

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

Zhian N. Kamvar¹, Jonah C. Brooks², Niklaus J. Grnwald^{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. Grnwald
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 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. 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 several new functionalities including the new function `mlg.filter` to define clone boundaries allowing for inspection and definition of what is a clonal lineage, a sliding-window analysis of the index of association, modular bootstrapping of any genetic distance, and analyses across any level of hierarchies.

Keywords: clonality, population genetics, bootstrap, index of association, hierarchical analysis, sliding window

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; Grnwald 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.

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) (Grnwald 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 (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. *Poppr* has been rapidly
48 adopted and applied to a range of studies including for example horizontal transmission in leukemia of
49 clams (Metzger et al., 2015), study of the vector-mediated parent- to-offspring transmission in an avian
50 malaria-like parasite (Chakarov et al., 2015), and characterization of the emergence of the invasive forest
51 pathogen *Hymenoscyphus pseudoalbidus* (Gross et al., 2014).

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

IMPLEMENTATIONS AND EXAMPLES

CLONAL IDENTIFICATION

57 As highlighted in previous work, clone correction is an important component of population genetic
58 analysis of organisms that have cryptic growth or are known to reproduce asexually (Kamvar et al., 2014b;
59 Milgroom, 1996; Grnwald et al., 2003). This method is a partial correction for bias that affects metrics
60 that rely on allele frequencies assuming panmixia. It was initially designed for data with only a handful
61 of markers. With the advent of large-scale sequencing and reduced-representation libraries, it has become
62 easier to sequence tens of thousands of markers from hundreds of individuals (Elshire et al., 2011; Davey
63 et al., 2011; Davey and Blaxter, 2010). With this larger number of markers, the genetic resolution is much
64 greater, but the chance of genotyping error is also greatly increased (Mastretta-Yanes et al., 2015). Taking
65 this fact and occasional somatic mutations into account, it would be impossible to separate true clones
66 from independent individuals by just comparing what multilocus genotypes are different. We introduce
67 a new method for collapsing unique multilocus genotypes determined by naive string comparison into
68 multilocus lineages utilizing any genetic distance given three different clustering algorithms: farthest
69 neighbor, nearest neighbor, and UPGMA (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 genclone objects 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. 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. 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. 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.

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 and spread to Europe in the mid 19th century (Goss et al., 2014; Li et al., 2013; Lees et al., 2006). *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). 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 dendograms 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. 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 1).

We utilized simulated data constructed 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., 2015). 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 fraction indicating that this method is robust to simulated populations.

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 *aedgenet* 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 (Kamvar et al., 2014b; Bruvo et al., 2004). To detect mlg clusters, the infoMAP community detection algorithm was applied with 10,000 trials as implemented in the R package *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).

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 information).

