Combining Field Observations and Genetic Data to Reconstruct the Invasion of *Phytophthora ramorum* in California[1]

Peter J. P. Croucher,[2] Silvia Mascheretti,[2] and Matteo Garbelotto[2]

Abstract

Although it has been convincingly shown that forest populations of the pathogen *Phytophthora ramorum* have undergone a significant bottleneck and reproduce exclusively asexually (Ivors et al. 2004, 2006; Mascheretti et al. 2008), objective results showing that nurseries were the original source of the introduction remain elusive (Mascheretti et al. 2008). A previous attempt to define routes of pathogen movement resulted in a largely unresolved network (Mascheretti et al. 2009), showing at best that populations from Santa Cruz and Marin Counties were important sources within California. Previous attempts at reconstructing the entire history of the sudden oak death (SOD) epidemic in California were limited by: 1) incomplete sampling; 2) the inability to include singleton samples; and 3) over-collapsing of non-spatially contiguous, yet genetically similar, samples into large meta-samples that confounded the coalescent analyses. Here, we employ a complete sampling coverage of 832 isolates of *P. ramorum* (the causative agent of SOD) from 60 California forests, genotyped at nine microsatellite loci.

The following microsatellite loci were genotyped: PrMS39a, PrMS39b, PrMS43a, PrMS43b, PrMS45 (Prospero et al. 2007), locus 18, locus 64 (Ivors et al. 2006), and loci ILVO145PrMS145 (a and c) (Vercauteren et al. 2010). Rather than using data simply based on number of microsatellite repeats, we employed Bruvo's distances as in previous studies (Bruvo et al. 2004). This metric is appropriate for analyses of populations that comprise closely related genotypes originated from the same founder genotypes because larger shifts in the number of repeats are weighed, not proportionally, but in terms of likelihood. Analysis of molecular variance (AMOVA) (Excoffier et al. 1992), as implemented by ARLEQUIN v.3.5 (Excoffier et al. 2005), was employed to generate pair-wise estimates of Φ_{ST} among all 62 *P. ramorum* forest and nursery populations. The Bruvo distance among each pair of unique multilocus genotypes (MGs) was estimated and fed to ARLEQUIN as an external file as the basis for the evolutionary distance in the AMOVA calculations.

Many pair-wise estimates of Φ_{ST} were low and not significantly different from zero (as evaluated by permuting individuals across locations 10,000 times). Populations were therefore recursively clustered by pooling the pair of populations or clusters that yielded the minimum Φ_{ST} at each round until no further insignificant clustering (i.e., minimum Φ_{ST} $P > 0.05$) was possible (Mascheretti et al. 2008, Mascheretti et al. 2009, Roewer et al. 2005). The algorithm was supervised by applying it only to populations from within the same county. Following the county-based clustering, the algorithm was continued to completion, permitting collapses among counties. Additionally, we re-ran the algorithm without supervision. The results were evaluated based upon knowledge from previous analyses of *P. ramorum* populations (Mascheretti et al. 2008, Mascheretti et al. 2009) and by subjecting the final set of populations from each run (by county, overall, and unsupervised) to a traditional hierarchical AMOVA: evaluating the Φ_{CT} value, which provides an unbiased way to judge population groupings (i.e., by maximizing the 'among group' variation whilst minimizing the 'within group variation') (Dupanloup et al. 2002). Examination of Φ_{CT} values indicated that the supervised clustering (by county $\Phi_{CT} = 0.3083$, $P < 0.001$) was superior to the unsupervised algorithm (unsupervised $\Phi_{CT} = 0.2879$, $P < 0.001$), but that allowing collapsing among counties offered little further improvement (overall $\Phi_{CT} = 0.3099$, $P < 0.001$). Since collapses among counties may reflect recent shared source populations, rather than direct migration between these populations, the 'by county' set of populations was conservatively chosen as the basis for subsequent analyses.

[1] A version of this paper was presented at the Sudden Oak Death Fifth Science Symposium, June 19-22, 2012, Petaluma, California.

[2] Department of Environmental Science, Policy and Management, 130 Mulford Hall, University of California, Berkeley, CA 94720-3114.
Corresponding author: matteog@berkeley.edu.

