MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2017) **17**, 1090–1103 doi: 10.1111/1755-0998.12639

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

LINDSAY V. CLARK* and ANDREA DRAUCH SCHREIER†

**Department of Crop Sciences, University of Illinois, Urbana-Champaign, 1201 W. Gregory Drive, Urbana, IL 61801, USA,*
†Department of Animal Science, University of California – Davis, Davis, CA 95616, USA

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 data set from a natural population of octoploid white sturgeon (*Acipenser transmontanus*). Our methodology is implemented in the R package POLYSAT version 1.6.

Keywords: polyploidy, R package, simple sequence repeat, sturgeon

Received 14 November 2015; revision received 6 November 2016; accepted 16 November 2016

Introduction

Polyploidy, both recent and ancient, is pervasive throughout the plant kingdom (Udall & Wendel 2006), and to a lesser extent, the animal kingdom (Gregory & 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 (for example, 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 (for example, AAAB vs. AABB vs. ABBB) (Clark & 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. ‘Disomic inheritance’ refers to situations in which 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 analyse 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

Correspondence: Lindsay V. Clark, Fax: +1-217-333-4582;
E-mail: lvclark@illinois.edu

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 segregation 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, meiotic pairing in allopolyploids may occur between both homologous and homeologous chromosome pairs, but at different rates based on sequence similarity (Obbard *et al.* 2006; Gaeta & Pires 2010), which often differs from locus to locus even within a species (Dufresne *et al.* 2014). Waples (1988) proposed a method for estimating allele frequencies 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 possessing 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 nonoverlapping 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 straightforward 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 can be useful if one parent is heterozygous for a particular allele and the other parent lacks that allele, in which case segregation may follow a 1:1 ratio and can be analysed under the diploid testcross model (Rousseau-Gueutin *et al.* 2008; Swaminathan *et al.* 2012; other ratios are possible, in which case the testcross model does not apply). 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, for example tests that look for departures from or make assumptions of Hardy–Weinberg Equilibrium (Clark & 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 (except in rare cases of multi-allelic SNPs), 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 data set, 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 the 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 data sets or data sets with rare alleles, it is likely that some alleles in the data set 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 data set. Our method is appropriate for natural populations of $\sim$100 individuals or more. It is also appropriate for certain mapping populations, including F_2, recombinant inbred lines, and doubled haploids. It can be used on organisms of any ploidy as long as each subgenome has the same ploidy, for example, octoploid species with four diploid subgenomes or two tetraploid subgenomes, but not two diploid subgenomes and one tetraploid subgenome. Negative correlations between allelic variables 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 analysed as diploid or polysomic markers. Our method 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 data set 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, is implemented in the R package POLYSAT version 1.6.

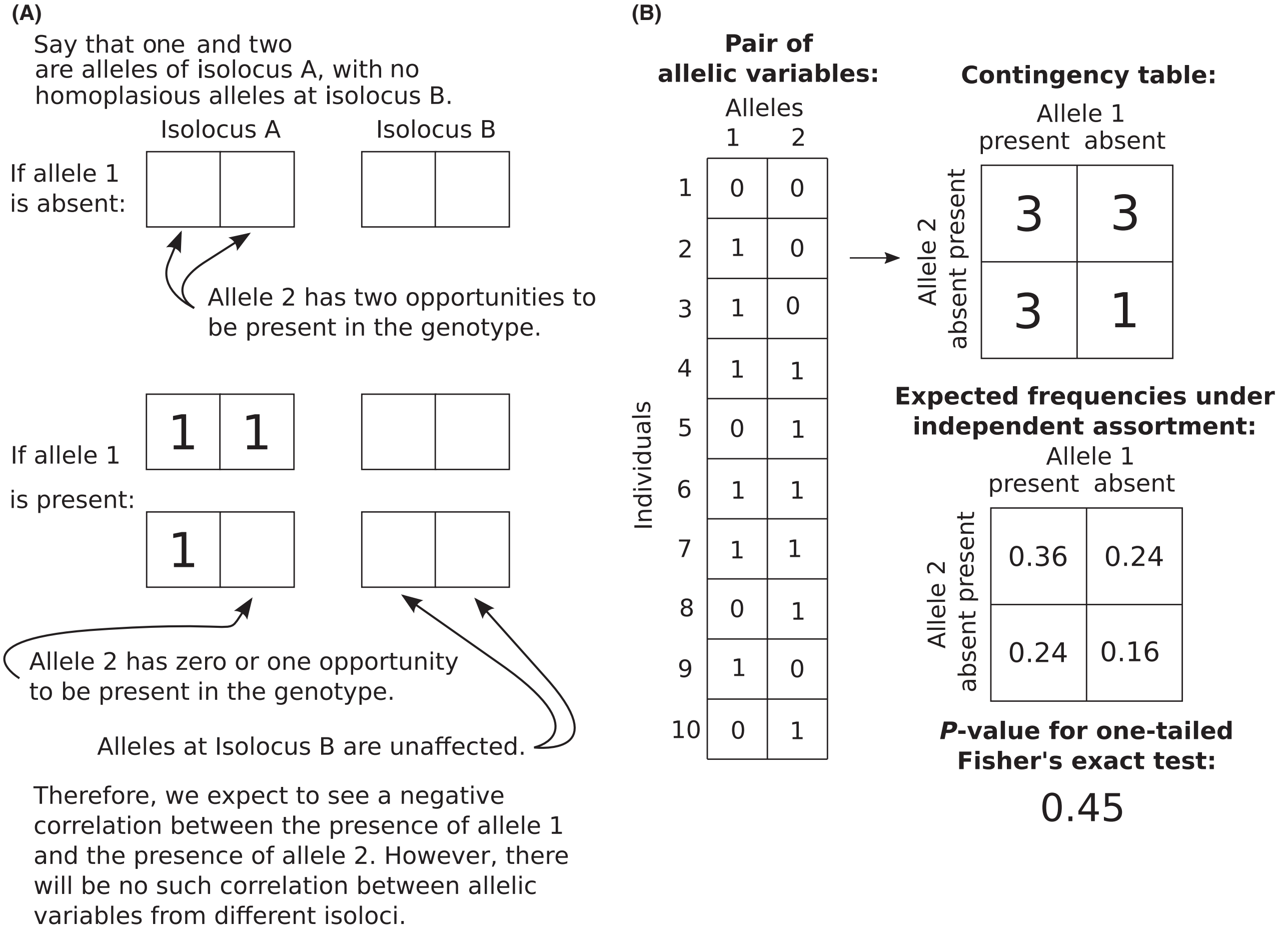
Materials and methods

Rationale

Say that a microsatellite data set 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 independent 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. 1A). The negative correlation can also be proved mathematically (Appendix S1, Supporting information). We use 'correlation' in a broad sense here; 'negative correlation' means that the presence of one allele is associated with the absence of another allele or vice versa.