BOOTSTRAPPING

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

164 To demonstrate calculating a dendrogram with bootstrap support, we used the *poppr* function *aboot*
 165 on population allelic frequencies derived from the data set *microbov* in the *adegenet* package with 1000
 166 bootstrap replicates (Jombart, 2008; Lalo et al., 2007). The resulting dendrogram shows bootstrap support
 167 values > 50% (Fig. 4).

```
library("poppr")
data("microbov", package = "adegenet")
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

168 Analysis of population genetics of clonal organisms often borrows from ecological methods such as
 169 analysis of diversity within populations (Milgroom, 1996; Arnaud-Hanod et al., 2007; Grnwald et al.,
 170 2003). When choosing markers for analysis, it is important to make sure that the observed diversity in your
 171 sample will not appreciably increase if an additional marker is added (Arnaud-Hanod et al., 2007). This
 172 concept is analogous to a species accumulation curve, obtained by rarefaction. The genotype accumulation
 173 curve in *poppr* is implemented in the function *genotype_curve*. The curve is constructed by randomly
 174 sampling x loci and counting the number of observed MLGs. This repeated r times for 1 locus up to $n - 1$
 175 loci, creating $n - 1$ distributions of observed MLGs.

176 The following code example demonstrates the genotype accumulation curve for data from Everhart and
 177 Scherm (2015) showing that these data reach a small plateau and have a greatly decreased variance with
 178 12 markers, indicating that there are enough markers such that adding more markers to the analysis will
 179 not create very many new genotypes (Fig. 5).

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

set.seed(20150428)
genotype_curve(monpop, 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

180 The index of association (I_A) is a measure of multilocus linkage disequilibrium that is most often used
 181 to detect clonal reproduction within organisms that have the ability to reproduce via sexual or asexual
 182 processes (Brown et al., 1980; Smith et al., 1993; Milgroom, 1996). It was standardized in 2001 as \bar{r}_d
 183 by Agapow and Burt (2001) to address the issue of scaling with increasing number of loci. This metric is
 184 typically applied to traditional dominant and co-dominant markers such as AFLPs, SNPs, or microsatellite
 185 markers. With the advent of high throughput sequencing, SNP data is now available in a genome-wide
 186 context and in very large matrices including thousands of SNPs. Thus, the likelihood of finding mutations
 187 within two individuals of a given clone increases and tools are needed for defining clone boundaries.
 188 For this reason, we devised two approaches using the index of association for large numbers of markers
 189 typical for population genomic studies. Both functions utilize *adegenet*'s "genlight" object class, which
