2017

Best Practices for Population Genetic Analyses

Niklaus J. Grünwald
USDA-ARS, Corvallis, OR, nik.grunwald@ars.usda.gov

Sydney E. Everhart
University of Nebraska-Lincoln, everhart@unl.edu

B. J. Knaus
USDA-ARS, Corvallis, OR

Zhian N. Kamvar
Oregon State University, zkamvar@unl.edu

Phytopathology
REVIEW

Best Practices for Population Genetic Analyses

N. J. Grünwald,† S. E. Everhart, B. J. Knaus, and Z. N. Kamvar

First and third authors: Horticultural Crop Research Unit, USDA-ARS, Corvallis, OR; and second and fourth authors: Department of Botany and Plant Pathology, Oregon State University, Corvallis.
Current address of second author: Department of Plant Pathology, University of Nebraska, Lincoln.
Accepted for publication 9 May 2017.

ABSTRACT

Population genetic analysis is a powerful tool to understand how pathogens emerge and adapt. However, determining the genetic structure of populations requires complex knowledge on a range of subtle skills that are often not explicitly stated in book chapters or review articles on population genetics. What is a good sampling strategy? How many isolates should I sample? How do I include positive and negative controls in my molecular assays? What marker system should I use? This review will attempt to address many of these practical questions that are often not readily answered from reading books or reviews on the topic, but emerge from discussions with colleagues and from practical experience. A further complication for microbial or pathogen populations is the frequent observation of clonality or partial clonality. Clonality invariably makes analyses of population data difficult because many assumptions underlying the theory from which analysis methods were derived are often violated. This review provides practical guidance on how to navigate through the complex web of data analyses of pathogens that may violate typical population genetics assumptions. We also provide resources and examples for analysis in the R programming environment.

Characterizing the population biology of microbes and parasites including genetically distinct groups such as insects, fungi, oomycetes, nematodes, bacteria, or viruses is a complex task requiring subtle skills that are often not explicitly stated in the scientific literature (Goodwin 1997; Grünwald and Goss 2011; McDonald 1997). For many reasons, population genetics is a tremendously useful discipline with a long history of application to plant pathology. For example, one can establish where the likely center of origin of a plant pathogen is located, which in turn allows harnessing of plant resistance genes (Goss et al. 2014; Grünwald and Flier 2005; Stukenbrock et al. 2007; Vleeshouwers and Oliver 2014; Vleeshouwers et al. 2011). Plant pathogens continue to emerge and reemerge and population genetic analyses can be used to infer genetic patterns (subdivision, bottlenecks, clonality, etc.) and processes (gene flow, migration, mutation, genetic drift, etc.) of pathogen emergence. For example, it is now clear that the sudden oak death pathogen, *Phytophthora ramorum*, emerged repeatedly on two continents, most likely via migration in association with shipments of ornamental plants (Goss et al. 2009b; Grünwald et al. 2012). The human pathogen *Cryptococcus gattii* arose recently in the North-Western United States (Byrnes et al. 2010). *Batrachochytrium dendrobatidis*, causal agent of the chytridiomycosis pandemic of amphibians, can be traced back to the outbreak of a single clonal lineage (James et al. 2009). Yet other pathogens have been shown to emerge via hybridization or speciation (Geiser et al. 1998; Giraud et al. 2010; Goss et al. 2011; Stukenbrock et al. 2007). Population genetics can also be useful to provide inferences on the degree of sexual outcrossing by studying linkage among markers (Atallah et al. 2010; Goss et al. 2014; Milgroom 1996; Milgroom et al. 2014).

This review is geared towards biologists who are interested in learning the fundamentals of conducting scientifically rigorous studies into the population genetics of microbial populations. This review assumes that the reader has a basic knowledge of genetics. A critical analysis of the preferred and minimum requirements for tools, techniques and analyses typically used to describe patterns and infer processes in populations of organisms is provided. At the same time, we acknowledge that within the space constraints of a review article not all aspects can be covered in detail. Finally, this review provides resources for reproducing some of the analyses shown here in the R programming environment (R Core Team 2016).

POPULATION GENETICS

Population genetics is the study of the distribution in space and time of allele frequencies (patterns) resulting from certain evolutionary forces or processes (Carbone and Kohn 2004; Milgroom 2015). Characterization of allele frequencies and distributions in a population enable inferences about processes (genetic drift, mutation, gene flow, and natural selection), which have shaped the patterns observed for a given population (e.g., clustering, differentiation, divergence, etc.)

†Corresponding author: N. J. Grünwald; E-mail: nik.grunwald@ars.usda.gov

(Hartl and Clark 1997). For example, one might find populations that are moderately differentiated (Fig. 1A), poorly differentiated (Fig. 1B), or clonal (Fig. 1C) in nature (Table 1). A good population analysis consists of asking relevant biological questions, sampling individuals, determining frequencies of alleles at loci and using statistical approaches to infer patterns and processes (Fig. 2). A rigorous population genetic study requires a series of iterative steps necessary for analysis (Fig. 2). We also refer the reader to a seminal textbook on the population biology of plant pathogens that provides a treasure of resources that go beyond the content of our review given the space limitations, yet are highly complementary (Milgroom 2015). Most fundamental to a good study is the ability to formulate and test biological hypotheses.

TEST BIOLOGICAL HYPOTHESES

Perhaps the single most important aspect of a good population genetic study is the testing of specific hypotheses. Formulated as questions, these might consist of: Are populations regionally differentiated? Is this population introduced? Is there gene flow among populations? Are there sink and source populations? What is the center of origin or diversity? Are populations clonal, sexual or mixed? Too many studies neglect the formulation of testable

TABLE 1
Basic population statistics for the three scenarios presented in Figure 1: highly differentiated populations (A), poorly differentiated (B), and clonal populations (C)

Population	*N*	MLG[a]	*Het*[b]	G_{ST}[c]
A	80	10	0.538	0.143
B	80	6	0.513	0.007
C	80	6	0.588	0.069

[a] MLG = number of multilocus genotypes.
[b] *Het* = observed heterozygosity.
[c] G_{ST} = Nei's measure of population differentiation for multiple allele cases (Nei 1973).