Algorithm for clustering alleles into isoloci

Preliminary clusters: the `alleleCorrelations` *function.* An overview of our algorithm is presented in



Say that one and two are alleles of isolocus A, with no homoplasious alleles at isolocus B.

Allele 2 has two opportunities to be present in the genotype.

Allele 2 has zero or one opportunity to be present in the genotype.

Alleles at Isolocus B are unaffected.

Therefore, we expect to see a negative correlation between the presence of allele 1 and the presence of allele 2. However, there will be no such correlation between allelic variables from different isoloci.

	1	2
1	0	0
2	1	0
3	1	0
4	1	1
5	0	1
6	1	1
7	1	1
8	0	1
9	1	0
10	0	1

	present	absent
present	3	3
absent	3	1

	present	absent
present	0.36	0.24
absent	0.24	0.16

Fig. 1 Negative correlation between two allelic variables at a locus. (A) Qualitative reasoning for the expectation of negative correlation between two allelic variables at the same isolocus. (B) Use of Fisher's exact test to identify negative correlation between a pair of allelic variables. Ten individuals are shown for the sake of illustration, but an ideal data set would have 100 or more individuals. In the allelic variables, the presence of an allele in an individual is indicated by 1, and the absence is indicated by 0.

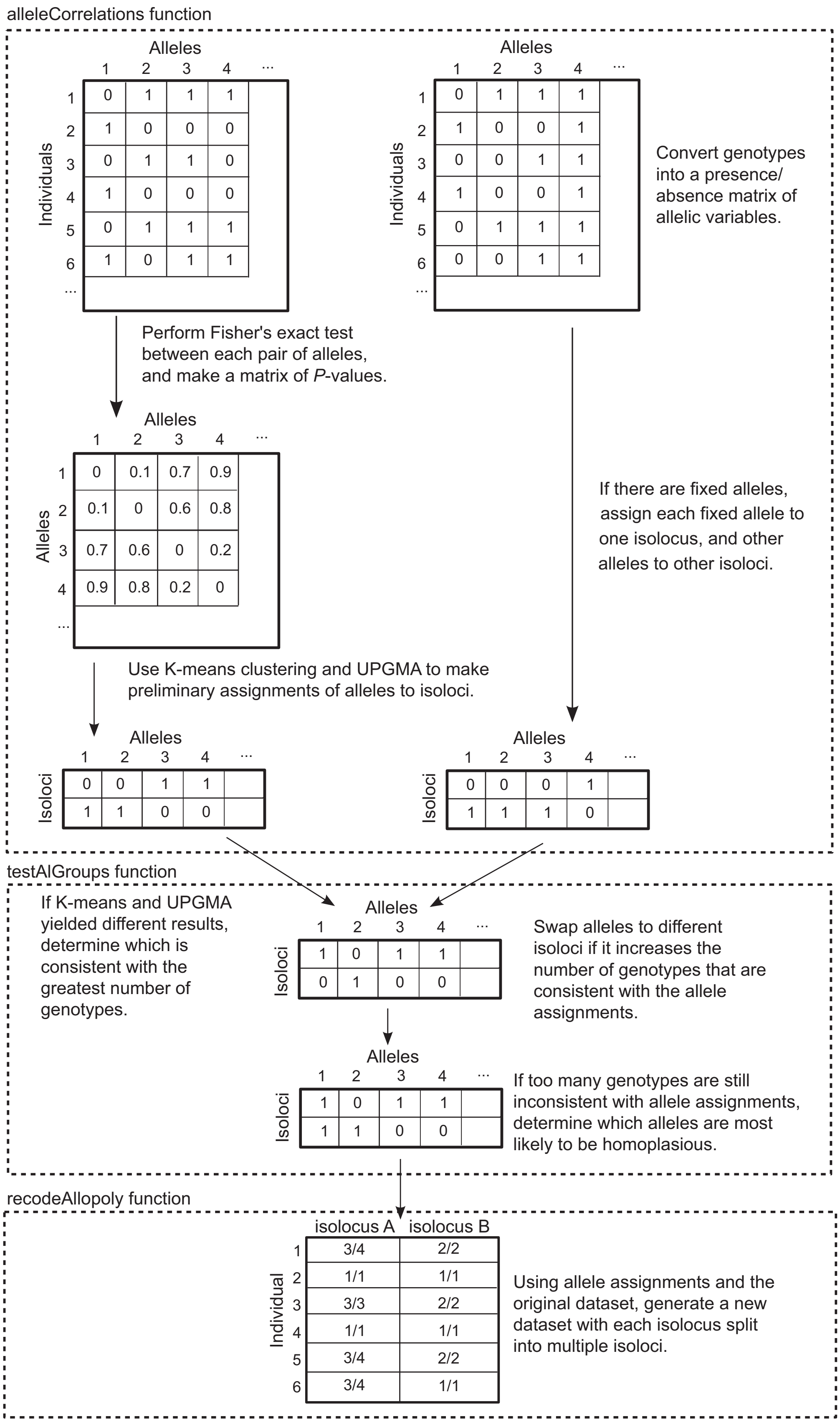
Fig. 2. To test independence of two binary allelic variables, we use Fisher's exact test as it is appropriate for small sample sizes, which are likely to occur in typical population genetics data sets when rare alleles are present. A two-by-two contingency table is generated for the test, with rows indicating the presence or absence of the first allele, columns indicating the presence or absence of the second allele and each cell indicating the number of individuals in that category (Fig. 1B). A one-tailed Fisher's exact test is used, with the alternative hypothesis being that more individuals have just one allele of the pair than would be expected if the allelic variables were independent, that is the alternative hypothesis is that the odds ratio is less than one, indicating a negative association between the presence of the first allele and the presence of the second 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, as the *P*-values are used as a dissimilarity statistic. The function `alleleCorrelations` in POLYSAT 1.6 produces such a matrix of *P*-values for a single microsatellite marker. The same function also produces two sets of preliminary assignments of alleles to isoloci, using UPGMA and the Hartigan & Wong (1979) method of *k*-means clustering, respectively. The `n.subgen` argument is used to specify how many subgenomes the organism has, that is into how many isoloci each locus should be split.