 190 efficiently stores 8 binary alleles in a single byte (Jombart and Ahmed, 2011). As calculation of the \bar{r}_d

191 requires distance matrices of absolute number of differences, we utilize a function that calculates these
192 distances directly from the compressed data called `bitwise.dist`.

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

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

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

DATA FORMAT UPDATES: POPULATION STRATA AND HIERARCHIES

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

219 To facilitate handling hierarchical and multilocus genotypic metadata, `poppr` version 1.1 introduced
220 a new S4 data object called “genclone”, extending `adegenet`’s “genind” object. The genclone object
221 formalized the definitions of multilocus genotypes and population hierarchies by adding two slots called
222 “mlg” and “hierarchy” that carried a numeric vector and a data frame, respectively. These new slots allow
223 for increased efficiency and ease of use by allowing these metadata to travel with the genetic data. The
224 hierarchy slot in particular contains a data frame where each column represents a separate hierarchical
225 level. This is then used to set the population factor of the data by supplying a hierarchical formula
226 containing one or more column names of the data frame in the hierarchy slot.

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

AVAILABILITY

232 As of this writing, the `poppr` R package version 2.0 containing all of the features described here
233 is located at <https://github.com/grunwaldlab/poppr/tree/2.0-rc>. It is necessary

234 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):

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

237 Several population genetics packages in R are currently going through a major upgrade and have not yet
238 been updated in CRAN. We will upload *poppr* 2.0 to CRAN once all other reverse dependent packages
239 have been updated.

DISCUSSION

240 Genomic data has become more readily accessible due to advances in low-cost sequencing technology.
241 Many tools have been developed or adapted to these data, but most of them were designed with
242 sexual populations in mind. We have presented here implementations of model-free analyses for clonal
243 organisms. Particularly important is the implementation of \bar{r}_d for genomic data (Agapow and Burt, 2001).
244 Random sampling of loci across the genome can give an expected distribution of \bar{r}_d , which is expected to
245 have a mean of zero for panmictic populations. Additionally, due to the fact that it acts on multiple
246 loci, is not affected by the number of loci sampled, is model free, and has the ability to detect population
247 structure, \bar{r}_d is well suited to sliding window analyses and has the potential to be applied to non-clonal
248 populations.

249 Clustering multilocus genotypes into multilocus lineages based on genetic distances is a non-trivial task.