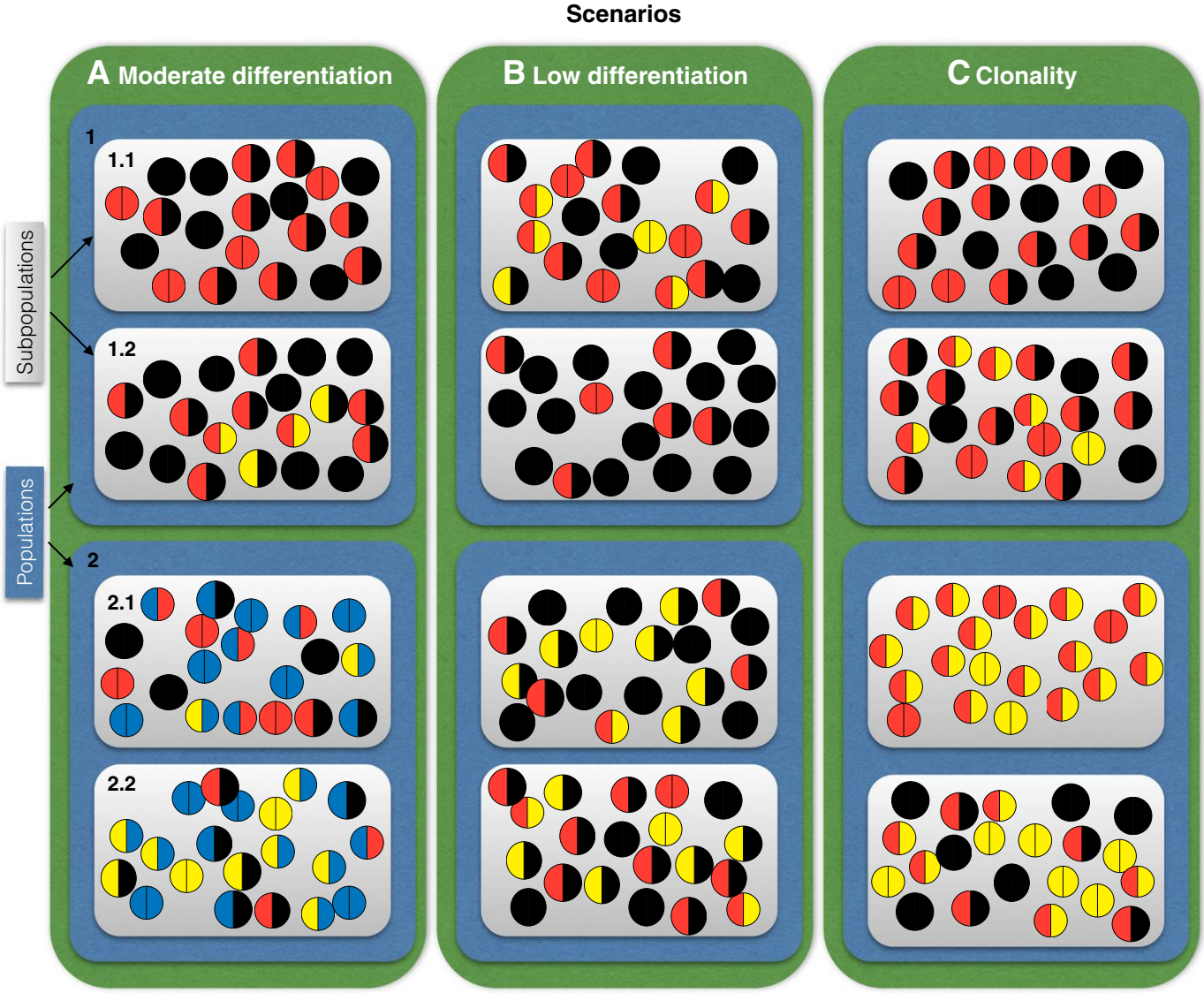

FIGURE 1
Three distinct scenarios depicting sampling of 20 individuals of a diploid organism with up to four alleles (black, blue, red, and yellow) per locus for populations that are **A,** moderately differentiated, **B,** poorly differentiated, or **C,** are clonal across a hierarchy consisting of subpopulations (gray) and populations (blue). Basic population genetic metrics on these three populations are provided in Table 1.

hypotheses and are thus rightly rejected in journals because they merely describe genetic or genotypic diversity without answering biological questions. To formulate testable hypotheses, we need to understand the natural history of the organism.

KNOW YOUR ORGANISM

Knowing the genetics and natural history of the organism one works with is fundamentally important (Fig. 2). For example, knowing the organism spreads via air over long distances as opposed to spreading more locally via water splash dispersal will result in different expectations for testing hypotheses and sampling individuals across the most appropriate spatial scales (Linde et al. 2002; Mundt et al. 2009, 2011).

Similarly, knowledge of ploidy of an individual will influence selection of an appropriate marker system or analysis method. For example, a codominant marker system such as microsatellites, sequences, or single nucleotide polymorphisms (SNPs) are a preferred choice for diploid or polyploid organism (Grünwald and Goss 2011; Grünwald et al. 2016b). In diploid, nuclear loci alleles can be sequenced and show one (homozygous) or two (heterozygous) alleles per locus. A pertinent example comes from sequencing nuclear loci of the sudden oak death pathogen *Phytophthora ramorum* to determine ancestry (Goss et al. 2009a). For haploid pathogens sequencing of genes is simplified by the fact that only one allele is found at a locus (Berbegal et al. 2013; Milgroom et al. 2014).

Observing sexual reproductive structures in nature will provide an expectation that the organism is sexual (Goss et al. 2014; Milgroom 1996; Peever et al. 2004). However, populations exhibiting a mixed mating system (e.g., involving both outcrossing and selfing) may purge the variation needed to infer a sexual mode of reproduction through inbreeding, resulting in a situation where it becomes difficult to detect if sexual reproduction occurs. Knowing that hybridization might occur will result in selection of different analysis methods (see below). Investigators should inform their analyses as much as possible by deriving methods, sampling schemes, and hypotheses from prior

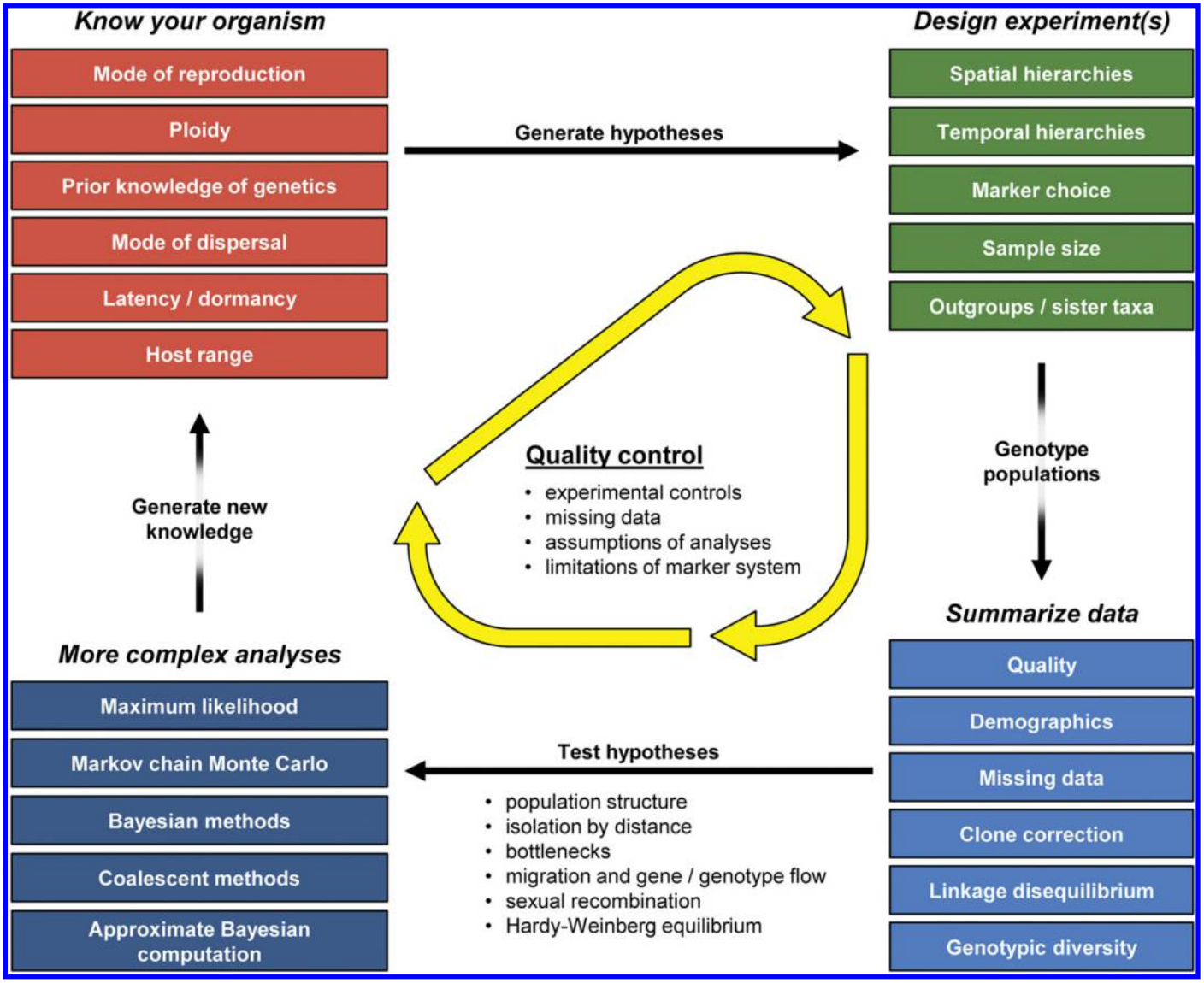

FIGURE 2

Summary of major steps in a population genetic analysis. The first step is to know your organism and generate hypotheses that inform the next step in the process, experimental design. After genotyping populations, data should be summarized and hypotheses tested with more complex analyses. While the focus of the analysis is to address biological questions, it is of central importance to consider quality control; ensure conclusions take into consideration quality and limitations of the data and analyses. Completion of these steps will hopefully conclude in generating new knowledge, whereupon new hypotheses may be generated.

knowledge about the biology of the organism. Often, a small pilot study will enable design of a more rigorous study.

HOW SHOULD I SAMPLE POPULATIONS?

All population genetic analyses start with sampling individuals randomly from a population (Fig. 1). One of the biggest pitfalls in conducting a population genetic analysis is to just pick a few isolates off the shelves of your culture collection and request a few more isolates from colleagues around the world without development of a sound sampling strategy geared toward testing a specific hypothesis or answering specific biological questions. Most importantly, the individual samples are expected to be a random sample of the population they should represent. Thus, sampling design can make or break a study.

