1

Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality

Zhian N. Kamvar [1], **Jonah C. Brooks** [2], **Niklaus J. Grünwald** [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. Grünwald
Horticultural Crops Research Laboratory USDA ARS
3420 NW Orchard Ave.
Corvallis, OR, 97330, `grunwaln@science.oregonstate.edu`

2 **ABSTRACT**

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

15 **Keywords: clonality, population genomics, bootstrap, index of association, hierarchical analysis, sliding window**

INTRODUCTION

16 To paraphrase Dobzhansky, nothing in the field of plant-microbe interactions makes sense except in
17 the light of population genetics (Dobzhansky, 1973). Genetic forces such as selection and drift act on
18 alleles in a population. Thus, a true understanding of how plant pathogens emerge, evolve and adapt to
19 crops, fungicides, or other factors, can only be elucidated in the context of population level phenomena
20 given the demographic history of populations (McDonald and Linde, 2002; Grünwald and Goss, 2011;
21 Milgroom et al., 1989). The field of population genetics, in the era of whole genome resequencing,
22 provides unprecedented power to describe the evolutionary history and population processes that drive

23 coevolution between pathogens and hosts. This powerful field thus critically enables effective deployment
24 of R genes, design of pathogen informed plant resistance breeding programs, and implementation of
25 fungicide rotations that minimize emergence of resistance.

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

32   Genotyping by sequencing and whole genome resequencing provide the unprecedented ability to
33 identify thousands of single nucleotide polymorphisms (SNPs) in populations (Elshire et al., 2011; Luikart
34 et al., 2003; Davey et al., 2011). With traditional marker data (e.g., SSR, AFLP) a clone was typically
35 defined as a unique multilocus genotype (MLG) (Grünwald and Hoheisel, 2006; Falush et al., 2003; Goss
36 et al., 2009; Cooke et al., 2012; Taylor and Fisher, 2003). Availability of large SNP data sets provides new
37 challenges for data analysis. These data are based on reduced representation libraries and high throughput
38 sequencing with moderate sequencing depth which invariably results in substantial missing data, error in
39 SNP calling due to sequencing error, lack of read depth or other sources of spurious allele calls (Mastretta-
40 Yanes et al., 2015). It is thus not clear what a clone is in large SNP data sets and novel tools are required
41 for definition of clone boundaries.

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

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

IMPLEMENTATIONS AND EXAMPLES

CLONAL IDENTIFICATION

60 As highlighted in previous work, clone correction is an important component of population genetic
61 analysis of organisms that are known to reproduce asexually (Kamvar et al., 2014b; Milgroom, 1996;
62 Grünwald et al., 2003). This method is a partial correction for bias that affects metrics that rely on
63 allele frequencies assuming panmixia and was initially designed for data with only a handful of markers.
64 With the advent of large-scale sequencing and reduced- representation libraries, it has become easier
65 to sequence tens of thousands of markers from hundreds of individuals (Elshire et al., 2011; Davey
66 et al., 2011; Davey and Blaxter, 2010). With this larger number of markers, the genetic resolution is
67 much greater, but the chance of genotyping error is also greatly increased and missing data is frequent

68 (Mastretta-Yanes et al., 2015). Taking this fact and occasional somatic mutations into account, it would
69 be impossible to separate true clones from independent individuals by just comparing what MLGs are
70 different. We introduce a new method for collapsing unique multilocus genotypes determined by naive
71 string comparison into multilocus lineages utilizing any genetic distance given three different clustering
72 algorithms: farthest neighbor, nearest neighbor, and ~~UPGMA~~ (Unweighted Pair Group Method with
73 Arithmetic Mean (UPGMA, average neighbor) (Sokal, 1958).

