Population structure and adaptation of a bacterial pathogen in California grapevines

Mathieu Vanhove, Anne Sicard, Jeffery Ezennia, Nina Leviten and Rodrigo P.P. Almeida

Department of Environmental Science, Policy and Management, University of California-Berkeley, Berkeley, CA, 94720.

Summary

*Xylella fastidiosa* subsp. *fastidiosa* causes Pierce’s disease of grapevine (PD) and has been present in California for over a century. A singly introduced genotype spread across the state causing large outbreaks and damaging the grapevine industry. This study presents 122 *X. fastidiosa* subsp. *fastidiosa* genomes from symptomatic grapevines, and explores pathogen genetic diversity associated with PD in California. A total of 5218 single-nucleotide polymorphisms (SNPs) were found in the dataset. Strong population genetic structure was found; isolates split into five genetic clusters divided into two lineages. The core/soft-core genome constituted 41.2% of the total genome, emphasizing the high genetic variability of *X. fastidiosa* genomes. An ecological niche model was performed to estimate the environmental niche of the pathogen within California and to identify key climatic factors involved in dispersal. A landscape genomic approach was undertaken aiming to link local adaptation to climatic factors. A total of 18 non-synonymous polymorphisms found to be under selective pressures were correlated with at least one environmental variable highlighting the role of temperature, precipitation and elevation on *X. fastidiosa* adaptation to grapevines in California. Finally, the contribution to virulence of three of the genes under positive selective pressure and of one recombinant gene was studied by reverse genetics.

Introduction

In agricultural ecosystems, bacterial plant pathogens offer largely untested models to measure a phenotypic trait such as virulence and map associated genomic loci. A long standing paradigm in crop–pathogen interactions is that hosts and pathogens are engaged in gene-for-gene co-evolutionary dynamics (Keen, 1990). However, with the development of high-throughput genomics, multiple studies reported that these interactions might also be influenced by abiotic factors acting on genetic loci (Croll and McDonald, 2017). While the study of microbial biogeography has expanded in the recent years, the local adaptation of plant pathogens has not been expansively examined, despite the fact that these are biologically amenable systems to study (Kraemer and Boynton, 2017). Furthermore, pathogens in agroecosystems can have devastating effects on crop yields and epidemics remain a major concern (Stukenbrock and McDonald, 2008; Fisher et al., 2012). Understanding pathogen evolution and the origin of pathogenicity and virulence remain central to mitigate impacts and risks of plant pathogens.

In the last two decades, landscape genetics has emerged as a discipline aimed at linking population genetics, spatial statistics and landscape ecology in order to quantify the effects of landscape features on gene flow and adaptation (Manel et al., 2003; Manel and Holderegger, 2013). The heterogeneous space or ‘landscape’ affects microevolutionary dynamics at various scales, leaving genomic signatures that may be identified (Biek and Real, 2010). To date, in this emerging field, Dudaniec and Tesson (2016) noted that little attention has been paid to the linkage between microorganism dispersal and environmental factors, despite evidence of non-random distributions and patterns of isolation by distance (IBD) within populations (Martiny et al., 2006). Because of their high dispersal abilities (Taylor et al., 2006), microbes were long thought to display little genetic biogeographic differentiation. However, numerous studies support the idea that microbial species indeed exhibit biogeographic patterns (Martiny et al., 2006).

Local adaptation occurs when different biotic and/or abiotic selective pressures lead to higher fitness in a...
focal population compared to other (Giraud et al., 2017).

Different considerations must be taken into account when investigating spatially structured microbial populations compared to macro-organisms, such as large population sizes, fast generation time, colonization bottlenecks and seasonal variations (Prosser et al., 2007; Hanson et al., 2012; Hahn et al., 2015). Recombination also influences local adaptation by introducing foreign genes and increasing genetic variance, either allowing for adaptation or disturbing locally adapted gene combinations (Bürger, 1999).

*Xylella fastidiosa* subsp. *fastidiosa* in California represents an opportunity to study local adaptation of a plant pathogen in an agroecosystem. This pathogen is responsible for Pierce’s disease of grapevine (PD), a devastating disease first described in 1892 in California (Pierce, 1892). Multiple epidemics have been reported across the state over the past century. The disease is caused by a blockage of xylem vessels, probably due both to bacterial populations and/or secretions and plant defence responses (Sicard et al., 2018). The following reduction in xylem sap flow leads to leaf scorch and stunted growth and can result in vine death. The only natural means of pathogen spread is via xylem sap-feeding insect vectors (Sicard et al., 2018). *X. fastidiosa* has been classified into four subspecies, namely, *fastidiosa*, *multiplex*, *pauca* and *sandyi*. A fifth subspecies isolated in mulberry, subsp. *morus*, is thought to be the results of inter-subspecific homologous recombination between subspp. *fastidiosa* and *multiplex* (Nunney et al., 2014; Vanhove et al., 2019). *X. fastidiosa* populations have been historically isolated due to geographical and host barriers, but the emergence of the pathogen in Europe in 2013 (subsp. *pauca*) is an example of the potential impacts associated with human-mediated invasions (Sicard et al., 2018). The clade of subsp. *fastidiosa* causing PD is not native to California; available data suggest a single introduction from Central America (Nunney et al., 2010). This singly introduced genotype then spread in Californian vineyards and is now found across the major grape-growing regions of the state (Tumber et al., 2014). With favourable climatic conditions plant pathogens are expected to extend their range (Garrett et al., 2006), and the impact of climate on *X. fastidiosa* has been extensively reported (Bosso et al., 2016a). Low winter temperatures are known to be the primary limitation of the geographical range of *X. fastidiosa* causing PD (Purcell, 1974), suggesting that the disease may increase its distribution due to climate change. As such, this disease system offers a broad set of spatial scales to study abiotic factors that affect plant pathogen distribution.