The genetic relationships among the final set of populations (post-clustering) were visualized by estimating the matrix of average pair-wise Bruvo genetic distances (Bruvo et al. 2004) among all populations with five or more samples and constructing the shortest neighbor-joining (NJ) tree using FASTME v.2.07 (Desper and Gascuel 2002). Bootstrap values (1000) were generated similarly and summarized using CONSENSE (PHYLIP package v.3.6) (Felsenstein 2005). All individuals and populations were also subject to genetic clustering analysis using STRUCTURE (Pritchard et al. 2000), and the posterior probability of membership in each of the identified genetic clusters was then examined for each population, singleton, and 'historical' isolate.

We have previously (Mascheretti et al. 2009) attempted to infer infection routes by estimating all bidirectional migration rates *M* among populations using MIGRATE-N (Beerli 2006, Beerli and Felsenstein 2001), but results were far from showing a clear pattern of spread, possibly due to the overcollapsing of all contiguous genetically similar populations into the same metapopulation. Additionally, it is very unlikely that most *P. ramorum* populations approach genetic equilibrium, violating the assumptions of MIGRATE-N and rendering estimates of *M* unreliable. Here, rather than estimate *M*, we used the output *ln* marginal likelihoods (*ln*(*ml*)) to choose among different migration models (Beerli and Palczewski 2010) and reconstruct the minimal, most probable, set of unidirectional migration routes among the new and highly resolved population dataset. Twenty-nine populations ($n \geq 5$) yield a computationally intractable number of possible models, and we therefore took a stepwise approach (Croucher et al. 2012). The number of possible models was limited by incorporating "epidemiological" data, disallowing models that included migration from a younger to an older population.

A total of 224 MGs were defined, more than half of the MGs (139 (62.1 percent); 66 (57.4 percent) were defined. The most frequent genotypes were MG38 (n = 95; 11.4 percent); MG42 (n = 113; 13.5 percent); and MG46 (n = 122; 14.6 percent); corresponding to the three frequent and possible founder MGs (#13, #14, and #15; respectively) originally identified in Mascheretti et al. (2008).

Iterative collapsing reduced the initial set of 43 populations ($n > 5$) to 29 (table 1). AMOVA results are given in table 1. Although the proportion of genetic variation within individual populations (68 percent) and among groups (31 percent) was high, variation among populations within groups (the final populations) was extremely low (1 percent) when clustering populations only within, but not between, counties. The NJ tree, rooted through the Nursery_SC1 sample, is shown in fig. 1. The tree accurately reflects the age of each infestation, with the oldest infestations (except SO4) all close to the root and the youngest infestations furthest from the root. Although bootstrap values were low (expected given few loci), the remarkable congruence between age and topology indicates that the tree accurately reflects the genetic relationships among populations and that these genetic relationships reflect the progression of the infestation from older to younger populations.

Table 1—Analysis of molecular variance results for the 29 populations defined by "within-county" collapsing

Group	df	SS	Variance Components	% Variation
Among Groups	28	17.975	0.0214	30.83
Among Populations Within Groups	15	0.871	0.0472	1.17
Within Populations	767	36.236	0.0695	68.00
SUM	810	55.081		
Φ_{SC}	0.0170^{NS}			
Φ_{ST}	0.3200***			
Φ_{CT}	0.3083***			

STRUCTURE (Pritchard et al. 2000) was used to analyze all populations. The *ln* probability of the data, for each replicate at each value of *k*, was summarized using STRUCTURE HARVESTER (Earl and vonHoldt 2011). Ln Pr(X|*k*) increased with each value of *k* and did not have a clear maximum. The method of Evanno et al. (2005) indicated an optimal ΔK at $k = 2$. This method cannot evaluate $k = 1$ and it is therefore probable that the 'true' $k = 1$ – suggesting that the entire epidemic may result from single introduced MG. A second maximum ΔK was observed at $k = 4$, and this value was therefore examined as an indicator of possible *sub-structure* in the data. The genetic clusters map to the tree by age, with cluster 1 representing the oldest infestations, cluster 3 the next oldest and the majority of populations, then cluster 4, and finally cluster 2 mapping to the youngest populations.

