

Novel tools for analyzing genome-wide data of clonal populations

Zhian N. Kamvar¹, Jonah C. Brooks², Niklaus J. Grunwald^{1,3*}

¹ Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA

² College of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA

³ Horticultural Crops Research Laboratory, USDA-Agricultural Research Service, Corvallis, OR, USA

Correspondence*:

Niklaus J. Grunwald
Horticultural Crops Research Laboratory USDA ARS
3420 NW Orchard Ave.
Corvallis, OR, 97330, grunwalg@science.oregonstate.edu

2 ABSTRACT

To gain a detailed understanding of how plant microbe evolve and adapt to host and other factors such as pesticides, knowledge of the population dynamics and evolutionary history of populations is crucial. With the advent of high throughput sequencing technologies, obtaining genomic sequences for representative populations has become easier than ever before. A move towards open, reproducible science has provided impetus for developing population genetic analysis tools in R. 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 analyses across any level of hierarchies, a new function to define clone boundaries we call `mlg.filter` allowing for inspection and definition of what is a clonal lineage, and the index of association for reduced representation genomic data, and modular bootstrapping of any genetic distance.

Keywords: clonality, population genetics, bootstrap, open source

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 evolve and adapt to crops, fungicides, or other factors, can only emerge in the context of population level phenomena given the demographic history of populations (McDonald and Linde, 2002; Grunwald and Goss, 2011; 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,

28 many plant pathogen populations are at least partially clonal if not completely clonal (Milgroom, 1996;
29 Anderson and Kohn, 1995). Thus, development of tools for analysis of clonal or polyploid populations is
30 needed.

31 Genotyping by sequencing and whole genome resequencing provide the unprecedented ability to
32 identify >1,000 single nucleotide polymorphisms (SNPs) in populations (Elshire et al., 2011; Luikart
33 et al., 2003; Davey et al., 2011). Availability of these large SNP data sets provides new challenges for
34 data analysis. For example, it is not clear what a clone is in large SNP data where the chance of observing
35 variation at a given SNP locus within independent samples of the same clone are substantial enough that
36 novel tools for definition of clone boundaries are required. With traditional marker data (e.g., SSR, AFLP)
37 a clone was typically defined as a unique multilocus genotype (MLG). However, with large SNP data a
38 measure of genetic distance is required to define the boundary of an MLG (e.g., clone) or the boundaries
39 of a clonal lineage. Definition of a clone is further complicated by the presence of missing data that is
40 typical for reduced representation libraries used in GBS or genome re-sequencing. If two individuals are
41 identical for all observed SNPs except for one missing allele, should they be considered different?

42 The research community using the R statistical and computing language (R Core Team, 2015) has
43 developed a plethora of new resources for population genetic analysis (Paradis, 2010; Jombart, 2008).
44 Recently, we introduced the R package *poppr* specifically developed for analysis of clonal populations
45 (Kamvar et al., 2014b). *Poppr* previously introduced several novel features including the ability to conduct
46 a hierarchical analysis across unlimited hierarchies, test for linkage association, graph minimum spanning
47 networks or provide bootstrap support for Bruvo's distance in resulting trees. It was well received by
48 the community, garnering 14 citations in its first year of publication. Since it's first release, however,
49 limitations with speed, ease of use, and efficiency became more apparent as genomic data became more
50 readily available.

51 In version 1.1, to address difficulties with handling hierarchical and multilocus genotypic metadata,
52 a new S4 object called "genclone" was defined to expand the genind object of *aedgenet*. The genclone
53 object formalized the definitions of multilocus genotypes and population hierarchies by adding two slots
54 called mlg and hierarchy that carried a numeric vector and a data frame, respectively. These new slots
55 allow for increased efficiency and ease of use by allowing these metadata to travel with the genetic data.
56 The addition of the population hierarchies has proved to be advantageous enough that they have recently
57 been adopted into the more central *aedgenet* package (Jombart, 2008).

58 In version 1, *poppr* was appropriate for traditional markers systems, but not well suited to population
59 genomic data resulting from high throughput sequencing methods. The raw size of these data made
60 it difficult to conduct traditional analyses. Here, we introduce *poppr* 2.0, which provides a significant
61 update to *poppr* including novel tools for analysis of clonal populations specifically for large SNP data.
62 Significant novel tools include functions for calculating clone boundaries and collapsing individuals into
63 user-specified clones based on genetic distance, sliding window analyses, genotype accumulation curves,
64 reticulations in minimum spanning networks, and bootstrapping for any genetic distance.

