Resolving microsatellite genotype ambiguity in populations of allopolyploid and diploidized autopolyploid organisms using negative correlations between alleles

Lindsay V. Clark
current email: lvclark@illinois.edu
permanent forwarding email: lindsay.v.clark.04@dartmouth.edu
Department of Crop Sciences
University of Illinois, Urbana-Champaign

Andrea Drauch Schreier
Department of Animal Science
University of California, Davis

May 25, 2015

1 Abstract

A major limitation in the analysis of genetic marker data from polyploid organisms is non-Mendelian segregation, particularly when a single marker yields allelic signals from multiple, independently segregating loci (isoloci). However, with markers such as microsatellites that detect more than two alleles, it is sometimes possible to deduce which alleles belong to which isoloci. Here we describe a novel mathematical property of codominant marker data when it is recoded as binary (presence/absence) allelic variables: under random mating in an infinite population, two allelic variables will be negatively correlated if they belong to the same locus, but uncorrelated if they belong to different loci. We present an algorithm to take advantage of this mathematical property, sorting alleles into isoloci based on correlations, then refining

the allele assignments after checking for consistency with individual genotypes. We demonstrate the utility of our method on simulated data, as well as a real microsatellite dataset from a natural population of octoploid white sturgeon (*Acipenser transmontanus*). Our methodology is implemented in the R package POLYSAT version 1.4.

2 Introduction

Polyploidy, both recent and ancient, is pervasive throughout the plant kingdom [Udall and Wendel, 2006], and to a lesser extent, the animal kingdom [Gregory and Mable, 2005]. However, genetic studies of polyploid organisms face considerable limitations, given that most genetic analyses were designed under the paradigm of diploid Mendelian segregation. In polyploids, molecular markers typically produce signals from all copies of duplicated loci, causing difficulty in the interpretation of marker data [Dufresne et al., 2014]. If signal (e.g. fluorescence in a SNP assay, or peak height of microsatellite amplicons in capillary electrophoresis) is not precisely proportional to allele copy number, partial heterozygotes may be impossible to distinguish from each other (*e.g.* AAAB vs. AABB vs. ABBB) [Clark and Jasieniuk, 2011, Dufresne et al., 2014]. However, under polysomic inheritance (all copies of a locus having equal chances of pairing with each other at meiosis), it is possible to deal with allele copy number ambiguity using an iterative algorithm that estimates allele frequencies, estimates genotype probabilities, and re-estimates allele frequencies until convergence is achieved [De Silva et al., 2005, Falush et al., 2007]. Genotypes cannot be determined with certainty using such methods, but population genetic parameters can be estimated.

The situation is further complicated when not all copies of a locus pair with each other with equal probability at meiosis (referred to as "disomic inheritance" assuming that the locus behaves as multiple independent diploid loci [Obbard et al., 2006]; similarly, one could refer to an octoploid locus as having "tetrasomic inheritance" if it behaved as two tetrasomic loci). In this manuscript we will refer to duplicated loci that do not pair with each other at meiosis (or pair infrequently) as "isoloci" after Obbard et al. [2006]. When a genetic marker consists of multiple isoloci, it is not appropriate to analyze that marker under the assumption of polysomic inheritance; for example, if allele A can be found at both isoloci but allele B is only found at one isolocus in a population, the genotypes AAAB and AABB are possible but ABBB is

not (excluding rare events of meiotic pairing between isoloci). Markers from autopolyploids that have undergone diploidization are likely to behave as multiple isoloci; a locus may still exist in multiple duplicated copies, but the chromosomes on which those copies reside may have diverged so much that they no longer pair at meiosis, or pair with different probabilities [Obbard et al., 2006]. This segregtion pattern is also typically the case in allopolyploids, in which homeologous chromosomes from two different parent species might not pair with each other during meiosis. Further, meotic pairing in allopolyploids may occur between both homologous and homeologous chromosome pairs, but at different rates based on sequence similarity [Gaeta and Pires, 2010, Obbard et al., 2006], which often differs from locus to locus even within a species [Dufresne et al., 2014]. Waples [1988] proposed a method for estimating allele freqencies in polyploids under disomic inheritance, although it requires that allele dosage can be determined in heterozygotes (in his example, by intensity of allozyme bands on a gel) and allows a maximum of two alleles per locus, with both isoloci posessing both alleles. De Silva et al. [2005] describe how their method for estimating allele frequencies under polysomic inheritance, allowing for multiple alleles, can be extended to cases of disomic inheritance, but require that isoloci have non-overlapping allele sets, and do not address the issue of how to determine which alleles belong to which isolocus.

Given that marker data do not follow straighforward Mendelian laws in polyploid organisms, they are often recoded as a matrix of ones and zeros reflecting the presence and absence of alleles (sometimes referred to as "allelic phenotypes") [Obbard et al., 2006]. In mapping populations such binary data are useful if one parent is heterozygous for a particular allele and the other parent lacks that allele, in which case segregation follows a 1:1 ratio and can be analyzed under the diploid testcross model [Swaminathan et al., 2012, Rousseau-Gueutin et al., 2008]. However, in natural populations, inheritance of dominant (presence/absence) markers typically remains ambiguous, and such markers are treated as binary variables that can be used to assess similarity among individuals and populations but are inappropriate for many population genetic analyses, *e.g.* tests that look for departures from or make assumptions of Hardy-Weinberg Equilibrium [Clark and Jasieniuk, 2011].

Microsatellites are a special case given that they have multiple alleles, allowing for the possibility of assigning alleles to isoloci, which would drastically reduce the complexity of interpreting genotypes in allopolyploids and diploidized autopolyploids. For example, if an allotetraploid individual has

alleles A, B, and C, and if A and B are known to belong to one isolocus and C to the other, the genotype can be recoded as AB at one isolocus and CC at the other isolocus, and the data can be subsequently processed as if they were diploid. If two isoloci are sufficiently diverged from each other, they may have entirely different sets of alleles. This is in contrast to other markers such as SNPs and AFLPs that only have two alleles, in which case isoloci must share at least one allele (or be monomorphic, and therefore uninformative). With microsatellites, one could hypothetically examine all possible combinations of allele assignments to isoloci and see which combination was most consistent with the genotypes observed in the dataset, but this method would be impractical in terms of computation time and so alternative methods are needed. Catalán et al. [2006] proposed a method for assigning microsatellite alleles to isoloci based on inspection of fully homozygous genotypes in natural populations. In their example with an allotetraploid species, any genotype with just two alleles was assumed to be homozygous at both isoloci, and therefore those two alleles could be inferred to belong to different isoloci. With enough unique homozygous genotypes, all alleles could be assigned to one isolocus or the other, and both homozygous and heterozygous genotypes could be resolved. However, their method made the assumption of no null alleles, and would fail if it encountered any homoplasy between isoloci (alleles identical in amplicon size, but belonging to different isoloci). Moreover, in small datasets or datasets with rare alleles, it is likely that some alleles in the dataset will never be encountered in a fully homozygous genotype. The method of Catalán et al. [2006] was never implemented in any software to the best of our knowledge, despite being the only published methodology for splitting polyploid microsatellite genotypes into diploid isoloci.

In this manuscript, we present a novel methodology for assigning microsatellite alleles to isoloci based on the distribution of alleles among genotypes in the dataset. Our method is appropriate for both mapping populations and natural populations, as long as the dataset can be split into reasonably-sized groups of individuals ($\sim$ 100 individuals or more) lacking strong population structure. Negative correlations between alleles are used to cluster alleles into putative isolocus groups, which are then checked against individual genotypes. If necessary, alleles are swapped between clusters or declared homoplasious so that the clusters agree with the observed genotypes within a certain error tolerance. Genotypes can then be recoded, with each marker split into two or more isoloci, such that isoloci can then be analyzed as diploid or polysomic markers. Our method still works when there are null

alleles, homoplasy between isoloci, or occasional meiotic recombination between isoloci, albeit with reduced power to find the correct set of allele assignments. We test our methodology on simulated allotetraploid, allohexaploid, and allo-octoploid (having two tetrasomic genomes) data, and compare its effectiveness to that of the method of Catalán et al. [2006]. We also demonstrate the utility of our method on a real dataset from a natural population of octoploid white sturgeon (*Acipenser transmontanus*). Our methodology, as well as a modified version of the Catalán et al. [2006] methodology, are implemented in the R package POLYSAT version 1.4.

