AN ABSTRACT OF THE DISSERTATION OF

Zhian N. Kamvar for the degree of Doctor of Philosophy in Botany and Plant Pathology presented on December 6, 2016.

Title: Development of Tools for Genetic Analysis of Clonal Populations and Applications

Abstract approved: _____

Niklaus J. Grünwald

Research in the population genetics of microbial plant pathogens requires the use of specialized analyses designed for clonal organisms to avoid violating the assumptions of traditional population genetic models. The software necessary for performing these analyses existed within several different programs that did not necessarily run on all computing platforms. This meant that researchers not only had to reshape their data into different formats for each analysis, but they also had to switch computing platforms, not only creating a drain in time, but also increasing the risk of propagating human error into the analysis. To address this problem, we created the software package *poppr*, written in the R statistical language, available on all computing platforms. This package is designed for analysis of clonal, partially clonal, and sexual populations, empowering researchers to perform their work in a reproducible manner. We additionally demonstrate the utility of *poppr* for both plant pathological and theoretical questions

by using real-world and simulated data. In chapter 4, we demonstrate evidence for at least two origins for the outbreak of the Sudden Oak Death pathogen, *Phytophthora ramorum* in Curry County, Oregon. In chapter 5, we use *poppr* to assess the power of the index of association with clone- correction, showing that clone-correction has the potential to reduce the power of detecting clonal reproduction. All of the software and analyses in this work were performed in an open and reproducible framework, serving as an example of the power of reproducible research in plant pathology.

Development of Tools for Genetic Analysis of Clonal Populations and Applications

by

Zhian N. Kamvar

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 6, 2016
Commencement June 2017

Doctor of Philosophy dissertation of Zhian N. Kamvar presented on December 6, 2016.

APPROVED:

Major Professor, representing Botany and Plant Pathology

Head of the Department of Botany and Plant Pathology

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Zhian N. Kamvar, Author

ACKNOWLEDGEMENTS

I would like to acknowledge... Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas vel eros sed mauris porttitor semper nec a orci. Nullam vestibulum mi nec condimentum posuere. Pellentesque eget diam id sapien aliquet ullamcorper. Pellentesque blandit nec lectus ut mollis. Praesent in facilisis justo. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Sed eget congue leo, sed consequat libero. In rutrum malesuada nisi. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Morbi sollicitudin tortor ut sem facilisis mollis.

CONTRIBUTION OF AUTHORS

The following people contributed to this dissertation:

Chapter 2

Javier F. Tabima assisted in writing the code and design, Niklaus J. Grünwald assisted in design, and editing of the manuscript.

Chapter 3

Zhian N. Kamvar, Meredeth M. Larsen, Alan M. Kanaskie, Everett M. Hansen, and Niklaus J. Grünwald...

Chapter 4

Jonah C. Brooks assisted in writing and testing the code. Niklaus J. Grünwald assisted in the design, coördination of the collaborative effort, and editing the manuscript.

Chapter 5

Niklaus J. Grünwald assisted in the design, and editing of the manuscript.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

This dissertation is for my grandfather, Franklin R. Hepner and for my mother, Jane H. Kamvar.

Chapter 1: Introduction

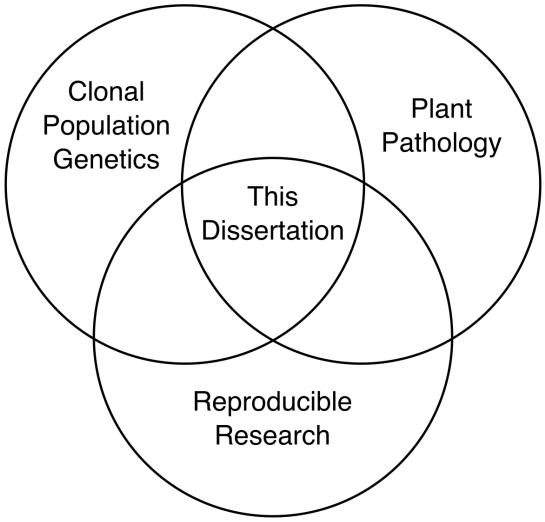


Figure 1.1: A graphical representation of the three themes governing the work presented in this dissertation.

The work presented here presents tools designed to answer questions related to clonal population genetics and plant pathology. The term 'tool' in this context is not in reference to new derivations of mathematical models or theory for population genetics of clonal organisms; rather, 'tool' refers to software code used to apply mathematical models or theory to population genetic data. The merit of this work lies within the context of reproducible science (Buckheit and Donoho, 1995). It provides the tools and demonstrates the usefulness and flexibility by applying these tools to the outbreak

of sudden oak death from 2001 to 2014 in Curry County, OR, USA and to simulation data, assessing the power of the index of association to detect clonal reproduction.

1.1 Clonal Population Genetics and Plant Pathology

Population genetics is the study of trying to understand the genetic variation within and among groups of organisms of the same species existing together in an evolutionary context (Milgroom and Fry, 1997).

Sex exists in all major groups of organisms (Heitman et al., 2012).

Pathogenic species can use reproductive mode to adapt to different environmental factors.

Knowing the reproductive mode of a pathogen can help you understand how to develop management strategies e.g. if you have a clonal population, you want to make sure you capture the diversity of genotypes before developing anti- microbial treatments (Taylor et al., 1999).

Population genetics in the context of plant pathology.

Life cycle concerns: overwintering oospores, sexual structures (image of *P. infestans life cycle*).

Sexual reproduction risk: *Puccinia* pathway.

Gene for Gene hypothesis

1.2 Scientific Software and Reproducible Research

1.2.1 Tools of Scientific Inquiry

> *Clearly, a discipline is defined by the questions asked, not the tools used.*
>
> – Milgroom and Fry (1997, 4)

The above quote by Michael Milgroom and William Fry in 1997 was in reference to the use of molecular markers in molecular biology as compared to population genetics. While it is undisputed that questions shape a field of inquiry, the notion that tools are not influential in disciplines is misleading. Tools are necessary for providing answers to the questions proposed; they are the vehicle by which we apply our scientific theory to the unknown world.

A tool, in this sense is any instrument, physical or analytical, that is used to collect, measure, manipulate, represent, or analyze data (Gigerenzer, 1991). This definition encompasses things like hammers, hand lenses, mass spectrometers, maps, axioms, algorithms, gel electrophoresis, equations, etc. All of these tools are used within a theoretical framework (e.g. gravity, refraction); any observations or results produced with a particular tool is ultimately tied to the theory employed by the scientist using it (and would thus invoke different interpretations under a different theoretical framework) (Kuhn, 1996). If all the assumptions of the theoretical framework are met, the tool will produce an observation or result that will help the scientist describe the natural phenomena accurately in terms according to theory.

These tools, however, should not simply be seen as a means to an end of answering

questions. Many tools will produce answers whether or not they are correct. A simple example of this concept was demonstrated by Anscombe, showing the need for graphical visualization in statistical analysis (1973). Reproduced in Fig. 1.2, four data sets are shown fitted with a trendline. Using linear regression, all four data sets produce the exact same result (slope, intercept, variance, correlation). Upon visual inspection, their differences are striking.

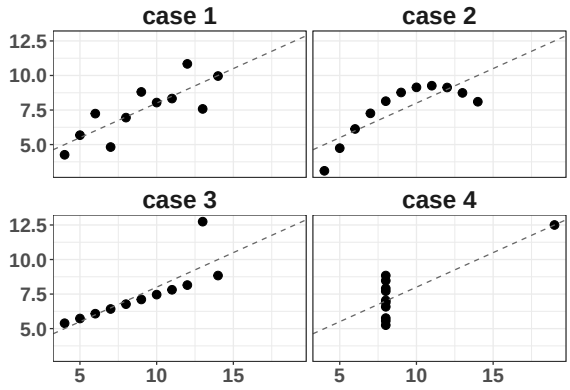


Figure 1.2: A reproduction of 'Anscombe's quartet' (Anscombe, 1973) demonstrating different situations in which linear regression would give the same answer.

If we imagine each data set as separate population and linear regression as our molecular marker, it wouldn't matter what our question was, because there would be no hope to detect any differentiation between these populations with the marker chosen.

Science moves forward by asking questions of natural phenomena and then investigating further the results of these questions, narrowing the scope of the succeeding questions to pin down an detailed mechanism that can explain the initial phenomena. The tools provide the observations and results that set the context for future questions

(Searls, 2010).

1.2.2 Scientific Software

The development of scientific software does not stand apart from science itself, but rather it serves as implementation of scientific theory (Baxter et al., 2006; Ouzounis and Valencia, 2003; Partridge et al., 1984; Searls, 2010)

Scientific software has been used to implement new theory (Agapow and Burt, 2001; Ali et al., 2016; Felsenstein, 1989).

Scientific software has been used to make accessible old theory and methods (Goudet, 1995).

Ultimately, publication and maintenance of scientific software exists to standardize the protocols in which we manipulate analyze our data.

1.2.3 Reproducible Research

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*
>
> – John Buckheit and David Donho paraphrasing John Claerbout (1995)

Buckheit and Donoho (1995) didn't invent the concept of reproducible research in scientific computing, but they did show computational research community that such

a concept was possible.

The ultimate goal of reproducible research is in the term itself: to ensure that research is produced in a manner such that future researchers can reproduce and verify the results (Goecks et al., 2010). This has been a problem in the past due to varying factors, including software (Ioannidis et al., 2008).

In terms of population genetics, software suites for analyses proliferated in the late 1990's and early 2000's (Adamack and Gruber, 2014; Excoffier and Heckel, 2006). The number of formats also proliferated.

Scientific software needs to be usable to be useful. Example in de Meeûs and Balloux (2004) using MULTILOCUS (Agapow and Burt, 2001).

The benefits of scientific software are many (McKiernan et al., 2016).

1.3 Goals

1.4 Part 1: Tools

1.4.1 Summary of Chapter 2

To address the lack of tools for reproducible research in clonal population genetics, we present the software package *poppr* in the R computing language (Kamvar et al., 2014b; R Core Team, 2016). Previously, tools necessary for analysis of clonal populations were available in several stand-alone software programs, each requiring different data input formats. Moreover, each program had different levels of documentation and limited support for all computing platforms. The novelty of *poppr* was to introduce indices of

multilocus genotype diversity, the index of association, and a fast implementation of Bruvo's genetic distance over unlimited levels of a user-specified population hierarchy (Agapow and Burt, 2001; Arnaud-Hanod et al., 2007; Bruvo et al., 2004). Because this was implemented in R these analyses could be performed in a reproducible manner on all computing platforms.

1.4.2 Summary of Chapter 3

The initial implementation of *poppr* contained basic tools for analysis of clonal populations (Kamvar et al., 2014b), but lacked tools for custom definitions of multilocus genotypes and performed poorly with genomic-scale data. Chapter 3 introduces an updated and improved *poppr* version 2.0. With high throughput sequencing (HTS) data, the amount of missing data and genotyping error increases, and the definition of a multilocus genotype becomes unclear. Moreover, the calculation of the index of association scales poorly with an increase in the number of loci. To address these limitations, we improved poppr with new functionalities to define multilocus genotypes based on genetic distance and calculate the index of association over random samples or windows of SNP loci.

1.5 Part 2: Applications

The end goal of part one was not simply the development of the software, but rather, we developed the software for the goal of analyzing our own data. Serendipitously, by

analyzing our own data with this software, we are able to not only demonstrate that it can be used, but also demonstrate that an entire analysis can be conducted in an open and reproducible manner. The next two chapters focus on answering practical and theoretical questions related to plant pathology and clonal population genetics, respectively.

1.5.1 Summary of Chapter 4

A newly-emerged disease of oak—called Sudden Oak Death—spread from California to the Southwest corner of Oregon in 2001. Because of intense management strategies, the epidemic was largely contained to Curry County for the next 15 years. In 2011, an isolated patch of disease appeared in Cape Sebastian, 12 miles from the nearest infected site. With microsatellite genotyping performed across 2 labs and 15 years, we sought to describe the spread of the epidemic in a population genetic context and ask the question of whether or not there was evidence for more than one introduction event. All the analyses were performed in an open-source and reproducible manner using R.

1.5.2 Summary of Chapter 5

The index of association is a measure of multilocus linkage disequilibrium, that is, a correlation coefficient across multiple loci. In sexual populations, loci are randomly assorting due to recombination, resulting in a near-zero value of the index of association.

In clonal populations, recombination is non-existent, meaning that loci are passed from parent to offspring in a non-independent fashion, resulting in a significantly non-zero value of the index of association. de Meeûs and Balloux (2004) showed that this index shows high variance with low levels of sexual reproduction, but due to limitations in software were not able to perform power analyses. We use *poppr* to investigate the power of the index of association to detect sexual reproduction in simulated data sets generated with microsatellite and genomic markers.

Chapter 2: *Poppr*: an R Package For Genetic Analysis of Populations With Clonal, Partially Clonal, and/or Sexual Reproduction

Zhian N. Kamvar, Javier F. Tabima, and Niklaus J. Grünwald

Journal: **PeerJ**
PO Box 614, Corte Madera, CA 94976, USA
Published 2014-03-04, Issue: **2**:e281, DOI: 10.7717/peerj.281

2.1 Abstract

Many microbial, fungal, or oomcyete populations violate assumptions for population genetic analysis because these populations are clonal, admixed, partially clonal, and/or sexual. Furthermore, few tools exist that are specifically designed for analyzing data from clonal populations, making analysis difficult and haphazard. We developed the R package poppr providing unique tools for analysis of data from admixed, clonal, mixed, and/or sexual populations. Currently, poppr can be used for dominant/codominant and haploid/diploid genetic data. Data can be imported from several formats including GenAlEx formatted text files and can be analyzed on a user-defined hierarchy that includes unlimited levels of subpopulation structure and clone censoring. New functions include calculation of Bruvo's distance for microsatellites, batch-analysis of the index of association with several indices of genotypic diversity, and graphing including dendrograms with bootstrap support and minimum spanning networks. While functions for genotypic diversity and clone censoring are specific for clonal populations, several functions found in poppr are also valuable to analysis of any populations. A manual with documentation and examples is provided. Poppr is open source and major releases are available on CRAN: http://cran.r-project.org/package=poppr. More supporting documentation and tutorials can be found under 'resources' at: http://grunwaldlab.cgrb.oregonstate.edu/.

2.2 Introduction

The Wright-Fisher model of populations is one of the oldest models utilized in population genetic theory. Populations in this model are characterized as having non-overlapping generations with a constant size free from any selective pressures (Hartl and Clark, 2007; Nielsen and Slatkin, 2013; Weir and Cockerham, 1996). Conceptually, these populations are represented as pools of alleles that are independently assorting where random mating is approximated by randomly sampling alleles with replacement from one generation to the next. Assumptions of this model, or related models, are implicitly assumed for common population genetic analysis tools. In clonal populations, however, alleles are not independently passed on from one generation to the next, and these assumptions are violated. Classical textbooks on population genetics do not provide much guidance on how to analyze clonal or mixed clonal and sexual populations. In reality, many populations are not strictly clonal or sexual, but can range from completely sexual to completely clonal and this is commonly observed for fungal, oomycete, or microbial populations (Anderson and Kohn, 1995; Milgroom, 1996). Currently, analysis of these populations is not straightforward as we lack the sophisticated tools and methods developed for model populations that are typically either haploid or diploid (Grünwald and Goss, 2011).

Inferring population structure with many commonly used model-based clustering approaches such as the program STRUCTURE (Pritchard et al., 2000) is inherently problematic for clonal populations. These approaches cannot be used as clonal populations violate basic assumptions of panmixia and Hardy-Weinberg equilibrium. Thus,

model free methods such as those relying on k-means clustering, dendrograms including bootstrap support for clades, or minimum spanning networks are more appropriate (Cooke et al., 2012; Goss et al., 2009; Mascheretti et al., 2008). Furthermore, analysis of mixed or clonal populations traditionally relies on calculation of diversity of genotypes observed and analysis of clone-censored versus non-censored populations (Grünwald and Hoheisel, 2006; McDonald, 1997; Milgroom, 1996). Clone censoring involves reduction of any population sample to a single observation for each multilocus genotype (MLG) in a population thereby approximating panmictic populations and removing the effect of genetic linkage (Milgroom, 1996). Analysis of diversity, in turn, involves calculation of the number of genotypes observed (richness), diversity, and evenness (Grünwald et al., 2003). Typical measures of genotypic diversity are borrowed from ecology and use either the Shannon-Wiener or Stoddart and Taylor index (Grünwald et al., 2003; Shannon, 2001; Stoddart and Taylor, 1988).

A critical aspect of analyzing clonal or mixed populations is testing a null hypothesis of panmixia (Milgroom, 1996). Testing of this hypothesis for potentially clonal populations typically relies on assessment of linkage disequilibrium among loci (Milgroom, 1996). This is achieved via calculation of the index of association or related indices in combination with resampling of the data to obtain a null distribution for the expectation of random mating (Brown et al., 1980; Burt et al., 1996; Milgroom, 1996; Smith et al., 1993). These approaches have, for example, been applied to *Pyrenophora teres* (Peever and Milgroom, 1994) and *Aphanomyces euteiches* (Grünwald and Hoheisel, 2006) and are routinely used in the analyses of clonal populations although they are not easily calculated given available software including MULTILOCUS, which is no

longer supported, and LIAN, which only works for haploids (Agapow and Burt, 2001; Haubold and Hudson, 2000).

Hierarchical sampling adds another layer of complexity to analysis of clonal populations. With microbial populations, the geographic structure of each population is not entirely clear, and it is often important to sample temporally to see if clones persist over time (Grünwald and Hoheisel, 2006). A common approach when faced with multiple levels of sampling is to create a separate data set for each level or combination of levels and to analyze them separately. However, the number of data sets undergo a factorial increase with each hierarchical level, therefore increasing the chances of human error in data reformatting or analysis. Thus, tools are needed for analysis of population data across hierarchies or subsets of data.

Here, we introduce the R package *poppr* that is specifically designed for analysis of populations that are clonal, admixed and/or sexual. *Poppr* complements and builds on previously existing R packages including *adegenet* and *vegan* (Jombart, 2008; Jombart and Ahmed, 2011; Oksanen et al., 2013) while implementing tools novel to R significantly facilitating data import, population genetic analyses, and graphing of clonal or partially clonal populations. These tools include among others: analysis across hierarchies of populations, subsetting of populations, clone-censoring, Bruvo's genetic distance Bruvo et al. (2004), the index of association and related statistics (Brown et al., 1980; Smith et al., 1993), and bootstrap support for trees based on Bruvo's distance. By providing a centralized suite of tools appropriate for many data types, this package represents a novel and useful resource specifically tailored for analysis of clonal populations.

2.3 Materials and Methods

2.3.1 Data import

Poppr allows import of data in several formats for dominant/codominant, haploid/diploid and geographic data. The R package *adegenet*, that defines the `genind` data structure that *poppr* utilizes, allows support for importing data natively from STRUCTURE, GENETIX, GENEPOP, and FSTAT. While these formats are very common and widely supported, these do not allow for import of geographic and/or regional data. Furthermore, *adegenet* will only handle diploids with this format, though manual import is possible. To aid in importing data, *poppr* has newly added the function `read.genalex()`, to read data from *GenAlEx* formatted text files into the `genind` data object of the package *adegenet* (Jombart, 2008; Jombart and Ahmed, 2011; Peakall and Smouse, 2006). *GenAlEx* is a popular add-in for MICROSOFT EXCEL that can handle data including codominant/dominant and haploid/diploid markers as well as geographic and regional data. This function further facilitates the import of haploid, geographic, and regional data.

Transferring data to new formats and manipulating data by hand, such as collapsing data into clones or subsetting data into different hierarchical levels, is tedious, creates redundancy, and can result in lost or misrepresented data. *Poppr* includes tools to automate such repetitive tasks. Many currently available data formats and software implementations allow analysis of only one or two levels of a population hierarchy. With *poppr* the user can import a single data set with an unlimited number of hierarchical levels. This is achieved by having the user combine the levels using a common delimiter

(e.g. "Year_Country_City"). These combined levels are then used as the defining population factor in the input file and can easily be manipulated within R.

2.3.2 Data analysis

Once data is imported into R, the user can dynamically access and manipulate the population hierarchy with the function `splitcombine()`, subset the data set by population with `popsub()`, and check for cloned multilocus genotypes using `mlg()`. For data sets that include clones, the *poppr* function `clonecorrect()` will censor exact clones with respect to any level of a population hierarchy by creating a new data set that includes only unique multilocus genotypes (MLGs) per population. A full list of functions available in *poppr* is provided in table 2.1.

Typical analyses in *poppr* start with summary statistics for diversity, rarefaction, evenness, MLG counts, and calculation of distance measures such as Bruvo's distance, providing a suitable stepwise mutation model appropriate for microsatellite markers (Bruvo et al., 2004). *Poppr* will define MLGs in your data set, show where they cross populations, and can produce graphs and tables of MLGs by population that can be used for further analysis with the R package *vegan* Oksanen et al. (2013). Many of the diversity indices calculated by the *vegan* function `diversity()` are useful in analyzing the diversity of partially clonal populations. For this reason, *poppr* features a quick summary table (Table 2.2) that incorporates these indices along with the index of association, I_A (Brown et al., 1980; Smith et al., 1993), and its standardized form, $\bar{r}_d$, which accounts for the number of loci sampled (Agapow and Burt, 2001).

Table 2.1: List of functions found in *poppr* and short descriptions.

Function	Description
Import/Export	
`getfile`	Provides a quick GUI to grab files for import
`read.genalex`	Read *GenAlEx* formatted csv files to a genind object
`genind2genalex`	Converts genind objects to *GenAlEx* formatted csv files
Manipulation	
`missingno`	Handles missing data
`clonecorrect`	Clone censors at a specified population hierarchy
`informloci`	Detects and removes phylogenetically uninformative loci.
`popsub`	Subsets genind objects by population
`shufflepop`	Shuffles genotypes at each locus using four different algorithms (details in table 2.3)
`splitcombine`	Manipulates population hierarchy
Analysis	
`bruvo.boot`	Produces dendrograms with bootstrap support based on Bruvo's distance
`bruvo.dist`	Calculates Bruvo's distance
`diss.dist`	Calculates the percent allelic dissimilarity
`ia`	Calculates the index of association
`mlg`	Calculates the number of multilocus genotypes
`mlg.crosspop`	Finds all multilocus genotypes that cross populations
`mlg.table`	Returns a table of populations by multilocus genotypes
`mlg.vector`	Returns a vector of a numeric multilocus genotype assignment for each individual
`poppr`	Returns a diversity table by population
`poppr.all`	Returns a diversity table by population for all compatible files specified
Visualization	
`greycurve`	Helper to determine appropriate parameters for adjusting the grey level for msn functions
`bruvo.msn`	Produces minimum spanning networks based off Bruvo's distance colored by population
`poppr.msn`	Produces a minimum spanning network for any pairwise distance matrix related to the data

Both measures of association can detect signatures of multilocus linkage and values significantly departing from the null model of no linkage among markers are detected via permutation analysis utilizing one of four algorithms described in table 2.3 (Agapow and Burt, 2001). The user can specify the number of samples taken from the observed data set to obtain the null distribution expected for a randomly mating population. Detailed examples of these analyses can be found in the *poppr* manual.

2.3.3 Visualizations

Poppr generates bar charts of MLG counts found within each population of your data set (Fig. 2.1). Histograms with rug plots for I_A and $\bar{r}_d$ allow visual assessment of the quality of the distribution derived from resampling to see if a higher number of replications are necessary (Fig. 2.2). *Poppr* automatically produces custom minimum spanning networks for Bruvo's or other distances using Prim's algorithm, as implemented in the package *igraph* Csardi and Nepusz (2006), with the functions `bruvo.msn()` for Bruvo's distance (Fig. 2.3) and `poppr.msn()` for any distance matrix. The combination of data structures from *adegenet* and *igraph* allow graphing that is color coded by population with vertices grouped by MLG (Csardi and Nepusz, 2006; Jombart, 2008; Jombart and Ahmed, 2011). *Poppr* also includes visualization of dendrograms using UPGMA Schliep (2011) and Neighbor-Joining Paradis et al. (2004) algorithms with bootstrap support for Bruvo's distance using the function `bruvo.boot()` (Fig. 2.4). Neither graphing of minimum spanning networks or dendrograms with bootstrap support are currently possible for populations in any other R packages.

Table 2.2: Summary table shown as it would appear in the R console produced by the `poppr()` function with 999 permutations to calculate I_A and $\bar{r}_d$ p-values from the `Aeut` data set in *poppr* from (Grünwald et al., 2003). `N`: census size, `MLG`: multilocus genotypes, `eMLG`: expected MLG based on rarefaction, `SE`: standard error from rarefaction, `H`: Shannon-Wiener Index, `G`: Stoddart and Taylor's Index, `Hexp`: Nei 1978 Expected Heterozygosity, `E.5`: Evenness (E_5), `Ia`: I_A, `p.Ia`: p-value for I_A, `rbarD`: $\bar{r}_d$, `p.rD`: p-value for $\bar{r}_d$. Table was obtained with the following code: `library(poppr); data(Aeut); poppr(Aeut, sample = 999)`.