A sampling strategy should be developed based on what is known about the biology of the pathogen (e.g., mode of spread, known sexual/asexual cycles, local/regional/global distribution, etc.) and the primary research question. For example, clonal populations do not require as many samples as sexual populations or populations that are highly structured. A recent study was based on knowledge that the brown rot pathogen *Monilinia fructicola* infects both blossoms and fruit. This raised the question as to whether all fruit infections are caused primarily by prior blossom infections (Everhart and Scherm 2015). This study required a hierarchical sampling at two dates during flowering and fruiting, respectively. To measure the relative contribution of immigration to populations, sampling over space and time is needed (Zhan et al. 1998). For most marker systems, a suggested sample might include around 10 to 20 individuals per sample unit, e.g., subpopulation, and upwards of 25 to 30 individuals per population (Milgroom 2015). Note, however, that sample size considerations depend on many factors including the biological question, ploidy, the marker system, the diversity expected, and the statistical power required (Hale et al. 2012). For example, statistical power to detect linkage disequilibrium requires larger sample sizes (Brown 1975; Milgroom 2015). User-friendly software for optimizing sampling strategy for common genetic study topics such as hybridization, temporal sampling, bottlenecks, connectivity and assignment is available (Hoban et al. 2013).

Another crucial aspect is the strategy of sampling across populations. Ideally, populations should be sampled hierarchically in order to assess population variation that may occur over different spatial or temporal scales (Everhart and Scherm 2015; Grünwald and Goss 2011; Grünwald and Hoheisel 2006; Kamvar et al. 2015b; McDonald 1997; Zhan et al. 1998). Without a hierarchical sample one cannot infer if variation is observed at the overall, population, or subpopulation level. Figure 1 illustrates a hierarchical sample with subpopulation and population level samples.

A less obvious aspect of sampling is whether we sampled a single individual or a mixture of individuals. In mycology or microbiology, this is typically attained by single-spore isolation or hyphal-tipping of strains. If sampling does not isolate a single individual, genotyping methods will result in a mixture of genotypes that cannot be differentiated. The fundamental assumption of genotyping is that an individual, rather than mixtures of individuals, are sampled and genotyped.

A final consideration might include the choice of outgroups or sister taxa to include that can provide information on the ancestral state for observed polymorphisms. For example, a recent study included all *Phytophthora* clade 1c taxa to infer the origin and evolutionary history of the Irish famine pathogen *Phytophthora infestans* (Goss et al. 2014). In the absence of a close relative the ancestral state of any locus cannot be inferred. Once a population sample is in hand, a suitable molecular marker needs to be chosen to genotype individuals and assess allelic diversities.

CHOOSING AND USING MOLECULAR MARKERS

Molecular markers need to be chosen appropriately to be neutral, reasonably polymorphic, reproducible, and provide insights at the right evolutionary scale (Grünwald and Goss 2011). Markers with high mutation rates such as microsatellites (also known as simple sequence repeats or SSRs) provide insights into recent divergence (Atallah et al. 2010; Baumgartner et al. 2010; Berbegal et al. 2013; Dunn et al. 2014; Dutech et al. 2010; Goss et al. 2009b), whereas mitochondrial, nuclear or other sequence loci provide inferences about the more distant evolutionary history given the slower mutation rates (Carbone et al. 2004; Goss et al. 2009a; Malvarez et al. 2007; Schoebel et al. 2014; Stukenbrock et al. 2007).

One important aspect of genotyping is the reproducibility of allele calling (Bonin et al. 2004). Often authors do not provide information on how they address genotyping error. Marker systems like random amplified polymorphic DNA, amplified fragment length polymorphism, simple sequence repeats (SSR), or genotyping by sequencing (GBS) differ in regards to reproducibility (Jones et al. 1997; Grünwald et al. 2016b). To avoid this pitfall, we recommend replication of analyses on independent DNA extractions, PCR runs, electrophoresis runs, and/or sequencing runs. This is ideally done on either all individuals sampled or a statistically representative random subset of individuals. Furthermore, a set of positive controls should be included in all steps of the analyses. Good positive controls may consist of a panel of strains with known genotypes that cover most of the range of alleles observed in the population. Negative controls should also be included, particularly for samples from obligate pathogens such as rusts or powdery mildews to assure that genotyping is specific to the organism and

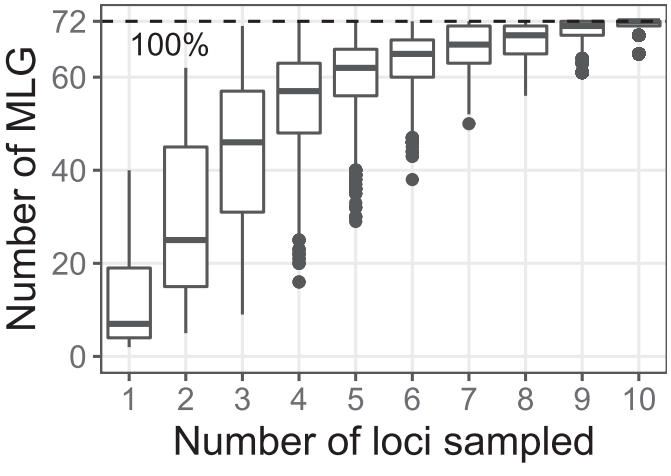

FIGURE 3

A genotype accumulation curve is useful in determining a threshold needed to discriminate among a given percentage of unique individuals given a random sample of *n* loci. *n* loci were randomly sampled 1,000 times in order to create the distribution using *poppr* in R with simple sequence repeats data for populations of the potato late blight pathogen *Phytophthora infestans* (Goss et al. 2014; Grünwald et al. 2016a; Kamvar et al. 2015a).

does not amplify fragments from contaminants such as epiphytic microbes. Inclusion of appropriate controls is just as important with current deep sequencing technologies, such as RADseq and GBS (Elshire et al. 2011; Baird et al. 2008; Grünwald et al. 2016b).

One practical issue arises with obligate pathogens such as downy mildews or rusts that cannot be cultured and thus do not allow for DNA extraction. These pathogens must be grown on a host and DNA extraction thus often includes at least low levels of contamination with nontarget organisms. Several approaches have been used to cope with this issue. DNA can be enriched for the target pathogen by washing spores off sporulating lesions to lower concentration of nontarget DNA. PCR primers can be developed for amplification of species specific loci (Ali et al. 2011). A reference genome of the organisms can be developed where contamination is removed by filtering based on sequence homology analysis (Baxter et al. 2010). GBS can readily be used on obligate pathogens, particularly if a reference genome is available (Summers et al. 2015).

Ascertainment bias should be avoided when developing markers and will improve the applicability of the markers for other studies (Schlötterer 2004). Ascertainment bias here refers to the fact that