3 Alleles in an unstructured population are negatively correlated if they belong to the same locus

Say that a microsatellite dataset is recoded as an "allelic phenotype" matrix, such that each row represents one individual, and each allele becomes a column (or an "allelic variable") of ones and zeros indicating whether that allele is present in that individual or not. Under Hardy-Weinberg equilibrium and in the absence of linkage disequilibrium, these allelic variables are expected to be uncorrellated if the alleles belong to different loci or different isoloci. However, if two alleles belong to the same locus (or isolocus), the allelic variables should be negatively correlated. This is somewhat intuitive given that the presence of a given allele means that there are fewer locus copies remaining in which the other allele might appear (Fig. 1). The negative correlation can also be proved mathematically.

Define a set of three or more alleles numbered $1 \cdots n$ that all belong to one isolocus. k and j are alleles in this set. Allele frequencies in the population being sampled are defined as $p_1 \cdots p_n$, where each allele frequency is between one and zero, and all allele frequencies sum to one. G_k means that allele k is present in a given individual's genotype.

In a diploid, the probability that allele k is present in an individual is

$$P(G_k) = p_k^2 + 2p_k * (1 - p_k) = 2p_k - p_k^2 \tag{1}$$

and the probability that allele k is absent in an individual is

$$P(-G_k) = (1 - p_k)^2 \tag{2}$$

Given that allele j is present, the conditional probability that allele k is also present is

$$P(G_k|G_j) = \frac{P(G_k \bigcap G_j)}{P(G_j)} = \frac{2p_k p_j}{2p_j - p_j^2} = \frac{2p_k}{2 - p_j} \tag{3}$$

and given that allele j is absent, the conditional probability that allele k is present is

$$P(G_k| - G_j) = \frac{P(G_k \bigcap -G_j)}{P(-G_j)} = \frac{2p_k - p_k^2 - 2p_k p_j}{(1 - p_j)^2} \tag{4}$$

It follows that

$$P(G_k| - G_j) - P(G_k|G_j) = \frac{p_k[2(1 - p_k - p_j) + p_k p_j]}{(1 - p_j)^2(2 - p_j)} \tag{5}$$

Given that $0 < p_k < 1$, $0 < p_j < 1$, and $p_k + p_j < 1$, $P(G_k| - G_j) - P(G_k|G_j) > 0$ and therefore $P(G_k| - G_j) > P(G_k|G_j)$. The magnitude of difference between these probabilities is dependent on allele frequencies, but the presence of allele j in a genotype always reduces the probability that allele k is also present. A proof of the same principal at a tetrasomic locus is provided in appendix 1.

4 Methods for clustering alleles into isoloci using negative correlations

As demonstrated in the previous section, the occurences of two different alleles across individuals in a panmictic population should be negatively correlated if the two alleles belong to the same isolocus, and uncorrelated if they belong to different isoloci. Therefore, tests for negative correlation can be used to guide the assignment of alleles to isoloci. For any pair of alleles, a two-by-two contingency table can be generated, containing counts of individuals that have both alleles, neither allele, the first allele but not the second, and the second allele but not the first. A statistical test is then performed to check for independence of the two allelic variables; we use Fisher's exact test here because it is appropriate for small sample sizes, which are likely to occur in typical population genetics datasets when rare alleles are

present. A one-tailed Fisher's exact test is used, with the alternative hypothesis being that a disproportionately high number of individuals will just have one allele of the pair, as opposed to both alleles or neither allele. This alternative hypothesis corresponds to the two alleles belonging to the same isolocus, whereas the null hypothesis is that they belong to different isoloci and therefore assort independently. The P-values from Fisher's exact test on each pair of allelic variables from a single microsatellite marker are then stored in a symmetric square matrix. We expect to see clusters of alleles with low P-values between them; alleles within a cluster putatively belong to the same isolocus. For clustering algorithms, zeros are inserted along the diagonal of the matrix, since the P-values are used as a dissimilarity statistic. The function `alleleCorrelations` in POLYSAT 1.4 produces such a matrix of P-values for a single microsatellite marker.

Multiple clustering algorithms exist for square matrices. Here we test K-means and hierarchical clustering for their ability to correctly assign alleles to isoloci based on P-values produced by Fisher's exact test. K-means clustering places rows of a matrix into clusters, where each row is more similar to the mean of its own cluster than to the mean of any other cluster. We use the Hartigan and Wong [1979] method of K-means clustering as it is the default method implemented in R, and we found that if we set the number of randomly chosen starting centroids high enough (`nstart = 50` in the `kmeans` function) it converged on the same answer as all other methods of K-means clustering (data not shown). One potential issue with K-means clustering is that since it only seeks to group similar rows together, it can group pairs of alleles with high P-values rather than low P-values. Hierarchical clustering, on the other hand, treats the values in the square matrix as a dissimilarity statistic, and seeks to cluster objects that are most similar. We used three methods of hierarchical clustering (complete linkage, single linkage, and UPGMA, which differ in how they use distances between elements of clusters to determine distances between clusters) on simulated tetraploid, hexaploid, and octoploid datasets, and found that UPGMA was the most accurate for assigning alleles to isoloci for all ploidies (Table 1). K-means was more accurate than UPGMA for all ploidies, and the method of Catalán et al. [2006] had similar accuracy to K-means for allotetraploid datasets, but performed much more poorly at higher ploidies (Table 1). Although K-means was more accurate overall than UPGMA, UPGMA sometimes found the correct answer when K-means found the incorrect answer, and therefore both results are output by the `alleleCorrelations` function in POLYSAT. To choose be-

tween K-means and UPGMA when they give different results, the function `testAlGroups` in POLYSAT checks every genotype in the dataset against both results. A genotype is consistent with a set of allele assignments if it has at least one allele belonging to each isolocus, and no more alleles belonging to each isolocus than the ploidy of that isolocus (*e.g.* two in an allotetraploid). The set of results that is consistent with the greatest number of genotypes is selected, or K-means in the event of a tie. Selecting the best results out of K-means and UPGMA improved the accuracy of allele assignments at all ploidies, particularly hexaploids (Table 1).

When the best set of allele assignments was chosen from K-means and UPGMA, 13% of tetraploid datasets, 31% of hexaploid datasets, and 41% of octoploid datasets still had incorrect allele assignments (Table 1). We expected that rare alleles would be the most likely to be assigned incorrectly, given that they would be present in the fewest genotypes and therefore there would be the least statistical power to detect correlations between them and other alleles. To correct the allele assignments, an algorithm was added to the `testAlGroups` function that individually swaps the assignment of each rare allele to the other isolocus (or isoloci) and then checks whether the new set of assignments is consistent with a greater number of genotypes than the old set of assignments. If an allele is successfully swapped, then every other rare allele is checked once again, until no more swaps are made. The maximum number of genotypes in which an allele must be present to be considered a rare allele is adjusted using the `rare.al.check` argument to the `testAlGroups` function. On the same set of datasets that were used to compare K-means and hierarchical clustering methods to the Catalán et al. [2006] method, we tested the accuracy of allele assignments when alleles present in $\leq 25\%$ or $\leq 50\%$ of genotypes were subjected to the swapping algorithm (Table 1). Note that the frequency of genotypes with a given allele will always be higher than the allele frequency itself, although a 50% threshold is still much higher than the cutoff for considering an allele to be "rare" in most population genetic analyses. For all ploidies, swapping rare alleles between isoloci resulted in a considerable improvement in accuracy.