	Pop	N	MLG	eMLG	SE	H	G	Hexp	E.5	Ia	p.Ia	rbarD	p.rD
	Athena	97	70	65.98	1.25	4.06	42.19	0.99	0.72	2.91	0.00	0.07	0.00
Mt.	Vernon	90	50	50.00	0.00	3.67	28.72	0.98	0.73	13.30	0.00	0.28	0.00
	Total	187	119	68.45	2.99	4.56	68.97	0.99	0.72	14.37	0.00	0.27	0.00

Table 2.3: Permutation algorithms in *poppr* implemented in the calculation of I_A and $\bar{r}_d$ p-values, iterated over all loci independently.

Method	Name	Units Sampled	With Replacement	Weight
1	`permutation`	alleles	No	-
2	`parametric bootstrap`	alleles	Yes	allele frequencies
3	`non-parametric bootstrap`	alleles	Yes	equal
4	`multilocus`	genotypes	No	-

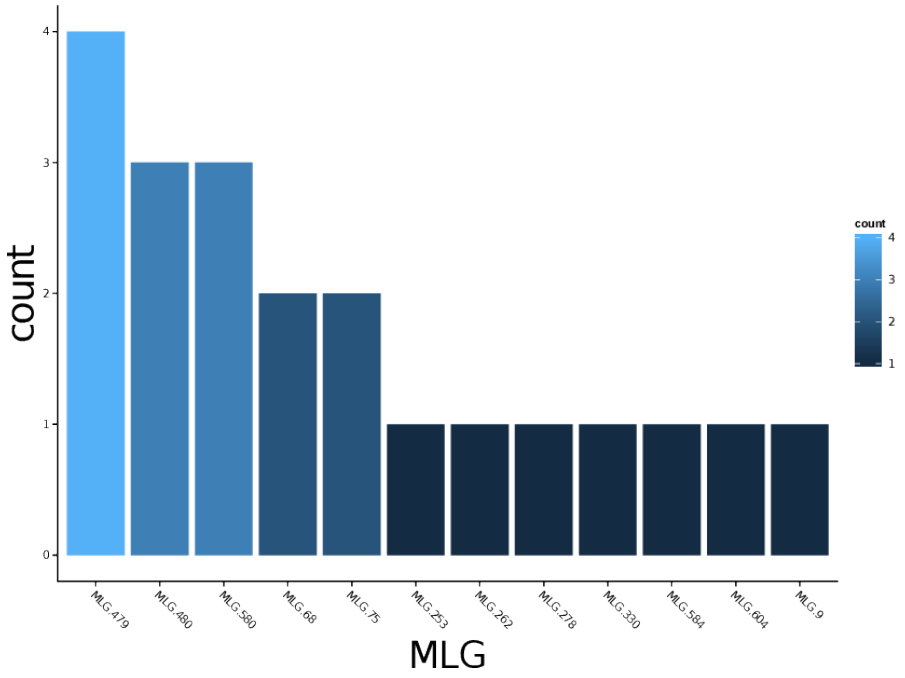


Figure 2.1: Distribution of 12 multilocus genotypes from the `Finland` population of the `H3N2` SNP data set (Jombart, 2008)

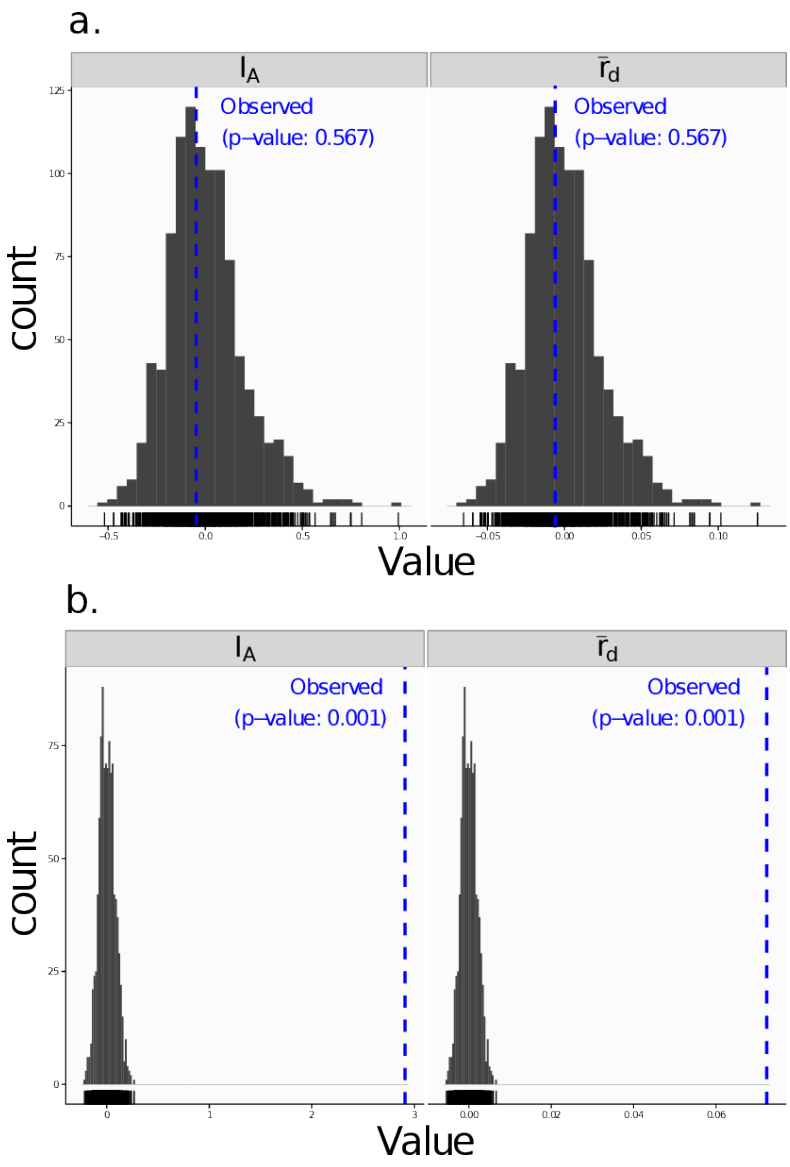


Figure 2.2: Visualizations of tests for linkage disequilibrium, where observed values (blue dashed lines) of I_A and $\bar{r}_d$ are compared to histograms showing results of 999 permutations using method 1 in table 2.1. Results are shown for the sexual population `5` of the `nancycats` data set (Jombart, 2008) (a) and for the clonal `Athena` population of the `Aeut` data set (Grünwald et al., 2003) (b).

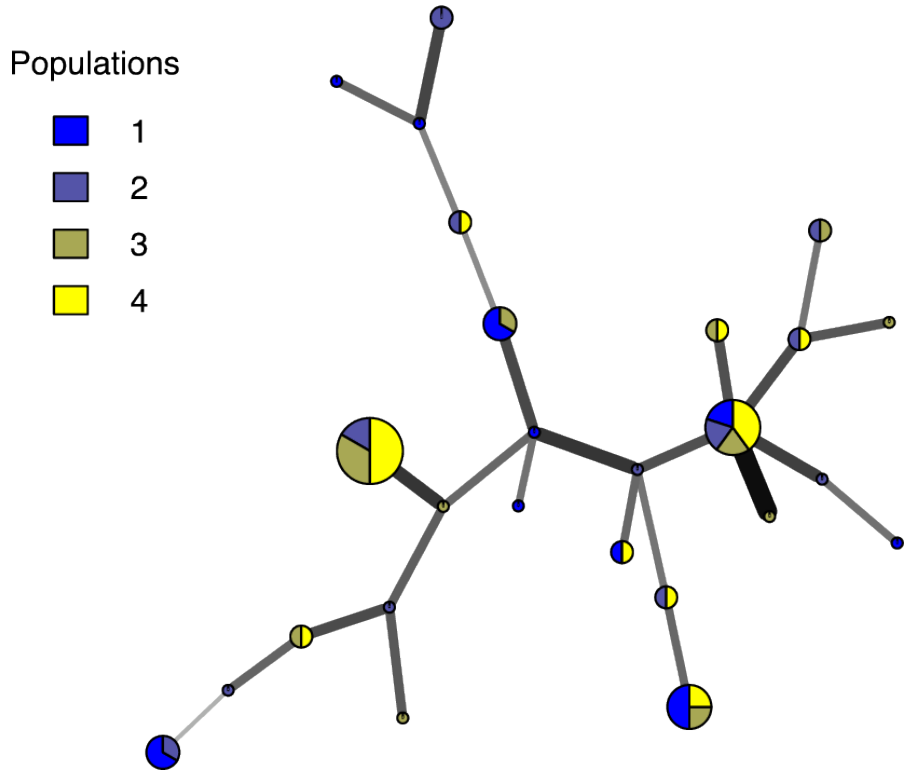


Figure 2.3: Example minimum spanning network using Bruvo's distance on a simulated partially clonal data set with 50 individuals genotyped over 10 microsatellite loci produced with the software SimuPOP v.1.0.8 (Peng and Amos, 2008). Each node represents a unique multilocus genotype. Node shading (colors) represent population membership, while edge widths and shading represent relatedness. Edge length is arbitrary.

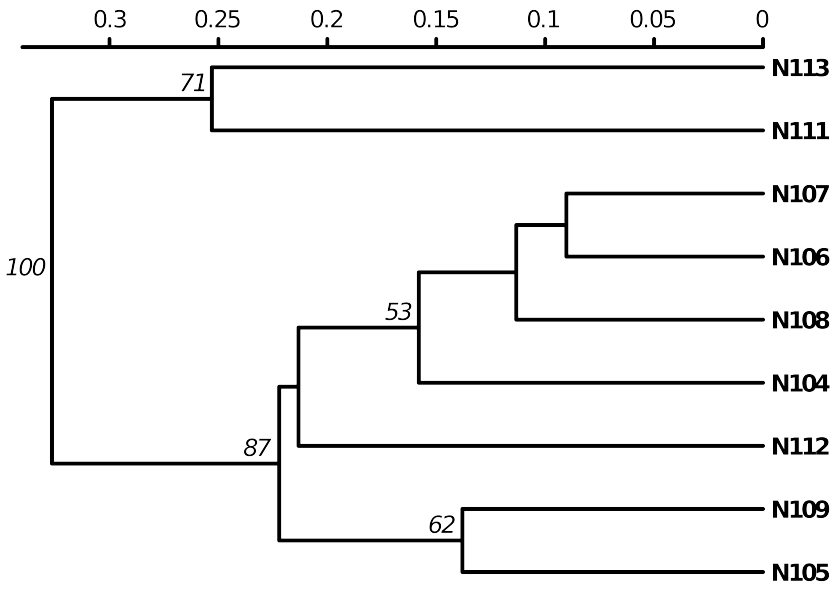


Figure 2.4: UPGMA tree produced from Bruvo's distance with 1000 bootstrap replicates (node values greater than 50% are shown). Data from population 9 of the `nancycats` data set (Jombart, 2008)

2.3.4 Performance

Most of the functions in *Poppr* were written and optimized for performance in R and are available for inspection and/or download at `https://github.com/grunwaldlab/poppr`. Algorithms of $\geq O(n^2)$ complexity were written in the byte-compiled C language to optimize runtime performance.

For comparisons of I_A and Bruvo's distance, we utilized the data set 'nancycats' (237 diploid individuals genotyped at nine microsatellite loci) from the *adegenet* package. Calculations were run independently 10 times and then averaged. Bruvo's distance was calculated on a machine with OSX 10.8.4 and a 2.9 GHz intel processor. The I_A

and $\bar{r}_d$ calculations were performed on a machine with OSX 10.5.8 and a 2.4 GHz intel processor due to the inability of the software MULTILOCUS to work on any later version of OSX.

2.4 Citation of methods implemented in *poppr*

Several of the methods implemented in *poppr* are described elsewhere. Users should refer to the original publications for interpretations and citation. See table 2.4 for a full list of citations. As with any R package, users should always cite the R Core Team (R Core Team, 2013).

Table 2.4: Citation of methods and indices implemented in *poppr*

Method/Index	**Citation**	**Function(s) in *poppr***
Expected MLG (rarefaction)	(Hurlbert, 1971), (Heck et al., 1975) (for std. err.)	`poppr()`
H	(Shannon, 2001)	`poppr()`
G	(Stoddart and Taylor, 1988)	`poppr()`
H_{exp}	(Nei, 1978)	`poppr()`
E_5	(Grünwald and Hoheisel, 2006), (Ludwig and Reynolds, 1988), (Pielou, 1975)	`poppr()`
I_A	(Brown et al., 1980),	`ia()`

Continued on next page...

Table 2.4 – continued from previous page

Method/Index	**Citation**	**Function(s) in *poppr***
	(Smith et al., 1993)	`poppr()`
$\bar{r}_d$	(Agapow and Burt, 2001)	`ia()`
		`poppr()`
Clone correction	(Grünwald and Hoheisel, 2006),	`clonecorrect()`
	(Grünwald et al., 2003),	`poppr()`
	(Milgroom, 1996)	
Minimum Spanning	(Csardi and Nepusz, 2006)	`poppr.msn()`
Networks		`bruvo.msn()`
Bruvo's Distance	(Bruvo et al., 2004)	`bruvo.dist()`
		`bruvo.msn()`
		`bruvo.boot()`
Bootstrapping	(Paradis et al., 2004)	`bruvo.boot()`
Neighbor Joining	(Paradis et al., 2004)	`bruvo.boot()`
UPGMA	(Schliep, 2011)	`bruvo.boot()`

2.5 Results and Discussion

Poppr provides significant, convenient tools for analysis of clonal and partially clonal populations available in one environment on all major operating systems. The ability to analyze data for multiple populations across a user-defined hierarchy and

clone-censoring provide novel functionality in R. Combined with R's graphing abilities, publication-ready figures are thus obtained conveniently.

2.5.1 New functionalities

Table 2.5: Comparison of programs that calculate I_A

	Haploids	**Diploids**	$\bar{r}_d$	**All Platforms**	**Batch Analysis**
poppr	Yes	Yes	Yes	Yes	Yes
LIAN	Yes	No	No	Yes	Yes
multilocus	Yes	Yes	Yes	No	No

Poppr implements several new functionalities. As of this writing, aside from *poppr*, there exist two programs that calculate I_A: LIAN (Haubold and Hudson, 2000) and MULTILOCUS (Agapow and Burt, 2001). LIAN can calculate I_A for haploid data and is only available online or for *nix systems with a C compiler such as OSX and Linux (Haubold and Hudson, 2000). MULTILOCUS implemented $\bar{r}_d$, a novel correction for I_A, but is no longer supported (Agapow and Burt, 2001). MULTILOCUS will only calculate index values for one data set at a time and LIAN requires the user to structure the data set with populations in contiguous blocks to analyze multiple populations within a single file. Thus *poppr* provides significant improvements for calculation of linkage disequilibrium, and handles both haploid and diploid data, works on all major operating systems, and is capable of batch analysis of multiple files and multiple populations

defined within a file including the possibility of clone correction and sub-setting. A comparison of the capabilities of these programs are summarized in Table 2.5.

To test significance for I_A and $\bar{r}_d$, *poppr* offers four permutation algorithms. Each one will randomly shuffle data at each locus, effectively unlinking the loci. The algorithm previously utilized by MUTLILOCUS is included. The MULTILOCUS-style algorithm shuffles genotypes, maintaining the associations between alleles at each locus (Agapow and Burt, 2001). More appropriately, alleles are expected to assort independently in panmictic populations. *Poppr* thus provides three new algorithms for permutation that allow for independent allele assortment at each locus. The default algorithm permutes the alleles at each locus and the remaining two will randomly sample alleles from a multinomial distribution parametrically and non-parametrically (Weir and Cockerham, 1996). Details of these algorithms are presented in table 2.3. Because the index of association is calculated using a binary measure of dissimilarity, we have also made this available as a distance measure called `diss.dist()`. This pairwise distance is based on the percent allelic differences.

Poppr also newly implements Bruvo's genetic distance that utilizes a stepwise mutation model appropriate for microsatellite data (Bruvo et al., 2004). While this distance is implemented in the program GENODIVE Meirmans and Van Tienderen (2004) and the R package *polysat* Clark and Jasieniuk (2011), there are a few caveats with these two implementations. GENODIVE is closed-source, and only implemented in OSX. Both *poppr* and *polysat* are open-source and available on all platforms, but *polysat*, being optimized for polyploid individuals with ambiguous allelic dosage, is inappropriate for analyzing diploids. *Polysat* will collapse homozygous individuals into a single

allele and attempt to infer the second allelic state in comparison with heterozygous individuals. Since haploid and diploid individuals show clear allelic dosage, this procedure creates a bias misrepresenting the true distance. Not only is *poppr* not subject to this bias, but it also newly introduces bootstrap support for this distance as shown in Figure 2.4.

2.5.2 Performance

Table 2.6: Comparison of performance on one data set of 237 individuals over nine loci. Each time point represents an average of 10 independent runs. Calculations of I_A are based on 100 permutations.

	I_A (seconds)	**Bruvo's distance (seconds)**
poppr	13.4	0.3
polysat	-	58.3
multilocus	547.2	-

Poppr reduces the amount of intermediate files and repetitive tasks needed for basic population genetic analyses and implements computationally intensive functions, such as Bruvo's distance and the index of association in C to improve performance. The *polysat* package calculation of Bruvo's distance took 58.3 seconds on average whereas *poppr*'s calculation was over 190 times faster, averaging 0.3 seconds (Table 2.6). For calculation of I_A and $\bar{r}_d$ with 100 permutations and Nei's genotypic diversity (Nei, 1978), MULTILOCUS required around 9.12 minutes on average, as compared to 13.4

seconds with *poppr*.

2.6 Conclusions

The R package *poppr* provides new functions and tools specifically tailored for analysis of data from clonal or partially clonal populations. No software currently available provides this set of tools. Novel capabilities include analysis across multiple populations at multiple levels of hierarchies, clone-censoring, and subsetting. These in combination with R's command line interface and scripting capabilities makes analyses of these populations more streamlined and tractable. By implementing computationally expensive algorithms such as Bruvo's distance and I_A in C, analyses of multiple populations that would normally take hours to complete can now be finished in a matter of minutes. This allowed us to expand the utility of these measures to convenient new graphing abilities such as automatically creating dendrograms with bootstrap support for Bruvo's distance and minimum spanning networks. While major releases of *poppr* are available on CRAN, we are continuing to develop this package to be able to efficiently handle genome-sized SNP data. Development versions are available on github at https://github.com/grunwaldlab/poppr.

2.7 Acknowledgements

The authors would like to thank Sydney Everhart and Corine Schoebel for invaluable alpha testing and Paul-Michael Agapow for providing the *multilocus* C++ source code

for reference. We thank Sydney Everhart and Brian Knaus for comments that significantly improved this manuscript. Mention of trade names or commercial products in this manuscript are solely for the purpose of providing specific information and do not imply recommendation or endorsement.

Chapter 3: Novel R Tools For Analysis of Genome-Wide Population Genetic Data With Emphasis on Clonality

Zhian N. Kamvar, Jonah C. Brooks, and Niklaus J. Grünwald

Journal: **Frontiers in Genetics**
EPFL Innovation Park, Building I, CH – 1015 Lausanne Switzerland
Published 2015-06-10. Issue: **6**, DOI: 10.3389/fgene.2015.00208

3.1 Abstract

To gain a detailed understanding of how plant microbes evolve and adapt to hosts, pesticides, and other factors, knowledge of the population dynamics and evolutionary history of populations is crucial. Plant pathogen populations are often clonal or partially clonal which requires different analytical tools. With the advent of high throughput sequencing technologies, obtaining genome-wide population genetic data has become easier than ever before. We previously contributed the R package *poppr* specifically addressing issues with analysis of clonal populations. In this paper we provide several significant extensions to *poppr* with a focus on large, genome-wide SNP data. Specifically, we provide several new functionalities including the new function `mlg.filter` to define clone boundaries allowing for inspection and definition of what is a clonal lineage, minimum spanning networks with reticulation, a sliding-window analysis of the index of association, modular bootstrapping of any genetic distance, and analyses across any level of hierarchies.

3.2 Introduction

To paraphrase Dobzhansky, nothing in the field of plant-microbe interactions makes sense except in the light of population genetics (Dobzhansky, 1973). Genetic forces such as selection and drift act on alleles in a population. Thus, a true understanding of how plant pathogens emerge, evolve and adapt to crops, fungicides, or other factors, can only be elucidated in the context of population level phenomena given the demographic history of populations (Grünwald and Goss, 2011; McDonald and Linde, 2002;

Milgroom et al., 1989). The field of population genetics, in the era of whole genome resequencing, provides unprecedented power to describe the evolutionary history and population processes that drive coevolution between pathogens and hosts. This powerful field thus critically enables effective deployment of R genes, design of pathogen informed plant resistance breeding programs, and implementation of fungicide rotations that minimize emergence of resistance.

Most computational tools for population genetics are based on concepts developed for sexual model organisms. Populations that reproduce clonally or are polyploid are thus difficult to characterize using classical population genetic tools because theoretical assumptions underlying the theory are violated. Yet, many plant pathogen populations are at least partially clonal if not completely clonal (Anderson and Kohn, 1995; Milgroom, 1996). Thus, development of tools for analysis of clonal or polyploid populations is needed.

Genotyping by sequencing and whole genome resequencing provide the unprecedented ability to identify thousands of single nucleotide polymorphisms (SNPs) in populations (Davey et al., 2011; Elshire et al., 2011; Luikart et al., 2003). With traditional marker data (e.g., SSR, AFLP) a clone was typically defined as a unique multilocus genotype (MLG) (Cooke et al., 2012; Falush et al., 2003; Goss et al., 2009; Grünwald and Hoheisel, 2006; Taylor and Fisher, 2003). Availability of large SNP data sets provides new challenges for data analysis. These data are based on reduced representation libraries and high throughput sequencing with moderate sequencing depth which invariably results in substantial missing data, error in SNP calling due to sequencing error, lack of read depth or other sources of spurious allele calls (Mastretta-Yanes et

al., 2014). It is thus not clear what a clone is in large SNP data sets and novel tools are required for definition of clone boundaries.

The research community using the R statistical and computing language (R Core Team, 2015) has developed a plethora of new resources for population genetic analysis. R is particularly appealing because all code is open source and functions can be evaluated and modified by any user. Recently, we introduced the R package *poppr* specifically developed for analysis of clonal populations (Kamvar et al., 2014b). *Poppr* previously introduced several novel features including the ability to conduct a hierarchical analysis across unlimited hierarchies, test for linkage association, graph minimum spanning networks or provide bootstrap support for Bruvo's distance in resulting trees. *Poppr* has been rapidly adopted and applied to a range of studies including for example horizontal transmission in leukemia of clams (Metzger et al., 2015), study of the vector-mediated parent-to-offspring transmission in an avian malaria-like parasite (Chakarov et al., 2015), and characterization of the emergence of the invasive forest pathogen *Hymenoscyphus pseudoalbidus* (Gross et al., 2014). It has also been used to implement real-time, online R based tools for visualizing relationships among unknown MLGs in reference databases
(http://phytophthora-id.org/) (Grünwald et al., 2011).

Here, we introduce *poppr* 2.0, which provides a major update to *poppr* (Kamvar et al., 2014b) including novel tools for analysis of clonal populations specifically addressing large SNP data. Significant novel tools include functions for calculating clone boundaries and collapsing individuals into clonal groups based on a user-specified genetic distance threshold, sliding window analyses, genotype accumulation curves, reticula-

tions in minimum spanning networks, and bootstrapping for any genetic distance.

3.3 Implementations and Examples

3.3.1 Clonal identification

As highlighted in previous work, clone correction is an important component of population genetic analysis of organisms that are known to reproduce asexually (Grünwald et al., 2003; Kamvar et al., 2014b; Milgroom, 1996). This method is a partial correction for bias that affects metrics that rely on allele frequencies assuming panmixia and was initially designed for data with only a handful of markers. With the advent of large-scale sequencing and reduced- representation libraries, it has become easier to sequence tens of thousands of markers from hundreds of individuals (Davey and Blaxter, 2010; Davey et al., 2011; Elshire et al., 2011). With this larger number of markers, the genetic resolution is much greater, but the chance of genotyping error is also greatly increased and missing data is frequent (Mastretta-Yanes et al., 2014). Taking this fact and occasional somatic mutations into account, it would be impossible to separate true clones from independent individuals by just comparing what MLGs are different. We introduce a new method for collapsing unique multilocus genotypes determined by naive string comparison into multilocus lineages utilizing any genetic distance given three different clustering algorithms: farthest neighbor, nearest neighbor, and UPGMA (average neighbor) (Sokal, 1958).

These clustering algorithms act on a distance matrix that is either provided by the

user or generated via a function that will calculate a distance from genetic data such as `bruvo.dist`, which in particular applies to any level of ploidy (Bruvo et al., 2004). All algorithms have been implemented in C and utilize the OpenMP framework for optional parallel processing (Dagum and Menon, 1998). Default is the conservative farthest neighbor algorithm (Fig. 3.1A), which will only cluster samples together if all samples in the cluster are at a distance less than the given threshold. By contrast, the nearest neighbor algorithm will have a chaining effect that will cluster samples akin to adding links on a chain where a sample can be included in a cluster if all of the samples have at least one connection below a given threshold (Fig. 3.1C). The UPGMA, or average neighbor clustering algorithm is the one most familiar to biologists as it is often used to generate ultra-metric trees based on genetic distance (Fig. 3.1B). This algorithm will cluster by creating a representative sample per cluster and joining clusters if these representative samples are closer than the given threshold.