250 Moreover, this has not previously been implemented for genomic data for clonal populations. Perhaps
251 highlighting that many of the features presented in this paper are not necessarily exclusive to genomic
252 data is the fact that this method of clonal assignment has been available in the programs GENCLONE
253 and GENODIVE (Arnaud-Hanod et al., 2007; Meirmans and Van Tienderen, 2004). Our method with
254 `mlg.filter` builds upon this idea and allows the user to choose between three different approaches
255 for clustering MLGs. As diagrammed in Fig. 1 and demonstrated in Fig. 2, it is clear that the choice
256 of clustering algorithm has an impact on the data, where a genetic distance cutoff of 0.1 would be the
257 difference between 14 MLLs and 17 MLLs for nearest neighbor and UPGMA clustering, respectively (Fig.
258 2). The option to choose the clustering algorithm gives the user the ability to choose what is biologically
259 relevant to their populations.

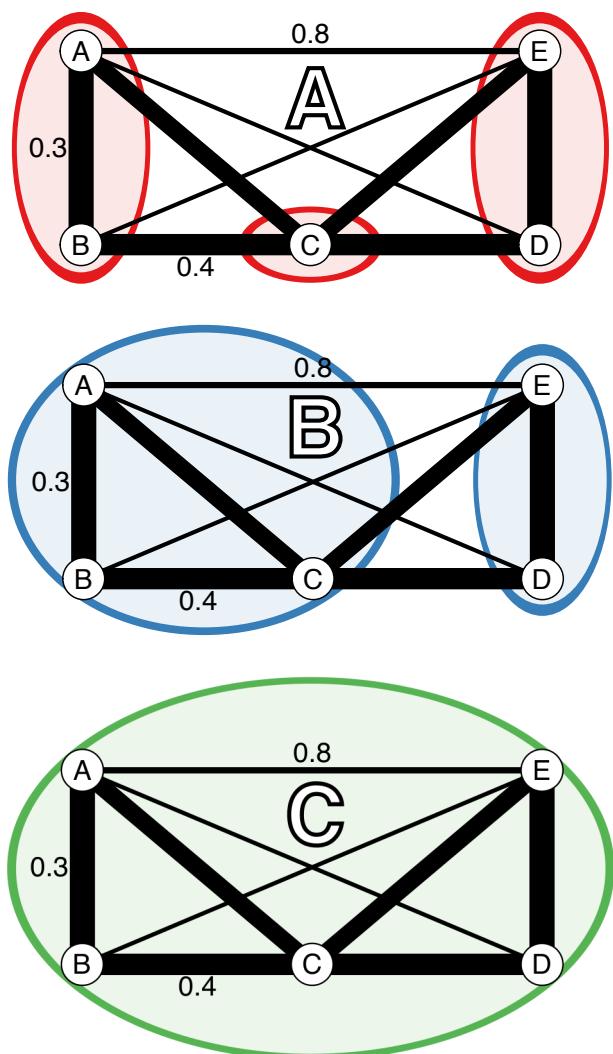
260 Multilocus genotypes that have been clustered can then be visualized in minimum spanning networks.
261 Reticulate minimum spanning networks are very important for clonal organisms where a minimum
262 spanning tree would become a chain, implying that the clones were derived in a progressive and linear
263 fashion. This presents but one potential scenario for clonal organisms, but does not account for any other
264 biologically relevant process. Reticulations in the minimum spanning networks allow for a representation
265 of uncertainty that goes along with clonal organisms. The current implementation in *poppr* has been
266 successfully used in analyses such as reconstruction of the *P. ramorum* epidemic in Curry County, OR
267 (Kamvar et al., 2014a, 2015). Reticulated networks also allow for the application of graph community
268 detection algorithms such as the infoMAP algorithm (Rosvall and Bergstrom, 2008). As shown in the *P.*
269 *ramorum* and H3N2 data, while it is possible to utilize these graph walking algorithms on non-reticulate
270 minimum spanning trees, the results derived from these are limited to explain populations derived from
271 serial cloning events.

272 Implementing these methods in R and hosting the code free and open on GitHub has given us the ability
273 to tailor our tools to the needs of the researchers who use them. We have spent a considerable amount of
274 time developing these methods in such a way that users without technical background would be able to
275 use and understand them without too much effort. As it is an open-source project, those with technical

276 knowledge are invited to contribute by raising issues or pull requests on our repository at <https://github.com/grunwaldlab/poppr/issues>.
277

FIGURES AND TABLES

FIGURE 1



278

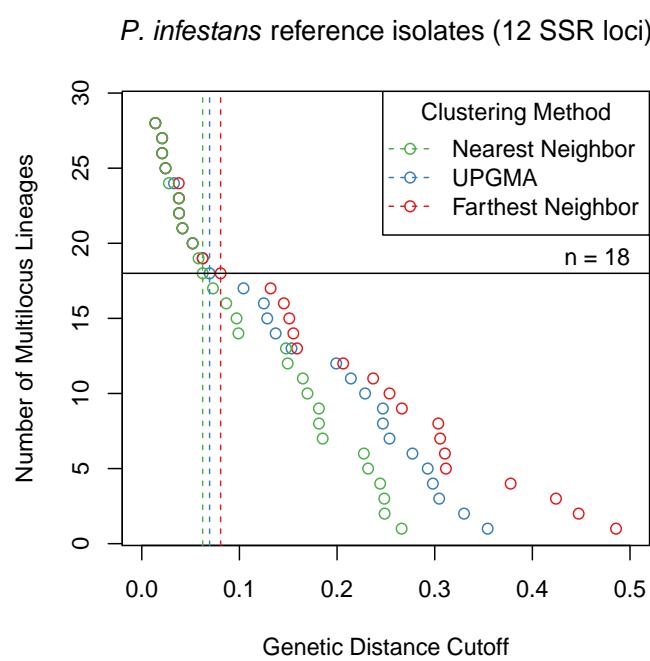
FIGURE 2

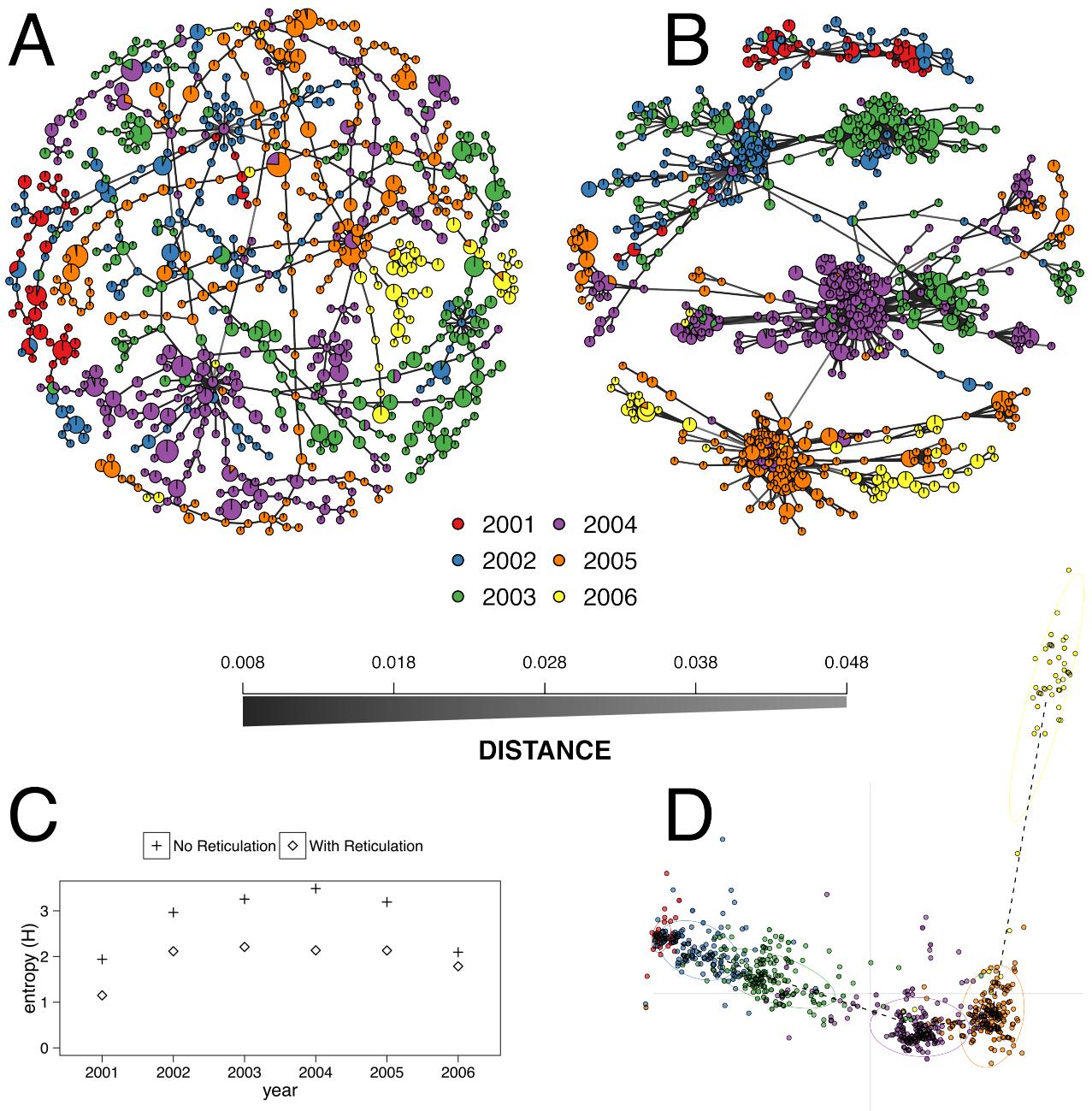
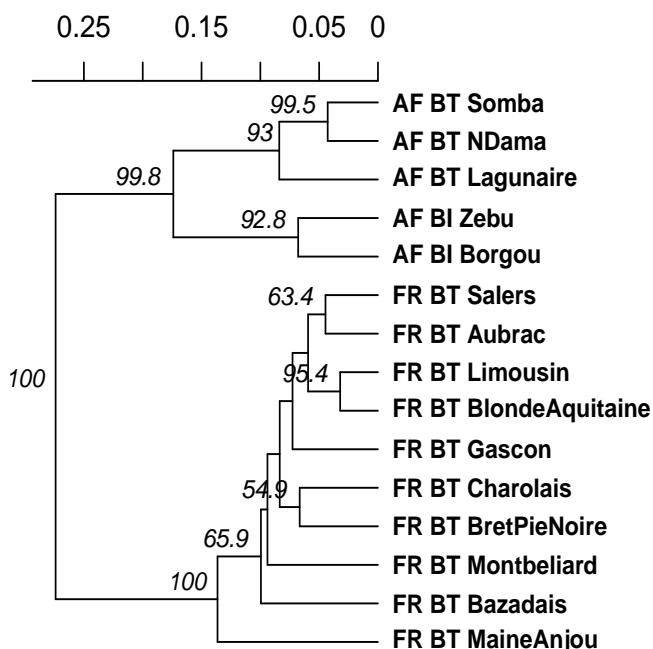
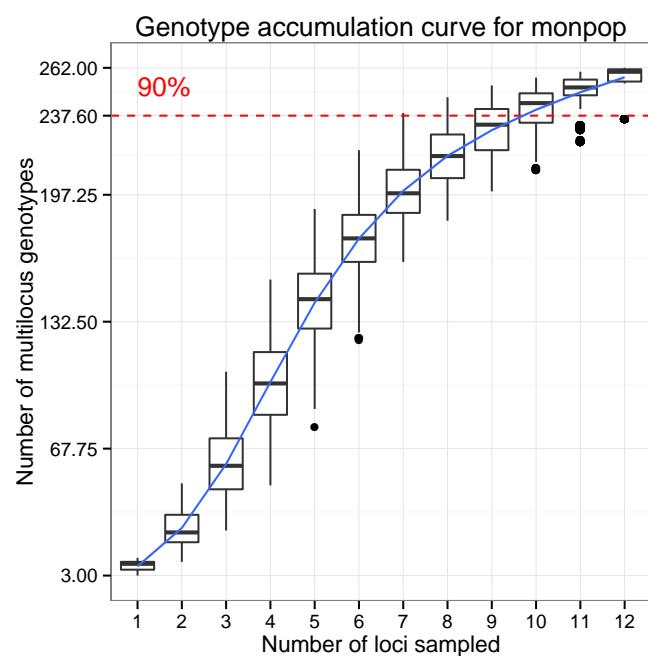
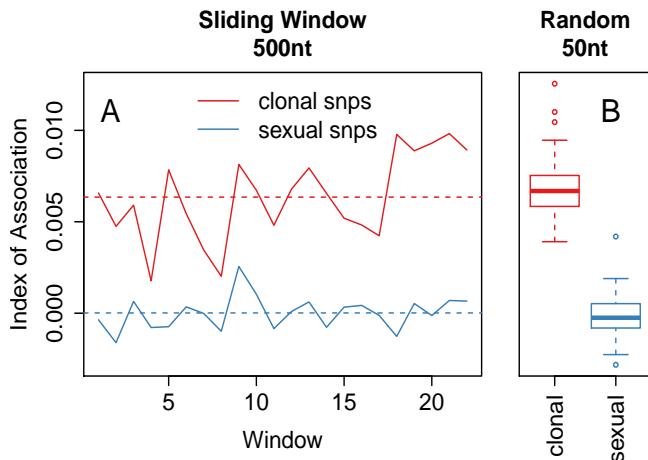
FIGURE 3

FIGURE 4