POPULATION STRATA AND HIERARCHIES

65 Assessments of population structure through methods such as hierarchical F_{st} and AMOVA benefit
66 greatly from multiple levels of population definition (Linde et al., 2002; Everhart and Scherm, 2015;
67 Grnwald and Hoheisel, 2006). With clonal organisms, basic practice has been to clone-censor data to
68 avoid downward bias in diversity due to duplicated genotypes that may or may not represent different
69 samples (Milgroom, 1996). Data structures for population genetic data mostly allow for only one level of
70 hierarchical definition. The impetus was placed on the researchers to provide the population hierarchies
71 for every step of the analysis. In *poppr* version 1.1, the hierarchy slot was introduced to allow unlimited
72 population hierarchies or stratifications to travel with the data. In practice, it is stored as a data frame
73 where each column represents a separate hierarchical level. This is then used to set the population factor

74 of the data by supplying a hierarchical formula containing one or more column names of the data frame
75 in the hierarchy slot. This functionality, developed in *poppr*, has been moved to the *adegenet* package in
76 version 2.0 and the slot and methods have been renamed to *strata*.

CLONAL IDENTIFICATION

77 As highlighted in previous work, clone correction is an important component of population genetic
78 analysis of organisms that have cryptic growth or are known to reproduce asexually (Kamvar et al.,
79 2014b; Milgroom, 1996; Grnwald et al., 2003). This method removes bias that would otherwise affect
80 metrics that rely on allele frequencies. It was initially designed for data with only a handful of markers.
81 With the advent of large-scale sequencing and reduced-representation libraries, it has become easier to
82 sequence tens of thousands of markers from hundreds of individuals (Elshire et al., 2011; Davey et al.,
83 2011; Davey and Blaxter, 2010). With this larger number of markers, the genetic resolution is much
84 greater, but the chance of genotyping error is also greatly increased (Mastretta-Yanes et al., 2015). Taking
85 this fact and occasional somatic mutations into account, it would be impossible to separate true clones
86 from independent individuals by just comparing what multilocus genotypes are different. We introduce
87 a new method for collapsing unique multilocus genotypes determined by naive string comparison into
88 multilocus lineages utilizing any genetic distance given three different clustering algorithms: farthest
89 neighbor, nearest neighbor, and UPGMA (average neighbor) (Sokal, 1958).

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

DEMONSTRATION DATA: *P. INFESTANS*

102 We utilize data from the microbe *Phytophthora infestans* to show how the *mlg.filter* function
103 collapses multilocus genotypes with Bruvo's distance assuming a genome addition model (Bruvo et al.,
104 2004). *P. infestans* is the causal agent of potato late blight originating from Mexico and spread to Europe
105 in the mid 19th century (Goss et al., 2014; Li et al., 2013; Lees et al., 2006). *P. infestans* reproduces
106 both clonally and sexually. The clonal lineages of *P. infestans* have been formally defined into 18 separate
107 clonal lineages using a combination of various molecular methods including AFLP and microsatellite
108 markers (Lees et al., 2006). For these data, we used *mlg.filter* to detect all of the distance thresholds
109 at which 18 multilocus lineages would be resolved. We used these thresholds to define multilocus lineages
110 and create contingency tables and dendograms to determine how well the multilocus lineages were
111 detected.

112 For the *P. infestans* population, the three algorithms were able to detect 18 multilocus lineages at
113 different distance thresholds (Fig. 1). Contingency tables between the described multilocus genotypes
114 and the genotypes defined by distance show that most of the 18 lineages were resolved, except for US-8,
115 which is polytomic (Table 1).

DEMONSTRATION DATA: SIMULATED DATA

116 We utilized simulated data constructed using the `glSim` function in *adegenet* (Jombart and Ahmed, 2011)
117 to obtain a SNP data set for demonstration. Two diploid data sets were created, each with 10k SNPs (25%
118 structured into two groups) and 200 samples with 10 ancestral populations of even sizes. Clones were
119 created in one data set by marking each sample with a unique identifier and then randomly sampling
120 with replacement. It is well documented that reduced- representation sequencing can introduce several
121 erroneous calls and missing data (Mastretta-Yanes et al., 2015). To reflect this, we mutated SNPs at a rate
122 of 10% and inserted an average of 10% missing data for each sample after clones were created, ensuring
123 that no two sequences were alike. The number of mutations and missing data per sample were determined
124 by sampling from a poisson distribution with $\lambda = 1000$. After pooling, 20% of the data set was randomly
125 sampled for analysis. Genetic distance was obtained with the function `bitwise.dist`, which calculates
126 the fraction of different sites between samples, counting missing data as equivalent in comparison.

