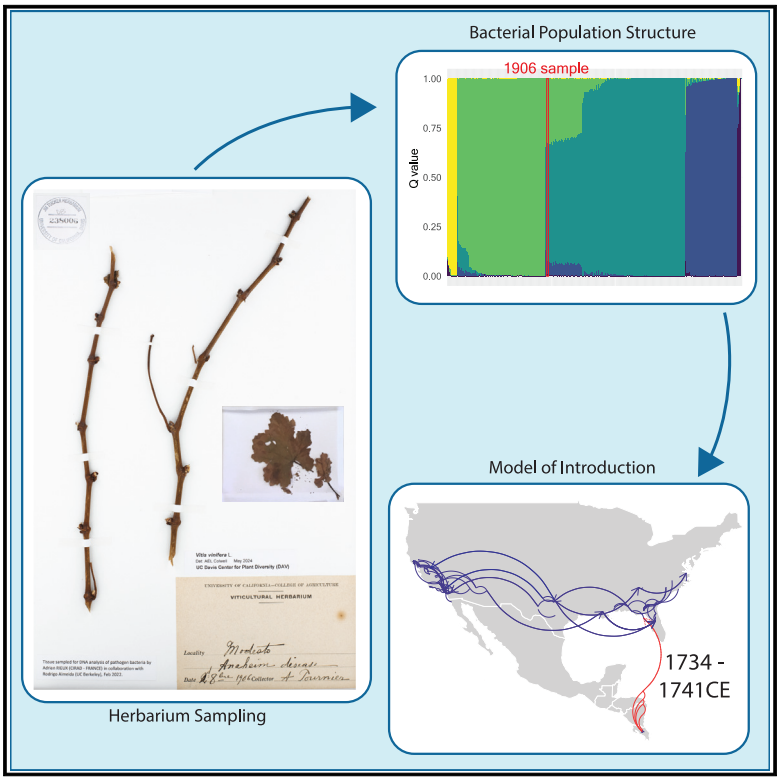
Article

Current Biology

Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas

Graphical abstract



Highlights

- The pathogen *X. fastidiosa* was recovered from a 1906 grapevine herbarium sample
- Deep sequencing enabled the *de novo* assembly of a high-quality aDNA pathogen genome
- The PD-causing pathogen was likely introduced multiple times into California
- Time calibration pushed back the date of introduction to the US to 1734–1741 CE

Authors

Monica A. Donegan, Alexandra K. Kahn, Nathalie Becker, ..., Martial Briand, Rodrigo P.P. Almeida, Adrien Rieux

Correspondence

rodrigoalmeida@berkeley.edu (R.P.P.A.), adrien.rieux@cirad.fr (A.R.)

In brief

Donegan et al. leverage the genome of a century-old specimen of the pathogen *Xylella fastidiosa* to unravel the emergence of Pierce's disease of grapevine across North America. The results recalibrate the timeline of pathogen introduction and challenge the assumption about a single introduction of *X. fastidiosa* from Central America through California.

Donegan et al., 2025, Current Biology *35*, 1–9
January 6, 2025
https://doi.org/10.1016/j.cub.2024.11.029

CellPress

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

Current Biology

CellPress

Article

Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas

Monica A. Donegan,[1,6] Alexandra K. Kahn,[1,6,8] Nathalie Becker,[2,6] Andreina Castillo Siri,[1] Paola E. Campos,[2,3] Karine Boyer,[3] Alison Colwell,[4] Martial Briand,[5] Rodrigo P.P. Almeida,[1,7,9,*] and Adrien Rieux[3,7,*]

[1]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA 94720, USA
[2]Institut de Systématique, Évolution, Biodiversité (ISyEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 50, 75005 Paris, France
[3]CIRAD, UMR PVBMT, La Réunion, 97410 Saint-Pierre, La Réunion, France
[4]Department of Plant Sciences, University of California, Davis, Davis, CA 95818, USA
[5]University of Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, Angers, France
[6]These authors contributed equally
[7]These authors contributed equally
[8]X (formerly Twitter): @alexzebra321
[9]Lead contact
*Correspondence: rodrigoalmeida@berkeley.edu (R.P.P.A.), adrien.rieux@cirad.fr (A.R.)
https://doi.org/10.1016/j.cub.2024.11.029

SUMMARY

Fossils and other preserved specimens are integral for informing timing and evolutionary history in every biological system. By isolating a plant pathogen genome from herbarium-preserved diseased grapevine material from 1906 (Herb_1906), we were able to answer questions about an enigmatic system. The emergence of Pierce's disease (PD) of grapevine has shaped viticultural production in North America; yet, there are uncertainties about the geographic origin of the pathogen (*Xylella fastidiosa* subsp. *fastidiosa*, *Xff*) and the timing and route of its introduction. We produced a high-quality, *de novo* genome assembly of this historical plant pathogen and confirmed degradation patterns unique to ancient DNA. Due to the inclusion of the Herb_1906 sample, we were able to generate a significant temporal signal in the genomic data. This allowed us to build a time-calibrated phylogeny, where we estimate the introduction of *Xff* into the US between 1734 and 1741 CE, an earlier time frame than previously inferred. In a large collection of >300 *Xff* genomes, the Herb_1906 sample was genetically most similar to a small population from Northern California but not basal to the entire *Xff* California clade. Based on phylogenetic placement and a phylogeographic reconstruction, our data support a single introduction of *Xff* into the Southeastern US from Central America, with multiple subsequent introductions into California.

INTRODUCTION

Sequencing methods have quickly advanced in recent years, resulting in vast genomic datasets for contemporary microbial pathogens. However, historical sequencing is typically limited to strains from the late 20th century, isolated in pure culture and preserved until next-generation sequencing became available. Sequencing these older cell cultures has led to improved scientific understanding of contemporary disease emergence and evolution but cannot take us very far into the past. Over the past decade, major advancements in molecular biology and genomics have increased our ability to recover and sequence microbial genomes from ancient host specimens.[1,2] As a result, there has been considerable progress in the reconstruction of historical pathogen outbreaks that caused diseases such as the bubonic plague (*Yersinia pestis*)[3] or the Irish potato famine (*Phytophthora infestans*).[4] In the field of plant pathology, herbarium collections are an important source of dated, identified, and preserved DNA, whose use in comparative genomics and phylogeography can shed light on the emergence and evolutionary history of plant pathogens.[5,6] For example, ancient DNA (aDNA) was recently used to elucidate the biogeographic history of the economically important citrus pathogen *Xanthomonas citri* pv. *citrii*.[7] Herbarium-derived DNA has been shown to bear biochemical characteristics such as high fragmentation levels and distinct patterns of DNA misincorporations, allowing authentication of the historical sequences generated.[6] Interestingly, damage decay of DNA in herbarium samples has been found to occur about six times faster than in bones, meaning that younger herbarium samples need to be treated as aDNA compared with museum specimens of other taxa.[8]

The globally emerging plant pathogen *Xylella fastidiosa* (*X. fastidiosa*) is a bacterial species composed of multiple subspecies with distinct geographic origins. One important

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

outbreak of this pathogen has resulted in the emergence of Pierce's disease (PD) of grapevines, caused by *X. fastidiosa* subsp. *fastidiosa (Xff)*. PD was first noted as a new disease in 1884 in the vineyards of Anaheim, California[9]; in the Southeastern United States (US), early 20th century attempts to establish vineyards were unsuccessful due to disease pressure.[10] There are two main hypotheses regarding the biogeographic history of this subspecies, with conflicting narratives. An early hypothesis proposed that the PD clade is native to the Southeastern US, inferred from the PD resistance of some native *Vitis* species, which varied along a temperature gradient.[10,11] However, there are many examples of naive hosts showing resistance to pathogens due to other physiological mechanisms aside from a co-evolutionary history with the pathogen. Muscadine grapevines, for example, are resistant to many non-co-evolved pests and diseases.[12] Additionally, multiple *Vitis* species native to Mexico also show PD resistance, suggesting that resistance to PD may be correlated with drought tolerance traits.[13] A more recent alternative hypothesis posits that *Xff* was introduced into North America from Central America at some point in the past few hundred years.[14] Genetic and genomic work has shown much greater genetic diversity of the subspecies across Costa Rica than across North America, suggesting a longer history of *Xff* in Costa Rica. Phylogenetic analyses show that the clade of *Xff* with PD strains is nested within a Costa Rican clade.[14] These results imply that the strains present in North America are the result of a genetic bottleneck that has not had sufficient time to diversify to the extent that it has in Central America.[15,16] A previous study performed on contemporary *Xff* genomes estimated the age of the most recent common ancestor (MRCA) of the California clade at 1960 CE (95% confidence interval [CI]: 1851–1976) and identified local adaptation across regions in California that provided support for a single introduction event.[17] There are two genetically distinct clades of *Xff* in the Southeastern US and west coast US (California), but several Southeastern US strains cluster with Californian *Xff* strains.[15] Additional *Xff* strains were recently discovered in Virginia that cluster with California strains.[18,19] One possibility is that, following a single introduction event from Central America, *Xff* populations rapidly diverged on the two coasts, or there may have been multiple introductions between grapevine-growing regions within the US.

In this study, we sequenced and assembled the first historical genome of *X. fastidiosa* from historical material. Besides expanding the temporal range of existing *Xff* genomic datasets (which date from the 1980s), adding ancient sequences helps to elucidate the past evolutionary history of the subspecies and resolve uncertainties about the timing and directionality of its movement in the US. For instance, we would expect a historical (~1900) California *Xff* strain to be basal to the California clade, given the history of PD emergence in the 1880s and previous MRCA date of 1960 CE. Alternatively, if *Xff* first established in the Southeastern US following an introduction event, there may have been one or more introductions into California resulting from plant trade. Understanding the history of PD in the US is particularly important because outbreaks of PD and other *X. fastidiosa*-caused diseases are emerging throughout Europe and the Mediterranean basin.

RESULTS

Detecting, sequencing, and authenticating *X. fastidiosa* from herbarium tissue

Ten samples of cultivated grapevine (*Vitis vinifera* [*V. vinifera*]) material were tested from a historic viticulture collection housed at the Center for Plant Diversity Herbarium at the University of California, Davis. These 10 specimens were selected based on disease annotations and visual symptoms (Table S1). One specimen, accession #DAV238006, tested positive for *X. fastidiosa* by qPCR (average C_T [cycle threshold] values of 23, corresponding to 52,500 copies/μL DNA). The C_T values were undetermined (>40) for the nine other samples, as well as for the two negative controls. Specimen DAV238006 was originally collected on August 2, 1906 in Modesto, California, which is around the reported start of grape and alfalfa growing and irrigated farming in that area.[20] The specimen retained some leaf tissue and several matchstick (without leaves) petioles, a typical symptom of PD, and was annotated "Anaheim disease" by its collector, A. Tournier (Figure 1A). Total DNA extracted from this specimen (yield of 2.6 ng per mg of petiole tissue) was converted into an Illumina library and sequencing generated 65 M paired-end reads with a base call accuracy of 99.90% to 99.96%. Importantly, no *V. vinifera* or *Xff*-specific DNA fragments were found in our *Xff*-negative herbarium-acquired *Coffea* sp. control, thus ruling out in-lab contamination. Because plant metagenomes contain DNA originating from multiple sources, we used a BLASTn approach to reconstruct the metagenomic composition of the reads obtained from the historical DAV238006 specimen. Identified sequences mostly consisted of the *Vitis* plant host genus (63.3%) and *Xff* (32.3%). Other identified taxa (0.60%) included *Homo sapiens* (0.5%) and a mixture of bacterial, viral, and fungal genera (totaling less than 0.1% of all reads). The remaining reads (3.8%) were not assigned to any known taxa (Figure 1B).

aDNA typically presents short fragments and cytosine deamination at fragment extremities.[21] We analyzed these degradation patterns using the dedicated tool mapDamage2.[22] *Xff* reads retrieved from historical DAV238006 specimen displayed mean fragments lengths of 125.31 ± 46 nt, and 3′G > A substitution rates at terminal nucleotides of 2.43%, decreasing exponentially along the DNA molecule (Figure 1C). Examining 5′C > T substitution rates gave similar results (not shown). We performed the same analyses for the reads mapping to the *V. vinifera* host genome. Mean read length was estimated at 135.48 ± 45 nt and 3′G >A substitution rate at terminal nucleotides averaged at 2.47%.