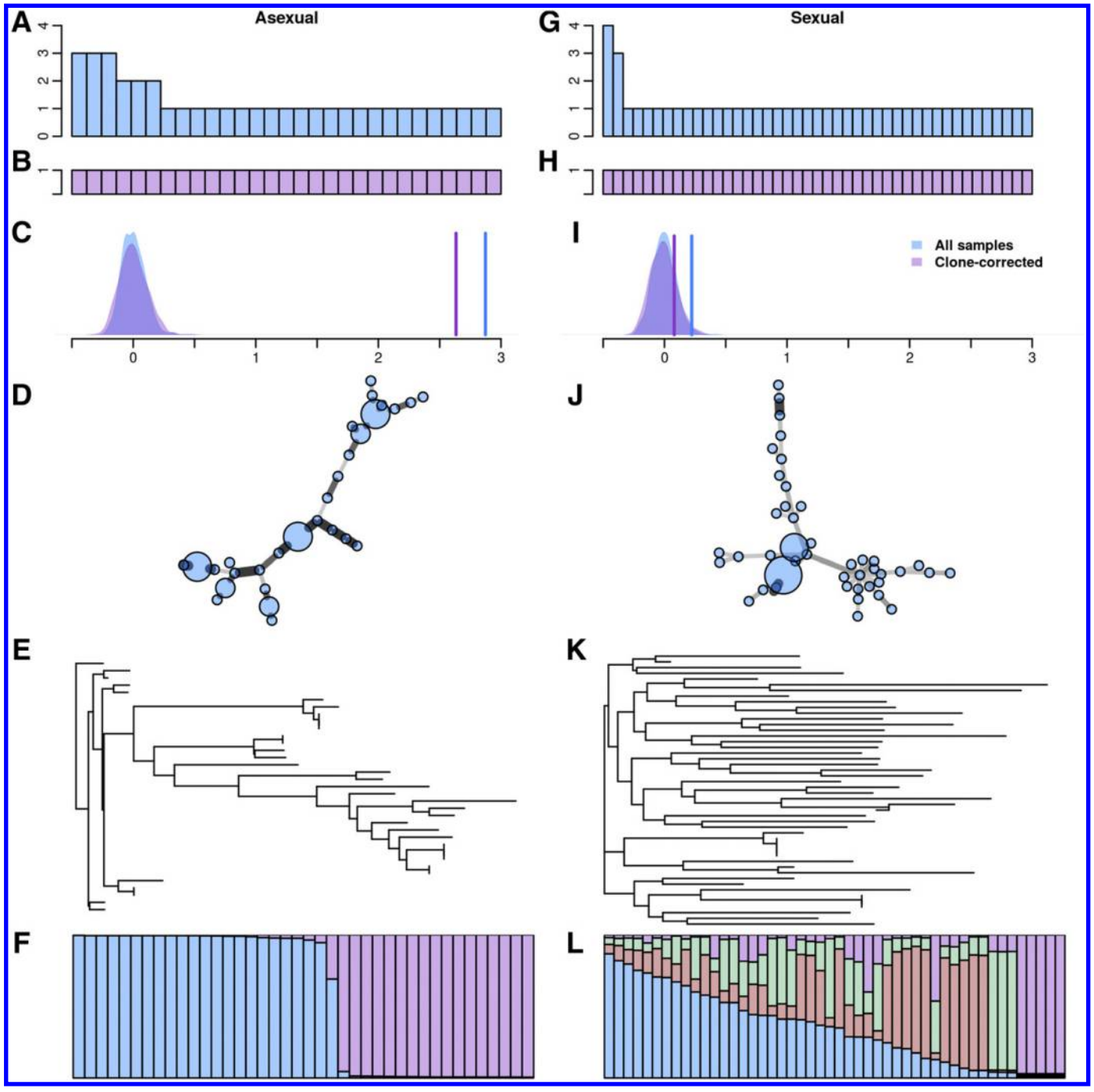

FIGURE 4

Contrasting **A to F,** clonal and **G to L,** sexual populations based on multilocus simple sequence repeats data. These data are taken from published, observed populations of the potato late blight pathogen *Phytophthora infestans* (Goss et al. 2014). From top to bottom we see histograms of total (**A and G**) and clone-corrected (**B and H**) multilocus genotypes, resampled and observed index of association for total and clone-corrected data (**C and I**), minimum spanning networks (**D and J**), dendrogram based on genetic distance (**E and K**), and clustering of individuals into populations using STRUCTURE (**F and L**). The code to reproduce this figure is provided on github (https://github.com/grunwaldlab/popgen_review_examples).

loci are chosen based on a nonrepresentative group of isolates such that they are not variable in all populations, thus biasing the population genetic inferences. Loci for any marker system should be evaluated on a geographically diverse group of isolates. This bias is typically avoided when knowledge of the natural history of an organism is taken into consideration, as mentioned above.

The minimum number of markers that should be used in a population genetic study varies with the genetic diversity of the population, scale of the study, and type of marker used. For example, clonal plant pathogens may exhibit low genetic diversity and require more markers to detect allelic variation. Similarly, fine-scale studies may require larger numbers of markers. The amount of allelic variation also varies with the type of marker used: SSR markers can have a large number of alleles at a locus while SNP markers have a fixed number of allelic states. Thus, fewer microsatellite markers may be needed when compared with SNP markers to achieve the same degree of resolution (Avise 1994; Grünwald et al. 2003; Schlötterer 2004). However, SSR markers are subject to significantly higher levels of homoplasy as compared with SNP markers. A genotype accumulation curve, available in the R packages *poppr* and *RClone*, is a useful tool for determining if more markers are needed (Arnaud-Haond et al. 2007; Bailleul et al. 2016; Kamvar et al. 2014, 2015a) (Fig. 3). When using whole genome or reduced representation genome sequencing, these considerations are less important (Elshire et al. 2011; Baird et al. 2008; Grünwald et al. 2016b; Luikart et al. 2003).

DATA QUALITY AND FORMATTING

One of the most challenging aspects of conducting analyses of populations is that there are many ways to conduct the analysis relying on many different data formats given the diversity of computational tools required. Moreover, after going down one route, further analysis might be needed and most good analyses are inherently iterative (Fig. 2). One open source approach might be to conduct most analyses in the R programming and statistical language (R Core Team 2016), for which many analysis methods are available in different packages. R provides many packages suitable for population genetics (Jombart 2008; Jombart and Ahmed 2011; Kamvar et al. 2014; Oksanen et al. 2013; Kamvar et al. 2015a), population genomics (Knaus and Grünwald 2017), and network analysis (Csardi and Nepusz 2006; Wickham 2009). While R is an extremely powerful and adaptable environment, it can be intimidating to new users and involves a steep learning curve. GenAlEx, a macro for Microsoft Excel, is commonly used for managing data and preliminary analyses (Peakall and Smouse 2012), but is limited in the scope of analyses that can be conducted as well as relying on proprietary software. The R package *poppr* (Kamvar et al. 2014, 2015a) can import GenAlEx formatted data providing a streamlined workflow into R. Further tools provided in *poppr* allow export of data in formats compatible with other commonly used population genetic software. A recent special issue on population genetics in R has been published in *Molecular Ecology Resources* includes novel tools for working with variant call format (VCF) files, large SNP data sets, conducting simulations, and educational resources (Archer et al. 2017; Kamvar et al. 2017; Knaus and Grünwald 2017; Paradis et al. 2017; Parobek et al. 2017; Stanley et al. 2017). R provides great publication-ready graphing tools as exemplified by Figure 3. Because of these tools, it has been possible to produce all the tables and figures necessary for manuscripts entirely within R (Kamvar et al. 2015b; Rojas et al. 2017). We also provide a primer on conducting analyses in R (Grünwald et al. 2016a; https://grunwaldlab.github.io/Population_Genetics_in_R/).

Despite best efforts, there will undoubtedly be some form of error when generating a data set. Before analysis, data should be checked carefully for quality issues, including spurious allele calls, private alleles, missing data or fixed alleles. Basic per locus summary statistics allow inspection of the behavior of individual loci, revealing potential abnormalities in the data.

Missing or null alleles, should be calculated per locus and may be indicative of technical error during amplification. Loci or samples with missing data might have to be removed from analyses. It is suggested that analyses be conducting by selectively removing loci and/or samples with missing data to assess if inferences change substantially. In some cases, missing data might be informative. This occurs if the null alleles are restricted to occurrence in a certain subpopulation, suggesting that lack of amplification is due to a heritable mutation in the primer-recognition site flanking the amplified locus. Missing data are also a considerable problem with GBS data where read depth per locus is very variable and many loci lack any sequence reads (Grünwald et al. 2016b). Given that this technology is rapidly evolving, it is too early for recommendations, but researchers should make sure missing data or imputed data do not bias their inferences.

GENOTYPIC DIVERSITY

A useful first analysis is the determination of basic diversity statistics including the calculation of genotypic diversity, evenness and richness. Diversity measures such Stoddart and Taylor's index G (Stoddart and Taylor 1988) or Shannon-Wiener index H' (Shannon and Weaver 1949) measure genotypic diversity, which combines both aspects of evenness and richness. Evenness indices measure how uniformly genotypes are distributed within a population, while richness is a measure of the number of genotypes observed regardless of relative abundance (Ludwig 1988). Genotypic diversity is an important metric for clonal or partially clonal populations often encountered with microbes.

Authors are often interested in comparing populations with unequal samples sizes and incorrectly divide a given index of diversity by sample size n as follows for the Shannon-Wiener or Stoddart and Taylor indices, respectively, $H'/\ln(n)$ or G/n. However, this scaling is inappropriate as the index decreases to 0 as n increases to infinity. A more appropriate correction is the number of genotypes observed (g) or expected by rarefaction (Grünwald et al. 2003): $H'/\ln(g)$ or Stoddard and Taylor's G/g. Since diversity comprises both richness and evenness, scaling by g thus reduces diversity to a simple measure of evenness.