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

131 Out of the 100 simulations run, we found that across all methods, detection of duplicated samples had
132 $\sim 98\%$ true positive fraction and $\sim 0.8\%$ false positive fraction indicating that this method is robust to
133 simulated populations.

MINIMUM SPANNING NETWORKS WITH RETICULATION

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

142 To demonstrate the utility of minimum spanning networks with reticulation, we used two clonal data
143 sets: H3N2 flu virus data from the *adegenet* package using years of each epidemic as the population
144 factor, and *Phytophthora ramorum* data from Nurseries and Oregon forests (Jombart et al., 2010; Kamvar
145 et al., 2014a). Minimum spanning networks were created with and without reticulation using the *poppr*
146 functions `diss.dist` and `bruvo.msn` for the H3N2 and *P. ramorum* data, respectively (Kamvar et
147 al., 2014b; Bruvo et al., 2004). To detect mlg clusters, the infoMAP community detection algorithm was
148 applied with 10,000 trials as implemented in the R package *igraph* version 0.7.1 utilizing genetic distance
149 as edge weights and number of samples in each MLG as vertex weights (Csardi and Nepusz, 2006; Rosvall
150 and Bergstrom, 2008).

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

157 The infoMAP algorithm revealed 63 communities with a maximum community size of 77 and $H = 3.56$
158 for the reticulate network of the H3N2 data and 117 communities with a maximum community size of

159 26 and $H = 4.65$ for the minimum spanning tree. The entropy across years was greatly decreased for all
 160 populations with the reticulate network compared to the minimum spanning tree (Fig. 2).

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

BOOTSTRAPPING

168 Calculating genetic distance for among samples and populations is very important method for assessing
 169 population differentiation through methods such as G_{st} , AMOVA, and Mantel tests (Nei, 1973; Excoffier
 170 et al., 1992; Mantel, 1967). Confidence in distance metrics is related to the confidence in the markers to
 171 accurately represent the diversity of the data. Especially true with microsatellite markers, a single hyper-
 172 diverse locus can make a population appear to have more diversity based on genetic distance. Using a
 173 bootstrapping procedure of randomly sampling loci with replacement when calculating a distance matrix
 174 gives confidence in hierarchical clustering. Because genetic data in a genind object is represented as a
 175 matrix with samples in rows and alleles in columns, bootstrapping is a non-trivial task as all alleles in
 176 a single locus need to be sampled together. To remedy this, we have created an internal S4 class called
 177 “bootgen”, which extends the internal “gen” class from *aedgenet*. This class can be created from any
 178 genind, genclone, or genpop object, and allows loci to be sampled with replacement. To further facilitate
 179 bootstrapping, a function called *aboot*, which stands for “any boot”, is introduced that will bootstrap
 180 any genclone, genind, or genpop object with any genetic distance that can be calculated from it.

181 To demonstrate calculating a dendrogram with bootstrap support, we used the *poppr* function *aboot*
 182 on population allelic frequencies derived from the data set *microbov* in the *aedgenet* package with 1000
 183 bootstrap replicates (Jombart, 2008; Lalo et al., 2007). The resulting dendrogram shows bootstrap support
 184 values > 50% (Fig. 3).

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

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

GENOTYPE ACCUMULATION CURVE

185 Analysis of population genetics of clonal organisms often borrows from ecological methods such as
 186 analysis of diversity within populations (Milgroom, 1996; Arnaud-Hanod et al., 2007; Grnwald et al.,
 187 2003). When choosing markers for analysis, it is important to make sure that the observed diversity in your
 188 sample will not appreciably increase if an additional marker is added (Arnaud-Hanod et al., 2007). This
 189 concept is analogous to a species accumulation curve, obtained by rarefaction. The genotype accumulation
 190 curve in *poppr* is implemented in the function *genotype_curve*. The curve is constructed by randomly
 191 sampling x loci and counting the number of observed MLGs. This repeated r times for 1 locus up to $n - 1$
 192 loci, creating $n - 1$ distributions of observed MLGs.