In the present study, five geographic locations were sampled across California resulting in the sequencing of 122 subsp. *fastidiosa* genomes obtained from symptomatic grapevines. The genetic diversity and population structure of the plant pathogen were quantified and the genomic basis of adaptation to abiotic factors at the scale of California was investigated by using two different approaches. The first approach consisted of detecting outlier loci that deviate from genome-wide patterns of diversity (Vitti et al., 2013). This approach uses large numbers of single nucleotide polymorphisms (SNPs) and detects markers that exhibit higher level of genetic differentiation than expected under neutrality (Holderegger and Wagner, 2008). The second approach, known as ecological association, detects significant statistical associations between potential genetic markers and environmental variables (Mita et al., 2013; Manel et al., 2016). The list of polymorphisms uncovered from these environmental associations and potentially involved in local adaptation can then be compared to genomic regions under selection to provide additional supports of environmental adaptation and reduce false positives (Branco et al., 2017).

We first modelled the ecological niche of the PD-causing bacterium in California, and used whole genome sequence data to explore the genetic structure of this population. Then, we examined patterns of selection by investigating genomic signatures of positive selection and correlate them with altitude, temperature and precipitation variables. While comparative genomic studies may reveal genes that contribute to pathogen virulence (Griswold, 2008), reverse genetics enable to experimentally test whether these genes are indeed involved in its pathogenesis. We also selected three genes under positive selective pressure and one recombinant gene and tested their effect on *X. fastidiosa* virulence on grapes.

Results

Population subdivision, variant detection and spatial structure

An average of 1 908 446 reads per isolate was obtained and the mapping of reads to the reference *X. fastidiosa* subsp. *fastidiosa* Temecula1 (ASM724v1) averaged 98.05% (Supporting Information Table S1), with a depth of coverage of 132.30 ± 60.2 SD. Genetic comparisons were possible due to a conservative variant-calling strategy resulting in a set of high confidence SNPs (see Materials and Methods). A total of 5218 SNPs were identified among the 122 isolates sequenced in California (Table 1). A Bayesian Analysis of Population Structure (BAPS) revealed the presence of two lineages and five genetic clusters: cluster 1 in Santa Barbara, cluster 2 in Temecula, cluster 3 in Bakersfield, cluster 4 in Sonoma and cluster 5 in Napa (Fig. 1, Supporting Information...
Table S1), harbouring different amount of genetic diversity ranging from 727 to 2875 SNPs (Table 1). Each cluster was roughly associated with its geographic origin, but a few outliers were present in each genetic cluster indicative of exchange among these subpopulations: cluster 1 (57.1% of isolates were isolated in Santa Barbara, 3 isolates were outliers), cluster 2 (Temecula, 94.7%, 1 outlier), cluster 3 (Bakersfield, 84.6%, 4 outliers), cluster 4 (Sonoma, 92.6%, 2 outliers) and cluster 5 (Napa, 88.3%, 5 outliers). Phylogenetic analysis revealed strongly supported clades. However, gene flow has occurred among subpopulations as portrayed by Wright’s FST (Table 2), with values ranging from 0.108 to 0.218. Signs of isolation by distance were also identified (Mantel r test = 0.392, P ≤ 0.001), indicative of genetic disparities over the landscape. The realized niche of subsp. fastidiosa infecting grapevines across California was predicted using MaxEnt, and selected environmental variables, precipitation in the coldest quarter (bio19; 46.0%) and altitude (41.8%) contributed the most to the model (Fig. 1).

Pan-genome analysis

To investigate the pan-genome of this population, we performed de novo assemblies on the entire dataset (Supporting Information Table S2). Isolates had an average of 2 517 953 bp, and a N50 and L50 of 51 903 bp and 20.97 respectively. The genetic diversity in the five different clusters ranged from 6.15 × 10^{-5} to 3.36 × 10^{-5} (Table 1). Due to quality issues, two isolates were removed (Je9 and Je17) from the dataset. Analysis of the core and accessory genomes revealed the presence of 4583 genes, with 1073 (≥99% of isolates) core genes and 816 (17.8% in 99%–95% of isolates) soft-core genes, 756 shell genes shared by 15%–95% of the population (16.5%) and 1938 cloud genes shared by less than 15% of the population (42.5%; Fig. 2C). Presence of