281

FIGURE 5

282

FIGURE 6

283

284 **Table 1** Contingency table comparing multilocus lineages assigned based on average neighbor clustering
 285 (columns) vs. multilocus lineages defined in Li et al. (2013) and Lees et al. (2006).

	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	2
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

FIGURE LEGENDS

286 **Figure 1** Diagrammatic representation of the three clustering algorithms implemented in `mlg.filter`.
 287 (A-C) Represent different clustering algorithms on the same imaginary network with a threshold of 0.451.
 288 Edge weights are represented in arbitrary units noted by the line thickness and numerical values next to
 289 the lines. All outer angles are 90 degrees, so the un-labeled edge weights can be obtained with simply
 290 geometry. Colored circles represent clusters of genotypes. (A) Farthest neighbor clustering does not cluster
 291 nodes B and C because nodes A and C are more than a distance of 0.451 apart. (B) UPGMA (average
 292 neighbor) clustering clusters nodes A, B, and C together because the average distance between them and

293 C is < 0.451. (C) Nearest neighbor clustering clusters all nodes together because the minimum distance
 294 between them is always < 0.451.

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

302 **Figure 3 (A-B)** Minimum spanning networks of the hemagglutinin (HA) segment of H3N2 viral DNA
 303 from the *aedegenet* package representing flu epidemics from 2001 to 2006 without reticulation (A) and with
 304 reticulation (B) (Jombart, 2008; Jombart et al., 2010). Each node represents a unique multilocus genotype,
 305 colors represent epidemic year, and edge color represents absolute genetic distance. (C) Shannon entropy
 306 values for population assignments compared with communities determined by the infoMAP algorithm on
 307 (A) and (B). (D) Graphic reproduced from Jombart et al. (2010) showing that the 2006 epidemic does not
 308 cluster neatly with the other years.

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

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

318 **Figure 6 (A)** Sliding window analysis of the standardized index of association (\bar{r}_d) across a simulated
 319 1.1×10^4 nt chromosome containing 1,100 variants among 100 individuals. Each window analyzed variants
 320 within 500nt chunks. The black line refers to the clonal and the blue line to the sexual populations. (B)
 321 boxplots showing 100 random samples of 50 variants to calculate a distribution of \bar{r}_d for the clonal (red)
 322 and sexual (blue) populations. Each box is centered around the mean, with whiskers extending out to 1.5
 323 times the interquartile range. The median is indicated by the center line. (A) and (B) are plotted on the
 324 same y-axis.

REFERENCES

- 325 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.
- 327 Anderson, J. B., and Kohn, L. M. (1995). Clonality in soilborne, plant-pathogenic fungi. *Annual review of phytopathology* 33, 369–391.
