61). The dataset was subdivided in two ways: a) based on isolates’ geographical origin (i.e. California, Southeast USA, Spain, and Taiwan), and b) based on isolates’ phylogenetic relationships (i.e. PD-I, PD-II, and PD-III; see results). All calculations described below were performed for both the (a) geographic and (b) phylogenetic subdivisions.

Genetic diversity was estimated by computing nucleotide diversity (π), Tajima’s D (62), and the Watterson’s estimator (θ) (63). Briefly, nucleotide diversity (π) measures the average number of nucleotide differences per site in pairwise comparisons among DNA sequences. Tajima’s D evaluates the frequency of polymorphism present in a population and compares that value to the expectation under neutrality. The Watterson θ estimator measures the mutation rate of a population. Population differentiation was estimated by calculating the Fixation Index (Fst) (64) within (a) geographic and (b) phylogenetic groups. In addition, the McDonald-Kreitman Test (MKT) (65) was used to estimate the rate of synonymous (syn-P) and non-synonymous (nonsyn-P) polymorphism, against the rate of fixed synonymous (syn-F) and non-synonymous (nonsyn-F) differences. In each instance, the Neutrality Index (NI) was calculated. NI > 1 suggests an excess of preserved polymorphism maintained via balancing selection. Alternatively, NI < 1 suggests population divergence via positive selection. Finally, the location and magnitude of selective sweeps was calculated using Nielsen’s composite-likelihood-ratio (CLR) (66). This test identifies regions with aberrant allele frequency spectra and estimates if the aberrant allele distribution fits the expectations of a selective sweep. The test was performed on a 1500bp sliding window across the length of the PD-causing core genome alignment.

Data Availability. The raw sequence data files for the newly published isolates were submitted to the NCBI Sequence Read Archive under accession number SAMN15732826 through SAMN15732849. All other used data has been previously published. All accession numbers are listed in Table S1.

ACKNOWLEDGMENTS

This work has received funding from the PD/GWSS Research Program, California Department of Food and Agriculture, and the European Union's Horizon 2020 research and innovation program under grant agreement N. 727987 "*Xylella fastidiosa* Active Containment Through a multidisciplinary-Oriented Research Strategy XF-ACTORS. Genome sequencing was performed at the UC Berkeley Vincent J. Coates Genomics Sequencing Laboratory, which is supported by an NIH instrumentation grant (S10 OD018174).

References

1. Pimentel D, Lach L, Zuniga R, Mprrison D. 2000. Environmental and Economic Costs of Nonindigenous Species in the United States. Bioscience 50:53.
2. Fletcher J, Bender C, Budowle B, Cobb WT, Gold SE, Ishimaru CA, Luster D, Melcher U, Murch R, Scherm H, Seem RC, Sherwood JL, Sobral BW, Tolin SA. 2006. Plant Pathogen Forensics: Capabilities, Needs, and Recommendations. Microbiol Mol Biol Rev 70:450–471.
3. Pyšek P, Jarošík V, Pergl J. 2011. Alien plants introduced by different pathways differ in invasion success: Unintentional introductions as a threat to natural areas. PLoS One 6.
4. Early R, Bradley BA, Dukes JS, Lawler JJ, Olden JD, Blumenthal DM, Gonzalez P, Grosholz ED, Ibañez I, Miller LP, Sorte CJB, Tatem AJ. 2016. Global threats from invasive alien species in the twenty-first century and national response capacities. Nat Commun 7.
5. Mable BK. 2019. Conservation of adaptive potential and functional diversity: integrating old and new approaches. Conserv Genet 20:89–100.
6. Baltrus DA, Nishimura MT, Dougherty KM, Biswas S, Mukhtar MS, Vicente J, Holub EB, Dangl JL. 2012. The molecular basis of host specialization in bean pathovars of *Pseudomonas syringae*. Mol Plant-Microbe Interact 25:877–888.
7. Karasov TL, Horton MW, Bergelson J. 2014. Genomic variability as a driver of plant–pathogen coevolution? Curr Opin Plant Biol 18:24–30.
8. Plissonneau C, Benevenuto J, Mohd-Assaad N, Fouché S, Hartmann FE, Croll D. 2017. Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. Front Plant Sci 8:1–15.
9. Mokryakov M V., Abdeev IA, Piruzyan ES, Schaad NW, Ignatov AN. 2010. Diversity of effector genes in plant pathogenic bacteria of genus *Xanthomonas*. Microbiology 79:58–65.
10. Rowntree JK, Cameron DD, Preziosi RF. 2011. Genetic variation changes the interactions between the parasitic plant-ecosystem engineer *Rhinanthus* and its hosts. Philos Trans R Soc B Biol Sci 366:1380–1388.
11. González R, Butković A, Elena SF. 2019. Role of host genetic diversity for susceptibility-to-infection in the evolution of virulence of a plant virus. Virus Evol 5:1–12.
12. Zhu Y, Chen H, Fan J, Wang Y, Li Y, Chen J, Fan JX, Yang S, Hu L, Leung H, Mew TW, Teng PS, Wang Z, Mundt CC. 2000. Genetic diversity and disease control in rice. Nature 406:718–722.
13. Escriu F. 2012. Diversity of Plant Virus Populations: A Valuable Tool Diversity of Plant Virus Populations: A Valuable Tool for Epidemiological Studies, p. 13. *In* IntechOpen.
14. Brown JKM. 2015. Durable Resistance of Crops to Disease: A Darwinian Perspective. Annu Rev Phytopathol 53:513–539.
15. Zhan J. 2016. Population Genetics of Plant Pathogens. eLS 1–7.
16. Giraud T, Gladieux P, Gavrilets S. 2010. Linking emergence of fungal plant diseases and ecological speciation. Trends Ecol Evol 25:387–395.
17. Mhedbi-Hajri N, Hajri A, Boureau T, Darrasse A, Durand K, Brin C, Saux MF Le, Manceau C, Poussier S, Pruvost O, Lemaire C, Jacques MA. 2013. Evolutionary History of the Plant Pathogenic Bacterium *Xanthomonas axonopodis*. PLoS One 8.
18. Zhan A, Hu J, Hu X, Zhou Z, Hui M, Wang S, Peng W, Wang M, Bao Z. 2009. Fine-scale

population genetic structure of zhikong scallop (chlamys farreri): Do local marine currents drive geographical differentiation? Mar Biotechnol 11:223–235.