<table>
<thead>
<tr>
<th>Cluster</th>
<th>SNP(^a)</th>
<th>(\eta)</th>
<th>(\pi)</th>
<th>(\theta)</th>
<th>Tajima’s (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Barbara (cluster 1, (n = 7))</td>
<td>727</td>
<td>297</td>
<td>5.97 × 10^{-5}</td>
<td>6.64 × 10^{-5}</td>
<td>−0.549</td>
</tr>
<tr>
<td>Temecula (cluster 2, (n = 16))</td>
<td>2316</td>
<td>219</td>
<td>3.38 × 10^{-5}</td>
<td>4.83 × 10^{-5}</td>
<td>−1.170</td>
</tr>
<tr>
<td>Bakersfield (cluster 3, (n = 26))</td>
<td>1362</td>
<td>184</td>
<td>6.15 × 10^{-5}</td>
<td>5.07 × 10^{-5}</td>
<td>0.842</td>
</tr>
<tr>
<td>Sonoma (cluster 4, (n = 28))</td>
<td>2875</td>
<td>211</td>
<td>5.76 × 10^{-5}</td>
<td>5.18 × 10^{-5}</td>
<td>0.450</td>
</tr>
<tr>
<td>Napa (cluster 5, (n = 43))</td>
<td>2479</td>
<td>138</td>
<td>3.36 × 10^{-5}</td>
<td>4.35 × 10^{-5}</td>
<td>−0.844</td>
</tr>
<tr>
<td>Total ((n = 120))</td>
<td>5218</td>
<td>5240</td>
<td>0.275</td>
<td>0.187</td>
<td>1.580</td>
</tr>
</tbody>
</table>

Note: A total of 5218 SNPs were identified. In the core genome Santa Barbara isolates (cluster 1) displayed the highest number of mutations (\(\eta = 297\)), where Napa only harboured 135 mutations. Each cluster displayed similar nucleotide diversity (\(\pi\)) and population mutation rates (\(\theta\)), but Tajima’s \(D\) values were negative for clusters 1, 2 and 5, indicative of a recent population expansion.

\(^a\). Single nucleotide polymorphism mapped to the \(X.\ fastidiosa\) subsp. \(fastidiosa\) Temecula1 reference.

Fig. 1. Distribution of and phylogenetic placement of the Pierce’s disease-causing \(Xylella\ fastidiosa\) subsp. \(fastidiosa\) populations within California. The predicted niche of the population was estimated using MAXENT; area values closer to 1 (red) indicate higher likelihood of pathogen occurrence. [Color figure can be viewed at wileyonlinelibrary.com]

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*, 22, 2625–2638
homologous recombination was investigated using ClonalFrameML and fastGEAR on the core genome alignment (725 750 bp). The relative effect of recombination to mutation was: \( \frac{r}{m} = 6.797 \) (i.e. recombination generated more substitutions than mutation), the relative rate of recombination to mutation was \( \frac{R}{\theta} = 0.524 \), and the average length of imports equal to \( \delta = 406 \) bp.

ClonalFrameML and fastGEAR analyses identified 98 and 47 recombining segments respectively (Supporting Information Fig. S6). Recombining elements were mapped to the reference genome; 64 and 28 of the recombining segments found respectively by ClonalFrameML and fastGEAR were identified and mapped to known genes (Supporting Information Tables S9 and S10). Ten of these genes were found using both methods, including an ABC transporter (cvbA), a cardiolipin synthase (cls), a transcriptional regulator (attO) and a cation acetate symporter (ppa_1).

Analysis of depth coverage variation has the potential to reveal duplications. A total of 30 genes were found to have an average coverage \( \geq 2 \) (Supporting Information Table S12). These genes encoded mostly for hypothetical proteins \( (n = 16) \) or phage-related proteins \( (n = 12) \). Interestingly, each cluster seemed to have a gene duplication for PD_0789, a resolvase/integrase-like protein (GO:0003677; GO:0000150; GO:0006310; Supporting Information Table S12). Additionally, Clusters 1 and 3, which are part of the same lineage, had a higher average coverage for PD_1184 (mean = 1.977 and 1.874 respectively), a toxin-like protein. We interpreted these as loci duplications conserved in these populations.

To investigate the temporal evolution of the grapevine genotype, an ML phylogeny was constructed using the core genome of the 120 isolates with an additional 24 previously published subsp. fastidiosa genomes. Published isolates include the following regions: eastern United States \( (n = 3) \) and Mexico \( (n = 2) \) with known isolation times for a time total time period dating from 1987 to 2015 (28 years, Supporting Information Fig. S7, Table S2). Tip-dating inference using BEAST led to the inference of a substitution rate of \( 6.37724 \times 10^{-7} \) per site per year (95 Confidence Interval (CI): \( 3.9277 \times 10^{-7}, \ 9.0912 \times 10^{-7} \)). The evolutionary rate was then extrapolated to the whole subspecies using BEAST (Fig. 2A).

The split between the Mexican outgroup and the rest of the USA isolates was estimated at 1269 CE (CI: 850 CE–1668 CE). A divergence between eastern and western isolates was estimated at 1827 CE (1676 CE–1946 CE). Based on these estimations, the time to most common recent ancestor (TMRCA) of subsp. fastidiosa in California dates to 1960 CE (1851 CE–1976 CE).