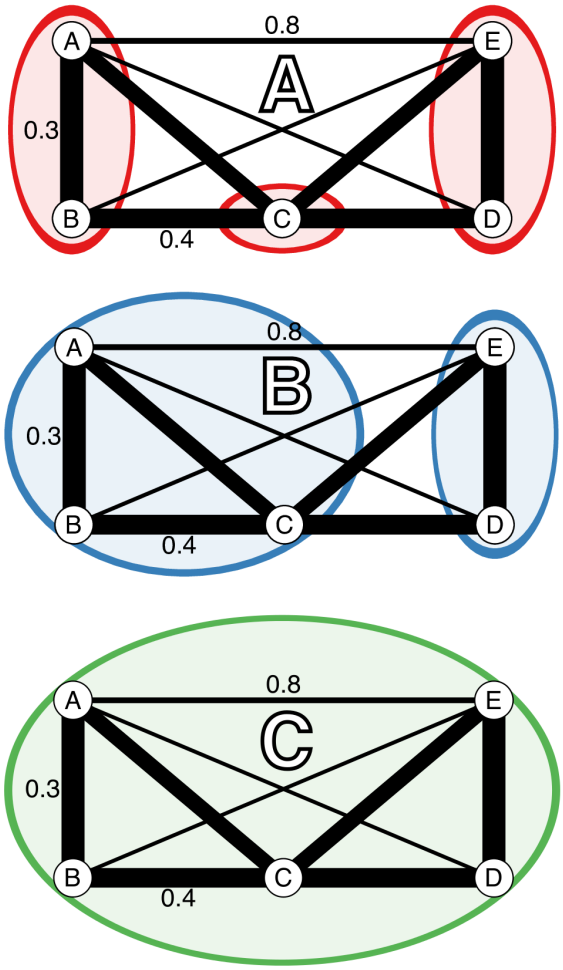


Figure 3.1: Diagrammatic representation of the three clustering algorithms implemented in `mlg.filter`. (**A**-**C**) Represent different clustering algorithms on the same imaginary network with a threshold of 0.451. Edge weights are represented in arbitrary units noted by the line thickness and numerical values next to the lines. All outer angles are 90 degrees, so the un-labeled edge weights can be obtained with simply geometry. Colored circles represent clusters of genotypes. (**A**) Farthest neighbor clustering does not cluster nodes B and C because nodes A and C are more than a distance of 0.451 apart. (**B**) UPGMA (average neighbor) clustering clusters nodes A, B, and C together because the average distance between them and C is < 0.451. (**C**) Nearest neighbor clustering clusters all nodes together because the minimum distance between them is always < 0.451.

We utilize data from the microbe *Phytophthora infestans* to show how the `mlg.filter` function collapses multilocus genotypes with Bruvo's distance assuming a genome addition model (Bruvo et al., 2004). *P. infestans* is the causal agent of potato late blight originating from Mexico that spread to Europe in the mid 19th century (Goss et al., 2014; Yoshida et al., 2013). *P. infestans* reproduces both clonally and sexually. The clonal lineages of *P. infestans* have been formally defined into 18 separate clonal lineages using a combination of various molecular methods including AFLP and microsatellite markers (Lees et al., 2006; Li et al., 2013). For these data, we used `mlg.filter` to detect all of the distance thresholds at which 18 multilocus lineages would be resolved. We used these thresholds to define multilocus lineages and create contingency tables and dendrograms to determine how well the multilocus lineages were detected.

For the *P. infestans* population, the three algorithms were able to detect 18 multilocus lineages at different distance thresholds (Fig. 3.2). Contingency tables between the described multilocus genotypes and the genotypes defined by distance show that most of the 18 lineages were resolved, except for US-8, which is polytomic (Table 3.1).

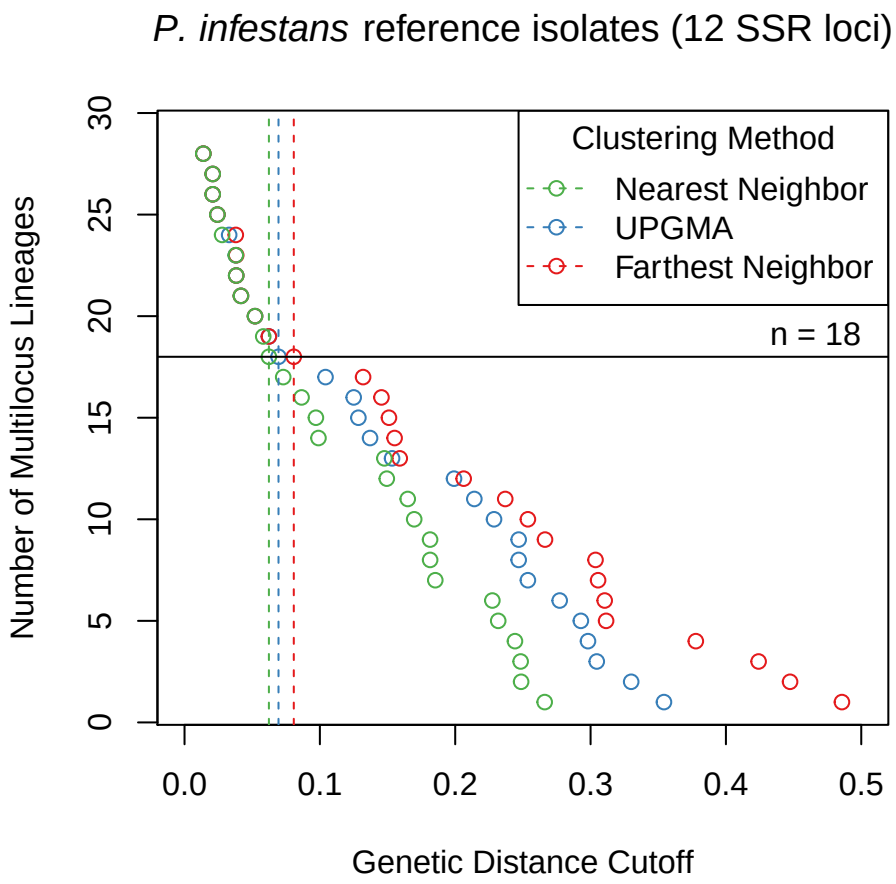


Figure 3.2: Graphical representation of three different clustering algorithms collapsing multilocus genotypes for 12 SSR loci from *Phytophthora infestans* representing 18 clonal lineages. The horizontal axis is Bruvo's genetic distance assuming the genome addition model. The vertical axis represents the number of multilocus lineages observed. Each point shows the threshold at which one would observe a given number of multilocus genotypes. The horizontal black line represents 18 multilocus genotypes and vertical dashed lines mark the thresholds used to collapse the multilocus genotypes into 18 multilocus lineages.

Table 3.1: Contingency table comparing multilocus lineages (MLL) defined in Li et al. (2013) and Lees et al. (2006) (rows) to MLLs inferred from Bruvo's genetic distance (columns) at a threshold of 0.07 with the average neighbor algorithm (Bruvo et al., 2004; Sokal, 1958). Values in the table represent the number of times any given inferred MLL matches with a previously defined MLL. For example, in our original data set, there were three genotypes previously defined as the US-24 MLL. All three genotypes were also determined to cluster into a single MLL by filtering. In contrast, US-8 was determined to cluster into three different MLLs by filtering.

	3	**4**	**5**	**6**	**8**	**10**	**12**	**15**	**16**	**17**	**18**	**20**	**21**	**22**	**24**	**25**	**27**	**28**
B	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.
C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.
D.1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
D.2	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.
EU-13	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.
EU-4	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.
EU-5	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.
EU-8	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
US-11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2
US-12	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
US-14	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
US-17	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.
US-20	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
US-21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.
US-22	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.
US-23	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.
US-24	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.
US-8	.	.	1	1	.	2	.	.	.	.	.	.	.	.	.	.	.	.

We utilized simulated data to evaluate the effect of sequencing error and missing data on MLG calling. We constructed the data using the `glSim` function in *adegenet* (Jombart and Ahmed, 2011) to obtain a SNP data set for demonstration. Two diploid data sets were created, each with 10k SNPs (25% structured into two groups) and 200 samples with 10 ancestral populations of even sizes. Clones were created in one data set by marking each sample with a unique identifier and then randomly sampling with replacement. It is well documented that reduced- representation sequencing can introduce several erroneous calls and missing data (Mastretta-Yanes et al., 2014). To reflect this, we mutated SNPs at a rate of 10% and inserted an average of 10% missing data for each sample after clones were created, ensuring that no two sequences were alike. The number of mutations and missing data per sample were determined by sampling from a Poisson distribution with $\lambda = 1000$. After pooling, 20% of the data set was randomly sampled for analysis. Genetic distance was obtained with the function `bitwise.dist`, which calculates the fraction of different sites between samples equivalent to Provesti's distance, counting missing data as equivalent in comparison (Prevosti et al., 1975).

All three filtering algorithms were run with a threshold of 1, returning a numeric vector of length $n - 1$ where each element represented a threshold at which two samples/clusters would join. Since each data set would have varying distances between samples, the clonal boundary threshold was defined as the midpoint of the largest gap between two thresholds that collapsed less than 50% of the data.

Out of the 100 simulations run, we found that across all methods, detection of duplicated samples had $\sim$ 98% true positive fraction and $\sim$ 0.8% false positive frac-

tion indicating that this method is robust to simulated populations (supplementary materials[1]).

3.3.2 Minimum Spanning Networks with Reticulation

In its original iteration, *poppr* introduced minimum spanning networks that were based on the *igraph* function `minimum.spanning.tree` (Csardi and Nepusz, 2006). This algorithm produces a minimum spanning tree with no reticulations where nodes represent individual MLGs. In other minimum spanning network programs, reticulation is obtained by calculating the minimum spanning tree several times and returning the set of all edges included in the trees. Due to the way *igraph* has implemented Prim's algorithm, it is not possible to utilize this strategy, thus we implemented an internal C function to walk the space of minimum spanning trees based on genetic distance to connect groups of nodes with edges of equal weight.

To demonstrate the utility of minimum spanning networks with reticulation, we used two clonal data sets: the H3N2 flu virus data from the *adegenet* package using years of each epidemic as the population factor, and *Phytophthora ramorum* data from Nurseries and Oregon forests (Jombart et al., 2010; Kamvar et al., 2014a). Minimum spanning networks were created with and without reticulation using the *poppr* functions `diss.dist` and `bruvo.msn` for the H3N2 and *P. ramorum* data, respectively (Bruvo et al., 2004; Kamvar et al., 2014b). To detect mlg clusters, the infoMAP community detection algorithm was applied with 10,000 trials as implemented in the R package

[1]Supplementary data available at https://github.com/grunwaldlab/supplementary-poppr-2.0; DOI: 10.5281/zenodo.17424

igraph version 0.7.1 utilizing genetic distance as edge weights and number of samples in each MLG as vertex weights (Csardi and Nepusz, 2006; Rosvall and Bergstrom, 2008).

To evaluate the results, we compared the number, size, and entropy (H) of the resulting communities as we expect a highly clonal organism with low genetic diversity to result in a few, large communities. We also created contingency tables of the community assignments with the defined populations and used those to calculate entropy using Shannon's index with the function `diversity` from the R package *vegan* version 2.2-1 (Oksanen et al., 2015; Shannon, 2001). A low entropy indicates presence of a few large communities whereas high entropy indicates presence of many small communities.

The infoMAP algorithm revealed 63 communities with a maximum community size of 77 and $H = 3.56$ for the reticulate network of the H3N2 data and 117 communities with a maximum community size of 26 and $H = 4.65$ for the minimum spanning tree. The entropy across years was greatly decreased for all populations with the reticulate network compared to the minimum spanning tree (Fig. 3.3). Note that the reticulated network (Fig. 3.3B) showed patterns corresponding with those resulting from a discriminant analysis of principal components (Fig. 3.3D) (Jombart et al., 2010).

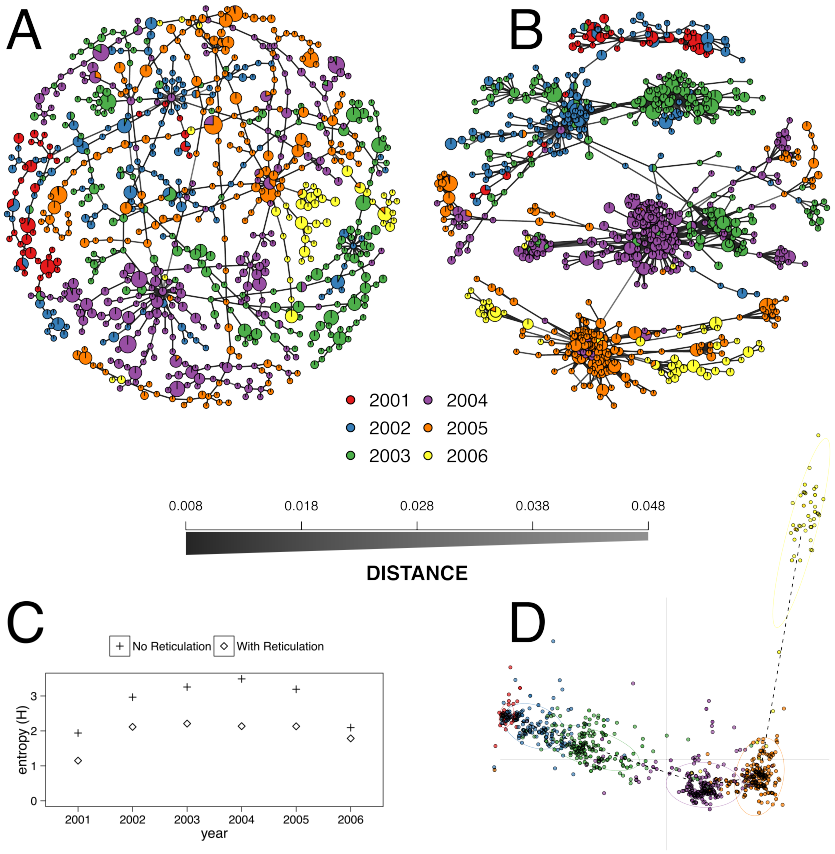


Figure 3.3: (**A**-**B**) Minimum spanning networks of the hemagglutinin (HA) segment of H3N2 viral DNA from the *adegenet* package representing flu epidemics from 2001 to 2006 without reticulation (**A**) and with reticulation (**B**) (Jombart, 2008; Jombart et al., 2010). Each node represents a unique multilocus genotype, colors represent epidemic year, and edge color represents absolute genetic distance. (**C**) Shannon entropy values for population assignments compared with communities determined by the infoMAP algorithm on (**A**) and (**B**). (**D**) Graphic reproduced from Jombart et al. (2010) showing that the 2006 epidemic does not cluster neatly with the other years via Discriminant Analysis of Principal Components. Horizontal axis represents the first discriminant component. Vertical axis represents the second discriminant component.

Graph walking of the reticulated minimum spanning network of *P. ramorum* by the infoMAP algorithm revealed 16 communities with a maximum community size of 13 and $H = 2.60$. The un-reticulated minimum spanning tree revealed 20 communities with a maximum community size of 7 and $H = 2.96$. In the ability to predict Hunter Creek as belonging to a single community, the reticulated network was successful whereas the minimum spanning tree separated one genotype from that community. The entropy for the reticulated network was lower for all populations except for the coast population (supplementary materials[2]).

3.3.3 Bootstrapping

Assessing population differentiation through methods such as G_{st}, AMOVA, and Mantel tests relies on comparing samples within and across populations (Excoffier et al., 1992; Mantel, 1967; Nei, 1973). Confidence in distance metrics is related to the confidence in the markers to accurately represent the diversity of the data. Especially true with microsatellite markers, a single hyper-diverse locus can make a population appear to have more diversity based on genetic distance. Using a bootstrapping procedure of randomly sampling loci with replacement when calculating a distance matrix provides support for clades in hierarchical clustering.

Data in genind and genpop objects are represented as matrices with individuals in rows and alleles in columns (Jombart, 2008). This gives the advantage of being able to use R's matrix algebra capabilities to efficiently calculate genetic distance.

[2]Supplementary data available at https://github.com/grunwaldlab/supplementary-poppr-2.0; DOI: 10.5281/zenodo.17424

Unfortunately, this also means that bootstrapping is a non- trivial task as all alleles at a single locus need to be sampled together. To remedy this, we have created an internal S4 class called "bootgen", which extends the internal "gen" class from *adegenet*. This class can be created from any genind, genclone, or genpop object, and allows loci to be sampled with replacement. To further facilitate bootstrapping, a function called `aboot`, which stands for "any boot", is introduced that will bootstrap any genclone, genind, or genpop object with any genetic distance that can be calculated from it.

To demonstrate calculating a dendrogram with bootstrap support, we used the *poppr* function `aboot` on population allelic frequencies derived from the data set `microbov` in the *adegenet* package with 1000 bootstrap replicates (Jombart, 2008; Laloë et al., 2007). The resulting dendrogram shows bootstrap support values $> 50\%$ (Fig. 3.4) and used the following code:

```
library("poppr");
data("microbov", package = "adegenet");
strata(microbov) <- data.frame(other(microbov));
setPop(microbov) <- ~coun/spe/breed;
bov_pop <- genind2genpop(microbov);


set.seed(20150428);
pop_tree <- aboot(bov_pop, sample = 1000, cutoff = 50);
```

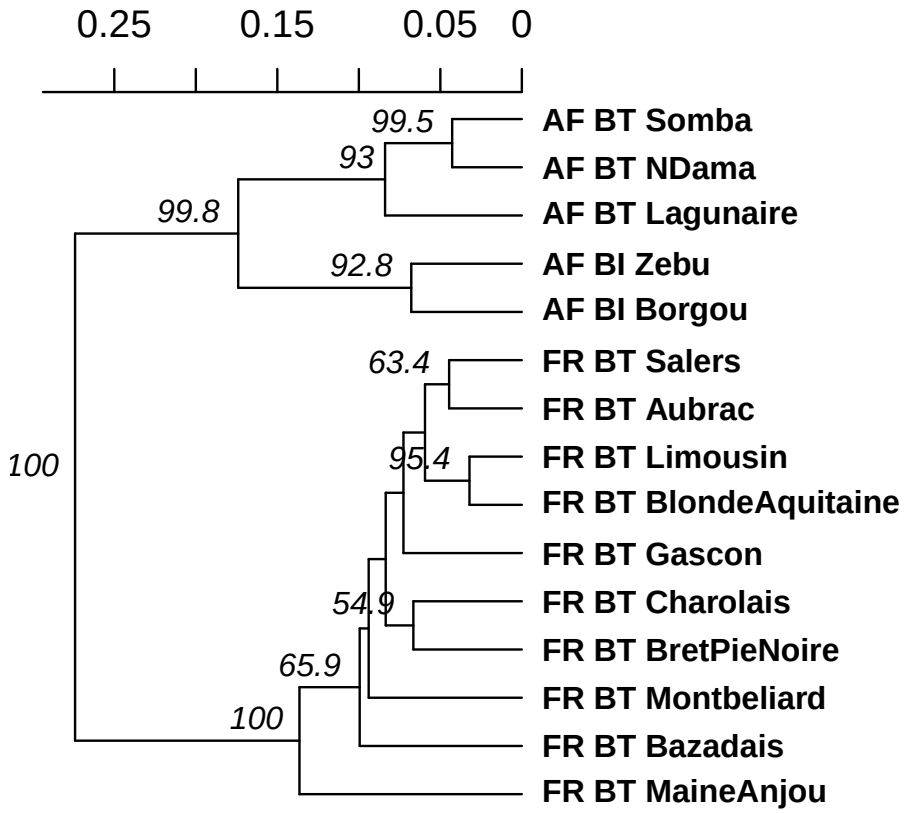


Figure 3.4: UPGMA dendrogram generated from Nei's genetic distance on 15 breeds of *Bos taurus* (BT) or *Bos indicus* (BI) from Africa (AF) or France (FR). These data are from Laloë et al. (2007). Node labels represent bootstrap support $> 50\%$ out of 1,000 bootstrap replicates.

3.3.4 Genotype Accumulation Curve

Analysis of population genetics of clonal organisms often borrows from ecological methods such as analysis of diversity within populations (Arnaud-Hanod et al., 2007; Grünwald et al., 2003; Milgroom, 1996). When choosing markers for analysis, it is important to make sure that the observed diversity in your sample will not appreciably increase if

an additional marker is added (Arnaud-Hanod et al., 2007). This concept is analogous to a species accumulation curve, obtained by rarefaction. The genotype accumulation curve in *poppr* is implemented in the function `genotype_curve`. The curve is constructed by randomly sampling x loci and counting the number of observed MLGs. This repeated r times for 1 locus up to $n - 1$ loci, creating $n - 1$ distributions of observed MLGs.

The following code example demonstrates the genotype accumulation curve for data from Everhart and Scherm (2015) showing that these data reach a small plateau and have a greatly decreased variance with 12 markers, indicating that there are enough markers such that adding more markers to the analysis will not create very many new genotypes (Fig. 3.5).

```
library("poppr");
library("ggplot2");
data("monpop", package = "poppr");


set.seed(20150428);
genotype_curve(monpop, sample = 1000);
p <- last_plot() + theme_bw();    # get the last plot
p + geom_smooth(aes(group = 1)); # plot with a trendline
```

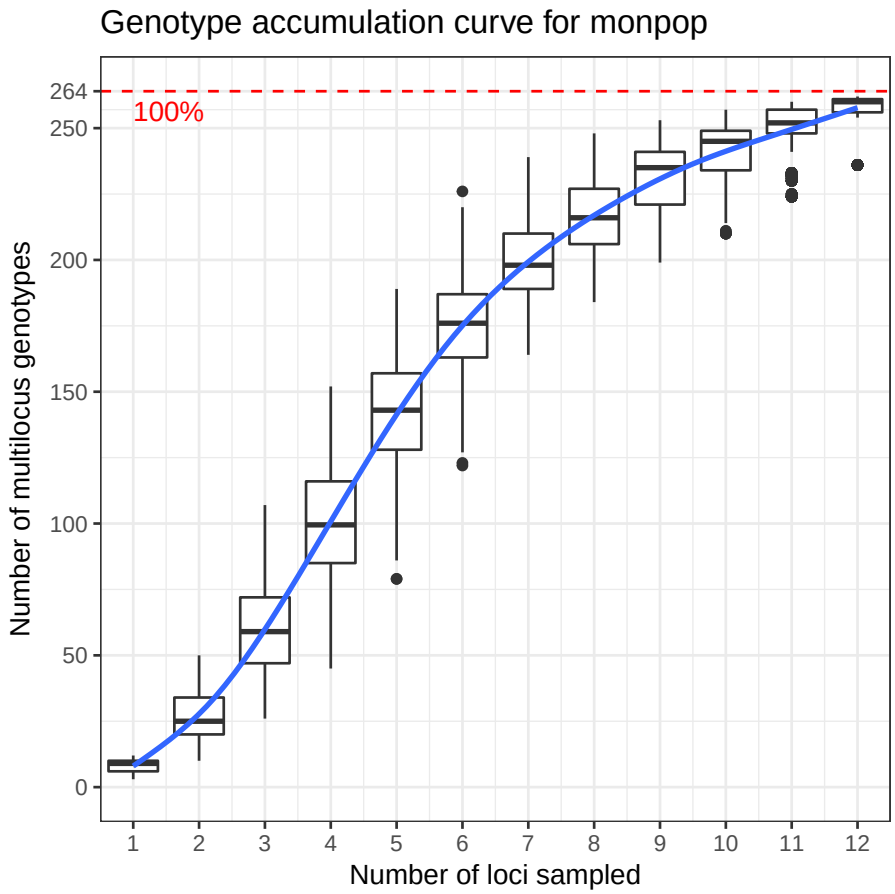


Figure 3.5: Genotype accumulation curve for 694 isolates of the peach brown rot pathogen, *Monilinia fructicola* genotyped over 13 loci from Everhart and Scherm (2015). The horizontal axis represents the number of loci randomly sampled without replacement up to *n - 1* loci, the vertical axis shows the number of multilocus genotypes observed, up to 262, the number of unique multilocus genotypes in the data set. The red dashed line represents 90% of the total observed multilocus genotypes. A trendline (blue) has been added using the *ggplot2* function `stat_smooth`.

3.3.5 Index of association

The index of association (I_A) is a measure of multilocus linkage disequilibrium that is most often used to detect clonal reproduction within organisms that have the ability to reproduce via sexual or asexual processes (Brown et al., 1980; Milgroom, 1996; Smith et al., 1993). It was standardized in 2001 as $\bar{r}_d$ by Agapow and Burt (2001) to address the issue of scaling with increasing number of loci. This metric is typically applied to traditional dominant and co-dominant markers such as AFLPs, SNPs, or microsatellite markers. With the advent of high throughput sequencing, SNP data is now available in a genome-wide context and in very large matrices including thousands of SNPs. For this reason, we devised two approaches using the index of association for large numbers of markers typical for population genomic studies. Both functions utilize *adegenet*'s "genlight" object class, which efficiently stores 8 binary alleles in a single byte (Jombart and Ahmed, 2011). As calculation of the $\bar{r}_d$ requires distance matrices of absolute number of differences, we utilize a function that calculates these distances directly from the compressed data called `bitwise.dist`.