Genome assembly confirms the presence of *X. fastidiosa* in 1906 grapevine specimen

A *de novo* assembly was produced after correcting and trimming the raw reads for DNA damage and filtering them by mapping to a large set of modern *X. fastidiosa* genomes (see STAR Methods). The resulting assembly (hereafter called Herb_1906) contained 195 contigs (max length: 390,697; mean length: 12,597), had extremely high (2,036×) coverage, an N_{50} (the sequence length of the shortest contig at 50% of the assembly) of 103,051, and an L_{50} of 7 (minimum number of contigs to cover 50% of the assembly), indicating overall good quality, similar to modern *Xff* draft genomes assembled from Illumina reads (Table 1). The Herb_1906 genome scored highly on genome completeness (99.638%) and included all seven MLST (multilocus sequence typing) genes

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

CellPress

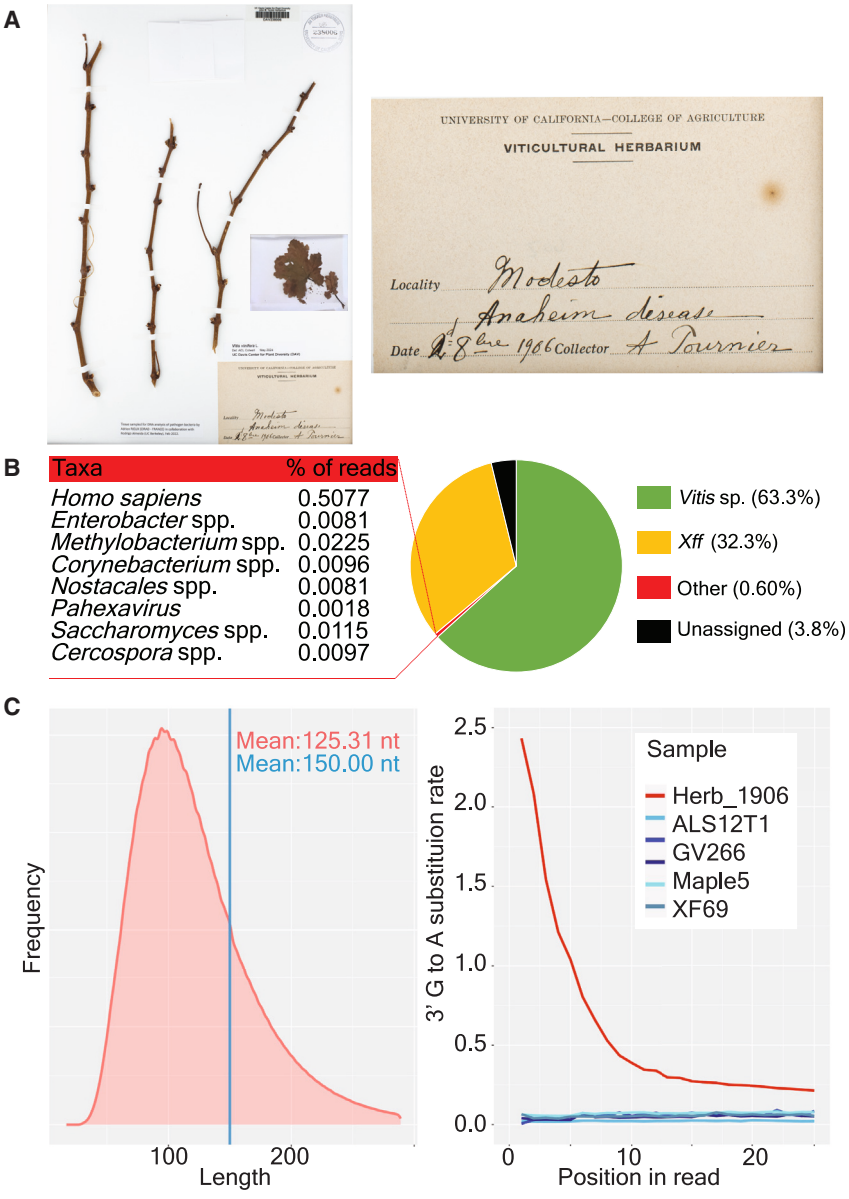


Figure 1. Discovery and validation of *Xff* from a 1906 herbarium specimen

(A) Photography of the original DAV238006 sample (Herbarium sample from which Herb_1906 was sequenced) including an inset (right) of the original photo with the annotation of "Anaheim disease" from Modesto.

(B) Metagenomic composition of Herb_1906 historical specimen obtained using a blastn assignment against the nucleotide (NCBI-nt) database.

(C) Herb_1906_*Xff* post-mortem DNA damage patterns measured both on the historical Herb_1906 genome (red) and five modern *Xff* strains (blue). Left: fragment length distribution (in nt). The blue line depicts read length from modern samples fragmented for sequencing with Illumina NovaSeq (150 bp). Right: deamination rates (G to A substitutions) of the first 25 nucleotides from the 3′ end.

used for typing *X. fastidiosa*.[23] Based on the MLST profile, Herb_1906 fits in ST1 (sequence type 1), which is common for California grapevine-infecting *Xff* strains.

Phylogenetic placement of Herb_1906 within the clade of California strains

We estimated ancestral populations within subspecies *fastidiosa* using sparse nonnegative matrix factorization (sNMF),[25] resulting in five ancestral clusters (k = 5) based on cross-entropy (Figure S1). Costa Rica strains split into two distinct clusters (V1 and V5), indicating high genetic diversity (Figure 2A). The herbarium strain was assigned to cluster V3 (q = 0.594), but showed some admixture as the coefficient of V2 was 0.329 (Figure 2A). A maximum likelihood tree (Figure S2) constructed from 331 *Xff* strains from the USA, Costa Rica, Spain, Taiwan, and Mexico revealed a high relatedness of Herb_1906 to a clade of 37 strains within Mendocino County, in Northern California (Figure 2D). Surprisingly, the herbarium strain was basal to this small clade from Mendocino but not basal to either of two major clades, including grapevine-infecting *Xff* in California (ancestral cluster V2 and V3; Figure 2D). That is, Herb_1906 was nested within California *Xff* and did not predate the split between the two major *Xff* clusters in California. Additionally, a strain isolated from grapevines in Florida collected in 1989 was basal to the Herb_1906/Mendocino clade (Figure 2D). This group, as well as related strains from Napa, Sonoma, and Temecula CA (V3), are more closely related to strains from Virginia (MAG_669) and from Georgia (XF51-CCPM1) than from other strains in California (V2). This suggests that California strains in group V3 (including Herb_1906) originated from a common ancestor shared with modern strains on the East Coast of the US. This common ancestor was not shared with all California *Xff* strains (i.e., strains in V2), indicating multiple introductions between the Southeastern US and California.

Tip-dated phylogeny indicates an earlier introduction than previously inferred

The maximum likelihood (ML) tree was used to test the presence of a temporal signal within the *Xff* clade. The linear regression test between root-to-tip distances and sampling ages displayed a significantly positive slope (value = 0.000054, adjusted R^2 = 0.011 with a p value = 0.033), reflecting the progressive accumulation of mutations over time. To specifically evaluate the contribution of historical genome Herb_1906 to the magnitude of temporal signal, we repeated the above test on a dataset containing modern genomes only, producing no temporal signal (p value = 0.37) (Figure 2B). In addition, the dataset passed the date-randomization test, as the inferred root age from the real versus date-

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

Table 1. Validation of Herb_1906 genome assembly

–	Herb_1906	Temecula1	Mendo20_D06
Assembly Size (bp)	2,456,318	2,521,148	2,481,353
GC %	51.45%	51.78%	51.487%
Predicted Coding Sequences	2,425	2,427	2,633
N50	103,051	2,519,802	97,460
L50	7	1	9
Completeness	99.638%	99.638%	99.638%
Contamination	0.023%	0%	0%
Average Fold Coverage	2,036×	26×[24]	223×

Genome assembly statistics. Herb_1906 is the genome recovered from an herbarium sample. Temecula1 is a reference closed genome and coverage was calculated in the original paper.[24] For Herb_1906 and Mendo20_D06, average fold coverage was estimated using Bbmap using Temecula1 as the reference; Mendo20_D06 is a modern Illumina-sequenced draft genome published in this study.

randomized datasets exhibited no overlap of the 95% highest posterior density (Figure S3). A Bayesian time-calibrated tree was built with BEAST (Bayesian Estimated Analysis by Sampling Trees) (Figure 2D), which was globally congruent (similar topology and node supports) with the ML tree (Figure S2). Importantly, posterior estimates of root age and substitution rate were robust to the choice of various demographic models (Figure S4). The root of the *Xff* clade was dated at 143 CE (95% HPD, highest posterior density: −398–624). The node splitting all Central American strains from strains in the US was estimated at 1734 CE (95% HPD: 1681–1800); the MRCA of all American strains was dated at 1741 CE (95% HPD: 1688–1806); the MRCA for the clade including all California subsp. *fastidiosa* strains, as well as some from the Southeastern US and Spain, was placed at 1824 CE (95% HPD: 1759–1868). Herb_1906 diverged from the modern population in Mendocino, CA in 1897. We estimated a mean substitution rate of 2.84×10^{-6} (95% HPD: $2.28 \times 10^{-6} - 3.84 \times 10^{-6}$) per site per year with a standard deviation for the uncorrelated log-normal clock of 1.99 [95% HPD: 1.80–2.17], suggesting high-rate heterogeneity within the *Xff* clade.

Phylogeographic diffusion model supports Central American origin of the pathogen and multiple California introductions

An ancestral-state reconstruction method was applied to the maximum likelihood phylogeny of North and Central American strains with Herb_1906 in order to make inferences about the geographic locations of ancestral nodes. The location of the ancestral node splitting US strains from Costa Rican strains was centered in Central America (Figure 3), although the 95% CI includes the southeast US. (Figure S5). Based on the inferred coordinates of ancestral nodes, the model suggests a single introduction event from Central America directly to the Southeastern US, although this may be influenced by missing data from across Mexico (see discussion). From the Southeastern US, *Xff* appears to have spread west-ward. Interestingly, the model includes multiple introductions into California (Figure 3), contradicting previous hypotheses about a single introduction to California.[17] Additional long-distance dispersal events include re-introductions of some clades of *Xff* back to the East Coast (i.e., CFBP8177).

Comparative genomics to modern strains

When comparing Herb_1906 with all the strains included in our study, a gene presence/absence analysis showed that no genes were uniquely absent, or present, in Herb_1906 in contrast to the rest of subspecies *fastidiosa*. We then focused on genetic variants between Herb_1906 and the closely related clade from Mendocino County, California (n = 37) and identified 17 gene gain and loss events. Eight coding regions were present in Herb_1906 and absent in the Mendocino clade (*prtR*, *cya*, and six hypothetical proteins), whereas nine coding regions were present in all 37 members of the Mendocino clade and absent in Herb_1906 (nine hypothetical proteins) (Table S4). The eight coding regions present in Herb_1906 were also absent in other strains than those of the Mendocino clade. Seven of these coding regions were totally absent from different combinations of one to three ancestry groups (V1 to V5), whereas *cya* displayed a more disparate presence/absence pattern among all ancestry groups (Figure S6).

DISCUSSION

In this study, we successfully reconstructed an ancient genome of the crop pathogen *X. fastidiosa* from an infected *V. vinifera* sampled in California in 1906 and preserved as an herbarium specimen. Previously, the oldest *Xff* genomes were produced by sequencing pure bacterial cultures isolated in the 1980s[26] and conserved until next-generation sequencing became available.[27,28] Using comparative and population genomics methods, we compared Herb_1906 with a set of 330 modern genomes representative of the bacterial subspecies global diversity, 73 of them having been generated during the course of this study. A better understanding of *X. fastidiosa's* evolutionary history is a subject of great interest because it may help to decipher how this devastating plant pathogen specializes on its hosts and diversifies while expanding its geographical range.