Population structure can also cause correlation between allelic variables, for example if two alleles are both common in one subpopulation and rare in another. Because correlation caused by population structure can potentially obscure the correlations that are used by our method, the `alleleCorrelations` function checks for significant positive correlations (after Holm–Bonferroni multiple testing correction) between allelic variables, which could only be caused by population structure, scoring error (such as stutter peaks being miscalled as alleles, and therefore tending to be present in the same genotypes as their corresponding alleles) or linkage disequilibrium (if two isoloci are part of a tandem duplication on the same chromosome, as opposed to duplication resulting from polyploidy), and prints a warning if such correlations are found.

If one or more alleles are present in all genotypes in a data set, it is not possible to perform Fisher's exact test to look for correlations between those fixed allelic variables 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 (for example, 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 (for example, 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 (that is 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.

Corrected clusters: the `testAlGroups` *function.* Although *k*-means was more accurate overall than UPGMA using simulated data (Table S1, Supporting information), UPGMA sometimes assigned alleles correctly when *k*-means assigned them incorrectly. To choose between *k*-means and UPGMA when they give different results, the function `testAlGroups` in POLYSAT checks every genotype in the data set against both results. Assuming no null alleles or homoplasy (which are dealt with later in the algorithm), 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 (for example, two in an allotetraploid). The ploidy of isoloci is specified using the `SGploidy` argument. 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 S1, Supporting information).

Given that allele assignments were still not always correct when *k*-means and UPGMA were used (Table S1, Supporting information), a second step was added to `testAlGroups` that randomly moves single alleles to different isoloci and tests whether the new set of assignments is consistent with a greater number of genotypes than the old set. We will refer to this as the 'allele swapping algorithm'. The `swap` argument of `testAlGroups` is set to `TRUE` to indicate that the allele swapping algorithm should be used, or `FALSE` to indicate that it should not. The allele swapping algorithm is based on the simulated annealing algorithm of Bertsimas & Tsitsiklis (1993). The cost function used is the



Individuals \ Alleles	1	2	3	4	...
1	0	1	1	1	
2	1	0	0	0	
3	0	1	1	0	
4	1	0	0	0	
5	0	1	1	1	
6	1	0	1	1	
...					

Individuals \ Alleles	1	2	3	4	...
1	0	1	1	1	
2	1	0	0	1	
3	0	0	1	1	
4	1	0	0	1	
5	0	1	1	1	
6	0	0	1	1	
...					

Convert genotypes into a presence/absence matrix of allelic variables.

Perform Fisher's exact test between each pair of alleles, and make a matrix of *P*-values.

Alleles	1	2	3	4	...
1	0	0.1	0.7	0.9	
2	0.1	0	0.6	0.8	
3	0.7	0.6	0	0.2	
4	0.9	0.8	0.2	0	
...					

If there are fixed alleles, assign each fixed allele to one isolocus, and other alleles to other isoloci.

Use K-means clustering and UPGMA to make preliminary assignments of alleles to isoloci.

Isoloci \ Alleles	1	2	3	4	...
	0	0	1	1	
	1	1	0	0	

Isoloci \ Alleles	1	2	3	4	...
	0	0	0	1	
	1	1	1	0	

If K-means and UPGMA yielded different results, determine which is consistent with the greatest number of genotypes.

Isoloci \ Alleles	1	2	3	4	...
	1	0	1	1	
	0	1	0	0	

Swap alleles to different isoloci if it increases the number of genotypes that are consistent with the allele assignments.

Isoloci \ Alleles	1	2	3	4	...
	1	0	1	1	
	1	1	0	0	

If too many genotypes are still inconsistent with allele assignments, determine which alleles are most likely to be homoplasious.

Individual	isolocus A	isolocus B
1	3/4	2/2
2	1/1	1/1
3	3/3	2/2
4	1/1	1/1
5	3/4	2/2
6	3/4	1/1

Using allele assignments and the original dataset, generate a new dataset with each isolocus split into multiple isoloci.

Fig. 2 Overview of functions in POLYSAT 1.6 for processing allopolyploid and diploidized autopolyploid data sets. Additionally, the processDatasetAllo function can be used to automatically run alleleCorrelations and testAlGroups on every locus in a data set. In the box representing the alleleCorrelations function, all alleles belonging to the locus on the left are variable in the data set, so Fisher's exact test is used to find correlations between allelic variables, then *k*-means and UPGMA are used to perform clustering. The locus on the right has one allele (4) that is present in all individuals, making it impossible to assign alleles to isoloci using Fisher's exact test. In the box representing the testAlGroups function, all steps are performed on all loci regardless of whether or not fixed alleles are present.

proportion of genotypes that is inconsistent with a set of allele assignments. If a set of allele assignments has a cost equal to or lower than that of the previous set, it is always retained. If it has a higher cost, it is retained with the probability exp (D/T), where D is the difference in cost between the new set of allele assignments and the old set (always a negative value) and T is the current 'temperature' of the algorithm. By default, the temperature starts at a value of one and is multiplied by 0.95 with each rep, and 100 attempts to move alleles are made in each rep. If an entire rep is performed without any alleles being successfully moved, the algorithm stops. Alternatively, if a set of allele assignments with a cost of zero is found, the algorithm immediately stops without finishing the rep. If the set of allele assignments found by *k*-means or UPGMA has a cost of zero, allele swapping is not performed.

Although our algorithm attempts primarily to sort alleles into nonoverlapping groups, there is always a possibility that different isoloci have some alleles with identical amplicon sizes (homoplasy). Therefore, we introduced a third step 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, to allow for meiotic or scoring error) 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 rechecked and the cycle is repeated until the desired level of agreement between allele assignments and genotypes is met.

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. 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 assign as 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, as it aims to improve overall agreement with the data set.)

