American Genetic Association

Journal of Heredity, 2018, 283–296
doi:10.1093/jhered/esy006
Invited Review
Advance Access publication 27 January 2018

OXFORD

Invited Review

The Analysis of Polyploid Genetic Data

Patrick G. Meirmans, Shenglin Liu, and Peter H. van Tienderen

From the Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, PO Box 94248, NL-1090 GE Amsterdam, the Netherlands (Meirmans and van Tienderen); and the [2]Department of Bioscience, Aarhus University, Aarhus, Denmark (Liu).

Address correspondence to Patrick G. Meirmans, University of Amsterdam, Science Park 904, 1098XH Amsterdam, the Netherlands, or e-mail: p.g.meirmans@uva.nl.

Received August 28, 2017; First decision October 20, 2017; Accepted January 20, 2018.

Corresponding Editor: Fred Allendorf

Abstract

Though polyploidy is an important aspect of the evolutionary genetics of both plants and animals, the development of population genetic theory of polyploids has seriously lagged behind that of diploids. This is unfortunate since the analysis of polyploid genetic data—and the interpretation of the results—requires even more scrutiny than with diploid data. This is because of several polyploidy-specific complications in segregation and genotyping such as tetrasomy, double reduction, and missing dosage information. Here, we review the theoretical and statistical aspects of the population genetics of polyploids. We discuss several widely used types of inferences, including genetic diversity, Hardy–Weinberg equilibrium, population differentiation, genetic distance, and detecting population structure. For each, we point out how the statistical approach, expected result, and interpretation differ between different ploidy levels. We also discuss for each type of inference what biases may arise from the polyploid-specific complications and how these biases can be overcome. From our overview, it is clear that the statistical toolbox that is available for the analysis of genetic data is flexible and still expanding. Modern sequencing techniques will soon be able to overcome some of the current limitations to the analysis of polyploid data, though the techniques are lagging behind those available for diploids. Furthermore, the availability of more data may aggravate the biases that can arise, and increase the risk of false inferences. Therefore, simulations such as we used throughout this review are an important tool to verify the results of analyses of polyploid genetic data.

Subject area: phylogeography,population structure
Keywords: autopolyploidy, genetic differentiation, genetic diversity, Hardy–Weinberg, population structure, tetrasomy

During their evolution, many plant and animal taxa have gone through one or several rounds of genome duplication (Ramsey and Schemske 1998). Even the genome of *Arabidopsis thaliana*, one of the smallest genomes among the angiosperms, shows evidence for several rounds of whole genome duplication (Bomblies and Madlung 2014). Polyploidy also plays a direct role in the speciation process: an estimated 2–4% of speciation events in flowering plants and 7% in ferns are thought to be the result of polyploidization (Otto and Whitton 2000). Furthermore, there are many species—especially in plants—in which multiple cytotypes are present. These cytotypes may show striking geographic distributions (Suda et al. 2007; Kolář et al. 2017) and may differ in life-history characteristics, such as mating system (Meirmans et al. 2006; Husband et al. 2008) and ecological preferences (Glennon et al. 2014; Parisod and Broennimann 2016).

At its most basic, evolution is simply a shift in allele frequencies. Therefore, the action of evolution depends heavily on the processes

that influence the distribution of genetic variation within species, most importantly genetic drift, breeding system, mode of inheritance, segregation patterns, migration, mutation, and selection. In polyploids, these processes work differently, resulting in different patterns in the distribution of genetic variation. So when we analyze such patterns to make inferences about the underlying processes, the results of the analysis needs to be interpreted differently for polyploids than for diploids. Though the theoretical work on the population genetics of polyploids goes back to Haldane (1930), there is little literature on the subject relative to what is available for diploids: even though there are multiple (sometimes book-length) overviews of the subject for diploids (e.g., Nei 1987; Holsinger and Weir 2009), a comprehensive overview of the theoretical population genetics of polyploids is lacking.

Apart from the theoretical aspects, there are practical problems that complicate the analysis of polyploid genetic data (Dufresne et al. 2014): the tools and theory that have been developed for diploids do not necessarily work for polyploids; when they do work the results may be biased and/or the interpretation of the results may be different. Biases in the analysis of polyploid genetic data may arise from several processes that are specific to the segregation during meiosis in polyploids and some practical issues that are only present in polyploids.

First, species may show segregation patterns that are partly disomic and partly polysomic (Ramsey and Schemske 2002; Stift et al. 2008; Chester et al. 2012). These segregation patterns may differ among chromosomes, or even along chromosomes, for example, in ancient allopolyploids that underwent partial re-diploidization (Allendorf et al. 2015; Limborg et al. 2017). In such "segmental allopolyploids" (Stebbins 1947), the estimates of genetic summary statistics may be biased, with the size of the bias depending on the frequency of tetrasomy (Meirmans and van Tienderen 2013). Second, segregation in polyploids can result in the occurrence of double reduction, where 2 copies of the same chromatid segment end up in the same gamete (Bever and Felber 1992; Ronfort et al. 1998). The most important effect of double reduction is that it increases the homozygosity at a locus: with double reduction even a tetraploid with a fully heterozygous genotype (*ABCD*) produces some homozygous gametes (*AA, BB, CC, DD*) next to the 6 expected heterozygous gametes. Third, species may show mixed ploidy levels; including these in a single data set may give rise to bias since the expectations are different for different ploidy levels. Most complicated in this respect are those species where there is ploidy variation within a single genome (Allendorf et al. 2015).

A practical problem specific to genetic analyses of polyploids is the difficulty to obtain the dosage of alleles for otherwise co-dominant markers, so that different partial heterozygous genotypes (e.g., AABC, ABBC, ABCC) cannot be distinguished. A recent review (Dufresne et al. 2014) gave an extensive overview of these and other issues, how they can be avoided, and which statistical tools are available. However, they did not go into great depth about how the theoretical expectations are different for polyploids nor about the extent of the biases that are present because of these polyploidy-specific complications.

In this article, we review the theoretical consequences of polyploidy on basic population genetic processes, and hence the interpretation of different analyses. Mostly focusing on autopolyploids, we will give theoretical expectations for basic diversity parameters such as H_S, F_{ST}, and related statistics and discuss how these can be estimated for polyploid data. Then, we outline how polyploidy affects some other popular types of inferences, such as the detection of population structure, principal components analysis, and assignment tests. Throughout, we discuss how those polyploidy-specific problems affect the results and interpretation of the analysis. Most importantly, we show when and how the results of certain analyses may be biased when applied to polyploid data, and whether it is possible to avoid or correct for such a bias.

Genetic Diversity

The Effect of Polyploidy

An autopolyploid population can harbor a larger amount of genetic diversity than an equally large diploid population. This is because the polyploid population contains a larger total number of chromosome copies than the diploid population. A tetraploid population of size N has a total of $4N$ chromosome copies at an autosomal locus, while diploids have $2N$ copies. Therefore, at a mutation rate of μ the number of mutations per generations is twice as high in the tetraploid population: $4N\mu$ versus $2N\mu$. Next to the increased number of mutations, the larger number of chromosome copies also affects the strength of genetic drift, which is reduced in a population of polyploids. Under mutation-drift equilibrium, the combination of both effects results in a higher level of genetic diversity in polyploids. The most widely used index for measuring the level of genetic diversity is the gene diversity, more commonly referred to as the expected heterozygosity (H_S), which is equal to the probability that 2 randomly picked alleles are not identical in state (Nei 1987). Under the assumptions that $N \gg 1$ and $kN\mu^2 \ll 1$, Moody et al. (1993) found the following generalization for the equilibrium value of H_S for any ploidy level k:

$$H_{S(k)} = 1 - \frac{1}{1 + 2kN\mu} \quad (1)$$

The relationship between $H_{S(k)}$ and $N\mu$ is shown in Figure 1a for 3 different ploidy levels. From this figure, we can see that the difference between ploidy levels is expected to be highest for small values of $N\mu$. This means that for markers with low mutation rates the level of diversity in a tetraploid population can be up to twice as high as that in a similar population of diploids. For markers with high mutation rates, diploids will only be slightly less diverse than tetraploids. When comparing lower ploidy levels, the differences tend to be larger than when comparing higher ploidy levels; the difference between diploids and tetraploids is larger than the difference between tetraploids and octaploids, even though both comparisons involve a doubling of the genome. It is also interesting to note that in diploid species with sex chromosomes, there can be a substantial difference between Y/W-chromosomes (k = ½) and autosomes (k = 2), even for markers with high mutation rates (Figure 1b).

Estimation and Bias in Diversity Estimates

The higher ploidy level has to be taken into account when estimating genetic diversity using the expected heterozygosity (H_S) and the observed heterozygosity (H_O). In diploids, it is common to calculate the expected heterozygosity at a locus by taking the complement of the expected frequency of homozygotes: $H_S = 1 - \sum p_i^2$. In polyploids with ploidy k, this approach would amount to $1 - \sum p_i^k$. However, this approach has little use in polyploids as it lumps all heterozygotes—both full and partial—into one single class. The result would also be of very limited use as an estimator for genetic diversity: the same set of allele frequencies would yield a higher diversity for tetraploids than for diploids. The solution is to estimate

the level of genetic diversity by calculating the expected heterozygosity always as if the species were diploid: $H_S = 1 - \sum p_i^2$. The rationale behind this is that the use of H_S as an estimator for genetic diversity is not tied to heterozygosity *per se*; the exact same equation is used in ecology to quantify the species diversity in ecological communities (Simpson 1949). For this reason, Nei (1987) preferred to refer to this statistic as the "gene diversity" to illustrate its independence of the ploidy level. For the estimation of H_S, it is important to include a correction to avoid bias stemming from small sample sizes; for diploids usually the method from Nei and Chesser (1983) is used; Hardy (2015) gives a generic method that is applicable to any ploidy level.

The presence of both full and partial heterozygotes also needs to be taken into account when calculating the observed heterozygosity. Here, the different heterozygous genotypes should be weighed by their degree of heterozygosity; the full heterozygote *ABCD* should have a higher weight than the partial heterozygote *AAAB*. For this, Moody et al. (1993) proposed calculating H_O using the concept of "gametic heterozygosity." For a tetraploid, the gametic heterozygosity of a genotype is defined as the frequency of heterozygotes among randomly sampled diploid gametes formed from the 4 allele copies present at a locus. So the fully heterozygous genotype *ABCD* has a gametic heterozygosity of 1 since—in the absence of double reduction—all gametes produced by this genotype will be heterozygous. In contrast, genotype *AAAB* has a gametic heterozygosity of 0.5 as only half the produced gametes will be heterozygous. The other 2 possible genotypes for a tetraploid, *AABB* and *AABC*, have gametic heterozygosities of 0.67 and 0.83, respectively. The concept of calculating gametic heterozygosity by drawing diploid gametes can be generalized to other ploidy levels, even though these ploidy levels do not actually produce diploid gametes. Even though this is not realistic for higher ploidy levels, the benefit of calculating the observed heterozygosity in this way is that it is possible to directly compare the observed heterozygosity to the expected heterozygosity as defined above.