Adopting a shotgun-based deep sequencing strategy allowed us to describe the metagenomic composition of our 1906 *V. vinifera* herbarium specimen. As expected, most of the assigned sequences (63.3%) originated from the host plant, in accordance with previous studies that performed deep sequencing of herbarium specimens.[29] One surprising result was the recovery of 32.3% of reads mapping to *X. fastidiosa*. For modern metagenomes of *X. fastidiosa* from a plant sample, the percentage of reads mapping to a *X. fastidiosa* reference has been <3% across studies done in grapevine, olive, and other host plants.[19,30,31] Such a result could be explained by differential rates of DNA degradation between vascular pathogens and plant hosts; in particular, pathogens embalmed in the xylem may be better preserved from postmortem DNA damage. However, when investigating this hypothesis, we did not find any significant difference in DNA damage between *Xff* and its host plant. Previous aDNA studies of herbarium samples recovered between 0.8%–27% of reads mapping to the bacterial pathogen *Xanthomonas citri*[7] and between 1%–20% in a study of *Phytophthora infestans*.[32] Our sample is above the upper range of these

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

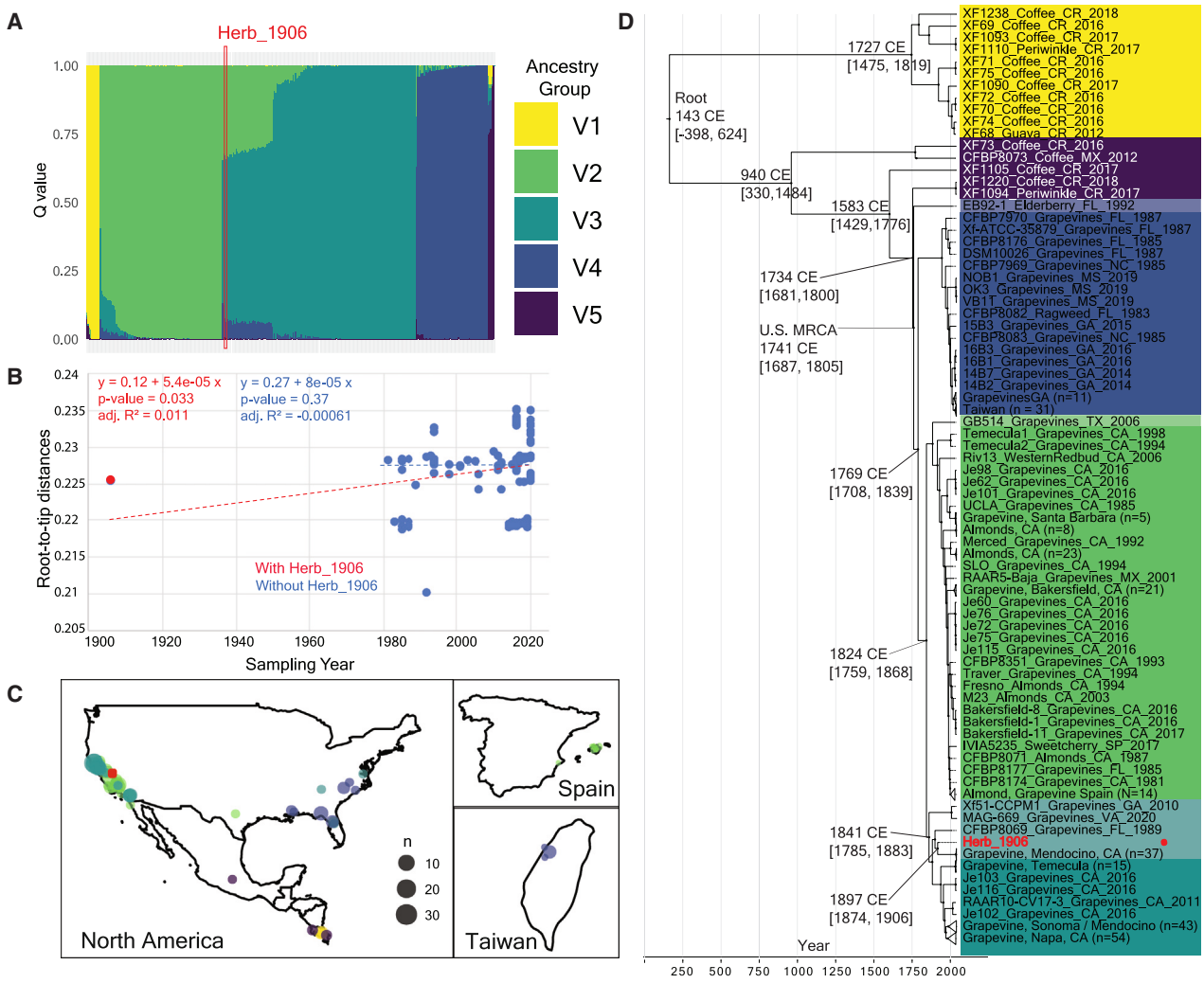


Figure 2. Phylogenomic analyses place Herb_1906 within a clade from California

(A) Ancestry coefficient (Q) bar plot for all 331 *Xff* genomes inferred with sNMF (k = 5; see Figures S1 and S7 for k selection). Ancestry groups (V1 through V5) align with the phylogenetic placements of strains in (D).

(B) When integrating Herb_1906, a significant root-to-tip linear regression is yielded ($p = 0.033$), plotted by a red dashed line. The blue points and trendline depict the dataset without Herb_1906, whereas the red trendline shows the dataset including Herb_1906.

(C) Locations of all 331 *Xff* strains in the dataset; nearby isolates were clustered, and size of the dot denotes sample size. Location of Herb_1906 in Modesto, CA is marked with a red dot.

(D) Bayesian time-calibrated tree built from a maximum likelihood tree including 331 *Xff* strains. Median node ages are shown for major nodes, as well as the 95% HPD. Herb_1906 is highlighted in red. Clades are color coded by assigned ancestry groups; strains assigned to a group with lower certainty (q = 0.5 to 0.7) are colored with a lighter shade of the same color.

See also Figures S1, S2, and S7.

studies and abnormal for a notably fastidious pathogen such as *X. fastidiosa* in particular, requiring further study. Fortunately, the high recovery rate likely contributed to our ability to reconstruct the whole historical genome with high coverage following a *de novo* assembly strategy.

The presence of temporal structure is an essential prerequisite to perform tip-calibrated inference.[33–35] Although the dataset containing contemporary genomes only (1983–2020) did not reveal the existence of any measurably evolving population, inclusion of the 1906 historical genome brought the required temporal signal within the *Xff* clade. This allowed us to build a time-calibrated phylogeny without making any underlying assumption on the age of any node in the tree, nor on the rate of evolution in order to propose new evolutionary scenarios for the origin of the pathogen. Previous studies that did not use historical strains had to first estimate a rate of evolution within a smaller clade displaying temporal signal before extrapolating the rate to the whole subspecies,[16,17] a strategy known to yield less accurate and robust estimates.[36] We inferred a mean substitution rate of 2.84×10^{-6} substitutions per site per year, a value around 4.5× faster than the one (6.37×10^{-7}) previously obtained at the Californian scale only.[17] We dated the MRCA of all *Xff* strains

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

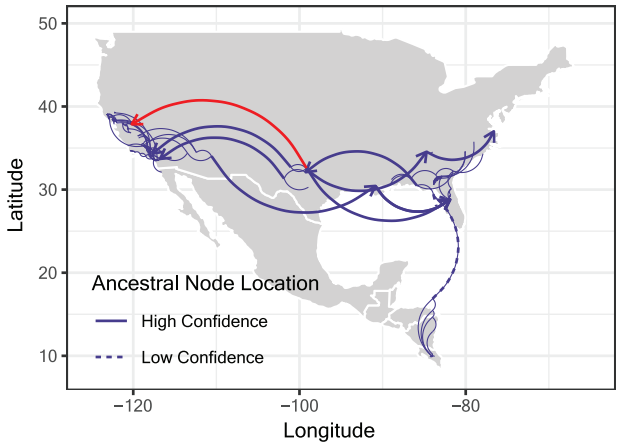


Figure 3. Ancestral-state reconstruction traces likely paths of *Xff* introductions

Ancestral-state reconstruction of geographic locations of internal nodes of maximum-likelihood phylogenies. Lines depict inferred locations of connected nodes in the phylogeny; arrows in bold depict long-distance dispersal events (see STAR Methods). Arrow line types depict confidence (area of the 95% CI for latitude/longitude) in the location of the parent node. The red arrow is the node preceding the Herb_1906 and Mendocino clade, indicating an introduction to Modesto.
See also Figure S5.

present in the USt 1741 CE (95% HPD: 1687–1805). This value, which predates the first report of the disease in California by approximately 150 years,[9] provides a slightly older time span than the US MRCA (1806) previously inferred.[17] Hence, our results show that *Xff* has been present in the US longer than expected but not long enough ago to support the hypothesis that *Xff* is endemic to the Gulf coast of the US; that is, not long enough for native grapevines to evolve resistance as a response to disease pressure.[11] The common ancestor of the whole *Xff* clade, including Costa Rican strains, by contrast, was dated to be approximately 2,000 years old, reflecting a long history of diversification among those strains. Although it is possible that additional effort to search for subspecies *fastidiosa* in poorly sampled regions may uncover novel and more basal phylogenetic clades, there is currently no evidence supporting the hypothesis that PD is endemic to the US.

There has been uncertainty about the history of pathogen dispersal within the US between the east and west coasts.[37] First reports of the disease, originally called Anaheim disease, were from the Santa Ana Valley in Southern California in ~1886.[38] In the following years, it was detected in the San Joaquin, Napa, and Sonoma Valleys. Corresponding with the order of detection, it was believed to have moved northward from one source within California. Based on the topological placement of the 1906 historical strain in the center of the phylogeny, as well as on the ancestral-state reconstruction, our results suggest that this was not the case. Instead, there were likely multiple introductions from the east coast/south central region of the US to locations across the state of California around the same time. Otherwise, we would expect the 1906 sample to fall near the base of the phylogenetic tree. The gold rush in the mid 1800s caused massive population and resource influx into Northern California, along with a boom of the grape industry, which quadrupled in size between the 1860s and the 1880s.[39] Many varieties were brought across the country at that time to establish vineyards in California, which could have caused multiple introductions in quick succession. It has been hypothesized that heavy rains precipitated the first PD epidemic in California from 1884–1894 by causing increased populations of insect vectors, triggering that first epidemic from pathogen load already present in the environment. Additionally, our updated hypothesis of an initial introduction of *Xff* to the Southeastern US, rather than California, fits the timeline of grapevine disease in the Southeastern US in the early 1800s.[10] Previous work[14,37,40] has also noted that their data could support an introduction from Central America to the Southeastern US.

The addition of a historical genome from California demonstrates that a seemingly straightforward hypothesis—i.e., a northward expansion of *Xff* after a single introduction in California—is inconsistent with the genomic data. Similarly, it is tempting to speculate about a Northern expansion of *Xff* from Costa Rica through Central America and into the US. However, the current whole-genome dataset suffers from lack of sampling from large geographic areas where *Xff* is known to occur. For example, only two *Xff* whole genomes are available from Mexico: CFBP8073 is a strain detected from a Robusta coffee plant (*Coffea canephora*) that was imported to France from Mexico,[41] although the location of origin within Mexico was unknown. This strain clusters with strains from Costa Rica (Figure 2), indicating northward expansion. Another *Xff* strain from Baja California, Mexico (RAAR5_Baja) was similar to strains from California, and the phylogeographic model indicated an introduction to Baja California from California. Aside from whole-genome data, researchers recently used MLST to detect additional diversity of *Xff* within vineyards in central Mexico, distinct from strains in Baja California.[42] Additional sampling will require thoughtful, international collaboration between researchers to unravel the historical routes of introduction.

Beyond the overarching introduction route of *Xff*, the historical genome allowed us to look at specific gene content changes in modern populations. In a closely related pathogen, gene gain and loss of a particular locus has recently been shown to act as a phenotypic switch, i.e., between vascular and nonvascular plant pathogenesis.[43] The phenotypic relevance of gene content changes in plant pathogens underscores the importance of looking for such patterns among *X. fastidiosa* ancestral groups. Among the eight genes present in Herb_1906 but absent in its sister clade from Mendocino County, *prtR* and *cya* have annotated functions. *prtR* encodes a transcriptional regulator, involved in multiple cellular processes, such as protease expression during cell growth at high temperatures in *Pseudomonas fluorescens*,[44] or at several levels during transitions between acute and chronic infections in *P. aeruginosa*.[45,46] *cya* has been previously identified as under positive selection in *X. fastidiosa*.[28] *prtR* and *cya* have also been found to modulate bacterial virulence by affecting type III secretion systems, which *X. fastidiosa* notably lacks.[47–49] *cya* was absent in 259/331 strains of this study, throughout all (V1 to V5) ancestry groups. *prtR* was even less present (8/331 strains), across V2 to V4 groups only (Figure S6). These two genes, identified as accessory in *X. fastidiosa*,[28] could be involved in adaptation to complex ecological conditions,[43] possibly local environments in California. Understanding the function of these two gene losses in specific *Xff* populations will require further study.