193 The following code example demonstrates the genotype accumulation curve for data from Everhart and
 194 Scherm (2015) showing that these data reach a small plateau and have a greatly decreased variance with
 195 12 markers, indicating that there are enough markers such that adding more markers to the analysis will
 196 not create very many new genotypes (Fig. 4).

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

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

INDEX OF ASSOCIATION

197 The index of association (I_A) is a measure of multilocus linkage disequilibrium that is most often used
 198 to detect clonal reproduction within organisms that have the ability to reproduce via sexual or asexual
 199 processes (Brown et al., 1980; Smith et al., 1993; Milgroom, 1996). It was standardized in 2001 as \bar{r}_d
 200 by Agapow and Burt (2001) to address the issue of scaling with increasing number of loci. This metric is
 201 typically applied to traditional dominant and co-dominant markers such as AFLPs, SNPs, or microsatellite
 202 markers. With the advent of high throughput sequencing, SNP data is now available in a genome-wide
 203 context and in very large matrices including thousands of SNPs. Thus, the likelihood of finding mutations
 204 within two individuals of a given clone increases and tools are needed for defining clone boundaries. For
 205 this reason, we devised two approaches using the index of association for large numbers of markers typical
 206 for population genomic studies.

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

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

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

AVAILABILITY

224 As of this writing, the `poppr` R package version 2.0 containing all of the features described here is located
 225 at <https://github.com/grunwaldlab/poppr>. It is necessary to install `aedegenet` 2.0 before

226 installing *poppr*. It can be found at <https://github.com/thibautjombart/adegenet>. Both
227 of these can be installed via the R package *devtools* (Wickham and Chang, 2015):

```
library("devtools")
install_github("thibautjombart/adegenet")
install_github("grunwaldlab/poppr")
```

DISCUSSION

228 We have presented here new model-free tools for the analysis of clonal populations with emphasis on
229 genomic-scale data. Especially important is `mlg.filter`, which

230 Creating structures like minimum spanning networks and dendrograms allow researchers to distill the
231 most important information from large distance matrices, revealing patterns that could support hypotheses
232 of differentiation or the lack thereof. Bifurcating dendrograms are most familiar to biologists as the
233 interpretation of them is straightforward and bootstrap confidence values can easily be obtained due to the
234 basic structure of the tree. Minimum spanning networks allow for a different view into populations, where
235 samples themselves can be treated as internal nodes connecting other samples, which could effectively
236 describe populations sampled through time. The drawback to these is that there is no clear method for a
237 bootstrap procedure to obtain confidence intervals.

238 Reticulate minimum spanning networks are very important for clonal organisms where a minimum
239 spanning tree would become a chain, implying that the clones were derived in a progressive and linear
240 fashion. This presents but one potential scenario for clonal organisms, but does not account for any other
241 biologically relevant process. Reticulations in the minimum spanning networks allow for a representation
242 of uncertainty that goes along with clonal organisms. The current implementation in *poppr* has been
243 successfully used in analyses such as reconstruction of the *P. ramorum* epidemic in Curry County, OR
244 (Kamvar et al., 2014a, 2015). Reticulated networks also allow for the application of graph community
245 detection algorithms such as the infoMAP algorithm (Rosvall and Bergstrom, 2008). As shown in the *P.*
246 *ramorum* and H3N2 data, while it is possible to utilize these graph walking algorithms on non-reticulate
247 minimum spanning trees, the results derived from these are limited to explain populations derived from
248 serial cloning events.

- 249
- bootstrapping methods encourage future developers to write distance implementations in common
250 format
 - moving towards open source, modular tools is the direction that population genetics and plant
251 pathology needs to go.
- 252