74 These clustering algorithms act on a distance matrix that is either provided by the user or generated via
75 a function that will calculate a distance from genetic data such as `bruvo.dist`, which in particular
76 applies to any level of ploidy (Bruvo et al., 2004). All algorithms have been implemented in C and
77 utilize the OpenMP framework for optional parallel processing (Dagum and Menon, 1998). Default is the
78 conservative farthest neighbor algorithm (Fig. 1A), which will only cluster samples together if all samples
79 in the cluster are at a distance less than the given threshold. By contrast, the nearest neighbor algorithm
80 will have a chaining effect that will cluster samples akin to adding links on a chain where a sample can
81 be included in a cluster if all of the samples have at least one connection below a given threshold (Fig.
82 1C). The UPGMA, or average neighbor clustering algorithm is the one most familiar to biologists as it is
83 often used to generate ultra-metric trees based on genetic distance (Fig. 1B). This algorithm will cluster by
84 creating a representative sample per cluster and joining clusters if these representative samples are closer
85 than the given threshold.

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

96 For the *P. infestans* population, the three algorithms were able to detect 18 multilocus lineages at
97 different distance thresholds (Fig. 2). Contingency tables between the described multilocus genotypes
98 and the genotypes defined by distance show that most of the 18 lineages were resolved, except for US-8,
99 which is polytomic (Table 1).

100 We utilized simulated data to evaluate the effect of sequencing error and missing data on MLG calling.
101 We constructed the data using the `glSim` function in *adegenet* (Jombart and Ahmed, 2011) to obtain a
102 SNP data set for demonstration. Two diploid data sets were created, each with 10k SNPs (25% structured
103 into two groups) and 200 samples with 10 ancestral populations of even sizes. Clones were created in one
104 data set by marking each sample with a unique identifier and then randomly sampling with replacement.
105 It is well documented that reduced- representation sequencing can introduce several erroneous calls and
106 missing data (Mastretta-Yanes et al., 2015). To reflect this, we mutated SNPs at a rate of 10% and inserted
107 an average of 10% missing data for each sample after clones were created, ensuring that no two sequences
108 were alike. The number of mutations and missing data per sample were determined by sampling from
109 a Poisson distribution with $\lambda = 1000$. After pooling, 20% of the data set was randomly sampled for
110 analysis. Genetic distance was obtained with the function `bitwise.dist`, which calculates the fraction
111 of different sites between samples equivalent to Provesti's distance, counting missing data as equivalent
112 in comparison (Prevosti et al., 1975).

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

117   Out of the 100 simulations run, we found that across all methods, detection of duplicated samples had
118 $\sim 98\%$ true positive fraction and $\sim 0.8\%$ false positive fraction indicating that this method is robust to
119 simulated populations (supplementary materials[1]).

MINIMUM SPANNING NETWORKS WITH RETICULATION

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

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

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

143   The infoMAP algorithm revealed 63 communities with a maximum community size of 77 and $H = 3.56$
144 for the reticulate network of the H3N2 data and 117 communities with a maximum community size of
145 26 and $H = 4.65$ for the minimum spanning tree. The entropy across years was greatly decreased for
146 all populations with the reticulate network compared to the minimum spanning tree (Fig. 3). Note that
147 the reticulated network (Fig. 3B) showed patterns corresponding with those resulting from a discriminant
148 analysis of principal components (Fig. 3D) (Jombart et al., 2010).

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

BOOTSTRAPPING

156 Assessing population differentiation through methods such as G_{st}, AMOVA, and Mantel tests relies on
157 comparing samples within and across populations (Nei, 1973; Excoffier et al., 1992; Mantel, 1967).

---

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

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

---

158 Confidence in distance metrics is related to the confidence in the markers to accurately represent the
159 diversity of the data. Especially true with microsatellite markers, a single hyper-diverse locus can make a
160 population appear to have more diversity based on genetic distance. Using a bootstrapping procedure of
161 randomly sampling loci with replacement when calculating a distance matrix provides support for clades
162 in hierarchical clustering.

163   Data in genind and genpop objects are represented as matrices with individuals in rows and alleles in
164 columns (Jombart, 2008). This gives the advantage of being able to use R's matrix algebra capabilities
165 to efficiently calculate genetic distance. Unfortunately, this also means that bootstrapping is a non- trivial
166 task as all alleles at a single locus need to be sampled together. To remedy this, we have created an internal
167 S4 class called "bootgen", which extends the internal "gen" class from *adegenet*. This class can be created
168 from any genind, genclone, or genpop object, and allows loci to be sampled with replacement. To further
169 facilitate bootstrapping, a function called `aboot`, which stands for "any boot", is introduced that will
170 bootstrap any genclone, genind, or genpop object with any genetic distance that can be calculated from it.