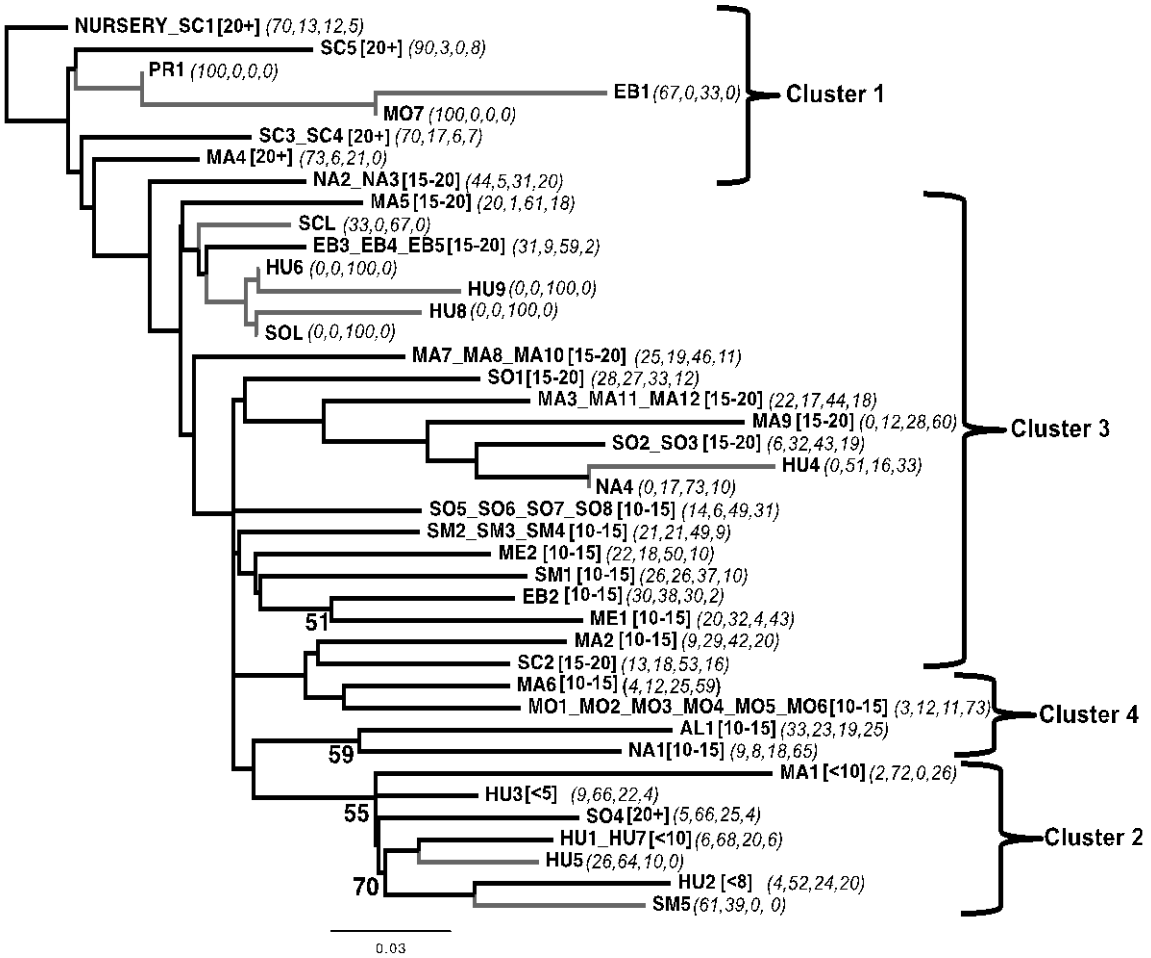

Figure 1—Neighbor-joining representations of the genetic relationships among *Phytophthora ramorum* populations. The numbers in parenthesis, following the years since infestation, indicate the percent posterior probability of membership in each of the genetic clusters (1,2,3,4) identified by STRUCTURE. Populations can be broadly identified with each of these clusters as indicated by the brackets. EB= East Bay, HU= Humboldt, MO= Monterey, NA= Napa, MA= Marin, SCL= Santa Clara, SC= Santa Cruz, AL= Southern Alameda, SM= San Mateo, PR= Presidio National Park, SO= Sonoma. Cluster 1 is consistently the oldest, while Cluster 2 is the youngest, based on field work.

The coalescent approach to migration model choice (Beerli and Palczewski 2010) identified for the first time, unambiguously and without subjectivity, that the Nursery_SC1 population was the original source population in the initial pair-wise analyses and prior to incorporating the historical data, thus corroborating independently that the SOD epidemic in California started in the Santa Cruz area and originally came from Nursery infestations (results not shown). The network shows a classical epidemic pattern with early spread to a few key localities - the forests around Santa Cruz, the San Francisco East Bay region, and the Golden Gate National Recreational Area in Marin County - followed by a multitude of infestation routes from these focal points. The analysis also revealed that, in several cases, multiple infestation routes involving different sources have affected the same counties, proving that many counties were each infested more than once during the brief history of the disease.