Recoding data sets based on allele assignments: the processDatasetAllo *and* recodeAllopoly *functions.* The function processDatasetAllo is a wrapper function that runs alleleCorrelations and testAlGroups in sequence on every marker in the data set. It tests several parameter sets for testAlGroups. If the data set was divided into subpopulations to prevent bias from population structure, allele assignments from the same parameter set are merged across subpopulations using the mergeAlleleAssignments function. processDatasetAllo generates a series of plots to indicate assignment quality, and selects a suggested best parameter set for each locus by first selecting the parameter set that results in the least amount of missing data when the genotypes are recoded, or in the case of a tie the parameter set that results in the fewest homoplasious alleles.

The list of allele assignments (output by processDatasetAllo) and the original data set are then passed to the recodeAllopoly function, which produces a new data set in which each marker is split into multiple isoloci. Missing data are substituted for genotypes that cannot be resolved due to homoplasy in the allele assignments. (For example, if alleles A and B belong to different isoloci, and C belongs to both, the allelic phenotype ABC could be genotype AA BC, AC BB or AC BC, assuming no null alleles.) An argument called allowAneuploidy lets the user specify whether to allow for apparent meiotic error. If allowAneuploidy = TRUE, 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.) Otherwise, missing data are inserted where there are too many alleles per isolocus.

Implementation of the Catalán method: the catalanAlleles *function.* POLYSAT 1.6 also includes an implementation of the algorithm of Catalán *et al.* (2006). One difference between our implementation and the original is that we allow ploidies higher than tetraploid, for example in a hexaploid, a genotype with three alleles is assumed to be fully homozygous. Additionally, after fully homozygous genotypes are examined, fully heterozygous genotypes are also examined if necessary for assigning alleles that were not present in any fully homozygous genotypes. The output of catalanAlleles can be passed directly to recodeAllopoly.

Simulated data sets

The function simAllopoly was added to POLYSAT to generate simulated data sets for testing the accuracy of allele assignment methods. It simulates one locus at a time and allows for adjustment of the number of isoloci, the ploidy of each isolocus, the number of alleles for each isolocus, the number of alleles that are homoplasious between isoloci, the number of null alleles (producing no amplicon), allele frequencies in the population, the meiotic error rate (frequency at which different isoloci pair with each other at meiosis) and the number of individual genotypes to output. By default, alleles from the first isolocus are labelled A1, A2, etc., alleles from the second isolocus labelled B1, B2, etc. and homoplasious alleles labelled H1, H2, etc.

For initial evaluation of clustering methods (Table S1, Supporting information), 10 000 simulated markers were generated for 100 individuals each for allotetraploid, allohexaploid and allo-octoploid (two tetrasomic isoloci) species under Hardy–Weinberg equilibrium. Although not included in the simulated data sets, note that it is also possible for an octoploid to possess four diploid subgenomes, as in strawberry. Each isolocus had a randomly chosen number of alleles between two and eight, and allele frequencies were generated randomly. A set of allele assignments for one marker was considered to be correct if no alleles were assigned incorrectly.

To evaluate the effect of sample size on assignment accuracy, 1000 additional markers were simulated for populations of 50, 100, 200, 400 and 800 individuals for allotetraploid, allohexaploid and allo-octoploid species.

To simulate population structure, 5000 simulated markers were generated for two populations of 50 allotetraploid individuals. Allele frequencies differed by five fixed amounts (Table 1) between the two populations, with 1000 markers simulated for each amount.

The effect of homoplasy on allele assignment methods was evaluated by simulating 1000 allotetraploid markers each for sample sizes of 50, 100, 200, 400 and 800, and homoplasious allele frequencies of 0.1, 0.2, 0.3, 0.4 and 0.5. For each correct set of allele assignments that was found, recodeAllopoly was run on the data set

Table 1 Percentages of simulated data sets with correct allele assignments under different levels of population structure. Two populations of 50 allotetraploid individuals were simulated under different allele frequencies and then merged into one data set 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 $(H_T - H_S)/H_T$, where H_S is the expected heterozygosity in each subpopulation, averaged across the two subpopulations, and H_T is the expected heterozygosity if the two subpopulations were combined into one population with random mating. Means and standard deviations across 1000 simulations are shown for F_{ST}. The third column shows the percentages of data sets 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 data set. The fourth, fifth and sixth columns indicate the percentages of data sets with correct allele assignments, using our methods and that of Catalán *et al.* (2006). 95% confidence intervals are given for percentages

Difference in allele frequency	F_{ST}	Significant positive correlations	*k*-means + UPGMA	*k*-means + UPGMA + swap	Catalán
0.0	0.000 ± 0.000	1% ± 1%	87% ± 2%	96% ± 1%	83% ± 2%
0.1	0.016 ± 0.004	2% ± 1%	97% ± 1%	100% ± 0%	89% ± 2%
0.2	0.062 ± 0.013	21% ± 3%	82% ± 2%	100% ± 0%	92% ± 2%
0.3	0.118 ± 0.022	63% ± 3%	82% ± 2%	100% ± 0%	98% ± 1%
0.4	0.176 ± 0.026	79% ± 3%	81% ± 2%	100% ± 0%	100% ± 0%

using `allowAneuploidy` = `FALSE` to determine which genotypes were resolvable.

To evaluate allele assignment when null alleles were present, 5000 markers were simulated for 100 allotetraploid individuals, with 1000 simulated markers at each null allele frequency of 0.1, 0.2, 0.3, 0.4 and 0.5.

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 too 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 nonaneuploid, but one isolocus will have more copies than expected and the other isolocus will have fewer (for example, three copies of one isolocus and one copy of the other isolocus in an allotetraploid). To evaluate the accuracy of allele assignment for isoloci that occasionally pair at meiosis, 4000 markers were simulated for 100 allotetraploid individuals, with 1000 simulated markers at each meiotic error rate of 0.01, 0.05, 0.10 and 0.20.

A custom script was written to simulate genotypes in allopolyploid mapping populations. Allotetraploid, allohexaploid and allo-octoploid (with two tetrasomic subgenomes) populations were simulated, with 200 individuals in each population. For each ploidy, 1000 loci were simulated for each generation spanning F_2 to F_8, assuming completely homozygous parents. Allele assignments were performed with the `alleleCorrelations` and `testAlGroups` functions, with `null.weight` = `1` and `swap` = `TRUE`.