19. McDonald BA, Linde C. 2002. The population genetics of plant pathogens and breeding strategies for durable resistance. Euphytica 124:163–180.

20. Slatkin M. 1985. Gene flow in natural populations. Annu Rev Ecol Syst Vol 16 393–430.

21. McDermott JM, McDonald BA. 1993. Gene flow in plant pathosystems. Annu Rev Phytopathol 31:353–373.

22. Pruvost O, Boyer K, Ravigné V, Richard D, Vernière C. 2019. Deciphering how plant pathogenic bacteria disperse and meet: Molecular epidemiology of *Xanthomonas citri* pv. *citri* at microgeographic scales in a tropical area of Asiatic citrus canker endemicity. Evol Appl 12:1523–1538.

23. EFSA. 2018. Update of the *Xylella* spp. host plant database. EFSA J 16:1–87.

24. Vanhove M, Retchless AC, Sicard A, Rieux A, Coletta-filho HD, Fuente LD La, Stenger DC, Almeida PP. 2019. Genomic Diversity and Recombination among *Xylella fastidiosa* Subspecies. Appl Environ Microbiol 85:1–17.

25. Bragard C, Dehnen-Schmutz K, Di Serio F, Gonthier P, Jacques MA, Jaques Miret JA, Justesen AF, MacLeod A, Magnusson CS, Milonas P, Navas-Cortés JA, Potting R, Reignault PL, Thulke HH, Van der Werf W, Vicent Civera A, Yuen J, Zappalà L, Makowski D, Delbianco A, Maiorano A, Muñoz Guajardo I, Stancanelli G, Guzzo M, Parnell S. 2019. Effectiveness of in planta control measures for *Xylella fastidiosa*. EFSA J 17.

26. Almeida RPP, De La Fuente L, Koebnik R, Lopes JRS, Parnell S, Scherm H. 2019. Addressing the New Global Threat of *Xylella fastidiosa*. Phytopathology 109:172–174.

27. Nunney L, Hopkins DL, Morano LD, Russell SE, Stouthamer R. 2014. Intersubspecific Recombination in *Xylella fastidiosa* Strains Native to the United States: Infection of Novel Hosts Associated with an Unsuccessful Invasion. Appl Environ Microbiol 80:1159–1169.

28. Nunney L, Yuan X, Bromley RE, Stouthamer R. 2012. Detecting Genetic Introgression: High Levels of Intersubspecific Recombination Found in *Xylella fastidiosa* in Brazil. Appl Environ Microbiol 78:4702–4714.

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.

30. Giampetruzzi A, Saponari M, Loconsole G, Boscia D, Savino VN, Almeida RPP, Zicca S, Landa BB, Chacón-Diaz C, Saldarelli P. 2017. Genome-Wide Analysis Provides Evidence on the Genetic Relatedness of the Emergent *Xylella fastidiosa* Genotype in Italy to Isolates from Central America. Phytopathology 107:816–827.

31. Saponari M, Giampetruzzi A, Loconsole G, Boscia D, Saldarelli P. 2018. *Xylella fastidiosa* in Olive in Apulia: Where We Stand. Phytopathology 109:175–186.

32. 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.

33. Castillo AI, Chacón-díaz C, Rodríguez-murillo N, Coletta- HD, Almeida RPP, Rica C. 2020. Impacts of local population history and ecology on the evolution of a globally dispersed pathogen . BMC Genomics 21:1–51.

34. Vanhove M, Sicard A, Ezennia J, Leviten N, Almeida RPP. 2020. Population structure

and adaptation of a bacterial pathogen in California grapevines. Env Microbiol.

35. Gomila M, Moralejo E, Busquets A, Segui G, Olmo D, Nieto A, Juan A, Lalucat J. 2018. Draft Genome Resources of Two Strains of *Xylella fastidiosa* XYL1732/17 and XYL2055/17 Isolated from Mallorca Vineyards. Phytopathology 109:222–224.

36. Castillo AI, Tuan S-J, Retchless AC, Hu F-T, Chang H-Y, Almeidaa RPP. 2019. Draft Whole-Genome Sequences of *Xylella fastidiosa* subsp. *fastidiosa* Strains TPD3 and TPD4, Isolated from Grapevines in Hou-li, Taiwan. Microbiology 8:1–3.

37. Schuenzel EL, Scally M, Stouthamer R, Nunney L. 2005. A Multigene Phylogenetic Study of Clonal Diversity and Divergence in North American Strains of the Plant Pathogen *Xylella fastidiosa*. Appl Environ Microbiol 71:3832–3839.

38. 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. Ecol Epidemiol 100:601–611.

39. Cella E, Angeletti S, Fogolari M, Bazzardi R, De L, Ciccozzi M, Cella E, Angeletti S, Fogolari M, Bazzardi R. 2018. Two different *Xylella fastidiosa* strains circulating in Italy: phylogenetic and evolutionary analyses. J Plant Interact 13:428–432.

40. Nunney L, Schuenzel EL, Scally M, Bromley RE, Stouthamerc R. 2014. Large-scale intersubspecific recombination in the plant-pathogenic bacterium *Xylella fastidiosa* is associated with the host shift to mulberry. Appl Environ Microbiol 80:3025–3033.

41. Nunney L, Ortiz B, Russell SA, Sánchez RR, 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.

42. Kabir P. Tumber JMA and KBF. 2014. Pierce's disease costs California $104 million per year. Calif Agric 68:20–29.

43. Hickey C. 2019. Pierce's Disease of Grape: Identification and Management. UGA Coop Ext Bull 1514:1–6.

44. Parker JK, Havird JC, De La Fuente L. 2012. Differentiation of *Xylella fastidiosa* strains via multilocus sequence analysis of environmentally mediated genes (MLSA-E). Appl Environ Microbiol 78:1385–1396.