Selection

In order to investigate signs of natural selection, two gene-based methods were used: \( \omega \) and the McDonald–Kreitman (MK) test. In addition, one univariate outlier test, XTX, based on high confidence SNPs, was also used. The core genome was generated based on de novo assemblies and the \( \omega \) ratio was estimated using codeml. The SNP outlier XTX method identified 190 SNPs that were mapped to the reference genome leading to the identification of 60 genes and a total of 64 non-synonymous mutations (Supporting Information Table S8). Some of the gene products under selection encoded for genes involved in pathogenesis (GO:0009405) such as a hemolysin-type calcium binding domain (cluster 3, 4 and 5), DNA recombination (GO:0006310) and proteolysis (GO:0006508; Supporting Information Table S11). Additionally, selection was investigated within each genetic cluster to assess whether the different clusters were under different selective pressures. The results of that analysis are summarized in the Supporting Information.

Selection, recombination and virulence

Three genes encoding for hypothetical proteins and displaying high values of \( \omega \) in at least one of the five clusters (PD_0516 > 2.8 in cluster 4 and 5; PD_2073 > 2.6 in cluster 4; PD_0616 > 1 in cluster 3) were knocked out to determine their effect on X. fastidiosa virulence in grapes. One recombinant gene with unknown function (PD_0579, Supporting Information Table S9) was also selected. The PD_2073 mutant did not survive on selective growth media pointing towards an essential physiological role. The other three knockout

© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., Environmental Microbiology, 22, 2625–2638
mutants (PD_0516, PD_0579 and PD_0616 mutants), and the virulent wild-type (WT) PD strain STL, were mechanically inoculated into susceptible grapevines. No difference in disease severity was observed among the different strains. A more detailed summary of the results is available as Supporting Information.

Association analysis of climatic variables

The association analyses between SNPs and environmental variables were performed using the Bayes’ factors and non-parametric Spearman’s Rho methods implemented in Bayenv2, and latent factor mixed model
Thirty genes, which were found to be responding to positive selective pressure, were significantly correlated with a climatic variable using two Environmental Association Analysis (EAA) methods. A total of 59 SNPs were significantly associated with at least one environmental variable, including 18 non-synonymous (NSY) mutations (Table 3, Fig. 3, Supporting Information Fig. S2). Among the 18 correlated NSY loci, four were associated with altitude, three with annual mean temperature (bio1), four with mean temperature in the warmest quarter (bio10) and three with mean temperature in the wettest quarter (bio8). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Precipitation variables were also correlated with 81 SNPs including four NSY mutations that were found to be associated with precipitation in the wettest month (bio13). Only one NSY mutation was detected with all three methods (snp_1295815, Fig. 3D) on gene PD_1095 (hypothetical protein) was associated with four NSY mutations that were found to be associated with precipitation in the wettest month (bio13).
Table 3. Non-synonymous mutations in the Pierce’s disease X. fastidiosa Californian population associated with climate variables.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Selection Method</th>
<th>Gene</th>
<th>Function</th>
<th>Landscape Genomics Methods</th>
<th>Bioclim</th>
<th>Variable</th>
<th>GO term</th>
<th>GO Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>161 654</td>
<td>X; ωC3</td>
<td>PD_0127</td>
<td>DUF2326 domain-containing</td>
<td>BF &amp; ρ*</td>
<td>bio8</td>
<td>Temperature</td>
<td>GO:0019068</td>
<td>virion assembly</td>
</tr>
<tr>
<td>463 139</td>
<td>X; ωC1; ωC2</td>
<td>PD_0378 (gp4)</td>
<td>Phage-related portal protein</td>
<td>BF &amp; LFMM</td>
<td>bio13</td>
<td>Precipitation</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>598 600</td>
<td>X; ωC1; ωC2</td>
<td>PD_0501</td>
<td>ABC transporter</td>
<td>BF &amp; ρ*</td>
<td>bio10</td>
<td>Temperature</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>627 294</td>
<td>X; ωC4; ωC5</td>
<td>PD_0515</td>
<td>Unknown</td>
<td>BF &amp; LFMM</td>
<td>bio8</td>
<td>Temperature</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>627 395</td>
<td>X; ωC4; ωC5</td>
<td>PD_0515</td>
<td>Unknown</td>
<td>ρ* &amp; LFMM</td>
<td>bio1</td>
<td>Temperature</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>627 509</td>
<td>X; ωC4; ωC5</td>
<td>PD_0515</td>
<td>Unknown</td>
<td>BF &amp; LFMM</td>
<td>Altitude</td>
<td>Temperature</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>628 191</td>
<td>ωC4; ωC5</td>
<td>PD_0516</td>
<td>Unknown</td>
<td>BF &amp; LFMM</td>
<td>bio1</td>
<td>Temperature</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>629 619</td>
<td>X; ωC1; ωC2</td>
<td>PD_0517</td>
<td>Arginine deaminase</td>
<td>BF &amp; LFMM</td>
<td>bio1</td>
<td>Temperature</td>
<td>GO:0016021</td>
<td>integral component of membrane</td>
</tr>
<tr>
<td>765 080</td>
<td>X; ωC1; ωC2</td>
<td>PD_0620</td>
<td>Glycine decarboxylase</td>
<td>BF &amp; ρ*</td>
<td>bio13</td>
<td>Precipitation</td>
<td>GO:0006546</td>
<td>glycine catabolic process</td>
</tr>
<tr>
<td>946 774</td>
<td>X; ωC1; ωC2</td>
<td>PD_0764 (int)</td>
<td>Phage-related integrase</td>
<td>ρ* &amp; LFMM</td>
<td>Altitude</td>
<td>Altitude</td>
<td>GO:0015074</td>
<td>DNA integration</td>
</tr>
<tr>
<td>977 362</td>
<td>X; ωC1; ωC2</td>
<td>PD_0789</td>
<td>Recombinase family</td>
<td>BF &amp; LFMM</td>
<td>bio10</td>
<td>Temperature</td>
<td>GO:0003677; GO:0000150; GO:0006310</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
</tr>
<tr>
<td>977 382</td>
<td>X; ωC1; ωC2</td>
<td>PD_0789</td>
<td>Recombinase family</td>
<td>BF &amp; LFMM &amp; ρ*</td>
<td>bio10</td>
<td>Temperature</td>
<td>GO:0003677; GO:0000150; GO:0006310</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
</tr>
<tr>
<td>978 167</td>
<td>X; ωC1; ωC2</td>
<td>PD_0790 (traC)</td>
<td>DNA primase</td>
<td>ρ* &amp; LFMM</td>
<td>bio18</td>
<td>Precipitation</td>
<td>GO:0005516; GO:0004683; GO:0006468</td>
<td>calmodulin binding; calmodulin-dependent protein kinase activity; protein phosphorylation</td>
</tr>
<tr>
<td>1 068 152</td>
<td>X; ωC1; ωC2</td>
<td>PD_0858</td>
<td>DUF4440 domain-containing</td>
<td>BF &amp; ρ*</td>
<td>bio13</td>
<td>Precipitation</td>
<td>GO:0000000150; GO:0000150; GO:0006310</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
</tr>
<tr>
<td>1 295 506</td>
<td>X; ωC1; ωC2</td>
<td>PD_1095</td>
<td>Unknown</td>
<td>BF &amp; LFMM</td>
<td>Altitude</td>
<td>Altitude</td>
<td>GO:0000150; GO:0006310</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
</tr>
<tr>
<td>1 295 815</td>
<td>X; ωC1; ωC2</td>
<td>PD_1095</td>
<td>Unknown</td>
<td>ρ* &amp; LFMM</td>
<td>bio10; bio13; bio14; bio19</td>
<td>Temperature</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
<td></td>
</tr>
<tr>
<td>1 446 174</td>
<td>X; ωC1; ωC2</td>
<td>PD_1243</td>
<td>DUF596 domain-containing</td>
<td>LFMM &amp; ρ*</td>
<td>bio8</td>
<td>Temperature</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
<td></td>
</tr>
<tr>
<td>1 446 184</td>
<td>X; ωC1; ωC2</td>
<td>PD_1243</td>
<td>DUF596 domain-containing</td>
<td>LFMM &amp; ρ*</td>
<td>Altitude; bio6</td>
<td>Temperature</td>
<td>DNA binding; recombinase activity; DNA recombination</td>
<td></td>
</tr>
</tbody>
</table>