Empirical data set

To demonstrate the usefulness of our allele assignment method on a real data set, 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 & May, 2002; Drauch Schreier *et al.*, 2011). We selected for allele assignment 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 performed a preliminary analysis of population structure using the `Lynch.distance` dissimilarity statistic in POLYSAT and principal coordinates analysis (PCoA) using the `cmdscale` function in R. Thirteen microsatellite markers were used for PCoA, including the eight used for allele assignment and five tetrasomic (present in four copies rather than eight) markers. Allele assignment methods were then tested on the whole data set and on a subpopulation identified by PCoA.

The `testAlGroups` function was run on the sturgeon data set 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`), 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.

To evaluate the accuracy and usefulness of allele assignments, we compared G_{ST} (Nei & Chesser 1983) estimates using the five tetrasomic loci to estimates using the putatively tetrasomic recoded isoloci. Pairs of isoloci were excluded from G_{ST} estimates if they had any homoplasious alleles. Allele frequencies for tetrasomic loci and isoloci were estimated using the method of De Silva *et al.* (2005) using the `deSilvaFreq` function in POLYSAT with the selfing rate set to 0.0001. Pairwise G_{ST} between sampling regions was then estimated with the `calcPopDiff` function in POLYSAT.

Results

Simulated natural populations

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. 3). 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. For octoploids, the success of the Catalán *et al.* (2006) method was near zero even with 800 individuals in the data set (due to the low probability of producing fully homozygous genotypes at tetrasomic isoloci), whereas our method had an accuracy of 93% with 800 octoploid individuals.

Both negative and positive correlations between allelic variables at different loci can occur when the assumption of random mating is violated by population structure, confounding the use of negative correlations for assigning alleles to isoloci. Although population structure negatively impacted our method when the allele swapping algorithm was not used, accuracy was at or near 100% with the allele swapping algorithm (Table 1). Interestingly, low levels of population

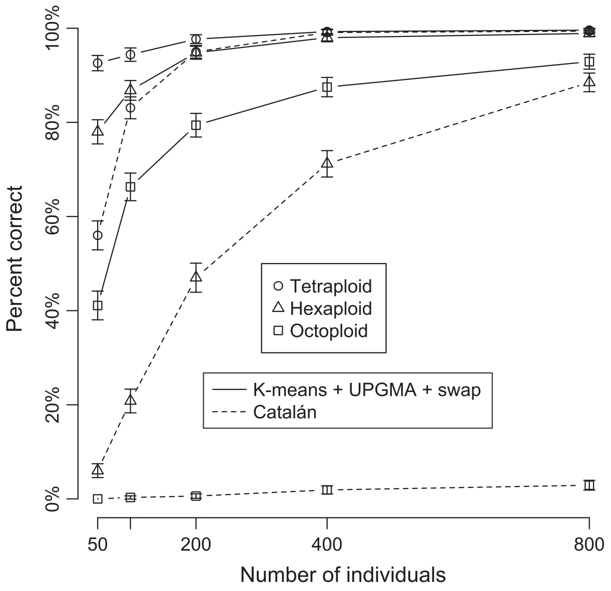


Fig. 3 Accuracy of allele assignments with different sample sizes. For each ploidy and sample size, 1000 simulations were performed. Octoploids were simulated with two tetraploid genomes. Whiskers indicate 95% confidence intervals. 'Swap' indicates that `testAlGroups` was used with `swap = TRUE`. The *y*-axis indicates the percentage of data sets for which allele assignments were completely correct.

structure ($F_{ST} \approx 0.02$) improved the accuracy of our method compared to $F_{ST} = 0$ (Table 1), probably as a result of an increase in the number of double homozygous genotypes. For this same reason, the Catalán *et al.* (2006) method had an improved success rate as population structure increased, but was not as accurate as our method except at $F_{ST} \approx 0.2$ (Table 1). In our simulations, significant positive correlations between allelic variables were found in most data sets that had moderate population structure (Table 1).

One advantage of our method over that of Catalán *et al.* (2006) is that our method allows for alleles belonging to different isoloci to have identical amplicon sizes (homoplasy). We tested the accuracy of allele assignments across several sample sizes and frequencies of homoplasious alleles, with and without the allele swapping algorithm (Fig. 4). 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. When allele assignments were correct, we tested the mean proportion of genotypes that were resolvable, given several frequencies of a homoplasious allele (Table 2). Although accuracy of assignment had been highest with a homoplasious allele frequency of 0.3, only 59% of genotypes could be resolved in such data sets (Table 2).

To test the effect of null alleles on the accuracy of our allele assignment method, we simulated data sets in which one isolocus had a null allele (Fig. 5). 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. We also found that the allele swapping algorithm improved the accuracy of allele assignments when the null allele was at a frequency of 0.1 in the population. However, at higher null allele frequencies (≥0.2), allele assignments were more accurate without allele swapping.

We simulated data sets in which gametes resulting in compensated aneuploidy (meiotic error) occured at a range of frequencies from 0.01 to 0.2 (Fig. 6). At all meiotic error rates, the allele swapping algorithm from `testAlGroups` improved the accuracy of allele assignment (Fig. 6). 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 data sets of 100 individuals with no homoplasy, null alleles or population structure (Fig. 6).

We also examined the effect of number of alleles on the accuracy of our method. Accuracy was highest when the number of alleles was similar among isoloci (Table S2, Supporting information).

Assignment of alleles to isoloci in octoploid sturgeon

When using principal coordinates analysis to test for genetic structure prior to performing allele assignment, we identified two major genetic groups (Table S3, Fig. S1, Supporting information) that were similar to the population structure previously observed (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 analysed by itself (data not shown). We therefore tested our method on Pop 1 (183 individuals) and on the combined set of 249 individuals.