The accuracy of allele assignment was dependent on the number of alleles belonging to the two isoloci in the simulated datasets. The assignments for tetraploids were very accurate (98-100%) when both isoloci had the same number of alleles, but decreased in accuracy as the difference in number of alleles between the two isoloci increased (Table 2). When one isolocus had two alleles and the other had eight alleles, the accuracy was 80% (Table 2).

The accuracy of allele assignments would be expected to increase with sample size for our method, given that power to detect correlation is very dependent on sample size. The Catalán et al. [2006] method should also improve with increasing sample size, given that with more individuals in the dataset, there is a greater probability of producing the double homozygotes that are needed to resolve the allele assignments. For all ploidies, we found that the accuracy of both our method and the Catalán et al. [2006] method was dependent on sample size, and that our method performed better than the Catalán et al. [2006] method at all sample sizes (Fig. 2). For tetraploids and hexaploids, the effect of sample size was greater on the Catalán et al. [2006] method than on our method, particularly at small sample sizes (Fig. 2). For octoploids, the success of the Catalán et al. [2006] method was near zero even with 800 individuals in the dataset (due to the low probabiltiy of producing fully homozygous genotypes at tetrasomic isoloci), whereas our method had an accuracy of 93% with 800 octoploid individuals.

5 Caveats of the method

Population structure. Both negative and positive correlations between alleles at different loci (or isoloci) can occur when the assumption of random mating is violated by population structure, confounding the use of negative correlations for assigning alleles to isoloci. We simulated allotetraploid datasets consisting of two populations of fifty individuals differing in allele frequency by a predetermined amount, and found that accuracy of our method remained high ($\sim 90\%$) even at moderate levels of F_{ST} (~ 0.2; Table 3). Interestingly, low levels of population structure ($F_{ST} \approx 0.02$) improved the accuracy of our method to 99%, compared to 94% when $F_{ST} = 0$ (Table 3), probably as a result of an increase in the number of double homozygous genotypes, which would have been informative during the allele swapping step. For this same reason, the Catalán et al. [2006] method, which depends on double homozygous genotypes, had an improved success rate as population structure increased, and exceeded our method in accuracy at moderate levels of F_{ST} (Table 3). However, accuracy of our method decreased with increasing F_{ST} when $F_{ST} > 0.02$ (Table 3), likely because correlations between alleles caused by population structure outweighed the benefits of increased homozygosity. Because our method can be negatively impacted by population structure, the `alleleCorrelations` function checks for significant positive correlations be-

tween alleles, which could only be caused by population structure or scoring error (such as stutter peaks being mis-called as alleles, and therefore tending to be present in the same genotypes as their corresponding alleles), and prints a warning if such correlations are found. In our simulations, significant postive correlations between alleles were found in most datasets that had moderate population structure (Table 3).

Because the user may want to split highly structured datasets into multiple populations for making allele assignments (to avoid the issue of allele correlations caused by population stucture), the function `mergeAlleleAssignments` is included in POLYSAT 1.4 in order to merge several sets of allele assignments into one. This is particularly useful if some alleles are found in some populations but not others, or if alleles with identical amplicon sizes are found belonging to different isoloci in different populations.

Homoplasy. Although our algorithm attempts primarily to sort alleles into non-overlapping groups, there is always a possibility that different isoloci have some alleles with identical amplicon sizes. Therefore, we introduced an algorithm to the `testAlGroups` function to check whether any genotypes were still inconsistent with the allele assignments after the allele swapping step, and assign alleles to multiple isoloci until all genotypes (or a particular proportion that can be adjusted with the `threshold` argument) are consistent with the allele assignments. The allele that could correct the greatest number of inconsistent genotypes (or in the event of a tie, the one with the lowest P-values from Fisher's exact test between it and the alleles in the other isolocus) is made homoplasious first, then all genotypes are re-checked and the cycle is repeated until the desired level of agreement between allele assignments and genotypes is met.

We tested the accuracy of allele assignments across several sample sizes and frequencies of homoplasious alleles, with and without allele swapping beforehand (Fig. 3). Allele assignments were most accurate when allele swapping was not performed before testing for homoplasious alleles, and when the homoplasious allele was at a frequency of 0.3 in both isoloci. In addition to the issue of assignment accuracy, there is the fact that many genotypes may not be unambiguously resolvable when they contain a homoplasious allele. For example, if alleles A and B belong to different isoloci, and C belongs to both, the genotype ABC could be AA BC, AC BB, or AC BC. In such cases, the POLYSAT function `recodeAllopoly`, which recodes genotypes from single loci into multiple isoloci, marks the entire genotype as missing. When allele assignments were correct, we tested the mean proportion of genotypes that

were resolvable, given several frequencies of a homoplasious allele (Table 4). Although accuracy of assignment had been highest with a homoplasious allele frequency of 0.3, only 57% of genotypes could be resolved in such datasets (Table 4).

Homoplasy between alleles within an isolocus is also possible, meaning that two or more alleles belonging to one isolocus are identical in amplicon size but not identical by descent. Although such homoplasy is an important consideration for analyses that determine similarity between individuals and populations, homoplasy within isoloci does not affect the allele assignment methods described in this manuscript.

Null alleles. Mutations in primer annealing sites are a common occurrence with microsatellite markers, and result in alleles that produce no PCR product, known as null alleles. To test the effect of null alleles on the accuracy of our allele assignment method, we simulated datasets in which one isolocus had a null allele (Fig. 4). One potential issue with null alleles is that, when homozygous, they can result in genotypes that do not appear to have any alleles from one isolocus. Such genotypes are used by the `testAlGroups` function as an indicator that alleles should be swapped or made homoplasious, which would be incorrect actions if the genotype resulted from a null allele rather than inaccuracy of allele assignment. We therefore added an argument to the `testAlGroups` function, `null.weight`, to indicate how genotypes with no apparent alleles for one isolocus should be prioritized for determining which alleles to make homoplasious. If null alleles are expected to be common, `null.weight` can be set to zero so that genotypes with no apparent alleles for one isolocus are not used for assigning homoplasy. The default value of 0.5 for `null.weight` will cause `testAlGroups` to use genotypes with no apparent alleles for one isolocus as evidence of homoplasy, but with lower priority than genotypes with too many alleles per isolocus. (No argument was added to adjust the allele swapping algorithm, since it only swaps alleles if the overall agreement with the dataset is improved.) We found that, when null alleles were present, the accuracy of the algorithm was greatly improved when genotypes lacking alleles for one isolocus were not used as evidence of homoplasy (Fig. 4). We also found that the allele swapping algorithm still improved the accuracy of allele assignments, particularly when the null allele was at a frequency greater than 0.25 in the population.

It is also possible for an entire isolocus to be null. This is often apparent when a marker has fewer alleles per genotype than expected, *e.g.* a maximum of two alleles per individual in a tetraploid. Such loci should be

excluded from the allele assignment analysis described in this manuscript. If they are included in analysis accidentally, they can be identified by weak K-means/UPGMA clustering of alleles (which can be evaluated from the `betweenss`, `totss`, and `heatmap.dist` outputs from `alleleCorrelations`) and by a high proportion of alleles appearing to be homoplasious.

Meiotic error due to intermediate inheritance patterns. As mentioned in the introduction, although homologous chromosomes are likely to preferentially pair with each other, occasional pairing between homeologous (in an allopolyploid) or paralogous (in an autopolyploid) chromosomes may occur during meiosis. As a result, offspring may be aneuploid, having too many or two few chromosomes from either homologous pair, or may have translocations between homeologous or paralogous chromosomes. Most commonly, the aneuploidy or translocations will occur in a compensated manner [Chester et al., 2015], meaning that for a given pair of isoloci, the total number of copies will be the same as in a non-aneuploid, but one isolocus will have more copies than expected and the other isolocus will have fewer (*e.g.* three copies of one isolocus and one copy of the other isolocus in an allotetraploid).

