Reconstructing the invasion history of *P. ramorum* in California

Croucher P., Mascheretti S., Garbelotto M.

Genetic epidemiology

- Understanding:
  - where introduction(s) occurred
  - invasion pathways
  - how organism is reproducing
  - microevolution and adaptation

Important to formulate

a)- future spread

b)- future impacts

c)- effective control strategies

d)- preventive regulations

Genetic epidemiology

- Understanding:
  - where introduction(s) occurred
  - invasion pathways
  - how organism is reproducing
  - microevolution and adaptation

- PROBLEMS (Fitzpatrick et al. Biological Invasions 2012)
  - Short time since introduction, genetic variability predates introduction
  - Lack of equilibrium
  - Most assumption of available population genetics approaches and software programs are violated

P. ramorum in California:

- Ivors et al. 2004 Mycol. Res. and 2006 Mol. Ecol.:
  - Only NA1 lineage present in California forests, but three lineages present in CA nurseries
  - Forest populations reproducing only clonally (Ia)
  - NA1 lineage shows signs of a very strong bottleneck, high genetic similarity and hard to find polymorphisms (Prospero et al 2006, Ivors et al. 2006)

P. ramorum in California:

- Mascheretti et al. 2008 & 2009 Mol. Ecol.:
  - Analytical power increased by discovering *2n* isolates were *mostly* either homozygous, or only *one* of the two alleles changed: can be treated as *n*
  - Generated a tree based on two different genetic similarity values (PHIst and MSN distances). In both cases, nursery populations were at one end, and recently established populations at the other, suggesting nurseries as a source
  - Used a completely different approach (MDS) to show that nursery populations, most of Santa Cruz Co., and one Marin Co. pop were tightly clustered, while recent pops were more distant

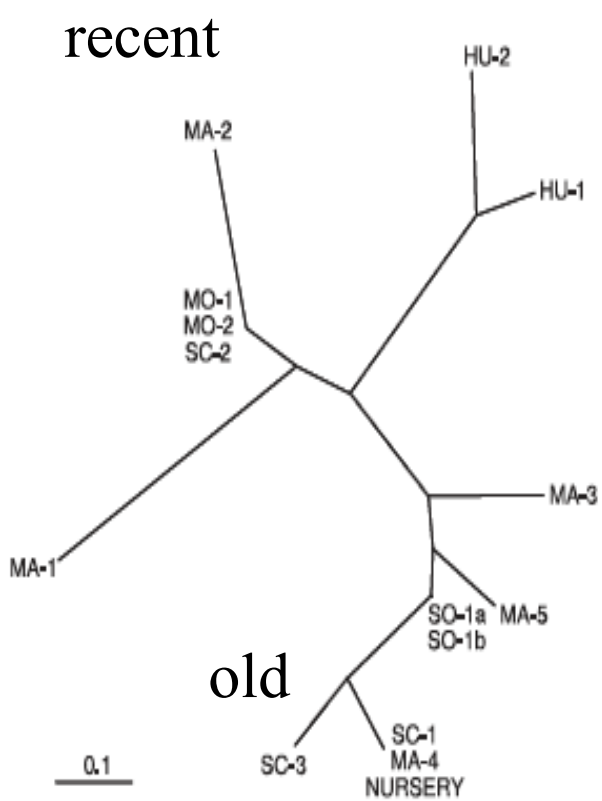
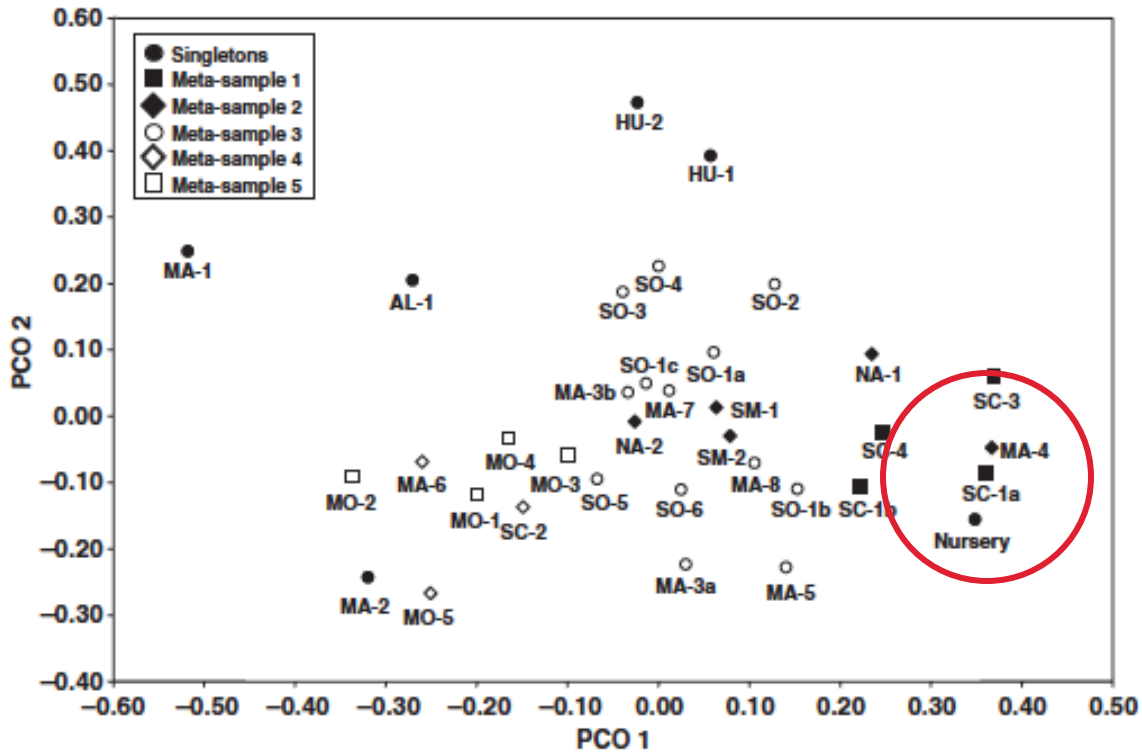


Fig. 3 Neighbour-joining tree illustrating the relationships between statistically significantly differentiated *Phytophthora ramorum* population samples and meta-samples, formed by clustering samples according to minimum Φ_{ST}.



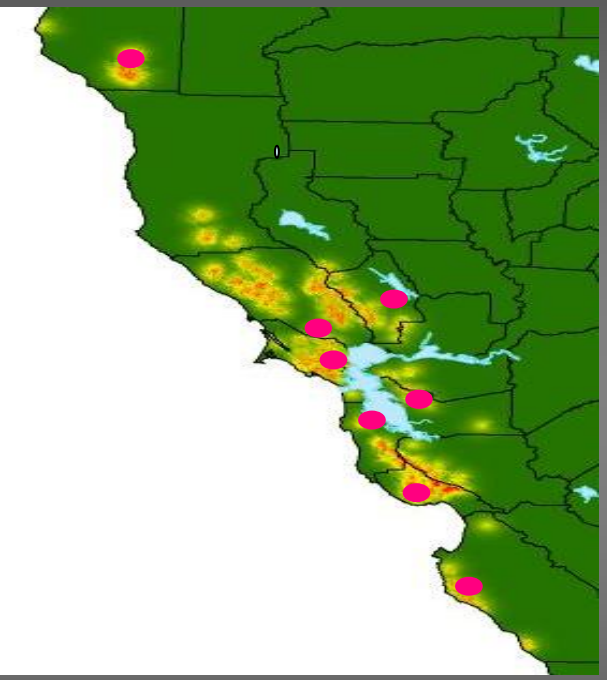
Multi Dimensional Scaling: US Nurseries, Scotts Valley (SC1), and Marin Mount Tam (MA4) are very close, most recent infestations are distant. This implies genetic differences are accumulating with time, but certain mutations could have happened early (AL-1) so interpretation not straightforward