For seven 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 3). Regardless of whether the whole data set or only Pop 1 was used for making allele assignments, eliminating the allele swapping algorithm caused a large increase in the number of loci with apparent homoplasy (Table 3). For the six loci that lacked homoplasy when the whole data set was used for making assignments, nearly all genotypes could be resolved unambiguously when using those assignments (Table 3). Assignments made using Pop 1 performed somewhat more poorly in Pop 2, resulting in 15–40% missing data

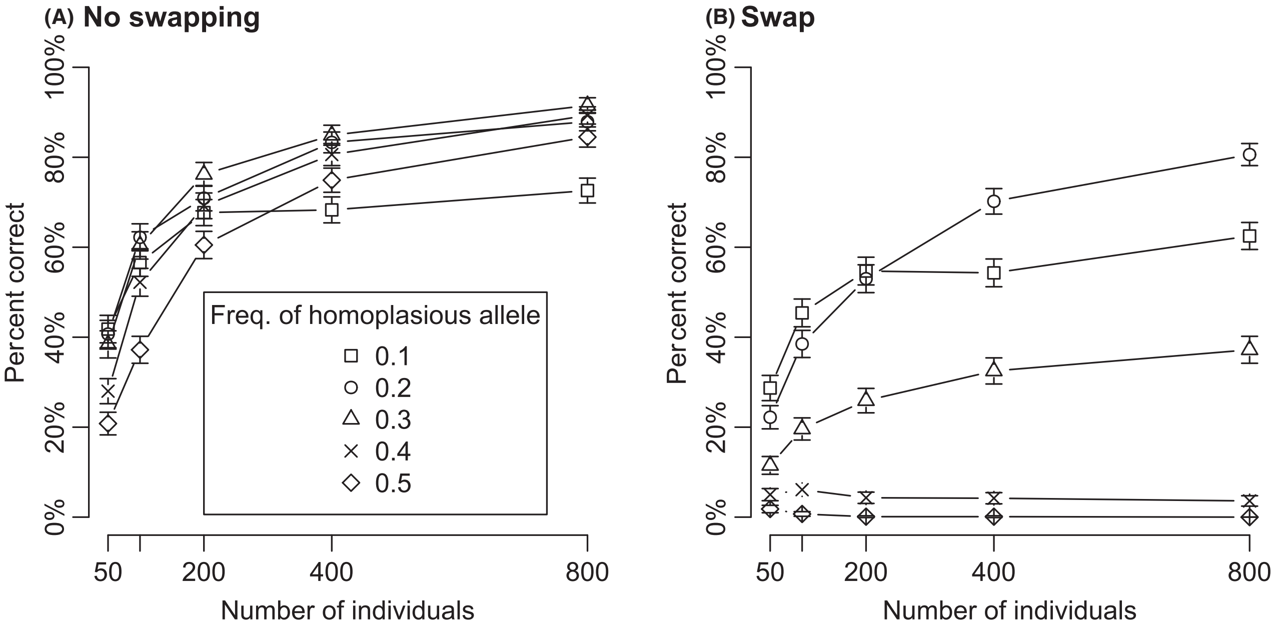


Fig. 4 Percentages of simulated data sets with correct allele assignments when homoplasious alleles are present. Whiskers indicate 95% confidence intervals. The *y*-axis indicates the percentage of data sets for which allele assignments were completely correct. Allotetraploid data sets were simulated with one pair of homoplasious alleles (alleles from two different isoloci, but with identical amplicon size) for each locus. The frequency of homoplasious alleles was identical at both isoloci in each data set 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 data sets were simulated. Allele assignments were made using two methods: *k*-means + UPGMA (A; `swap = FALSE`) and *k*-means + UPGMA + swap (B; `swap = TRUE`); 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.

Table 2 For data sets from Fig. 4 with correct allele assignments at `swap = FALSE` (no swapping), percentages of genotypes that could be unambiguously resolved. Means and standard deviations are shown

Freq. of homoplasious allele	Mean percentage of genotypes that could be resolved
0.1	85.4% ± 5.8%
0.2	71.2% ± 8.2%
0.3	59.1% ± 9.5%
0.4	51.8% ± 8.8%
0.5	48.5% ± 7.1%

(Table 3). Given these results, we retained for further analysis allele assignments for AciG110, As015, AciG35, Atr117, AciG52 and Atr107 that were made using the whole data set with allele swapping, and allele assignments for Atr1173 that were made using Pop 1 with allele swapping.

By recoding allo-octoploid markers as tetrasomic isoloci, we were able to estimate allele frequencies, which would not have been possible otherwise. We were then able to use allele frequencies to estimate pairwise G_{ST} between white sturgeon sampling regions. G_{ST} estimates using recoded isoloci were very similar to estimates obtained using known tetrasomic microsatellite markers (Fig. S2, Supporting information), suggesting that allele assignments were accurate. The two isoloci recoded from Atr1173 yielded G_{ST} estimates similar to those from other loci (Fig. S2, Supporting information), despite high missing data rates in Pop 2 (Table 3) suggesting any bias in allele frequencies caused by the missing data was negligible.

Simulated mapping populations

Negative correlations between allelic variables at the same isolocus can also occur in certain types of mapping populations, enabling the use of our algorithm to assign alleles to isoloci in these populations. There are several requirements that must be met however. (i) To prevent correlations between unlinked allelic variables, all individuals in the population must be equally related to each other. Pedigrees, nested association mapping (NAM) populations and multiple-cross mating designs are therefore not appropriate. (ii) No allele should be present in all individuals in the population. Our method therefore cannot be used on backcross or near isogenic line (NIL) populations, which are expected to segregate only AB and BB genotypes. (iii) All alleles belonging to one

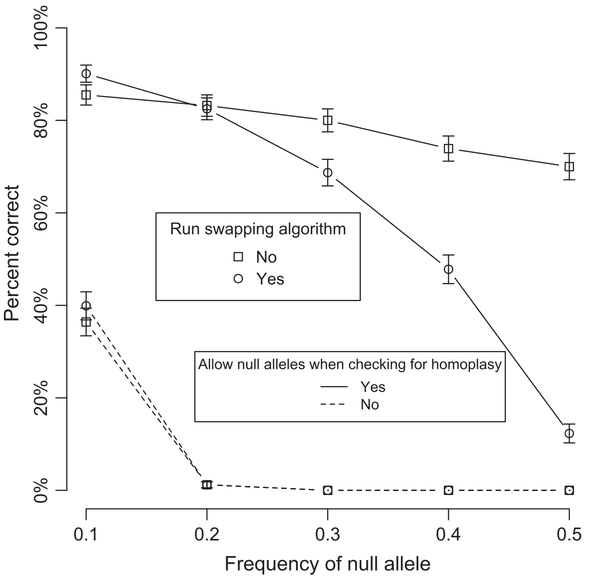


Fig. 5 Percentages of simulated data sets with correct allele assignments when one isolocus has a null allele. Whiskers indicate 95% confidence intervals. The *y*-axis indicates the percentage of data sets for which allele assignments were completely correct. Allotetraploid data sets were simulated, and frequency of the null allele was set at one of five levels (*x*-axis). A total of 1000 data sets were simulated at each null allele frequency. Two parameters for `testAlGroups` were adjusted: `swap` at values of false or true (corresponding to the methods *k*-means + UPGMA and *k*-means + UPGMA + swap, 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).