We simulated datasets in which gametes resulting in compensated aneuploidy (meiotic error) occured at a range of frequencies from 0.01 to 0.2 (Fig. 5). At all meiotic error rates, the allele swapping algorithm from `testAlGroups` improved the accuracy of allele assignment (Fig. 5). Meiotic error did not have a large impact on the success of our method; even at a meiotic error rate of 0.2 (where 0.5 would be fully autopolyploid), our algorithm still had an accuracy of 62% on datasets of 100 individuals with no homoplasy, null alleles, or population structure (Fig. 5). Although the allele swapping algorithm assumes fully disomic (or tetrasomic in the case of an octoploid with two subgenomes) inheritance in the dataset, allele swapping still improves accuracy of allele assignments when inheritance is not fully disomic (Fig. 5).

To avoid calling alleles as homoplasious when meiotic error or intermediate inheritance causes some genotypes to have too many alleles per isolocus (or none, as with null alleles), the `tolerance` argument was included for the `testAlGroups` function. This argument indicates the proportion of genotypes that can still be in disagreement with the allele assignments when the algorithm stops adding homoplasious alleles. (For example, if `tolerance` is zero, all genotypes must be consistent with the allele assignments, if `tolerance` is one the algorithm will not check for homoplasy at all, and at the default of 0.05, the algorithm will keep adding homoplasious alleles

until no more than 5% of genotypes disagree with the allele assignments.) Allowing for a few genotypes to disagree with the allele assignments is also expected to make the algorithm more robust to occasional scoring error. The function `recodeAllopoly` also has an argument, `allowAneuploidy`, to allow for meiotic error when splitting markers into multiple isoloci. For genotypes with too many alleles for one isolocus, the function will adjust the recorded ploidy for the relevant samples and isoloci. (Ploidy is used by other POLYSAT functions, such as those that estimate allele frequency.)

Fixed alleles. If one or more alleles are present in all genotypes in a dataset, it is not possible to perform Fisher's exact test to look for correlations between those fixed alleles and any others. The function `alleleCorrelations` therefore checks for fixed alleles before performing Fisher's exact test. Each fixed allele is assigned to its own isolocus. If only one isolocus remains, all remaining alleles are assigned to it. If no isoloci remain (*e.g.* in an allotetraploid with two fixed alleles and several variable alleles), then all remaining alleles are assigned as homoplasious to all isoloci. If multiple isoloci remain (*e.g.* in an allohexaploid with one fixed allele), then Fisher's exact test, K-means clustering and UPGMA are performed to assign the alleles to the remaining isoloci. It is possible that an allele with a very high frequency may be present in all genotypes but not truly fixed (*i.e.* some genotypes are heterozygous). However, allele swapping performed by `testAlGroups` can assign alleles to an isolocus even if that isolocus already has an allele assigned to it that is present in all individuals.

6 Assignment of alleles to isoloci in octoploid sturgeon

To demonstrate the usefulness of our allele assignment method on a real dataset, we used previously published data from natural populations of octoploid white sturgeon (*Acipenser transmontanus*) [Drauch Schreier et al., 2012]. Previous studies of inheritance patterns in this species suggested that it possesses two tetrasomic subgenomes, at least for portions of its genome [Rodzen and May, 2002, Drauch Schreier et al., 2011]. We selected for analysis the eight microsatellite markers that, based on number of alleles per genotype, appeared to be present in eight copies rather than four.

Because population structure can impact allele clustering, we first per-

formed a preliminary analysis of population structure using a simple genetic dissimilarity statistic and principal coordinates analysis (Fig. 6). The two major genetic groups that we identified (Table 5) were similar to the population structure previously observed, which divides the Fraser River white sturgeon population into two genetic groups that exist on either side of a natural obstruction in the river [Drauch Schreier et al., 2012]. The smaller group (Pop 2) consisted of only 66 individuals and, likely due to small sample size, produced poor quality allele assignments with high levels of homoplasy when analyzed by itself (data not shown). We therefore tested our method on Pop 1 (183 individuals) and on the combined set of 249 individuals. Because simulations indicated that the allele swapping algorithm reduced the accuracy when there was homoplasy in the dataset (Fig. 3), we tried allele assignment with and without allele swapping. In checking for homoplasy, we allowed up to 5% of genotypes to disagree with allele assignments in anticipation of meiotic error, scoring error, or genotypes homozygous for null alleles (`tolerance = 0.05` in `testAlGroups`), and to allow for null alleles at low frequency we set `null.weight = 0.5` so that genotypes with too many alleles per isolocus would be used for assignment of homoplasy first, before genotypes with no alleles for one of their isoloci.

For four out of eight loci, our algorithm found allele assignments devoid of homoplasy when only Pop 1 was used for assignment and when the allele swapping algorithm was used (Table 6). Eliminating the allele swapping algorithm or using the whole dataset for allele assignment increased the number of apparent homoplasious alleles in most cases, and did not reduce the number of apparent homoplasious alleles for any locus (Table 6). For the four loci with homoplasy, most genotypes in the dataset could not be assigned unambiguously (Table 6). For the four loci with no apparent homoplasy, nearly all genotypes in Pop 1 could be assigned unambiguously, and approximately three quarters of the genotypes in Pop 2 (which was not used for creating the assignments) could be assigned unambiguously (Table 6). Despite the fact that Pop 1 was previously determined to consist of three subpopulations with Phi-PT values ranging from 0.06 to 0.17 [Drauch Schreier et al., 2012], allele correlations resulting from population structure did not appear to prevent us from obtaining reasonable assignments of alleles to isoloci. Significant positive correlations between alleles were found at one and two out of eight loci when Pop 1 and the whole dataset were used to make assignments, respectively (data not shown).

7 Conclusions

Here we introduce the R package POLYSAT version 1.4, with several new functions applicable to the analysis of allopolyploids and diploidized autopolyploids. These include `simAllopoly`, which generates simulated datasets; `catalanAlleles`, which uses the the Catalán et al. [2006] method to assign alleles to isoloci; `alleleCorrelations`, which performs Fisher's exact test between each pair of alleles from a marker, and then uses K-means clustering and UPGMA to make initial assignments of alleles to isoloci; `testAlGroups`, which checks the consistency of allele assignments with individual genotypes, chooses between the K-means and UPGMA method, swaps alleles to different isoloci if it improves consistency, and identifies homoplasious alleles; `mergeAlleleAssignments`, which merges the allele assignments from two different populations using the same microsatellite marker; and `recodeAllopoly`, which uses allele assignments to recode the dataset, splitting each microsatellite marker into multiple isoloci. An overview of the data analysis workflow is given in Fig. 7.

We found that, with simulated data, the accuracy of our allele assignment algorithm was impacted by issues such as homoplasy and null alleles, and that the optimal parameters for the algorithm depended on which of these issues were present in the dataset. This suggests, since most users will not know whether their dataset has homoplasy or null alleles, that the `testAlGroups` function should initially be run with several different parameter sets, and for each locus, the results with the fewest homoplasious alleles should be chosen. A heatmap of the P-values generated from Fisher's exact test can also serve as a qualitative visual indicator of how well the alleles can be separated into isolocus groups. We also found that, although our algorithm was negatively impacted by meiotic error (pairing of non-homologous chromosomes during meiosis) and moderate population structure, its accuracy remained fairly high in both cases. Sensitivity to population structure is the biggest drawback of our method in comparison to that of Catalán et al. [2006], which actually has improved results as population structure increases. However, even low frequencies of null alleles, homoplasy, or meiotic error can prevent the method of Catalán et al. [2006] from working entirely.