P. ramorum in California (continued-1):

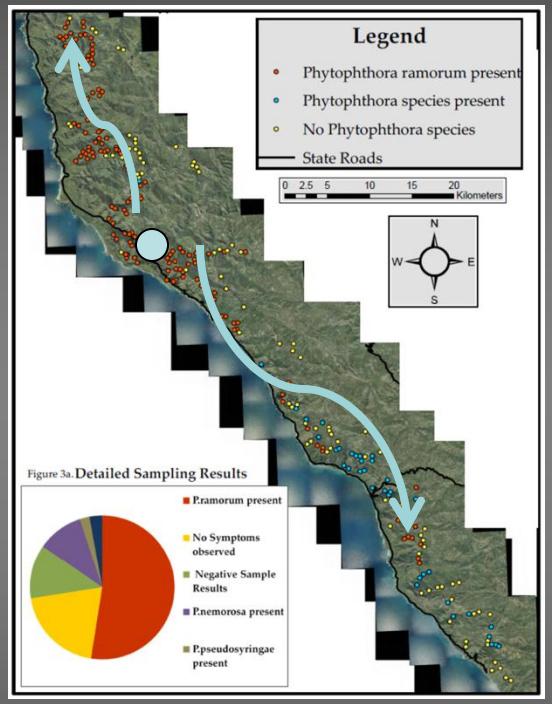
- Mascheretti et al. 2009 Mol. Ecol.:
  - Genetically identical populations that are disjunct and at significant distances imply human-mediated movement. Identified at least 6 of these on top of Santa Cruz and Marin Counties initial introductions, bringing the number of quasi simultaneous introductions of Pram to a minimum of 8 different locations
  - Many populations are undistinguishable and contiguous over a large area. Some of them are likely the result of spread from a single introduction (Big Sur), some others could actually be the results of multiple introductions from the same source (Marin and West Sonoma), but we cannot tell

Human-mediated transport: identical but disjunct pops

Identical and contiguous pops may represent two very different scenarios



Multiple introductions explain why range so large in spite of limited dispersal ability of pathogen. Sampling incomplete, for sure more introductions occurred

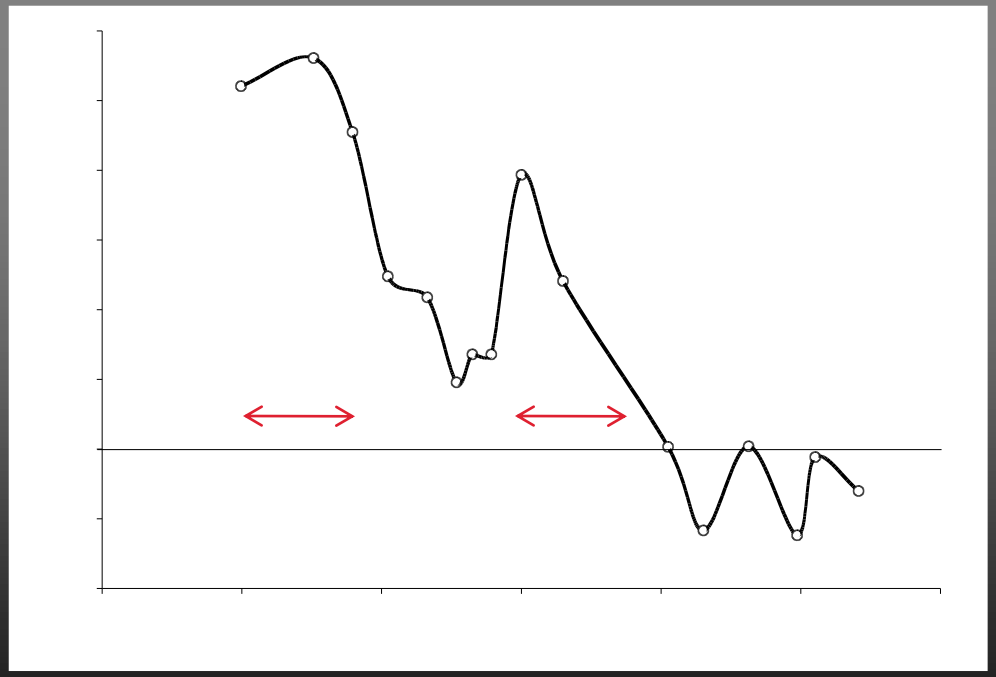


Big Sur: outbreaks were observed to progress in time from the Pfeiffer Big Sur area and to be genetically identical

Western Sonoma and Central Marin: are they really the result of a single introduction or genetic identity is due to same source?

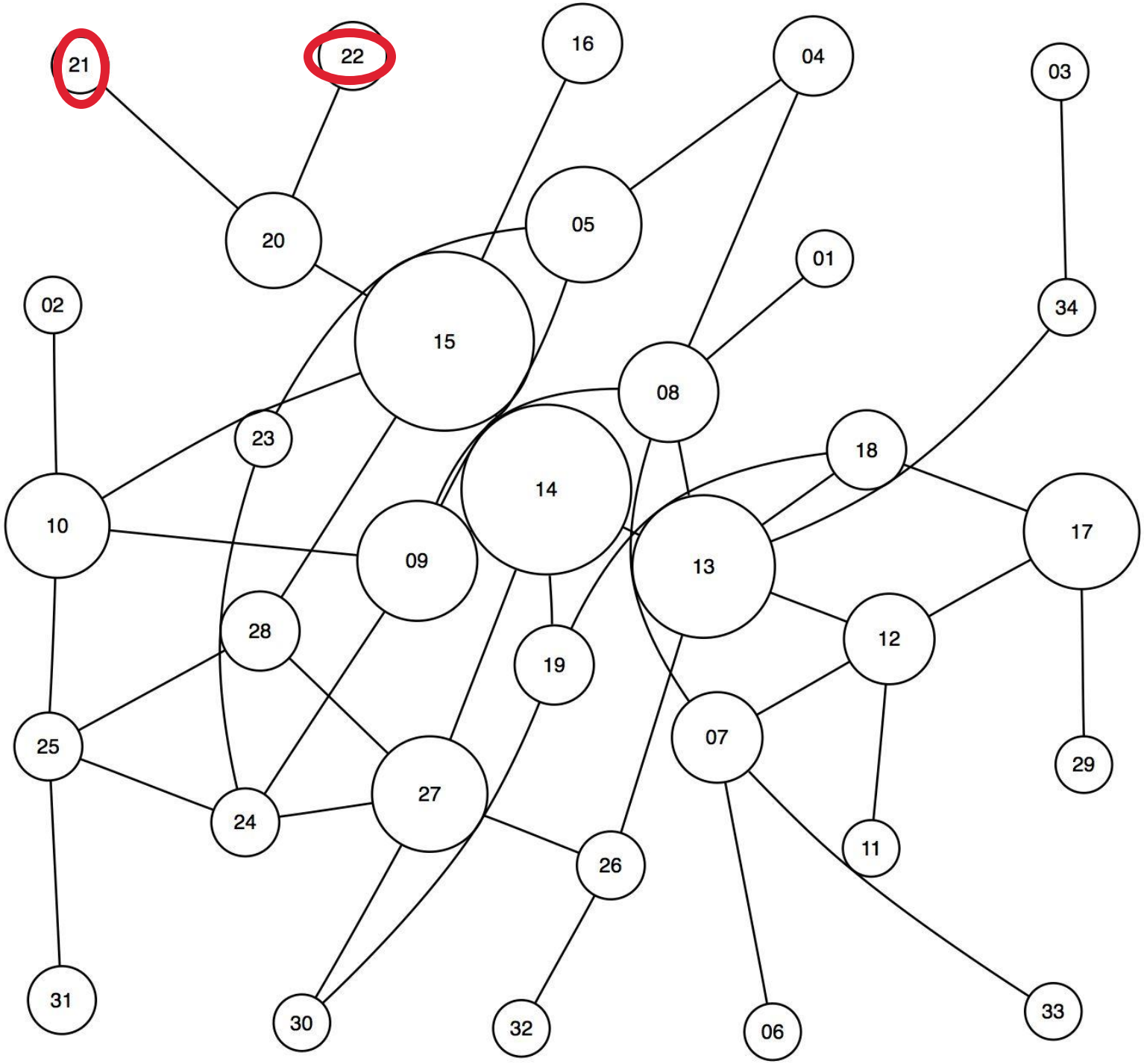
P. ramorum in California (continued-1):