45. Francis M, Lin H, Rosa JC La, Doddapaneni H, Civerolo EL. 2006. Genome-based PCR primers for specific and sensitive detection and quantification of *Xylella fastidiosa*. Eur J Plant Pathol 115:203–213.

46. Andrews S, Wingett SW, Hamilton RS. 2018. FastQ Screen : A tool for multi-genome mapping and quality control [version 2 ; referees : 4 approved] Referee Status : F10000research 1–13.

47. Ewels P, Lundin S, Max K. 2016. Data and text mining MultiQC : summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048.

48. Marcel M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB.netJournal 17:5–7.

49. Bankevich A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

50. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, Prjibelski A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel S, Woyke T, McLean J, Lasken R, Tesler G, Alekseyev M, Pevzner P. 2013. Assembly single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol 20:714–737.

51. 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.

52. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

53. 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.

54. Stamatakis A. 2014. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

55. Paradis E. 2010. Pegas: An R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420.

56. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. 2017. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. Mol Biol Evol 34:1167–1182.

57. Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular visualization in R. Bioinformatics 30:2811–2812.

58. Guttman DS, Dykhuizen DE. 1994. Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science (80-) 266:1380–1383.

59. Didelot X, Wilson DJ. 2015. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLoS Comput Biol 11:1–18.

60. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME : gain loss mapping engine. Bioinformatics 26:2914–2915.

61. Pfeifer B, Wittelsbu U, Ramos-onsins SE, Lercher MJ. 2014. PopGenome : An Efficient Swiss Army Knife for Population Genomic Analyses in R. Mol Biol Evol 31:1929–1936.

62. Tajima F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genet Soc Am 595:585–595.

63. Watterson GA. 1975. On the numer of Segregating Sites in Genetical Models without Recombination. Theor Popul Biol 276:256–276.

64. Wrigth S. 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution (N Y) 19:395–420.

65. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

66. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15:1566–1575.

67. 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.

68. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin E V. 2014. Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. BMC Med 12:1–19.

69. Iranzo J, Wolf YI, Koonin E V., Sela I. 2019. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. Nat Commun 10.

70. Hartmann FE, Croll D. 2017. Distinct trajectories of massive recent gene gains and losses in populations of a microbial eukaryotic pathogen. Mol Biol Evol 34:2808–2822.

71. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. 2007.

Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3:2515–2528.

72. Moulana A, Anderson RE, Fortunato CS, Huber JA. 2020. Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus *Sulfurovum* in Geographically Distinct Deep-Sea Hydrothermal Vents . mSystems 5:1–18.

73. Frick DN, Richardson CC. 2001. DNA Primases. Annu Rev Biochem 70:39–80.

74. Browning DF, Busby SJW. 2016. Local and global regulation of transcription initiation in bacteria. Nat Rev Microbiol 14:638–650.

75. Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res 19:1450–1454.

76. Albalat R, Cañestro C. 2016. Evolution by gene loss. Nat Rev Genet 17:379–391.

77. Nunney L, Vickerman DB, Bromley RE, Russell SA, Hartman JR, Morano LD, Stouthamer R. 2013. Recent Evolutionary Radiation and Host Plant Specialization in the *Xylella fastidiosa* Subspecies Native to the United States. Appl Environ Microbiol 79:2189–2200.

78. Kandel PP, Almeida RPP, Cobine PA, De La Fuente L. 2017. Natural Competence Rates Are Variable Among *Xylella fastidiosa* Strains and Homologous Recombination Occurs In Vitro Between Subspecies *fastidiosa* and *multiplex*. Mol Plant-Microbe Interact 30:589–600.

79. Potnis N, Kandel PP, Merfa M V., Retchless AC, Parker JK, Stenger DC, Almeida RPP, Bergsma-Vlami M, Westenberg M, Cobine PA, De La Fuente L. 2019. Patterns of inter- and intrasubspecific homologous recombination inform eco-evolutionary dynamics of *Xylella fastidiosa*. ISME J 13:2319–2333.

80. Schmutzer M, Barraclough TG. 2019. The role of recombination, niche-specific gene pools and flexible genomes in the ecological speciation of bacteria. Ecol Evol 9:4544–4556.

81. Barbosa RL, Rinaldi FC, Guimarães BG, Benedetti CE. 2007. Crystallization and preliminary X-ray analysis of BigR, a transcription repressor from *Xylella fastidiosa* involved in biofilm formation. Acta Crystallogr Sect F Struct Biol Cryst Commun 63:596–598.

82. 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.

83. Guimarães BG, Barbosa RL, Soprano AS, Campos BM, De Souza TA, Tonoli CCC, Leme AFP, Murakami MT, Benedetti CE. 2011. Plant pathogenic bacteria utilize biofilm growth-associated repressor (BigR), a novel winged-helix redox switch, to control hydrogen sulfide detoxification under hypoxia. J Biol Chem 286:26148–26157.

84. De Lira NPV, Pauletti BA, Marques AC, Perez CA, Caserta R, De Souza AA, Vercesi AE, Paes Leme AF, Benedetti CE. 2018. BigR is a sulfide sensor that regulates a sulfur transferase/dioxygenase required for aerobic respiration of plant bacteria under sulfide stress. Sci Rep 8:1–13.

85. Federici MT, Marcondes JA, Picchi SC, Stuchi ES, Fadel AL, Laia ML, Lemos MVF, Lemos EGM. 2012. *Xylella fastidiosa*: An *in vivo* system to study possible survival strategies within citrus xylem vessels based on global gene expression analysis. Electron J Biotechnol 15.

86. Da Silva Neto JF, Koide T, Gomes SL, Marques M V. 2007. The single extracytoplasmic-function sigma factor of *Xylella fastidiosa* is involved in the heat shock response and

presents an unusual regulatory mechanism. J Bacteriol 189:551–560.

87. Lee KM, Go J, Yoon MY, Park Y, Kim SC, Yong DE, Yoon SS. 2012. Vitamin B 12-Mediated restoration of defective anaerobic growth leads to reduced biofilm formation in *Pseudomonas aeruginosa*. Infect Immun 80:1639–1649.