Using a real microsatellite dataset from natural populations of white sturgeon, we found that our method was useful for recoding half of the markers into two indpendently segregating loci each. Given that white sturgeon are octoploid with two tetrasomic subgenomes [Drauch Schreier et al., 2011], we

expected this dataset to be problematic; having tetrasomic isoloci as opposed to disomic isoloci would reduce the magnitude of the negative correlations between alleles, and was observed in simulations to reduce the accuracy of assignment using our method, although not nearly as severly as the reduction in efficacy of the Catalán et al. [2006] method (Table 1, Fig. 2). In population genetic studies, we expect that microsatellite markers that can be recoded using our method could then be used for analyses requiring polysomic or disomic inheritance (for example, Structure [Falush et al., 2007], estimation of allele frequency, or tests of Hardy-Weinberg Equilibrium), while the remaining markers will still be useful for other analysis (for example, Mantel tests using simple dissimilarity statistics). Additionally, we found that the allele assignments that we made were still fairly useful for recoding genotypes in a population that was not used for making the assignments.

Although inappropriate for biallelic marker systems such as single nucleotide polymorphisms (SNPs) and dominant marker systems such as AFLPs, the method that we have described could theoretically used to assign alleles to isoloci in any marker system in which multiple alleles are the norm. Allozymes, although rarely used in modern studies, are one such system. Although data from genotyping-by-sequencing (GBS, and the related technique restriction site-associated DNA sequencing, or RAD-seq) are typically processed to yield biallelic SNP markers, in the future as typical DNA sequencing read lengths increase, it may become common to find multiple SNPs within the physical distance covered by one read. In that case, haplotypes may be treated as alleles, and negative correlations between haplotypes may be used to assign them to isoloci.

8 Obtaining polysat 1.4

To obtain POLYSAT, first install the most recent version of R (available at `http://www.r-project.org`), launch R, then at the prompt type:

```
install.packages("polysat")
```

In the "doc" subdirectory of the package installation, PDF tutorials are available for POLYSAT as a whole and for the methodology described in this manuscript.

9 Supplementary files

- polysat_1.4-0.tar.gz: Source package for POLYSAT 1.4.
- polysat_1.4-0.zip: Microsoft Windows installation file for POLYSAT 1.4.
- allopolyVignette.pdf: Tutorial for creating and using allele assignments in POLYSAT.
- tables_figs.R, sturgeontest.R: R scripts for reproducing the analyses in this manuscript.
- sturgeon.csv: White sturgeon microsatellite dataset used in strugeontest.R.
- appendix1.pdf: Proof that alleles belonging to the same tetrasomic locus are negatively correlated in an infinite, unstructured population.

10 References

References

P. Catalán, J. G. Segarra-Moragues, M. Palop-Esteban, C. Moreno, and F. González-Candelas. A Bayesian approach for discriminating among alternative inheritance hypotheses in plant polyploids: the allotetraploid origin of genus *Bordera* (Dioscoreaceae). *Genetics*, 172(3):1939–1953, 2006.

M. Chester, R. K. Riley, P. S. Soltis, and D. E. Soltis. Patterns of chromosomal variation in natural populations of the neoallotetraploid *Tragopogon mirus* (Asteraceae). *Heredity*, 114(3):309–317, 2015.

L. V. Clark and M. Jasieniuk. Polysat: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, 11(3):562–566, 2011.

H. N. De Silva, A. J. Hall, E. Rikkerink, M. A. McNeilage, and L. G. Fraser. Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity*, 95(4):327–334, 2005.

575 A. Drauch Schreier, D. Gille, B. Mahardja, and B. May. Neutral mark-
576 ers confirm the octoploid origin and reveal spontaneous autopolyploidy in
577 white sturgeon, *Acipenser transmontanus*. *Journal of Applied Ichthyology*,
578 27(Suppl. 2):24–33, 2011.

579 A. Drauch Schreier, B. Mahardja, and B. May. Hierarchical patterns of pop-
580 ulation structure in the endangered fraser river white sturgeon (*Acipenser*
581 *transmontanus*) and implications for conservation. *Canadian Journal of*
582 *Fisheries and Aquatic Sciences*, 69(12):1968–1980, 2012.

583 F. Dufresne, M. Stift, R. Vergilino, and B. K. Mable. Recent progress and
584 challenges in population genetics of polyploid organisms: an overview of
585 current state-of-the-art molecular and statistical tools. *Molecular Ecology*,
586 23(1):40–69, 2014.

587 D. Falush, M. Stephens, and J. K. Pritchard. Inference of population struc-
588 ture using multilocus genotype data: dominant markers and null alleles.
589 *Molecular Ecology Notes*, 7(4):574–578, 2007.

590 R. T. Gaeta and J. C. Pires. Homeologous recombination in allopolyploids:
591 the polyploid ratchet. *New Phytologist*, 186(1):18–28, 2010.

592 T. R. Gregory and B. K. Mable. Polyploidy in animals. In T. R. Gregory,
593 editor, *The Evolution of the Genome*, chapter 8, pages 427–517. Elsevier,
594 San Diego, 2005.

595 J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Journal of*
596 *the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100–108,
597 1979.

598 D. J. Obbard, S. A. Harris, and J. R. Pannell. Simple allelic-phenotype
599 diversity and differentiation statistics for allopolyploids. *Heredity*, 97(4):
600 296–303, 2006.

601 J. A. Rodzen and B. May. Inheritance of microsatellite loci in white sturgeon
602 (*Acipenser transmontanus*). *Genome*, 45(6):1064–1076, 2002.

603 M. Rousseau-Gueutin, E. Lerceteau-Köhler, L. Barrot, D. J. Sargent,
604 A. Monfort, D. Simpson, P. Arús, G. Guérin, and B. Denoyes-Rothan.
605 Comparative genetic mapping between octoploid and diploid *Fragaria*

species reveals a high level of colinearity between their genomes and the essentially disomic behavior of the cultivated octoploid strawberry. *Genetics*, 179(4):2045–2060, 2008.

K. Swaminathan, W. B. Chae, T. Mitros, K. Varala, L. Xie, A. Barling, K. Glowacka, M. Hall, S. Jezowski, R. Ming, M. Hudson, J. A. Juvik, D. S. Rokshar, and S. P. Moose. A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. *BMC Genomics*, 13:142, 2012.

J. A. Udall and J. F. Wendel. Polyploidy and crop improvement. *Crop Science*, 46(S1):S3–S14, 2006.

R. S. Waples. Estimation of allele frequencies at isoloci. *Genetics*, 118(2): 371–384, 1988.

11 Tables and Figures

Clustering method	4x = 2x + 2x	6x = 2x + 2x + 2x	8x = 4x + 4x
Catalán	83.3%	21.0%	0.4%
K-means	84.0%	59.0%	57.9%
Complete linkage	59.1%	32.9%	25.9%
Single linkage	34.3%	12.5%	6.8%
UPGMA	68.8%	49.7%	33.3%
K-means + UPGMA	86.6%	68.6%	59.3%
K-means + UPGMA + swap ≤ 0.25	95.0%	83.7%	62.8%
K-means + UPGMA + swap ≤ 0.50	95.0%	84.3%	64.4%

Table 1: Percentages of simulated datasets with correct assignments of alleles to isoloci using different clustering methods. For each type of ploidy tested, including tetraploid (4x = 2x + 2x), hexaploid (6x = 2x + 2x + 2x), and octoploid (8x = 4x + 4x), 10,000 datasets were simulated with 100 individuals each, with two to eight alleles at each isolocus and randomly generated allele frequencies. Data were simulated without homoplasy, null alleles, population structure, or meiotic error. Datasets were simulated using the `simAllopoly` function in POLYSAT 1.4. The Catalán et al. [2006] method, as implemented in the `catalanAlleles` function in POLYSAT 1.4, was used to attempt to assign alleles to isoloci by examining fully-homozygous genotypes. Note that if the Catalán method fails to find the correct answer, it will not produce any results, unlike the other methods listed in this table. For all other methods in the table, a square matrix was calculated for each dataset, containing P-values from Fisher's exact test for negative correlation between each pair of alleles, using the R function `fisher.test` with `alternative = "less"`. K-means clustering was performed using the Hartigan and Wong [1979] method as implemented in the R function `kmeans`, with 50 random sets chosen (`n.start = 50`) to ensure convergence on the most optimal set of clusters. Hierarchical clustering was performed in R with the `hclust` (`method = "complete"` for complete linkage, `method = "single"` for single likage, and `method = "average"` for UPGMA) and `cutree` functions. For "K-means + UPGMA", both methods were performed, and if they produced different results, the results that were consistent with the greatest number of genotypes were retained. For the "swap ≤ 0.25" and "swap ≤ 0.50" methods, alleles occuring in $\leq 25\%$ or $\leq 50\%$, respectively, of individuals were tested to see if agreements between assignments and genotypes were improved when those alleles were individually swapped to different isoloci, in which case the swaps were retained. In POLYSAT 1.4, the Fisher's exact test, K-means clustering, and UPGMA steps are included in the `alleleCorrelations` function, whereas the comparison of K-means and UPGMA results as well as allele swapping are implemented in the `testAlGroups` function.