- Mascheretti et al. 2008 & 2009 Mol. Ecol.:
  - Showed through spatial autocorrelation analyses that allelic composition changes with distance and same alleles found up to 500m and then between 1 and 4 Km



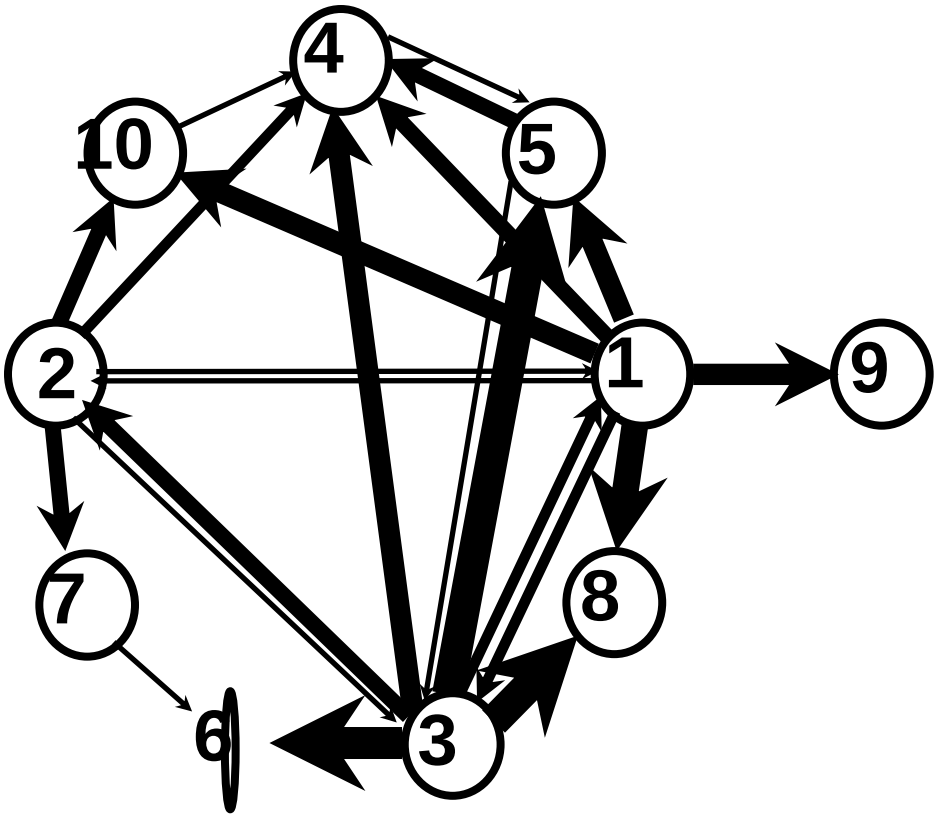
P. ramorum in California (continued-2):

- Mascheretti et al. 2008 & 2009 Mol. Ecol.:
  - Statistically showed that new genotypes are arising locally, but locally generated genotypes come and disappear. Three sites analyzed at 4 years interval had the same genetic composition. Three most abundant genotypes in CA thus are the *likely* founder progenitors of CA infestation

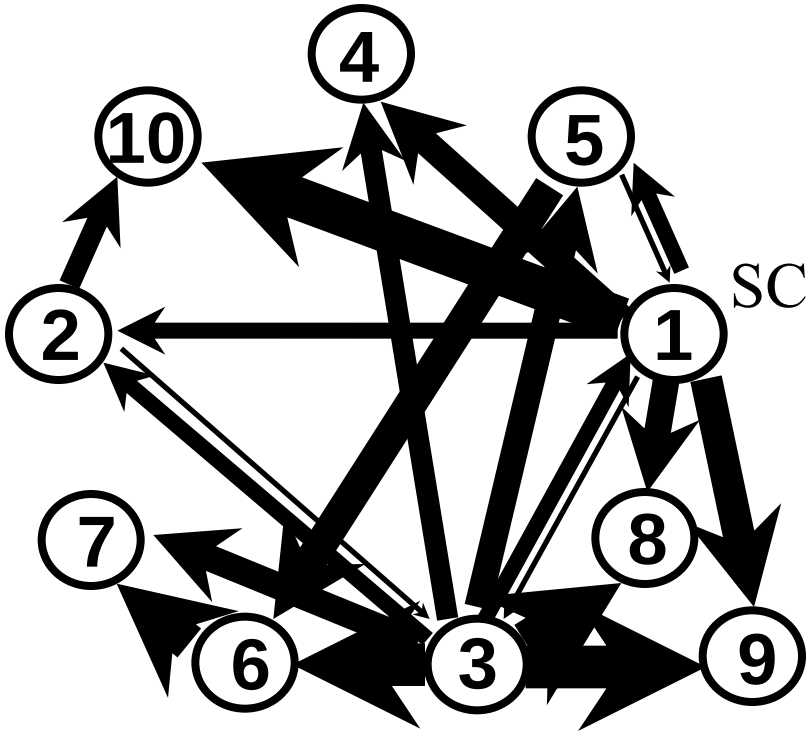

21
22
16
04
03
20
05
01
02
15
34
08
23
18
14
10
09
13
17
28
12
19
25
07
29
27
24
11
26
31
30
32
06
33

Story still very incomplete…

- Sampling lacked several important infestations (East Bay, Mendocino, Northern Humboldt, Santa Clara)
- Large meta-populations were confounding the analyses, also many isolates were excluded because could not be treated as haploids
- Only populations could be analyzed, but single isolates could not be placed: problem for young or small outbreaks
- We could not reconstruct spread pathways through coalescent analysis (MIGRATE N), except for saying that Santa Cruz and Marin infestations were important. Some linkages looked dubious, too many holes in sampling. Basically an impasse on the most important deliverable
- We could not unequivocally prove nurseries were source of forest infestation. We suspected a nursery in Scott's valley as a primary source but we had no evidence
- We could not prove our hypothesis of common genotypes being founders was correct and thus we could not reconstruct microevolutionary history of pathogen



A



B

Marin

"Migration" events:

Sampling lacked several important infestations (East Bay, Mendocino, Northern Humboldt, Santa Clara)

- We included samples from all infestations known as of 2011
- Analyzed 813 samples from 60 locations in 13 counties
- Analyzed 14 samples from US nurseries and 5 samples from Scott's valley nursery in CA

Large meta-populations were confounding the analyses, also many isolates were excluded because could not be treated as haploids

- Use Bruvo distances among genotypes as input for most analyses: this metric penalizes multiple mutations (larger allelic changes). It is the most suitable metric when analyzing a population of extremely closely related individuals (smaller changes should be more likely than bigger changes, if we are confident source is uniform). Now, we can use all isolates

- Used ANOVA to compare validity of metapopulations generated by pooling populations with statistically insignificant PHIst values, with those with statistically insignificant PHIst values but only within the same county and…