Until this study, there has been no clear understanding of the presence of four distinct clusters in California forests. The Nursery-associated cluster 1 appears in all analyses as the most ancestral one of the four (data not shown), with clusters 3 and 4 derived from it by a single repeat change at one locus, while cluster 2 is more closely related to cluster 3. Interestingly, the founder cluster 1 remains associated almost exclusively with nurseries and with forests neighboring Santa Cruz nurseries, where multiple lines of evidence indicate the original outbreak started. The other three clusters are much more widespread than cluster 1 (fig. 2). It is possible that clusters other than 1 may be better adapted to colonize forests. With the ability to differentiate the four clusters, this hypothesis can now be tested by further research.

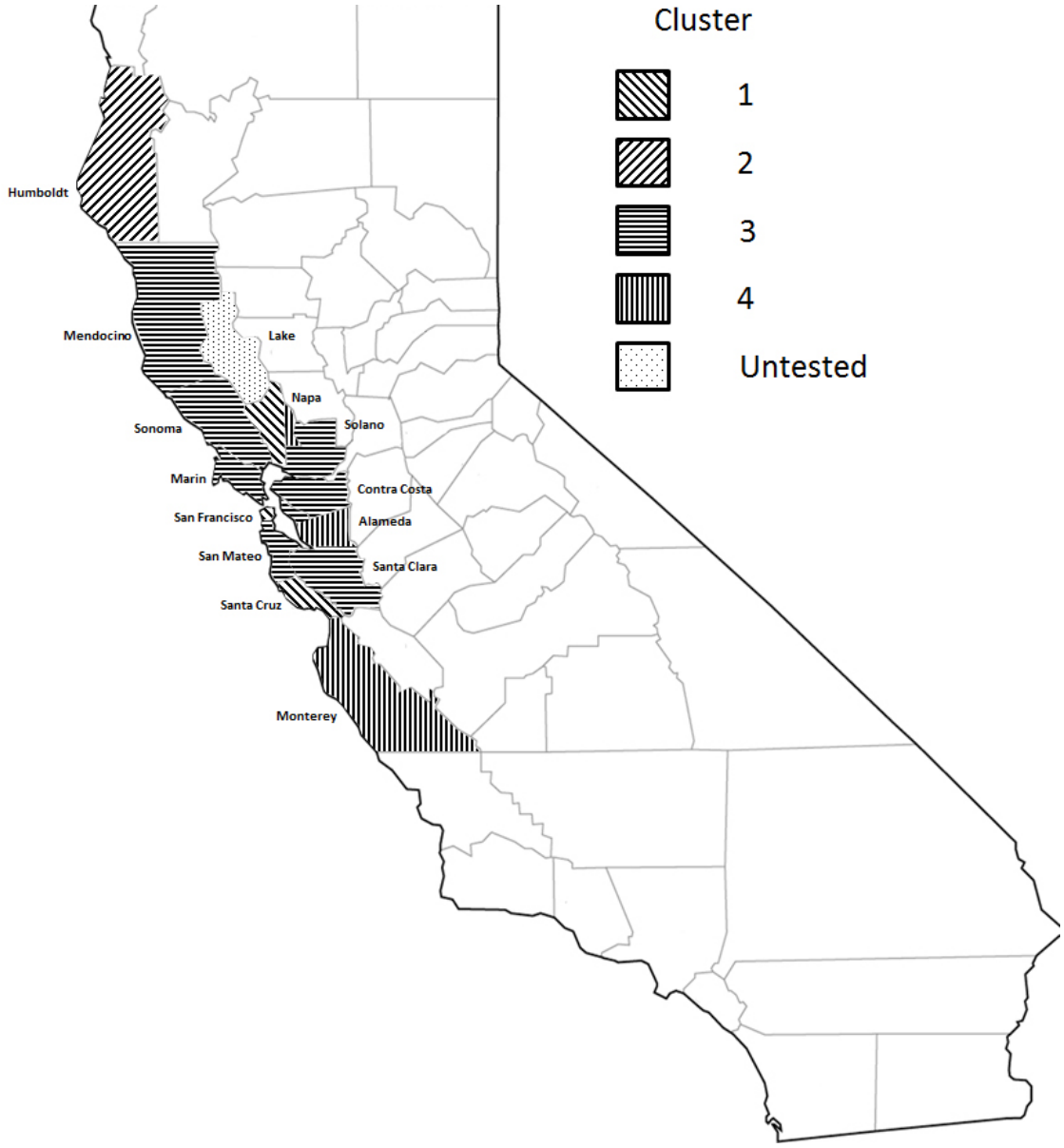

Figure 2—Distribution of the four genetic clusters of *Phytophthora ramorum* in California wildlands. Cluster 1 is the most ancestral cluster and is associated with nurseries; however, it is not as widespread as the other clusters, suggesting a microevolutionary trajectory possibly driven by adaptation.