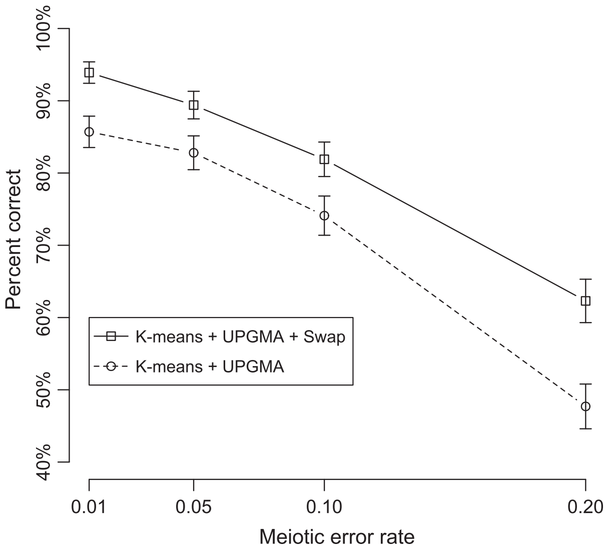


Fig. 6 Percentages of simulated data sets with correct allele assignments when meiotic error causes compensated aneuploidy. Whiskers indicate 95% confidence intervals. The *y*-axis indicates the percentage of data sets for which allele assignments were completely correct. Meiotic error was simulated 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 data set was otherwise simulated for an allotetraploid organism with 100 individuals. Meiotic error rate, as shown in the *x*-axis, was controlled using the `meiotic.error.rate` argument of `simAllopoly`. For each error rate, 1000 data sets were simulated. For the `testAlGroups` function, the `tolerance` argument was set to 1 to prevent the function from checking for homoplasy, and `swap` was set to false or true (corresponding to the methods *k*-means + UPGMA and *k*-means + UPGMA + swap, respectively). Each data set was tested for both values of `swap`.

isolocus should have had the opportunity to pair with each other at meiosis. This eliminates F_1 populations, where an individual with genotype AB might be crossed to an individual with genotype CD. However, allele assignments in F_2 populations, as well as related populations such as recombinant inbred line (RIL) and doubled haploid (DH), can be performed with very high accuracy using our algorithm.

Accuracy of allele assignment was 100% for allotetraploids and allohexaploids for all population types tested (F_2 to F_8; Table 4). Due to the highly heterozygous nature of tetrasomic loci, accuracy was 46% for allo-octoploids in the F_2 generation. However, accuracy for allo-octoploids increased to 98% in the F_3 and 100% in F_4 and higher populations, due to increased homozygosity from selfing.

Discussion

Here, we introduce the R package POLYSAT version 1.6, with several new functions applicable to the analysis of allopolyploids and diploidized autopolyploids. These include `simAllopoly`, which generates simulated data sets; `catalanAlleles`, which uses the Catalán *et al.* (2006) method to assign alleles to isoloci; `alleleCorrelations`, which performs Fisher's exact test between each pair of allelic variables 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; `processDatasetAllo`, which runs `alleleCorrelations`, `testAlGroups` (with multiple parameter sets) and `mergeAlleleAssignments` on an entire data set; and `recodeAllopoly`, which uses allele assignments to

Table 3 Assignment of alleles from eight microsatellite markers to two tetrasomic genomes in octoploid white sturgeon (*Acipenser transmontanus*). Alleles were assigned using negative correlations, 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', swap = FALSE intestAlGroups) and with allele swapping ('Swap', swap = TRUE). 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 data set of 249 individuals ('whole set') or a subset of 183 individuals based on population structure ('Pop 1', Table S3 and Fig. S1, Supporting information). The assignments from Pop 1 with Swap were then used to split the data set into isoloci using the recodeAllopoly function. Genotypes that could not be unambiguously determined were coded as missing data; percentages of missing data in each of two isoloci in Pop 1 and Pop 2 are shown

		Number of homoplasious alleles				Per cent missing data			
		Whole set used for assignment		Pop 1 used for assignment		Whole set used for assignment		Pop 1 used for assignment	
Marker	Number of alleles	No swapping	Swap	No swapping	Swap	Pop 1	Pop 2	Pop 1	Pop 2
AciG110	20	3	0	0	0	0%, 1%	0%, 0%	0%, 1%	24%, 26%
As015	18	3	0	2	0	0%, 0%	0%, 0%	0%, 0%	15%, 17%
AciG35	18	2	0	1	0	1%, 0%	0%, 0%	0%, 0%	23%, 21%
Atr109	25	6	4	4	3	75%, 77%	70%, 77%	79%, 67%	67%, 67%
Atr117	22	1	0	0	0	0%, 0%	0%, 0%	0%, 0%	41%,42%
AciG52	22	4	0	0	0	0%, 1%	0%, 0%	0%,1%	33%, 33%
Atr107	24	3	0	0	0	1%, 1%	2%, 2%	1%, 1%	38%, 40%
Atr1173	18	3	1	3	0	49%, 55%	36%, 50%	3%, 2%	39%, 44%

Table 4 Accuracy of allele assignment in mapping populations. Percentages of data sets with accurate allele assignments are shown. 95% confidence intervals are indicated. A total of 1000 loci were simulated, each with 200 individuals

Generation	Allotetraploid	Allohexaploid	Allo-octoploid
F_2	100% ± 0%	100% ± 0%	45.9% ± 3.1%
F_3	100% ± 0%	100% ± 0%	97.5% ± 1.0%
F_4	100% ± 0%	100% ± 0%	100% ± 0%
F_5	100% ± 0%	100% ± 0%	100% ± 0%
F_6	100% ± 0%	100% ± 0%	100% ± 0%
F_7	100% ± 0%	100% ± 0%	100% ± 0%
F_8	100% ± 0%	100% ± 0%	100% ± 0%

recode the data set, splitting each microsatellite marker into multiple isoloci. An overview of the data analysis workflow is given in Fig. 2. Previous versions of POLYSAT (1.3 and earlier) were restricted in that estimation of allele frequency and certain interindividual distance metrics could only be performed on autopolyploids. With the ability to assign alleles to isoloci, these parameters may now be estimated for allopolyploids as well.