When the dosage information is missing, and only phenotypes are available, calculating estimates of genetic diversity becomes more difficult. For H_S, the allele frequencies can be used when calculated with a proper correction for missing dosage (De Silva et al. 2005). This approach is implemented in the programs Polysat (Clark and Jasieniuk 2011) and GenoDive (Meirmans and van Tienderen 2004), and generally works well to remove the bias incurred from the

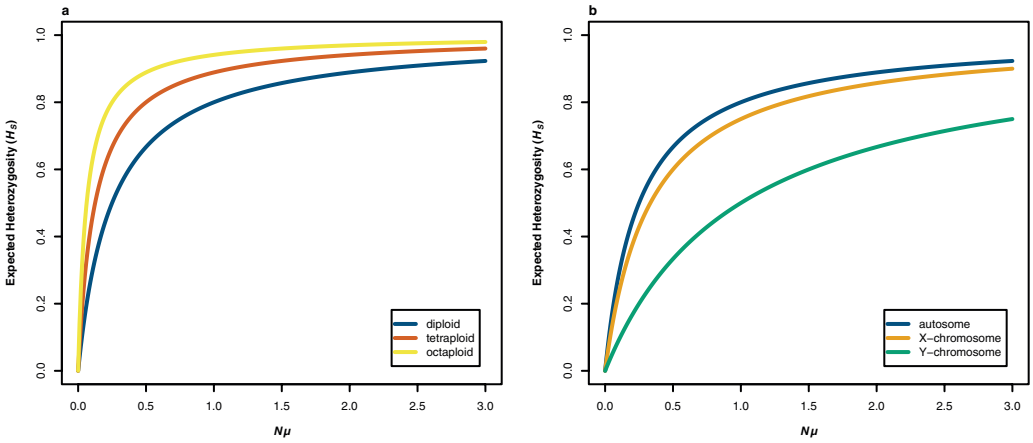

Figure 1. Theoretical values of the expected heterozygosity H_S of a population in mutation-drift equilibrium as a function of $N\mu$ for different ploidy levels (**a**) or for autosomes and sex chromosomes in a species with XY sex determination (**b**). See the Supplementary Data for the used R-script.

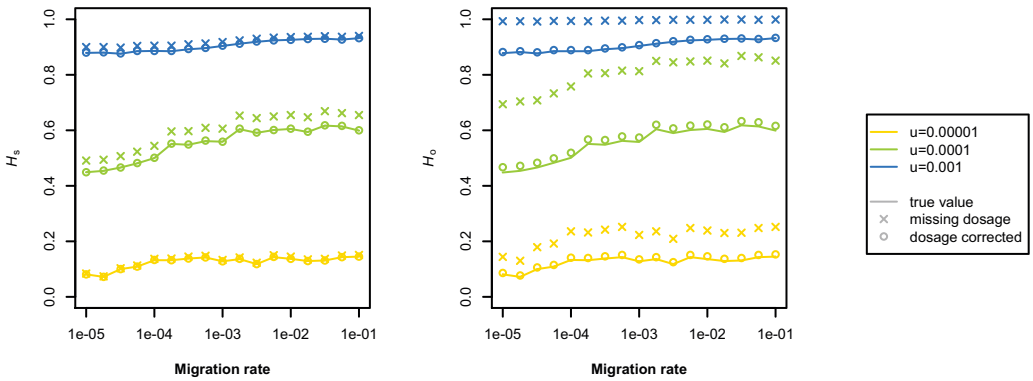

Figure 2. Bias from missing dosage in the estimation of H_S and H_O for different mutation rates and different migration rates. Data were simulated based on an island model of migration with 2 randomly mating tetraploid populations of size 1000. Migration and drift were simulated for 20000 generations at 200 loci, each with a K-allele mutation model with a maximum of 100 alleles. Afterwards, genotypes were drawn for 100 individuals per population and data sets were created both with and without dosage information. Statistics were calculated using GenoDive (Meirmans and van Tienderen 2004). For H_S, missing dosage was corrected using the method of DeSilva et al. (2005); for H_O the method of Hardy (2015) was used. See the Supplementary Data for the R-script used for the simulations.

missing data (Figure 2a). For H_O, Hardy (2015) suggested to average the values over the different partial heterozygotes that are possible for a given phenotype. For example, in a tetraploid the phenotype AB would have a gametic heterozygosity of 0.56, which is the average of the gametic heterozygosities for the genotypes *AAAB* (0.5), *AABB* (0.67), and *ABBB* (0.5). Using simulated data, Hardy (2015) then showed that the value of H_O calculated for phenotypic data only has a slightly higher error than when it was calculated for genotypic data. This is also shown in Figure 2b, based on simulated data from 2 tetraploid populations: without dosage correction (crosses), the estimates of H_O are much higher than the known true values (line); with dosage correction (circles) the values closely match the true values. However, since the simulations used here have only a limited scope of parameters—only random mating was simulated—one should always be careful in the interpretation of H_O values when dosage information is missing and preferably try to quantify the extent of any bias and/or error by performing simulations that match the study system.

Hardy–Weinberg Equilibrium

Hardy–Weinberg equilibrium (HWE) is the situation in which the observed heterozygosity in a population is equal to the expected heterozygosity. Testing for HWE is relevant as a test for random mating in a population, and also because conformation to HW-expectations is an important assumption for many other population genetic analyses (Waples 2015). Quantifying departure from HWE is generally done using Wright's inbreeding coefficient $F_{IS} = (H_S - H_O)/H_S$. This same equation can be used for polyploids using the H_O and H_S values calculated with the approaches discussed above. For diploids, it is well-known that, unless there are differences in allele frequencies between males and females, a single generation of random mating suffices to restore the genotype frequencies in a population to Hardy–Weinberg proportions ($F_{IS} = 0$). In polyploids, the situation is less simple. Imagine that a population of autotetraploids is completely lacking heterozygotes, consisting for 50% of genotype *AAAA* and 50% of genotype *BBBB*. The gametes produced by this population are all either *AA* or *BB*. A single generation of random mating then leads to a population with genotypes *AAAA*, *AABB*, and *BBBB*. Obviously, this population is not in HWE, since genotypes *AAAB* and *ABBB* are still completely missing. So instead of restoring HWE in a single generation, in polyploids random mating leads to a decay of the degree of HW disequilibrium. Under full polysomy and without double reduction, the value of F_{IS} at generation t is a function of the value at generation 0 and the ploidy level k:

$$F_{IS(t)} = F_{IS(0)} \cdot \left(\frac{0.5k-1}{k-1}\right)^t \quad (2)$$

Figure 3 shows that when starting with a complete lack of heterozygotes, about 6 generations of random mating will bring F_{IS} down to a value very close to zero. However, the rate of this decay depends on the rate of double reduction and the segregation pattern (Bever and Felber 1992). With double reduction, HWE will not be reached and the value of F_{IS} will remain higher than zero, even when mating is random among individuals. When the segregation is not full polysomic, the decay in F_{IS} will be slower than shown in Figure 3; in the extreme case of fully disomic inheritance, HWE is reached for the homologs but not for the homeologs.

Statistical testing of conformation to HW expectations is difficult in polyploids because the methods that are most frequently used for this in diploids are not available for polyploids. Chi-square and log-likelihood tests can often not be used because these tests do not work well when some of the expected genotype frequencies are very low, which is almost always the case in polyploids especially for the homozygous genotypes. Note that this problem of low expected frequencies is also present in diploids when multi-allelic markers such as microsatellites are used; using such markers in polyploids indeed exacerbates the problem. In diploids, it is common to instead use the Markov chain method of Guo and Thompson (1992), implemented in the program GenePop (Raymond and Rousset 1995). However, no software to perform such tests is available for polyploids. The best method to test for HWE in polyploids is to use F_{IS} as a test statistic in a Monte Carlo permutation test, constructing new data sets by randomly combining alleles into genotypes and recalculating F_{IS} for each generated data set. This generates a null distribution of F_{IS}, which can be used to assess the significance of the original F_{IS} value. Though HWE tests are a central tenet of population genetics, it is surprising that there are only 2 programs that can perform such a permutation test for polyploids: SPAGeDi (Hardy and Vekemans 2002) and GenoDive (Meirmans and van Tienderen 2004).

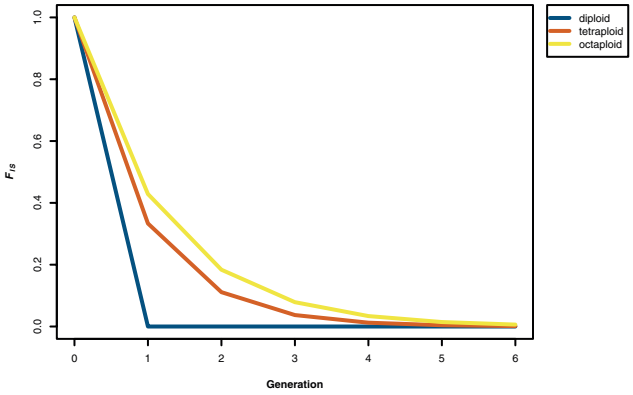

Figure 3. The decay in deviation from HWE as measured by F_{IS} as a function of the number of generations of random mating, starting from a complete lack of heterozygotes. See the Supplementary Data for the used R-script.

When dosage information is missing, statistical testing for HWE-conformation is not possible for polyploids. This is because all methods revolve, in some way or another, on information about the allele frequencies in the sampled populations. Getting unbiased estimates of the allele frequencies with missing dosage requires either an assumption of random mating or setting a fixed rate of selfing (De Silva et al. 2005), which both make the testing for HWE redundant. Even the Monte Carlo permutation test, which is often a dependable tool when parametric tests fail, is not applicable here as it is not possible to get permuted data sets under the null hypothesis of random mating without knowledge of the dosage. Even when dosage is known, interpreting the results of HWE tests is difficult in polyploids as the presence of double reduction at loci may cause deviation from HWE expectations even under random mating. As a result, the results cannot be used to draw conclusions about either mating system or genotyping errors, which are the 2 main reasons why researchers perform HWE tests (Waples 2015).