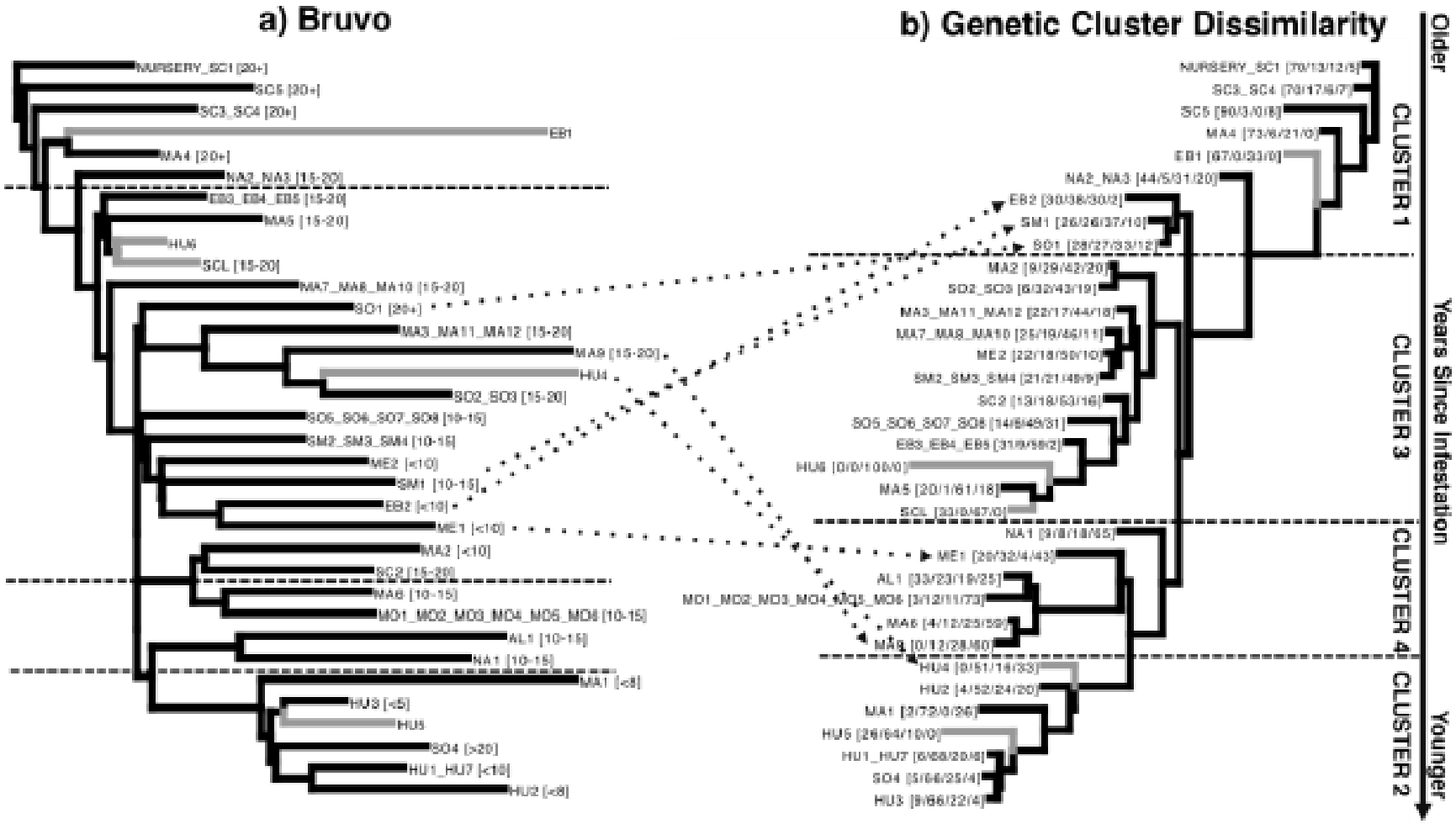
Results of AMOVAs

- **Iterative within-county collapsing reduced the initial set of 43 populations (*n* > 5) to 29 metapopulations with no significant genetic differences**

- **Metapopulations within county maximized genetic variance across metapopulations (30%, P=0.0001) and minimized genetic variance within metapopulations (1%, non significant)**

NJ Analysis using Bruvo distances and metapopulations grouped by county: caldes mostly match groups obtained by Bayesian grouping generated by STRUCTURE analysis

Fig. 2

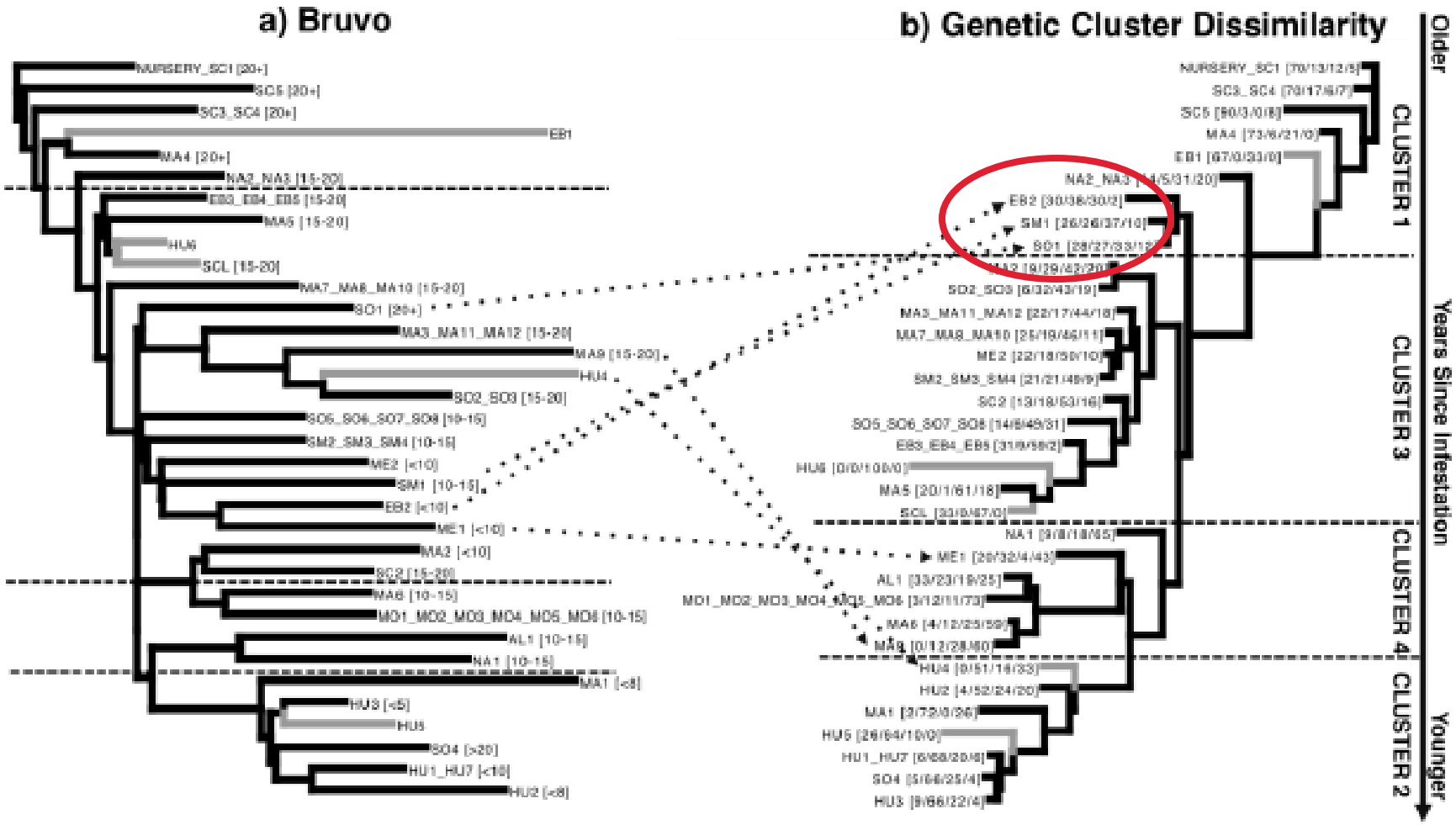


Clusters are interestingly associated with outbreaks <u>of different age</u>, nursery cluster 1 only significant in Santa Cruz county