POPULATION STRUCTURE

When we talk about population structure we are interested in describing the pattern of genetic relatedness among different

BOX 1
HARDY-WEINBERG EQUILIBRIUM

In diploid populations experiencing random sexual mating, the allele frequencies in the population can be used to predict genotype frequencies at single loci (i.e., combinations of alleles) (Hartl and Clark 1997). This equilibrium is met in populations where mutation, migration and natural selection are absent and whose size is large enough such that genetic drift is inconsequential. Like the Wright-Fisher model, one assumes that generations are nonoverlapping and focusing on a single locus, which has two alleles. A strength of this equilibrium is that a population may deviate from this equilibrium for many generations, but with one generation of satisfying these assumptions, this equilibrium can be restored. This equilibrium was originally formulated independently by G. H. Hardy and W. Weinberg and published in 1908.

perceived groups (e.g., populations or subpopulations). These groups can be defined a priori based on our knowledge of the biology or can be inferred a posteriori based on various clustering methods. It is suggested to use both methods as they get at the same question from two sides of the same coin based on either our prior knowledge of the system or assuming no prior knowledge. When defining groupings a priori one often uses geographic regions, treatments (such as different hosts or fungicide applications), or time periods (e.g., years). Populations are then analyzed for genetic variation within and among these predefined groups. Typical analyses across a spatial or temporal hierarchy might include analysis of molecular variance (Excoffier et al. 1992) or population differentiation based on fixation indices (Weir and Cockerham 1984). Below, we provide a discussion of the relative merits and trade-offs of using model-based, distance-based, and ordination-based clustering methods.

Several methods exist to infer clusters or subpopulations from a sample a posteriori. Some of these tools make the assumptions that populations are sexual in nature (e.g., STRUCTURE; Pritchard et al. 2000), while other methods are model free (e.g., K-means clustering, ordination, etc. discussed below). The software STRUCTURE has become a popular tool to make inferences about the number of groups which may exist in a sample of genotyped individuals, as well as what proportion of each individual originated from these groups (i.e., admixed individuals) (Pritchard et al. 2000). STRUCTURE is a model based statistical clustering method implemented in a Bayesian framework to infer the population membership of individuals (Falush et al. 2003; Hubisz et al. 2009; Pritchard et al. 2000). STRUCTURE plots, showing genetic admixture as stacked bar charts for an individual based on inferred grouping, are now a common component of population genetic analyses (Fig. 4F and L). The optimal number of clusters may be determined based on a post hoc test (Pritchard et al. 2000) or a more formal test (Evanno et al. 2005). The model employed assumes Hardy-Weinberg equilibrium (Box 1) as well as a lack of linkage disequilibrium among markers, such that each allele at each locus of each genotype is an independent draw from its frequency distribution (Milgroom 1996; Grünwald and Goss 2011). For clonal or populations with mixed reproduction (e.g., sexual and asexual), we suggest using model free clustering methods.

Several model free alternatives to STRUCTURE are available and are generally the preferred choice for inferring population structure in clonal populations. These methods range from clustering individuals based on a genetic distance (Box 2) to some form of ordination. Dendrograms based on pairwise genetic distance with bootstrap support are the most traditional of these techniques (Fig. 4E and K). Minimum spanning networks (MSN) are another way to visualize genetic relatedness among individual multilocus genotypes (MLGs) or haplotypes represented by size of the node and genetic distances among MLGs shown with connecting branches (Fig. 4D

BOX 2
CHOOSING A GENETIC DISTANCE

Clustering methods including minimum spanning networks and bootstrapped dendrograms (or 'trees') rely on the calculation of a genetic distance matrix that summarizes the relatedness between individuals or populations. The choice of genetic distance to use can have a profound effect on the result (Hartl and Clark 1997; Kosman and Leonard 2005; Weir 1996). To properly choose a genetic distance, one must consider the scope (individuals or populations), marker type, ploidy of the study organism, and any underlying assumptions about biological processes that lead to variation such as drift and mutation.

Distances between individuals: Inter-individual distances are computed without any knowledge of population structure and can be used for clustering, creating dendrograms, and variance partitioning methods such as analysis of molecular variance. For a model-free distance applicable for haploids and diploids with codominant markers, Kosman and Leonard (2005) provide a detailed comparison of the Dice, Jaccard, and simple mismatch distance. They recommend the use of a simple mismatch distance (aka Manhattan distance), which is represented as the fraction of mismatched alleles between two individuals, and is implemented in several programs including the R packages *poppr*, *mmod*, and *PopGenReport* (Adamack and Gruber 2014; Kamvar et al. 2014; Winter 2012). The main advantage of this distance is that it is easily interpretable, despite it not having any biological meaning. It should be noted that GenAlEx also calculates a form of this distance, but it differs in that homozygotes that share no alleles at a locus are further apart than a homozygote and heterozygote sharing no alleles at a locus (Kosman and Leonard 2005; Smouse and Peakall 1999). Model-based distances depend primarily on the type of marker used. For dominant markers derived from restriction fragments, the restdist program in Phylip is appropriate as it assumes a nucleotide substitution model for the restriction sites (Felsenstein 2004; Nei and Li 1979). Distances for sequence data are easily calculated using one of the nucleotide substitution models present in the dist.dna() function in the *ape* R package (Paradis et al. 2004), but are only applicable to haplotype data. For simple sequence repeats (SSR) data, if one can assume that the alleles mutate in a stepwise fashion, Bruvo's distance, as implemented in *poppr*, *polysat*, and Genodive is appropriate for all ploidy levels (Bruvo et al. 2004; Clark and Jasieniuk 2011; Kamvar et al. 2014; Meirmans and Van Tienderen 2004). Calculating genetic distance for polyploid data are challenging because of ambiguity of heterozygous genotypes (e.g., a tetraploid heterozygote with alleles A and T could be ATTT, AATT, or AAAT). Bruvo et al. (2004) addressed this with three models that account for allelic ambiguity, all of which are implemented in both *polysat* and *poppr*. For polyploid SSR data that doesn't follow a stepwise mutation model one can use the methods implemented in the meandist.matrix2() function in *polysat* or turn Bruvo's distance into the simple mismatch distance by multiplying all fragment lengths by 1,000 as demonstrated in Metzger et al. (2015).

Distances between populations: Population-based distances first assume that you have some knowledge of the population structure. Model-free methods include Prevosti's distance (Prevosti et al. 1975) (absolute differences between alleles) or Rogers' distance (Rogers 1972) (Euclidean), whereas Edwards' angular distance (Edwards 1971) (Euclidean), Reynolds' coancestry distance (Reynolds et al. 1983) (Euclidean), and Nei's 1978 distance (Nei 1972, 1978) all assume differences arise via genetic drift. Note that Nei's distance additionally assumes that mutations arise via an infinite alleles model. These methods are available in the R package *adegenet* and *poppr* (Jombart 2008; Kamvar et al. 2015a). All the population-level distances are based on population-level allele frequencies, which are readily obtained for haploids and diploids in several programs. For clonal populations, because population frequencies can be skewed due to the presence of repeated genotypes, it is recommended to calculate allele frequencies from clone-corrected data or by using a round-robin approach as implemented in both *RClone* and *poppr* (Arnaud-Haond et al. 2007; Bailleul et al. 2016; Kamvar et al. 2014, 2015a). For autopolyploid data, one can calculate frequencies in the *polysat* R package.

and J). An MSN provides a network of MLGs (called nodes or vertices) connected by lines (or edges) that reflect the genetic distance between nodes (Excoffier and Smouse 1994). Both reticulation, in cases with several possible connections due to homoplasy, recombination or hybridization, complicate rendering of MSNs and more detail can be found in the book authored by Milgroom (2015). Networks with reticulation provide the ability to show clusters of genetically like individuals (Kamvar et al. 2015a).

A more recent, rapidly adopted model-free ordination technique called discriminant analysis of principal components, can be used to visualize population structure (Jombart et al. 2010). This analysis can begin with identification of groups within the data based on K-means clustering. By reducing the dimensionality of the data to select principal components, one can then determine how well the data explain predefined groups via discriminant analysis on an unlimited number of markers, maximizing the variation among groups, while minimizing the within-group variation. Results are typically shown in a two-dimensional ordination scatter plot for the first two principal components.

CLONAL POPULATIONS VIOLATE SOME COMMON ASSUMPTIONS

Clonal or partially clonal species provide some complications for population genetic analysis commonly encountered in plant pathology and microbial ecology. Most importantly, many assumptions made during analyses might be violated. Genetic theory was built on idealized population such as the Wright-Fisher model that make the mathematical analyses feasible (Box 3). For many plant pathogens that may have both clonal and sexual reproduction, analyzing data both before and after clone correction can provide insight into the contributions of each reproductive mode to genetic structure of the resulting population. Figure 4 reveals some differences of populations of the potato late blight pathogen *Phytophthora infestans* in clonal (left) versus sexual (right) populations (Goss et al. 2014). Clone correction collapses samples into one observation per multilocus genotype (Fig. 4B and H). These types of comparisons are recommended for all microbial populations that may have both a sexual and asexual mode of reproduction. Several groups have developed tools for working with clonal organisms (Ali et al. 2016; Arnaud-Haond et al. 2007; Bailleul et al. 2016; Kamvar et al. 2014, 2015a).