Comparing Ploidy Levels

The genetic diversity in polyploids depends on the number of independent evolutionary origins of the polyploids (Soltis and Soltis 1999; Beck et al. 2011) as well as population genetic and demographic processes that have led to their current distribution. In species with multiple ploidy levels, the cytotypes may show differences in mating

system (Meirmans et al. 2006; Neiman et al. 2014), geographical distribution (Menken et al. 1995; Mráz et al. 2007a), environmental niche (Verduijn et al. 2004; Bretagnolle and Thompson 1996), and demographic history (Mráz et al. 2007b), which all can impact the amount of genetic diversity (Kolář et al. 2017). For example, when there are niche differences among cytotypes, the one with the most abundant habitat will be expected to have a larger population size and thus a higher level of genetic diversity. On the other hand, when there is ongoing gene flow among cytotypes—in spite of any reproductive barriers that may exist between them—any differences in genetic diversity may quickly get eroded. Comparing the levels of genetic diversity among the cytotypes may therefore reveal some of these processes.

When comparing the genetic diversity of cytotypes, it is possible to correct for the expected differences between them given Equation 1. One way to do this is to calculate, at a locus for a given cytotype, the level of diversity that can be predicted given the diversity at the same locus in the other cytotype. Say, for example, we want to predict the level of heterozygosity in a *k*-ploid ($H_{S(k)}$), relative to the diversity ($H_{S(2)}$) of a conspecific diploid. When assuming mutation-drift equilibrium and equal effective population sizes in the 2 cytotypes, this relationship is:

$$H_{S(k)} = \frac{k \cdot H_{S(2)}}{2 + (k-2) \cdot H_{S(2)}} \quad (3)$$

Luttikhuizen et al. (2007) used this method to compare the genetic diversity in diploid and tetraploid *Rorippa amphibia*. They visualized their results by plotting the genetic diversity in diploids on the *x* axis and the genetic diversity of the same loci in tetraploids on the *y* axis, and added a line for the predicted relationship between the two (Figure 4). They concluded that the levels of diversity matched very well between the 2 cytotypes. Such a correction may be useful for comparing different cytotypes, but also for comparing regions in a genome with different ploidy levels. In the latter cases, this comparison is actually more straightforward since factors such as population size and demographic history are the same for different regions within the genome.

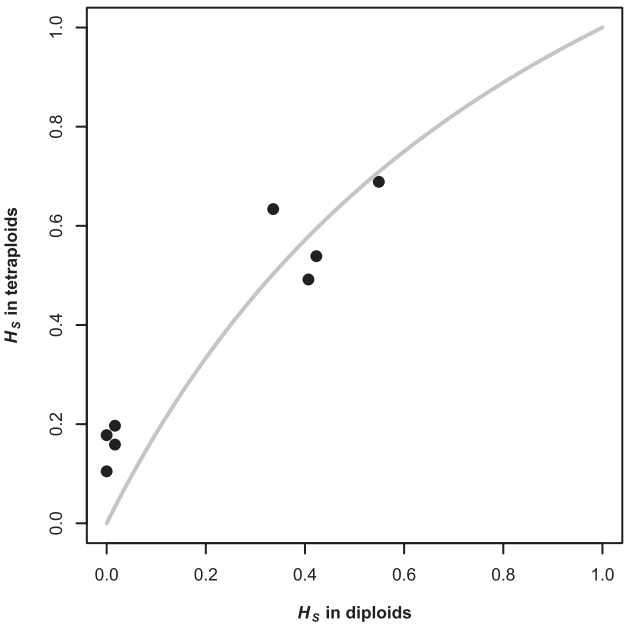

Figure 4. Comparison of expected heterozygosity at 8 microsatellite loci in diploid and tetraploid *Rorippa amphibia*. Though the tetraploids generally have a higher genetic diversity than the diploids, the difference matches the theoretical expectation (gray line) very well. Redrawn from Luttikhuizen et al. (2007).

In practice, it may be difficult to decide whether to interpret the results with or without correction for cytotype. The levels of genetic diversity in different cytotypes depend on many factors, of which the cytotype population size is one and the rate of gene flow among cytotypes another. When this rate is high, the cytotypes will share most of their genetic diversity; applying a correction will make it seem like the higher ploidy level is less diverse than the lower ploidy level.

Mixed ploidy samples exacerbate the problems of both double reduction and missing dosage information for the genetic markers; and this is particularly the case for analyses of heterozygosity and HWE. Double reduction can introduce bias in analyses of mixed ploidy levels since the rate of double reduction at a locus can be different in different cytotypes: in diploids, double reduction is not possible at all, while in polyploids the maximum possible rate increases with ploidy. Therefore, the degree of deviation from HWE due to double reduction is expected to be stronger for higher ploidy levels, making it difficult to study differences in breeding system among cytotypes. This remains problematic since there are many species where a shift in ploidy level coincided with a shift in breeding system, for example, asexual reproduction or a breakdown of self-incompatibility (Dufresne et al. 2014).

Population Differentiation

Measuring Differentiation with F_{ST}

Estimating the degree of differentiation among populations is a key concept in population genetics from which inferences can be made about the demography, history, and connectivity of populations. Traditionally, the degree of divergence is measured using the F_{ST}-statistic, though many alternatives have been developed (e.g., Hedrick 2005; Jost 2008). The value of F_{ST} is determined by various factors, most importantly the migration rate (m), the mutation rate (μ), and the population size (N). For neutral markers under the simplest population differentiation model, the Island Model, there is a rather simple relationship between the equilibrium value of F_{ST} and these 3 parameters (but see Whitlock and McCauley 1999). For diploids, this relationship is:

$$F_{ST} = \frac{1}{1 + 4Nm + 4N\mu} \quad (4)$$

For polyploids—as usual—the situation is slightly different. Because polyploid populations have a higher total number of chromosome copies than similarly sized diploid populations, the impacts of migration and mutation are different. Like we saw above for the level of genetic diversity, there are more mutation events in a polyploid population, which cause a decrease in the equilibrium value of F_{ST}; the size of this decrease depends on the mutation rate. The impact of migration is also higher in polyploids than in diploids, since a polyploid migrant carries more allele copies with it than a diploid migrant. In polyploids, migration will therefore lead to a better evening out of the allele frequencies than in diploids. Furthermore, the force of genetic drift—which is the ultimate force leading to population differentiation—is weaker in polyploids. These effects on F_{ST} scale with the ploidy level, which means that the relationship between F_{ST}, migration, mutation, effective population size, and the ploidy level can be generalized as follows:

$$F_{ST} = \frac{1}{1 + 2kNm + 2kN\mu} \quad (5)$$

This means that under the exact same population model, the expected value of F_{ST} is lower in polyploids than in diploids (Figure 5a). This complicates comparisons of the levels of F_{ST} between species with different ploidy levels, but also between cytotypes within a single species. In fact, this also applies to differences in ploidy within a genome: autosomes, X, and Y chromosomes all have different expectations for F_{ST}. Further examples are genomes with ancient genome duplication, such as in the salmonid fishes, where there is residual tetrasomy in parts of the genome, but disomy in the rest of the genome (Allendorf et al. 2015; Waples et al. 2017).

In the last few years, there has been a discussion about the usefulness of F_{ST} for quantifying genetic differentiation (Hedrick 2005; Meirmans 2006; Whitlock 2011; Wang 2015). The reason is that the maximum value that F_{ST} can attain depends on the level of diversity within populations, H_S, with $F_{ST(max)} = 1 - H_S$ (Meirmans and Hedrick 2011). As a result, when populations have a high value of H_S—which is often the case for highly variable markers such as microsatellites—the value of F_{ST} will be low, even when there is no gene flow among populations. This effect will be stronger in polyploids because H_S is generally higher in polyploids. However, finding a low value of F_{ST} with highly valuable markers does not necessarily mean that the degree of population differentiation has been underestimated. This is because the value of F_{ST} does not only depend on the mutation rate but also on other parameters, such as the migration rate and the effective population size (Meirmans and Hedrick 2011; Wang 2015).

To overcome these and other shortcomings of F_{ST}, several alternative statistics have been suggested (Ronfort et al. 1998; Hedrick 2005; Jost 2008), whose suitability for polyploid data we discuss below.

Other Differentiation Statistics

The F_{ST} alternative that is most relevant for polyploids is the ρ statistic. This statistic was developed by Ronfort et al. (1998) especially for autotetraploids, but it can be calculated for any ploidy level (Supplementary Appendix 1). For diploids, ρ is equivalent to the average relatedness of individuals within populations compared to the whole (manuals of Fstat & SPAGeDi; Goudet 1995; Hardy and Vekemans 2002), when calculated using the approach of Queller and Goodnight (1989). The great advantage of ρ is that it is designed to be comparable between ploidy levels (Figure 5d) and is also independent of the rate of double reduction. When expressed in terms of heterozygosities ρ takes the following form:

$$\rho = \frac{H_T - H_S}{H_T - H_O \cdot (k-1)/k} \quad (6)$$

The equilibrium value of ρ for any ploidy level is given by (Supplementary Appendix 2):

$$\rho = \frac{1}{1 + 2Nm + 2N\mu} \quad (7)$$

Comparison with Equation 5 shows that for a haploid organism, ρ is equivalent to the value of F_{ST}. For nonhaploids, ρ is higher than F_{ST}—except when both have a value of either zero or one.