	2	3	4	5	6	7	8
2	98.0%	97.0%	94.9%	89.0%	90.8%	83.7%	80.1%
3		100.0%	98.8%	97.5%	94.0%	93.3%	89.3%
4			98.5%	98.2%	96.6%	95.7%	94.6%
5				99.0%	99.0%	96.8%	96.6%
6					98.0%	97.8%	98.3%
7						98.2%	99.1%
8							97.9%

Table 2: Percentages of allele assignments for allotetraploid datasets from Table 1 that were correct using K-means + UPGMA + swap ≤ 0.50, by the number of alleles at each of two isoloci.

Difference in allele frequency	F_{ST}	Significant positive correlations	K-means + UPGMA + swap ≤ 0.50	Catalán
0.0	0.000 ± 0.000	0%	94%	84%
0.1	0.016 ± 0.004	2%	99%	89%
0.2	0.062 ± 0.013	21%	93%	94%
0.3	0.117 ± 0.021	62%	88%	99%
0.4	0.176 ± 0.026	82%	88%	100%

Table 3: Percentages of simulated datasets with correct allele assignments under different levels of population structure. Allotetraploid datasets were simulated as in Table 1, but instead of one population of 100 individuals, two populations of 50 individuals were simulated under different allele frequencies, then merged into one dataset that was then used for making allele assignments. The value shown in the leftmost column was randomly added or subtracted from the frequency of each allele in the first population to generate the allele frequencies of the second population. For isoloci with odd numbers of alleles, one allele had the same frequency in both populations. For each difference in allele frequency, 1000 simulations were performed (5000 total). F_{ST} was calculated from allele frequencies as $(H_T - H_S)/H_T$, and means and standard deviations across 1000 simulations are shown. The third column shows the percentages of datasets in which significant positive correlations were detected between any pair of alleles; positive correlations can be used as an indication that there is population structure in the dataset. The fourth and fifth columns indicate the percentages of datasets with correct allele assignments, using two methods described in Table 1.

Freq. of homoplasious allele	Mean percentage of genotypes that could be resolved
0.1	87.4%
0.2	73.0%
0.3	57.0%
0.4	43.9%
0.5	40.9%

Table 4: For datasets from Fig. 3 with correct allele assignments, percentages of genotypes that could be unambiguously resolved.

Sampling region	Pop 1	Pop 2
SG-1	3	35
SG-2	16	22
SG-3	39	1
UFR	46	1
NKO	46	4
SL	26	3
FL	7	0

Table 5: Distribution of sampling regions of white sturgeon among two groups of individuals (Pop 1 and Pop 2) determined by principal coordinate analysis (Fig 6). Region codes refer to those used by [Drauch Schreier et al., 2012]. Pop 1 was used for assigning alleles to isoloci.

Marker	Number of alleles	Number of homoplasious alleles: Whole set used for assignment, No swapping	Whole set used for assignment, Swap ≤ 0.5	Pop 1 used for assignment, No swapping	Pop 1 used for assignment, Swap ≤ 0.5	Percent missing data in recoded dataset: Pop 1	Pop 2
AciG110	20	3	1	0	0	0%, 1%	29%, 29%
As015	18	3	1	2	1	58%, 83%	59%, 77%
AciG35	18	2	0	1	0	0%, 1%	23%, 23%
Atr109	25	6	3	4	2	58%, 63%	55%, 52%
Atr117	22	1	1	0	0	0%, 0%	35%, 35%
AciG52	22	4	1	1	0	0%, 1%	24%, 24%
Atr107	24	3	1	2	1	64%, 66%	53%, 53%
Atr1173	18	3	2	3	2	61%, 76%	61%, 86%

Table 6: Assignment of alleles from eight microsatellite markers to two tetrasomic genomes in octoploid white sturgeon (*Acipenser transmontanus*). Alleles were assigned using the K-means + UPGMA method from Table 1, with the exception of Atr117 in Pop 1 due to a fixed allele in that locus and population. Assignments were performed without allele swapping ("No swapping", `rare.al.check = 0` in `testAlGroups`) and with allele swapping ("Swap ≤ 0.5", `rare.al.check = 0.5`). In testing for homoplasy `testAlGroups` was run with the defaults of `tolerance = 0.05` to allow for 5% of genotypes to disagree with allele assignments, and `null.weight=0.5` to allow for the possibility of null alleles. Assignments were performed using the whole dataset of 249 individuals ("whole set") or a subset of 183 individuals based on population structure ("Pop 1", Table 5 and Fig. 6). The assignments from Pop 1 with Swap ≤ 0.5 were then used to split the dataset into isoloci using the `recodeAllopoly` function. Genotypes that could not be unambiguously determined were coded as missing data; percentages of missing data in Pop 1 and Pop 2 are shown.

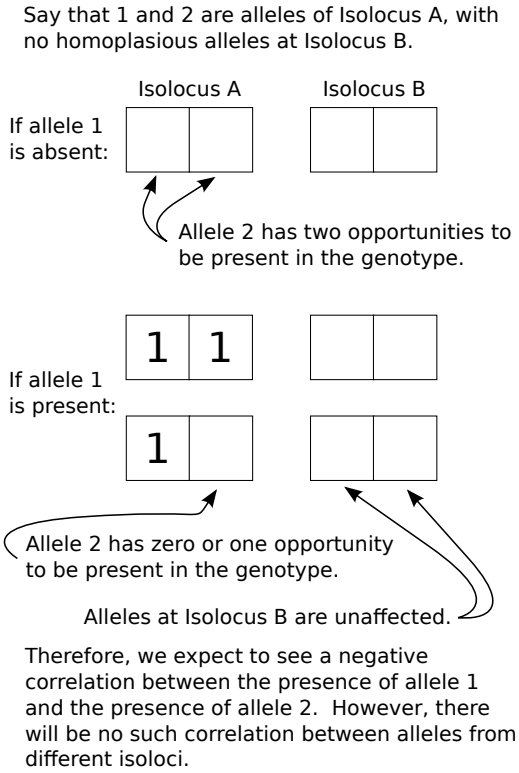

Figure 1: Qualitative reasoning for the expectation of negative correlation between two alleles at the same isolocus.