BOX 3
THE WRIGHT-FISHER MODEL

A basic concept in population genetics is the Wright-Fisher model (Hartl and Clark 1997; Wakeley 2008). Many of the statistical analyses employed in population genetics make assumptions that a population is evolving as a Wright-Fisher population or test for deviations from this model. In its simplest form, it consists of a single locus with two alleles in a population of constant finite size and does not incorporate selection or mutation. At a specified time interval (i.e., one season, one generation) all individuals randomly mate and then die, leaving only the new, nonoverlapping generation. Through the process of random mating one allele may produce more offspring in one generation due to drift. Due to the finite population size, this means that the other allele must decrease in abundance to accommodate the other's increase. Through time one allele eventually might become extinct resulting in the system moving to a state of fixation, where only one allele exists. Because this is a stochastic model it is frequently proposed as a null hypothesis when testing for other scenarios, such as whether selection has favored an allele. This model is named after Sewall Wright and Ronald Fisher, two individuals who set much of our foundation in theoretical population genetics.

THE SPECIAL CASES OF RECOMBINATION AND HYBRIDS

Recombination is the independent assortment of DNA sequences between different genomes, typically through sexual recombination, but potentially also through hybridization or horizontal gene transfer. When recombination occurs, a single haplotype can have two DNA regions with different ancestry where there is no single evolutionary history of descent. Methods to detect recombination include phylogenetic incompatibility testing between loci using the four-gamete test (Hudson and Kaplan 1985) or network analysis with emphasis on finding reticulation indicative of different ancestries (Milgroom et al. 2014; Posada and Crandall 2001). In all cases, several methods should be used to determine recombination including tree and network methods (Woolley et al. 2008).

BOX 4
SUGGESTED APPROACHES BASED ON COMMON PITFALLS ENCOUNTERED IN POPULATION GENETIC STUDIES

- **Test hypotheses.** Mere description of genotypic diversity does not suffice for publication in most journals. Researchers should strive toward testing rigorous, biological hypotheses based on the current knowledge of the biology of the organism.
- **Know your organism and design sampling strategies accordingly.** A hierarchical sampling is typically most informative for sexual populations and provides critical insights as to how genetic diversity is partitioned among (sub)populations (or strata, clusters, groups, etc.). In contrast, sampling clonal populations will not necessarily require a hierarchical sampling.
- **Provide evidence that allele calls are reproducible.** This is particularly important for any marker systems that might be subject to contamination, PCR error, electrophoresis error (e.g., obligate pathogens, random amplified polymorphic DNA, amplified fragment length polymorphism, etc.), or bioinformatic allele calling errors (genotyping-by-sequencing, RADseq, resequencing). This is particularly important for the current technology, namely whole genome or reduced-representation genome sequencing, given issues with missing allele calls, sequencing error, and imputation.
- **Understand the assumptions of each analysis method.** For example, STRUCTURE (Pritchard et al. 2000) assumes Hardy-Weinberg equilibrium (Box 1) and that markers are unlinked, an assumption that is violated in clonal populations. While assumptions can be violated to some degree, authors should investigate if results are sensitive to violation of assumptions by using independent model free methods (e.g., minimum spanning networks). Clonality requires analyses with and without clone-correction.
- **Avoid redundant analyses.** Showing the same figures or tables analyzed in several different ways for analogous methods is redundant. For example, it is not necessary to show analogous clustering methods (K-means clustering, genetic distance-based dendrogram, STRUCTURE plot, minimum-spanning-network, etc.) in multiple figures and/or tables. Ideally, each figure or table should test non-identical (but if appropriate, related), specific hypothesis. Conversely, during exploratory data analysis, conducting all possible analyses is encouraged.

In this review, we define a hybrid organism as offspring derived from two different species (Stukenbrock 2016). Hybrids can in most cases be recognized for diploid or polyploid species by having two (instead of one) most common recent ancestors. This hybrid state can be detected by sequencing and cloning genetic loci and showing polyphyletic ancestry. For example, the hybrid species *Phytophthora andina* shows independent segregation of haplotypes at nuclear loci that do not share the same most recent ancestor which can only be explained by a hybrid origin (Goss et al. 2011). Increased heterozygosity is typically also observed. The story can however be more complex. Following the process of hybridization, any hybrid can segregate into different lineages that gradually loose portions of the parental genomes and might eventually revert to a diploid state. At this state, detection of a hybrid might be more difficult as the signal of alleles deriving from different ancestral parents might be lost across large parts of the genome. Finally, horizontal gene transfer (HGT) can be seen as a similar case where a DNA sequence with unrelated ancestry is inserted into a non-homologous genome.

RESOURCES

This review is accompanied with web-based resources to reproduce some of the analyses discussed above that can be conducted in R. The code to reproduce Figure 4 is provided on github (https://github.com/grunwaldlab/popgen_review_examples). A more extensive primer on conducting population genetics in R is available online providing numerous examples on how to reproduce other aspects of the work presented in this paper (Grünwald et al. 2016a; https://grunwaldlab.github.io/Population_Genetics_in_R/).

CONCLUSIONS AND OUTLOOK

In closing, it is hoped that the reader finds this review useful in designing and conducting rigorous analyses of the genetic structure of populations. Including hypothesis-driven research, a good sampling strategy based on knowledge of the biology, combined with rigorous controls and careful data analysis will provide the basis for good population genetic research. Box 4 provides a quick overview of the most notable pitfalls encountered most often in the literature.

Two aspects that were not covered here are gaining rapid prominence. First, coalescent analyses provide powerful tools that go beyond the scope of this review yet provide complementary approaches to those discussed here (Carbone et al. 2004; Carbone and Kohn 2001; Drummond et al. 2005; Goss 2015; Goss et al. 2009a; Grünwald and Goss 2011; Hudson 1990). Second, the field of population genetics is now increasingly relying on high throughput sequencing technologies to genotype individuals at thousands of SNPs using for example GBS or RADseq (Andrews and Luikart 2014; Davey et al. 2011; Elshire et al. 2011; Grünwald et al. 2016b; Luikart et al. 2003; Vinatzer et al. 2014; Weigel and Nordborg 2015). With GBS, scientists must rethink their toolbox given a range of new challenges such as removal of linked SNPs, massive amounts of missing data, imputation, ensuring appropriate allele calls, and use of reference genomes among others (Grünwald et al. 2016b).

ACKNOWLEDGMENTS

We thank three reviewers for providing insightful and constructive feedback on our first draft that much improved our review. This work was supported by the U.S. Department of Agriculture (USDA) Agricultural Research Service Grant 5358-22000-039-00D, USDA National Institute of Food and Agriculture Grant 2011-6800430154, the USDA-ARS Floriculture Nursery Initiative, the Oregon Department of Agriculture/Oregon Association of Nurseries (ODA-OAN) research programs, and a USDA NIFA 2012-67012-19844 to SEE.

LITERATURE CITED

Adamack, A. T., and Gruber, B. 2014. PopGenReport: Simplifying basic population genetic analyses in R. Methods Ecol. Evol. 5:384-387.

Ali, S., Gautier, A., Leconte, M., Enjalbert, J., and de Vallavieille-Pope, C. 2011. A rapid genotyping method for an obligate fungal pathogen, *Puccinia striiformis* f. sp. *tritici*, based on DNA extraction from infected leaf and multiplex PCR genotyping. BMC Res. Notes 4:240.

Ali, S., Soubeyrand, S., Gladieux, P., Giraud, T., Leconte, M., Gautier, A., Mboup, M., Chen, W., de Vallavieille-Pope, C., and Enjalbert, J. 2016. *Cloncase*: Estimation of sex frequency and effective population size by clonemate resampling in partially clonal organisms. Mol. Ecol. Resour. 16: 845-861.

Andrews, K. R., and Luikart, G. 2014. Recent novel approaches for population genomics data analysis. Mol. Ecol. 23:1661-1667.

Archer, F. I., Adams, P. E., and Schneiders, B. B. 2017. *Stratag*: An R package for manipulating, summarizing and analysing population genetic data. Mol. Ecol. Resour. 17:5-11.