The first approach is a sliding window analysis implemented in the function `win.ia`. It utilizes the position of markers in the genome to calculate $\bar{r}_d$ among any number of SNPs found within a user-specified windowed region. It is important that this calculation utilize $\bar{r}_d$ as the number of loci will be different within each window (Agapow and Burt, 2001). This approach would be suited for a quick calculation of linkage disequilibrium across the genome that can detect potential hotspots of LD that could be investigated further with more computationally intensive methods assuming that

the number of samples $<<$ the number of loci.

As it would necessarily focus on loci within a short section of the genome that may or may not be recombining, a sliding window approach would not be good for utilizing $\bar{r}_d$ as a test for clonal reproduction. A remedy for this is implemented in the function `samp.ia`, which will randomly sample m loci, calculate $\bar{r}_d$, and repeat r times, thus creating a distribution of expected values of $\bar{r}_d$.

To demonstrate the sliding window and random sampling of $\bar{r}_d$ with respect to clonal populations, we simulated two populations containing 1,100 neutral SNPs for 100 diploid individuals under the same initial seed. One population had individuals randomly sampled with replacement, representing the clonal population. After sampling, both populations had 5% random error and 1% missing data independently propagated across all samples. On average, we obtained a higher value of $\bar{r}_d$ for the clonal population compared to the sexual population for both methods (Fig. 3.6).

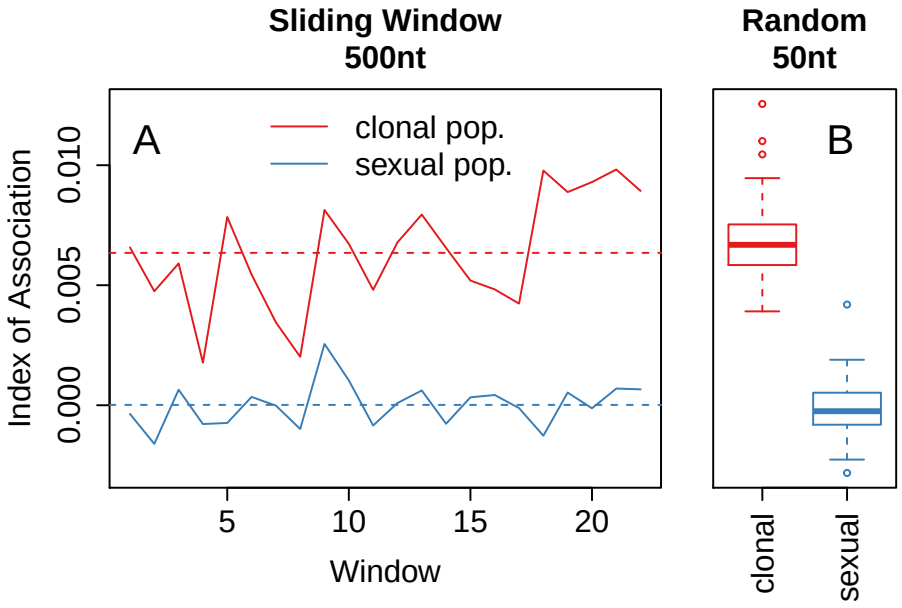


Figure 3.6: **(A)** Sliding window analysis of the standardized index of association ($\bar{r}_d$) across a simulated 1.1×10^4 nt chromosome containing 1,100 variants among 100 individuals. Each window analyzed variants within 500nt chunks. The black line refers to the clonal and the blue line to the sexual populations. **(B)** boxplots showing 100 random samples of 50 variants to calculate a distribution of $\bar{r}_d$ for the clonal (red) and sexual (blue) populations. Each box is centered around the mean, with whiskers extending out to 1.5 times the interquartile range. The median is indicated by the center line. **(A)** and **(B)** are plotted on the same y-axis.

3.3.6 Data format updates: population strata and hierarchies

Assessments of population structure through methods such as hierarchical F_{st} (Goudet, 2005) and AMOVA (Michalakis and Excoffier, 1996) require hierarchical sampling of populations across space or time (Everhart and Scherm, 2015; Grünwald and Hoheisel, 2006; Linde et al., 2002). With clonal organisms, basic practice has been to clone-

censor data to avoid downward bias in diversity due to duplicated genotypes that may or may not represent different samples (Milgroom, 1996). This correction should be performed with respect to a population hierarchy to accurately reflect the biology of the organism. Traditional data structures for population genetic data in most analysis tools allow for only one level of hierarchical definition. The investigator thus had to provide the data set for analysis at each hierarchical level.

To facilitate handling hierarchical and mutlilocus genotypic metadata, *poppr* version 1.1 introduced a new S4 data object called "genclone", extending *adegenet*'s "genind" object (Kamvar and Grünwald, unpublished). The genclone object formalized the definitions of multilocus genotypes and population hierarchies by adding two slots called "mlg" and "hierarchy" that carried a numeric vector and a data frame, respectively. These new slots allow for increased efficiency and ease of use by allowing these metadata to travel with the genetic data. The hierarchy slot in particular contains a data frame where each column represents a separate hierarchical level. This is then used to set the population factor of the data by supplying a hierarchical formula containing one or more column names of the data frame in the hierarchy slot.

The functionality represented by the hierarchy slot has now been migrated from the *poppr* to the *adegenet* package version 2.0 to allow hierarchical analysis in *adegenet*, *poppr*, and other dependent packages. The prior *poppr* `hierarchy` slot and methods have now been renamed `strata` in *adegenet*. A short example of the utility of these methods can be seen in the code segment under **Bootstrapping**, above. This migration provides end users with a broader ability to analyze data hierarchically in R across packages.

3.4 Availability

As of this writing, the *poppr* R package version 2.0 containing all of the features described here is located at https://github.com/grunwaldlab/poppr/tree/2.0-rc. It is necessary to install *adegenet* 2.0 before installing *poppr*. It can be found at https://github.com/thibautjombart/adegenet. Both of these can be installed via the R package *devtools* (Wickham and Chang, 2015). More information and example code can be found in the supplementary materials[3].

3.4.1 Requirements

- R version 3.0 or better
- A C compiler. For windows, it can be obtained via Rtools (http://cran.r-project.org/bin/windows/Rtools/). On OSX, it can be obtained via Xcode. For parallel support, gcc version 4.6 or better is needed.

[3]Supplementary data available at https://github.com/grunwaldlab/supplementary-poppr-2.0; DOI: 10.5281/zenodo.17424

3.4.2 Installation

From within R, *poppr* can be installed via:

```
install.packages("devtools")
library("devtools")
install_github("thibautjombart/adegenet")
install_github("grunwaldlab/poppr@2.0-rc")
```

Several population genetics packages in R are currently going through a major upgrade following the 2015 R hackathon on population genetics (https://github.com/NESCent/r-popgen-hackathon) and have not yet been updated in CRAN. We will upload *poppr* 2.0 to CRAN once all other reverse dependent packages have been updated.

3.5 Discussion

Given low cost and high throughput of current sequencing technologies we are entering a new era of population genetics where large SNP data sets with thousands of markers are becoming available for large populations in a genome- wide context. This data provides new possibilities and challenges for population genetic analyses. We provide novel tools that enable analysis of this data in R with a particular emphasis on clonal organisms.

Particularly useful is the implementation of $\bar{r}_d$ in a genomic context (Agapow and Burt, 2001). Random sampling of loci across the genome can give an expected distri-

bution of $\bar{r}_d$, which is expected to have a mean of zero for panmictic populations. This metric is not affected by the number of loci sampled, is model free, and has the ability to detect population structure. $\bar{r}_d$ is also implemented for sliding window analyses that are useful to detect candidate regions of linkage disequilibrium for further analysis.

Clustering multilocus genotypes into multilocus lineages based on genetic distances is a non-trivial task given large SNP data sets. Moreover, this has not previously been implemented for genomic data for clonal populations. Clonal assignment has previously been available in the programs GenClone and Genodive for classical markers (Arnaud-Hanod et al., 2007; Meirmans and Van Tienderen, 2004). Our method with `mlg.filter` builds upon this idea and allows the user to choose between three different approaches for clustering MLGs. The choice of clustering algorithm has an impact on the data (Fig. 3.1, 3.2), where for example a genetic distance cutoff of 0.1 would be the difference between 14 multilocus lineages (MLLs) and 17 MLLs for nearest neighbor and UPGMA clustering, respectively (Fig. 3.2). The option to choose the clustering algorithm gives the user the ability to choose what is biologically relevant to their populations. While there is not one optimal procedure for defining boundaries in clonal lineages, our tool provides a means of exploring the potential MLG or MLL boundary space.

Minimum spanning networks are a useful tool to analyze the relationships between individuals in a population, because it reduces the complexity of a distance matrix to the connections that are strongest. By default, these networks are drawn without reticulations, but for clonal organisms where many of the connections between samples are equivalent, the minimum spanning network appears as a chain and reduces the

information that can be communicated. This is problematic because the ability to detect population structure with one instance of a minimum spanning network is limited. Adding reticulation into the minimum spanning network thus presents all equivalent connections and allows population structure to be more readily detectable. As shown in Fig. 3.3, population structure is apparent both visually and by graph community detection algorithms such as the infoMAP algorithm (Rosvall and Bergstrom, 2008). Additionally, the current implementation in *poppr* has been successfully used in analyses such as reconstruction of the *P. ramorum* epidemic in Oregon forests (Kamvar et al., 2014a, 2015c).

Poppr 2.0 is open source and available on GitHub. Members of the community are invited to contribute by raising issues or pull requests on our repository at https://github.com/grunwaldlab/poppr/issues.

3.6 Acknowledgements

We thank Ignazio Carbone for discussions on the index of association; David Cooke, Sanmohan Baby, and Jens Hansen for beta testing; and Thibaut Jombart for allowing us to incorporate the `strata` slot and related methods in *adegenet*. We also thank all the members of the 2015 R hackathon on population genetics in Durham, NC for their advice and input (https://github.com/NESCent/r-popgen-hackathon). This work was supported in part by US Department of Agriculture (USDA) Agricultural Research Service Grant 5358-22000-039-00D, USDA National Institute of Food and Agriculture Grant 2011-68004-30154, USDA APHIS, the USDA-ARS Floriculture Nursery Initiative, and the USDA-Forest Service Forest Health Monitoring Program (to NJG).

Chapter 4: Spatial and Temporal Analysis of Populations of the Sudden Oak Death Pathogen in Oregon Forests

Zhian N. Kamvar, Meredeth M. Larsen, Alan M. Kanaskie, Everett M. Hansen, and Niklaus J. Grünwald

Journal: **Phytopathology**
3340 Pilot Knob Rd, St Paul, MN 55121, USA
Published 2015-07, Volume 105, Issue: **7**, DOI: 10.1094/PHYTO-12-14-0350-FI

4.1 Abstract

Sudden oak death caused by the oomycete *Phytophthora ramorum* was first discovered in California toward the end of the 20th century and subsequently emerged on tanoak forests in Oregon before its first detection in 2001 by aerial surveys. The Oregon Department of Forestry has since monitored the epidemic and sampled symptomatic tanoak trees from 2001 to the present. Populations sampled over this period were genotyped using microsatellites and studied to infer the population genetic history. To date, only the NA1 clonal lineage is established in this region, although three lineages exist on the North American west coast. The original introduction into the Joe Hall area eventually spread to several regions: mostly north but also east and southwest. A new introduction into Hunter Creek appears to correspond to a second introduction not clustering with the early introduction. Our data are best explained by both introductions originating from nursery populations in California or Oregon and resulting from two distinct introduction events. Continued vigilance and eradication of nursery populations of *P. ramorum* are important to avoid further emergence and potential introduction of other clonal lineages.

4.2 Introduction

Sudden oak death (SOD) emerged as a severe epidemic disease on coast live oak (*Quercus agrifolia*) and tanoak (*Notholithocarpus densiflorus*) in California in the mid 1990s and reemerged shortly thereafter on tanoak in Oregon in the early 2000s (Everhart et al., 2014; Grünwald et al., 2008a; Hansen et al., 2008; Rizzo et al., 2005).

SOD is caused by *Phytophthora ramorum* Werres, De Cock & Man in't Veld, and is considered to be one of the top two oomycete pathogens based on its scientific and economic importance (Kamoun et al., 2014; Werres et al., 2001). The Oregon epidemic was first detected during aerial surveys in 2001 on tanoak but likely derived from initial introductions in the late 1990s. The Oregon Department of Forestry has since monitored the epidemic and sampled symptomatic tanoaks since 2001 (Hansen et al., 2008). Strains sampled from infected sites in forest or nursery environments have been genotyped in several labs using a range of microsatellite loci (Grünwald et al., 2009; Ivors et al., 2006; Prospero et al., 2004, 2009, 2007).

P. ramorum has emerged repeatedly around the world as 4 distinct clonal lineages found in North America (lineages NA1, NA2, and EU1) and Europe (EU1 and EU2) (Grünwald et al., 2012; Ivors et al., 2006; Poucke et al., 2012). The lineages have been named by the continent on which they first appeared, i.e. North America (= NA) or Europe (= EU) and are numbered in order of discovery (Grünwald et al., 2009). The NA1 clonal lineage was first discovered in California causing SOD on tanoak and coast live oak and is the one currently found in Curry County, Oregon, USA (Mascheretti et al., 2008). The EU1 and NA2 populations were discovered later in nursery environments and are currently only found in California, Oregon, Washington and British Columbia while the NA1 clone has been shipped with nursery plants from the West to the Southern and Southwestern US (Goss et al., 2009, 2011; Grünwald et al., 2012; Ivors et al., 2006; Mascheretti et al., 2008; Prospero et al., 2009). The EU1 clonal lineage is the one first discovered in Europe, but in 2007 the new EU2 lineage emerged in Northern Ireland and since migrated to Western Scotland (Poucke

et al., 2012; Werres et al., 2001). EU1 was first introduced to Europe and eventually migrated to the Pacific Northwest of North America (Goss et al., 2011).

P. ramorum populations sampled in Oregon forests to date belong exclusively to the NA1 clonal lineage (Hansen et al., 2008; Prospero et al., 2007). Given that NA2 and/or EU1 clones have been found in California, Oregon, Washington, and/or British Columbia in association with nursery plant movements, introduction of NA2 or EU1 from nursery environments to Curry County forests is a plausible scenario (Goss et al., 2009, 2011; Grünwald et al., 2012; Prospero et al., 2009, 2007). Our present work thus monitors populations and potential emergence of novel lineages in Oregon forests.

Our main objectives here are to describe the spatial and temporal pattern of the populations and clonal dynamic of the SOD pathogen in Curry County in southwestern Oregon from 2001 to the present. Specifically, we asked (1) if novel lineages have been introduced into the forests in Curry County, (2) if multiple introductions occurred, and (3) whether introduction might have come from nursery populations. We sampled infected tanoaks between 2001-2014 and characterized populations using microsatellite analysis.

4.3 Materials and Methods

4.3.1 Location

The SOD infested areas are located in the Siskiyou Mountains of Curry County in south western Oregon near the town of Brookings (42.0575° N, 124.2864° W) on

the coast (Figure 4.1) (Prospero et al., 2007). The Siskiyou mountains form part of the Klamath Mountain range (Franklin and Dyrness 1988). The vegetation in SOD infested areas is a mosaic of different vegetation types including mixed-evergreen, redwood (*Sequoia sempervirens*) and Douglas-fir (*Pseudotsuga menziesii*) forests with tanoak as the dominant SOD host.

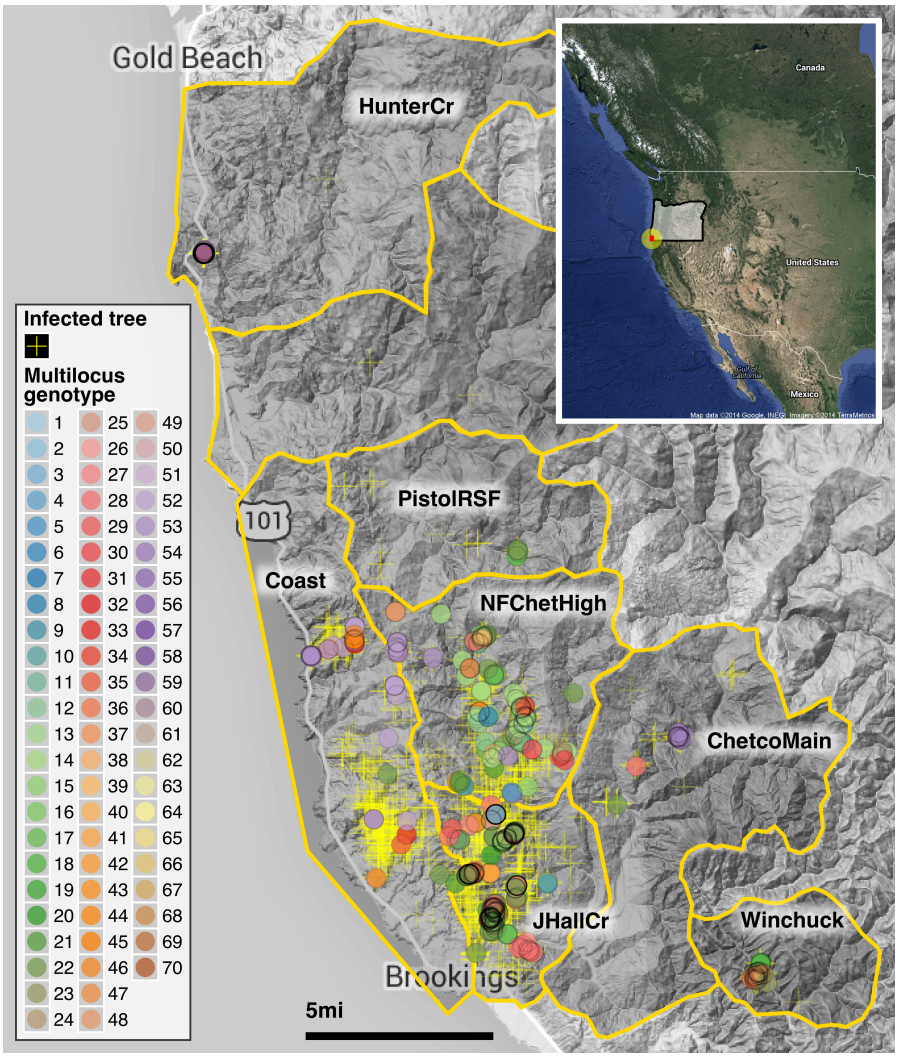


Figure 4.1: Spatial distribution of the SOD epidemic and multilocus genotypes of *Phytophthora ramorum* in Curry County, Oregon. The yellow crosses mark tanoak trees found positive for *P. ramorum* during aerial surveys. A total of 70 multilocus genotypes have been identified between 2001-2014 and are marked by color as shown in the legend. The abbreviations for regions shown in the map are explained in Table 4.1. The inset shows the placement of Curry county (red dot) in SW Oregon.

4.3.2 Sampling

Commencing in 2001, 2-4 aerial surveys per year were conducted over the tanoak range by the Oregon Department of Forestry and the USDA Forest Service in Curry County. The survey detects recently killed tanoaks based on the reddish-brown color of foliage (Hansen et al., 2008). All trees identified by aerial surveys were ground checked and geographically referenced using a hand-held GPS instrument (Garmin GPS 12XL or 60CX, Garmin International, Olathe, KS). Bark or foliage samples were collected for determination of *P. ramorum* presence by culturing in the field and laboratory. Host plants within the area of the delimitation survey, generally 300 feet, were also inspected and sampled if they were symptomatic. Maps of distribution were prepared using ArcView GIS version 3.3 and ArcMap version 10.2 (Environmental Systems Research Institute, Redlands, CA).

4.3.3 Isolation, identification and DNA extraction

Isolations were made from symptomatic plant tissue onto selective CARP agar (Difco corn meal agar, 10 ppm natamycin, 200pm NA-ampicillin, and 10 ppm rifampicin) (Prospero et al., 2007). Candidate *Phytophthora* cultures were transferred onto corn meal agar with 30 ppm β-sitosterol. *P. ramorum* identification was confirmed by microscopic inspection for presence of characteristic chlamydospores and deciduous sporangia (Werres et al., 2001). Genomic DNA was extracted using either the FastDNA SPIN kit (MP Biomedicals, LLC; 116540600) (Goss et al., 2009) or the cetyltrimethyl ammonium bromide (CTAB)-chloroform-isopropanol method (Winton and Hansen,

Table 4.1: Summary of *P. ramorum* isolates sampled in Oregon forests and multilocus genotypes (MLG)observed across regions and years.

Abbreviation	**Region name**	**Year**	**Number of isolates**	**MLGs detected (region specific)**
JHallCr	Joe Hall Creek	2001, 2002, 2003, 2004, 2005, 2013, 2014	244	30 (19)
NFChetHigh	North Fork Chetco	2003, 2012, 2013, 2014	114	35 (19)
Coast	Coastal Region	2006, 2010, 2011, 2012, 2013, 2014	34	12 (7)
HunterCr	Hunter Creek; Cape Sebastian	2011	66	4 (4)
Winchuck	...	2012, 2013	35	9 (3)
ChetcoMain	...	2013, 2014	16	7 (1)
PistolRSF	Pistol River South Fork	2013	4	2 (0)
Total	-	**2001-2014**	**513**	**70 (53)**

2001). Table 4.1 provides an overview of strains collected by year and region following regions as shown in figure 4.1.

4.3.4 Genotyping, data validation, and harmonization

Five microsatellite loci were utilized in this analysis: PrMS6, Pr9C3, PrMS39, PrMS45, and PrMS43 (Grünwald et al., 2009, 2008b; Prospero et al., 2004, 2007). Genotyping (see specific protocols in supplementary text) of *P. ramorum* strains collected 2001-2012 occurred over several years and in several laboratories, with different protocols and sequencers. Consequently, a concerted effort was made to create a comprehensive dataset with identical allele calls. To detect errors, allele calls from all five genotyped loci were generated for a subsample of 40 isolates representing the most common multilocus genotypes from the culture collection, and then compared to data from participating laboratories. Three of the five loci, PrMS6, Pr9C3, and PrMS39 had identical allele calls between laboratories for the subsampled isolates. The remaining two loci, PrMS45 and PrMS43, had allele calls that differed by a single bp between laboratories. Data from PrMS45 and PrMS43 were therefore corrected to allow consistent comparisons of allele calls. Given the varied nature of the genotyping data described above genotyping of *P. ramorum* strains consists of two datasets including either 5 loci (2001-14) or a newly developed, multiplexed method including 14 loci (samples 2013-14) (Table 4.2). Details on both genotyping methods can be found in the supplementary text 4.7.1 and figure 4.S1.

Table 4.2: (Caption on Next Page)

SSR Locus	**Dye**	**Product (bp)**[c]	**Primer sequence**[b]	**Final conc. (μM)**	**Rxn**
ILVOPrMS145abc[f]	6-FAM	167-257	Fwd6FAM-TGGCAGTGTTCTTCAACAGC Rev-*GTTT*ATTCCCGTGAACAGCGTATC	0.04	8-plex
PrMS39[d]	NED	130-258	FwdNED-GCACGGCCAGAGATTGATAG Rev-*GTTT*ATCTGCCGACGTGAAGAAGT	0.07	8-plex
PrMS9C3[d]	PET	210-226	FwdVIC-TCACACGAAGCAGCAACTCT Rev-*GTTT*AGCGGCACTACGGAATACAT	0.04	8-plex
ILVOPrMS79[af]	6-FAM	342-396	Fwd6FAM-AGGCGGAAAACGTCAGAAC Rev-*GTTT*CTCGAGAGGCTGGAAGTACG	0.15	8-plex
KI18[e]	VIC	217-279	FwdPET-TGCCATCACAACACAAATCC Rev-*GTTT*TGTGCTATCTTTCCTGAACGG	1.0	8-plex
KI64[e]	NED	342-401	FwdNED-GCGCTAAGAAAGACACTCCG Rev-*GTTT*CAACATGTAGCCATTGCAGG	0.35	8-plex
PrMS45[d]	VIC	138-186	FwdVIC-CGTGCTGCATCTGGTGTAGT Rev-GAAAGTCCGGATTTGCGTTA	0.15	8-plex
PrMS6[d]	PET	165-168	FwdPET-AATCGATCTCTCGGCTTTGA Rev-TATAGCCCCAGCTGCAACA	0.15	8-plex
ILVOPrMS131[f]	VIC	146-414	FwdVIC-CGGCCGTTTTTGTAAGTTTG Rev-*GTTT*CAGATCAAACCAAAATCTGCTC	0.2	2-plex
KI82ab[e]	NED	95-243	FwdNED-CCACGTCATTGGGTGACTTC Rev-*GTTT*CGTACAAGTCACGACTCCCC	0.2	2-plex
PrMS43[d]	6-FAM	122-493	Fwd6FAM-AAATATGCAAAAAGGCAGGA Rev-*GTTT*CCGCGTAACCTAGTCTGCTC	0.3	Simplex

Table Caption 4.C2: (Caption for Table 4.2) Newly multiplexed protocol for *P. ramorum* primer sequences of simple sequence repeat (SSR) loci and final concentrations used to determine multilocus genotypes for four clonal lineages. PrMS6, Pr9C3, PrMS39, PrMS45, and PrMS43 were utilized in this study as they were commonly genotyped across all laboratories. [a] ILVOPrMS79 amplifies three alleles in the NA1 lineage. The first two alleles are fixed and the third is polymorphic. [b] Reverse (Rev) primer includes PIG tail addition except for PrMS45 and PrMS6. Indicated in italic. [c] Product size range is for four lineages (EU1, EU2, NA1, NA2). Only the NA1 lineage has been reported in Curry County,OR forests. [d] Described by Prospero et al. (2004) and/or Prospero et al. (2007). [e] Described by Ivors et al. (2006). [f] Described by Vercauteren et al. (2010), and Vercauteren et al. (2011).