171   To demonstrate calculating a dendrogram with bootstrap support, we used the *poppr* function `aboot`
172 on population allelic frequencies derived from the data set `microbov` in the *adegenet* package with
173 1000 bootstrap replicates (Jombart, 2008; Laloë et al., 2007). The resulting dendrogram shows bootstrap
174 support values $> 50\%$ (Fig. 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)
```

GENOTYPE ACCUMULATION CURVE

175 Analysis of population genetics of clonal organisms often borrows from ecological methods such as
176 analysis of diversity within populations (Milgroom, 1996; Arnaud-Hanod et al., 2007; Grünwald et al.,
177 2003). When choosing markers for analysis, it is important to make sure that the observed diversity in your
178 sample will not appreciably increase if an additional marker is added (Arnaud-Hanod et al., 2007). This
179 concept is analogous to a species accumulation curve, obtained by rarefaction. The genotype accumulation
180 curve in *poppr* is implemented in the function `genotype_curve`. The curve is constructed by randomly
181 sampling x loci and counting the number of observed MLGs. This repeated r times for 1 locus up to $n-1$
182 loci, creating $n-1$ distributions of observed MLGs.

183   The following code example demonstrates the genotype accumulation curve for data from Everhart and
184 Scherm (2015) showing that these data reach a small plateau and have a greatly decreased variance with
185 12 markers, indicating that there are enough markers such that adding more markers to the analysis will
186 not create very many new genotypes (Fig. 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
```

INDEX OF ASSOCIATION

187 The index of association (I_A) is a measure of multilocus linkage disequilibrium that is most often used
188 to detect clonal reproduction within organisms that have the ability to reproduce via sexual or asexual
189 processes (Brown et al., 1980; Smith et al., 1993; Milgroom, 1996). It was standardized in 2001 as
190 $\bar{r}_d$ by Agapow and Burt (2001) to address the issue of scaling with increasing number of loci. This
191 metric is typically applied to traditional dominant and co-dominant markers such as AFLPs, SNPs, or
192 microsatellite markers. With the advent of high throughput sequencing, SNP data is now available in a
193 genome-wide context and in very large matrices including thousands of SNPs. For this reason, we devised
194 two approaches using the index of association for large numbers of markers typical for population genomic
195 studies. Both functions utilize *adegenet*'s "genlight" object class, which efficiently stores 8 binary alleles
196 in a single byte (Jombart and Ahmed, 2011). As calculation of the $\bar{r}_d$ requires distance matrices of absolute
197 number of differences, we utilize a function that calculates these distances directly from the compressed
198 data called `bitwise.dist`.

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

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

210   To demonstrate the sliding window and random sampling of $\bar{r}_d$ with respect to clonal populations, we
211 simulated two populations containing 1,100 neutral SNPs for 100 diploid individuals under the same
212 initial seed. One population had individuals randomly sampled with replacement, representing the clonal
213 population. After sampling, both populations had 5% random error and 1% missing data independently
214 propagated across all samples. On average, we obtained a higher value of $\bar{r}_d$ for the clonal population
215 compared to the sexual population for both methods (Fig. 6).

DATA FORMAT UPDATES: POPULATION STRATA AND HIERARCHIES

216 Assessments of population structure through methods such as hierarchical F_{st} (Goudet, 2005) and
217 AMOVA (Michalakis and Excoffier, 1996) require hierarchical sampling of populations across space
218 or time (Linde et al., 2002; Everhart and Scherm, 2015; Grünwald and Hoheisel, 2006). With clonal
219 organisms, basic practice has been to clone-censor data to avoid downward bias in diversity due to
220 duplicated genotypes that may or may not represent different samples (Milgroom, 1996). This correction
221 should be performed with respect to a population hierarchy to accurately reflect the biology of the
222 organism. Traditional data structures for population genetic data in most analysis tools allow for only
223 one level of hierarchical definition. The investigator thus had to provide the data set for analysis at each
224 hierarchical level.

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

232 supplying a hierarchical formula containing one or more column names of the data frame in the hierarchy
233 slot.