In conclusion, this study emphasizes the power of herbarium collection to elucidate the history of pathogen spread. Sequencing and analyzing the first historical genome of *Xff* shifted the US introduction date of this pathogen back 70 years earlier than previously suggested. In addition, phylogenetic inferences suggest that there may have been multiple introductions of this pathogen into California, where previously there was only evidence of a single introduction. However, there is still considerable uncertainty regarding the global spread of *X. fastidiosa*, due to large geographic and chronological gaps in sample collection. We hope that this work will encourage further herbarium sampling to increase both the temporal and spatial coverage of *X. fastidiosa* datasets, which, in turn, may improve the accuracy of the reconstruction of pathogen evolutionary history.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed toward the lead contact, Rodrigo P.P. Almeida (rodrigoalmeida@berkeley.edu).

Materials availability

There are restrictions to the availability of modern strains because of the lack of an external centralized repository for their distribution and our need to maintain the stock. We are glad to share strains with reasonable compensation by requestor for its processing and shipping, along with the appropriate APHIS (Animal and Plant Health Inspection Service) permits.

Data and code availability

The authors confirm that all data used in this study are fully available without restriction. Raw reads and consensus genomes were deposited to the NCBI Sequence Read Archive (SRA) and GenBank, respectively, under accession numbers listed in the supplemental information. Data for the historical sample (assembly and reads) can be found in NCBI: PRJNA1114123; other newly generated data are deposited in NCBI: PRJNA786005 and NCBI: PRJNA722088. Accession numbers of any previously published data used in this study are also listed in supplemental information. Images and label data for the herbarium specimens used in this study are found on the Consortium of California Herbaria website www.cch2.org using the accession numbers listed in Table S1. Code for the bioinformatics pipeline is available at Github: https://github.com/MonicaDonegan/aDNA-xylella-herbarium.

ACKNOWLEDGMENTS

We are grateful to O. Pruvost, P. Lefeuvre, E. Clark, A. Brown, K. Newman, P. Lee, and I. McCabe for valuable comments and discussions. We thank Brent Mishler and Ana Penny (The University and Jepson Herbaria) for allowing us to consult and sample historical plant specimens. Collection of any plant material used in this study complies with institutional, national, and international guidelines. Permission to collect and analyze each historical specimen included in this study was provided by the herbarium institutions (and their curators) from which they were sampled. This work was financially supported by Agence Nationale de la Recherche (JCJC MUSEOBACT contract ANR-17-CE35-0009-01), the Thomas Jefferson Fund (TJF19-21), the European Regional Development Fund (ERDF contract GURDT I2016-1731-0006632), the Conseil Régional de La Réunion, the Centre de Coopération Internationale en Recherche Agronomique pour le Développement(CIRAD), the CDFA PD/GWSS Board, and the BeXyl Project (HORIZON ID: 101060593). Computational work was performed on both the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley and on the MESO@LR-Platform at the University of Montpellier (https://hal.umontpellier.fr/MESO). Laboratory work on the historical sample was conducted on the Plant Protection Platform (3P, IBISA).

AUTHOR CONTRIBUTIONS

Conceptualization, N.B., A.C.S., P.E.C., R.P.P.A., and A.R.; resources, A.C.; investigation, M.A.D., A.K.K., N.B., K.B., R.P.P.A., and A.R.; formal analysis, M.A.D., A.K.K., N.B., M.B., and A.R. writing—original draft, M.A.D. and A.K.K.; writing—review & editing, M.A.D., A.K.K., N.B., A.C., R.P.P.A., and A.R.; supervision, R.P.P.A. and A.R.; funding acquisition, N.B., A.C.S., A.K.K., R.P.P.A., and A.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Sampling of historical specimens
  - Sampling, Sequencing, and Assembly of 77 Modern Isolates
- METHOD DETAILS
  - DNA Extraction and qPCR
  - Sequencing and Reads Initial Trimming
  - Ancient DNA Damage Assessment and Correction
  - Taxonomic Assessment
  - *De novo* Assembly of the Historical Genome
  - Population Genomics and Phylogenetics
  - Molecular Clock
  - Phylogeographic Ancestral State Reconstruction
  - Comparative genomics
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.11.029.

Received: June 13, 2024
Revised: October 1, 2024
Accepted: November 18, 2024
Published: December 16, 2024

REFERENCES

1. Spyrou, M.A., Bos, K.I., Herbig, A., and Krause, J. (2019). Ancient pathogen genomics as an emerging tool for infectious disease research. Nat. Rev. Genet. *20*, 323–340. https://doi.org/10.1038/s41576-019-0119-1.
2. Duchêne, S., Ho, S.Y.W., Carmichael, A.G., Holmes, E.C., and Poinar, H. (2020). The recovery, interpretation and use of ancient pathogen genomes. Curr. Biol. *30*, R1215–R1231. https://doi.org/10.1016/j.cub.2020.08.081.
3. Swali, P., Schulting, R., Gilardet, A., Kelly, M., Anastasiadou, K., Glocke, I., McCabe, J., Williams, M., Audsley, T., Loe, L., et al. (2023). *Yersinia pestis* genomes reveal plague in Britain 4000 years ago. Nat. Commun. *14*, 2930. https://doi.org/10.1038/s41467-023-38393-w.
4. Martin, M.D., Cappellini, E., Samaniego, J.A., Zepeda, M.L., Campos, P.F., Seguin-Orlando, A., Wales, N., Orlando, L., Ho, S.Y.W., Dietrich, F.S., et al. (2013). Reconstructing genome evolution in historic samples of the Irish potato famine pathogen. Nat. Commun. *4*, 2172. https://doi.org/10.1038/ncomms3172.
5. Yoshida, K., Burbano, H.A., Krause, J., Thines, M., Weigel, D., and Kamoun, S. (2014). Mining herbaria for plant pathogen genomes: back to the future. PLoS Pathog. *10*, e1004028. https://doi.org/10.1371/journal.ppat.1004028.

6. Burbano, H.A., and Gutaker, R.M. (2023). Ancient DNA genomics and the renaissance of herbaria. Science *382*, 59–63. https://doi.org/10.1371/journal.ppat.1004028.

7. Campos, P.E., Pruvost, O., Boyer, K., Chiroleu, F., Cao, T.T., Gaudeul, M., Baider, C., Utteridge, T.M.A., Becker, N., Rieux, A., et al. (2023). Herbarium specimen sequencing allows precise dating of *Xanthomonas citri* pv. *citri* diversification history. Nat. Commun. *14*, 4306. https://doi.org/10.1038/s41467-023-39950-z.

8. Weiß, C.L., Schuenemann, V.J., Devos, J., Shirsekar, G., Reiter, E., Gould, B.A., Stinchcombe, J.R., Krause, J., and Burbano, H.A. (2016). Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R. Soc. Open Sci. *3*, 160239. https://doi.org/10.1098/rsos.160239.

9. Pierce, N.B. (1892). The California vine disease: a preliminary report of investigations. US Government Printing Office.

10. Hewitt, W.B. (1958). The probable home of Pierce's disease virus. Am. J. Enol. Vitic. *9*, 94–98. https://doi.org/10.5344/ajev.1958.9.2.94.

11. Ruel, J.J., and Walker, M.A. (2006). Resistance to Pierce's disease in *Muscadinia rotundifolia* and other native grape species. Am. J. Enol. Vitic. *57*, 158–165. https://doi.org/10.5344/ajev.2006.57.2.158.

12. Bouquet, A. (1983). Etude de la résistance au phylloxéra radicicole des hybrides Vitis vinifera x *Muscadinia rotundifolia*. Vitis *22*, 311–323.

13. Riaz, S., Tenscher, A.C., Heinitz, C.C., Huerta-Acosta, K.G., and Walker, M.A. (2020). Genetic analysis reveals an east-west divide within North American *Vitis* species that mirrors their resistance to Pierce's disease. PLoS ONE *15*, e0243445. https://doi.org/10.1371/journal.pone.0243445.

14. Nunney, L., Yuan, X., Bromley, R., Hartung, J., Montero-Astúa, M., Moreira, L., Ortiz, B., and Stouthamer, R. (2010). Population genomic analysis of a bacterial plant pathogen: novel insight into the origin of pierce's disease of grapevine in the U.S. PLOS One *5*, e15488. https://doi.org/10.1371/journal.pone.0015488.

15. Castillo, A.I., Chacón-Díaz, C., Rodríguez-Murillo, N., Coletta-Filho, H.D., and Almeida, R.P.P. (2020). Impacts of local population history and ecology on the evolution of a globally dispersed pathogen. BMC Genomics *21*, 369. https://doi.org/10.1186/s12864-020-06778-6.

16. Vanhove, M., Retchless, A.C., Sicard, A., Rieux, A., Coletta-Filho, H.D., De La Fuente, L., Stenger, D.C., and Almeida, R.P.P.A. (2019). Genomic diversity and recombination among *Xylella fastidiosa* subspecies. Appl. Environ. Microbiol. *85*, 1–17. https://doi.org/10.1128/AEM.02972-18.

17. Vanhove, M., Sicard, A., Ezennia, J., Leviten, N., and Almeida, R.P.P. (2020). Population structure and adaptation of a bacterial pathogen in California grapevines. Environ. Microbiol. *22*, 2625–2638. https://doi.org/10.1111/1462-2920.14965.

18. Abdelrazek, S., Bush, E., Oliver, C., Liu, H., Sharma, P., Johnson, M.A., Donegan, M.A., Almeida, R.P.P., Nita, M., and Vinatzer, B.A. (2024). A survey of *Xylella fastidiosa* in the US state of Virginia reveals wide distribution of subspecies *fastidiosa* and infection of grapevine by subspecies *multiplex*. Phytopathology *114*, 35–46. https://doi.org/10.1094/PHYTO-06-23-0212-R.

19. Johnson, M.A., Liu, H., Bush, E., Sharma, P., Yang, S., Mazloom, R., Heath, L.S., Nita, M., Li, S., and Vinatzer, B.A. (2022). Investigating plant disease outbreaks with long-read metagenomics: sensitive detection and highly resolved phylogenetic reconstruction applied to *Xylella fastidiosa*. Microb. Genom. *8*, 1–14. https://doi.org/10.1099/mgen.0.000822.

20. Barnes, D.H. (1987) The Greening of Paradise Valley. Modesto Irrigation District. 442–452.

21. Dabney, J., Meyer, M., and Pääbo, S. (2013). Ancient DNA damage. Cold Spring Harb. Perspect. Biol. *5*, a012567. https://doi.org/10.1101/cshperspect.a012567.

22. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193.

23. Scally, M., Schuenzel, E.L., Stouthamer, R., and Nunney, L. (2005). Multilocus sequence type system for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation to clonal diversity. Appl. Environ. Microbiol. *71*, 8491–8499. https://doi.org/10.1128/AEM.71.12.8491-8499.2005.

24. Van Sluys, M.A., de Oliveira, M.C., Monteiro-Vitorello, C.B., Miyaki, C.Y., Furlan, L.R., Camargo, L.E., da Silva, A.C., Moon, D.H., Takita, M.A., Lemos, E.G., et al. (2003). Comparative analyses of the complete genome sequences of pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J. Bacteriol. *185*, 1018–1026. https://doi.org/10.1128/JB.185.3.1018-1026.2003.

25. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. Genetics *196*, 973–983. https://doi.org/10.1534/genetics.113.160572.

26. Wells, J.M., Raju, B.C., Hung, H.Y., Weisburg, W.G., Mandelco-Paul, L., and Brenner, D.J. (1987). *Xylella fastidiosa* gen. nov., sp. nov: Gram-negative, Xylem-Limited, Fastidious plant bacteria related to *Xanthomonas* spp. Int. J. Syst. Bacteriol. *37*, 136–143. https://doi.org/10.1099/00207713-37-2-136.

27. Denancé, N., Briand, M., Gaborieau, R., Gaillard, S., and Jacques, M.A. (2019). Identification of genetic relationships and subspecies signatures in *Xylella fastidiosa*. BMC Genomics *20*, 239. https://doi.org/10.1186/s12864-019-5565-9.

28. Batarseh, T.N., Morales-Cruz, A., Ingel, B., Roper, M.C., and Gaut, B.S. (2022). Using genomes and evolutionary analyses to screen for host-specificity and positive selection in the plant pathogen *Xylella fastidiosa*. Appl. Environ. Microbiol. *88*, e0122022. https://doi.org/10.1128/aem.01220-22.