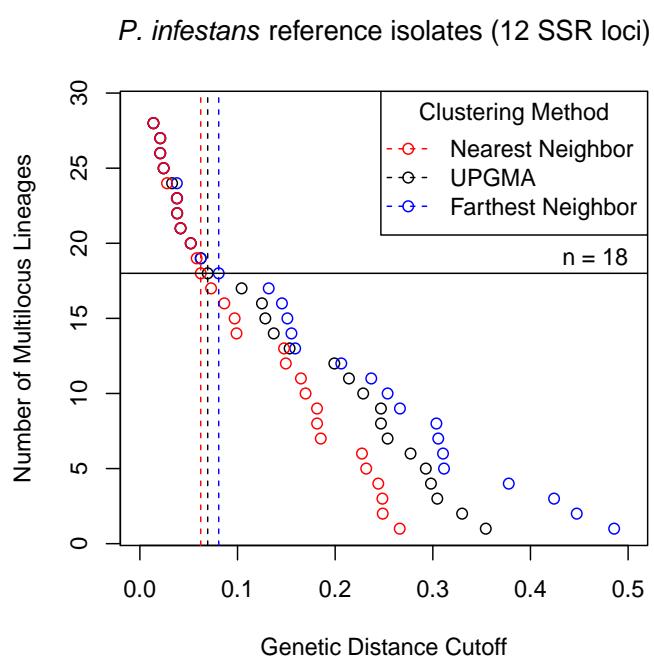
FIGURES AND TABLES**FIGURE 1**

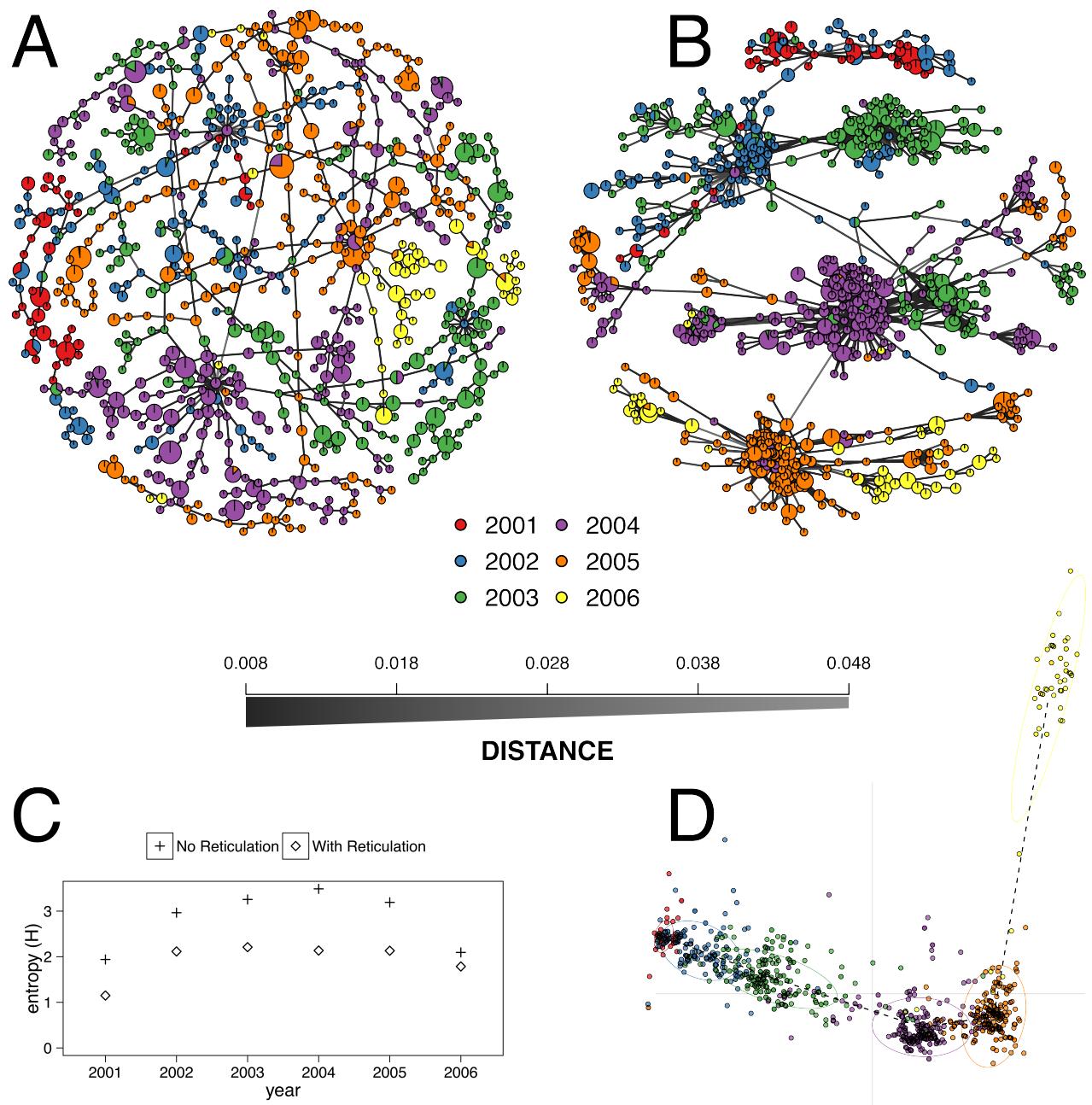
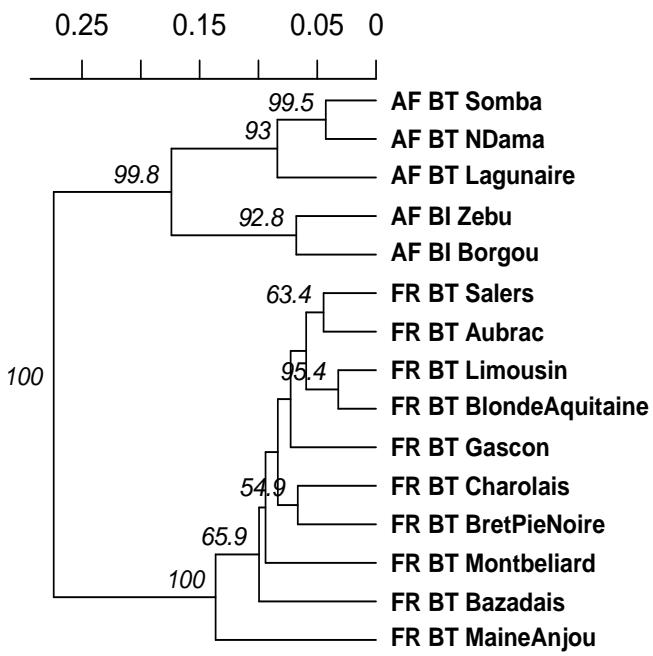
FIGURE 2