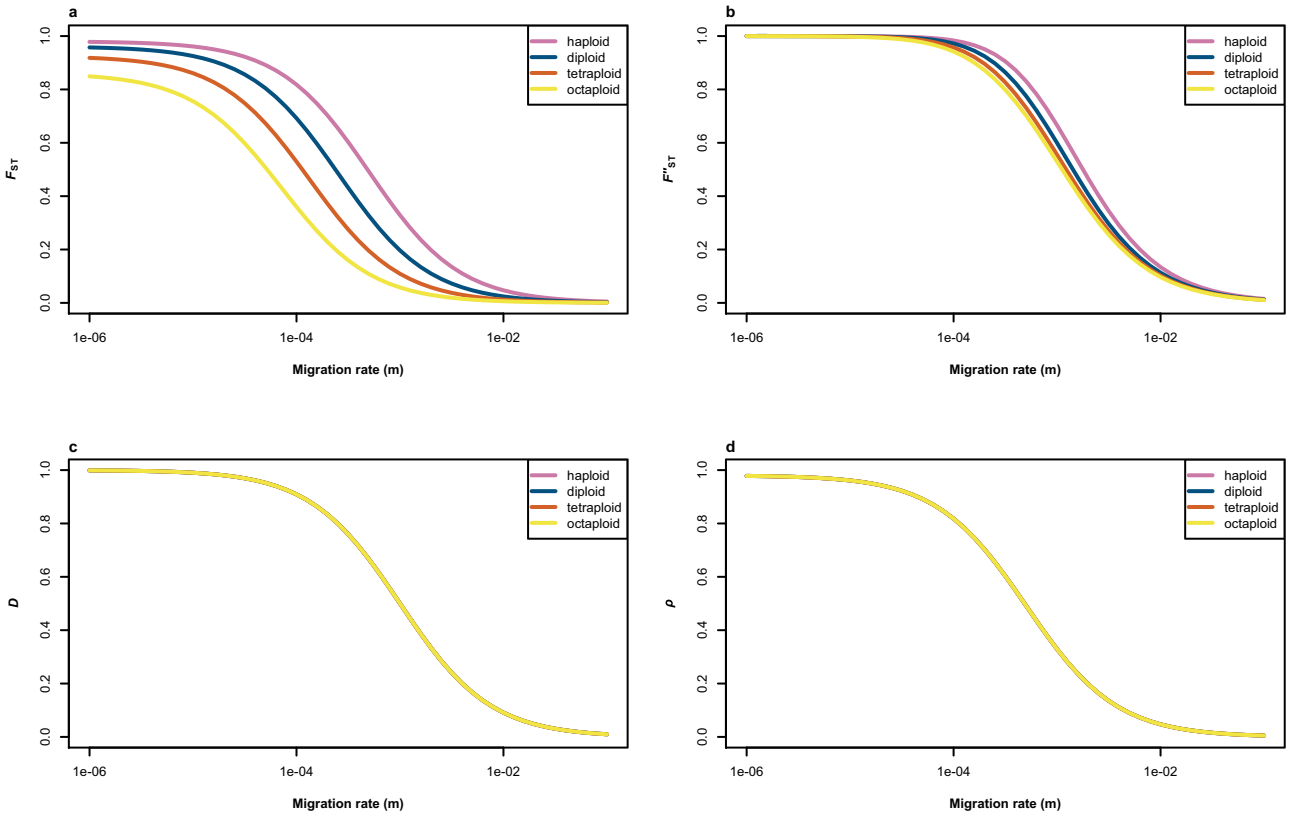

Figure 5. Theoretical expectations for 4 different estimators of genetic differentiation as a function of the migration rate for different ploidy levels. Expectations were calculated following Equations 5, 7, 9, and 11, under an island model of migration with 100 randomly mating populations of size N = 1000. Mutation followed an infinite alleles model with μ = 0.001. See the Supplementary Data for the used R-script. Note that in **c** and **d**, all lines are overlapping indicating that these summary statistics have the same expected value for every ploidy level.

Two alternative statistics have been suggested to overcome the dependence of F_{ST} on H_S, namely F'_{ST} (Hedrick 2005) and D (Jost 2008). In this review, these 2 statistics are represented by their unbiased estimators F''_{ST} (Meirmans and Hedrick 2011) and D_{est} (Jost 2008), but in the literature the names of the statistics and their estimators are used interchangeably. F''_{ST} is defined as F_{ST} scaled by the maximum value it can attain given the observed value of H_S. This should make sure that the statistic always has a value of 1 in cases where there is no sharing of alleles among populations. Furthermore, Hedrick (2005) stated that this should make F''_{ST} better suited than F_{ST} for comparisons among species with different population sizes.

F''_{ST} can be defined as a function of the number of populations, the within population diversity (H_S) and the total diversity (H_T) (Meirmans and Hedrick 2011):

$$F''_{ST} = \frac{r \cdot (H_T - H_S)}{(r \cdot H_T - H_S) \cdot (1 - H_S)} \tag{8}$$

Using recurrence equations and the assumptions that N is large and μ and m are small, equations for the equilibrium values for H_S and H_T for any ploidy level k can be obtained for the finite island model (Supplementary Appendix 2). These equations can then be used to obtain the equilibrium value of F''_{ST} (Supplementary Appendix 2):

$$F''_{ST} = \frac{1}{1 + \frac{m}{r\mu}} \cdot \left(1 + \frac{m}{r\mu \cdot (1 + 2kN\mu + 2kNm)}\right) \tag{9}$$

From Figure 5b, we see that F''_{ST} is not fully independent of the ploidy level, especially at intermediate migration rates the expected value decreases with the ploidy level. The dependence on the ploidy level also varies with the mutation rate: when the mutation rate is high the expected values are very similar for all ploidy levels, while for low mutation rates the expected values can differ between ploidy levels.

The D_{est}-statistic (Jost 2008) was developed based on different concepts of genetic diversity and population differentiation than are used for the classic F_{ST}. Rather than using the expected heterozygosity within and among populations (or a related concept Weir and Cockerham 1984), Jost uses the effective number of alleles to quantify diversity. Nevertheless, D_{est} can still be expressed in terms of H_S and H_T:

$$D_{est} = \left(\frac{r}{r-1}\right)\left(\frac{H_T - H_S}{1 - H_S}\right) \tag{10}$$

Substituting the equilibrium values of H_S and H_T for any ploidy k into the equation above gives following the equilibrium value of D_{est} (Supplementary Appendix 2):

$$D_{est} = \frac{1}{1 + m/(r\mu)} \tag{11}$$

This equation is the same as what was obtained by Jost (2008) for diploids, showing that the equilibrium value of D_{est} is not only independent of the population size (as was noted before), but also of the ploidy level (Figure 5c). Despite its ploidy-independence, the use of D_{est} is not recommended since the time needed by D_{est} to reach its equilibrium value can be very long (Ryman and Leimar 2009; Meirmans and Hedrick 2011).

Of the F_{ST} alternatives that we discussed here, ρ should be the statistic of choice for studies of population structure in polyploids. Not only is it independent of ploidy level and double reduction, its close relatedness with F_{ST} means that much of the vast literature of F_{ST} is also applicable to ρ. Furthermore, ρ has the same rapid approach to equilibrium as F_{ST} (Liu S and Meirmans PG, unpublished results). For diploids, ρ is also a suitable statistic as it is independent of the level of inbreeding and therefore permits easier comparison among species that differ in their mating system (when the impact of the mating system on differentiation is not of interest). A downside of ρ is that, like F_{ST}, its value may be underestimated with highly polymorphic markers. D_{est} may be useful in such cases, when studying polyploids, since its value is also ploidy independent.

Estimation and Bias

There are 2 main methods for estimating F_{ST} and associated statistics: heterozygosity-based estimates (Nei 1987) and ANOVA-based estimates (Weir and Cockerham 1984). Both methods can be used for polyploids, but again some modifications have to be applied to allow for the multiple homologous chromosomes within individuals. Heterozygosity-based estimates can simply be obtained using Equations 6, 8, and 10 with the estimates of H_O, H_S, and H_T. ANOVA-based estimates of F_{ST} and ρ can be obtained using the method from Ronfort et al. (1998), or by adapting the analysis of molecular variance framework (AMOVA; Excoffier et al. 1992; Michalakis and Excoffier 1996). The AMOVA can also be used to get an estimate of F''_{ST}, by adapting the method of Meirmans (2006) for polyploidy. Multiple software packages are available to perform these calculations for a large range of different ploidy levels (see Table 1 in Dufresne et al. 2014).

As we saw above, missing dosage information leads to an overestimation of the genetic diversity within population, which consequently leads to an underestimation of the degree of population differentiation. Though this bias is mostly very small, there are some important differences between statistics and between the heterozygosity-based and ANOVA-based methods of estimation. Figure 6 shows the extent of this bias for simulated data of 2 randomly mating tetraploid populations connected by varying rates of migration, and with 3 different mutation rates. For F_{ST} (Figure 6; top row), the bias is small, and this bias can completely be corrected for under the assumption of HWE when the method of De Silva et al. (2005) is used. For ρ (Figure 6; second row), the heterozygosity-based estimates (using Equation 6) are nearly unbiased, but correction generally leads to a large overestimation and even increases the value above the theoretical maximum of 1 (these values are not shown in Figure 7 to keep all y-axes on the same scale). These overestimations are caused by small errors in the corrected value of H_O. The ANOVA-based estimate of ρ has a larger bias, especially for the intermediate mutation rate, but here the correction does work correctly. F''_{ST} (Figure 6; third row) has a small bias with the heterozygosity-based estimator and a very large bias with the ANOVA-based estimator. In both cases, the bias can be corrected for under the assumption of HWE using the method from De Silva et al. (2005). D_{est} can only be calculated using a heterozygosity-based method, but then is nearly unbiased. Note that these results do not take selfing or double reduction into account. More extensive simulations are needed to assess the degree of bias under such conditions.

For estimating the differentiation statistics, it is important to know whether the species has disomic or tetrasomic inheritance; unknowingly analyzing disomic data as if it were tetrasomic can result in a large bias (Meirmans and van Tienderen 2013). The degree of this bias differs between statistics: it is largest for F_{ST} and F'_{ST}, relatively small for D_{est} and completely absent for ρ (Meirmans and van Tienderen 2013). For segmental allopolyploids, the degree