- 329 Arnaud-Hanod, S., Duarte, C. M., Alberto, F., and Serro, E. A. (2007). Standardizing methods to address
 330 clonality in population studies. *Molecular Ecology* 16, 5115–5139.
- 331 Brown, A., Feldman, M., and Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum*
 332 *spontaneum*. *Genetics* 96, 523–536. Available at: <http://www.genetics.org/content/96/2/523.abstract>.
- 334 Bruvo, R., Michiels, N. K., D'Souza, T. G., and Schulenburg, H. (2004). A simple method for the
 335 calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology* 13, 2101–
 336 2106.

- 337 Chakarov, N., Linke, B., Boerner, M., Goesmann, A., Krger, O., and Hoffman, J. I. (2015). Apparent
338 vector-mediated parent-to-offspring transmission in an avian malaria-like parasite. *Molecular ecology* 24,
339 1355–1363.
- 340 Cooke, D. E., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J., Deahl, K. L.,
341 Farrer, R. A., Gilroy, E. M., Goss, E. M., et al. (2012). Genome analyses of an aggressive and invasive
342 lineage of the Irish potato famine pathogen. *PLoS pathogens* 8, e1002940.
- 343 Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research.
344 *InterJournal Complex Systems*, 1695. Available at: <http://igraph.org>.
- 345 Dagum, L., and Menon, R. (1998). OpenMP: An industry standard API for shared-memory
346 programming. *Computational Science & Engineering, IEEE* 5, 46–55.
- 347 Davey, J. W., and Blaxter, M. L. (2010). RADSeq: Next-generation population genetics. *Briefings in
348 Functional Genomics* 9, 416–423. doi:10.1093/bfgp/elq031.
- 349 Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L.
350 (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature
351 Reviews Genetics* 12, 499–510.
- 352 Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American
353 Biology Teacher* 75, 87–91.
- 354 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E.
355 (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one
356* 6, e19379.
- 357 Everhart, S., and Scherm, H. (2015). Fine-scale genetic structure of *Monilinia fructicola* during brown
358 rot epidemics within individual peach tree canopies. *Phytopathology* 105, 542–549.
- 359 Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from
360 metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data.
361 *Genetics* 131, 479–491.
- 362 Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus
363 genotype data: Linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. Available at:
364 <http://www.genetics.org/content/164/4/1567.abstract>.
- 365 Goss, E. M., Larsen, M., Chastagner, G. A., Givens, D. R., and Grnwald, N. J. (2009). Population
366 genetic analysis infers migration pathways of *Phytophthora ramorum* in US nurseries. *PLoS pathogens* 5,
367 e1000583.
- 368 Goss, E. M., Tabima, J. F., Cooke, D. E., Restrepo, S., Fry, W. E., Forbes, G. A., Fieland, V. J., Cardenas,
369 M., and Grnwald, N. J. (2014). The Irish potato famine pathogen *Phytophthora infestans* originated in