88. Cordonnier C, Le Bihan G, Emond-Rheault JG, Garrivier A, Harel J, Jubelin G. 2016. Vitamin B12 uptake by the gut commensal bacteria bacteroides thetaiotaomicron limits the production of shiga toxin by enterohemorrhagic *Escherichia coli*. Toxins (Basel) 8.

89. Chen H, De La Fuente L. 2020. Calcium transcriptionally regulates movement, recombination and other functions of *Xylella fastidiosa* under constant flow inside microfluidic chambers. Microb Biotechnol 13:548–561.

90. Cursino L, Li Y, Zaini PA, De La Fuente L, Hoch HC, Burr TJ. 2009. Twitching motility and biofilm formation are associated with tonB1 in *Xylella fastidiosa*. FEMS Microbiol Lett 299:193–199.

91. Santos-Beneit F. 2015. The Pho regulon: A huge regulatory network in bacteria. Front Microbiol 6:1–13.

92. Singh J, Khan A. 2019. Distinct patterns of natural selection determine sub-population structure in the fire blight pathogen, Erwinia amylovora. Sci Rep 9:1–13.

93. Ørsted M, Hoffmann AA, Sverrisdóttir E, Nielsen KL, Kristensen TN. 2019. Genomic variation predicts adaptive evolutionary responses better than population bottleneck history. PLoS Genet 15:1–18.

94. Holderegger R, Kamm U, Gugerli F. 2006. Adaptive vs. neutral genetic diversity: Implications for landscape genetics. Landsc Ecol 21:797–807.

95. Moutinho AF, Bataillon T, Dutheil JY. 2019. Variation of the adaptive substitution rate between species and within genomes. Evol Ecol.

96. Kung SH, Almeida RPP. 2011. Natural competence and recombination in the plant pathogen *Xylella fastidiosa*. Appl Environ Microbiol 77:5278–5284.

FIGURES LEGENDS

Figure 1. Maximum Likelihood (ML) tree and haplotype network showing phylogenetic and geographic diversification of worldwide PD-causing subsp. *fastidiosa* isolates. Color represents isolates from the same geographical location: California (Red), Texas (Pink), Georgia (Green), North Carolina (Dark green), Florida (Yellow), Spain (Light blue), and Taiwan (Dark blue). PD-causing strains have been divided into three phylogenetically supported clades (PD-I, PD-II, PD-III). **a.** Haplotype network of PD-causing subsp. *fastidiosa* isolates. Haplotypes belonging to each PD-causing clade are shown within black circles. Roman numbers identify detected haplotypes (I-CXLI). Size of the circle indicates the number of isolates belonging to each haplotype. **b.** Maximum likelihood (ML) tree of PD-causing subsp. *fastidiosa* isolates. Tree was built using the core genome alignment without removing recombinant segments. Bootstrap values mark branch support. Arrows point towards the base of PD-causing clades (-I to -III).

Figure 2. Phylogeographic analysis showing diversification of PD-causing isolates within the contiguous USA. Color represents isolates from the same geographical location: California (Red), Texas (Pink), Georgia (Green), North Carolina (Dark green), and Florida (Yellow). Coordinates were recorded during field sampling. In absence of this information, coordinates referred to the city or vineyard closest to the sample site were used. Florida coordinates were not available, the location shown in the map represents central Florida. Isolates from Southern and Northern California are shown within pale red circles. PD-causing strains were divided into three phylogenetically supported clades: PD-I (Southeast USA isolates exclusively), PD-II (Southern California and Texas isolates), and PD-III (both Southern and Northern California isolates, and three Georgia isolates). Tree was built using the core genome alignment without removing recombinant segments. Bootstrap values mark branch support.

Figure 3. Venn diagram and maps showing population linked gene gain/loss events among PD-causing isolates. Color represents isolates from the same geographical location: California (Red), Texas (Pink), Georgia (Green), North Carolina (Dark green), Florida (Yellow), Spain (Light blue), and Taiwan (Dark blue). **a.** Venn diagram shows both the number of genes shared between geographic PD-causing populations and genes unique to each population. Size of the oval represents sample size. **b.** Estimated number of genes gained and lost between geographical locations and following introduction events. Arrows point from the source population to its descendant following introduction events. California isolates belong to the phylogenetically distinct clades PD-II and PD-III. Included Southeast isolates belong to the phylogenetically distinct PD-I and PD-III clades. All maps were publicly available from Wikimedia commons.

Figure 4. Frequency and location of recombination events in fastGEAR identified lineages. Analysis shows results for: **a.** the California/Spain population and **b.** the Southeast USA/Taiwan population. FastGEAR's recombination plots show two distinct lineages on each population (red, PD-III in California/Spain and PD-II/PD-III in Southeast USA/Taiwan; blue, PD-II in California/Spain and PD-I in Southeast USA/Taiwan). The recombination events are shown across the length of the core genome alignment. Larger areas represent recipient sequences while shorter segments of different color within those areas represent donor sequences from another lineage. Recombinant segments from unidentified lineages are shown in black. Maximum Likelihood (ML) trees showing the phylogenetic relationship of isolates within each intra-population cluster identified by fastGEAR are also included. Trees were built using the core genome alignment without removing recombinant segments for the California/Spain and Southeast USA/Taiwan populations. Bootstrap values mark branch support.

Figure 5. Line plot showing variations in Nielsen's composite likelihood ratio (CLR) across the length of the core genome alignment (1500 bp window size). The CLR identifies regions with aberrant allele frequency and determines if their distribution matches those expected from a selective sweep. Peaks represent higher CLR values at that position, which is indicative of a putative selective sweep. Color represents isolates from the same geographical location or phylogenetic cluster. **a.** Lines indicate distinct geographic population: California (Red), Southeast USA (Green), Spain (Light blue), and Taiwan (Dark blue). **b.** Lines indicated distinct phylogenetic clusters: PD-I (Teal), PD-II (Yellow), and PD-III (Purple).

Table 1. List of genes gained/lost among geographic and phylogenetic PD-causing groups.