Arnaud-Haond, S., Duarte, C. M., Alberto, F., and Serrão, E. A. 2007. Standardizing methods to address clonality in population studies. Mol. Ecol. 16: 5115-5139.

Atallah, Z. K., Maruthachalam, K., du Toit, L., Koike, S. T., Michael Davis, R., Klosterman, S. J., Hayes, R. J., and Subbarao, K. V. 2010. Population analyses of the vascular plant pathogen *Verticillium dahliae* detect recombination and transcontinental gene flow. Fungal Genet. Biol. 47:416-422.

Avise, J. C. 1994. Molecular Markers, Natural History and Evolution: Natural History and Evolution. Springer.

Bailleul, D., Stoeckel, S., and Arnaud-Haond, S. 2016. *RClone*: A package to identify multilocus clonal lineages and handle clonal datasets in R. Methods Ecol. Evol. 7:966-970.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376.

Baumgartner, K., Travadon, R., Bruhn, J., and Bergemann, S. E. 2010. Contrasting patterns of genetic diversity and population structure of *Armillaria mellea* sensu stricto in the eastern and western United States. Phytopathology 100:708-718.

Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., Thines, M., Ah-Fong, A., Anderson, R., Badejoko, W., Bittner-Eddy, P., Boore, J. L., Chibucos, M. C., Coates, M., Dehal, P., Delehaunty, K., Dong, S., Downton, P., Dumas, B., Fabro, G., Fronick, C., Fuerstenberg, S. I., Fulton, L., Gaulin E., Govers, F., Hughes, L., Humphray, S., Jiang, R. H., Judelson, H., Kamoun, S., Kyung, K., Meijer, H., Minx, P., Morris, P., Nelson, J., Phuntumart, V., Qutob, D., Rehmany, A., Rougon-Cardoso, A., Ryden, P., Torto-Alalibo, T., Studholme, D., Wang, Y., Win, J., Wood, J., Clifton, S. W., Rogers, J., Van den Ackerveken, G., Jones, J. D., McDowell, J. M., Beynon, J., and Tyler, B. M.2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. Science 330:1549-1551.

Berbegal, M., Perez-Sierra, A., Armengol, J., and Grünwald, N. J. 2013. Evidence for multiple introductions and clonality in Spanish populations of *Fusarium circinatum*. Phytopathology 103:851-861.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., and Taberlet, P. 2004. How to track and assess genotyping errors in population genetics studies. Mol. Ecol. 13:3261-3273.

Brown, A. H. D. 1975. Sample sizes required to detect linkage disequilibrium between two or three loci. Theor. Popul. Biol. 8:184-201.

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. Mol. Ecol. 13:2101-2106.

Byrnes, E. J., Li, W., Lewit, Y., Ma, H., Voelz, K., Ren, P., Carter, D. A., Chaturvedi, V., Bildfell, R. J., May, R. C., and Heitman, J. 2010. Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. PLoS Pathog. 6:e1000850.

Carbone, I., and Kohn, L. 2004. Inferring process from pattern in fungal population genetics. Appl. Mycol. Biotechnol. 4:29-58.

Carbone, I., and Kohn, L. M. 2001. A microbial population-species interface: Nested cladistic and coalescent inference with multilocus data. Mol. Ecol. 10:947-964.

Carbone, I., Liu, Y.-C., Hillman, B. I., and Milgroom, M. G. 2004. Recombination and migration of *Cryphonectria* hypovirus 1 as inferred from gene genealogies and the coalescent. Genetics 166:1611-1629.

Clark, L. V., and Jasieniuk, M. 2011. *Polysat*: An R package for polyploid microsatellite analysis. Mol. Ecol. Resour. 11:562-566.

Csardi, G., and Nepusz, T. 2006. The *igraph* software package for complex network research. InterJournal Complex Syst. 1695.

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. Nat. Rev. Genet. 12:499-510.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22:1185-1192.

Dunn, A. R., Bruening, S. R., Grünwald, N. J., and Smart, C. D. 2014. Evolution of an experimental population of *Phytophthora capsici* in the field characterized by mutation and recombination. Phytopathology 104:1107-1117.

Dutech, C., Fabreguettes, O., Capdevielle, X., and Robin, C. 2010. Multiple introductions of divergent genetic lineages in an invasive fungal pathogen, *Cryphonectria parasitica*, in France. Heredity 105:220-228.

Edwards, A. W. F. 1971. Distances between populations on the basis of gene frequencies. Biometrics 27:873-881.

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.

Evanno, G., Regnaut, S., and Goudet, J. 2005. Detecting the number of clusters of individuals using the software *STRUCTURE*: A simulation study. Mol. Ecol. 14:2611-2620.

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

Excoffier, L., and Smouse, P. E. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics 136:343-359.

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.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Geiser, D. M., Pitt, J. I., and Taylor, J. W. 1998. Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*. Proc. Nat. Acad. Sci. 95:388-393.

Giraud, T., Gladieux, P., and Gavrilets, S. 2010. Linking the emergence of fungal plant diseases with ecological speciation. Trends Ecol. Evol. 25: 387-395.

Goodwin, S. B. 1997. The population genetics of *Phytophthora*. Phytopathology 87:462-473.

Goss, E. M. 2015. Genome-enabled analysis of plant-pathogen migration. Annu. Rev. Phytopathol. 53:121-135.

Goss, E. M., Carbone, I., and Grünwald, N. J. 2009a. Ancient isolation and independent evolution of the three clonal lineages of the exotic sudden oak death pathogen *Phytophthora ramorum*. Mol. Ecol. 18:1161-1174.

Goss, E. M., Cardenas, M. E., Myers, K., Forbes, G. A., Fry, W. E., Restrepo, S., and Grünwald, N. J. 2011. The plant pathogen *Phytophthora andina* emerged via hybridization of an unknown *Phytophthora* species and the Irish potato famine pathogen, *P. infestans*. PLoS One 6:e24543.

Goss, E. M., Larsen, M., Chastagner, G. A., Givens, D. R., and Grünwald, N. J. 2009b. Population genetic analysis infers migration pathways of *Phytophthora ramorum* in U.S. nurseries. PLoS Pathog. 5:e1000583.

Goss, E. M., Tabima, J. F., Cooke, D. E. L., Restrepo, S., Fry, W. E., Forbes, G. A., Fieland, V. J., Cardenas, M., and Grünwald, N. J. 2014. The Irish potato famine pathogen *Phytophthora infestans* originated in central Mexico rather than the Andes. Proc. Natl. Acad. Sci. 111:8791-8796.

Grünwald, N. J., and Flier, W. G. 2005. The biology of *Phytophthora infestans* at its center of origin. Annu. Rev. Phytopathol. 43:171-190.

Grünwald, N. J., Garbelotto, M., Goss, E. M., Heungens, K., and Prospero, S. 2012. Emergence of the sudden oak death pathogen *Phytophthora ramorum.* Trends Microbiol. 20:131-138.

Grünwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. 2003. Analysis of genotypic diversity data for populations of microorganisms. Phytopathology 93:738-746.

Grünwald, N. J., and Goss, E. M. 2011. Evolution and population genetics of exotic and re-emerging pathogens: Novel tools and approaches. Annu. Rev. Phytopathol. 49:249-267.

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

Grünwald, N. J., Kamvar, Z. N., and Everhart, S. E. 2016a. Population Genetics in R. doi:10.5281/zenodo.160588

Grünwald, N. J., McDonald, B. A., and Milgroom, M. G. 2016b. Population genomics of fungal and oomycete pathogens. Annu. Rev. Phytopathol. 54: 323-346.

Hale, M. L., Burg, T. M., and Steeves, T. E. 2012. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. PLoS One 7:e45170.

Hartl, D. L., and Clark, A. G. 1997. Principles of Population Genetics, 3rd ed. Sinauer Associates, Sunderland, MA.

Hoban, S., Gaggiotti, O., Congress Consortium, and Bertorelle, G. 2013. Sample planning optimization tool for conservation and population genetics (*SPOTG*): A software for choosing the appropriate number of markers and samples. Methods Ecol. Evol. 4:299-303.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. 2009. Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. 9:1322-1332.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxford Surveys Evol. Biol. 7:44.

Hudson, R. R., and Kaplan, N. L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147-164.

James, T. Y., Litvintseva, A. P., Vilgalys, R., Morgan, J. A. T., Taylor, J. W., Fisher, M. C., Berger, L., Weldon, C., du Preez, L., and Longcore, J. E. 2009. Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. PLoS Pathog. 5:e1000458.