370 central Mexico rather than the Andes. *Proceedings of the National Academy of Sciences* 111, 8791–8796.
- 371 Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular
372 Ecology Notes* 5, 184–186.
- 373 Gross, A., Hosoya, T., and Queloz, V. (2014). Population structure of the invasive forest pathogen
374 *Hymenoscyphus pseudoalbidus*. *Molecular ecology* 23, 2943–2960.
- 375 Grnwald, N. J., and Goss, E. M. (2011). Evolution and population genetics of exotic and re-emerging
376 pathogens: Novel tools and approaches. *Annual Review of Phytopathology* 49, 249–267.
- 377 Grnwald, N. J., and Hoheisel, G.-A. (2006). Hierarchical analysis of diversity, selfing, and genetic
378 differentiation in populations of the oomycete *Aphanomyces euteiches*. *Phytopathology* 96, 1134–1141.

- 379 Grnwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. (2003). Analysis of genotypic
380 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>.
- 382 Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers.
383 *Bioinformatics* 24, 1403–1405. doi:10.1093/bioinformatics/btn129.
- 384 Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide SNP
385 data. *Bioinformatics* 27, 3070–3071.
- 386 Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A
387 new method for the analysis of genetically structured populations. *BMC genetics* 11, 94.
- 388 Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grnwald, N. J. (2015). Spatial and
389 temporal analysis of populations of the sudden oak death pathogen in Oregon forests. *Phytopathology*, in
390 press.
- 391 Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grnwald, N. J. (2014a).
392 Sudden_Oak_Death_in_Oregon_Forests: Spatial and temporal population dynamics of the sudden oak
393 death epidemic in Oregon Forests. doi:10.5281/zenodo.13007.
- 394 Kamvar, Z. N., Tabima, J. F., and Grnwald, N. J. (2014b). Poppr: An R package for genetic analysis of
395 populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2, e281.
- 396 Lalo, D., Jombart, T., Dufour, A.-B., and Moazami-Goudarzi, K. (2007). Consensus genetic structuring
397 and typological value of markers using multiple co-inertia analysis. *Genetics Selection Evolution* 39, 1–23.
- 398 Lees, A., Wattier, R., Shaw, D., Sullivan, L., Williams, N., and Cooke, D. (2006). Novel microsatellite
399 markers for the analysis of *Phytophthora infestans* populations. *Plant Pathology* 55, 311–319.
- 400 Li, Y., Cooke, D. E., Jacobsen, E., and Lee, T. van der (2013). Efficient multiplex simple sequence repeat
401 genotyping of the oomycete plant pathogen *Phytophthora infestans*. *Journal of microbiological methods*
402 92, 316–322.
- 403 Linde, C., Zhan, J., and McDonald, B. (2002). Population structure of *Mycosphaerella graminicola*:
404 From lesions to continents. *Phytopathology* 92, 946–955.
- 405 Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of
406 population genomics: From genotyping to genome typing. *Nature Reviews Genetics* 4, 981–994.
- 407 Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*
408 27, 209–220.