Annotation
Genes absent in the Taiwan population (Found in the contiguous USA)
Site-specific DNA-methyltransferase (QIS25725.1); ko:K00571,ko:K00590,ko:K07319 (adenine-specific DNA-methyltransferase); PF01555
Hypothetical protein (QIS26419.1)
Peptidoglycan DD-metalloendopeptidase family protein (QIS26766.1); PF06594, PF00353 (RTX calcium-binding nonapeptide repeat)
Site-specific DNA-methyltransferase (QIS25737.1); ko:K00571,ko:K00590,ko:K07319 (adenine-specific DNA-methyltransferase); PF01555
Pseudogene
Genes absent in the Spanish population (Found in the contiguous USA)
CPD*: LacZ, Beta-galactosidase/beta-glucuronidase; ko:K01192(beta-mannosidase)
Glutamate 5-kinase (AAO28181.1); ko:K00931; PF00696
PD-II and PD-III exclusive genes in the California population
Alpha/beta fold hydrolase (QIS25057.1); ko:K02170,ko:K07002 (pimeloyl-[acyl-carrier protein] methyl ester esterase)
Hypothetical protein (QJP55224.1); PF04014 (Antidote-toxin recognition MazE, bacterial antitoxin)
Hypothetical protein (AAO28982.1)
PD-I and PD-III exclusive genes in the Southeast population, excluding the Taiwanese clade
Hypothetical protein (QIS26118.1)
Phage head morphogenesis protein (QIS26295.1); PF04233
Genes gained after introduction into Taiwan
CPD*: OM_channels Superfamily, Porin superfamily
CPD*: DUF769 Superfamily; ko:K15125 (filamentous hemagglutinin)
CPD*: entero_EhxA Superfamily
Hypothetical protein (QIS25070.1); RTX toxin (QIS25071.1)
Genes gained after introduction into Spain
Pseudogene: Glycoside hydrolase family 125 protein; ko:K09704 (uncharacterized protein); PF06824 (Metal-independent alpha-mannosidase)
DUF596 domain-containing protein (QIS26773.1); PF04591
Hypothetical protein (QID15519.1) Hemagglutinin (QID15518.1); ko:K02014 (iron complex outer membrane receptor protein)

ko: KEGG orthology; PF: Pfam database entry ID

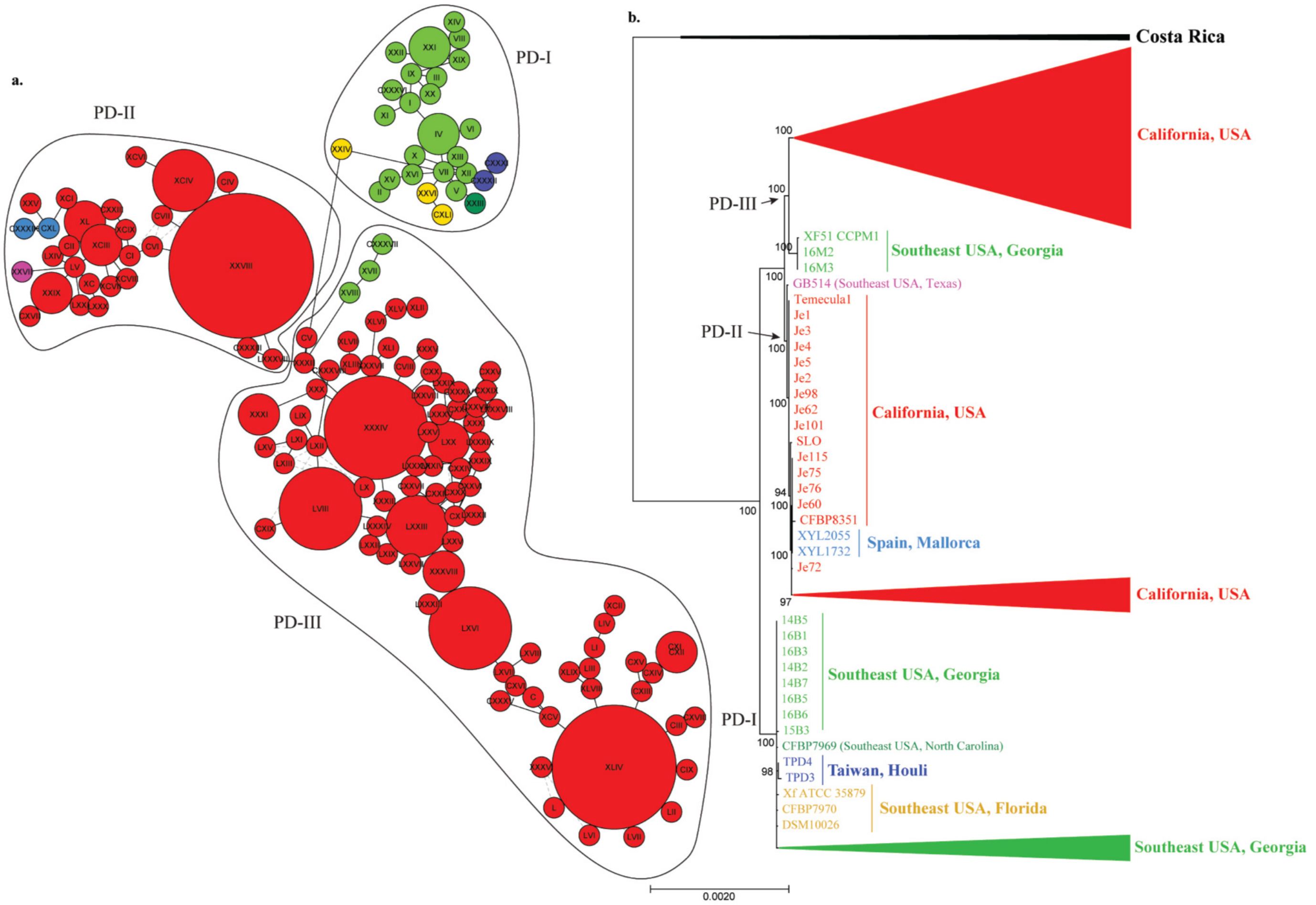
Table 2. Diversity and neutrality statistics of PD-causing isolates. a) Geographically divided populations, and b) Phylogenetically divided populations.