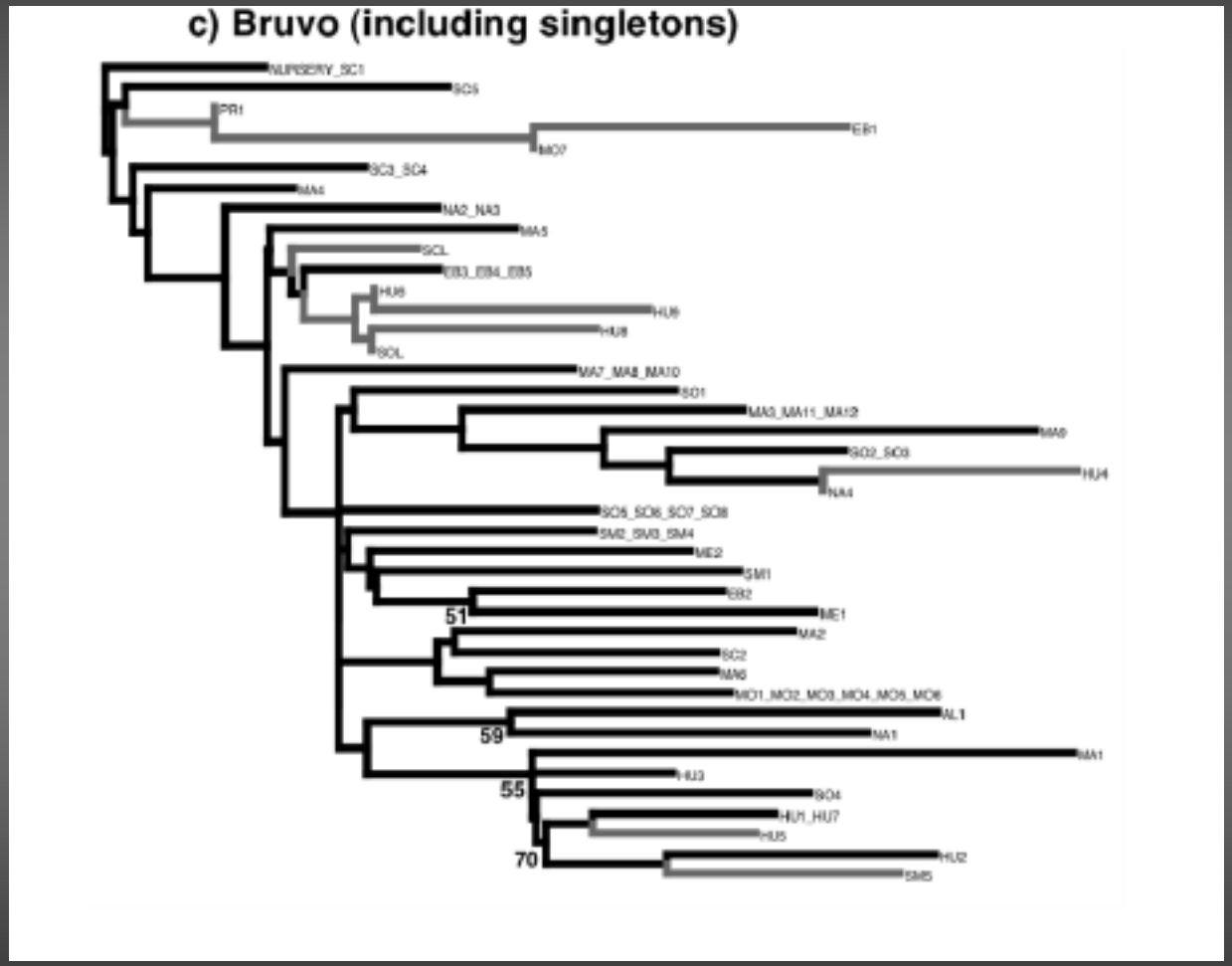
- Congruence between analyses shows the presence of four distinct genetic clusters in California forests
- Cluster 1 is the oldest, predominantly linked to nurseries and surrounding forest, but it is not well represented anywhere else
- Cluster 3 is the most represented one: mostly Marin and Sonoma
- Cluster 4 Big Sur, and potentially other nursery-associated outbreaks such as Castro Valley Alameda, Skyline Napa
- Cluster 2 Sonoma Mountain- Humboldt

The few incongruent placements are normally associated with smaller outbreaks and may indicate an early reversion to Cluster 1 from Cluster 3 (supported by coalescent analysis)

Fig. 2



Can we assign single isolates using the backbone NJ tree: YES!



Identification of 4 clusters and assignment of individual genotypes allows to:

1- Identify nursery-related outbreaks (Cluster 1)

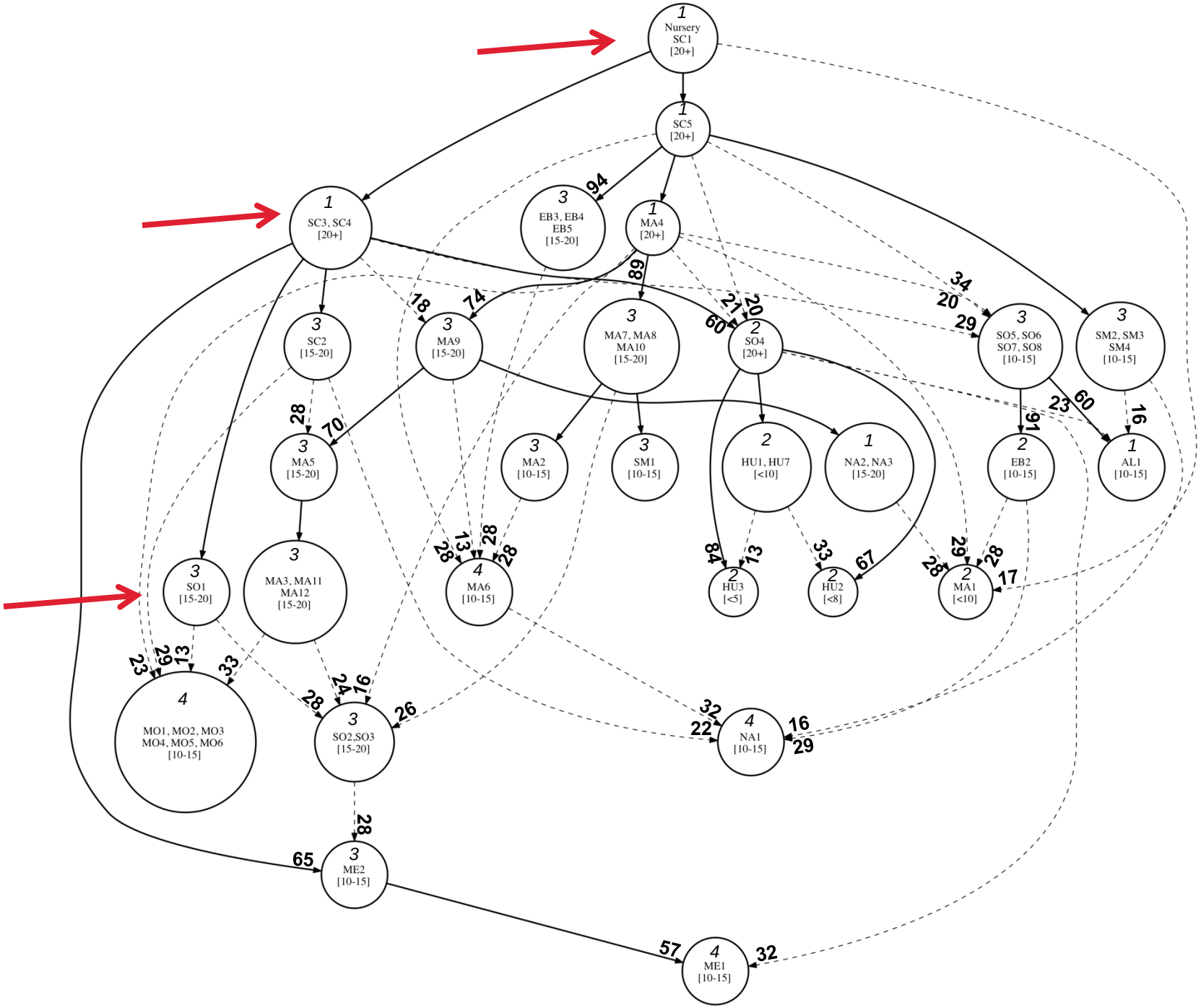
- Scott's Valley (SC Co.)- old
- Mount Tam (Marin)- old
- East Bay (Contra Costa)- intermediate
- Prunedale (Monterey)- intermediate
- Crystal Springs SFPUC (San Mateo)- interm.
- Jack London (Sonoma)- interm.
- Presidio San Francisco- recent