FIGURE 3

255

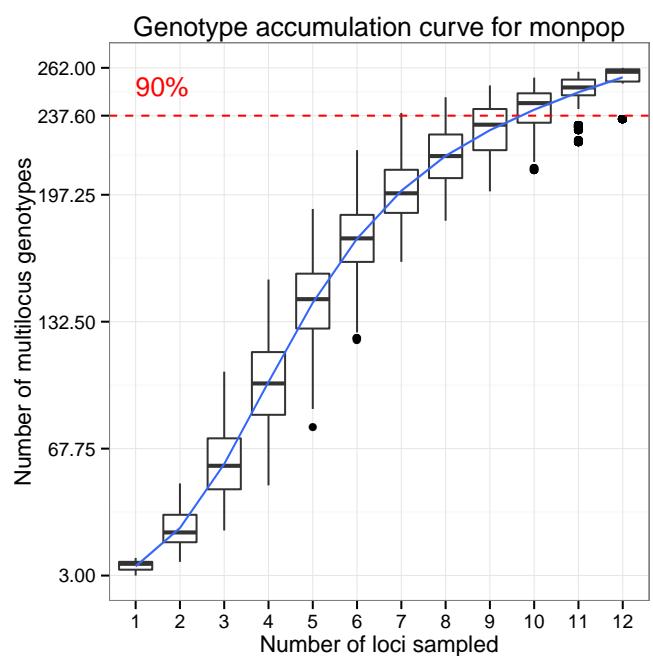
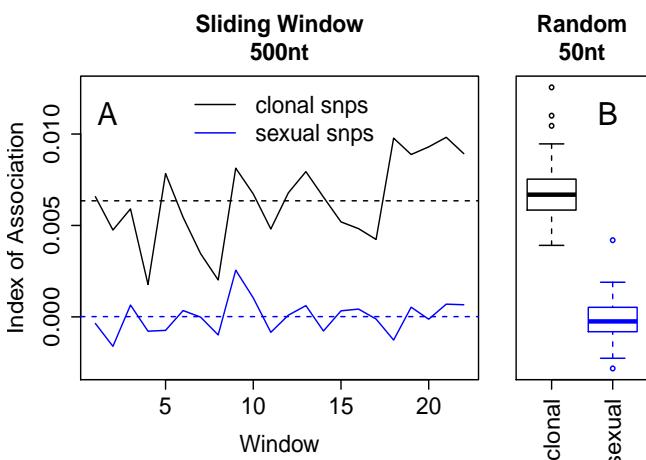
FIGURE 4