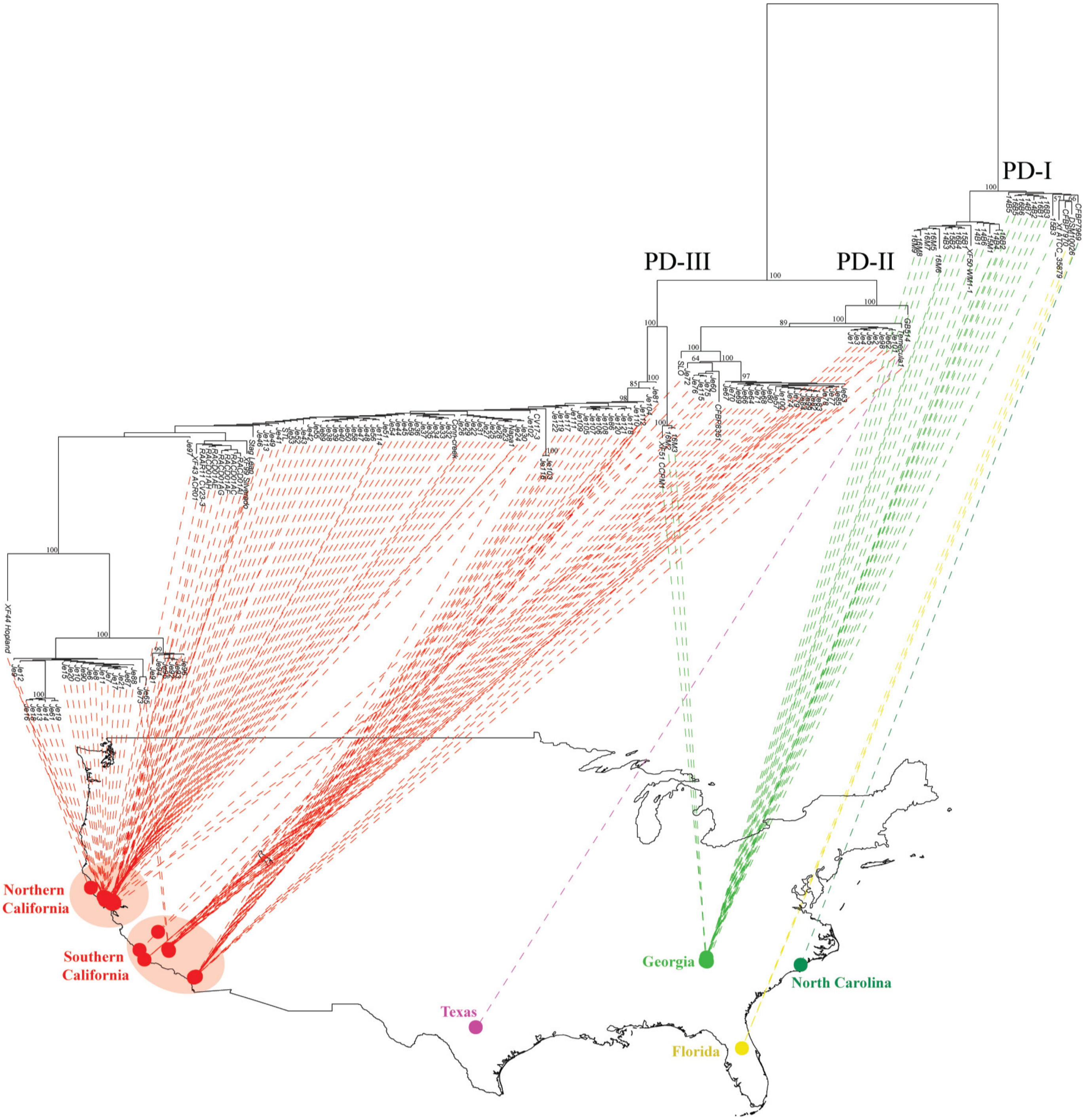
Population (a)	Core (nt)	SNPs	π	θ	Tajima's D
California (140)	14,446,213	458	3.22x10e^{-06}	1.64x10e^{-05}	-1.448
Southeast USA (31)		947	1.36x10e^{-05}	5.75x10e^{-06}	-0.658
Spain (2)		2	1.38x10e^{-07}	1.38x10e^{-07}	*
Taiwan (2)		6	4.15x10e^{-07}	4.15x10e^{-07}	*