- 409 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piero, D., and Emerson, B.
410 (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly
411 optimization for population genetic inference. *Molecular ecology resources* 15, 28–41.
- 412 McDonald, B. A., and Linde, C. (2002). The population genetics of plant pathogens and breeding
413 strategies for durable resistance. *Euphytica* 124, 163–180. doi:10.1023/A:1015678432355.
- 414 Meirmans, P. G., and Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: Two programs for
415 the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4, 792–794.
- 416 Metzger, M. J., Reinisch, C., Sherry, J., and Goff, S. P. (2015). Horizontal transmission of clonal cancer
417 cells causes leukemia in soft-shell clams. *Cell* 161, 255–263.
- 418 Michalakis, Y., and Excoffier, L. (1996). A generic estimation of population subdivision using distances
419 between alleles with special reference for microsatellite loci. *Genetics* 142, 1061–1064.
- 420 Milgroom, M. G. (1996). Recombination and the multilocus structure of fungal populations. *Annual review of phytopathology* 34, 457–477.

- 422 Milgroom, M. G., Levin, S. A., and Fry, W. E. (1989). Population genetics theory and fungicide
423 resistance. *Plant disease epidemiology* 2, 340–367.
- 424 Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National
425 Academy of Sciences* 70, 3321–3323.
- 426 Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L.,
427 Solymos, P., Stevens, M. H. H., and Wagner, H. (2015). *Vegan: Community ecology package*. Available
428 at: <http://CRAN.R-project.org/package=vegan>.
- 429 Paradis, E. (2010). Pegas: an R package for population genetics with an integrated–modular approach.
430 *Bioinformatics* 26, 419–420.
- 431 Prevosti, A., Ocaa, J., and Alonso, G. (1975). Distances between populations of *Drosophila subobscura*,
432 based on chromosome arrangement frequencies. *Theoretical and Applied Genetics* 45, 231–241.
- 433 R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R
434 Foundation for Statistical Computing Available at: <http://www.R-project.org/>.
- 435 Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal
436 community structure. *Proceedings of the National Academy of Sciences* 105, 1118–1123.
- 437 Shannon, C. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing
438 and Communications Review* 5, 3–55. Available at: <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>.
- 440 Smith, J. M., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria?
441 *Proceedings of the National Academy of Sciences* 90, 4384–4388. doi:10.1073/pnas.90.10.4384.
- 442 Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 38,
443 1409–1438.
- 444 Taylor, J. W., and Fisher, M. C. (2003). Fungal multilocus sequence typing — it's not just for bacteria.
445 *Current opinion in microbiology* 6, 351–356.
- 446 Wickham, H., and Chang, W. (2015). *Devtools: Tools to make developing R packages easier*. Available
447 at: <http://CRAN.R-project.org/package=devtools>.