FIGURE 5

257

TABLE 1

| | 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 | . | . | 2 | . | . |
| US-22 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| US-23 | . | . | . | . | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . |
| US-24 | . | . | 1 | 1 | . | 2 | . | . | . | . | . | . | . | . | . | . | . | . |
| US-8 | . | . | 1 | 1 | . | 2 | . | . | . | . | . | . | . | . | . | . | . | . |

FIGURE AND TABLE LEGENDS**FIGURE 1**

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

FIGURE 2

264 (A-B) Minimum spanning networks of the hemagglutinin (HA) segment of H3N2 viral DNA from the
265 *adegenet* package representing flu epidemics from 2001 to 2006 with (B) and without (A) reticulations
266 (Jombart, 2008; Jombart et al., 2010). Each node represents a unique multilocus genotype, colors represent
267 epidemic year, and edge color represents absolute genetic distance. (C) Shannon entropy values for
268 population assignments compared with communities determined by the infoMAP algorithm on (A) and
269 (B). (D) Graphic reproduced from Jombart et al. (2010) showing that the 2006 epidemic does not cluster
270 neatly with the other years.

FIGURE 3

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

FIGURE 4

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

FIGURE 5

280 (A) Sliding window analysis of the standardized index of association (\bar{r}_d) across a simulated 1.1×10^4 nt
281 chromosome containing 1,100 variants among 100 individuals. Each window analyzed variants within
282 100nt chunks. The black line indicates clonal population, the blue line indicates sexual. (B) 100 random
283 samplings of 50 variants to calculate a distribution of \bar{r}_d for the clonal (black) and sexual (blue) population.
284 (A) and (B) are plotted on the same y-axis.

TABLE 1

285 Contingency table comparing multilocus lineages assigned based on average neighbor clustering
286 (columns) vs. multilocus lineages defined in Li et al. (2013) and Lees et al. (2006).

REFERENCES

- 287 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.
- 288 Anderson, J. B., and Kohn, L. M. (1995). Clonality in soilborne, plant-pathogenic fungi. *Annual review of phytopathology* 33, 369–391.
- 291 Arnaud-Hanod, S., Duarte, C. M., Alberto, F., and Serro, E. A. (2007). Standardizing methods to address
292 clonality in population studies. *Molecular Ecology* 16, 5115–5139.
- 293 Brown, A., Feldman, M., and Nevo, E. (1980). MULTILocus sTRUCTURE oF nATURAL
294 pOPULATIONS oF *Hordeum spontaneum*. *Genetics* 96, 523–536. Available at: <http://www.genetics.org/content/96/2/523.abstract>.

- 296 Bruvo, R., Michiels, N. K., D'Souza, T. G., and Schulenburg, H. (2004). A simple method for the
297 calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology* 13, 2101–
298 2106.
- 299 Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research.
300 *InterJournal Complex Systems*, 1695. Available at: <http://igraph.org>.
- 301 Dagum, L., and Menon, R. (1998). OpenMP: An industry standard aPI for shared-memory
302 programming. *Computational Science & Engineering, IEEE* 5, 46–55.
- 303 Davey, J. W., and Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in
304 Functional Genomics* 9, 416–423. doi:10.1093/bfgp/elq031.
- 305 Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L.
306 (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature
307 Reviews Genetics* 12, 499–510.
- 308 Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American
309 Biology Teacher* 75, 87–91.
- 310 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E.
311 (2011). A robust, simple genotyping-by-sequencing (gBS) approach for high diversity species. *PloS one*
312 6, e19379.
- 313 Everhart, S., and Scherm, H. (2015). Fine-scale genetic structure of *Monilinia fructicola* during brown
314 rot epidemics within individual peach tree canopies. *Phytopathology* 105, 542–549.
- 315 Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from
316 metric distances among dNA haplotypes: Application to human mitochondrial dNA restriction data.
317 *Genetics* 131, 479–491.
- 318 Goss, E. M., Tabima, J. F., Cooke, D. E., Restrepo, S., Fry, W. E., Forbes, G. A., Fieland, V. J., Cardenas,
319 M., and Grnwald, N. J. (2014). The irish potato famine pathogen *Phytophthora infestans* originated in
320 central mexico rather than the andes. *Proceedings of the National Academy of Sciences* 111, 8791–8796.
- 321 Grunwald, N. J., and Goss, E. M. (2011). Evolution and population genetics of exotic and re-emerging
322 pathogens: Novel tools and approaches. *Annual Review of Phytopathology* 49, 249–267.
- 323 Grnwald, N. J., and Hoheisel, G.-A. (2006). Hierarchical analysis of diversity, selfing, and genetic
324 differentiation in populations of the oomycete aphanomyces euteiches. *Phytopathology* 96, 1134–1141.
- 325 Grnwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. (2003). Analysis of genotypic
326 diversity data for populations of microorganisms. *Phytopathology* 93, 738–46. Available at: <http://apsjournals.apsnet.org/doi/abs/10.1094/PHYTO.2003.93.6.738>.
- 327 Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers.
329 *Bioinformatics* 24, 1403–1405. doi:10.1093/bioinformatics/btn129.
- 330 Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide sNP
331 data. *Bioinformatics* 27, 3070–3071.
- 332 Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A
333 new method for the analysis of genetically structured populations. *BMC genetics* 11, 94.
- 334 Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grnwald, N. J. (2014a).
335 Sudden_Oak_Death_in_Oregon_Forests: Spatial and temporal population dynamics of the sudden oak
336 death epidemic in Oregon Forests. doi:10.5281/zenodo.13007.
- 337 Kamvar, Z. N., Larsen, M. M., Kanaskie, A., Hansen, E., and Grnwald, N. J. (2015). Spatial and
338 temporal analysis of populations of the sudden oak death pathogen in oregon forests. *Phytopathology*, in
339 press.