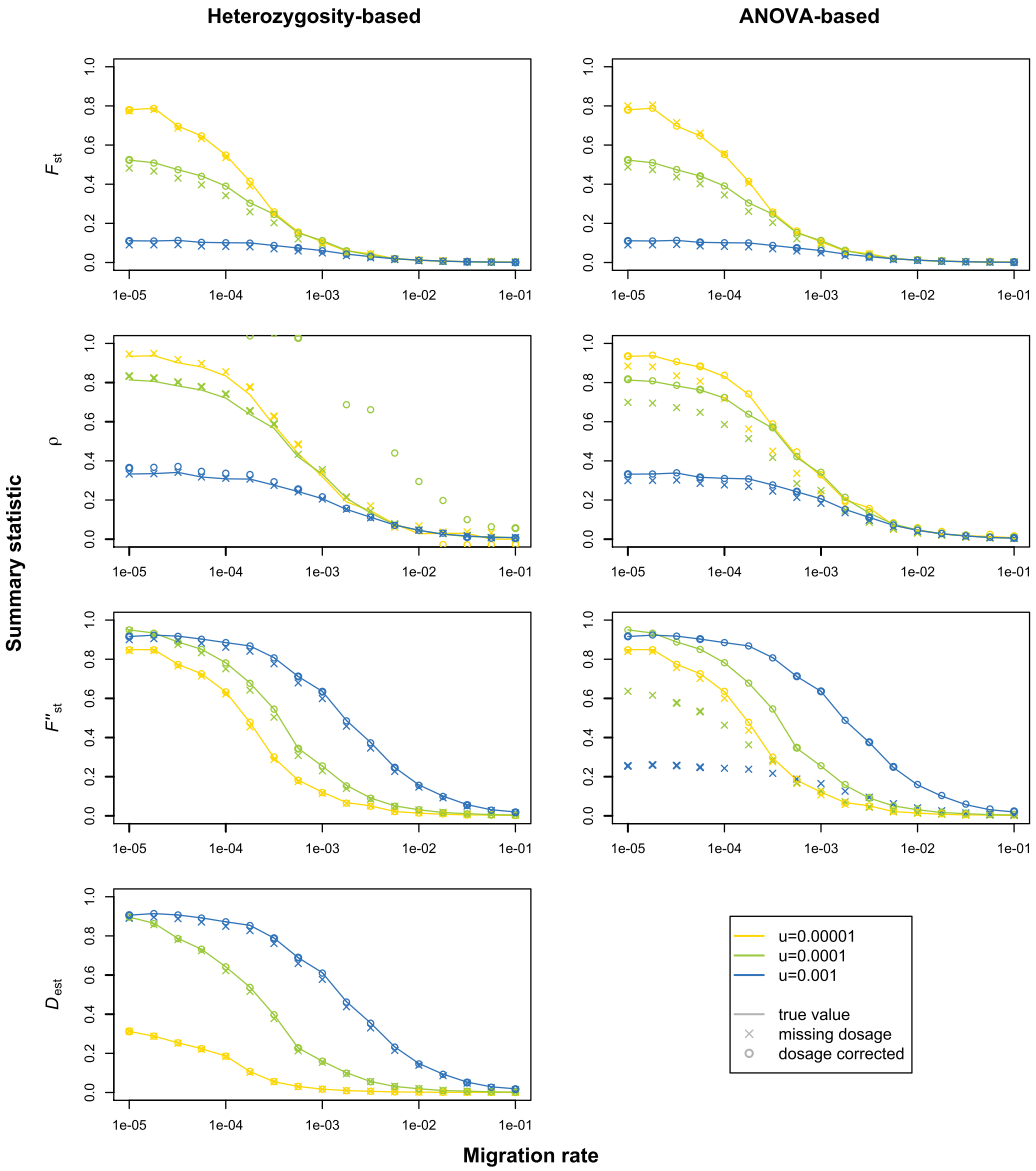

Figure 6. Bias from missing dosage in the estimation of F_{ST}, ρ, F'_{ST}, and D_{est} for tetraploid data with different mutation rates and different migration rates. The same simulated data were used as for Figure 2. All statistics, except D_{est}, were estimated both using the heterozygosity-based framework of Nei (1987) and the ANOVA framework of Weir and Cockerham (1984). For the heterozygosity-based estimates, missing dosage information was corrected for using the method of DeSilva et al. (2005); in addition, for the estimation of ρ the method of Hardy (2015) was used for H_O. For the ANOVA-based estimates, missing dosage information was corrected for by replacing missing dosage with randomly drawn alleles, based on population allele frequencies as calculated using the same method of DeSilva et al. (2005). See the Supplementary Data for the R-script used for the simulations.

of the bias depends on the frequency of tetrasomic inheritance resulting in the exchange of alleles among homoeologous chromosomes. Simulations have shown that in general low rates of tetrasomy are sufficient to even out the allele frequencies between the subgenomes: about one exchange event per generation is enough to remove most of the bias (Meirmans and van Tienderen 2013). This is analogous to the rule-of-thumb of population differentiation, where it is said that one migrant per generation suffices to prevent divergence of allele frequencies among populations (Whitlock and McCauley 1999).

Mixed Ploidy

There is no single summary statistic that is ideal for all analyses of population differentiation in data sets with mixed ploidy levels, even when the allele dosage is known. This is because there are different ways in which to compare data from multiple ploidy levels, and the different statistics differ in how suitable they are for different types of comparisons. One way in which data from multiple cytotypes can be compared is to look at the population differentiation in each cytotype separately, and then compare their strength. A second way to compare cytotypes is to calculate the degree of differentiation between them, which allows making inferences about the degree of gene flow between cytotypes.

For the first type of comparison, looking at the population differentiation in each cytotype separately, ρ is the most suitable statistic since its value is independent of the ploidy level, when the population sizes, mutation rates, and migration rates are the same in the compared cytotypes. Therefore, ρ allows analysis of whether any of those factors are different in the different

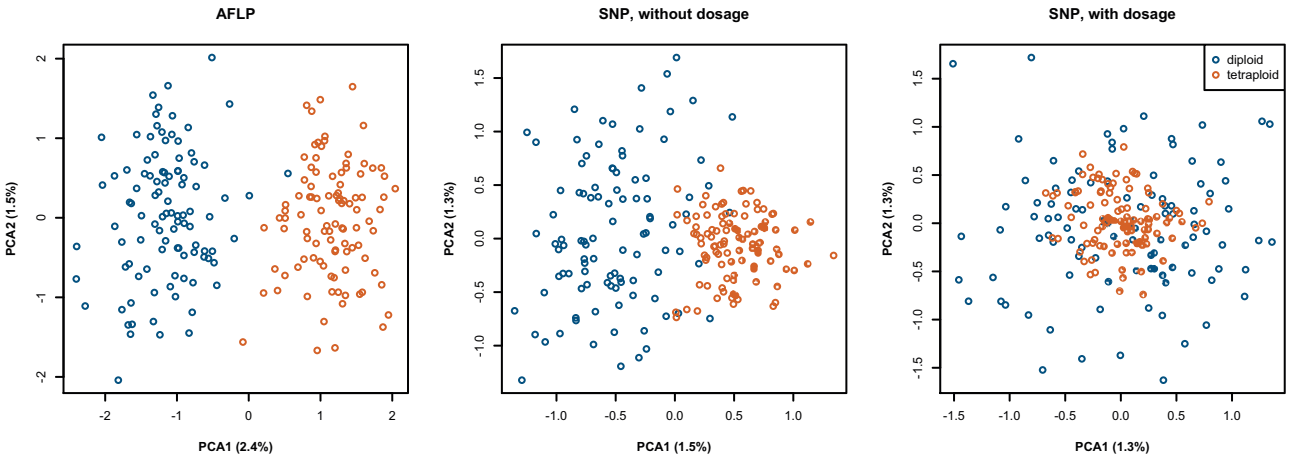

Figure 7. Spurious clustering arising from dominance and missing dosage information in a principal components analysis of diploid and tetraploid individuals drawn from a single population. A set of allele frequencies at 100 SNP loci was generated by drawing random numbers from a uniform distribution. Based on these allele frequencies, 100 diploid and 100 tetraploid genotypes were then generated and analyzed with and without dosage information and after conversion to dominant AFLP data. See the Supplementary Data for the used R-script.

cytotypes. F_{ST} and F''_{ST} are not suitable for this type of comparison as their values are not ploidy-independent. In species where there is ploidy variation among loci within the genome (Allendorf et al. 2015), a genome scan for selected loci (Waples et al. 2017) should be based on ρ instead of F_{ST}, since otherwise the parts of the genome with residual tetrasomy may be tagged as outliers with lower F_{ST} values.

A closer look at how ρ is calculated shows that this statistic is less suited for the second type of comparison, calculating the degree of differentiation between cytotypes. To give similar values for different ploidy levels for the exact same model of population structure, ρ weighs the differentiation by ploidy. This is necessary, because otherwise one gets (for the same population size and migration rate) a different value for diploids than for tetraploids. However, the way that ρ is calculated has the implicit side effect that for a given set of population allele frequencies, ρ will give different values for diploids and tetraploids. This is because in diploids these allele frequencies require a different population size, mutation rate, and/or migration rate than for tetraploids. For calculating the degree of differentiation between cytotypes, F_{ST} is the better choice as it always gives a similar value for a given set of allele frequencies, regardless of the ploidy level.

In conclusion, ρ is the best choice for comparing the strength of the population structure among studies, among species, or among cytotypes, provided that the statistic is calculated for each cytotype separately (so when there are multiple data sets; one for each cytotype). F_{ST} is the best choice for calculating the degree of divergence between cytotypes (so when there is a single data sets that combines multiple cytotypes). When in doubt about which statistic is best for certain usage scenarios, it is advisable to use simulations such as presented above.

Detecting Population Structure

Clustering Approaches

An important part of population genetic studies is the detection of population structure using a clustering approach. Such a clustering analysis can reveal the impact of past events on the current genetic structure of populations (Cornille et al. 2016), but also the impact of anthropogenic disturbance (van Hengstum et al. 2012), habitat fragmentation (van der Meer and Jacquemyn 2015), and hybridization (Clark and Jasieniuk 2012). In polyploids, clustering analyses are used to study the relationships between different ploidy levels (e.g., Kolář et al. 2016) and to test hypotheses concerning multiple origins of polyploids. Many different statistical approaches are available for detecting population structure, including classical multivariate analyses (principal components analysis, redundancy analysis, *K*-means), multivariate analyses adapted to genetic data (DAPC Jombart et al. 2010; AMOVA-based *K*-means Meirmans 2012a) and model based inferences (Structure Pritchard et al. 2000; InStruct Gao et al. 2007; Admixture Alexander et al. 2009). Several of these can be applied to polyploids whereas others are restricted to diploids.

Multivariate Analyses

The use of classical multivariate analyses such as PCA or *K*-means clustering is widespread and goes back to the first large-scale analyses of population structure (Menozzi et al. 1978). Such analyses are often performed on a set of allele frequencies, either population allele frequencies, for analyses at the population level, or individual allele frequencies, for analyses at the individual level. The latter can be calculated by taking the number of occurrences of an allele in an individual and dividing that by the ploidy level. Therefore, these frequencies can only take a limited set of values: for diploids, the possible values are 0, 0.5, and 1; for tetraploids, these are 0, 0.25, 0.5, 0.75, and 1. If those individual and population allele frequencies can be calculated correctly for polyploids, it is straightforward to use them for multivariate analyses in the exact same way as is done for diploids. However, there is a problem when the allele frequencies cannot be calculated correctly for polyploids, for example, because of missing dosage information.