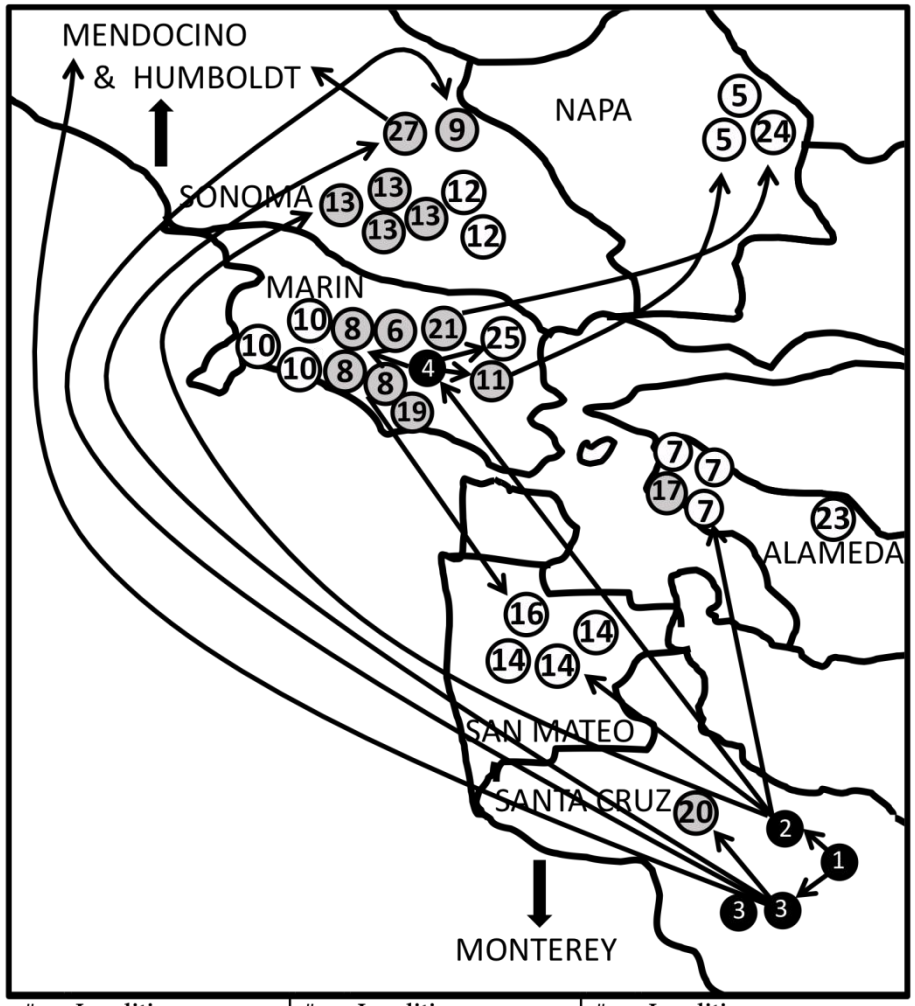
2- Identify multiple introductions in several counties (more than one cluster present)

- Marin
- East Bay counties
- Sonoma
- Humboldt
- Napa county
- San Mateo

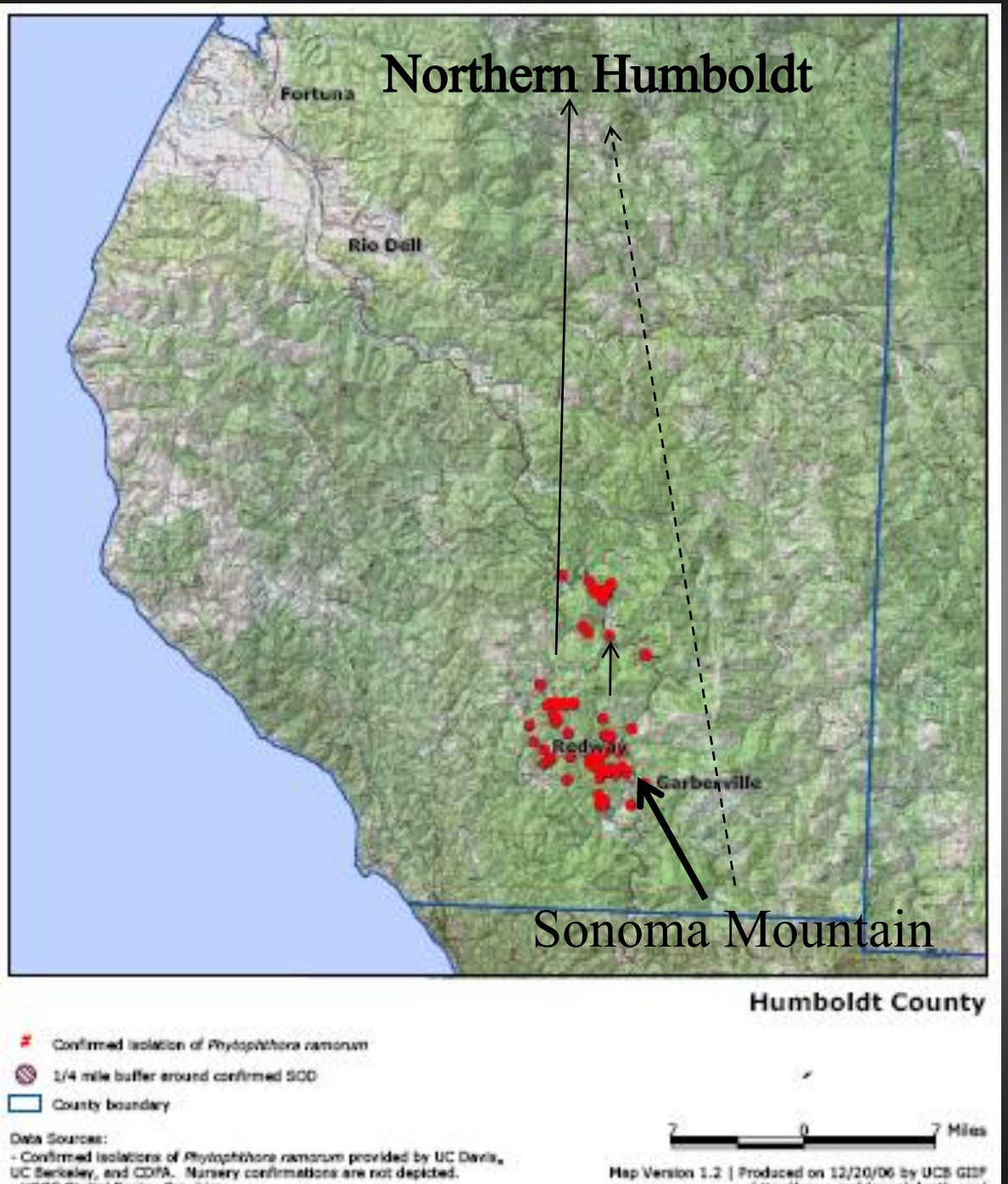
Coalescent analysis with these many pops computationally intractable