29. Campos, P.E., Groot Crego, C.G., Boyer, K., Gaudeul, M., Baider, C., Richard, D., Pruvost, O., Roumagnac, P., Szurek, B., Becker, N., et al. (2021). First historical genome of a crop bacterial pathogen from herbarium specimen: insights into citrus canker emergence. PLoS Pathog. *17*, e1009714. https://doi.org/10.1371/journal.ppat.1009714.

30. Román-Reyna, V., Dupas, E., Cesbron, S., Marchi, G., Campigli, S., Hansen, M.A., Bush, E., Prarat, M., Shiplett, K., Ivey, M.L.L., et al. (2021). Metagenomic sequencing for identification of *Xylella fastidiosa* from leaf samples. mSystems *6*, e0059121. https://doi.org/10.1128/mSystems.00591-21.

31. Sicard, A., Saponari, M., Vanhove, M., Castillo, A.I., Giampetruzzi, A., Loconsole, G., Saldarelli, P., Boscia, D., Neema, C., and Almeida, R.P.P. (2021). Introduction and adaptation of an emerging pathogen to olive trees in Italy. Microb. Genom. *7*, 00735. https://doi.org/10.1099/MGEN.0.000735.

32. Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F.N., Kamoun, S., Krause, J., et al. (2013). The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. eLife *2*, e00731. https://doi.org/10.7554/eLife.00731.

33. Rieux, A., and Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. Mol. Ecol. *25*, 1911–1924. https://doi.org/10.1111/mec.13586.

34. Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics *161*, 1307–1320. https://doi.org/10.1093/genetics/161.3.1307.

35. Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R., and Rodrigo, A.G. (2003). Measurably evolving populations. Trends Ecol. Evol. *18*, 481–488. https://doi.org/10.1016/S0169-5347(03)00216-7.

36. Rieux, A., Eriksson, A., Li, M., Sobkowiak, B., Weinert, L.A., Warmuth, V., Ruiz-Linares, A., Manica, A., and Balloux, F. (2014). Improved calibration of the human mitochondrial clock using ancient genomes. Mol. Biol. Evol. *31*, 2780–2792. https://doi.org/10.1093/molbev/msu222.

37. Castillo, A.I., Bojanini, I., Chen, H., Kandel, P.P., De La Fuente, L., and Almeida, R.P.P. (2021). Allopatric plant pathogen population divergence following disease emergence. Appl. Environ. Microbiol. *87*, 1–19. https://doi.org/10.1128/AEM.02095-20.

38. Gardner, M.W., and Hewitt, W.B. (1974). Pierce's Disease of the Grapevine: the Anaheim Disease and the California Vine Disease (University of California and Davis), p. 225.
39. Simpson, J. (2008). California and the creation of a modern wine industry: 1860–1919. Work. Pap. Econ. Hist. *37*, 08–14.
40. Yuan, X., Morano, L., Bromley, R., Spring-Pearson, S., Stouthamer, R., and Nunney, L. (2010). Multilocus sequence typing of *Xylella fastidiosa* causing Pierce's disease and oleander leaf scorch in the United States. Phytopathology *100*, 601–611. https://doi.org/10.1094/PHYTO-100-6-0601.
41. Jacques, M.A., Denancé, N., Legendre, B., Morel, E., Briand, M., Mississipi, S., Durand, K., Olivier, V., Portier, P., Poliakoff, F., et al. (2015). New coffee plant-infecting *Xylella fastidiosa* variants derived via homologous recombination. Appl. Environ. Microbiol. *82*, 1556–1568. https://doi.org/10.1128/AEM.03299-15.
42. Aguilar-Granados, A., Hernández-Macías, B., Santiago-Martínez, G., Ruiz-Medrano, R., Kameyama-Kawabe, L., Hinojosa-Moya, J., Del Carmen Montes-Horcasitas, M., and Xoconostle-Cázares, B. (2021). Genetic diversity of *Xylella fastidiosa* in Mexican vineyards. Plant Dis. *105*, 1490–1494. https://doi.org/10.1094/PDIS-09-20-1900-RE.
43. Gluck-Thaler, E., Cerutti, A., Perez-Quintero, A.L., Butchacas, J., Roman-Reyna, V., Madhavan, V.N., Shantharaj, D., Merfa, M.V., Pesce, C., Jauneau, A., et al. (2020). Repeated gain and loss of a single gene modulates the evolution of vascular plant pathogen lifestyles. Sci. Adv. *6*, eabc4516. https://doi.org/10.1126/sciadv.abc4516.
44. Burger, M., Woods, R.G., McCarthy, C., and Beacham, I.R. (2000). Temperature regulation of protease in *Pseudomonas fluorescens* LS107d2 by an ECF sigma factor and a transmembrane activator. Microbiology (Reading) *146*, 3149–3155. https://doi.org/10.1099/00221287-146-12-3149.
45. Matsui, H., Sano, Y., Ishihara, H., and Shinomiya, T. (1993). Regulation of pyocin genes in Pseudomonas aeruginosa by positive (prtN) and negative (prtR) regulatory genes. J. Bacteriol. *175*, 1257–1263. https://doi.org/10.1128/jb.175.5.1257-1263.1993.
46. Jiao, H., Li, F., Wang, T., Yam, J.K.H., Yang, L., and Liang, H. (2021). The pyocin regulator PrtR regulates virulence expression of *Pseudomonas aeruginosa* by modulation of Gac/RSM system and c-di-GMP signaling pathway. Infect. Immun. *89*, e0060220. https://doi.org/10.1128/IAI.00602-20.
47. Kim, Y.R., Kim, S.Y., Kim, C.M., Lee, S.E., and Rhee, J.H. (2005). Essential role of an adenylate cyclase in regulating *Vibrio vulnificus* virulence. FEMS Microbiol. Lett. *243*, 497–503. https://doi.org/10.1016/j.femsle.2005.01.016.
48. Danchin, A., Guiso, N., Roy, A., and Ullmann, A. (1984). Identification of the *Escherichia coli* cya gene product as authentic adenylate cyclase. J. Mol. Biol. *175*, 403–408. https://doi.org/10.1016/0022-2836(84)90356-5.
49. Wu, W., and Jin, S. (2005). PtrB of *Pseudomonas aeruginosa* suppresses the Type III Secretion System under the stress of DNA damage. J. Bacteriol. *187*, 6058–6068. https://doi.org/10.1128/JB.187.17.6058-6068.2005.
50. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.
51. Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147. https://doi.org/10.1371/journal.pone.0011147.
52. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
53. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
54. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.
55. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
56. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
57. DOE Joint Genome Institute. (2020). BBTools. https://jgi.doe.gov/data-and-tools/software-tools/.
58. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
59. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
60. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.
61. Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics *12*, 385. https://doi.org/10.1186/1471-2105-12-385.
62. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A., Gladstone, R.A., Lo, S., Beaudoin, C., Floto, R.A., et al. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. *21*, 180. https://doi.org/10.1186/s13059-020-02090-4.
63. Didelot, X., and Wilson, D.J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLOS Comput. Biol. *11*, e1004041. https://doi.org/10.1371/journal.pcbi.1004041.
64. Crispell, J., Balaz, D., and Gordon, S.V. (2019). Homoplasyfinder: A simple tool to identify homoplasies on a phylogeny. Microb. Genom. *5*, e000245. https://doi.org/10.1099/mgen.0.000245.
65. Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., and Harris, S.R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genom. *2*, e000056. https://doi.org/10.1099/mgen.0.000056.
66. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
67. Frichot, E., and François, O. (2015). LEA: an R package for landscape and ecological association studies. Methods Ecol. Evol. *6*, 925–929. https://doi.org/10.1111/2041-210X.12382.
68. Doizy, A., Prin, A., Cornu, G., Chiroleu, F., and Rieux, A. (2023). Phylostems: a new graphical tool to investigate temporal signal of heterochronous sequences datasets. Bioinform. Adv. *3*, vbad026. https://doi.org/10.1093/bioadv/vbad026.
69. Rieux, A., and Khatchikian, C.E. (2017). TIPDATINGBEAST:an R package to assist the implementation of phylogenetic tip-dating tests using BEAST. Mol. Ecol. Resour. *17*, 608–613. https://doi.org/10.1111/1755-0998.12603.
70. Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973. https://doi.org/10.1093/molbev/mss075.
71. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. *67*, 901–904. https://doi.org/10.1093/sysbio/syy032.
72. Drummond, A.J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214. https://doi.org/10.1186/1471-2148-7-214.
73. Revell, L.J. (2016). Package 'phytools'. R Top. Doc. 165. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5b679e40fa12cd9bc925dd9b5865b8a79b891de9.
74. Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. *17*, 1–9. https://doi.org/10.1186/s13059-016-1132-8.

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Xylella fastidiosa isolated from Almond (24)	This study	Table S2
Xylella fastidiosa isolated from Grapevine (53)	This study	Table S2
Biological samples		
Herbarium *Vitis vinifera* samples	U.C. Davis Plant Center for Plant Diversity Herbarium	DAV238006, see Table S1.
Critical commercial assays		
DNeasy Blood & Tissue kit	Qiagen	ID 69504
GoTaq Probe qPCR Master Mix	Promega	A6101
Deposited data		
Herb_1906 Assembly and Raw Reads	NCBI	PRJNA1114123
Illumina data from 53 grapevine-isolated *Xff*	NCBI	PRJNA722088
Illumina data from 24 almond-isolated *Xff*	NCBI	PRJNA786005
Additional 253 *Xff* strains	NCBI	Table S3
Vitis vinifera host genome	NCBI	GCA_927798595.1
Oligonucleotides		
Xf-riM-F-PEI primer (GTGAAATCAAGATAGAGTCTTG)	This study	N/A
Xf-riM-R-PEI primer (CGCATCCCGTGGCTCAGTC)	This study	N/A
Xf-riM-S-PEI probe (CGCATCCCGTGGCTCAGTC)	This study	N/A
Software and algorithms		
SPAdes	Bankevich et al.[50]	https://github.com/ablab/spades
ProgressiveMauve	Darling et al.[51]	https://darlinglab.org/mauve/user-guide/progressivemauve.html
Prokka	Seemann[52]	https://github.com/tseemann/prokka
Quast	Gurevich et al.[53]	https://github.com/ablab/quast
CheckM	Parks et al.[54]	https://github.com/Ecogenomics/CheckM
BWA	Li and Durbin[55]	https://github.com/lh3/bwa
BLAST	Altschul et al.[56]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
BBTools	DOE JGI[57]	https://jgi.doe.gov/data-and-tools/software-tools/bbtools/
Trimmomatic	Bolger et al.[58]	https://github.com/usadellab/Trimmomatic
Mapdamage	Jónsson et al.[22]	https://github.com/ginolhac/mapDamage
Bowtie2	Langmead and Salzberg[59]	https://github.com/BenLangmead/bowtie2
Cutadapt	Martin[60]	https://github.com/marcelm/cutadapt
krona	Ondov et al.[61]	https://github.com/marbl/Krona
Panaroo	Tonkin-Hill et al.[62]	https://github.com/gtonkinhill/panaroo
ClonalFrameML	Didelot and Wilson[63]	https://github.com/xavierdidelot/ClonalFrameML
HomoplasyFinder	Crispell et al.[64]	https://github.com/JosephCrispell/homoplasyFinder
snp-sites	Page et al.[65]	https://github.com/sanger-pathogens/snp-sites
RAxML	Stamatakis[66]	https://cme.h-its.org/exelixis/web/software/raxml/
LEA	Frichot and François[67]	https://github.com/bcm-uga/LEA
Phylostems	Doizy et al.[68]	https://gitlab.com/cirad-apps/phylostems

(*Continued on next page*)

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
TipDatingBeast	Rieuz and Khatchikian[69]	https://cran.r-project.org/web/packages/TipDatingBeast/index.html
BEAST	Drummond et al.[70]	https://beast.community/
Tracer	Rambaut et al.[71]	https://github.com/beast-dev/tracer
TreeAnnotator	Drummond and Rambaut[72]	https://beast.community/treeannotator
phytools	Revell[73]	https://github.com/liamrevell/phytools
Scoary	Brynildsrud et al.[74]	https://github.com/AdmiralenOla/Scoary
Other		
NovaSeq 6000	Illumina	N/A
NextSeq 500	Illumina	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Sampling of historical specimens

The historical viticulture specimen collection in the Center for Plant Diversity Herbarium at University of California, Davis (https://herbarium.ucdavis.edu/) was prospected in February 2022. Ten specimens labeled as diseased (Table S1) were sampled for this study. Petiole sections were excised using sterile equipment and transported back to our laboratory in Réunion inside individual envelopes where they were stored at 17°C in vacuum-sealed boxes until use.