Jombart, T. 2008. *Adegenet*: An R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.

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

Jombart, T., Devillard, S., and Balloux, F. 2010. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC Genet. 11:94.

Jones, C. J., Edwards, K. J., Castaglione, S., Winfield, M. O., Sala, F., and van de Wiel, C. 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol. Breed. 3:381-390.

Kamvar, Z. N., Brooks, J. C., and Grünwald, N. J. 2015a. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. Front. Genet. 6:208.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. 2015b. Spatial and temporal analysis of populations of the sudden oak death pathogen in Oregon forests. Phytopathology 105:982-989.

Kamvar, Z. N., López-Uribe, M. M., Coughlan, S., Grünwald, N. J., Lapp, H., and Manel, S. 2017. Developing educational resources for population genetics in R: An open and collaborative approach. Mol. Ecol. Resour. 17: 120-128.

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

Knaus, B. J., and Grünwald, N. J. 2017. *VcfR*: An R package to manipulate and visualize VCF format data. Mol. Ecol. Resour. 17:44-53.

Kosman, E., and Leonard, K. J. 2005. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. Mol. Ecol. 14:415-424.

Linde, C. C., Zhan, J., and McDonald, B. A. 2002. Population structure of *Mycosphaerella graminicola*: from lesions to continents. Phytopathology 92:946-955.

Ludwig, J. A. 1988. Statistical Ecology: A Primer on Methods and Computing. Wiley, New York.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. 4:981-994.

Malvarez, G., Carbone, I., Grünwald, N. J., Subbarao, K. V., Schafer, M., and Kohn, L. M. 2007. New populations of *Sclerotinia sclerotiorum* from lettuce in California and peas and lentils in Washington. Phytopathology 97: 470-483.

McDonald, B. A. 1997. The population genetics of fungi: Tools and techniques. Phytopathology 87:448-453.

Meirmans, P. G., and Van Tienderen, P. H. 2004. Genotype and Genodive: Two programs for the analysis of genetic diversity of asexual organisms. Mol. Ecol. Notes 4:792-794.

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

Milgroom, M. G. 1996. Recombination and the multilocus structure of fungal populations. Annu. Rev. Phytopathol. 34:457-477.

Milgroom, M. G. 2015. Population Biology of Plant Pathogens: Genetics, Ecology, and Evolution. American Phytopathological Society, St. Paul, MN.

Milgroom, M. G., Jiménez-Gasco, M. M., Olivares García, C., Drott, M. T., and Jiménez-Díaz, R. M. 2014. Recombination between clonal lineages of the asexual fungus *Verticillium dahliae* detected by genotyping by sequencing. PLoS One 9:e106740.

Mundt, C. C., Sackett, K. E., and Wallace, L. D. 2011. Landscape heterogeneity and disease spread: experimental approaches with a plant pathogen. Ecol. Appl. 21:321-328.

Mundt, C. C., Sackett, K. E., Wallace, L. D., Cowger, C., and Dudley, J. P. 2009. Long-distance dispersal and accelerating waves of disease: Empirical relationships. Am. Nat. 173:456-466.

Nei, M. 1972. Genetic distance between populations. Am. Nat. 106:283-292.
Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. 70:3321-3323.
Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583-590.
Nei, M., and Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Nat. Acad. Sci. 76: 5269-5273.
Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., Simpson, G. L., et al. 2013. *Vegan*: Community ecology package. R package version 2.0-7. https://cran.r-project.org
Paradis, E., Claude, J., and Strimmer, K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289-290.
Paradis, E., Gosselin, T., Goudet, J., Jombart, T., and Schliep, K. 2017. Linking genomics and population genetics with R. Mol. Ecol. Resour. 17: 54-66.
Parobek, C. M., Archer, F. I., DePrenger-Levin, M. E., Hoban, S. M., Liggins, L., and Strand, A. E. 2017. *SkeleSim*: An extensible, general framework for population genetic simulation in R. Mol. Ecol. Resour. 17:101-109.
Peakall, R., and Smouse, P. E. 2012. *GenAlEx* 6.5: Genetic analysis in Excel. Population genetic software for teaching and research–an update. Bioinformatics 28:2537-2539.
Peever, T. L., Salimath, S. S., Su, G., Kaiser, W. J., and Muehlbauer, F. J. 2004. Historical and contemporary multilocus population structure of *Ascochyta rabiei* (teleomorph: *Didymella rabiei*) in the Pacific Northwest of the United States. Mol. Ecol. 13:291-309.
Posada, D., and Crandall, K. A. 2001. Intraspecific gene genealogies: Trees grafting into networks. Trends Ecol. Evol. 16:37-45.
Prevosti, A., Ocaña, J., and Alonso, G. 1975. Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. Theor. Appl. Genet. 45:231-241.
Pritchard, J., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.
R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reynolds, J., Weir, B. S., and Cockerham, C. C. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. Genetics 105:767-779.
Rogers, J. S. 1972. Measures of genetic similarity and genetic distances. Pages 145-153 in: Studies in Genetics. University of Texas Publishers, TX.
Rojas, A. J., Jacobs, J. L., Napieralski, S., Karaj, B., Bradley, C. A., and Chase, T. 2017. Oomycete species associated with soybean seedlings in North America. Part I: Identification and pathogenicity characterization. Phytopathology 107:280-292.
Schlötterer, C. 2004. The evolution of molecular markers–just a matter of fashion? Nat. Rev. Genet. 5:63-69.
Schoebel, C. N., Stewart, J., Grünwald, N. J., Rigling, D., and Prospero, S. 2014. Population history and pathways of spread of the plant pathogen *Phytophthora plurivora*. PLoS One 9:e85368.
Shannon, C. E., and Weaver, W. 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.
Smouse, P. E., and Peakall, R. 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561-573.
Stanley, R. R. E., Jeffery, N. W., Wringe, B. F., DiBacco, C., and Bradbury, I. R. 2017. *Genepopedit*: a simple and flexible tool for manipulating multilocus molecular data in R. Mol. Ecol. Resour. 17:12-18.
Stoddart, J. A., and Taylor, J. F. 1988. Genotypic diversity: Estimation and prediction in samples. Genetics 118:705-711.
Stukenbrock, E. H. 2016. The role of hybridization in the evolution and emergence of new fungal plant pathogens. Phytopathology 106:104-112.
Stukenbrock, E. H., Banke, S., Javan-Nikkhah, M., and McDonald, B. A. 2007. Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. Mol. Biol. Evol. 24:398-411.
Summers, C. F., Gulliford, C. M., Carlson, C. H., Lillis, J. A., Carlson, M. O., and Cadle-Davidson, L. 2015. Identification of genetic variation between obligate plant pathogens *Pseudoperonospora cubensis* and *P. humuli* using RNA sequencing and genotyping-by-sequencing. PLoS One 10:e0143665.
Vinatzer, B. A., Monteil, C. L., and Clarke, C. R. 2014. Harnessing population genomics to understand how bacterial pathogens emerge, adapt to crop hosts, and disseminate. Annu. Rev. Phytopathol. 52:19-43.
Vleeshouwers, V. G. A. A., and Oliver, R. P. 2014. Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. Mol. Plant-Microbe Interact. 27:196-206.
Vleeshouwers, V. G. A. A., Raffaele, S., Vossen, J. H., Champouret, N., Oliva, R., and Segretin, M. E. 2011. Understanding and exploiting late blight resistance in the age of effectors. Annu. Rev. Phytopathol. 49:507-531.
Wakeley, J. 2008. Coalescent Theory. Roberts & Company, Greenwood Village, CO.
Weigel, D., and Nordborg, M. 2015. Population genomics for understanding adaptation in wild plant species. Annu. Rev. Genet. 49:315-338.
Weir, B. S. 1996. Genetic data analysis II: Methods for discrete population genetic data. Sinauer Associates, Sunderland, MA.
Weir, B. S., and Cockerham, C. C. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.
Wickham, H. 2009. *Ggplot2*: Elegant graphics for data analysis. Springer.
Winter, D. J. 2012. *Mmod*: An R library for the calculation of population differentiation statistics. Mol. Ecol. Resour. 12:1158-1160.
Woolley, S. M., Posada, D., and Crandall, K. A. 2008. A comparison of phylogenetic network methods using computer simulation. PLoS One 3:e1913.
Zhan, J., Mundt, C. C., and McDonald, B. A. 1998. Measuring immigration and sexual reproduction in field populations of *Mycosphaerella graminicola.* Phytopathology 88:1330-1337.