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

AVAILABILITY

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

REQUIREMENTS

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

INSTALLATION

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

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

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

DISCUSSION

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

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

---

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

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 1, 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. 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, 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, 2015).

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`.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT

The authors declare no observable conflict of interest.

AUTHOR CONTRIBUTIONS

ZNK and JCB wrote and tested the code. ZNK maintains the code. ZNK and NJG conceived, discussed implications, and wrote the manuscript. NJG coordinated the collaborative effort.

REFERENCES

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.

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

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.

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`.

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.

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.

Cooke, D. E., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J., Deahl, K. L., Farrer, R. A., Gilroy, E. M., Goss, E. M., et al. (2012). Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS pathogens* 8, e1002940.

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.

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

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* 6, e19379.

Everhart, S., and Scherm, H. (2015). Fine-scale genetic structure of *Monilinia fructicola* during brown rot epidemics within individual peach tree canopies. *Phytopathology* 105, 542–549.

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, 479–491.

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`.

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 pathogens* 5, e1000583.

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

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

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

350 Grünwald, N. J., and Goss, E. M. (2011). Evolution and population genetics of exotic and re-emerging
351 pathogens: Novel tools and approaches. *Annual Review of Phytopathology* 49, 249–267.

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

354 Grünwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. (2003). Analysis of genotypic
355 diversity data for populations of microorganisms. *Phytopathology* 93, 738–46. Available at: `http://`
356 `apsjournals.apsnet.org/doi/abs/10.1094/PHYTO.2003.93.6.738`.

357 Grünwald, N. J., Martin, F. N., Larsen, M. M., Sullivan, C. M., Press, C. M., Coffey, M. D., Hansen,
358 E. M., and Parke, J. L. (2011). Phytophthora-ID. org: a sequence-based *Phytophthora* identification tool.
359 *Plant Disease* 95, 337–342.

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

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

364 Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A
365 new method for the analysis of genetically structured populations. *BMC genetics* 11, 94.

366 Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. (2015). Spatial and
367 temporal analysis of populations of the sudden oak death pathogen in Oregon forests. *Phytopathology*, in
368 press.

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

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

374 Laloë, D., Jombart, T., Dufour, A.-B., and Moazami-Goudarzi, K. (2007). Consensus genetic structuring
375 and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution* 39, 1–23.

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

378 Li, Y., Cooke, D. E., Jacobsen, E., and Lee, T. van der (2013). Efficient multiplex simple sequence repeat
379 genotyping of the oomycete plant pathogen *Phytophthora infestans*. *Journal of microbiological methods*
380 92, 316–322.

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

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

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

387 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., and Emerson, B.
388 (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly
389 optimization for population genetic inference. *Molecular ecology resources* 15, 28–41.

390 McDonald, B. A., and Linde, C. (2002). The population genetics of plant pathogens and breeding
391 strategies for durable resistance. *Euphytica* 124, 163–180. doi:10.1023/A:1015678432355.

392 Meirmans, P. G., and Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: Two programs for
393 the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4, 792–794.

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

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

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

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

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

404 Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L.,
405 Solymos, P., Stevens, M. H. H., and Wagner, H. (2015). *Vegan: Community ecology package*. Available
406 at: `http://CRAN.R-project.org/package=vegan`.

407 Prevosti, A., Ocaña, J., and Alonso, G. (1975). Distances between populations of *Drosophila*
408 *subobscura*, based on chromosome arrangement frequencies. *Theoretical and Applied Genetics* 45,
409 231–241.

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

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

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

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

419 Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 38,
420 1409–1438.