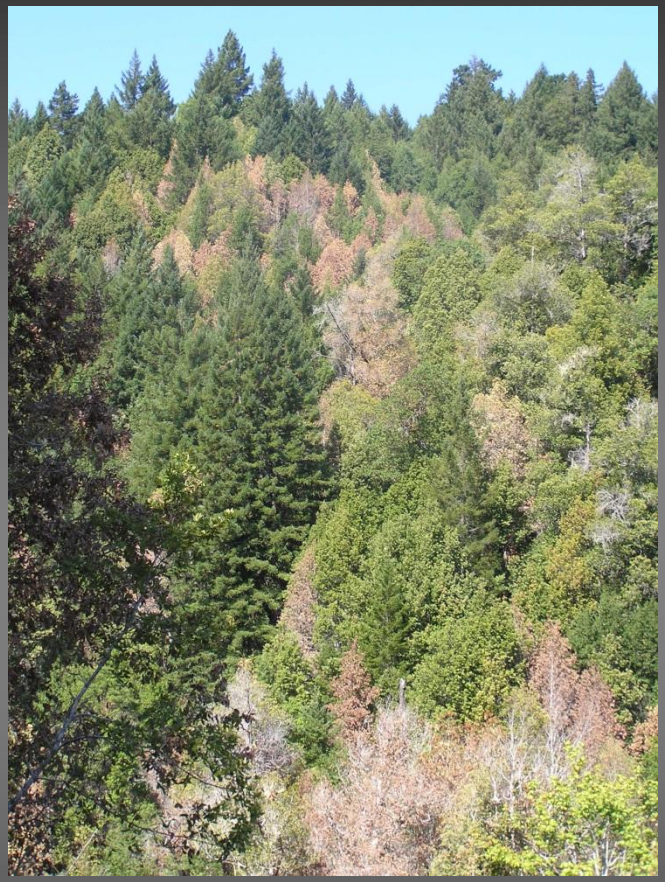
- Rather than estimate *M*, we used the output *ln* marginal likelihoods (*ln*(*ml*)), repeated 100 times
- We incorporated field epidemiological data by not allowing younger infestations to be a source for older infestations
- Analysis robustly identified source for 79% of populations, another 10% can be selected using cumulative information from this and previous studies


1 Nursery SC1 [20+]
1 SC5 [20+]
1 SC3, SC4 [20+]
3 EB3, EB4 EB5 [15-20]
1 MA4 [20+]
3 SC2 [15-20]
3 MA9 [15-20]
3 MA7, MA8 MA10 [15-20]
2 SO4 [20+]
3 SO5, SO6 SO7, SO8 [10-15]
3 SM2, SM3 SM4 [10-15]
3 MA5 [15-20]
3 MA2 [10-15]
3 SM1 [10-15]
2 HU1, HU7 [<10]
1 NA2, NA3 [15-20]
2 EB2 [10-15]
1 AL1 [10-15]
3 SO1 [15-20]
3 MA3, MA11 MA12 [15-20]
4 MA6 [10-15]
2 HU3 [<5]
2 HU2 [<8]
2 MA1 [<10]
4 MO1, MO2, MO3 MO4, MO5, MO6 [10-15]
3 SO2,SO3 [15-20]
4 NA1 [10-15]
3 ME2 [10-15]
4 ME1 [10-15]



#	Localities	#	Localities	#	Localities
1.	Nursery + SC1	11.	MA9	21.	MA6
2.	SC5	12.	SO2, SO3	22.	MO1, MO2, MO3, MO4, MO5, MO6 (N/S)
3.	SC3, SC4	13.	SO6, SO6, SO7, SO8	23.	AL1
4.	MA4	14.	SM2, SM3, SM4	24.	NA1
5.	NA2, NA3	15.	ME2 (N/S)	25.	MA1
6.	MA5	16.	SM1	26.	HU3 (N/S)
7.	EB3, EB4, EB5	17.	EB2	27.	SO4
8.	MA7, MA8, MA10	18.	ME1 (N/S)	28.	HU1, HU7 (N/S)
9.	SO1	19.	MA2	29.	HU2 (N/S)
10.	MA3, MA11, MA12	20.	SC2		


Northern Humboldt
Fortuna
Rio Dell
Redway
Garberville
Sonoma Mountain
Humboldt County
Confirmed isolation of Phytophthora ramorum
1/4 mile buffer around confirmed SOD
County boundary
Data Sources:
- Confirmed isolations of Phytophthora ramorum provided by UC Davis, UC Berkeley, and CDFA. Nursery confirmations are not depicted.
7 0 7 Miles
Map Version 1.2 | Produced on 12/20/06 by UCB GIIF