- 340 Kamvar, Z. N., Tabima, J. F., and Grnwald, N. J. (2014b). Poppr: An r package for genetic analysis of
341 populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2, e281.
- 342 Lalo, D., Jombart, T., Dufour, A.-B., and Moazami-Goudarzi, K. (2007). Consensus genetic structuring
343 and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution* 39, 1–23.
- 344 Lees, A., Wattier, R., Shaw, D., Sullivan, L., Williams, N., and Cooke, D. (2006). Novel microsatellite
345 markers for the analysis of phytophthora infestans populations. *Plant Pathology* 55, 311–319.
- 346 Li, Y., Cooke, D. E., Jacobsen, E., and Lee, T. van der (2013). Efficient multiplex simple sequence repeat
347 genotyping of the oomycete plant pathogen phytophthora infestans. *Journal of microbiological methods*
348 92, 316–322.
- 349 Linde, C., Zhan, J., and McDonald, B. (2002). Population structure of mycosphaerella graminicola:
350 From lesions to continents. *Phytopathology* 92, 946–955.
- 351 Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of
352 population genomics: From genotyping to genome typing. *Nature Reviews Genetics* 4, 981–994.
- 353 Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer*
354 research 27, 209–220.
- 355 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piero, D., and Emerson, B.
356 (2015). Restriction site-associated dNA sequencing, genotyping error estimation and de novo assembly
357 optimization for population genetic inference. *Molecular ecology resources* 15, 28–41.
- 358 McDonald, B. A., and Linde, C. (2002). The population genetics of plant pathogens and breeding
359 strategies for durable resistance. *Euphytica* 124, 163–180. doi:10.1023/A:1015678432355.
- 360 Milgroom, M. G. (1996). Recombination and the multilocus structure of fungal populations. *Annual*
361 *review of phytopathology* 34, 457–477.
- 362 Milgroom, M. G., Levin, S. A., and Fry, W. E. (1989). Population genetics theory and fungicide
363 resistance. *Plant disease epidemiology* 2, 340–367.
- 364 Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National*
365 *Academy of Sciences* 70, 3321–3323.
- 366 Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L.,
367 Solymos, P., Stevens, M. H. H., and Wagner, H. (2015). Vegan: Community ecology package. Available
368 at: <http://CRAN.R-project.org/package=vegan>.
- 369 Paradis, E. (2010). Pegas: an R package for population genetics with an integrated-modular approach.
370 *Bioinformatics* 26, 419–420.
- 371 R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R
372 Foundation for Statistical Computing Available at: <http://www.R-project.org/>.
- 373 Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal
374 community structure. *Proceedings of the National Academy of Sciences* 105, 1118–1123.
- 375 Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing*
376 *and Communications Review* 5, 3–55. Available at: <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>.
- 377 Smith, J. M., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria?
378 *Proceedings of the National Academy of Sciences* 90, 4384–4388. doi:10.1073/pnas.90.10.4384.
- 379 Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 38,
380 1409–1438.

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