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

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

425 Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin,
426 F. N., Kamoun, S., Krause, J., et al. (2013). The rise and fall of the *Phytophthora infestans* lineage that
427 triggered the Irish potato famine. *Elife* 2, e00731.
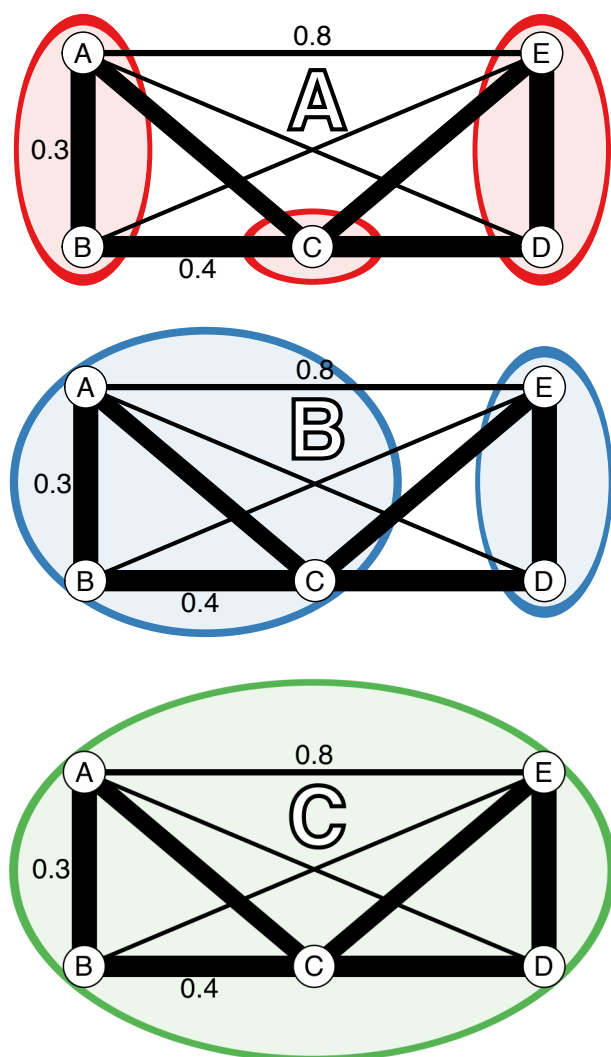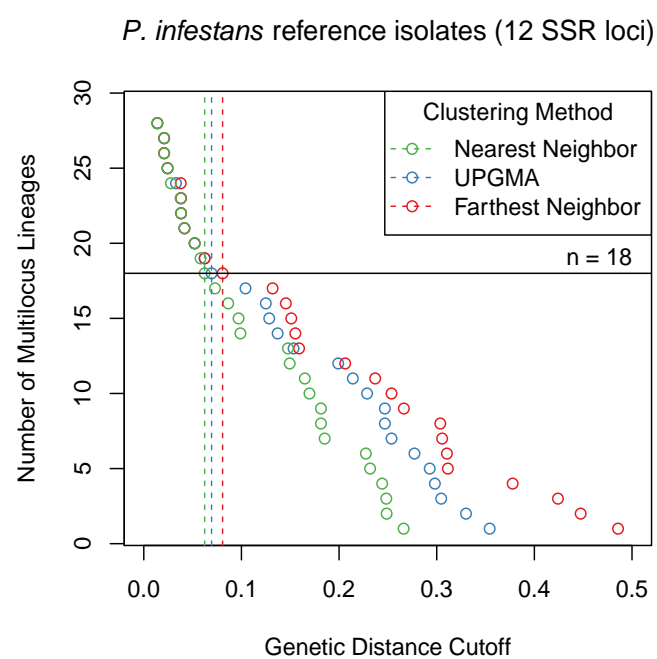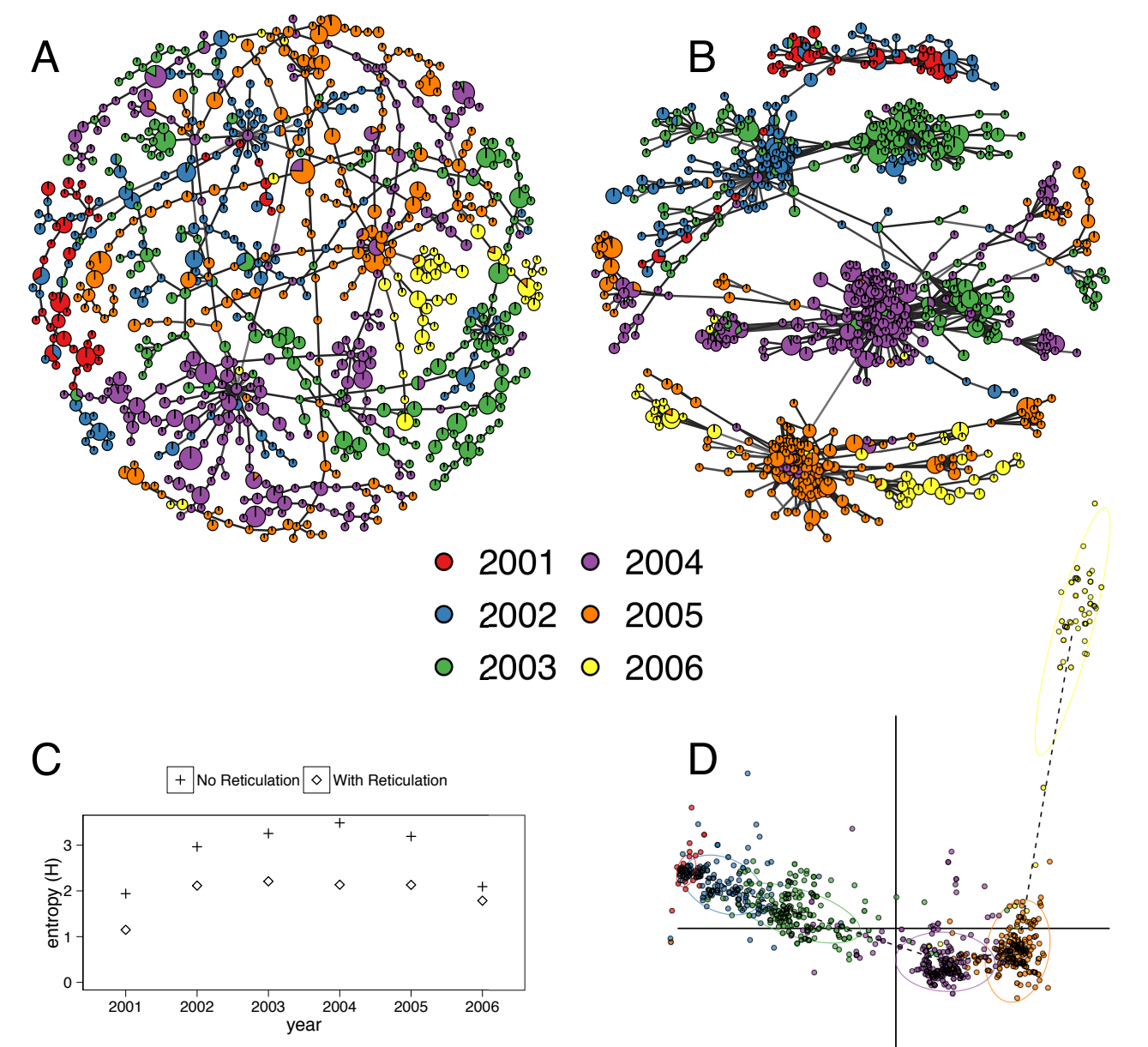
FIGURES AND TABLES