Sampling, Sequencing, and Assembly of 77 Modern Isolates

In the fall of 2020, we isolated 77 *X. fastidiosa* subsp. *fastidiosa* strains: 53 from four vineyards in Hopland (Mendocino County, California) and 24 from 15 almond orchards across California. Metadata on host plant and geographic location is included in Table S2. Petioles were brought back to the lab for culturing on PD3 and PWG solid medium. Positive samples were triple cloned to isolate single genotypes; live cells were stored in liquid PD3 medium with 30% glycerol. DNA from pure colonies was extracted using a Qiagen DNeasy Blood & Tissue kit and submitted for Illumina NovaSeq paired-end sequencing at the QB3 Vincent J. Coates Genomics Sequencing Laboratory, which returns de-multiplexed raw reads. Genome assemblies were performed as previously described,[16,17] briefly, we assembled using SPAdes v. 3.15.2[50] reordered contigs using progressiveMauve[51] using Temecula1 as a reference, and annotated all assemblies using Prokka v. 1.14.6.[52] Genome quality was assessed using Quast v. 5.0.2[53] as well as checkM v. 1.2.2.[54] Four assemblies (Mendo20_A12, Mendo20_A13, ALS17T2. and ALS17T13) were discarded for analyses due to high L50 or contamination issues detected from CheckM. Genome assembly statistics for all 77 modern isolates are included in Table S2, but only those used in the downstream analyses (n=73) are included in Table S3.

METHOD DETAILS

DNA Extraction and qPCR

Total DNA was extracted in a bleach-cleaned facility lab with no prior exposure to modern *Xff* DNA, precluding the risk of contamination with contemporary DNA sequences. We followed the custom CTAB protocol described in Campos et al.[7] including both blank extractions (sample buffers only) and one *Coffea* sp. herbarium sample (known not to be *Xff*-infected), as negative controls. Reads generated from the negative control sample were sequentially mapped to reference sequence genomes of *Vitis vinifera* (GCA_910591555.1) and *Xff* (GCA_000007245.1) using BWA-aln 0.7.15.[55] Mapped reads were blasted against the nucleotide database using the blastn command of NCBI BLAST 2.2.31.[56] A qPCR (quantitative polymerase chain reaction) assay was run in duplicate on all extractions with a detection threshold of C_T = 33 (200 oligonucleotide copies), using 200 nM primers Xf-riM-F-PEI (GTGAAATCAAGATAGAGTCTTG), Xf-riM-R-PEI (CGCATCCCGTGGCTCAGTC), 150 nM TaqMan probe Xf-riM-S-PEI (CGCATCCCGTGGCTCAGTC), 0.3 mg/ml BSA and GoTaq Probe qPCR Master Mix, 40 cycles with an annealing temperature of 62°C. Primers were re-designed based on an initial reference (ANSES/ LSV / MA 039 version 5, July 2020) with a shortened target of 58 bp more suitable for aDNA, matching the rimM reference sequence with 100% identity in *fastidiosa*, *pauca* and *sandyi* subspecies, and with one mismatch for *multiplex* subspecies. A corresponding synthetic oligonucleotide target of 58 nt, in triplicates containing 200, 2,000, 20,000, 200,000 or 2,000,000 copies each, respectively, was used to validate a standard curve, and to estimate copy numbers of the samples.

Sequencing and Reads Initial Trimming

Samples that tested positive by qPCR for *X. fastidiosa* were sent to the Fasteris facility (https://www.fasteris.com/en-us/) along with negative controls. DNA extractions were converted into libraries following the TrueSeq DNA Nano protocol. Sequencing was performed in a paired-end 2×150 cycles configuration on a NextSeq flow cell. Artefactual homopolymer sequences were removed

from libraries when presenting entropy less than 0.6 using BBDuk from BBMap 37.92.[57] Adapters were trimmed using the Illuminaclip option from Trimmomatic 0.36.[58] Reads were processed into the post-mortem DNA damage assessment pipeline detailed in the section below. Additional quality trimming was performed with Trimmomatic based on base quality (*leading:15; trailing:15; slidingwindow:5:15*) and read length (minlen:*30*).

Ancient DNA Damage Assessment and Correction

Post-mortem DNA damage was measured by DNA fragment length distribution and terminal deamination patterns using mapDamage 2.2.1.[22] Alignments required for mapDamage were performed with Bowtie 2[59] (options–non-deterministic–very-sensitive), using the *Xff* reference strain Temecula1 genome (NC_004556.1). Reads were both corrected (with downscaled quality scores being associated to likely post-mortem damaged bases) using mapDamage (–rescale option) and trimmed by 5 base pairs at each end using Cutadapt[60] to remove the most heavily damaged sites from subsequent analyses. A similar investigation was performed for reads mapping to the *Vitis vinifera* host genome, using reference assembly GCA_927798595.1.

Taxonomic Assessment

The historical sample metagenomic composition was assessed using a blast-n assignment against the nucleotide (NCBI-nt) database.[56] Only top hits with an e-value below 0.001 were saved and graphically summarized by interactive charts of taxonomic abundance produced with krona[61] (Figure 1B).

De novo Assembly of the Historical Genome

Reads were first mapped against a reference dataset of all available contemporary *X. fastidiosa (Xf)* genomes using bowtie2.[59] Mapped reads were then assembled *de novo* using SPAdes.[50] Coverage was determined by mapping the raw reads to the assembled historical genome using bbmap[57] and assembly statistics and quality control were performed with CheckM[54] and Quast.[53] We checked the multilocus sequence type of this assembly using command-line BLAST v. 2.7.1 (NCBI) and publicly available fasta files of known allele sequences for the seven MLST genes (PubMLST: https://pubmlst.org/bigsdb?db=pubmlst_xfastidiosa_seqdef). For additional contemporary genomes used in this study, we imposed the following cutoffs to ensure high quality genomes: N50 of >40,000 bp and <15 missing lineage-specific marker genes (from CheckM, indicative of completeness) and <15 duplicated (2-5 copies) of lineage-specific marker genes (CheckM, indicative of contamination). While genome quality cutoffs can be arbitrary, they were applied because the vast set of *Xf* genomes used here included strains sequenced and assembled from various platforms and pipelines.

Population Genomics and Phylogenetics

Herb_1906 was annotated using Prokka[52] and aligned with 330 publicly available and new modern sequences (detailed above) from *Xff* using Panaroo.[62] All additional genomes used are documented in Table S3. Core genome alignments were checked for recombination using ClonalFrameML[63] and for homoplasic sites using HomoplasyFinder.[64] All recombinant and homoplasic sites were removed from the entire alignment using an in-house Python script. Snp-sites[65] was then used to extract single nucleotide polymorphisms (SNPs) from the non-recombinant alignment which was in turn used to generate a maximum likelihood phylogenetic tree using RAxML[66] with the GTR (general time reversible) cat parameter and 100 iterations. We ran the above pipeline (Panaroo > ClonalFrameML / HomoplasyFinder > Recombination Removal > Snp-sites > RAxML) on 330 subsp. *fastidiosa* strains with 1 subsp. *multiplex* outgroup (Red Oak 2: NCBI SAMN16582182), either including (n=332) or excluding (n=331) the historical strain Herb_1906. In addition to phylogenetic analyses, we estimated ancestry coefficient matrices from SNP data using the sparse nonnegative matrix factorization (sNMF) algorithm implemented in the LEA package in R.[67] The subsp. *multiplex* outgroup strain was dropped for these analyses. To determine the number of ancestral populations (k), we tested k from 1 to 15 with 10 repetitions for each k; the cross-entropy criterion was used to determine a k=5 for this dataset (Figure S1). We visually assessed this graph to choose the 'elbow': the inflection point after which there are only marginal decreases in the cross-entropy criterion. For k=5, all strains were assigned to a cluster with a minimum coefficient of q > 0.5. Following the determination of k, we re-ran the sNMF algorithm with k=5 and 10 repetitions; the repetition with the lowest entropy score was used for the final ancestry coefficient matrix (Q). We also re-ran the sNMF algorithm from k to 2-7 using 10 repetitions to examine clusters graphically (Figure S7). Using k=4, a cluster of 4 strains from Costa Rica and 1 from Mexico were not assigned to a group with q > 0.5, whereas in k=5 this group forms its own cluster (V5). Using k=6, strains including the historical strain, some modern CA strains from Mendocino, a strain from Georgia (Xf51_CCPM1), a strain from Virginia (MAG_669), and a strain from Florida (CFBP8069) formed a separate cluster from other modern CA strains from Napa, Sonoma, and Mendocino. However, these strains all form a monophyletic cluster, so this confirmed our choice of k=5.

Molecular Clock

As a requirement to perform tip-calibrated phylogenetic inferences,[33] the existence of a temporal signal was investigated using both a linear regression test between sample age and root-to-tip distances with Phylostems[68] and a date-randomization test (DRT) with the R package TipDatingBeast.[69] Temporal signal was considered present at nodes displaying a significant positive correlation and when there was no overlap between the inferred root height 95% highest posterior density (95% HPD) of the initial dataset and that of 20 date-randomized datasets (Figure S3). Tip-dating calibration Bayesian inferences were performed using BEAST 1.8.4[70] on a subset of 10,000 randomly subsampled SNPs (single nucleotide polymorphisms). The xml input file was manually edited with the number

Please cite this article in press as: Donegan et al., Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.11.029

on invariant A,T,C,Gs to correct rates for the fact that SNPs only were being used, as advised by Rieux and Balloux.[33] Leaf heights were constrained to be equal to sample ages. We also considered a GTR substitution model with a Γ distribution and invariant sites (GTR + G + I), an uncorrelated relaxed log-normal clock to account for variations between lineages, and a tree prior for demography of coalescent extended Bayesian skyline. Five independent chains were run for 200 million steps and sampled every 20,000 steps, discarding the first 20,000 steps as burn-in. Convergence to the stationary, sufficient sampling (effective sample size > 200) and mixing were checked by inspecting posterior samples with Tracer 1.7.[71] Final parameters estimation was based on the combination of the different chains. Maximum clade credibility method in TreeAnnotator[72] was used to determine the best-supported tree of the combined chains. To assess the influence of various demographic models on parameter inferences, we compared the performance of the following demographic models (Bayesian Skyplot, Constant size, Extended Bayesian Skyline Plot, Exponential growth and Skygrid) on both root age and substitution rate posterior estimates (Figure S4).

Phylogeographic Ancestral State Reconstruction

We estimated latitude and longitude separately at all nodes of the Maximum Likelihood phylogeny using a Maximum Likelihood ancestral state reconstruction "fastAnc" function implemented in the phytools package in R[73]. Strains strains lacking precise metadata were trimmed from the phylogeny as well as strains from Spain and Taiwan to focus on movement in the Americas (Table S3). For visualizing dispersal events, we focused on long-distance movement. We calculated inferred distance along each edge of the phylogenetic tree (i.e. between a node / node or node / tip) by taking the absolute value of the difference in latitude and the absolute value of the difference in longitude. A long-distance dispersal event was defined as > 10 decimal degree difference in either latitude or longitude. Of 558 edges, 8 were defined as long-distance dispersal events. To determine confidence in each event, we calculated the area of an ellipse for the 95% confidence at the parent (origin) node for each of the 8 events (ellipse area = π * (latitude - lower 95% C.I. latitude) * (longitude - lower 95% C.I. longitude). Seven of the eight long-distance dispersal events had 95% C.I. ellipse areas of < 600 square decimal degrees, while the eighth had a 95% C.I. ellipse area of 4,813.5 square d.d (Figure S5). We used a cutoff of 600 d.d.[2] at the parent node to define "high confidence" dispersal events.

Comparative genomics

We used Scoary[74] to organize gene content unique to the ancient genome in contrast to the clade to which the strain belongs. The gene presence absence matrix produced by Panaroo was evaluated for genes that are unique to the ancient isolate in comparison to the rest of the dataset, as well as any genes that are completely absent in the ancient isolate. The default parameters were used for Panaroo, wherein the sequence identity threshold and length difference cutoffs are both set at 0.98. Differences are visualized in Figure S6, This was also conducted as a comparison between Herb_1906 and the clade including 37 strains most closely related to it, which is included in Table S4. This process is constrained by the mapping conducted for genome assembly - only genes that have been sequenced across the species *X. fastidiosa* will be detectable, so if a gene from the ancient isolate was unique to that extent, it would not be detectable.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistics on DNA fragment length distribution and terminal deamination patterns were calculated using mapDamage[22] and can be found in Figure 1C. Genome assembly statistics (N_{50}, L_{50}, GC%, etc) were calculated with both Quast[53] and CheckM[54] and can be found for Herb_1906 in the results section and Table 1, and for all strains included in the study can be found in Table S2. The estimation of posterior densities for node heights were performed in BEAST[70] and described in detail in the "molecular clock" section of the method details. 95% confidence intervals for time estimation of major nodes can be found in Figure 2D. The full maximum likelihood phylogeny including bootstrap support for each node was generated using RaxML[66] and can be visualized in Figure S2. Confidence intervals for the ancestral state reconstruction were calculated with the phytools[73] package in R and can be visualized in Figure S5.