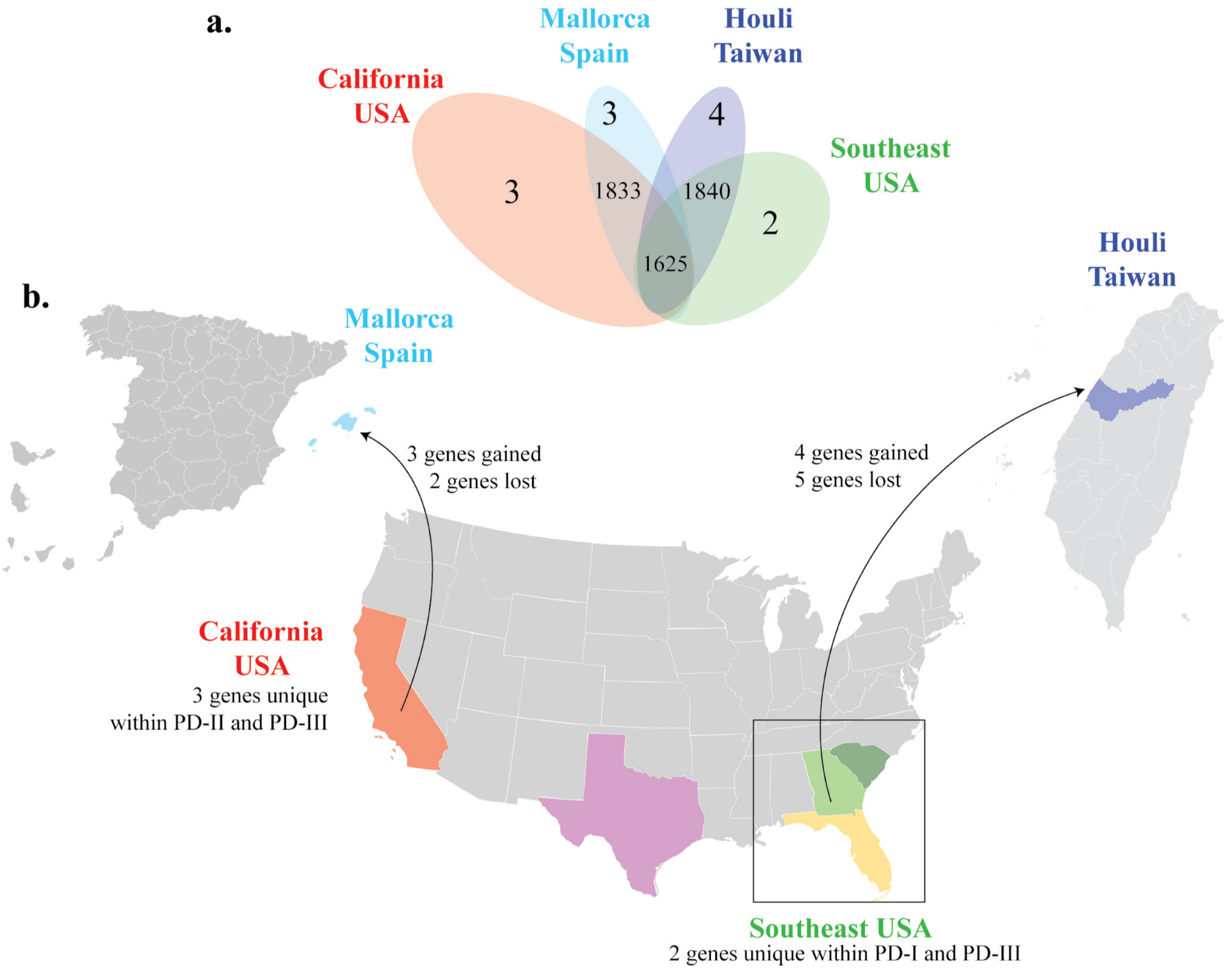
* Spain isolates were not included.

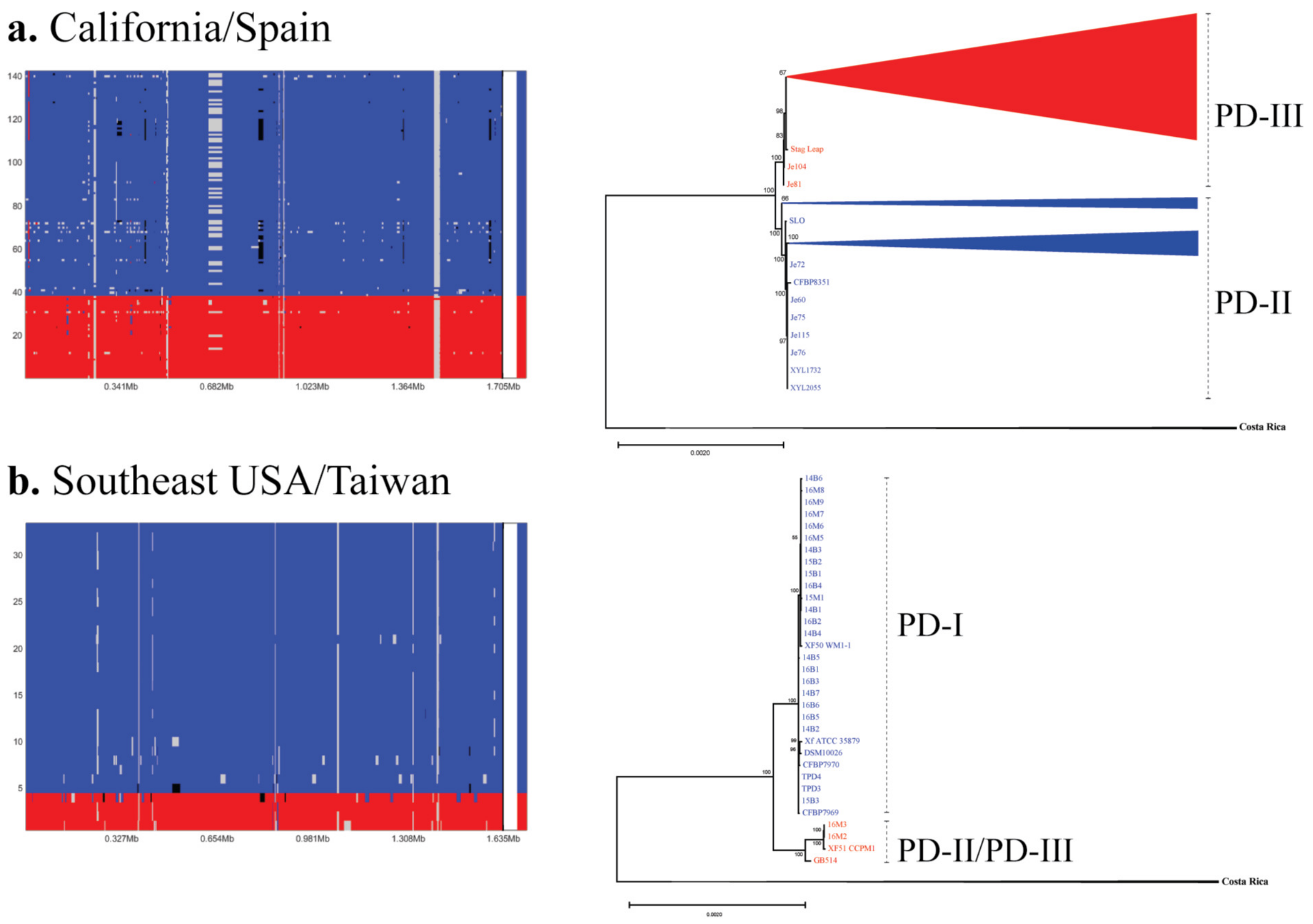
* Taiwan isolates were not included.