FIGURE 1

FIGURE 2

P. infestans reference isolates (12 SSR loci)

FIGURE 3

2001   2004
2002   2005
2003   2006
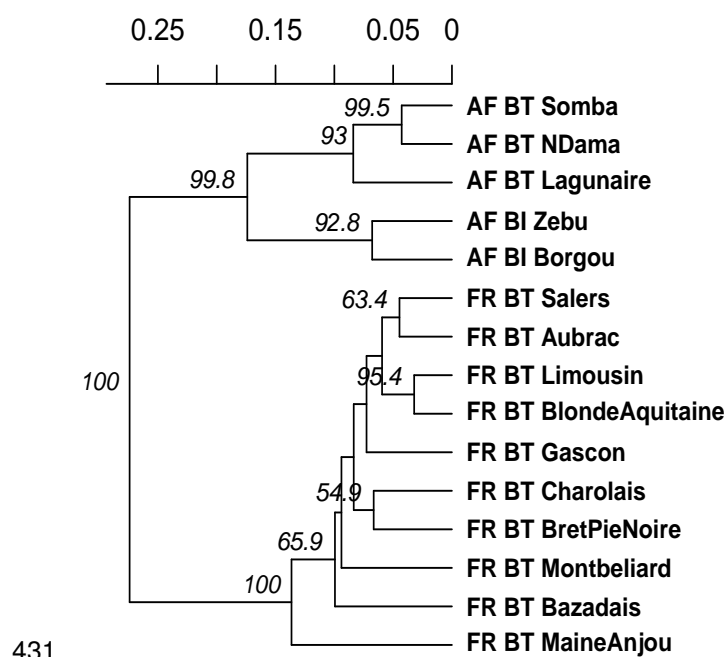
430

FIGURE 4

431

FIGURE 5

432

FIGURE 6

433
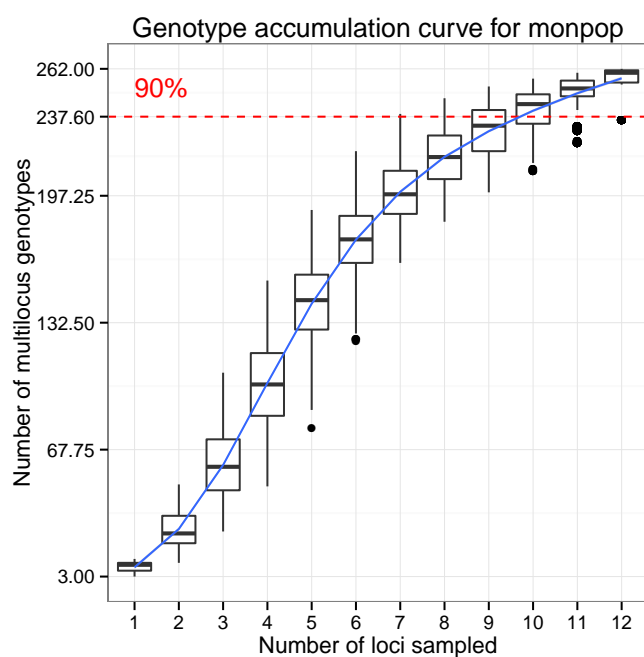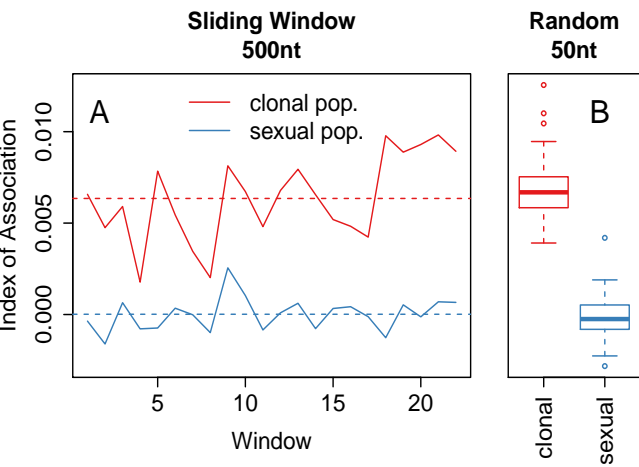
Table 1 Contingency table comparing multilocus lineages ~~assigned based on average neighbor clustering (columns) vs. multilocus lineages~~ (MLL) defined in Li et al. (2013) and Lees et al. (2006) ~~.~~ (rows) to MLLs inferred from Bruvo's genetic distance (columns) at a threshold of 0.07 with the average 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.

P. infestans MLL	Inferred MLL											
	3	4	5	6	8	10	12	15	16	17	18	21
B	.	.	.	.	.	.	.	.	.	.	.	.
C	.	.	.	.	.	.	.	.	.	.	.	.
D.1	.	.	.	.	.	.	.	.	.	.	.	.
D.2	.	.	.	.	.	.	.	.	.	.	.	.
EU-13	.	.	.	.	.	.	.	.	.	1	.	.
EU-4	.	.	.	.	.	.	.	.	.	.	1	.
EU-5	.	.	.	.	.	.	.	.	.	.	.	2
EU-8	.	.	.	.	.	.	1	.	.	.	.	.
US-11	.	.	.	.	.	.	.	.	.	.	.	.
US-12	.	1	.	.	.	.	.	.	.	.	.	.
US-14	.	.	.	.	.	1	.	.	.	.	.	.
US-17	.	.	.	.	.	.	.	.	.	.	.	.
US-20	2	.	.	.	.	.	.	.	.	.	.	.
US-21	.	.	.	.	.	.	.	.	.	.	.	.
US-22	.	.	.	.	.	.	.	.	.	.	.	.
US-23	.	.	.	.	.	.	.	3	.	.	.	.
US-24	.	.	.	.	3	.	.	.	.	.	.	.
US-8	.	.	1	1	.	2	.	.	.	.	.	.