Current Biology, Volume *35*

Supplemental Information

Century-old herbarium specimen provides insights into Pierce's disease of grapevines emergence in the Americas

Monica A. Donegan, Alexandra K. Kahn, Nathalie Becker, Andreina Castillo Siri, Paola E. Campos, Karine Boyer, Alison Colwell, Martial Briand, Rodrigo P.P. Almeida, and Adrien Rieux

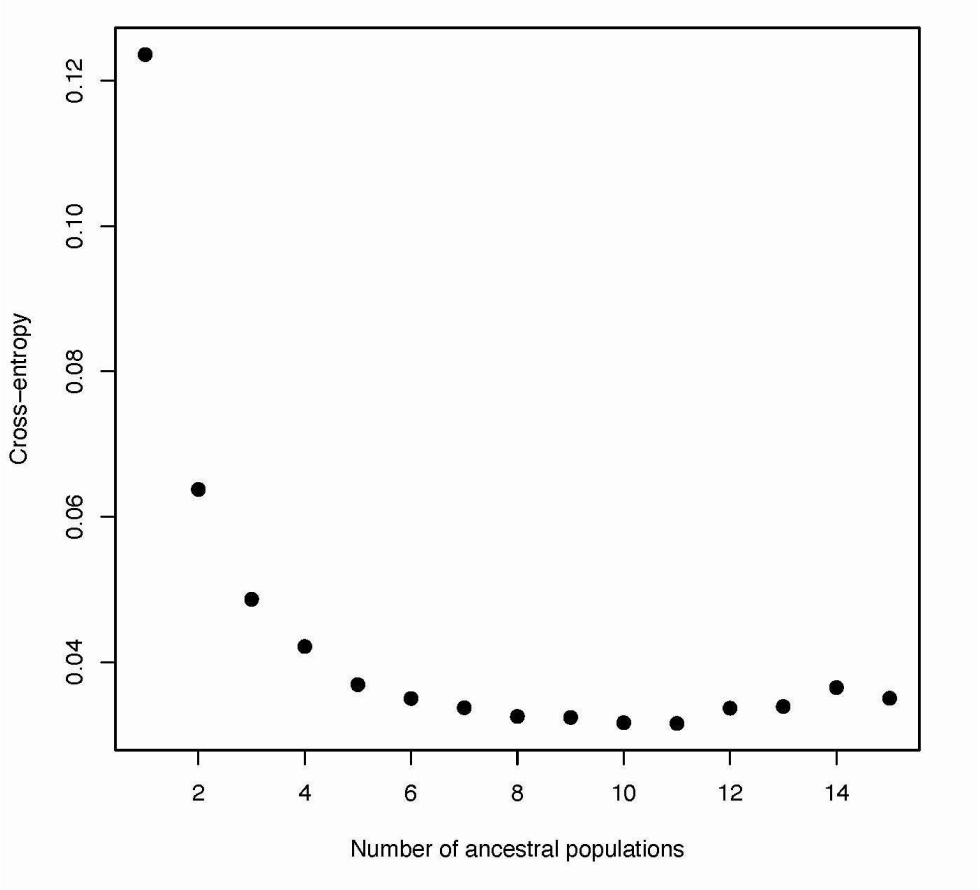


Figure S1 Cross-entropy scores to estimate k ancestral populations, Related to Figure 2. Cross-entropy scores from sNMF to estimate the number of ancestral populations (k); we tested k from 1 to 15 with 10 repetitions. The choice of k=5 reflects the inflection point where there are only marginal decreases to the cross-entropy criterion by adding more complexity (clusters) to the model.

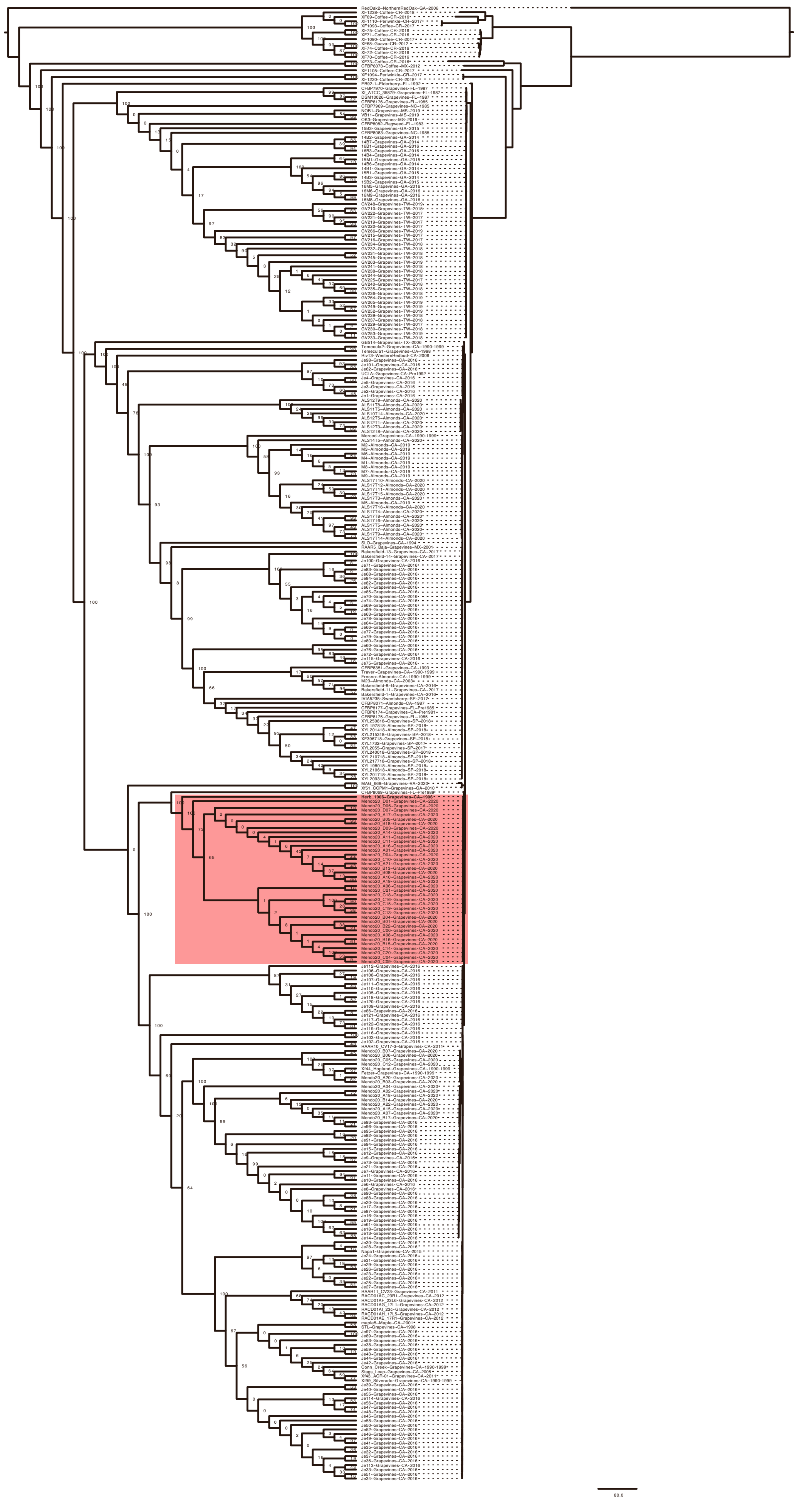


Figure S2 Maximum likelihood phylogeny and cladogram of all strains, Related to Figure 2. A maximum likelihood phylogeny of the non-recombinant core genome alignment was generated using RaxML. A cladogram of the phylogeny (left) is displayed to more clearly visualize bootstrap support and tree topology. A phylogenetic tree (right) is displayed so as to accurately visualize branch lengths, with a scale bar to show the substitution rate across the tree. Herb_1906 and its closest relatives are highlighted with a pink box.

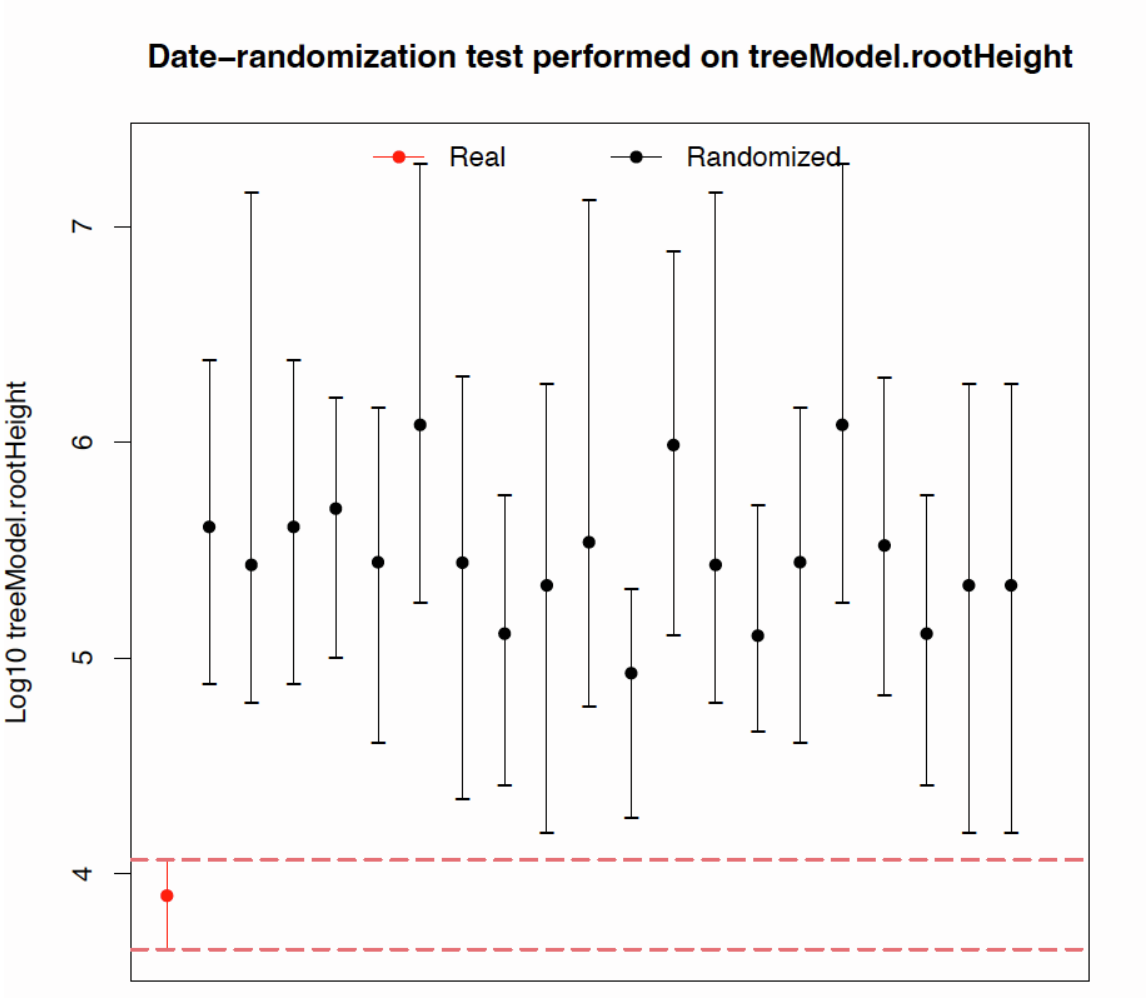


Figure S3. Temporal signal evaluation using the date-randomization test (DRT), Related to STAR Methods.