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 data set. This suggests, as most users will not know whether their data set 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 to maximize the number of genotypes that can be resolved by recodeAllopoly. Although the true solution may involve more homoplasy than the minimum amount identified by our method, when the solution is uncertain, it is practical to choose the solution that will be of the greatest use, that is the solution that allows the greatest number of genotypes to be resolved. 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 allele assignment algorithm was negatively impacted by meiotic error (pairing of nonhomologous chromosomes during meiosis), its accuracy remained fairly high. Assuming correct allele assignments in a population with meiotic error, recodeAllopoly is able to identify some but not all individuals with meiotic error; for example, if alleles A, B and C belonged to one isolocus and D to another, an ABC D individual would be correctly recoded, where as an ABB D individual would be incorrectly recoded as AB DD. Otherwise, recodeAllopoly should give 100% accurate results if allele assignments are correct. Our method is therefore superior to that of Catalán *et al.* (2006), which fails completely even with low frequencies of null alleles, homoplasy or meiotic error.

When discussing homoplasy with respect to our algorithm, we have referred exclusively to homoplasy between alleles belonging to different isoloci. It is important to note that 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. Additionally, when discussing null alleles, we have assumed that non-null alleles still exist for all isoloci. 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, for example 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 an analysis accidentally, they can be identified by weak *k*-means/UPGMA clustering of alleles (which can be evaluated from the graphical output of `processDatasetAllo`) and by a high proportion of alleles appearing to be homoplasious.

Using a real microsatellite data set from natural populations of white sturgeon, we found that our method was useful for recoding most of the markers into two independently segregating isoloci each. Given that white sturgeon are octoploid with two tetrasomic subgenomes (Drauch Schreier *et al.*, 2011), we expected this data set to be problematic; having tetrasomic isoloci as opposed to disomic isoloci would reduce the magnitude of the negative correlations between allelic variables, and was observed in simulations to reduce the accuracy of assignment using our method, although not nearly as severely as the reduction in efficacy of the Catalán *et al.* (2006) method (Table S1, Supporting information, Fig. 3). In population genetic studies, we expect that microsatellite markers that can be recoded using our method could be used for analyses requiring polysomic or disomic inheritance [for example, estimation of allele frequency and population differentiation (Fig. S2, Supporting information), structure (Falush *et al.* 2007) 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 moderately useful for recoding genotypes in a population that was not used for making the assignments. Despite the introduction of missing data into Pop 2 when its genotypes at Atr1173 were recoded, G_{ST} estimates were similar to those obtained from nonrecoded tetrasomic microsatellites in the same populations (Fig. S2, Supporting information). We do however recommend caution when interpreting results from loci where our method has introduced missing data for a large portion of individuals. Such results can be confirmed by comparison to results from loci with little or no missing data.

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 be 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.

Obtaining POLYSAT *1.6*

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. Source code is available at https://github.com/lvclark/polysat/ under a GNU GPL-2 licence.

Acknowledgements

Author LC was supported by the DOE Office of Science, Office of Biological and Environmental Research (grant number DE-SC0012379). We thank Subject Editor Frederic Austerlitz and five anonymous reviewers for feedback on earlier versions of this manuscript.

References

Bertsimas D, Tsitsiklis J (1993) Simulated annealing. *Statistical Science*, **8**, 10–15.

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

Chester M, Riley RK, Soltis PS, Soltis DE (2015) Patterns of chromosomal variation in natural populations of the neoallotetraploid *Tragopogon mirus* (Asteraceae). *Heredity*, **114**, 309–317.

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

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

Drauch Schreier A, Gille D, Mahardja B, May B (2011) Neutral markers confirm the octoploid origin and reveal spontaneous autopolyploidy in white sturgeon, *Acipenser transmontanus*. *Journal of Applied Ichthyology*, **27**, 24–33.

Drauch Schreier A, Mahardja B, May B (2012) Hierarchical patterns of population structure in the endangered Fraser river white sturgeon (*Acipenser transmontanus*) and implications for conservation. *Canadian Journal of Fisheries and Aquatic Sciences*, **69**, 1968–1980.

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. *Molecular Ecology*, **23**, 40–69.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Gaeta RT, Pires JC (2010) Homeologous recombination in allopolyploids: the polyploid ratchet. *New Phytologist*, **186**, 18–28.

Gregory TR, Mable BK (2005) Polyploidy in animals. In: *The Evolution of the Genome* (ed. Gregory TR), Chapter 8, pp. 427–517. Elsevier, San Diego.

Hartigan JA, Wong MA (1979) A K-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **28**, 100–108.

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. *Annals of Human Genetics*, **47**, 253–259.

Obbard DJ, Harris SA, Pannell JR (2006) Simple allelic-phenotype diversity and differentiation statistics for allopolyploids. *Heredity*, **97**, 296–303.

Rodzen JA, May B (2002) Inheritance of microsatellite loci in white sturgeon (*Acipenser transmontanus*). *Genome*, **45**, 1064–1076.

Rousseau-Gueutin M, Lerceteau-Köhler E, Barrot L *et al.* (2008) 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**, 2045–2060.

Swaminathan K, Chae WB, Mitros T *et al.* (2012) A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. *BMC Genomics*, **13**, 142.

Udall JA, Wendel JF (2006) Polyploidy and crop improvement. *Crop Science*, **46**, S3–S14.

Waples RS (1988) Estimation of allele frequencies at isoloci. *Genetics*, **118**, 371–384.

L.C. wrote and tested the software and drafted the manuscript. A.S. provided the white sturgeon data and gave critical feedback on the manuscript.

Data accessibility

POLYSAT is available from CRAN (http://cran.r-project.org). All data sets and scripts used in this manuscript are provided as Supporting Information.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Supplementary materials and methods, tables, and figures.

Appendix S2 Tutorial for creating and using allele assignments in POLYSAT.

Appendix S3 and S4 R scripts for reproducing the analyses in this manuscript.

Appendix S5 White sturgeon microsatellite data set used in strugeontest.R.