4.3.5 Nursery populations

To determine if forest populations cluster with different nursery populations from Oregon or California, we used previously published data from our work to determine relationships among nursery and Curry County forest populations (Goss et al., 2009, 2011; Grünwald et al., 2009; Prospero et al., 2009, 2007).

4.3.6 Data analysis

All individuals genotyped for this effort belonged to the NA1 clonal lineage (Grünwald et al., 2009). Thus, all analyses presented here focused on describing the clonal dynamic using model-free approaches that avoid violation of population genetic theory. Samples were grouped into different multilocus genotypes (MLGs) defined by the unique combination of alleles across all observed loci from the consensus five SSR loci genotyped across all years. For identification purposes, unique MLGs were then assigned an arbitrary number from 1 to the total number of observed MLGs. Population genetic analysis was conducted using the computer and statistical language R (R Core Team, 2014) using various packages as well as R functions written specifically for this project (see github link below). Graphs and figures were created using the R packages *ggplot2, ape, igraph, ggmap,* and *poppr* (Csardi and Nepusz, 2006; Kahle and Wickham, 2013; Kamvar et al., 2014b; Paradis et al., 2004; Wickham, 2009). Within-locus allelic diversity was analyzed across and within years and regions using the function `locus_table()` from the R package *poppr* (Table **??**) (Kamvar et al., 2014b). To address the temporal and spatial aspects of the data, populations were analyzed both

by year isolated and watershed region (Table 4.1; Fig. 4.1). Watershed regions were drawn with ArcMap version 10.2 (Environmental Systems Research Institute, Redlands, CA). The regions represent drainages or portions of drainages in which infected trees were discovered as the disease progressed over time. In most cases, ridgelines dividing drainages formed the boundary of a region. These regions were saved as shapefiles and imported into R with *rgdal* (Bivand et al., 2014).

Genotypic diversity was analyzed within and across years and populations, with the Shannon-Wiener index (H) and the Stoddard and Taylor's index (G), (Shannon, 2001; Stoddart and Taylor, 1988). Both G and H measure genotypic diversity, combining richness and evenness. If all genotypes are equally abundant, then the value of G will be the number of MLGs and the value of H will be the natural log of the number of MLGs. Both G and H are used as they weigh more or less abundant MLGs more heavily, respectively (Grünwald et al., 2003). Evenness was calculated as E_5, which is an estimator of evenness that utilizes both H and G that gives a ratio of the number of abundant genotypes to rare genotypes (Grünwald et al., 2003; Ludwig and Reynolds, 1988; Pielou, 1975). These were calculated with the R packages *poppr* and *vegan* (Kamvar et al., 2014b; Oksanen et al., 2013). Confidence intervals were calculated using the R package *boot* with 9,999 bootstrap resamplings (Canty and Ripley, 2015). Richness, or the expected number of MLGs ($eMLG$), was calculated using rarefaction from the R packages *poppr* and *vegan* (Heck et al., 1975; Hurlbert, 1971). Some statistics (AMOVA, genotypic diversity, index of association, allelic diversity, and Nei's distance) were also performed on clone-censored data where each MLG was represented once per population hierarchy.

Because the analysis of genotypic diversity, richness and evenness is agnostic to specific alleles within MLGs, assessment of genetic relatedness between MLGs was performed using the function `bruvo.dist()` using *poppr*, which calculates Bruvo's genetic distance, utilizing a stepwise mutation model for microsatellite loci (Bruvo et al., 2004; Kamvar et al., 2014b). This distance thus gives a more fine-scale picture of relationships between individuals than band-sharing models. These relationships were visualized with minimum spanning networks generated using the R packages *igraph* and *poppr* (Csardi and Nepusz, 2006; Kamvar et al., 2014b).

If the epidemic had a single origin, a correlation between genetic and geographic distance would be expected as populations acquire mutations over time and clonally diverge regardless of rates of spread. Divergence is affected by rates and distance of spread where long-distance dispersal or low rates of mutation would lead to less divergence and thus lower correlations between genetic and geographic distances. This was tested by performing Mantel tests across all hierarchical levels in the data set utilizing the function `mantel.randtest()` in the R package *ade4* between Bruvo's distance as described above and Euclidean distances between geographic coordinates (Dray and Dufour, 2007; Mantel, 1967). *P*-values were calculated using 99,999 bootstrap replicates.

As the eradication efforts destroy the immediate habitat in an infected area, one question that we wanted to address was whether or not genotypes were clustering to specific regions or if they were evenly spread throughout Curry County (Prospero et al., 2007). This was tested using three methods: bootstrap analysis of Nei's genetic distance, Analysis of MOlecular VAriance (AMOVA), and Discriminant Analysis of Prin-

cipal Components (DAPC) in the R packages *poppr*, *ade4*, and *adegenet* (Excoffier et al., 1992; Jombart et al., 2010; Kamvar et al., 2014b). The bootstrap analysis utilized 10,000 bootstrap replicates treating loci as independent units with the function `aboot()` in *poppr* and was visualized as an unrooted neighbor-joining tree in figtree v. 1.4.2 (Figure 4.5). AMOVA utilizes a distance matrix between genotypes for which hierarchical partitions are defined and attempts to analyze the variation within samples, between samples, between subpopulations within populations and finally between populations. In this case, we used both the hierarchies of samples within years within regions and samples within regions within years. DAPC is a multivariate, model-free approach to clustering based on prior population information (Jombart et al., 2010). This allows us to analyze the population structure by assessing how well samples can be reassigned into previously defined populations. Both DAPC and AMOVA were run with and without Hunter Creek and Pistol River South Fork due to isolated genotypes and small sample size, respectively. For the DAPC analysis, these removed populations had their origins predicted from the DAPC object using the function predict.dapc in the R package *adegenet* (Jombart et al., 2010).

Since DAPC is sensitive to the number of principal components used in analysis, the function `xvalDapc()` from the R package *adegenet* was used to select the correct number of principal components with 1,000 replicates using a training set of 90% of the data. The number of principal components was chosen based on the criteria that it had to produce the highest average percent of successful reassignment and lowest root mean squared error (Jombart et al., 2010). Significant deviations from random population structure was tested in AMOVA utilizing the function `randtest()` from

the R package *ade4* with 9,999 bootstrap replicates (Dray and Dufour, 2007).

All data and R scripts to reproduce the analyses shown here are deposited publicly on github (https://github.com/grunwaldlab/Sudden_Oak_Death_in_Oregon_Forests) and citable (DOI: 10.5281/zenodo.13007).

4.4 Results

4.4.1 Demographic pattern and genetic diversity

The epidemic has expanded over time from the initial focus in Joe Hall Creek NE of Brookings, Oregon mostly north (first to N Fork Chetco High) and northwest (Coast, Pistol River South Fork), but also east (Chetco Main and Winchuck) (Fig. 4.1; Table 4.1). To date a total of 70 multilocus genotypes have been found in forest populations (Table 4.1). MLG 22 is most abundant and the only MLG detected across the whole period (although it was not sampled in every year) (Fig. 4.2). MLG 59, the second most abundant MLG, was only detected in 2011 and has a high frequency due to the sampling design applied: all 2011 strains were sampled in one concentrated area in the northwestern sampling range geographically distant from any other location (Fig. 4.2). Given that sampling strategies for some years were not comprehensive, samples from some years have to be interpreted with caution (e.g., 2005-6, 2010-11). Samples from 2013 and 2014 are sampled from all regions and can be considered more representative.

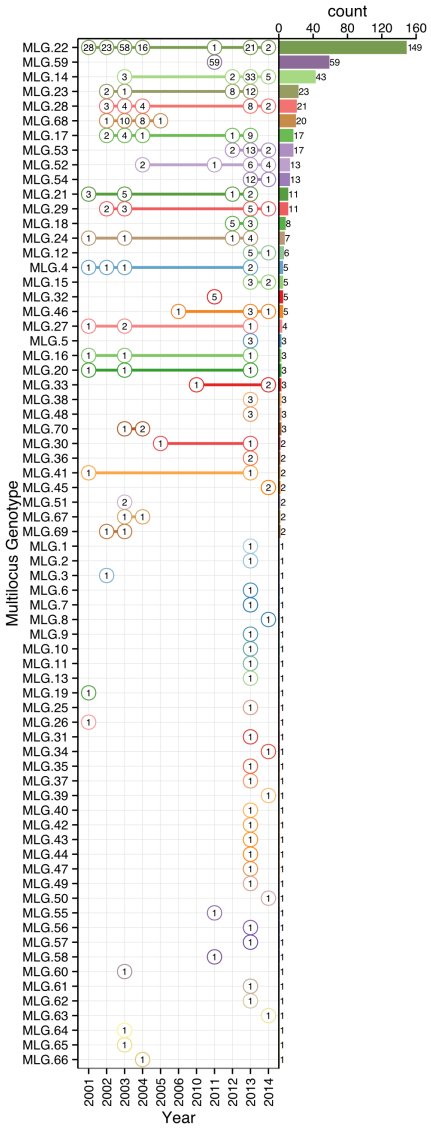


Figure 4.2: Rank distribution of multilocus genotypes (MLGs) of *P. ramorum* and recovery per year. The vertical axis denotes unique MLGs detected in the whole data set with decreasing abundance as indicated by the barplot on the right side. The horizontal axis indicates year of sampling. Each numbered circle represents the number of observations of each MLG with lines connecting genotypes found in multiple years.

Allelic and genotype diversity within loci revealed that PrMS43 had, on average, the highest number of alleles (n = 18). All other loci had 5 or fewer alleles with a moderate to high amount of diversity (Table 4.S2). Nevertheless, the genotype accumulation curve showed a slight plateau, indicating that we have enough power in our data to describe a significant number of MLGs (Fig. 4.S2). Genotypic diversity ($H = 2.98$, $G = 8.64$), evenness ($E_5 = 0.41$), and richness ($eMLG = 7$) were low as expected for a clonal population slowly accumulating mutations over space and time (Table 4.S3). A pattern of increasing diversity across years (with number of MLGs not fewer than 10) was also observed (Table 4.S3). The minimum spanning network showed that MLGs 17, 22, and 28 clustered in the center of the network and had the highest number of connections to other genotypes in the forest populations (Fig. 4.3). Most genotypes were connected to their immediate neighbors by a genetic distance of 0.05 or the equivalent of one mutational step across 5 diploid loci.

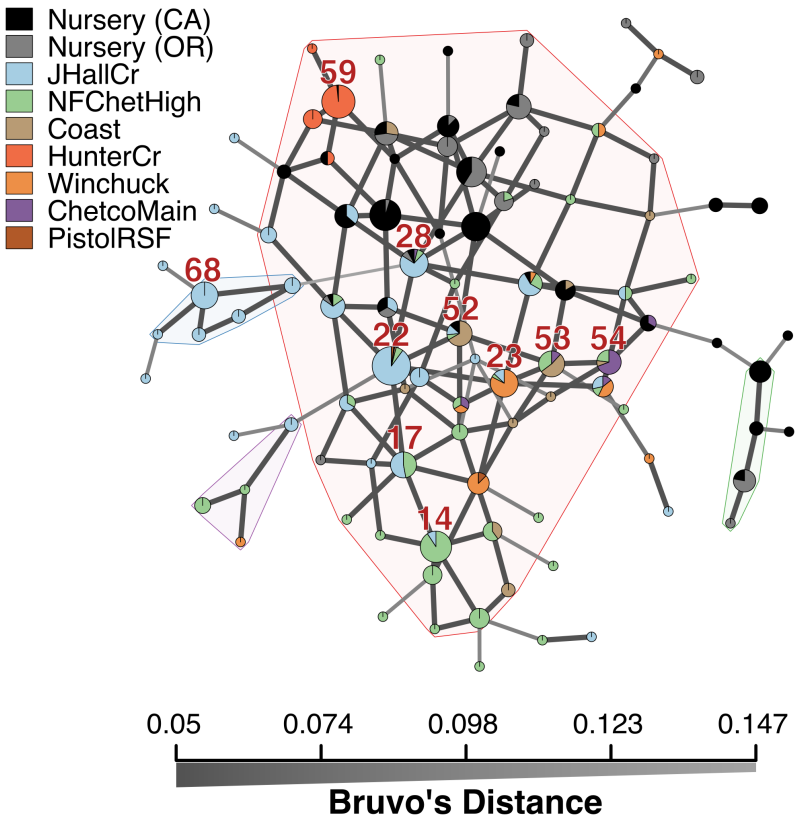


Figure 4.3: Minimum spanning network based on Bruvo's genetic distance for microsatellite markers for *P. ramorum* populations. Nodes (circles) represent individual multilocus genotypes. The 10 most abundant forest genotypes are labeled with their MLG designation. Node colors represent population membership proportional to the pie size. Node sizes are relatively scaled to $log_{1.75}n$, where n is the number of samples in the nodes to avoid node overlap. Edges (lines) represent minimum genetic distance between individuals determined by Prim's algorithm. Nodes that are more closely related will have darker and thicker edges whereas nodes more distantly related will have lighter and thinner edges or no edge at all. Reticulation was introduced by finding exact ties in genetic distance after Prim's algorithm was run. Subgroups of >3 MLGs where all nodes are no more than one mutational step (d = 0.05) away from its neighbors, are highlighted in arbitrary colors.

4.4.2 Spatial Correlation

A Mantel test revealed significant correlations of genetic distance and geographic distance for most samples collected after 2003 (Table 4.3). When partitioned by year, this correlation appears to increase and become more pronounced with the progression of the epidemic. When partitioned by region, those that are closer to the origin of the epidemic (Joe Hall Creek and N Fork Chetco River) show significant correlation. When the overall mantel test was run without Hunter Creek, the correlation coefficient was reduced (0.175), but was still significant ($p = 0.0001$).

4.4.3 Population differentiation

Cluster analysis of populations with respect to year using Nei's genetic distance showed no significant (>70%) bootstrap support for any clades, but does show that these tend to cluster by region as opposed to year (Fig. 4.S3). AMOVA analysis revealed significant population structure between regions on both clone-corrected (with respect to hierarchy) and uncorrected data sets (Table 4.4). Significant structure was only found between years within regions on the uncorrected data set. Both patterns were observed without Hunter Creek and Pistol River South Fork isolates. DAPC clustering showed that the first discriminant component separated Hunter Creek from all other regions and the second discriminant component shows a gradient from Joe Hall Creek to the coast (Fig. 4.4). This distinction was reflected in the percent of correct posterior assignment of isolates to their original populations. Over the whole data set there was an 81.5% assignment-success rate. Hunter Creek received 100% successful reassign-

Table 4.3: Table of correlation coefficients generated across forest regions and years of *P. ramorum* isolates using the Mantel test. Euclidean distances were calculated from geographic coordinates while genetic distance was based on Bruvo's distance. Significance of values are based on 99,999 Monte-Carlo permutations and marked as follows: ^ $\leq$ 0.05, ~ $\leq$ 0.01, * $\leq$ 0.001, - = no data, NaN = insufficient data for analysis.

	2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2010**	**2011**	**2012**	**2013**	**2014**	**Pooled**
JHallCr	0.06	0.24	0.14*	0.28*	NaN	-	-	-	-	0.18~	NaN	0.14*
NFChetHigh	-	-	NaN	-	-	-	-	-	0.68	0.41*	-0.23	0.35*
Coast	-	-	-	-	-	NaN	NaN	NaN	NaN	0.55^	-0.25	0.13
HunterCr	-	-	-	-	-	-	-	0.06	-	-	-	0.06
Winchuck	-	-	-	-	-	-	-	-	0.41~	0.03	-	0.11
ChetcoMain	-	-	-	-	-	-	-	-	-	0.53	NaN	0.63*
PistolRSF	-	-	-	-	-	-	-	-	-	0.94	-	0.94
Pooled	0.06	0.24	0.13*	0.28*	NaN	NaN	NaN	0.87*	0.59*	0.15*	0.14^	**0.52***

ment. Joe Hall Creek, Winchuck, and Coast all had >85% successful reassignment whereas Chetco Main, North Fork of the Chetco, and Pistol River South Fork all had <69% successful reassignment (Fig. 4.S4). The isolation of the Hunter Creek isolates in the DAPC analysis was found to be mainly driven by allele 493 at locus PrMS43 (Fig. 4.S5). The only other population to share this allele was Joe Hall Creek where it was present in a total of 4 isolates, and only isolates found in the coastal region or North Fork Chetco contained the allele 489, which is one mutational step away in a stepwise mutation model of a tetranucleotide repeat locus. When DAPC was run without Hunter Creek and Pistol River South Fork data, percent successful reassignment for all regions did not change significantly. Prediction of sources for the Hunter Creek data revealed that over 98% of the genotypes were assigned to the Coast with a 99% probability.

Table 4.4: AMOVA table generated comparing *P. ramorum* isolates for two different hierarchies, year within region and region within year, respectively. Results are rounded to three significant figures. Clone corrected results are provided in parentheses. P values are based on 9,999 permutations.

Heirarchy	**df**	**Sum of squares**	**Variation (%)**	***P***	ϕ **statistic**
Region by year					
Between region	10 (10)	160 (21)	11.6 (3)	0.366 (0.175)	0.448 (0.101)
Between year within region	12 (12)	59.5 (19.1)	33.3 (7.07)	1e-04 (2e-04)	0.376 (0.0729)
Within year within region	490 (129)	281 (141)	55.2 (89.9)	1e-04 (1e-04)	0.116 (0.03)
Year by region					
Between year	6 (6)	197 (23)	45 (12.3)	1e-04 (1e-04)	0.496 (0.12)
Between region within year	16 (16)	22.5 (17.2)	4.56 (-0.283)	1e-04 (0.446)	0.0829 (-0.00323)
Within region within year	490 (129)	281 (141)	50.4 (88)	1e-04 (1e-04)	0.45 (0.123)

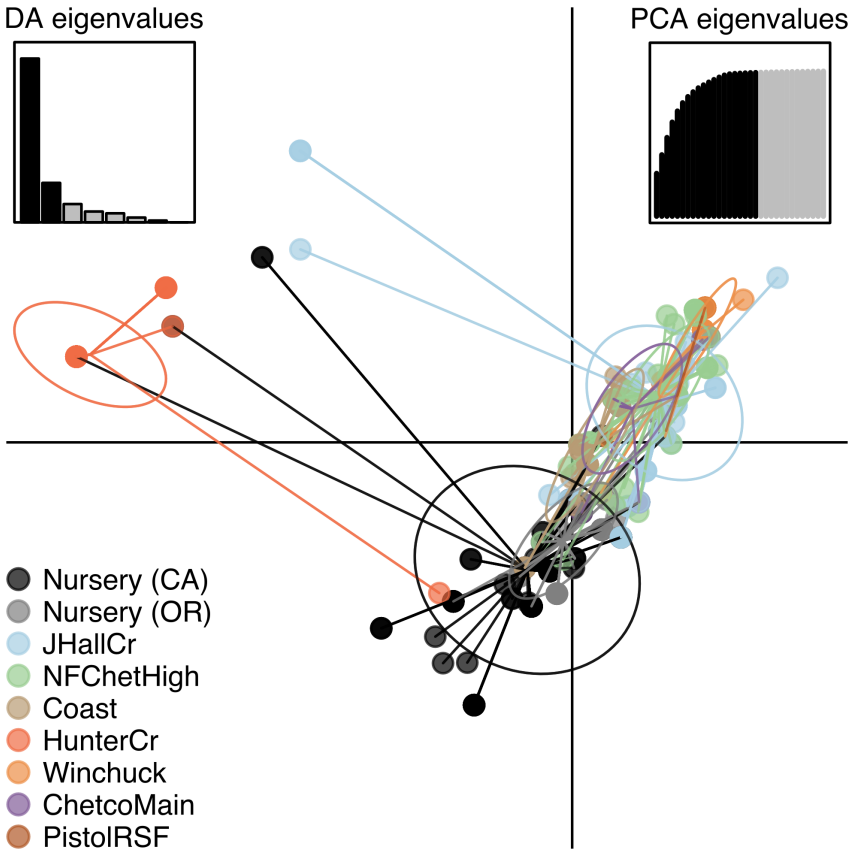


Figure 4.4: Scatterplot from DAPC of the first two principal components discriminating *P. ramorum* populations by regions. Points represent individual observations. Colors and lines represent population membership. Inertia ellipses represent an analog of a 67% confidence interval based on a bivariate normal distribution.

4.4.4 Clustering of forest with nursery populations

We used previously published data to determine if nursery populations in California or Oregon could have been source populations for the Oregon forest epidemic (Goss et al., 2009, 2011; Grünwald et al., 2009; Prospero et al., 2009, 2007). Nursery data included 40 MLGs across 216 samples of NA1 clones. Of these 40, 12 MLGs matched the forest sample and 28 MLGs were unique to the nurseries. The only region that did not contain genotypes that matched those found in nurseries was Pistol River South Fork. When considering those 12 genotypes that were present in both data sets, with the exception of Joe Hall Creek, all genotypes were first isolated from nurseries before discovery in the forest. At the most variable locus, PrMS43, both nursery populations had the allele 281 at frequencies of 4.5% and 4.9% for CA and OR, respectively. This allele was not observed in the forest population. Both populations contained allele 489 at $>$10% frequency and the CA nursery population contained allele 493 at a frequency of 1.4%.

When nursery genotypes were added to the minimum spanning network, MLGs found at Hunter Creek, previously isolated in the network, connected by only a single MLG from the coast, gained more connections to nursery MLGs. Clustering with Nei's distance revealed the Nursery isolates from CA consistently clustering closest with Hunter Creek isolates in both clone-corrected and uncorrected data sets (Fig. 4.5). DAPC clustering revealed a decrease in overall assignment-success rate at 78%. The nursery isolates received 74% and 83% assignment success for CA and OR nurseries, respectively.

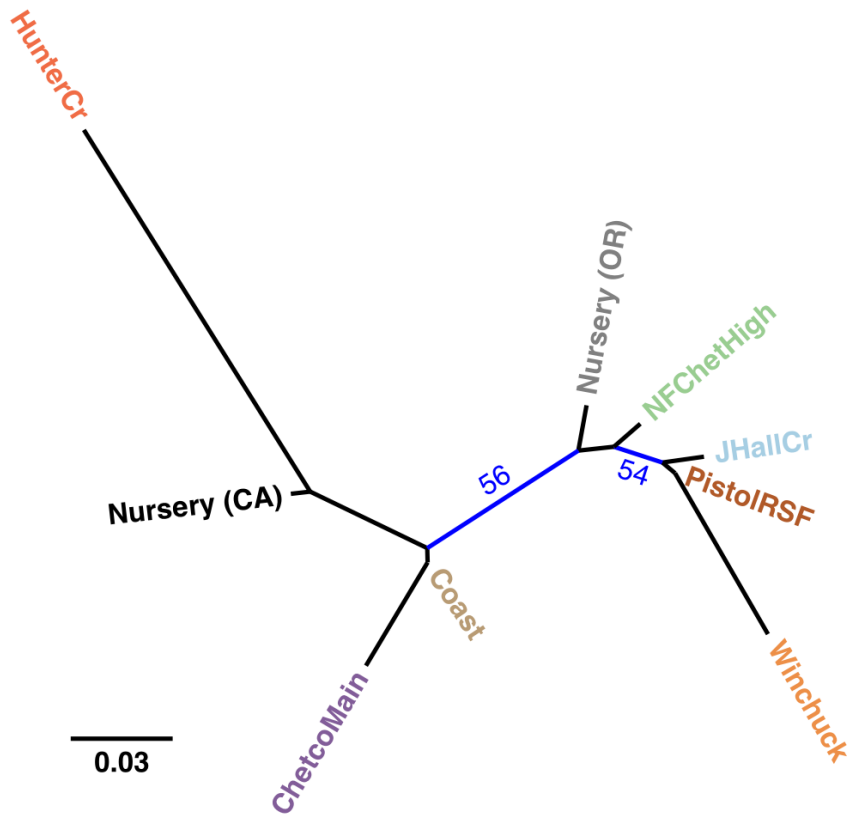


Figure 4.5: Unrooted, neighbor-joining tree with 10,000 bootstrap replicates of Nei's genetic distance for *P. ramorum* populations defined by region. Tip labels are colored by region. Branches with bootstrap values greater than 50% are shown in blue. Nursery populations are shown as originating from California (CA) or Oregon (OR).