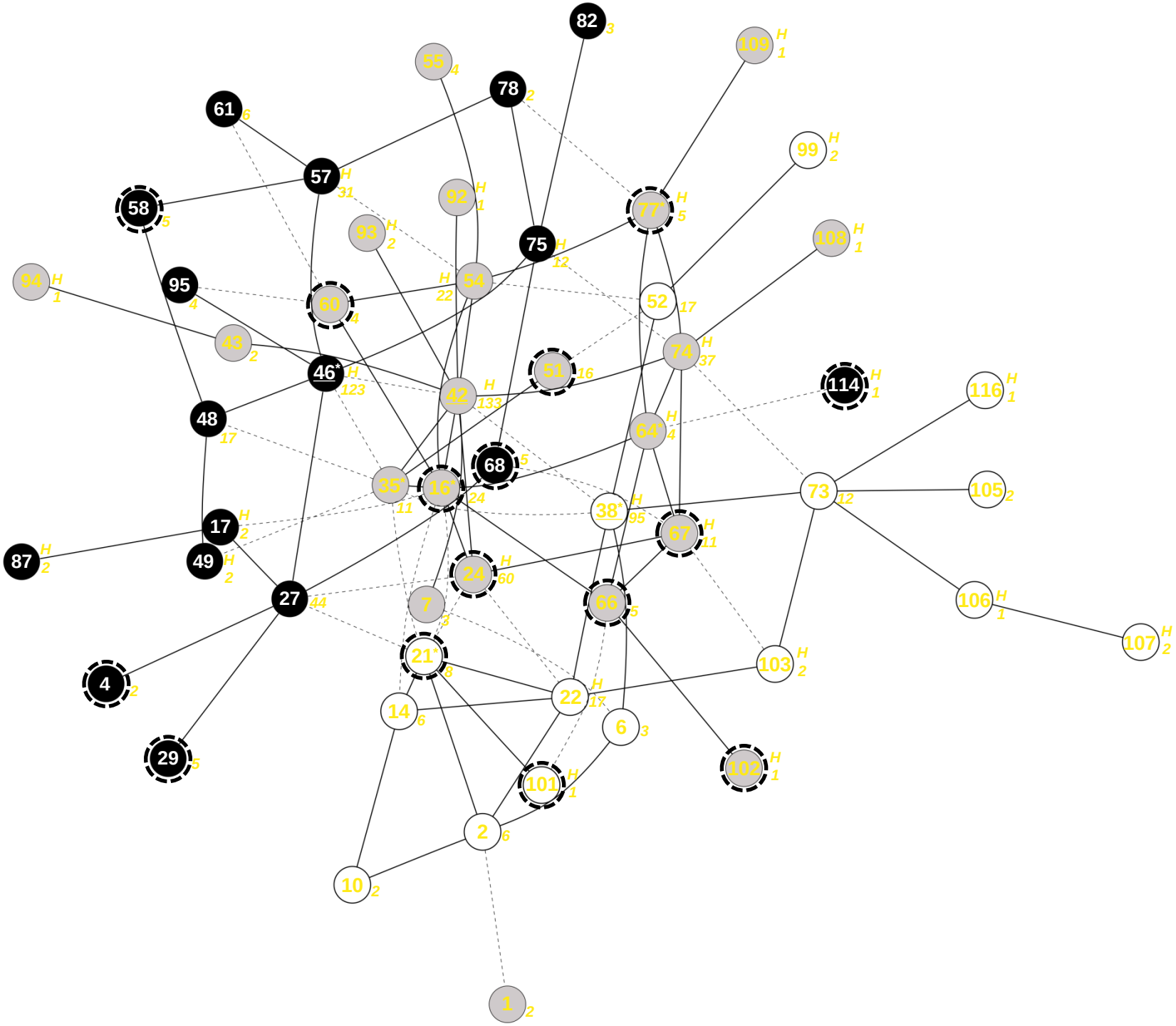


Coalescent analysis

- Reconstructs pathways of spread highlighting both the presence of both long distance human mediated spread and local natural spread
- Nurseries are placed as the source of entire California infestation without any assumptions
- Multiple introductions in several counties
- Large wild outbreaks are the most important sources of further infestations: size of pathogen population matters
- Many outbreaks are equally aged but disease incidence drastically different, suggesting strong different ecological constraints

Microevolution

- Are the most abundant genotypes the ones that were introduced or was there a shift in frequency?
- Do the four clusters we identified correspond to four founder genotypes?
- What is the evolutionary relationship between the clusters (genotypes)?


82 3
109 H 1
55 4
78 2
61 6
99 H 2
57 H 31
92 H 1
77* H 5
58 5
93 H 2
75 H 12
108 H 1
94 H 1
95 4
54 H 22
60 4
52 17
43 2
74 H 37
46* H 123
51 16
114 H 1
116 H 1
42 H 133
48 17
64* H 4
68 5
35* 11
16* 24
73 12
105 2
38* H 95
17 H 2
67 H 11
87 H 2
49 H 2
24 H 60
27 44
7 3
66 5
106 H 1
107 H 2
21* 8
103 H 2
4 2
22 H 17
14 6
6 3
29 5
102 H 1
101 H 1
2 6
10 2
1 2

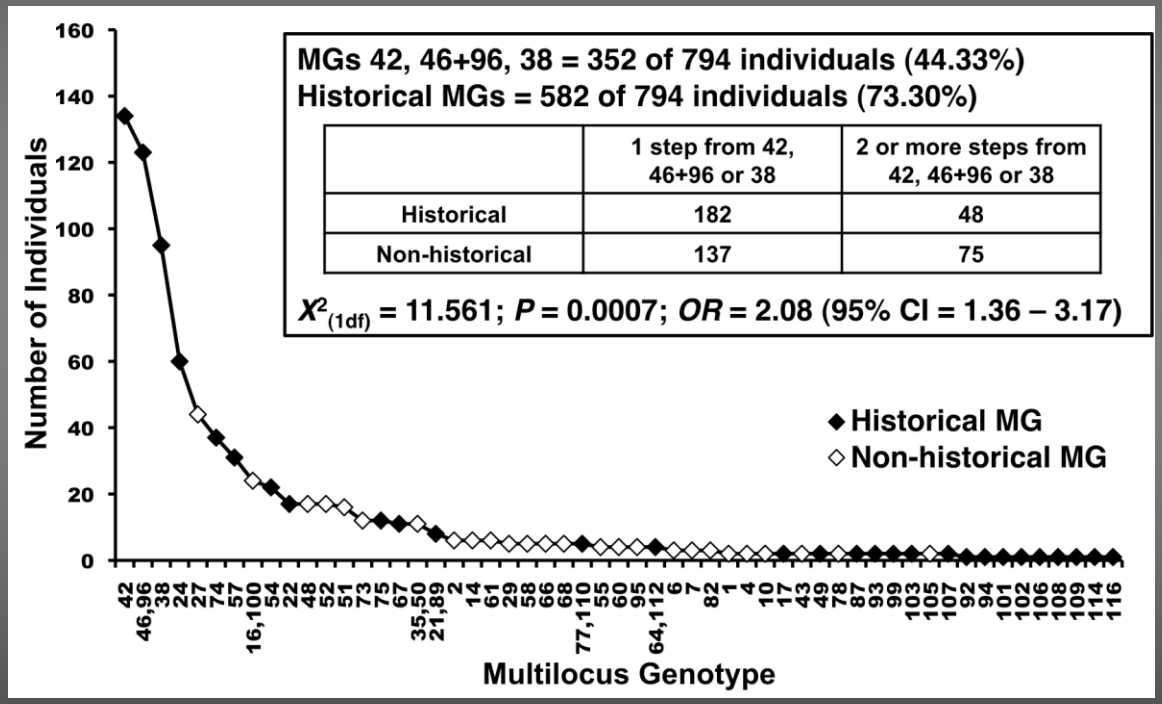
- Three of the four STRUCTURE clusters (1, 3, and 4) correspond to three linked subnetworks of connected genotypes

- Number of repeats at locus MS43b associated with each of the three clusters
  - Cluster 1= repeat no.= 93 (ancestral)
  - Cluster 3= repeat no.= 92
  - Cluster 4= repeat no.= 91

- Cluster 2 not contiguous, contains genotypes derived from all three other clusters= most recent

- All four subnetworks have one central (or most represented for 2) genotype. The central genotypes for clusters 1, 3,4 correspond to the most abundant genotypes in Mascheretti et al. 2009

In order to test the validity of our assumption we tested whether these putative founder genotypes were more abundant in historical isolates collected 2000-2005



CONCLUSION-1

one of the best reconstructions for a forest invasion

- Success in reconstructing the invasion history of *P. ramorum* in California due to the use of Bruvo distances, appropriate for populations evolving from closely related individuals
- Congruence of different analyses (NJ, STRUCTURE, MSN) strengthens validity of results
- Despite short age of invasion, lack of sexual recombination and short dispersal range allow for a reconstruction of its history
- Intractable coalescent analysis made tractable by including field epidemiological data (age, splitting of genetically identical pops based on geography)
- Complete sampling (thanks to collaborators and citizen scientists)

CONCLUSIONS-2

- Nursery population confirmed as primary source without assumptions
- Four clusters identified, corresponding to three founder genotypes derived from a single nursery genotype
- Nursery genotype not as widespread as derived ones: adaptation or drift?
- Large wild populations as major sources: attempt to mitigate pathogen population size may reduce further spread
- At least one nursery escape in recent times (Presidio): this trickle effect from nurseries is hidden by huge "naturalized" populations but can lead to significant problem if accidental release of a different lineage
- Coalescent analysis confirms long distance, human-mediated, spread and also depicts local progressive spread
- Many infestations have comparable age but disease progressed at dramatically different rates. Multiple introductions in most counties
- Cluster 2 is complex, recent and may mark shift from original founder events
- We can use coalescent reconstruction to see what we could have prevented by lowering population sizes at different times (affecting different sources or nodes). Does that benefit justify the cost? As invasion progresses costs increase

Pacific Southwest US Forest Service

The Betty & Gordon Moore Foundation

The Koret Foundation

The Packard Foundation

NSF-NIH: the ecology of infectious diseases program