FIGURE LEGENDS

442 **Figure 1** Diagrammatic representation of the three clustering algorithms implemented in `mlg.filter`.
443 (**A-C**) Represent different clustering algorithms on the same imaginary network with a threshold of 0.451.
444 Edge weights are represented in arbitrary units noted by the line thickness and numerical values next to
445 the lines. All outer angles are 90 degrees, so the un-labeled edge weights can be obtained with simply
446 geometry. Colored circles represent clusters of genotypes. (**A**) Farthest neighbor clustering does not cluster
447 nodes B and C because nodes A and C are more than a distance of 0.451 apart. (**B**) UPGMA (average
448 neighbor) clustering clusters nodes A, B, and C together because the average distance between them and
449 C is < 0.451. (**C**) Nearest neighbor clustering clusters all nodes together because the minimum distance
450 between them is always < 0.451.

451 **Figure 2** Graphical representation of three different clustering algorithms collapsing multilocus
452 genotypes for 12 SSR loci from *Phytophthora infestans* representing 18 clonal lineages. The horizontal
453 axis is Bruvo's genetic distance assuming the genome addition model. The vertical axis represents the
454 number of multilocus lineages observed. Each point shows the threshold at which one would observe a
455 given number of multilocus genotypes. The horizontal black line represents 18 multilocus genotypes and
456 vertical dashed lines mark the thresholds used to collapse the multilocus genotypes into 18 multilocus
457 lineages.

458 **Figure 3** (**A-B**) Minimum spanning networks of the hemagglutinin (HA) segment of H3N2 viral DNA
459 from the *adegenet* package representing flu epidemics from 2001 to 2006 without reticulation (**A**) and with
460 reticulation (**B**) (Jombart, 2008; Jombart et al., 2010). Each node represents a unique multilocus genotype,
461 colors represent epidemic year, and edge color represents absolute genetic distance. (**C**) Shannon entropy
462 values for population assignments compared with communities determined by the infoMAP algorithm on
463 (**A**) and (**B**). (**D**) Graphic reproduced from Jombart et al. (2010) showing that the 2006 epidemic does not
464 cluster neatly with the other years via Discriminant Analysis of Principal Components. Horizontal axis
465 represents the first discriminant component. Vertical axis represents the second discriminant component.

466 **Figure 4** UPGMA dendrogram generated from Nei's genetic distance on 15 breeds of *Bos taurus* (BT)
467 or *Bos indicus* (BI) from Africa (AF) or France (FR). These data are from Laloë et al. (2007). Node labels
468 represent bootstrap support > 50% out of 1,000 bootstrap replicates.

469 **Figure 5** Genotype accumulation curve for 694 isolates of the peach brown rot pathogen, *Monilinia*
470 *fructicola* genotyped over 13 loci from Everhart and Scherm (2015). The horizontal axis represents the
471 number of loci randomly sampled without replacement up to $n-1$ loci, the vertical axis shows the number
472 of multilocus genotypes observed, up to 262, the number of unique multilocus genotypes in the data set.
473 The red dashed line represents 90% of the total observed multilocus genotypes. A trendline (blue) has
474 been added using the *ggplot2* function `stat_smooth`.

475 **Figure 6** (**A**) Sliding window analysis of the standardized index of association ($\bar{r}_d$) across a simulated
476 1.1×10^4 nt chromosome containing 1,100 variants among 100 individuals. Each window analyzed
477 variants within 500nt chunks. The black line refers to the clonal and the blue line to the sexual populations.
478 (**B**) boxplots showing 100 random samples of 50 variants to calculate a distribution of $\bar{r}_d$ for the clonal
479 (red) and sexual (blue) populations. Each box is centered around the mean, with whiskers extending out
480 to 1.5 times the interquartile range. The median is indicated by the center line. (**A**) and (**B**) are plotted on
481 the same y-axis.