The assignment successes for the regions before or after inclusion of nursery data changed less than 5% for all regions except for the coast, which saw a decrease of 17.6% when nursery populations were included (Fig. 4.S4). Prediction of sources for nursery genotypes against the forest data revealed that 48% of these nursery isolates were predicted to share membership with the coast at $\geq$ 95% probability (Fig. 4.S6, 4.S7). A total of 3.2% of the isolates were predicted to share membership with Hunter Creek at $\geq$ 99.9% membership probability. Furthermore, 21.75% of the nursery data could not be assigned to any of the forest populations at >60% probability.

Since 3.2% of the nursery isolates had a very strong signal for Hunter Creek, we predicted sources for Hunter Creek isolates when considering nursery isolates. This approach determines if Hunter Creek isolates cluster more readily with nursery or coast populations. Indeed, 92% of the Hunter Creek isolates were predicted to share membership with California nurseries at $\geq$ 99% membership probability (Fig. 4.S8). No Hunter Creek isolate was predicted to share membership with a forest population at >0.45% membership probability.

4.5 Discussion

To date populations monitored from 2001-14 show presence of only the NA1 clonal lineage observed previously (Prospero et al., 2009, 2007). The fact that individuals belonging to the EU1 and NA2 clonal lineages have not been found in Oregon forests, despite their presence on the west coast from British Columbia to California is welcome news (Goss et al., 2009, 2011; Grünwald et al., 2012). The lack of EU1 or NA2 isolates

provides evidence that monitoring for *P. ramorum* in nurseries by federal and state agencies is helping avoid emergence of new clones in Oregon's Forests.

Our analysis provides support for a most parsimonious scenario of two introductions into Curry county from nurseries: one initial introduction into Curry County sometime before detection of the first infected tanoaks in 2001 from California (or less likely Oregon) nurseries followed by a second introduction into the Hunter Creek area again from nurseries. The relative position of the nursery populations in the minimum spanning network and DAPC scatter plot (Fig. 4.3, 4.4) suggest that the introductions from nurseries were rare, though more even sampling and migration models could disprove this hypothesis. Since 2001, the epidemic has spread clonally throughout Southwestern Curry County mostly north, but also west, towards the coast, and southeast. This clonal spread of the pathogen from the Joe Hall area is supported partially by Mantel tests showing significant levels of isolation by distance in years following 2002 (Table 4.3) along with significant AMOVA results across regions (Table 4.4). The populations sampled in 2011 in Hunter Creek (Cape Sebastian) appear to have originated from a new source and cluster into a distinct group based on DAPC (Fig. 4.4). Based on the minimum spanning network, this population would appear isolated in the epidemic if it were not for MLG 32, which is connected with MLG 33 (found on the coast in 2010) by one mutational step at locus PrMS43. This, in turn, is connected with the other MLGs from Hunter Creek by one mutational step at locus PrMS39. When considering clustering via Bruvo's distance in combination with data from nursery populations, however, these genotypes from Hunter Creek appear to be more similar to California nursery populations (Fig. 4.3) than Oregon forest populations. Predictions

based on DAPC place samples from Hunter Creek as coming from California nurseries (Fig. 4.S8). This, in combination with the observation that purely forest genotypes (i.e. those only found in the forest) are connected to Hunter Creek genotypes through nursery genotypes, indicates a possible contribution from nursery populations to the epidemic. This is supported by the observation that all population level clustering, with and without clone correction, places the Hunter Creek isolates adjacent to the nursery isolates from CA (Fig. 4.4, 4.5). This appears to be driven by the high frequency of allele 246 at locus PrMS39, which interestingly appears to segregate in an east to west fashion and is increasing in frequency over time (Fig. 4.S9, 4.S10). This, along with the results from the DAPC clustering and subsequent prediction (Fig. 4.S6, 4.S7) provide weak support for a potential third introduction into the coastal region from nurseries sometime after the Hunter Creek introduction event.

An interesting aspect is the observation that there appeared to be more than one cluster of genotypes introduced into the Joe Hall area during the early stages of the epidemic. The two dominant clusters that appeared were the ones that contained MLG 22 and MLG 68. The former has been found in the most recent sampling year, whereas the latter has not been observed since 2005 or beyond the Joe Hall area. This latter group was also the most distantly related group overall, more distant than some nursery genotypes. While it is clear that the eradication effort has not been entirely successful, there is some evidence that it is having an effect as a major genotype cluster has effectively been eradicated, although disappearance of MLGs could also be explained by being less fit than lineages dominating now.

The Curry County epidemic is in many ways different from the epidemic in Cali-

fornia. When introduced into California in the mid 1990's, the causal agent of sudden oak death was unknown and thus gave it time to clonally expand and diversify as management strategies in natural forest systems were limited (Rizzo et al., 2002). With the foresight of the epidemic in central California, the ODF was able to implement a quarantine effort against the import of hosts as soon as the causal agent was known (A Kanaskie, pers. comm.). This quarantine along with aggressive eradication efforts have affected the spread of *P. ramorum* (Mascheretti et al., 2008). Drawing conclusions from previous population studies in California and applying them to the Oregon epidemic should be undertaken with great care given the drastically differing management scenarios (Mascheretti et al., 2009, 2008).

Our work has some inherent drawbacks. Given the cost of aerial surveys and subsequent ground crew work, and the fact that trees are eradicated once found, populations are not hierarchically sampled across all years. The destructive nature of the management approach means that it was not possible to conduct controlled ecological experiments focusing on effects of climate and rainfall on the spread of disease as was possible in California trials (Eyre et al., 2013). In addition, most of our work only used 5 microsatellite loci for genotyping. Ideally, more loci should have been used as was done in other studies (Croucher et al., 2013). Although only 5 loci were used, clear patterns of population dynamics in space and time emerged and the MLG accumulation curve supported the fact that loci are informative. Finally, the populations genotyped here are clonal and belong to the NA1 clonal lineage. Thus, much of the analytical power provided by population genetic theory does not apply given that basic assumptions would be violated (Grünwald and Goss, 2011). Our work

uses appropriate methods to infer patterns that are model free, yet informative such as spatial clustering. Thus, we believe that this work provides novel and important insights into the *P. ramorum* population biology in the Siskiyou forest. Our data indicates that there might have been at least two introductions into Oregon forests from nurseries. The nature of the data does not allow inference of directional migrations given the uneven sampling strategy and moderate number of loci used across all years. We are currently exploring genotyping-by-sequencing (GBS) as a method that could provide further detail on how these populations evolved over space and time (Elshire et al., 2011). GBS can provide richer detail by providing codominant SNP data across several thousand loci sampling the whole genome.

4.6 Acknowledgements

This work was supported in part by US Department of Agriculture (USDA) Agricultural Research Service Grant 5358-22000-039-00D, USDA APHIS, the USDA-ARS Floriculture Nursery Initiative, the Oregon Department of Agriculture/Oregon Association of Nurseries (ODA-OAN) and the USDA-Forest Service Forest Health Monitoring Program.

4.7 Supplementary Material

4.7.1 Supplementary Text

For the years up to and including 2012, the multilocus genotype (MLG) of each *P. ramorum* strain was determined based on microsatellite analysis of five loci, PrMS6, Pr9C3, PrMS39, PrMS45 and PrMS43, using previously published protocols (Grünwald et al., 2009, 2008b; Prospero et al., 2004, 2007). Multilocus genotyping of *P. ramorum* strains collected in 2013 and 2014 included an extra nine loci, KI18, KI64, KI82a, KI82b, ILVOPrMS79, ILVOPrMS131, ILVOPrMS145a, ILVOPrMS145b, ILVOPrMS145c which are amplified by an additional six primer pairs (Ivors et al., 2006; Vercauteren et al., 2010, 2011). The locus ILVOPrMS79, amplifies up to three alleles, however two separate loci have yet to be described (Vercauteren et al., 2011). The addition of nine loci to the genotyping assay coincided with the discovery by the Oregon Department of Agriculture of an EU1 *P. ramorum* isolate in a Curry County nursery in 2012. Preceding 2012, only NA1 isolates had been found in Curry County. Because different loci are polymorphic for different clonal lineages, the entire panel of 14 loci was necessary to adequately describe the *P. ramorum* population in the event that multiple lineages were discovered in the forest.

Methods for genotyping the 2013 and 2014 *P. ramorum* strains with all 14 loci use new multiplex protocol. Previously published primers were modified by the addition of a 5' PIG tail "GTTT" to reverse primers in an effort to reduced stutter peaks and hence to better facilitate allele scoring (Table 4.2) (Brownstein et al., 1996). In two cases (PrMS45, PrMS6), a PIG tail was not added to reverse primers to simplify

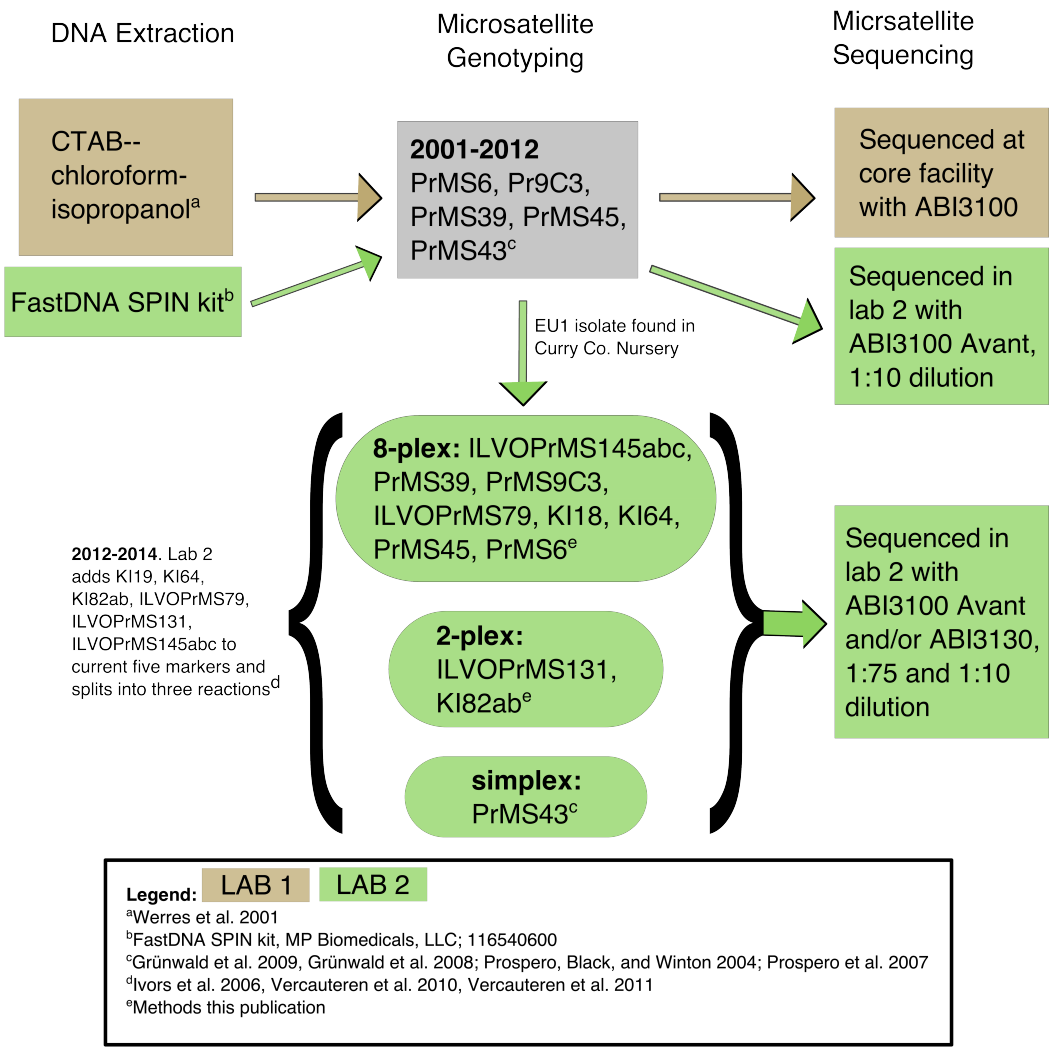
scoring of overlapping alleles. Also, where a T residue was already present at the 5' end of a reverse primer, as in the case of KI18, a "GTT" was added instead of "GTTT". Forward primers were assigned fluorescent labels, 6-FAM, NED, VIC, or PET, to facilitate separation of overlapping markers (Table 4.2). Primer concentrations were determined by visual inspection of electropherograms (Table 4.2).

Amplification of all 14 loci was separated into three reactions (8-plex, 2-plex and simplex). The simplex reaction amplified the PrMS43 locus using methods described earlier (Grünwald et al., 2009; Prospero et al., 2007). The 8-plex and 2-plex amplified the remaining loci and were performed under identical conditions with the exception of primers and primer concentrations (Table 4.2). For the multiplex reactions, the QIAGEN Type-it Mutation Detect PCR Kit (QIAGEN, 206343, Valencia, CA) was used. Multiplex PCR reactions were performed in 5μl volumes with 10ng template DNA and1X final buffer concentration. Amplifications were run on a Veriti thermal cycler (Life Technologies, Grand Island, NY) with an initial denaturation at 95 °C for 5 min, followed by 33 cycles of 95 °C for 30 s, 60 °C for 90 s, and 72 °C for 20 s, and a final extension at 60 °C for 30 min. Genotyping prior to 2012 included three reference DNA lineages (EU1, NA1, and NA2). After 2012, a fourth lineage (EU2) was added as a reference (Poucke et al., 2012).

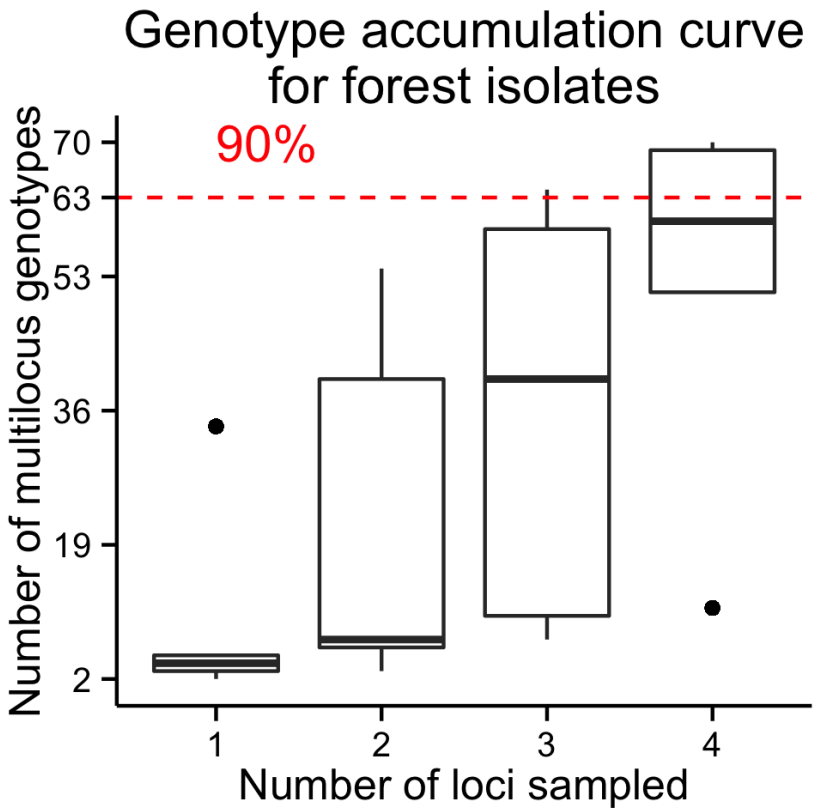
Electrophoresis and visualization of all microsatellites were performed on ABI3100, ABI3100 Avant, or ABI3130 genetic analyzers (Applied Biosystems). For evaluation of the loci, genotyped prior to 2013, the PCR products were diluted 10 times in ultrapure H_20 and 1.5 μl of diluted product was added to both 8.5 μl of Hi-Di^{TM} Formamide (Applied Biosystems, 4311320) and 0.25 μl of GeneScan^{TM} 500 LIZ^{TM} size standard

(Applied Biosystems, 4322682). The simplex reaction (PrMS43) was also diluted 10 times while the 8-plex and 2-plex products were diluted 75 times. After dilution, 2.5 μl of the 8-plex, 2-plex and simplex products were added to 7.5 μl of Hi-Di^{TM} Formamide containing GeneScan^{TM} 500 LIZ^{TM} size standard at a ratio of 6 μl size standard to 1 ml Hi-Di^{TM} Formamide. Allele sizing was determined using GeneMapper® v3.7 and v5.0 software (Applied Biosystems).

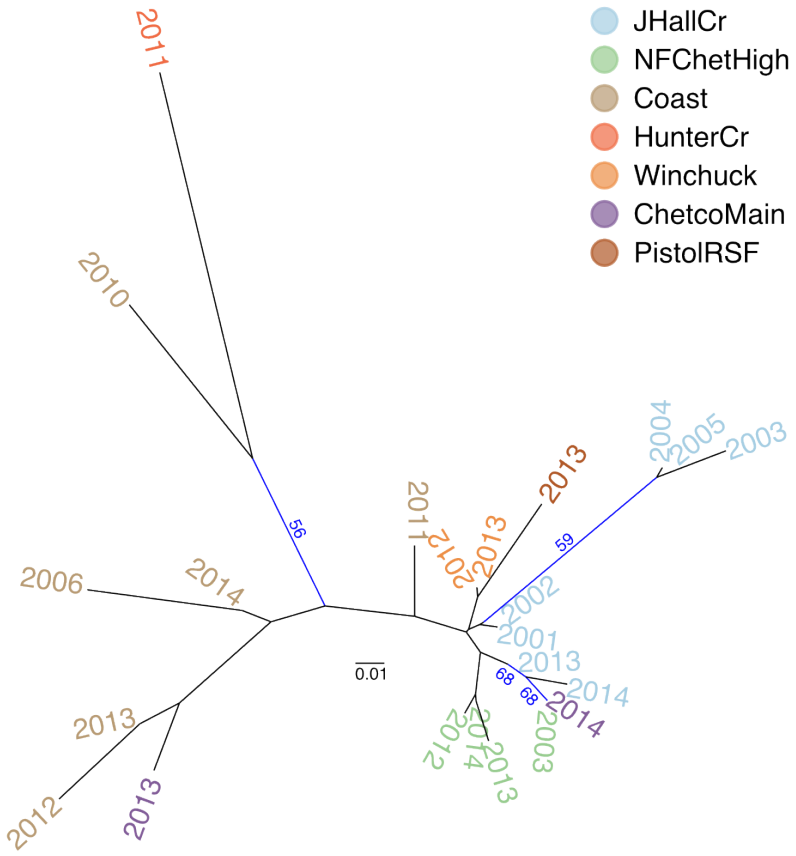
4.7.2 Supplementary Figures



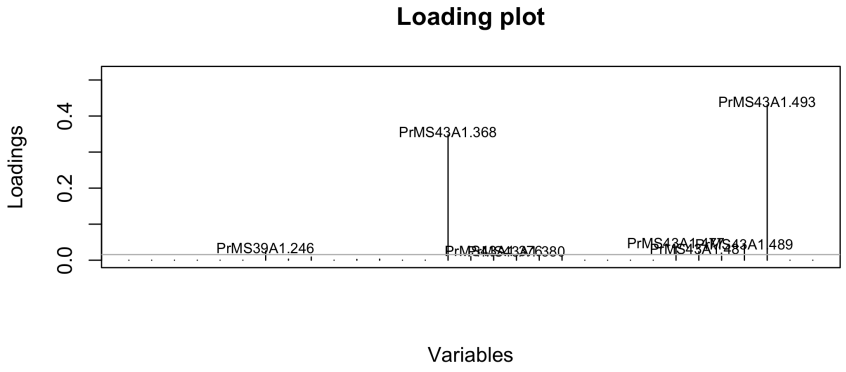
Supplementary Figure 4.S1: Diagram of DNA extraction, genotyping, and sequencing protocols utilized by two labs from 2001 to 2014. See supplementary text for details.



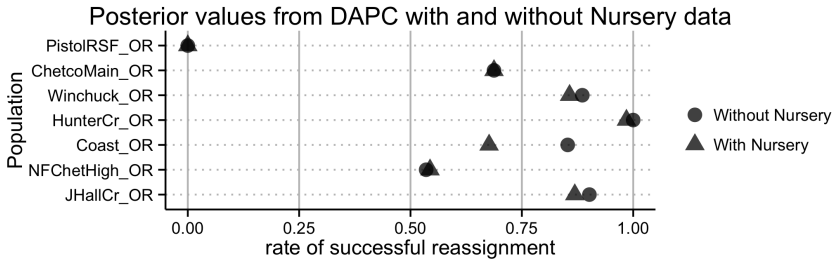
Supplementary Figure 4.S2: Genotype accumulation curve for OR forest *P. ramorum* isolates. The vertical axis denotes the number of observed MLGs, from 0 to the observed number of MLG in the forest populations, for a number of loci, indicated on the horizontal axis, randomly sampled without replacement. Each boxplot contains 1,000 random samples representing different possible combinations of *n* loci. The horizontal red dashed line represents 90% of MLG resolution.



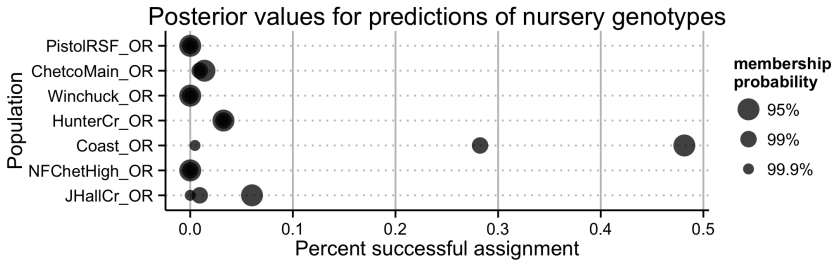
Supplementary Figure 4.S3: Neighbor joining tree based on Nei's distance of the forest *P. ramorum* isolates by region with respect to year. Bootstrap values $>$ 50% of 10,000 replicates are shown in blue.



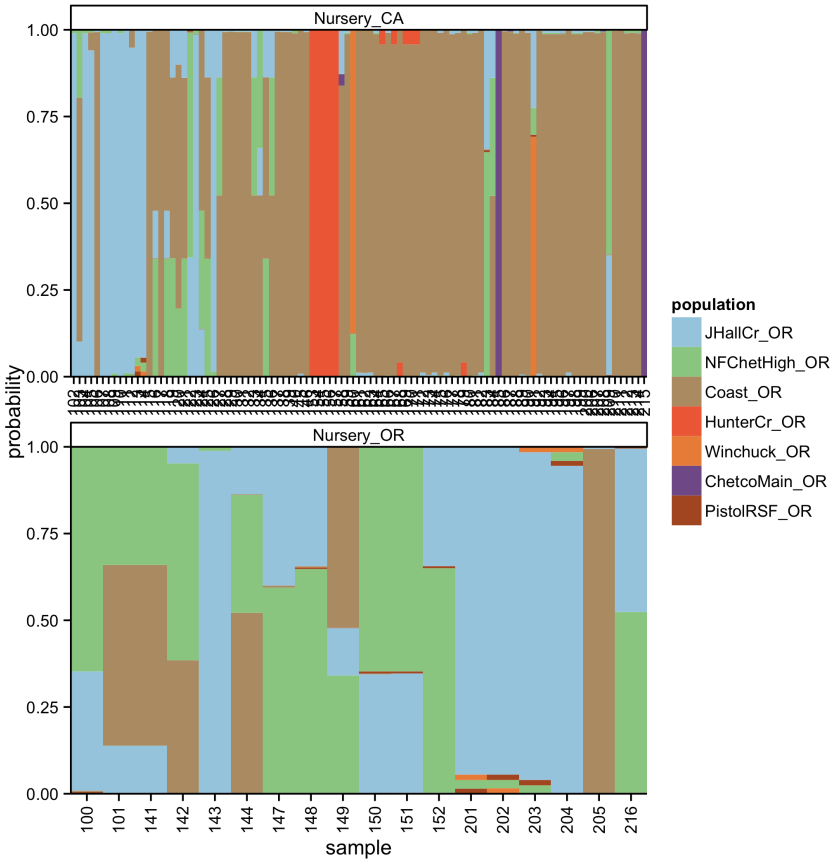
Supplementary Figure 4.S4: Fractions of posterior population assignments from DAPC clustering of *P. ramorum* isolates from forest populations. The horizontal axis represents the fraction of samples whose posterior group membership matched their prior group membership on the vertical axis. Shape indicates presence or absence of nursery populations in the DAPC.