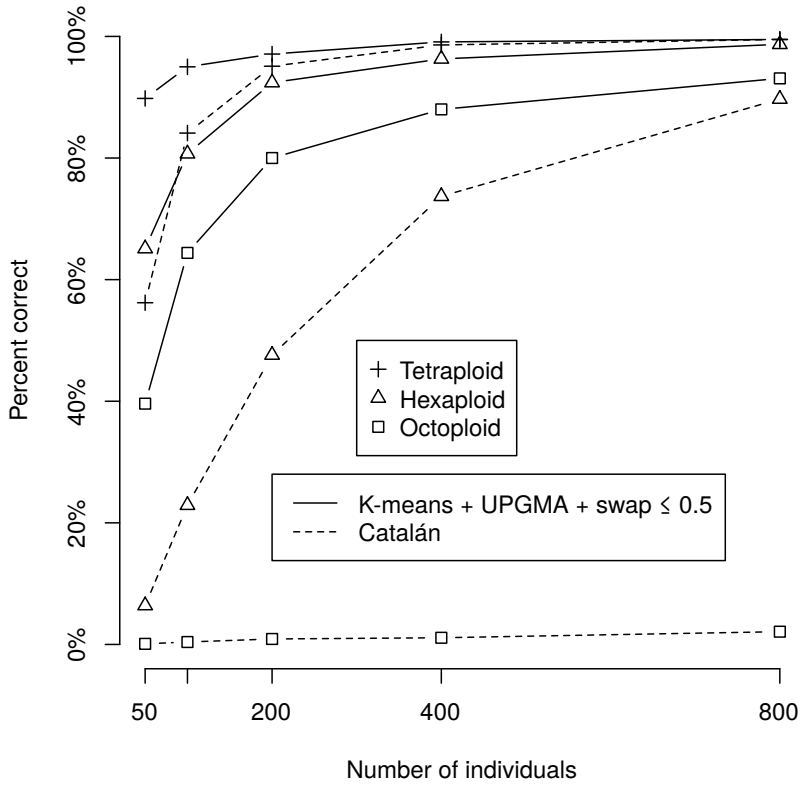

Figure 2: Accuracy of allele assignments with different sample sizes. For each ploidy and sample size, 1000 simulations were performed. Octoploids were simulated with two tetraploid genomes. Simulations and allele assignments were performed as in Table 1.

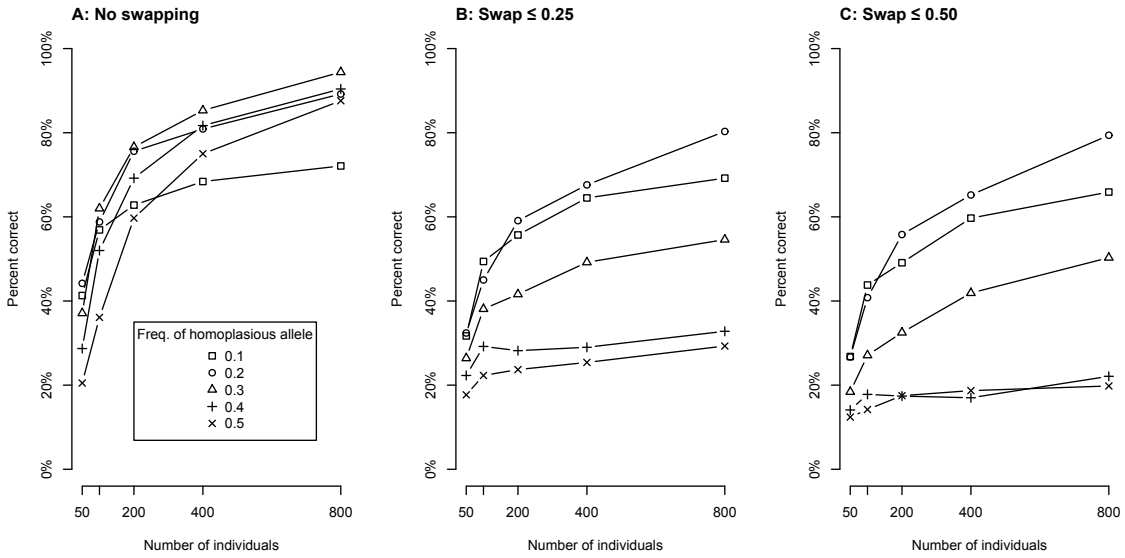

Figure 3: Percentages of simulated datasets with correct allele assignments when homoplasious alleles are present. Allotetraploid datasets were simulated as in Table 1, but in each dataset one pair of homoplasious alleles (alleles from two different isoloci, but with identical amplicon size) was simulated. The frequency of homoplasious alleles was identical at both isoloci in each dataset, and was set at five different levels (0.1 through 0.5). Five different sample sizes were tested (50, 100, 200, 400, and 800). For each homoplasious allele frequency and sample size, 1000 datasets were simulated. Allele assignments were made using three methods from Table 1: K-means + UPGMA (A), K-means + UPGMA + swap ≤ 0.25 (B), or K-means + UPGMA + swap ≤ 0.50 (C); plus an algorithm in the function `testAlGroups` that identifies the alleles most likely to be homoplasious, and assigns alleles as homoplasious until all genotypes are consistent with allele assignments.

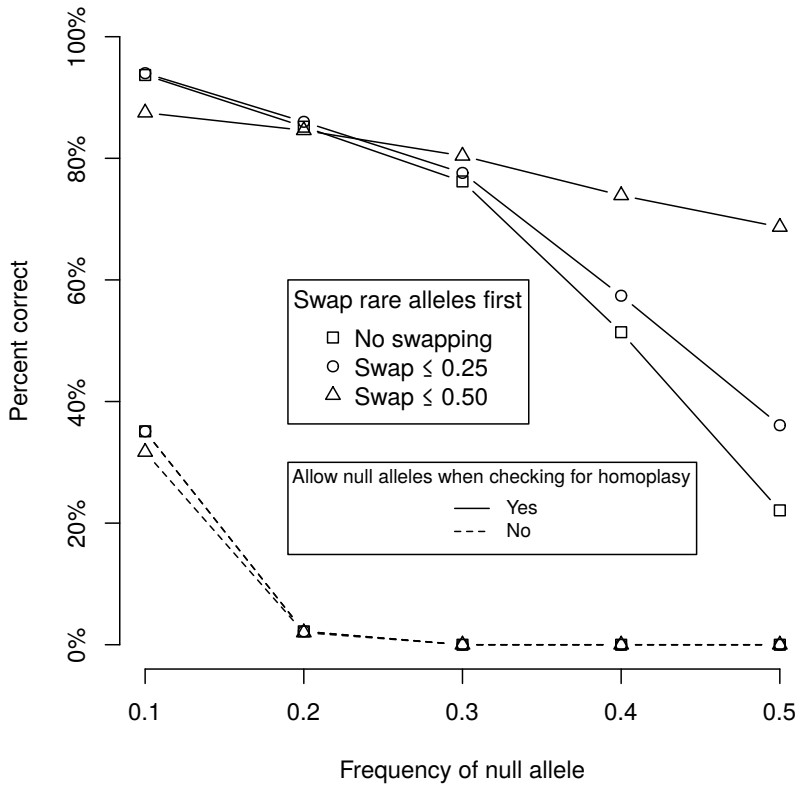

Figure 4: Percentages of simulated datasets with correct allele assignments when one isolocus has a null allele. Allotetraploid datasets were simulated as in Table 1, and frequency of the null allele was set at one of five levels (x-axis). 1000 datasets were simulated at each null allele frequency. Two parameters for `testAlGroups` were adjusted: `rare.al.check` at values of zero, 0.25, and 0.5 (corresponding to the methods K-means + UPGMA, K-means + UPGMA + swap ≤ 0.25, and K-means + UPGMA + swap ≤ 0.50, respectively); and `null.weight` at values of zero (null alleles are allowed when checking for evidence of homoplasy) and 0.5 (genotypes lacking alleles belonging to a given isolocus are taken as evidence that their other alleles are homoplasious).