Some multivariate analyses have been specially adapted to genetic data. Discriminant analysis of principal components (DAPC; Jombart et al. 2010) uses a combination of PCA, *K*-means clustering and discriminant analysis to detect and visualize population structure. Since it is based on a matrix of within-individual allele frequencies, it can be easily applied to polyploids. AMOVA-based *K*-means clustering (Meirmans 2012a) uses the AMOVA framework to calculate among-cluster sum of squares. Therefore, this analysis results in the clustering with the maximum possible F_{ST}-value among clusters. Since the AMOVA can be extended to polyploid data, the same holds for AMOVA-based clustering.

In multivariate analyses of mixed ploidy levels, a bias can easily arise as a result of missing dosage information. Unfortunately, the presence of this bias is often not realized and multivariate analyses remain widely used for assessing the genetic relationships between cytotypes. To illustrate this, we performed a PCA on a simulated data set of SNP genotypes of diploid and tetraploid individuals drawn from a single gene pool, so without any differentiation between the cytotypes. These individuals were created by first drawing a set of 100 biallelic SNP loci with random allele frequencies. Based on these frequencies, 100 diploid and 100 tetraploid individuals were created by random sampling of alleles and combining these into genotypes.

When the SNP data was converted to dominant AFLP marker data and a PCA was performed on the resulting data set, a clear separation was visible between the 2 cytotypes (Figure 7a), even though they were drawn from the same gene pool. When the SNP data were kept co-dominant, but without dosage information for the genotypes, the spurious separation was less strong but still clearly visible (Figure 7b). For the correct data, codominant SNPs with dosage information, there was no spurious clustering (Figure 7c). However, it is worth pointing out that there was more spread in the diploid data than in the tetraploid data; this is because even though there are more genotype classes in tetraploids than in diploids, the tetraploids individuals are on average more similar to each other in within-individual allele frequencies since the most extreme genotypes (the homozygotes) are much more frequent in diploids than in tetraploids.

Clearly, one should be cautious about the use of multivariate methods, especially when used on a data set with multiple ploidy levels. Other multivariate methods use a matrix of pairwise differences—either between individuals or between populations—as their primary input; the most important example of this is a principal coordinates analysis (PCoA). Here, one also has to be cautious as the suitability of the analysis and interpretation of the results depend on the choice of distance metric and not all distance metrics are equally suitable for polyploid data (see "Resemblance between individuals" section).

Structure

The most widely used clustering analysis is Structure (Pritchard et al. 2000), which applies population assignments in a Bayesian context. The great appeal of Structure is that it performs "soft" instead of "hard" clustering; individuals can be partly assigned to several clusters. This makes it ideal for studying the occurrence of admixture between populations or, in our case, cytotypes. Structure is well-suited for analyzing polyploids (e.g., Anderson et al. 2017), as it can take data from any ploidy level. A drawback is that the description in the manual of how to input polyploid data is rather cryptic and it is easy to end up with incorrectly read data without noticing. Even when the dosage of alleles is unknown, the input file should contain complete genotypes instead of marker phenotypes; for this, the unknown allele copies can be filled in ad libitum. When the RECESSIVEALLELES flag in the file with parameters is set to 1, Structure will discard all dosage information in the input file and then perform the analysis correcting for the unknown dosage. Unfortunately, there is no way to enter data from species where there is ploidy variation within the genome; for such species, we suggest to use only the loci that show disomic inheritance, or those that show polysomic inheritance. Simulation studies have shown that Structure generally performs well in diploids (Evanno et al. 2005) though there may be some biases when there is isolation by distance (Meirmans 2012b) or unbalanced sampling (Puechmaille 2016). No simulation studies have been performed as yet to test whether there is any bias in Structure when used with polyploids and with mixed ploidy data sets.

Several other clustering algorithms have been developed that are based on similar principles as Structure. One such method is implemented in the program InStruct (Gao et al. 2007), which performs a simultaneous inference of population structure and the degree of inbreeding. InStruct can take either diploid or tetraploid data, but has no support for higher ploidy levels or mixed diploid/tetraploid data sets. Other methods that are based on similar principles as Structure are Admixture (Alexander et al. 2009), FastStructure (Raj et al. 2014), and FineStructure (Lawson et al. 2012). These methods are all geared toward use with large biallelic SNP data sets and all perform much faster than Structure. Unfortunately, these 3 programs only work with diploid data and there is no information on how they perform when polyploid data is shoehorned to fit this requirement. Therefore, we recommend the use of Structure for polyploid data sets, even with a large number of loci (though computationally demanding) and with the precautions expressed above.

Resemblance between Individuals

Matrices of resemblance between individuals have various uses in analyses of genetic data, for example, for analyzing spatial autocorrelation within populations, for constructing dendrograms that indicate the relationship between individuals, or for use in matrix-based analyses such as principal coordinates analysis. In all these cases, the results of the analysis can depend heavily on the choice of resemblance metric and therefore a large number of metrics has been developed. Two main classes of metrics are widely used in genetics: distance metrics and relatedness coefficients. Distance metrics simply try to estimate how different individuals are from each other; relatedness coefficients try to estimate an actual biological property—the degree of relatedness between individuals (e.g., the relatedness between a parent and its offspring is 0.5).

We used some simple simulations to test how 8 resemblance metrics perform when calculated for diploid, tetraploid, and mixed ploidy data. For this, we selected 4 distance metrics—Euclidean, Smouse and Peakall (1999), Chord (Cavalli-Sforza and Edwards 1967), Bruvo et al. (2004)—and 4 relatedness metrics—Huang et al. (2014b), Ritland (1996), Loiselle (1995), Weir (1996). We selected only metrics that can be calculated for polyploids based on the full genotypes, leaving out any metrics based on allele presence–absence. Genetic data were simulated by drawing diploid and tetraploid microsatellite genotypes from a single gene pool, both with and without information on the dosage of alleles. The software used for calculating the resemblance matrices was GenoDive (Meirmans and van Tienderen 2004; Euclidean, Smouse and Peakall, Chord), Poppr (Kamvar et al. 2014; Bruvo, using the default settings), and PolyRelatedness (Huang et al. 2014a; all other metrics). In Figure 8, the presence of the 8 metrics can be gauged from the drawn lines. The 3 vertical dotted lines indicate the mean values for diploid–diploid, diploid–tetraploid, and tetraploid–tetraploid comparisons, when dosage is known. When these 3 lines are close together, the metric yields identical means for diploids, tetraploids, and mixed ploidy. The 3 horizontal dotted lines indicate the same, but then without any dosage information. Below, the results of these simulations are discussed separately for the distance/dissimilarity metrics and for the relatedness coefficients.

Distance/Dissimilarity Metrics

When comparing the 4 distance metrics (Figure 8 top row), it is clear that all of them have a bias, even when the dosage is known. Notably, this is also the case for the Bruvo distance (Bruvo et al.

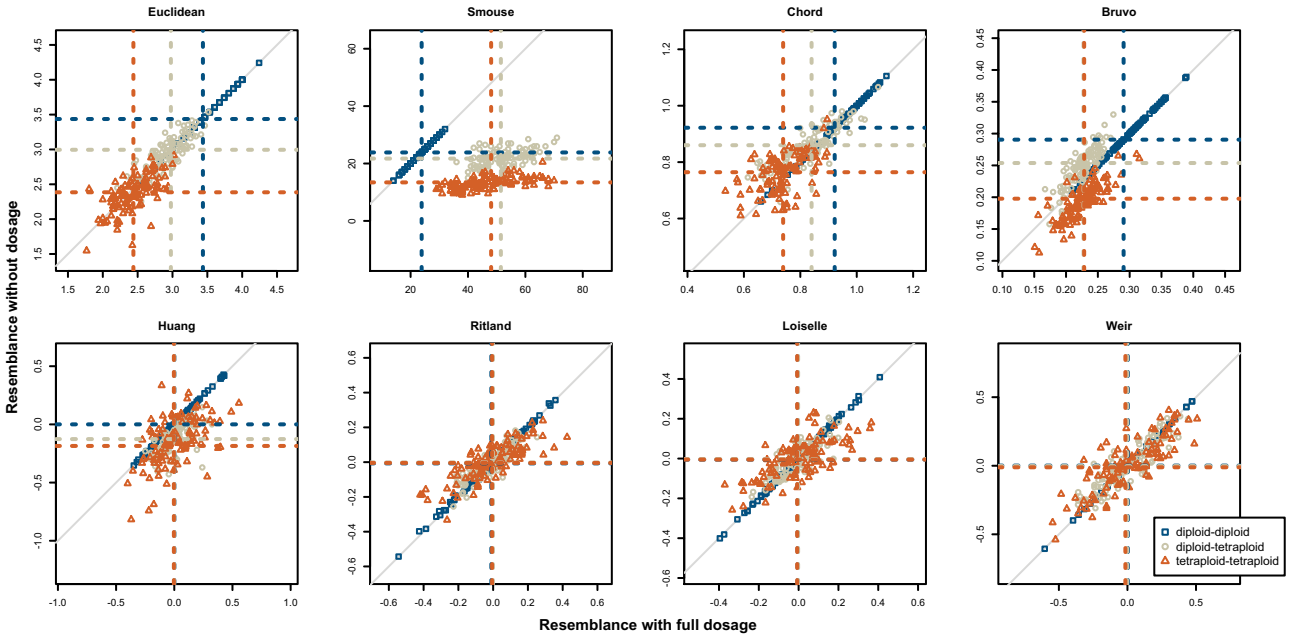

Figure 8. The effect of ploidy and missing dosage information on 8 metrics of pairwise resemblance between individuals. A set of allele frequencies for 20 loci was simulated for a single diploid population of size 10000, with a K-allele mutation model with 100 alleles and rate 0.0001. Based on these frequencies, 100 diploid and 100 tetraploid genotypes were created, between which the resemblance metrics were calculated, for the tetraploids both with and without dosage information.

2004), which is surprising since this metric was especially developed for analyzing microsatellite data from polyploids. Diploid–tetraploid and tetraploid–tetraploid comparisons have the same mean Bruvo distance, but the mean of diploid–diploid comparisons is notably higher. In short, none of the distance metrics should be used to compare ploidy levels, or with mixed ploidy data; use within a single ploidy level, whether diploid or polyploid, should not be problematic. However, there is a considerable amount of deviation between the results with and without dosage information, which can be seen from the spread around the $x=y$ line. This means that one should be careful when interpreting the results when dosage is unknown, even when these metrics are used within a single ploidy level.