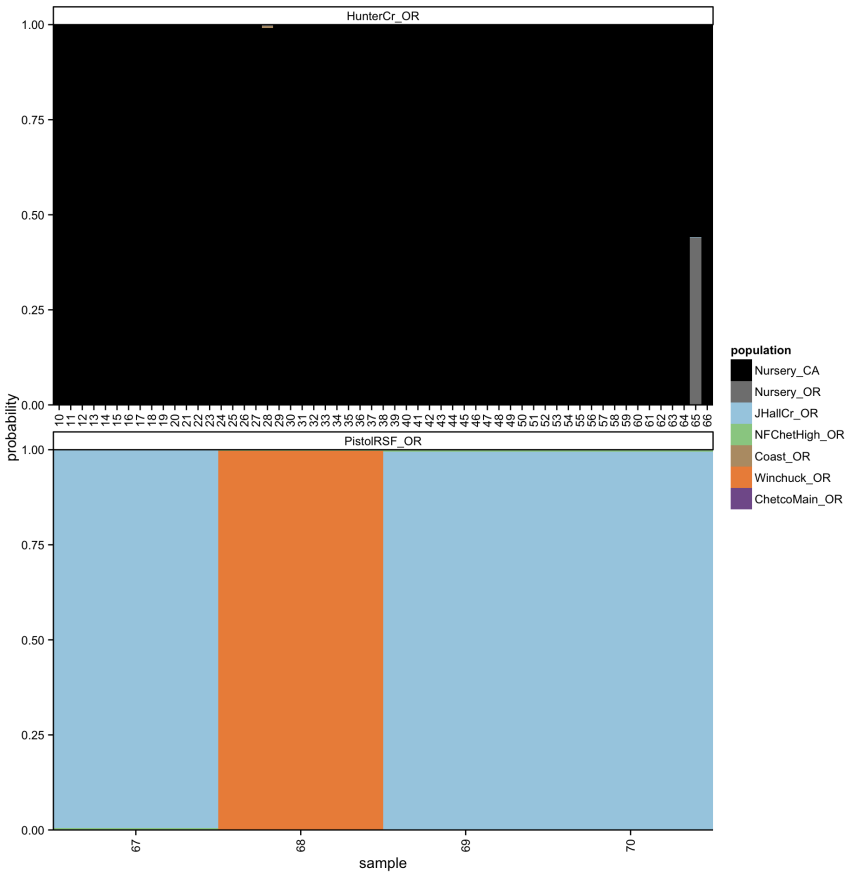
Supplementary Figure 4.S5: Loading plot from DAPC of *P. ramorum from* forest populations showing the contribution of alleles to the first DAPC eigenvalue separating Hunter Creek isolates from all other regions.



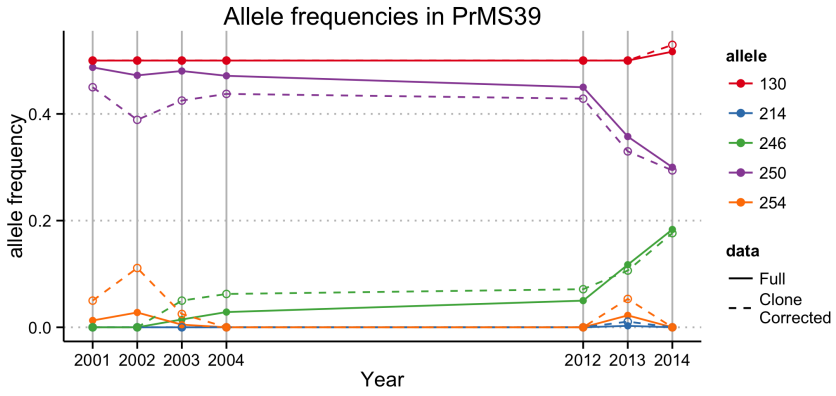
Supplementary Figure 4.S6: Prediction of nursery genotypes of *P. ramorum* into forest watershed regions. The horizontal axis indicates the fraction of nursery genotypes to be predicted to be similar to the populations on the vertical axis with a 95, 99, and 99.9% probability as indicated by the size of the points.



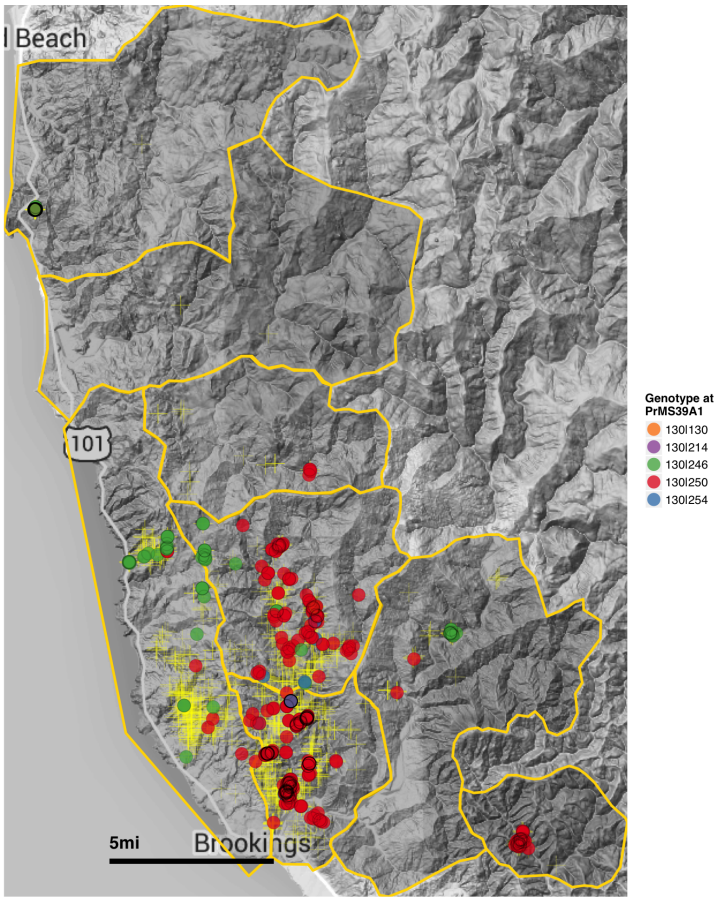
Supplementary Figure 4.S7: Graphical representation of prediction of nursery isolates of *P. ramorum* into forest watershed regions. Each column represents a different isolate. Colors within the columns represent membership probabilities from forest populations.



Supplementary Figure 4.S8: Graphical representation of predicted membership of *P. ramorum* isolates from Hunter Creek and Pistol River South Fork in forest and nursery populations. Each column represents a different isolate. Colors within the columns represent membership probabilities from the populations indicated in the legend.



Supplementary Figure 4.S9: Allele frequencies of locus PrMS39 of *P. ramorum* across years of the forest populations. Years 2005 through 2011 have been omitted due to small sample sizes and outlier genotypes.



Supplementary Figure 4.S10: Map of the infected area in Curry county showing the *P. ramorum* genotypes at locus PrMS39. Each colored circle represents a different forest isolate while each yellow cross represents a sampled tree. Yellow borders denote different regions.

4.7.3 Supplementary Tables

Supplementary Table 4.S1: (Caption on next page)

Population	MLG	alleles	1-D	Hexp	Evenness
ChetcoMain	7	3.0	0.58	0.97	0.94
Coast	12	3.8	0.59	0.97	0.96
HunterCr	4	2.6	0.56	0.96	0.95
JHallCr	30	4.4	0.57	0.91	0.89
NFChetHigh	35	4.8	0.60	0.90	0.89
PistolRSF	2	2.4	0.55	1.00	1.00
Winchuck	9	3.4	0.56	0.94	0.92
2001	10	3.6	0.56	0.94	0.91
2002	9	3.0	0.56	0.94	0.93
2003	20	4.2	0.56	0.90	0.89
2004	8	2.8	0.50	0.84	0.85
2005	2	2.4	0.50	0.90	0.92
2006	1	2.0	0.50	1.00	1.00
2010	1	2.0	0.50	1.00	1.00
2011	6	3.0	0.58	0.98	0.96
2012	7	3.2	0.58	0.96	0.95
2013	47	5.4	0.60	0.89	0.90
2014	17	3.8	0.59	0.97	0.93
pooled	70	6.0	0.60	0.89	0.90

Supplementary Table Caption 4.C1: (Caption for Table 4.S1) Mean allelic diversity metrics of clone-corrected populations of *Phytophthora ramorum* sampled in Curry County, Oregon between 2001-14 causing sudden oak death. MLG = Number of Multilocus Genotypes; alleles = Average number of alleles across 5 loci; 1-D = Simpson's Index averaged across 5 loci; Hexp = Nei's 1978 expected heterozygosity; Evenness = Evenness averaged across 5 loci

Supplementary Table 4.S2: Allelic diversity metrics for each locus of clone-corrected *Phytophthora ramorum* data in Curry County, Oregon between 2001-14 causing sudden oak death. alleles = number of observed alleles; 1-D = Simpson's Index; Hexp = Nei's 1978 expected heterozygosity; E.5 = Evenness

locus	alleles	1-D	Hexp	E.5
PrMS6	2	0.50	0.99	0.99
Pr9C3	2	0.50	0.99	0.99
PrMS39	5	0.62	0.78	0.78
PrMS45	3	0.51	0.76	0.95
PrMS43	18	0.89	0.95	0.78
mean	6	0.60	0.89	0.90

Supplementary Table 4.S3: (Caption on Next Page)

Population	N	MLG	eMLG	SE	H	G	E.5	rbarD	p.rD
2001	39	10	3.65	1.16	1.19 (0.61-1.48)	1.9 (1.38-2.81)	0.4 (0.4-0.57)	-0.07	1.00
2002	36	9	4.20	1.13	1.37 (0.79-1.64)	2.34 (1.52-3.52)	0.45 (0.43-0.65)	0.19	0.00
2003	102	20	4.75	1.30	1.81 (1.39-2)	2.92 (2.19-3.96)	0.38 (0.36-0.49)	0.14	0.00
2004	35	8	4.56	1.01	1.57 (1.12-1.76)	3.53 (2.3-4.77)	0.66 (0.56-0.84)	0.18	0.00
2005	2	2	2.00	0.00	0.69 (0-0.69)	2 (1-2)	1 (1-1)		
2006	1	1	1.00	0.00	0 (0-0)	1 (1-1)	NaN (NA)		
2010	1	1	1.00	0.00	0 (0-0)	1 (1-1)	NaN (NA)		
2011	68	6	2.15	0.83	0.56 (0.26-0.79)	1.32 (1.13-1.59)	0.42 (0.38-0.56)	0.18	0.00
2012	20	7	5.01	0.92	1.62 (1.03-1.78)	4 (2.2-5.26)	0.74 (0.59-0.92)	-0.17	1.00
2013	179	47	7.74	1.20	3.15 (2.83-3.17)	13.8 (10.2-16.3)	0.57 (0.54-0.7)	0.03	0.00
2014	30	17	8.09	1.03	2.67 (2.08-2.61)	12.2 (6.16-12.2)	0.83 (0.69-0.93)	-0.01	0.56
JHallCr	244	30	4.89	1.33	1.97 (1.69-2.12)	3.13 (2.57-3.83)	0.34 (0.33-0.41)	0.10	0.00
NFChetHigh	114	35	6.81	1.34	2.74 (2.3-2.82)	7.07 (4.8-9.88)	0.42 (0.4-0.59)	0.07	0.00
Coast	34	12	5.91	1.17	2.05 (1.49-2.16)	5.56 (3.36-7.22)	0.68 (0.61-0.85)	0.28	0.00
HunterCr	66	4	1.88	0.69	0.42 (0.18-0.62)	1.24 (1.1-1.46)	0.46 (0.39-0.6)	-0.05	1.00
Winchuck	35	9	4.29	1.07	1.47 (0.95-1.7)	2.88 (1.89-4.15)	0.56 (0.49-0.77)	-0.01	0.76
ChetcoMain	16	7	5.00	0.93	1.45 (0.69-1.69)	2.84 (1.49-4.74)	0.56 (0.49-0.88)	0.40	0.00
PistolRSF	4	2	2.00	0.00	0.56 (0-0.69)	1.6 (1-2)	0.8 (0.8-1)		
Total	513	70	6.95	1.33	2.98 (2.78-3.04)	8.64 (7.19-10.1)	0.41 (0.39-0.48)	0.08	0.00

Supplementary Table Caption 4.C3: (Caption for Table 4.S3) Genotypic diversity metrics or populations of *Phytophthora ramorum* sampled in Curry County, Oregon between 2001-14 causing sudden oak death. Pop = Population name (Total == Pooled) N = Census population size MLG = Number of unique multilocus genotypes (MLG) observed eMLG = Number of expected MLG based on rarefaction at smallest N >= 10 SE = Standard error of rarefaction analysis H = Shannon-Wiener Index of MLG diversity (95% CI in parentheses) G = Stoddart and Taylor's Index of MLG diversity (95% CI in parentheses) E.5 = Evenness (95% CI in parentheses) rbarD = Standardized index of association p.rbarD = p-value for the standardized index of association based on 999 permutations NaN = Insufficient data for analysis

Chapter 5: Factors Influencing Recombination Inference in Diploid Populations

Zhian N. Kamvar and Niklaus J. Grünwald

Target Journal: **Molecular Ecology**

5.1 Abstract

TBD...

5.2 Introduction

Population genetic theory is largely based on the neutral Wright-Fisher model in which populations are infinitely large, with discrete generations, randomly assorting alleles, no migration, and no mutation (Hartl and Clark, 2007; Nielsen and Slatkin, 2013). By using this model, population geneticists are able to ask fundamental questions and test hypotheses about evolutionary processes that could lead to structured populations.

This neutral model, however, cannot be applied to populations whose life history violates the fundamental assumption of random assortment of alleles, such as populations that undergo clonal reproduction (Milgroom, 1996; Orive, 1993). For many clonal populations, the contribution of genetic variation from mutation is greater than that of recombination. While this increases the risk of the accumulation of deleterious mutations, the two-fold cost of sex is drastically reduced, meaning that selectively advantageous combinations of alleles are maintained (Heitman et al., 2012). Detecting recombination in populations of pathogenic microorganisms is therefore important for management strategies as a prevalence of sexual reproduction could create resistant genotypes (de Meeûs et al., 2006; Goss et al., 2014; Milgroom, 1996; Nieuwenhuis and James, 2016; Smith et al., 1993).

Several studies have been conducted in an attempt to quantify the amount of sex in populations that undergo clonal reproduction (Ali et al., 2016; Balloux et al., 2003;

de Meeûs and Balloux, 2004; Nieuwenhuis and James, 2016; Smith et al., 1993). For populations with well-defined sexual and clonal phases occurring at separate times, such as rust type fungi, methods like *CloNcaSe* are effective for estimating the rate of sexual reproduction and effective population size (Ali et al., 2016). However, this method cannot be applied to populations where the reproductive cycle is not partitioned into discrete generations.

Simply detecting the presence of clonal reproduction, however can be useful in and of itself (Milgroom, 1996). A method commonly used to assess this is the index of association (I_A), and its standardized version, $\bar{r}_d$, which measure multilocus linkage disequilibrium (Agapow and Burt, 2001; Brown et al., 1980; de Meeûs and Balloux, 2004; Haubold et al., 1998; Kamvar et al., 2014b; Smith et al., 1993). The value of I_A, as shown in equation (5.1), is measured as the ratio of observed variance (V_O) and expected variance (V_E) in genetic distance between samples (Agapow and Burt, 2001; Smith et al., 1993).

$$I_A = \frac{V_O}{V_E} - 1 \tag{5.1}$$

The expected variance is practically modeled as the sum of the variances over *m* loci: $V_E = \sum^m var_j$ (Agapow and Burt, 2001; Haubold et al., 1998). If the differences between samples are randomly distributed (linkage equilibrium), we can expect the value of I_A to be zero (Agapow and Burt, 2001; Smith et al., 1993). Under scenarios of non-random mating (e.g. population structure or clonal reproduction), the observed variance would be greater than the expected due to a multi-modal distribution of

distances, and I_A would be greater than zero (Agapow and Burt, 2001; Milgroom, 2015; Smith et al., 1993). Agapow and Burt (2001) noted that this metric does not have an upper limit and increases with the number of loci. To correct this, they developed $\bar{r}_d$ (equation (5.2)), which has a similar structure to a correlation coefficient and ranges from 0 (no linkage) to 1 (complete linkage).

$$\begin{aligned} \bar{r}_d &= \frac{\sum\sum cov_{j,k}}{\sum\sum \sqrt{var_j \cdot var_k}} \\ &= \frac{V_O - V_E}{2\sum\sum \sqrt{var_j \cdot var_k}} \end{aligned} \tag{5.2}$$

de Meeûs and Balloux (2004) investigated the effect of increasing levels of sexual reproduction on $\bar{r}_d$ (noted in their publication as $\bar{r}_D$). They found that very little (1%) sexual reproduction is required to produce a value of $\bar{r}_d$ close to zero. This indicates that $\bar{r}_d$ alone might not be well suited as a measure of clonal reproduction. Prugnolle and de Meeûs (2010) tested the effect of sampling design on $\bar{r}_d$, finding that the its value is drastically reduced when clones from multiple populations are sampled, leading to an over-estimation of the level of recombination.

These studies laid the groundwork for understanding the behavior of $\bar{r}_d$ under different scenarios of non-random mating in diploid organisms, but there were some limitations in available technology that prevented deep analysis from being performed. For both studies, the only software available for calculation of $\bar{r}_d$ for diploid organisms was MULTILOCUS, which could only take one data set at a time (Agapow and Burt, 2001; de Meeûs and Balloux, 2004; Kamvar et al., 2014b; Prugnolle and de Meeûs,

2010). This constrained the researchers to only analyze a minimal set of populations (20) per scenario.

Since the distribution of I_A and $\bar{r}_d$ are not known, the safest way to test for significance are random permutation tests. These tests effectively create unlinked populations by shuffling individuals at each locus, independently and re-calculating I_A and $\bar{r}_d$ (Agapow and Burt, 2001; Haubold et al., 1998; Smith et al., 1993). A one-sided *t*-test of significance was then used to see if the observed statistic was greater than the observed distribution.

While significance testing is available in MULTILOCUS in the form of random permutations, it's computationally expensive, and can only take one data set at a time (Agapow and Burt, 2001; Kamvar et al., 2014b). As a result, power analysis of $\bar{r}_d$ to detect clonal reproduction has not yet been performed (de Meeûs and Balloux, 2004). Our current study aims to evaluate the power and specificity of assessing $\bar{r}_d$ via permutation analysis to detect non-random mating.

In the years since the studies conducted by de Meeûs and Balloux (2004) and Prugnolle and de Meeûs (2010), reduced-representation, high-throughput sequencing methods such as Genotyping-By-Sequencing (GBS) and RAD-seq have rapidly become popular tools for population genetic analysis (Davey and Blaxter, 2010; Davey et al., 2011; Elshire et al., 2011). These methods have the capability to generate thousands of unlinked markers at a fraction of the cost and time necessary to develop high quality microsatellite markers. These marker systems are also prone to missing data and high error rates (Mastretta-Yanes et al., 2014). The index of association was developed for multiple loci in a time when obtaining even 100 unlinked markers posed a significant

challenge. With the advent of these current technologies, how does marker choice and genotyping error affect the index of association?

In 2014, we developed the R package *poppr* for analysis of clonal populations, removing the limitations of data input and computational expense of analyzing the index of association, and in 2015, we expanded this to analysis of genome- wide SNP data (Kamvar et al., 2015b, 2014b; R Core Team, 2016). With these tools we expand on previous studies by asking how sample size, marker choice, clone-correction, and the assumption of homogeneous mutation rates affect our ability to detect clonal reproduction in diploid populations. Our objectives to answer these questions are to (1) re-analyze $\bar{r}_d$ against increasing rates of sexual reproduction and different levels of population mixture in both microsatellite and SNP data sets, (2) perform a power analysis of $\bar{r}_d$ to assess sensitivity and specificity, (3) assess how genotypic and allelic evenness and diversity affects $\bar{r}_d$. Because studies have observed significantly negative values of the I_A and $\bar{r}_d$ (p $\geq$ 0.95), we additionally seek to determine what factors result in negative $\bar{r}_d$ values.

5.3 Methods

Initial sets of simulations were created for different levels of sexual reproduction for each marker type. All simulations were performed with the python package simuPOP version 1.1.7 in python version 3.4(Peng and Amos, 2008). For each scenario, 100 simulations with 10 replicates were created with a census size of 10,000 diploid individuals with equal mating type proportions evolved over 10,000 generations. The simulated

populations were first stored in the native simuPOP format and then transferred to feather format using the python and R packages *feather* version 0.3.0 for downstream analyses. Microsatellite simulations were performed on Ubuntu Linux version 14.04; SNP simulations were performed on CentOS Linux version 6.8. During downstream analysis, 10, 25, 50, and 100 individuals were sampled without replacement for each replicate in R version 3.2 with the package *poppr* version 2.2.1 (Kamvar et al., 2015b, 2014b; R Core Team, 2016). Analyses (described below) were performed on both full and clone-corrected data sets. All downstream analyses were run on the OSU CGRB Core Computing Facility (supplementary information).

5.3.1 Simulating Microsatellite Loci

Each population was simulated with 21 co-dominant, unlinked loci containing 6 to 10 alleles per locus with frequencies drawn from a uniform distribution and subsequently normalized. Before mating, mutations occurred at each locus at a rate of 1e-5 mutations/generation with the exception of the first locus, at which the mutation rate was set to 1e-3. All mutations were applied in a stepwise manner using the `StepwiseMutator()` operator in simuPOP.

5.3.2 GBS Simulations

Simulations of 10,000 binary loci spread evenly over 10 chromosomal fragments were simulated with a mutation rate of 1e-5 mutations per generation for forward and

backward mutations using the `SNPMutator()` operator and and a recombination rate of 1e-5 between adjacent loci using the `Recombinator()` operator in simuPOP.

5.3.3 Mating

Simulations of sexual reproduction were conducted at 10 rates of sexual reproduction on a log scale (0.0, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.1, 0.5, 1.0) reflecting the fraction of individuals in generation *t+1* produced via sexual reproduction. One to three offspring could be produced at each mating event. For sexual events, two parents were chosen randomly from the population with the `RandomSelection()` operator and offspring genotypes were created via the `MendelianGenoTransmitter()` operator. The clonal fraction was created by randomly sampling individuals from the population and duplicating their genotypes with the `CloneGenoTransmitter()` operator. If one mating type was lost before 10,000 generations, the simulation would continue to completion with only clonal reproduction.

5.3.4 Analysis of Microsatellite Data

The standardized index of association ($\bar{r}_d$) was calculated for full and clone-corrected data using the *poppr* function `ia()`. Tests for significance were performed by randomly permuting the alleles at each locus independently and then assessing $\bar{r}_d$. This was done 999 times for each replicate population. This was done for both full and clone-corrected data. The p-values reflect the proportion of observations greater than

the observed statistic. Estimates of genotypic diversity were assessed with the *poppr* function `diversity_boot()` with 999 bootstrap replicates, recording the estimate and variance. The genotypic diversity statistics we calculated were Shannon's Index, $H = -\sum p_i ln p_i$ (Shannon, 2001), Stoddart and Taylor's Index, $G = 1/\sum p_i^2$ (Stoddart and Taylor, 1988), and Evenness, $E_5 = (G-1)/(e^H - 1)$ (Pielou, 1975) where p_i is the frequency of the *i*th genotype.

We additionally Nei's 1978 expected heterozygosity, also known as gene diversity (Nei, 1978), calculated mean allelic evenness $E_{5A} = (1/m)\sum E_{5l}$ where *m* is the number of loci and E_{5l} is the evenness of the alleles at locus *l*.

5.3.5 Analysis of SNP Data

Because GBS data are associated with high error rates, we additionally wanted to assess the effect of missing data on analysis. To do this, we used scripts written for Kamvar et al. (2015b) to randomly insert missing data via the `pop_NA()` function (Kamvar et al., 2015a) at rates of 1%, 5%, and 10% each.

The overall value of $\bar{r}_d$ was calculated for each simulation with the *poppr* function `bitwise.ia()`. Significance was first assessed by randomly shuffling genotypes at each locus independently and then, to preserve existing background linkage structure, at each chromosome independently. This was done 999 times for each replicate population. P-values represent the proportion of random samples greater than or equal to the observed statistic.

5.3.6 Power Analysis

Permutation analysis of the index of association is commonly used as a diagnostic for detecting non-random mating. If the observed value of $\bar{r}_d$ is greater than 95% of the results from the permutations, then the population is declared to be clonal. Standard practice for analyzing microbial populations is to perform this test on both the whole data set (wd) and clone-corrected data (cc), where each multilocus genotype is represented only once per population to avoid signatures of linkage that arise from re-sampling the same individual. (Goss et al., 2014; McDonald, 1997; Milgroom, 1996). If the p-value for $\bar{r}_d$ is significant after clone-correction, then the population is considered clonal. Because knowing a pathogen's mode of reproduction can affect management strategies, it's important to know how sample size, rate of sexual reproduction, mutation rate homogeneity, and clone-correction affect this method's sensitivity and specificity.

The Receiver Operating Characteristic (ROC) curve is a method of assessing the the balance between sensitivity and specificity of a diagnostic method (Metz, 1978). This is done by simultaneously assessing the true positive fraction of tests to a false positive fraction along a gradient of thresholds of increasing leniency. Briefly, if a method has perfect explanatory power, the area under the ROC curve will be equal to 1. If a method has no explanatory power, the area under the ROC curve will be equal to 0.5. We used this method to assess the efficacy of $\bar{r}_d$ to detect non-random mating.

To calculate the ROC curve, we first have to define defined what a positive value is. We define it as a significant value of $\bar{r}_d$ based on a threshold, α, where α is a value

in [0, 1]. Therefore, any randomly mating population (sexual reproduction is equal to one) with a significant value of $\bar{r}_d$ is considered a false positive, and subsequently, any non-randomly mating population with a significant value of $\bar{r}_d$ is considered a true positive (Table 5.1).