Each SNP listed was found to be under positive selection with least one method (MK test, dN/dS or X;X) and was found to be correlated with one environmental variable by two of the three EEA methods: Bayes’ factors and non-parametric Spearman’s Rho methods implemented in Bayenv2 or latent factor mixed model (LFMM). ρ*, non-parametric Spearman’s Rho (ρ) distribution.
changes and modifications to farming practices, or transportation and establishment of novel genotypes in various regions (Sicard et al., 2018). Regardless, while the inferred dates and respective CI are ecologically reasonable, considering the history of PD in the USA, additional sampling efforts from other locations are expected to provide more robust data on the origin of PD in the USA.

The core/soft-core genome constituted 41.2% of the 4583 genes in the population, similar to what has been observed for other bacteria and X. fastidiosa (Lapierre and Gogarten, 2009; Mira et al., 2010). The large accessory genome, 1938 cloud genes shared by <15% of the population (42.3%), appeared similar to other study. In a study of 205 multidrug-resistant (MDR) Serratia marcescens in the United Kingdom and Ireland, Moradigaravand et al. (2016) found an accessory genome value of 61.3%. Analysis of depth of coverage variation revealed duplications that could contribute to pathogenesis. In the first lineage, 10 genes had double coverage including PD_1184, a toxin protein and PD_0789 (an integrase involved in DNA recombination). On the other hand, depth of coverage variation was more pronounced in the second lineage, with over 35 genes found in each of the three genetic clusters. These are expected to be conserved duplications, and are potentially important for host (plant or insect) colonization.

Various selective forces are acting on the population of subsp. fastidiosa analysed. Tajima’s D values were negative in genetic clusters 1, 2 and 5. This metric searches for genomic regions undergoing a selective sweep (Tajima, 1989). Negative Tajima’s D values are indicative of a surplus of rare alleles, signatures of positive selection or recent population expansion (Charlesworth, 2006). ClonalFrameML and fastGEAR analyses revealed the

Fig. 3. Six SNPs displaying the highest differences between the population carrying the reference nucleotide and the population with the mutation are shown. Each map is composed of the environmental variable associated with the correlated SNP. The change in aminocaid is indicated and density plots portray the values of the environmental variable for the two populations of isolates (wild-type and mutant). The full maps of the 18 non-synonymous mutations associated with a climatic variable can be found on Supporting Information Fig. S2. [Color figure can be viewed at wileyonlinelibrary.com]
exchange of genetic material occurring among the genetic clusters. Recombination is recognized as a main driver of genetic diversity in this species (Nunney et al., 2014). In this population, the relative effect of recombination and mutation to substitution accumulation was 6.797. Previous work found that recombination contributed twice as much (Vanhove et al., 2019), indicating that recombination within subsp. fastidiosa in California is more frequent than in other populations of X. fastidiosa studied. It is possible that allele exchange has limited fitness consequences, given the population is host-limited and relatively young. Additionally, the singly-introduced genotype affecting grapevines in California has recombined with endemic subsp. multiplex, potentially facilitating subsp. fastidiosa adaptation to grapevines and environmental conditions (Vanhove et al., 2019).