Relatedness Coefficients

In contrast with the distance metrics, the relatedness coefficients (Figure 8 bottom row) do not show a bias with respect to ploidy: regardless of the ploidy comparison, they always show an average of about zero, which is the expected level given the random sampling used to construct these genotypes. The relatedness coefficients also do not show any bias related to ploidy when the dosage information is missing, with the exception of the Methods-of-Moment estimator of Huang et al. (2014b), even though that was especially developed for polyploids. However, all kinship coefficients are affected by the missing dosage in the sense that there is still considerable spread around the $x=y$ line. This residual variance is smallest for the estimator from Ritland (1996), meaning that this estimator is least affected by missing dosage information. In practice, calculation of the relatedness coefficients relies on the population allele frequencies, so the degree of dependence on the dosage information may differ between implementations, depending on whether the allele frequencies are corrected for the missing dosage or not. Furthermore, the simulations used here are very simple, and cannot be used to tell whether the resulting relatedness coefficients are actually correct; all that we can conclude is that they are unbiased with respect to ploidy, under the simulated conditions.

Other Inferences

Detection of Loci under Selection

The availability of large genetic data sets has made it possible to screen population samples for loci that are putatively under selection. There are 2 main approaches to this: looking for outlier loci that show higher or lower F_{ST}-values than expected (e.g., Beaumont and Nichols 1996), or looking for associations between genetic variation and environmental variables, while correcting for neutral spatial patterns (e.g., Frichot et al. 2015). A lot of different methods have been developed for these types of analyses (see Vatsiou et al. 2015 for an overview), each with its own assumptions and drawbacks. One drawback that they all have in common is that they only work with diploid data; as far as we are aware, none of these methods allows entry of polyploid data. This is a major lacuna in the toolbox for polyploid data analysis that prevents polyploidy researchers from addressing pressing questions around the impact of polyploidy on the efficacy of selection.

Assignment Tests

In an assignment test, the putative population of origin is determined for individuals of unknown provenance. In the original implementation of Paetkau et al. (1995), the genotype of the individual is compared to the allele frequencies at a number of populations. For every population, the likelihood is calculated that the individual comes from that population, assuming HWE within populations. The population with the highest likelihood score is then selected as the putative source population. This simple likelihood analysis has subsequently been appended, among others with a Monte Carlo permutation approach (Rannala and Mountain 1997), and by being placed in a Bayesian framework (Cornuet et al. 1999). The use of assignment tests mainly lies in forensic types of analysis, that is, the identification of material of unknown provenance. Though they are also frequently used for measuring migration rates, they are less suited for this (Christie et al. 2017).

When the dosage information is available and there is no double reduction, the calculation of the likelihood for an assignment test can relatively simply be adjusted for polyploidy. All that is needed is the expected frequency of the observed (single-locus) genotype given HWE. These expected frequencies are readily calculated for polyploids, though they get a bit cumbersome for higher ploidy levels. Nevertheless, WhichRun (Banks and Eichert 2000), the most commonly used and most versatile software for assignment tests, only works with diploids. For polyploids, the program GenoDive (Meirmans and van Tienderen 2004) can perform assignment tests, but only when dosage is known and there is no double reduction. However, the software AutoPoly (Field et al. 2017) can take missing dosage and double reduction into account by combining the likelihoods for all possible genotypes for a phenotype. A test with simulated data has shown that this method performs well even when there is little differentiation among the possible source populations. Another software that can perform population assignments is GSI_sim, which was originally developed for genetic stock identification in fisheries but can be put to more general use (Anderson et al. 2008). GSI_sim has the unique ability that the ploidy level can be specified independently for each locus, and is therefore ideally suited for analyzing species with ploidy variation within the genome (see Anderson et al. 2017 for an application).

Parentage Analysis

In a parentage analysis, the putative parents of offspring (e.g., in the form of seeds, juveniles, larvae, or recruits) is determined based on their respective genotypes. Most such analyses rely on the principle of exclusion: adults that, at a locus, do not contain any of the alleles carried by the offspring are excluded as possible parents. When combined over multiple loci, this approach can be very powerful, with a low rate of false positives (Christie et al. 2017). The exclusion principle can also be applied to polyploids, but with the consideration that each parent must contribute multiple alleles. Since this is more restrictive than with diploids, it follows logically that for polyploids less loci will be needed for equal discriminatory power. However, since polyploidy specific complications such as missing dosage and double reduction will also affect the performance of such analyses, this remains to be formally tested. There are 2 programs that have been especially developed to perform parentage analyses in polyploids and both allow for missing dosage information: Orchard (Spielmann et al. 2015), which has a graphical user interface but takes only tetraploid data, and Polypatex (Zwart et al. 2016), which is implemented as an R-package and can accommodate data from any ploidy level.

Analysis of Mating Systems

Many polyploids show a different mating system than their diploid relatives; this can be an increased level of self fertilization (Mable 2004) or even parthenogenetic reproduction (Menken et al. 1995; Neiman et al. 2011). In fact, most parthenogenetically reproducing plants and animals are polyploid (Dufresne et al. 2014). Genetic data can be very insightful for analysing mating systems as it can be used for estimating selfing rates or for the detection of clonal lineages. In diploids, the most straightforward way to test for nonrandom mating is to perform a test for HWE conformation. As explained above, missing allele dosage and double reduction make this difficult in polyploids. For tetraploids, Ritland developed the software MLTET (updated from Ritland 1990) to estimate inbreeding coefficients and selfing rates. The program takes data from both population samples and progeny arrays, but does require complete genotypes and the absence of double reduction. Recently, Hardy (2015) developed a method—implemented in the software SPAGeDi (Hardy and Vekemans 2002)—to estimate the rate of self-fertilization that combines information from multiple loci. This new method was found to be robust both when allelic dosage was missing and in the presence of double reduction. Detection of clonal lineages in population samples of any ploidy level can be done by the software GenoDive (Meirmans and van Tienderen 2004).

Conclusions

It should be clear from this review that there are still many biases and pitfalls when it comes to the analysis of polyploid genetic data. On the other hand, the statistical toolbox that is available to polyploidy researchers is flexible and still expanding. Modern sequencing techniques can work around some of the limitations that are outlined above; for example, when the sequencing depth is high enough the problem of missing dosage can be solved. The new techniques also bring new challenges; many genomic tools for assembling, scoring, and quality-checking of sequencing data are developed for diploids and adapting them for use with polyploids may not be straightforward. It is also important to realize that in cases where polyploidy leads to a bias in the analysis, adding more loci does not help to resolve the bias, but may even exacerbate it; a spurious pattern (e.g., in differentiation between diploids and polyploids; Figure 7) may be hardly visible when only a handful of loci is used, but become hugely problematic when tens of thousands of loci are available (Meirmans 2015).

In cases where a bias is suspected, simulation of genetic data is an indispensible part of the analysis of genetic data. Such simulations do not necessarily need to be very complex to be insightful: the simulations in Figure 7 were simply done by drawing random numbers and using these as allele frequencies. In other cases, a simple Island model of migration suffices to obtain a set of genetic marker data that can be used to test the performance of different summary statistics (Figure 6). We hope that the set of R-scripts that we used to produce some of the figures in this paper may provide a starting point for simulating additional scenarios. Unfortunately, specialized simulation software such as exists for diploids in great number (Hoban et al. 2012) does not yet exist for polyploids. Such a polyploid simulation program would ideally incorporate the effects of double reduction and variable rates of polysomic inheritance. Simulations also provide the backbone for Approximate Bayesian Computation (Beaumont et al. 2002), which offers a very flexible framework for testing evolutionary scenarios using population genetic data and can be adapted for use with polyploid species (St Onge et al. 2011). Because of this flexibility, ABC probably is the way forward in the analysis of polyploid genetic data sets of increasingly large sizes.

Supplementary Material

Supplementary data are available at *Journal of Heredity* online.

Acknowledgments

We would like to thank Marc Stift, Filip Kolář, and the participants of the 2017 polyploidy workshop in Prague for stimulating discussions.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19:1655–1664.

Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE. 2015. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J Hered*. 106:217–227.

Anderson EC, Ng TC, Crandall ED, Garza JC. 2017. Genetic and individual assignment of tetraploid green sturgeon with SNP assay data. *Conserv Genet*. 18:1119–1130.

Anderson EC, Waples RS, Kalinowski ST. 2008. An improved method for predicting the accuracy of genetic stock identification. *Can J Fish Aquat Sci*. 65:1475–1486.

Banks MA, Eichert W. 2000. WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *J Hered*. 91:87–89.

Beaumont MA, Nichols R. 1996. Evaluating loci for use in the genetic analysis of population structure. *P Roy Soc B Biol Sci*. 263:1619–1626.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics*. 162:2025–2035.

Beck JB, Windham MD, Pryer KM. 2011. Do asexual polyploid lineages lead short evolutionary lives? A case study from the fern genus *Astrolepis*. *Evolution*. 65:3217–3229.

Bever JD, Felber F. 1992. The theoretical population genetics of autopolyploidy. *Oxf Surv Evol Biol*. 8:185–217.

Bomblies K, Madlung A. 2014. Polyploidy in the Arabidopsis genus. *Chromosome Res*. 22:117–134.

Bretagnolle F, Thompson J. 1996. An experimental study of ecological differences in winter growth between sympatric diploid and autotetraploid *Dactylis glomerata*. *J Ecol*. 84:343–351.

Bruvo R, Michiels NK, D'Souza TG, Schulenburg H. 2004. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Mol Ecol*. 13:2101–2106.

Cavalli-Sforza LL, Edwards AW. 1967. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet*. 19:233–257.

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc Natl Acad Sci U S A*. 109:1176–1181.

Christie MR, Meirmans PG, Gaggiotti OE, Toonen RJ, White C. 2017. Disentangling the relative merits and disadvantages of parentage analysis and assignment tests for inferring population connectivity. *ICES J Mar Sci*. 74:1749–1762.

Clark LV, Jasieniuk M. 2011. POLYSAT: an R package for polyploid microsatellite analysis. *Mol Ecol Resour*. 11:562–566.

Clark LV, Jasieniuk M. 2012. Spontaneous hybrids between native and exotic Rubus in the Western United States produce offspring both by apomixis and by sexual recombination. *Heredity (Edinb)*. 109:320–328.

Cornille A, Salcedo A, Kryvokhyzha D, Glémin S, Holm K, Wright SI, Lascoux M. 2016. Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Mol Ecol*. 25:616–629.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*. 153:1989–2000.