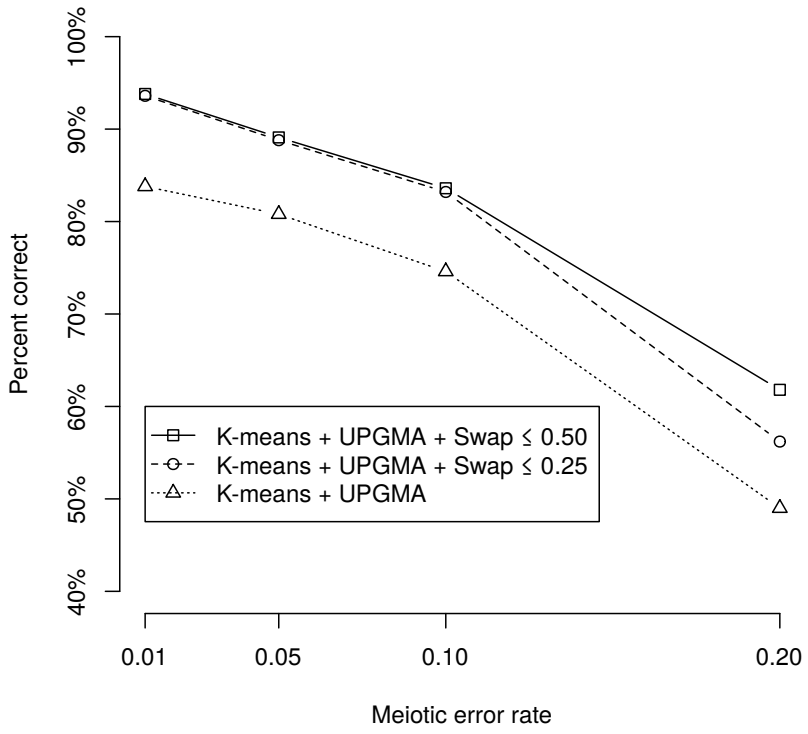

Figure 5: Percentages of simulated datasets with correct allele assignments when meiotic error causes compensated anueoploidy. Meiotic error was simulated in the `simAllopoly` function on a per-gamete basis, with each error causing an allele from one isolocus to be substituted with an allele from the other isolocus. Each dataset was otherwise simulated for an allotetraploid organism with 100 individuals as in Table 1. Meiotic error rate, as shown in the x-axis, was controlled using the `meiotic.error.rate` argument of `simAllopoly`. For each error rate, 1000 datasets were simulated. For the `testAlGroups` function, the `tolerance` argument was set to 1 to prevent the function from checking for homoplasy, and `rare.al.check` was set to zero, 0.25, or 0.5 (corresponding to the methods K-means + UPGMA, K-means + UPGMA + swap ≤ 0.25, and K-means + UPGMA + swap ≤ 0.50, respectively). Each dataset was tested for all three values of `rare.al.check`.

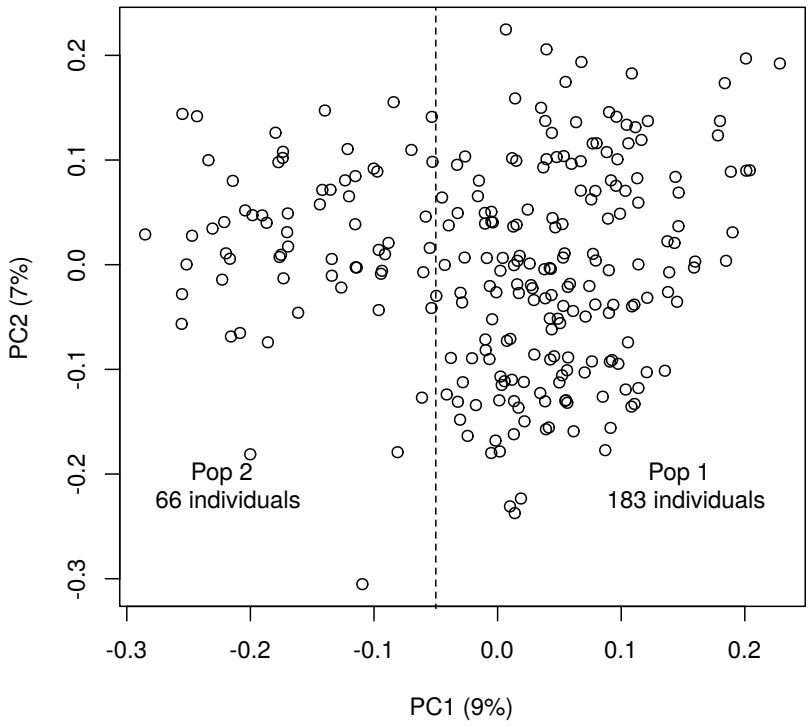

Figure 6: Principal coordinates analysis of 249 white sturgeon individuals, based on genotypes at eight microsatellite loci from Drauch Schreier et al. [2012]. Inter-individual distances were calculated using the `Lynch.distance` function in POLYSAT. Percentages of variation explained by the first two axes are shown. The dashed line indicates the cutoff for dividing the set into two groups (Pop 1 and Pop 2) based on population structure.

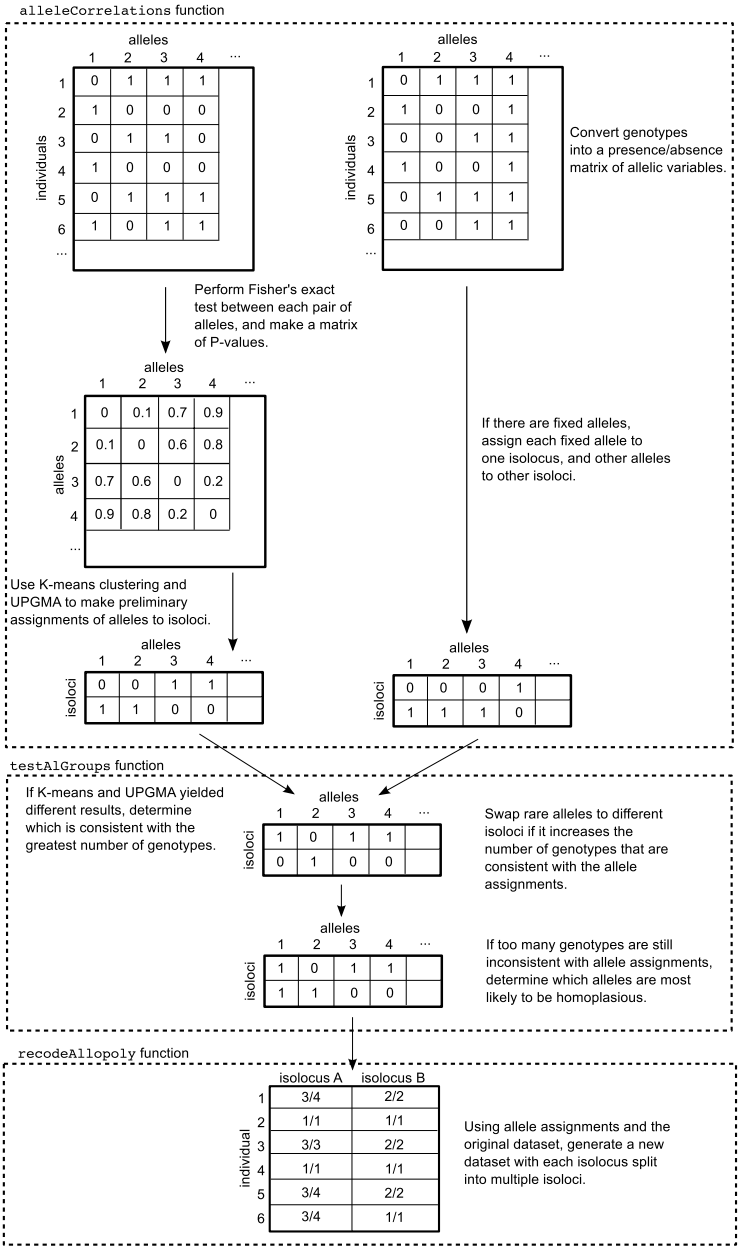

Figure 7: Overview of functions in POLYSAT 1.4 for processing allopolyploid and diploidized autopolyploid datasets.