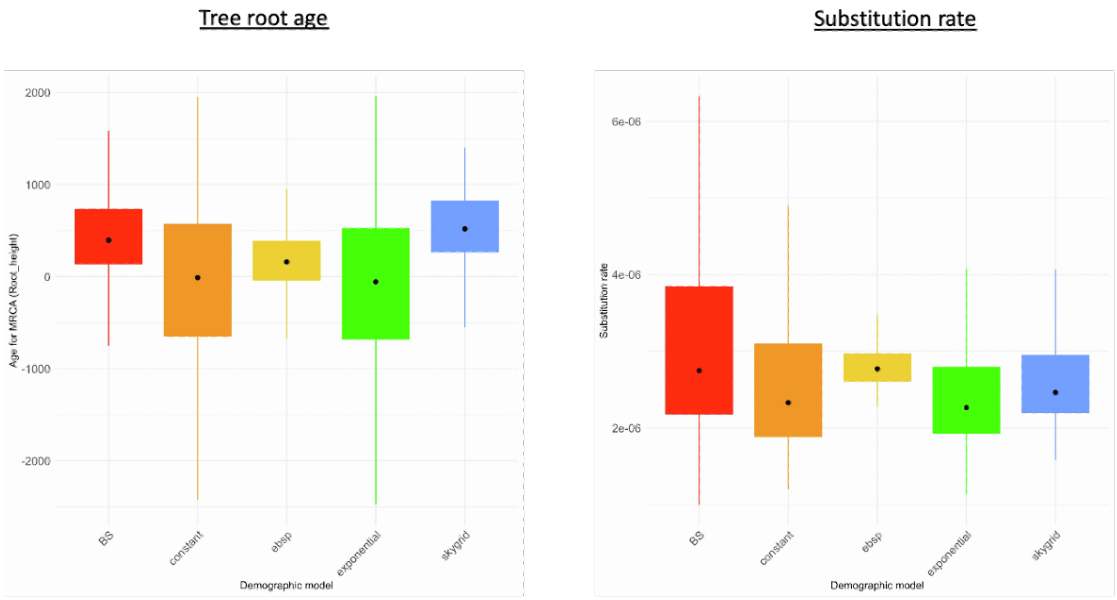


Figure S4 Comparison of BEAST demographic models, Related to STAR Methods.

Influence of different BEAST demographic models on posterior estimates of both root age and substitution rate, respectively. Tested models: BS: Bayesian Skyplot, Constant size, EBSP: Extended Bayesian Skyline Plot, Exponential growth and Skygrid models. Boxplots display the five following summary statistics: minimum, first quartile, median, third quartile and maximum of the 95% highest posterior density intervals

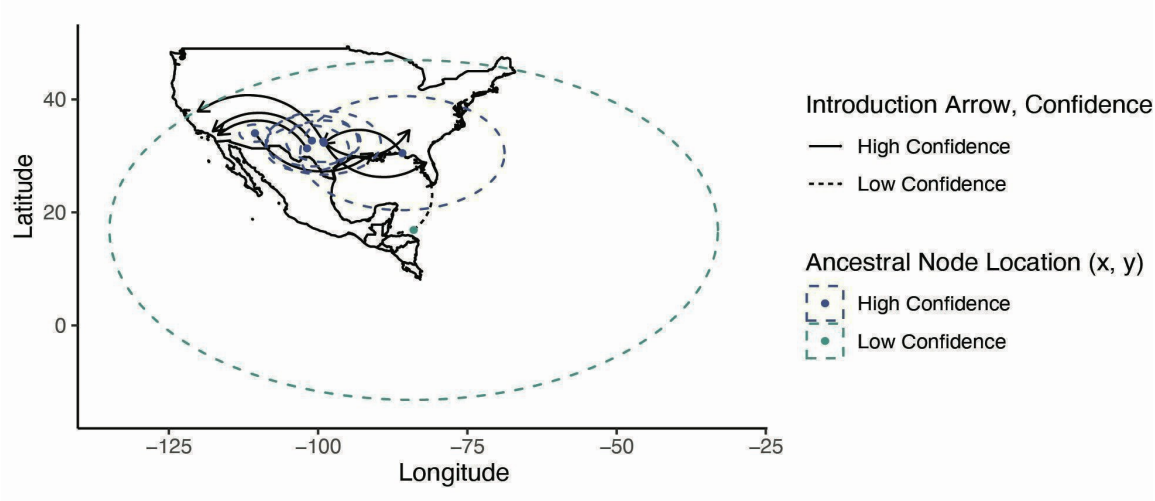


Figure S5 Confidence Intervals for Ancestral State Reconstruction, Related to Figure 3.
Ellipse / 95% CI for nodes in phylogeographic diffusion. Estimated latitude and longitude from ancestral state reconstructions were visualized with the ggforce and ggplot2[S1,S2] packages in R.

Ancestry group	Strain, or merged clades with strain numbers	group-1650 (27/331)	prtR (8/331)	group_840 (8/331)	cya (72/331)	group_2099 (13/331)	group_1158 (46/331)	group_421 (14/331)	group_1913 (109/331)
V1	XF1110_CR				■	■	■		
V1	XF71_CR4							■	
V1	XF75_CR8					■		■	
V1	XF1090_CR					■		■	
V1	XF72_CR5					■		■	
V1	XF70_fixed_CR					■		■	
V1	XF74_CR7					■		■	
V1	XF68_CR1					■		■	
V1	XF1093_CR				■	■			
V1	XF69				■	■		■	
V1	XF1238							■	
V5	XF73_CR6							■	
V5	XF1105_CR				■	■		■	
V5	XF1220_CR						■	■	
V5	XF1094_CR							■	
V4	EB92-1_FL				■		■		
V4	DSM10026_FL	■	■	■	■				
V4	CFBP7969				■				
V4	CFBP8082				■				
V4	VB11		■	■					
V4	NOB1	■							
V4	Xf_ATCC_35879_FL	■							
V4	DSM10026_FL		■	■	■				
V4	CFBP8176	■	■	■			■		
V4	CFBP7970		■	■	■		■		
V4	14B2_GA						■		
V4	GrapevinesGA (1/11 strain)	■					■		
V4	GrapevinesGA (2/11 strains)				■		■		
V4	GrapevinesGA (5/11 strains)						■		
V4	GrapevinesGA (3/11 strains)				■				
V4	Taiwan (10/31 strains)	■					■		
V4	Taiwan (13/31 strains)						■		
V4	Taiwan (5/31 strains)				■		■		
V4	Taiwan (2/31 strains)	■			■		■		
V4	Taiwan (1/31 strain)						■		
V2	GB514_TX		■	■		■			■
V2	Temecula2	■							
V2	Riv13_WesternRedbud_CA	■							
V2	Grapevine, Bakersfield, CA (2/21 strains)	■							
V2	Grapevine, Bakersfield, CA (8/21 strains)				■				
V2	Je72_XF2_52_CA				■				
V2	Je115_XF2_95_CA				■				
V2	CFBP8351				■				
V2	M23_Almonds_CA	■							
V2	Bakersfield8_CA	■							
V2	Bakersfield1_CA	■							
V2	Bakersfield11_CA	■							
V2	IVIA5235_Sweetcherry_SP				■				
V2	Almond, Grapevine Spain (11/14)				■				
V2	CFBP8071				■				
V3	MAG_669		■	■					
V3	**Herb_1906**	■	■	■	■	■	■	■	■
V3	Grapevine, Temecula (15/15)								■
V3	Je103_XF2_83_CA								■
V3	Je116_XF2_96_CA								■
V3	RAAR10_CV17-3_CA								■
V3	Je102_XF2_82_CA								■
V3	Grapevine, Sonoma Mendocino (1/43)	■							■
V3	Grapevine, Sonoma Mendocino (2/43)				■				
V3	Grapevine, Sonoma Mendocino (22/43)								■
V3	Grapevine, Sonoma Mendocino (7/43)				■				■
V3	Grapevine, Napa, CA (12/54)				■				■
V3	Grapevine, Napa, CA (33/54)								■
V3	Grapevine, Napa, CA (2/54)				■				

Figure S6 Gene presence and absence, Related to STAR Methods. List of strains (second column) presence (gray cells) of at least one of the eight genes (third to tenth column) specifically identified as present in the HERB-1906 sequence, while absent from all strains from its closest Mendocino clade (n=37). Gene names, or group identifiers, are followed by their frequency among the total number of strains (331). Strain names are similar to the ones displayed Figure 2D, from top to bottom in the same order as on the tree.

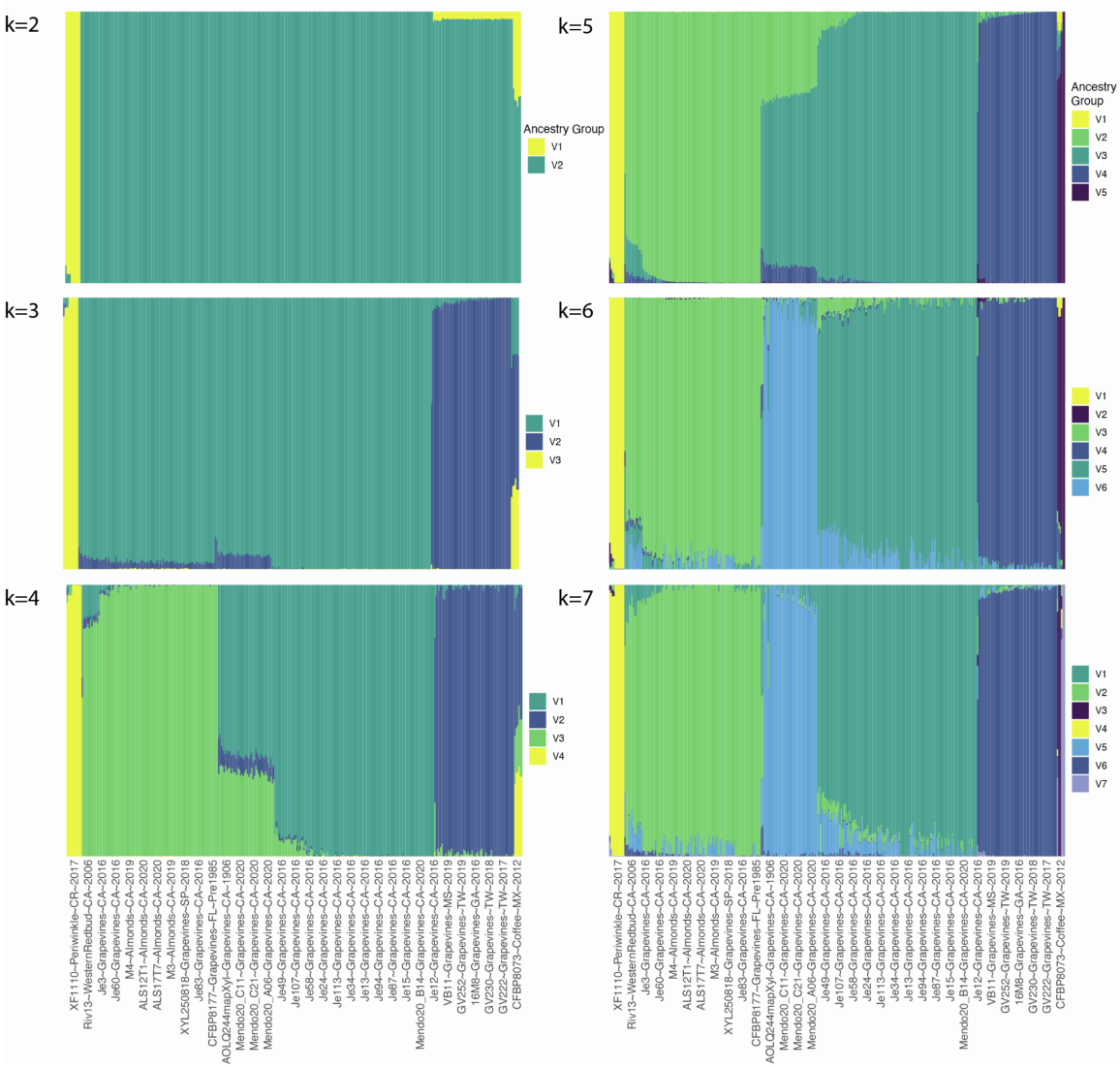


Figure S7 sNMF analysis from k=2 to k=7, Related to Figure 2. All 331 strains included in the analysis are visualized, but only 1 of 10 labels are plotted on the x-axis for simplicity.

Supplemental References

S1. Pedersen, T. (2024). ggforce: Accelerating ‘ggplot2’. R package version 0.5.0..
S2. Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis. ggplot2* (Springer-Verlag, New York). doi:10.1007/978-0-387-98141-3.