De Silva HN, Hall AJ, Rikkerink E, McNeilage MA, Fraser LG. 2005. Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity (Edinb)*. 95:327–334.

Dufresne F, Stift M, Vergilino R, Mable BK. 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol Ecol*. 23:40–69.

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

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. 131:479–491.

Field DL, Broadhurst LM, Elliott CP, Young AG. 2017. Population assignment in autopolyploids. *Heredity (Edinb)*. 119:389–401.

Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O. 2015. Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity (Edinb)*. 115:22–28.

Gao H, Williamson S, Bustamante CD. 2007. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*. 176:1635–1651.

Glennon KL, Ritchie ME, Segraves KA. 2014. Evidence for shared broad-scale climatic niches of diploid and polyploid plants. *Ecol Lett*. 17:574–582.

Goudet J. 1995. FSTAT (Version 1.2): a computer program to calculate *F*-statistics. *J Hered*. 86:485–486.

Guo SW, Thompson EA. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*. 48:361–372.

Haldane JBS. 1930. Theoretical genetics of autopolyploids. *J Genet*. 22:359–372.

Hardy O, Vekemans X. 2002. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes*. 2:618–620.

Hardy OJ. 2015. Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Mol Ecol Resour*. 16:103–117.

Hedrick PW. 2005. A standardized genetic differentiation measure. *Evolution*. 59:1633–1638.

Hoban S, Bertorelle G, Gaggiotti OE. 2012. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet*. 13:110–122.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting *F*(ST). *Nat Rev Genet*. 10:639–650.

Huang K, Guo ST, Shattuck MR, Chen ST, Qi XG, Zhang P, Li BG. 2014a. A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity (Edinb)*. 114:133–142.

Huang K, Ritland K, Guo S, Dunn DW, Chen D, Ren Y, Qi X, Zhang P, He G, Li B. 2014b. Estimating pairwise relatedness between individuals with different levels of ploidy. *Mol Ecol Resour*. 15:772–784.

Husband BC, Ozimec B, Martin SL, Pollock L. 2008. Mating consequences of polyploid evolution in flowering plants: current trends and insights from synthetic polyploids. *Int J Plant Sci*. 169:195–206.

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

Jost L. 2008. G(ST) and its relatives do not measure differentiation. *Mol Ecol*. 17:4015–4026.

Kamvar ZN, Tabima JF, Grünwald NJ. 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*. 2:e281.

Kolář F, Čertner M, Suda J, Schönswetter P, Husband BC. 2017. Mixed-ploidy species: progress and opportunities in polyploid research. *Trends Plant Sci*. 22:1041–1055.

Kolář F, Fuxová G, Záveská E, Nagano AJ, Hyklová L, Lučanová M, Kudoh H, Marhold K. 2016. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model *Arabidopsis arenosa*. *Mol Ecol*. 25:3929–3949.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet*. 8:e1002453.

Limborg MT, Larson WA, Seeb LW, Seeb JE. 2017. Screening of duplicated loci reveals hidden divergence patterns in a complex salmonid genome. *Mol Ecol*. 26:4509–4522.

Loiselle B, Sork V, Nason J, Graham C. 1995. Spatial genetic-structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot*. 82:1420–1425.

Luttikhuizen PC, Stift M, Kuperus P, van Tienderen PH. 2007. Genetic diversity in diploid vs. tetraploid *Rorippa amphibia* (Brassicaceae). *Mol Ecol*. 16:3544–3553.

Mable BK. 2004. Polyploidy and self-compatibility: is there an association? *New Phytol*. 162:803–811.

Meirmans PG. 2012a. AMOVA-based clustering of population genetic data. *J Hered*. 103:744–750.

Meirmans PG. 2012b. The trouble with isolation by distance. *Mol Ecol*. 21:2839–2846.

Meirmans PG. 2006. Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution*. 60:2399–2402.

Meirmans PG. 2015. Seven common mistakes in population genetics and how to avoid them. *Mol Ecol*. 24:3223–3231.

Meirmans PG, Den Nijs JC, van Tienderen PH. 2006. Male sterility in triploid dandelions: asexual females vs. asexual hermaphrodites. *Heredity (Edinb)*. 96:45–52.

Meirmans PG, Hedrick PW. 2011. Assessing population structure: F(ST) and related measures. *Mol Ecol Resour*. 11:5–18.

Meirmans PG, van Tienderen PH. 2004. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Notes*. 4:792–794.

Meirmans PG, van Tienderen PH. 2013. The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity (Edinb)*. 110:131–137.

Menken SBJ, Smit E, Nijs HJCMD. 1995. Genetical population structure in plants: gene flow between diploid sexual and triploid asexual dandelions (*Taraxacum section Ruderalia*). *Evolution*. 49:1108–1118.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science*. 201:786–792.

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

Moody ME, Mueller LD, Soltis DE. 1993. Genetic variation and random drift in autotetraploid populations. *Genetics*. 134:649–657.

Mráz P, Gaudeul M, Rioux D, Gielly L, Choler P, Taberlet P. 2007a. Genetic structure of *Hypochaeris uniflora* (Asteraceae) suggests vicariance in the Carpathians and rapid post-glacial colonization of the Alps from an eastern Alpine refugium. *J Biogeogr*. 34:2100–2114.

Mráz P, Singliarová B, Urfus T, Krahulec F. 2007b. Cytogeography of *Pilosella officinarum* (Compositae): altitudinal and longitudinal differences in ploidy level distribution in the Czech Republic and Slovakia and the general pattern in Europe. *Ann Bot*. 101:59–71.

Nei M. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.

Nei M, Chesser RK. 1983. Estimation of fixation indices and gene diversities. *Ann Hum Genet*. 47:253–259.

Neiman M, Paczesniak D, Soper DM, Baldwin AT, Hehman G. 2011. Wide variation in ploidy level and genome size in a New Zealand freshwater snail with coexisting sexual and asexual lineages. *Evolution*. 65:3202–3216.

Neiman M, Sharbel TF, Schwander T. 2014. Genetic causes of transitions from sexual reproduction to asexuality in plants and animals. *J Evol Biol*. 27:1346–1359.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet*. 34:401–437.

Paetkau D, Calvert W, Stirling I, Strobeck C. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol*. 4:347–354.

Parisod C, Broennimann O. 2016. Towards unified hypotheses of the impact of polyploidy on ecological niches. *New Phytol*. 212:540–542.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945–959.

Puechmaille SJ. 2016. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol Ecol Resour*. 16:608–627.

Queller DC, Goodnight KF. 1989. Estimating relatedness using genetic markers. *Evolution*. 43:258–275.

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*. 197:573–589.

Ramsey J, Schemske D. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Ann Rev Ecol Syst*. 29:467–501.

Ramsey J, Schemske D. 2002. Neopolyploidy in flowering plants. *Ann Rev Ecol Syst*. 33:589–639.

Rannala B, Mountain JL. 1997. Detecting immigration by using multilocus genotypes. *Proc Natl Acad Sci U S A*. 94:9197–9201.

Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered*. 86:248–249.

Ritland K. 1990. A series of Fortran computer-programs for estimating plant mating systems. *J Hered*. 81:235–237.

Ritland K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res*. 67:175–185.

Ronfort J, Jenczewski E, Bataillon T, Rousset F. 1998. Analysis of population structure in autotetraploid species. *Genetics*. 150:921–930.

Ryman N, Leimar O. 2009. G(ST) is still a useful measure of genetic differentiation—a comment on Jost's D. *Mol Ecol*. 18:2084–2087.

Simpson EH. 1949. Measurement of diversity. *Nature*. 163:688.

Smouse PE, Peakall R. 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity (Edinb)*. 82 (Pt 5):561–573.

Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends Ecol Evol*. 14:348–352.

Spielmann A, Harris SA, Boshier DH, Vinson CC. 2015. orchard: Paternity program for autotetraploid species. *Mol Ecol Resour*. 15:915–920.

St Onge KR, Källman T, Slotte T, Lascoux M, Palmé AE. 2011. Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. *Mol Ecol*. 20:3306–3320.

Stebbins GL Jr. 1947. Types of polyploids; their classification and significance. *Adv Genet*. 1:403–429.

Stift M, Berenos C, Kuperus P, van Tienderen PH. 2008. Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics*. 179:2113–2123.

Suda J, Weiss-Schneeweiss H, Tribsch A, Schneeweiss GM, Trávnícek P, Schönswetter P. 2007. Complex distribution patterns of di-, tetra-, and hexaploid cytotypes in the European high mountain plant *Senecio carniolicus* (Asteraceae). *Am J Bot*. 94:1391–1401.

van der Meer S, Jacquemyn H. 2015. Genetic diversity and spatial genetic structure of the grassland perennial *Saxifraga granulata* along two river systems. *PLoS One*. 10:e0130463.

van Hengstum T, Lachmuth S, Oostermeijer JGB, Den Nijs JCM, Meirmans PG, Van Tienderen PH. 2012. Human-induced hybridization among congeneric endemic plants on Tenerife, Canary Islands. *Plant Syst Evol*. 298:1119–1131.

Vatsiou AI, Bazin E, Gaggiotti OE. 2015. Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol Ecol*. 25:89–103.

Verduijn M, Van Dijk P, Van Damme J. 2004. Distribution, phenology and demography of sympatric sexual and asexual dandelions (*Taraxacum officinale* s.l.): geographic parthenogenesis on a small scale. *Biol J Linn Soc*. 82:205–218.

Wang J. 2015. Does GST underestimate genetic differentiation from marker data? *Mol Ecol*. 24:3546–3558.

Waples RS. 2015. Testing for Hardy–Weinberg proportions: have we lost the plot? *J Hered*. 106:1–19.

Waples RK, Seeb JE, Seeb LW. 2017. Congruent population structure across paralogous and nonparalogous loci in Salish Sea chum salmon (*Oncorhynchus keta*). *Mol Ecol*. 26:4131–4144.

Weir BS. 1996. *Genetic data analysis II: methods for discrete population genetic data*. Sunderland: Sinauer Associates.

Weir BS, Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution*. 38:1358–1370.

Whitlock MC. 2011. G'_{ST} and D do not replace F_{ST}. *Mol Ecol*. 20:1083–1091.

Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: F_{ST} not equal to 1/(4Nm + 1). *Heredity (Edinb)*. 82 (Pt 2):117–125.

Zwart AB, Elliott C, Hopley T, Lovell D, Young A. 2016. Polypatex: an R package for paternity exclusion in autopolyploids. *Mol Ecol Resour*. 16:694–700.