Table 5.1: Definitions of false positive and true positive values for ROC analysis of simulations. *p* is the p-value for $\bar{r}_d$, α is a threshold value in [0, 1] Random Mating populations are simulated with a sex rate of 1. Non-Random Mating populations are simulated with a sex rate less than one.

Reproductive Mode	$p \leq \alpha$	$p > \alpha$
Non-Random Mating	True Positive	False Negative
Random Mating	False Positive	True Negative

We constructed each curve by assessing the ROC over values of α in increments of 0.01 from 0 to 1 and plotting the true positive fraction on the y axis and false positive fraction on the x axis. Curves were calculated hierarchically by rate of sexual reproduction (< 1) and unique seed used to generate the populations. For each hierarchical level, separate curves were calculated for sample size, mutation rate, and clone-correction. The area under the ROC curves was calculated using the `auc()` function in the R package *flux* version 0.3-0 (Jurasinski et al., 2014).

5.4 Results

5.4.1 Microsatellite Data

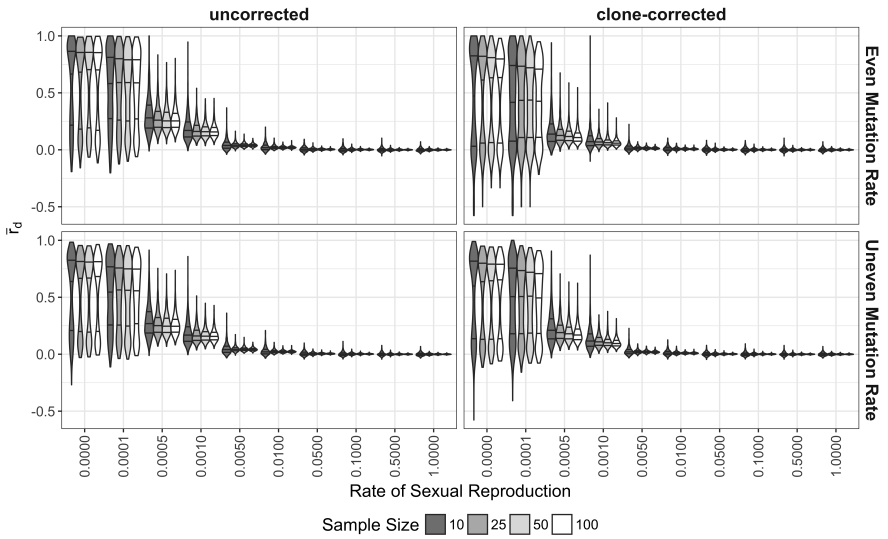


Figure 5.1: Violin plots showing the decay of the Index of Association measured with declining rates of sexual reproduction for data sets simulated with even and uneven mutation rates over 20 and 21 loci, respectively. $\bar{r}_d$ was calculated for both whole and clone corrected data sets. Each violin plot contains 1000 unique data sets. Black lines mark the 25, 50, and 75th percentile.

5.4.2 SNP Data

Because of the computational intensity of the simulations we were only able to run 24 unique seeds over all rates of sexual reproduction with 10 replicates per seed. Because some replicates failed to save, we randomly sampled five replicates per seed per rate

of sexual reproduction, giving us a total of 1200 total populations for analysis.

Bibliography

Adamack, A. T., and Gruber, B. (2014). PopGenReport: Simplifying basic population genetic analyses in r. *Methods in Ecology and Evolution* 5, 384–387. doi:10.1111/2041-210x.12158.

Agapow, P.-M., and Burt, A. (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes* 1, 101–102. doi:10.1046/j.1471-8278.2000.00014.x.

Ali, S., Soubeyrand, S., Gladieux, P., Giraud, T., Leconte, M., Gautier, A., et al. (2016). Cloncase: Estimation of sex frequency and effective population size by clonemate resampling in partially clonal organisms. *Molecular Ecology Resources*. doi:10.1111/1755-0998.12511.

Anderson, J. B., and Kohn, L. M. (1995). Clonality in soilborne, plant-pathogenic fungi. *Annual review of phytopathology* 33, 369–391.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician* 27, 17. doi:10.2307/2682899.

Arnaud-Hanod, S., Duarte, C. M., Alberto, F., and Serrão, E. A. (2007). Standardizing methods to address clonality in population studies. *Molecular Ecology* 16, 5115–5139. doi:10.1111/j.1365-294X.2007.03535.x.

Balloux, F., Lehmann, L., and de Meeûs, T. (2003). The population genetics of clonal and partially clonal diploids. *Genetics* 164, 1635–1644.

Baxter, S. M., Day, S. W., Fetrow, J. S., and Reisinger, S. J. (2006). Scientific software development is not an oxymoron. *PLoS Comput Biol* 2, e87. doi:10.1371/journal.pcbi.0020087.

Bivand, R., Keitt, T., and Rowlingson, B. (2014). *rgdal: Bindings for the Geospatial Data Abstraction Library*. Available at: https://CRAN.R-project.org/package=rgdal.

Brown, A., Feldman, M., and Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96, 523–536. Available at: http:

//www.genetics.org/content/96/2/523.abstract.

Brownstein, M. J., Carpten, J. D., and Smith, J. R. (1996). Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *BioTechniques* 20, 1004–6, 1008–10.

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

Buckheit, J. B., and Donoho, D. L. (1995). "WaveLab and reproducible research," in *Wavelets and statistics* (Springer), 55–81. doi:10.1007/978-1-4612-2544-7_5.

Burt, A., Carter, D. A., Koenig, G. L., White, T. J., and Taylor, J. W. (1996). Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proceedings of the National Academy of Sciences* 93, 770–773.

Canty, A., and Ripley, B. D. (2015). *Boot: Bootstrap r (s-plus) functions*. Available at: https://CRAN.R-project.org/package=boot.

Chakarov, N., Linke, B., Boerner, M., Goesmann, A., Krüger, O., and Hoffman, J. I. (2015). Apparent vector-mediated parent-to-offspring transmission in an avian malaria-like parasite. *Molecular ecology* 24, 1355–1363.

Clark, L., and Jasieniuk, M. (2011). Polysat: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources* 11, 562–566. doi:10.1111/j.1755-0998.2011.02985.x.

Cooke, D. E. L., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J., et al. (2012). Genome analyses of an aggressive and invasive lineage of the irish potato famine pathogen. *PLoS Pathog* 8, e1002940. doi:10.1371/journal.ppat.1002940.

Croucher, P. J. P., Mascheretti, S., and Garbelotto, M. (2013). Combining field epidemiological information and genetic data to comprehensively reconstruct the invasion history and the microevolution of the sudden oak death agent *Phytophthora ramorum* (Stramenopila: Oomycetes) in California. *Biological Invasions* 15, 2281–2297. doi:10.1007/s10530-013-0453-8.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network

research. *InterJournal* Complex Systems, 1695. Available at: http://igraph.org.

Dagum, L., and Menon, R. (1998). OpenMP: An industry standard API for shared-memory programming. *Computational Science & Engineering, IEEE* 5, 46–55.

Davey, J. W., and Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics* 9, 416–423. doi:10.1093/bfgp/elq031.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12, 499–510. doi:10.1038/nrg3012.

de Meeûs, T., and Balloux, F. (2004). Clonal reproduction and linkage disequilibrium in diploids: A simulation study. *Infection, Genetics and Evolution* 4, 345–351. doi:10.1016/j.meegid.2004.05.002.

de Meeûs, T., Lehmann, L., and Balloux, F. (2006). Molecular epidemiology of clonal diploids: a quick overview and a short DIY (do it yourself) notice. *Infection, Genetics and Evolution* 6, 163–170. doi:10.1016/j.meegid.2005.02.004.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* 75, 87–91.

Dray, S., and Dufour, A.-B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* 22. doi:10.18637/jss.v022.i04.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6, e19379. doi:10.1371/journal.pone.0019379.

Everhart, S. E., and Scherm, H. (2015). Fine-scale genetic structure of *Monilinia fructicola* during brown rot epidemics within individual peach tree canopies. *Phytopathology* 105, 542–549. doi:10.1094/phyto-03-14-0088-r.

Everhart, S. E., Tabima, J. F., and Grünwald, N. J. (2014). "*Phytophthora ramorum*," in *Genomics of plant-associated fungi and oomycetes: Dicot pathogens* (Berlin, Heidelberg: Springer), 159–174. doi:10.1007/978-3-662-44056-8_8.

Excoffier, L., and Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Reviews Genetics* 7, 745–758.

doi:10.1038/nrg1904.

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131, 479–91.

Eyre, C. A., Kozanitas, M., and Garbelotto, M. (2013). Population Dynamics of Aerial and Terrestrial Populations of *Phytophthora ramorum* in a California Forest Under Different Climatic Conditions. *Phytopathology* 103, 1141–1152. doi:10.1094/phyto-11-12-0290-r.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. Available at: http://www.genetics.org/content/164/4/1567.abstract.

Felsenstein, J. (1989). PHYLIP-phylogeny inference package (version 3.2). *cladistics* 5, 163–166.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review* 98, 254–267. doi:10.1037/0033-295x.98.2.254.

Goecks, J., Nekrutenko, A., Taylor, J., and Team, T. G. (2010). Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11, R86. doi:10.1186/gb-2010-11-8-r86.

Goss, E. M., Larsen, M., Chastagner, G. A., Givens, D. R., and Grünwald, N. J. (2009). Population genetic analysis infers migration pathways of *Phytophthora ramorum* in US nurseries. *PLoS Pathog* 5, e1000583. doi:10.1371/journal.ppat.1000583.

Goss, E. M., Larsen, M., Vercauteren, A., Werres, S., Heungens, K., and Grünwald, N. J. (2011). *Phytophthora ramorum* in Canada: Evidence for Migration Within North America and from Europe. *Phytopathology* 101, 166–171. doi:10.1094/phyto-05-10-0133.

Goss, E. M., Tabima, J. F., Cooke, D. E., Restrepo, S., Fry, W. E., Forbes, G. A., et al. (2014). The Irish potato famine pathogen *Phytophthora infestans* originated in central Mexico rather than the Andes. *Proceedings of the National Academy of Sciences* 111, 8791–8796.

Goudet, J. (1995). FSTAT (version 1.2): A computer program to calculate f-statistics. *Journal of heredity* 86, 485–486. Available at: http://jhered.oxfordjournals.

org/content/86/6/485.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* 5, 184–186.

Gross, A., Hosoya, T., and Queloz, V. (2014). Population structure of the invasive forest pathogen *Hymenoscyphus pseudoalbidus*. *Molecular ecology* 23, 2943–2960.

Grünwald, N. J., and Goss, E. M. (2011). Evolution and Population Genetics of Exotic and Re-Emerging Pathogens: Novel Tools and Approaches. *Annu. Rev. Phytopathol.* 49, 249–267. doi:10.1146/annurev-phyto-072910-095246.

Grünwald, N. J., and Hoheisel, G.-A. (2006). Hierarchical analysis of diversity, selfing, and genetic differentiation in populations of the oomycete *Aphanomyces euteiches*. *Phytopathology* 96, 1134–1141.

Grünwald, N. J., Garbelotto, M., Goss, E. M., Heungens, K., and Prospero, S. (2012). Emergence of the sudden oak death pathogen *Phytophthora ramorum*. *Trends in Microbiology* 20, 131–138. doi:10.1016/j.tim.2011.12.006.

Grünwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. (2003). Analysis of Genotypic Diversity Data for Populations of Microorganisms. *Phytopathology* 93, 738–746. doi:10.1094/phyto.2003.93.6.738.

Grünwald, N. J., Goss, E. M., and Press, C. M. (2008a). *Phytophthora ramorum*: a pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals. *Molecular Plant Pathology* 9, 729–740. doi:10.1111/j.1364-3703.2008.00500.x.

Grünwald, N. J., Goss, E. M., Ivors, K., Garbelotto, M., Martin, F. N., Prospero, S., et al. (2009). Standardizing the Nomenclature for Clonal Lineages of the Sudden Oak Death Pathogen, *Phytophthora ramorum*. *Phytopathology* 99, 792–795. doi:10.1094/phyto-99-7-0792.

Grünwald, N. J., Kitner, M., McDonald, V., and Goss, E. M. (2008b). Susceptibility in *Viburnum* to *Phytophthora ramorum*. *Plant Disease* 92, 210–214. doi:10.1094/pdis-92-2-0210.

Grünwald, N. J., Martin, F. N., Larsen, M. M., Sullivan, C. M., Press, C. M., Coffey, M. D., et al. (2011). Phytophthora-ID.org: a sequence-based *Phytophthora*

identification tool. *Plant Disease* 95, 337–342.

Hansen, E. M., Kanaskie, A., Prospero, S., McWilliams, M., Goheen, E. M., Osterbauer, N., et al. (2008). Epidemiology of *Phytophthora ramorum* in Oregon tanoak forests. *Canadian Journal of Forest Research* 38, 1133–1143. doi:10.1139/x07-217.

Hartl, D. L., and Clark, A. G. (2007). *Principles of population genetics*. Sunderland, MA, USA: Sinauer Assoc.

Haubold, B., and Hudson, R. R. (2000). LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics* 16, 847–849. doi:10.1093/bioinformatics/16.9.847.

Haubold, B., Travisano, M., Rainey, P. B., and Hudson, R. R. (1998). Detecting linkage disequilibrium in bacterial populations. *Genetics* 150, 1341–8.

Heck, K. L., van Belle, G., and Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* 56, 1459–1461. doi:10.2307/1934716.

Heitman, J., Sun, S., and James, T. Y. (2012). Evolution of fungal sexual reproduction. *Mycologia* 105, 1–27. doi:10.3852/12-253.

Hurlbert, S. H. (1971). The nonconcept of species diversity: A critique and alternative parameters. *Ecology* 52, 577–586. doi:10.2307/1934145.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A., et al. (2008). Repeatability of published microarray gene expression analyses. *Nature Genetics* 41, 149–155. doi:10.1038/ng.295.

Ivors, K., Garbelotto, M., Vries, I. D. E., Ruyter-spira, C., Hekkert, B. T., Rosenzweig, N., 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. doi:10.1111/j.1365-294x.2006.02864.x.

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi:10.1093/bioinformatics/btn129.

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations.

BMC Genet 11, 94. doi:10.1186/1471-2156-11-94.

Jurasinski, G., Koebsch, F., Guenther, A., and Beetz, S. (2014). *flux: Flux rate calculation from dynamic closed chamber measurements*. Available at: https://CRAN.R-project.org/package=flux.

Kahle, D., and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal* 5, 144–161. Available at: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Kamoun, S., Furzer, O., Jones, J. D. G., Judelson, H. S., Ali, G. S., Dalio, R. J. D., et al. (2014). The top 10 oomycete pathogens in molecular plant pathology. *Molecular Plant Pathology* 16, 413–434. doi:10.1111/mpp.12190.

Kamvar, Z. N., Brooks, J. C., and Grunwald, N. J. (2015a). Supplementary Material for Frontiers Plant Genetics and Genomics 'Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality'. doi:10.5281/zenodo.17424.

Kamvar, Z. N., Brooks, J. C., and Grünwald, N. J. (2015b). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics* 6. doi:10.3389/fgene.2015.00208.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. (2014a). Sudden_Oak_Death_in_Oregon_Forests: Spatial and temporal population dynamics of the sudden oak death epidemic in Oregon Forests. doi:10.5281/zenodo.13007.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. (2015c). Spatial and temporal analysis of populations of the sudden oak death pathogen in oregon forests. *Phytopathology* 105, 982–989. doi:10.1094/phyto-12-14-0350-fi.

Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014b). Poppr : an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2, e281. doi:10.7717/peerj.281.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. University of Chicago Press doi:10.7208/chicago/9780226458106.001.0001.

Laloë, D., Jombart, T., Dufour, A.-B., and Moazami-Goudarzi, K. (2007). Consensus genetic structuring and typological value of markers using multiple co-inertia

analysis. *Genetics Selection Evolution* 39, 1–23.

Lees, A., Wattier, R., Shaw, D., Sullivan, L., Williams, N., and Cooke, D. (2006). Novel microsatellite markers for the analysis of *Phytophthora infestans* populations. *Plant Pathology* 55, 311–319.

Li, Y., Cooke, D. E., Jacobsen, E., and Lee, T. van der (2013). Efficient multiplex simple sequence repeat genotyping of the oomycete plant pathogen *Phytophthora infestans*. *Journal of microbiological methods* 92, 316–322.

Linde, C., Zhan, J., and McDonald, B. (2002). Population structure of *Mycosphaerella graminicola*: From lesions to continents. *Phytopathology* 92, 946–955.

Ludwig, J. A., and Reynolds, J. F. (1988). *Statistical ecology: A primer in methods and computing*. John Wiley & Sons.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* 4, 981–994.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research* 27, 209–220.

Mascheretti, S., Croucher, P. J. P., Kozanitas, M., Baker, L., and Garbelotto, M. (2009). Genetic epidemiology of the Sudden Oak Death pathogen *Phytophthora ramorum* in California. *Molecular Ecology* 18, 4577–4590. doi:10.1111/j.1365-294x.2009.04379.x.

Mascheretti, S., Croucher, P. J. P., Vettraino, A., Prospero, S., and 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. doi:10.1111/j.1365-294x.2008.03773.x.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., and Emerson, B. C. (2014). Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Molecular Ecology Resources* 15, 28–41. doi:10.1111/1755-0998.12291.

McDonald, B. A. (1997). The population genetics of fungi: tools and techniques. *Phytopathology* 87, 448–453.

McDonald, B. A., and Linde, C. (2002). The population genetics of plant

pathogens and breeding strategies for durable resistance. *Euphytica* 124, 163–180. doi:10.1023/A:1015678432355.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., et al. (2016). How open science helps researchers succeed. *eLife* 5. doi:10.7554/elife.16800.

Meirmans, P. G., and Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: Two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4, 792–794. doi:10.1111/j.1471-8286.2004.00770.x.

Metz, C. E. (1978). Basic principles of ROC analysis. in *Seminars in nuclear medicine* (Elsevier), 283–298. doi:10.1016/S0001-2998(78)80014-2.

Metzger, M. J., Reinisch, C., Sherry, J., and Goff, S. P. (2015). Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell* 161, 255–263.

Michalakis, Y., and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142, 1061–1064.

Milgroom, M. G. (1996). Recombination and the multilocus structure of fungal populations. *Annual review of phytopathology* 34, 457–477.

Milgroom, M. G. (2015). *Population biology of plant pathogens: Genetics, ecology, and evolution*. 3340 Pilot Knob Road, St. Paul, MN 55121, USA: APS Press.

Milgroom, M. G., Levin, S. A., and Fry, W. E. (1989). Population genetics theory and fungicide resistance. *Plant disease epidemiology* 2, 340–367.

Milgroom, M., and Fry, W. (1997). "Contributions of population genetics to plant disease epidemiology and management," in *Advances in botanical research* (Elsevier BV), 1–30. doi:10.1016/s0065-2296(08)60069-5.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* 70, 3321–3323.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89, 583–590. Available at: http://www.genetics.org/content/89/3/583.abstract.

Nielsen, R., and Slatkin, M. (2013). *An introduction to population genetics: Theory and applications*. Sinauer Associates, Incorporated Available at: http://books.

google.com/books?id=Iy08kgEACAAJ.

Nieuwenhuis, B. P. S., and James, T. Y. (2016). The frequency of sex in fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 20150540. doi:10.1098/rstb.2015.0540.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *Vegan: Community ecology package*. Available at: https://CRAN.R-project.org/package=vegan.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2015). *Vegan: Community ecology package*. Available at: http://CRAN.R-project.org/package=vegan.

Orive, M. E. (1993). Effective population size in organisms with complex life-histories. *Theoretical Population Biology* 44, 316–340. doi:10.1006/tpbi.1993.1031.

Ouzounis, C. A., and Valencia, A. (2003). Early bioinformatics: The birth of a discipline–a personal view. *Bioinformatics* 19, 2176–2190. doi:10.1093/bioinformatics/btg309.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in r language. *Bioinformatics* 20, 289–290. doi:10.1093/bioinformatics/btg412.

Partridge, D., Lopez, P. D., and Johnston, V. S. (1984). Computer programs as theories in biology. *Journal of Theoretical Biology* 108, 539–564. doi:10.1016/s0022-5193(84)80079-x.

Peakall, R., and Smouse, P. E. (2006). GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6, 288–295.

Peever, T. L., and Milgroom, M. G. (1994). Genetic structure of *Pyrenophora teres* populations determined with random amplified polymorphic DNA markers. *Canadian Journal of Botany* 72, 915–923.

Peng, B., and Amos, C. I. (2008). Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics* 24, 1408–1409. doi:10.1093/bioinformatics/btn179.

Pielou, E. (1975). Ecological diversity.

Poucke, K. V., Franceschini, S., Webber, J. F., Vercauteren, A., Turner, J. A., McCracken, A. R., et al. (2012). Discovery of a fourth evolutionary lineage of *Phytophthora ramorum*: EU2. *Fungal Biology* 116, 1178–1191.

doi:10.1016/j.funbio.2012.09.003.

Prevosti, A., Ocaña, J., and Alonso, G. (1975). Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. *Theoretical and Applied Genetics* 45, 231–241.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Prospero, S., Black, J. A., and Winton, L. M. (2004). Isolation and characterization of microsatellite markers in *Phytophthora ramorum*, the causal agent of sudden oak death. *Molecular Ecology Notes* 4, 672–674. doi:10.1111/j.1471-8286.2004.00778.x.

Prospero, S., Grünwald, N. J., Winton, L. M., and Hansen, E. M. (2009). Migration Patterns of the Emerging Plant Pathogen *Phytophthora ramorum* on the West Coast of the United States of America. *Phytopathology* 99, 739–749. doi:10.1094/phyto-99-6-0739.

Prospero, S., Hansen, E. M., Grünwald, N. J., and 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. doi:10.1111/j.1365-294x.2007.03343.x.

Prugnolle, F., and de Meeûs, T. (2010). Apparent high recombination rates in clonal parasitic organisms due to inappropriate sampling design. *Heredity* 104, 135–140. doi:10.1038/hdy.2009.128.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: http://www.R-project.org/.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.R-project.org/.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: http://www.R-project.org/.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.R-

project.org/.

Rizzo, D. M., Garbelotto, M., and Hansen, E. M. (2005). *Phytophthora ramorum*: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests. *Annu. Rev. Phytopathol.* 43, 309–335. doi:10.1146/annurev.phyto.42.040803.140418.

Rizzo, D. M., Garbelotto, M., Davidson, J. M., Slaughter, G. W., and Koike, S. T. (2002). *Phytophthora ramorum* as the Cause of Extensive Mortality of Quercus spp. and Lithocarpus densiflorus in California. *Plant Disease* 86, 205–214. doi:10.1094/pdis.2002.86.3.205.

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 1118–1123.

Schliep, K. (2011). Phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593.

Searls, D. B. (2010). The roots of bioinformatics. *PLoS Comput Biol* 6, e1000809. doi:10.1371/journal.pcbi.1000809.

Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5, 3–55. Available at: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

Smith, J. M., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences* 90, 4384–4388. doi:10.1073/pnas.90.10.4384.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 38, 1409–1438.

Stoddart, J. A., and Taylor, J. F. (1988). Genotypic diversity: Estimation and prediction in samples. *Genetics* 118, 705–11.

Taylor, J. W., and Fisher, M. C. (2003). Fungal multilocus sequence typing – it's not just for bacteria. *Current opinion in microbiology* 6, 351–356.

Taylor, J. W., Geiser, D. M., Burt, A., and Koufopanou, V. (1999). The evolutionary biology and population genetics underlying fungal strain typing. *Clinical Microbiology Reviews* 12, 126–146.

Vercauteren, A., Dobbelaere, I. D., Grünwald, N. J., Bonants, P., Bockstaele, E. V.,

Maes, M., et al. (2010). Clonal expansion of the Belgian *Phytophthora ramorum* populations based on new microsatellite markers. *Molecular Ecology* 19, 92–107. doi:10.1111/j.1365-294x.2009.04443.x.

Vercauteren, A., Larsen, M., Goss, E., Grunwald, N. J., Maes, M., and Heungens, K. (2011). Identification of new polymorphic microsatellite markers in the NA1 and NA2 lineages of *Phytophthora ramorum*. *Mycologia* 103, 1245–1249. doi:10.3852/10-420.

Weir, B. S., and Cockerham, C. (1996). Genetic data analysis ii: Methods for discrete population genetic data.

Werres, S., Marwitz, R., veld, W. A. M. I., Cock, A. W. D., Bonants, P. J., Weerdt, M. D., et al. (2001). *Phytophthora ramorum sp. nov.*, a new pathogen on Rhododendron and Viburnum. *Mycological Research* 105, 1155–1165. doi:10.1016/s0953-7562(08)61986-3.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York Available at: `http://ggplot2.org`.

Wickham, H., and Chang, W. (2015). *Devtools: Tools to make developing R packages easier*. Available at: `http://CRAN.R-project.org/package=devtools`.

Winton, L. M., and Hansen, E. M. (2001). Molecular diagnosis of *Phytophthora lateralis* in trees, water, and foliage baits using multiplex polymerase chain reaction. *Forest Pathol* 31, 275–283. doi:10.1046/j.1439-0329.2001.00251.x.

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., et al. (2013). The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *eLife* 2, e00731. doi:10.7554/eLife.00731.001.