Recombining genes that persist within populations may confer benefits to the strains that harbour them. We here tested whether a recombining gene, PD_0579, had an effect on X. fastidiosa virulence. The deletion of this gene did not have any effect on multiplication or movement of the bacterium within grapevines or on disease symptom development, pointing towards its lack of effect on X. fastidiosa virulence in this plant species. Similarly, the three genes under positive selection tested biologically had no effect on disease symptom development or in planta movement. These strains might have already been selected for their fast multiplication and movement in grapes and as a consequence, may already be well adapted to this host. These genes may be associated with insect colonization, for example, or abiotic stresses among other possibilities. A NSY mutation in one of the candidate genes, PD_0516, was significantly associated with the annual mean temperature, suggesting that genes involved in local adaptation might not necessarily be associated with pathogenicity. PD_2073 encodes a hypothetical protein within an operon including six other genes with homology to the type I restriction and modification system (ProOpDB; Taboada et al., 2011), its mutant did not grow in vitro. This gene may also be part of this large, multi-functional enzyme complex.

Selective pressures imposed by biotic and abiotic factors may provide higher fitness to local populations in relation to those from other regions. Comparing the rate of synonymous and non-synonymous mutations has been widely used in microbial genomics (Giraud et al., 2017). Other studies have previously identified genes in bacterial plant pathogens using this approach. For instance, Richard et al. (Richard et al., 2017) reported selection in genes involved in resistance to copper-based insecticide in Xanthomonas citri pv. citri. In the present study, the reverse-ecology approaches identified signatures of adaptation in genes involved in several functions including response to antibiotics (mrcB), pathogenesis (btaE and upaG), heme transport (ccmC) and cell adhesion (pilA_1). These findings highlight the adaptive potential of X. fastidiosa, a concern for regions outside its historical range in the Americas associated with recent disease outbreaks (Sicard et al., 2018).

We used an ecological niche modelling approach to understand the abiotic niche requirements of subsp. fastidiosa in California grapevines. The predictive environmental model described where PD has been reported over past years (Tumber et al., 2014). These analyses confirmed the influence of climate on the chronic establishment of the bacterium, as previously suggested (Purcell, 1997). This approach has been previously used to predict the X. fastidiosa ecological niche in Italy and across the Mediterranean basin (Bosso et al., 2016a). Minimum temperature in the coldest month (Bio6) and altitude were identified as important factors in shaping X. fastidiosa distribution. A colder winter climate has been previously shown to be a limiting factor for X. fastidiosa survivorship in grapevines (Purcell, 1977).

The ecological determinants leading to subsp. fastidiosa adaptation to grapevines in California were studied in relation to the spatial genomic structure observed in this dataset. This approach aimed to characterize the association between genetic loci, selection pressures and abiotic factors. Eighteen NSY mutations were detected using EAAs methods. GO term analyses revealed that these genetic changes were primarily associated with genes involved in recombination, glycine catabolism and protein phosphorylation. A SNP in a recombinase protein (PD_0789) was found using all three landscape genomic methods and may play a role in the ability of the pathogen to adapt to its environment. Both an ABC transporter (PD_0501) and an arginine deaminase (PD_0517) had SNPs correlated with a temperature variable. In these cases, few isolates acquired the mutation, which then spread within focal populations. One NSY (snp_1446184) was significantly correlated with altitude and minimum temperature in the coldest month, which appeared as a potential limiting factor for pathogen expansion. PD has been reported in areas where winter temperature reaches 1–4°C (Purcell, 1997) but freezing exposures led to the elimination of the disease (Purcell, 1980). These findings provide evidence that epidemiological dynamics can be directly influenced by abiotic factors over short timescales as suggested by Biek and Real (2010). The role of abiotic factors in agroecosystems might be relevant to pathogen adaptation and future studies should consider such interactions.

The present study revealed the genomic structure of a PD-causing X. fastidiosa subsp. fastidiosa population in California, and shed light onto the environmental factors associated with the adaptation of this plant pathogen. Evidence of local adaptation was observed and despite
the presence of a robust population structure, the study supports the single-introduction hypothesis (Nunney et al., 2010). The relatively young age of this population resulted in the formation of distinct genetic clusters displaying signs of homologous recombination. The study of selection and gene–environment associations revealed the presence of traits associated with climatic variables (Branco et al., 2017). Whether local adaptation is favourable or detrimental to pathogenicity remains to be examined.