Acknowledgments

This study was funded by an Ecology of Infectious Diseases grant (EF-062654), provided jointly by NSF and NIH, and by the Gordon and Betty Moore Foundation. The collection of samples predating the NSF-NIH grant was made possible through funding from USDA-APHIS and by the USDA FS, State and Private Forestry.

Literature Cited

Beerli, P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics. 22: 341–345.

Beerli, P.; Felsenstein, J. 2001. Maximum-likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences of the United States of America. 98: 4563–4568.

Beerli, P.; Palczewski, M. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics. 185: 313–326.

Bruvo, R.; Michiels, N.K.; D'Souza, T.G.; Schulenburg, H. 2004. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. Molecular Ecology. 13: 2101–2106.

Croucher, P.J.P.; Oxford, G.S.; Lam, A.; Mody, N.; Gillespie, R.G. 2012. Colonization history and population genetics of the color-polymorphic Hawaiian happy-face spider *Theridion grallator* (Araneae, Theridiidae). Evolution. 66(9): 2815–2833.

Desper, R.; Gascuel, O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. Journal of Computational Biology. 9: 687–705.

Dupanloup, I.; Schneider, S.; Excoffier, L. 2002. A simulated annealing approach to define the genetic structure of populations. Molecular Ecology. 11: 2571–2581.

Earl, D.A.; vonHoldt, B.M. 2011. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. Published online Oct 13, 2011. DOI: 10.1007/s12686-011-89548-7

Evanno, G.; Regnaut, S.; Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE a simulation study. Molecular Ecology. 14: 2611–2620.

Excoffier, L.; Laval, G.; Schneider, S. 2005. Arlequin (Version 3.01): An Integrated Software Package for Population Genetics Data Analysis. . Computation and Molecular Population Genetics Lab (GMPG). Institute of Zoology University of Berne, Berne, Switzerland.

Excoffier, L.; Smouse, P.; Quattro, J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics.131: 479–491.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ivors, K.; Garbelotto, M.; Vries, I.D. et al. 2006. Microsatellite markers identify three lineages of *Phytophthora ramorum* in US nurseries, yet single lineages in US forest and European nursery populations. Molecular Ecology. 15: 1493–1505.

Ivors, K.L.; Hayden, K.J.; Bonants, P.J.; Rizzo, D.M.; Garbelotto, M. 2004. AFLP and phylogenetic analyses of North American and European populations of *Phytophthora ramorum*. Mycological Research. 108: 378–392.

Mascheretti, S.; Croucher, P.J.P.; Kozanitas, M.; Baker, L.; Garbelotto, M. 2009. Genetic epidemiology of the sudden oak death pathogen *Phytophthora ramorum* in California. Molecular Ecology. 18: 4577–4590.

Mascheretti, S.; Croucher, P.J.P.; Vettraino, A.; Prospero, S.; Garbelotto, M. 2008. Reconstruction of the Sudden Oak Death epidemic in California through microsatellite analysis of the pathogen *Phytophthora ramorum*. Molecular Ecology. 17: 2755–2768.

Pritchard, J.K.; Stephens, M.; Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155: 945–959.

Prospero, S.; Hansen, E.M.; Grunwald, N.J.; Winton, L.M. 2007. Population dynamics of the sudden oak death pathogen *Phytophthora ramorum* in Oregon from 2001 to 2004. Molecular Ecology. 16: 2958–2973.

Roewer, L.; Croucher, P.J.: Willuweit, S. et al. 2005. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Human Genetics, 116: 279–291.

Vercauteren, A.; Van Bockstaele, E.; De Dobbelaere, I. et al. 2010. Clonal expansion of the Belgian *Phytophthora ramorum* populations based on new microsatellite markers. Molecular Ecology. 19: 92–107.