Population (b)	Core (nt)	SNPs	π	θ	Tajima's D
PD-I (29)	14,446,213	93	7.58x10e^{-07}	1.64x10e^{-06}	-2.0604
PD-II (40)		114	9.65x10e^{-07}	1.87x10e^{-06}	-1.7813
PD-III (106)		509	3.25x10e^{-06}	6.72x10e^{-06}	-1.7425


a.
PD-I
PD-II
PD-III
XXI
IV
XXVIII
XCIV
XXXIV
LVIII
LXVI
XLIV
XXXI
LXXIII
XXIX
XCIII
XL
b.
Costa Rica
100
California, USA
100
PD-III
100
XF51 CCPM1
16M2
16M3
Southeast USA, Georgia
100
GB514 (Southeast USA, Texas)
Temecula1
Je1
Je3
Je4
Je5
Je2
Je98
Je62
Je101
SLO
Je115
Je75
Je76
Je60
CFBP8351
California, USA
PD-II
100
100
94
100
100
XYL2055
XYL1732
Spain, Mallorca
100
Je72
97
California, USA
14B5
16B1
16B3
14B2
14B7
16B5
16B6
15B3
Southeast USA, Georgia
PD-I
100
CFBP7969 (Southeast USA, North Carolina)
98
TPD4
TPD3
Taiwan, Houli
Xf ATCC 35879
CFBP7970
DSM10026
Southeast USA, Florida
Southeast USA, Georgia
0.0020

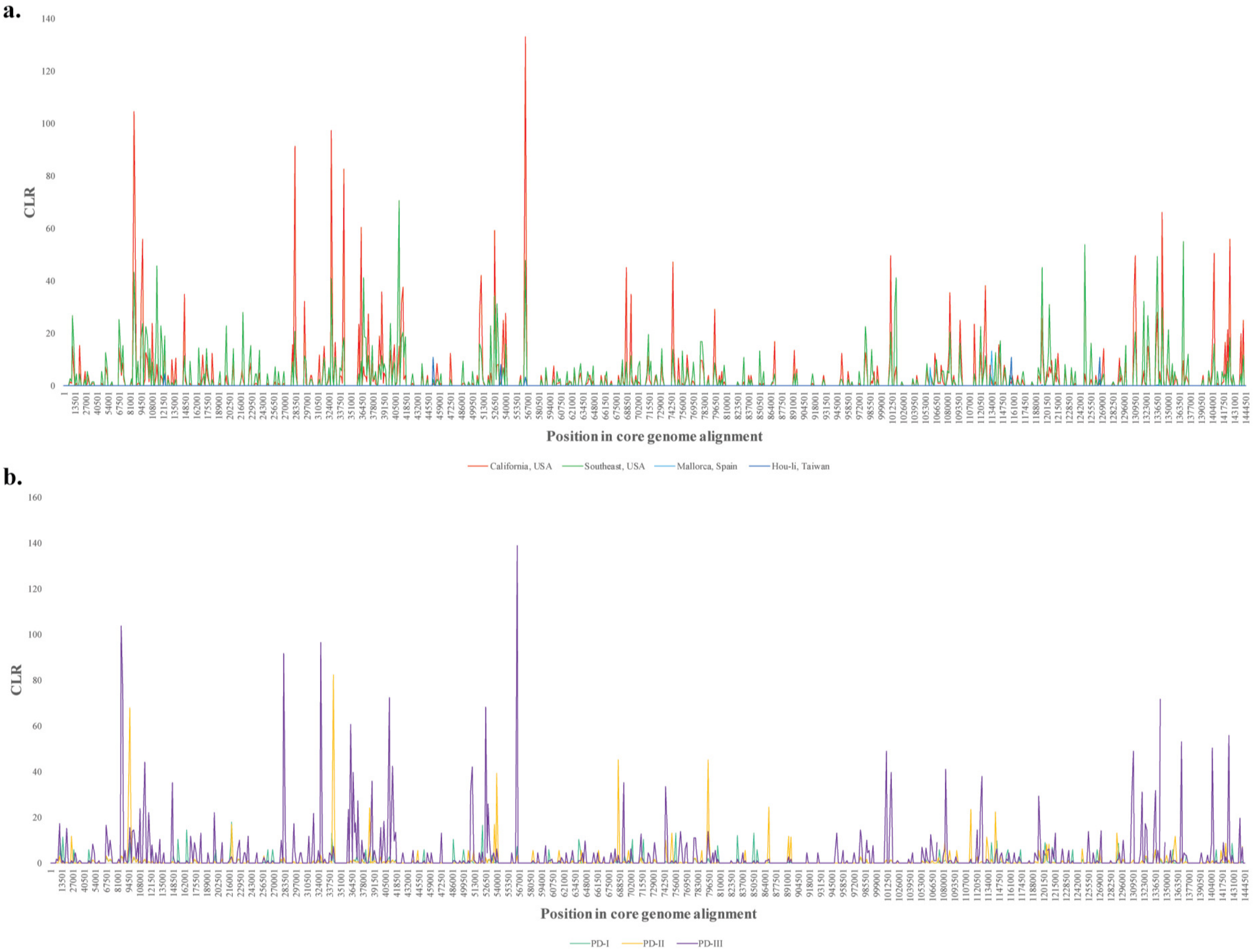

PD-I
PD-II
PD-III
XF44 Hopland
Northern California
Southern California
Texas
Georgia
North Carolina
Florida


a.
Mallorca
Spain
Houli
Taiwan
California
USA
Southeast
USA
3
3
4
1833
1840
2
1625
b.
Mallorca
Spain
Houli
Taiwan
3 genes gained
2 genes lost
4 genes gained
5 genes lost
California
USA
3 genes unique
within PD-II and PD-III
Southeast USA
2 genes unique within PD-I and PD-III



a. California/Spain

b. Southeast USA/Taiwan


a.
140
120
100
80
60
40
20
0
CLR
Position in core genome alignment
California, USA
Southeast, USA
Mallorca, Spain
Hou-li, Taiwan
b.
160
140
120
100
80
60
40
20
0
CLR
Position in core genome alignment
PD-I
PD-II
PD-III