Experimental procedures

Environmental sampling

We collected samples of European grapevine (Vitis vinifera) plants expressing Pierce’s disease symptoms in commercial vineyards across California (Supporting Information Table S2). A total of 122 isolates were obtained across 900 km in California from five different counties, Temecula (n = 23), Santa Barbara (n = 5), Bakersfield (n = 27), Napa (n = 41) and Sonoma (n = 28). The age of the vines, when available, ranged from 1994 to 2014; all samples were collected in 2015. Isolation in the laboratory was performed on PD3 solid medium (Davis et al., 1981a), followed by triple cloning on PD3. For long-term storage, strains were stored at −80°C in PW broth (Davis et al., 1981b) with 30% glycerol. DNA extraction was performed using commercial kit (DNeasy Blood & Tissue Kit; Qiagen) according to instructions by the manufacturer.

Whole-genome sequencing, SNP calling and ploidy analysis

High molecular weight DNA was extracted, and DNA libraries were prepared for Illumina MiSeq paired-end sequencing. DNA from all samples was sent for sequencing at the QB3 Vincent J. Coates Genomics Sequencing Laboratory. Raw reads and other information regarding each isolate have been submitted to the NCBI database (MiSeq project: SUB3867588). Reads were mapped to X. fastidiosa subsp. fastidiosa Temecula1 (ASM724v1; ENA assembly: GCA_000007245.1). Alignments were performed with the Burrows-Wheeler Aligner (BWA) 0.7.15 aln (Li and Durbin, 2009) with a quality threshold of 15 (Rhodes et al., 2014). FastQs were converted to SAM format using BWA and converted to BAM files, and the BAM files were then sorted and indexed with SAMTOOLS version 1.3.1 (Li et al., 2009). Duplicate reads were marked with PICARD TOOLS (v.2.4.1). The BAM files were processed around insertions or deletions (INDELS) using the GATK RealignerTargetCreator and IndelRealigner (McKenna et al., 2010). Single nucleotide polymorphisms (SNPs) and INDELS were identified using GATK UNIFIEDGENOTYPER version 3.6 in haploid mode (DePristo et al., 2011; Auwera et al., 2013). SNPs and INDELS were filtered to call only high-confidence variants, according to whether they were present in 80% of reads. Resulting variants were mapped to genes using VCF-annotator (Broad Institute, Cambridge, MA) and the latest release of X. fastidiosa subsp. fastidiosa Temecula1 (ASM724v1.36). Mapped reads for each isolate are given on Supporting Information Table S1.

De novo assembly and pan-genome analyses

Data processing was similar as previously done (Vanhove et al., 2019). Genomes were assembled using SPAdes 3.6.0 using the careful parameter (Bankevich et al., 2012); total contig length is summarized in Table S2. progressiveMauve was used to order contigs (Darling et al., 2010) and Prokka used for annotation (Seemann, 2014). A pan genome of the 122 isolates was constructed using Roary (Page et al., 2015). Recombination events were identified by finding regions of enriched SNP density by using ClonalFrameML (Didelot and Wilson, 2015) and fastGEAR (Mostowy et al., 2017). To detect change in ploidy (i.e. duplications), the mean coverage for each isolate was determined using ‘DepthOfCoverage’ from the GATK pipeline under default setting; the subsp. fastidiosa Temecula1 genome was used as a reference and coverage was normalized and averaged over a 500 bp window. Average coverage for each genetic cluster was computed and regions displaying a normalized coverage ≥2 were considered diploid events. The ontology of genes of interests was investigated. GO terms were then assigned using Blast2GO (Conesa and Götz, 2008) using a minimum E value of 1 × 10−10.

Phylogeny, population assignment and molecular dating

Whole-genome SNP files were converted to Nexus and Phylip formats. A maximum likelihood tree was generated with RAxML (1000 bootstrap replicates and a generalized time reversible (GTR) substitution matrix (Stamatakis, 2006), and visualized with FIGTREE v. 1.4 (Rambaut, 2012). Bayesian Analysis of Population Structure (BAPS; Corander et al., 2004) was used to assign isolates to genetic clusters. We investigated the presence of a temporal signal in the data set by using our 120 isolates and an additional 24 previously published subsp. fastidiosa genomes. The non-recombining core genome of strains with known isolation dates (1987–2015, 27 years of evolution) was obtained and used for this analysis (Supporting Information Fig. S7; https://localtempo ralsignal.shinyapps.io/LocalTemporalSignal/). We refer to Vanhove et al. (2019) for details on this analysis, as the
procedures used here are the same as performed in that study.

Population genetics and detection of regions under positive selection

Population level statistics were generated for each *X. fastidiosa* subsp. *fastidiosa* genetic cluster. The number of segregating sites (\(S\)), total number of mutations (\(\pi\)), nucleotide diversity (\(\theta\)), Waterson’s estimator (\(D\)) and Tajima’s \(D\) were estimated using VARISCAN V.2.0 (Vilella et al., 2005) on the core genome alignment without the recombination regions (Table 1). Fixation index (\(F_{ST}\)) statistics were calculated using BEDASSLE (Bradburd et al., 2013). In addition, several statistical methods are available (Vitti et al., 2013) to detect signatures of Darwinian selection. Candidate loci under selection were investigated using \(X_r^2\)X, a population differentiation statistic analogous to \(F_{ST}\) that accounts for variance–covariance of the population using Bayenv2 (Günther and Coop, 2013). On the other hand, gene-base methods compare the rate of synonymous (\(G\)) and nonsynonymous (\(N\)) mutations in protein-coding genes (Yang and Bielawski, 2000). The ratio of the rates of synonymous and non-synonymous substitution (\(d_N/d_S\)) is commonly used to characterize microbial adaptation (Hurst, 2002). The McDonald–Kreitman (MK) test identifies patterns of selection by comparing the number of silent (\(d_N\)) and non-silent substitutions (\(d_S\)) with the number of silent (\(p_S\)) and non-silent polymorphism (\(p_N\)) of an outgroup species (Stoletzki and Eyre-Walker, 2011). A \(d_N/d_S\) analysis and a McDonald–Kreitman (MK) test (McDonald and Kreitman, 1991) were performed using annotated high-confidence SNP mapped to the reference strain Temecula1. SNPs with allele frequency <20% were removed and only genes with \(\geq 5\) SNPs were considered to improve test performance as recommended by Liti et al. (2009)). The *X. fastidiosa* subsp. *multiplex* M12 strain (Chen et al., 2010) was used as an outgroup for the MK test. The \(d_N/d_S\) (\(\omega\)) ratio was estimated using de novo assemblies after removing recombination events (identified with ClonalFrameML), and Codeml from the PAML4.1 package (runmode 0, model 0 was used assuming constant \(d_N/d_S\); Yang, 2007); gene clusters were generated by Roary using gene annotation from Prokka (Seemann, 2014).

Ecological niche modelling, biogeography and environmental association analysis

California provides an appropriate setting to model the geographic distribution of a species (Phillips and Dud’ik, 2008). The software uses presence-only data and climatic variables. The ecological niche of the *X. fastidiosa* distribution was modelled using WorldClim layers v.2 and altitude, which were obtained from the WORLDCLIM database at 30 arc-seconds resolution (Fick and Hijmans, 2017). Each variable was tested for colinearity using a Pearson’s \(r\) \(\leq 0.80\) implemented in the R package ppcor as described by Bosso et al. (2016b). To test model prediction, 25% of the samples were randomly set aside (Supporting Information Fig. S1).

The Mantel test was used to assess the association between genetic and geographic distance among individuals, and to detect spatial autocorrelation (Mantel, 1967). Genetic variation was calculated as the Bray–Curtis distances between loci. The geographic distances were the Euclidean distances between the sampling localities. Mantel tests were performed using the ecodist package (Goslee and Urban, 2007) in R using 10,000 permutations. Environmental factors were extracted from WorldClimv.2.0 layers (Fick and Hijmans, 2017). The information for each sample was extracted in R (version 3.1.1) using the raster (Hijmans and van Etten, 2012) and dismo (Hijmans et al., 2012) packages.

Detection of loci correlated with physical variables was performed using Latent Factor Mixed Model (LFMM; Frichot et al., 2013) and Bayenv2 (Coop et al., 2010). LFMM is a Bayesian approach used to detect selection in landscape genomics. The method investigates the influence of population structure on allele frequencies by introducing unobserved variables as latent factors (Stucki et al., 2014). LFMM provides a way to investigate signatures of local adaptation by identification of high degrees of correlation between polymorphism and environmental variables. To detect signatures of selection, a positive false discovery rate of 0.05 was also applied using the qvalue package (Dabney et al., 2004) in R. BayEnv2 was also used to detect selection using Bayes’ factors (BF; BF \(\geq 3\) and within top 5%) and non-parametric Spearman’s \(R\) distribution (top 5%). To estimate the covariance matrix, three replicates were performed and averaged using 100,000 Monte Carlo Markov Chain (MCMC). To ensure independence between SNPs (Bayenv2 Manual) when computing the covariance matrix, loci identified using LFMM and loci found using the program LDhat, which identifies patterns of linkage disequilibrium using Hudson’s composite likelihood method (McVean et al., 2004), were removed. For both methods, LFMM and BayEnv2, three independent runs were performed using 100,000 MCMC cycles and resulting scores were averaged for each of the climate variable. To perform environmental association analysis (EAA), each variable was averaged, standardized and mean-centred across the population as described in the
Bayenv2 manual. Potential SNP candidates were mapped to the Temecula1 reference genome (ASM724v1) using a Basic Local Alignment Search Tool (BLAST) service obtained from the Universal Protein Resource (UniProt ID: 183190; UniProt, 2017) and Ensembl (Aken et al., 2016).

Biological testing of mutant strains of genes under positive selection or with evidence of recombination

The Materials and Methods section associated with these experiments is described in the Supporting Information.

Acknowledgements

We thank farmers and colleagues Mark Battany, Monica Cooper, Matthew Daugherty, David Haviland, Rhonda Smith, Lucia Varela, from University of California Cooperative Extension for assistance in site selection and sample collection. Funding supporting this research was provided by the California Department of Food and Agriculture PD/GWSS program and Horizon 2020 XF-ACTORS consortium. Anne Sicard is supported by the European Union’s Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 707013. 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


© 2020 Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*, **22**, 2625–2638
large-scale prokaryote pan genome analysis. Bioinformatics 31: 3691–3693.


Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

Appendix S1: Supporting Information