Heredity (2017) 119, 389–401

www.nature.com/hdy

ORIGINAL ARTICLE

Population assignment in autopolyploids

DL Field[1], LM Broadhurst[2], CP Elliott[3] and AG Young[2]

Understanding the patterns of contemporary gene dispersal within and among populations is of critical importance to population genetics and in managing populations for conservation. In contrast to diploids, there are few studies of gene dispersal in autopolyploids, in part due to complex polysomic inheritance and genotype ambiguity. Here we develop a novel approach for population assignment for codominant markers for autotetraploids and autohexaploids. This method accounts for polysomic inheritance, unreduced gametes and unknown allele dosage. It can also utilise information regarding the origin and genotype of one parent for population assignment of maternal or paternal parents. Using simulations, we demonstrate that our approach achieves high levels of accuracy for assignment even when population divergence is low ($F_{ST} \sim 0.06$) and with only 12 microsatellite loci. We also show that substantially higher accuracy is achieved when known maternal information is utilised, regardless of whether allele dosage is known. Although this novel method exhibited near identical levels of accuracy to *Structure* when population divergence was high, it performed substantially better for most parameters at moderate ($F_{ST} = 0.06$) to low levels of divergence ($F_{ST} = 0.03$). These methods fill an important gap in the toolset for autopolyploids and pave the way for investigating contemporary gene dispersal in a widespread group of organisms.
Heredity (2017) **119,** 389–401; doi:10.1038/hdy.2017.51; published online 4 October 2017

INTRODUCTION

Understanding contemporary patterns of dispersal are of crucial importance in evolution, ecology and conservation biology. Since the advent of diverse polymorphic markers (for example, microsatellites), assignment methods (Paetkau *et al.*, 1995; Rannala and Mountain, 1997; Cornuet *et al.*, 1999; Paetkau *et al.*, 2003) have allowed for rapid estimation of dispersal that would otherwise be difficult and time-consuming to obtain through direct observations (Berry *et al.*, 2004). Population assignment is one such tool that aims to identify the source population for specific individuals or assign them to multiple populations in the case of recent admixture. These methods have become important in forensics for identifying the provenance of material of unknown origin, determining the frequency of hybridization between species and the degree of connectivity among recently fragmented populations (for example, Cain *et al.*, 2000; Manel *et al.*, 2003; Paetkau *et al.*, 2003; Berry *et al.*, 2004). Although several methods for population assignment exist for diploid organisms, there are currently limited options for polyploids (but see Meirmans and Van Tienderen, 2004; Falush *et al.*, 2007). Polyploidy is a widespread phenomenon of major ecological and evolutionary importance in plants and animals (Otto and Whitton, 2000; Mable, 2004; Soltis *et al.*, 2004; Wood *et al.*, 2009). However, few population genetic studies of dispersal in polyploids have been conducted owing, in part, to a lack of methods that appropriately account for the complexities of polyploid data.

Although population assignment for diploids is relatively straightforward, several unique features of polyploids continue to provide significant challenges for implementing these techniques in natural populations. These partly depend on the presumed origin of whole-genome duplication (Ramsey and Schemske, 1998). Polyploids are commonly categorized broadly as either allopolyploid (derived from interspecific hybridization) or autopolyploid (derived from chromosomal doubling of the same genome). In allopolyploids, bivalents are mostly formed between pairs of homologous chromosomes (for example, A1/A2, B1/B2), resulting in disomic inheritance similar to that of diploids (Ronfort *et al.*, 1998). In contrast, segregation patterns in autopolyploids are considerably more complex because chromosomes either pair at random or form multivalents during meiosis. Polysomic inheritance in autopolyploids can result in two alternative segregation patterns. First, random chromosome segregation (RCeS), where gametes arise from any random assortment of homologous chromosomes but sister chromatids always end up in different gametes. Alternatively, maximum equational segregation and random chromatid segregation (RCdS) may occur where sister chromatids behave independently and distribute into the same gamete, a process that can result in double reduction (Bever and Felber, 1992). For example, consider an autotetraploid individual with four distinct alleles (*abcd*) at a locus. There are six possible gametes where sister chromatids distribute to different gametes (*ab*, *ac*, *ad*, *bc*, *bd* and *cd*) and four derived from double reduction (*aa*, *bb*, *cc* and *dd*). It is therefore important to consider these complexities given that they can influence segregation ratios, as well as expected gametic and genotype frequencies, at the population level.

A further challenge for polyploids is genotype ambiguity such that, for codominant markers (for example, microsatellites), allele dosage (copy number) cannot be reliably determined (Obbard *et al.*, 2006). Molecular markers are only able to detect which alleles are present but not how many of each there are. For example, in the case of a

[1]Department of Botany and Biodiversity Research, University of Vienna, Faculty of Life Sciences, Vienna, Austria; [2]CSIRO Plant Industry, Canberra, ACT, Australia and [3]Biodiversity Conservation Centre, Kings Park, Western Australia, Australia
Correspondence: Dr DL Field, Department of Botany and Biodiversity Research, University of Vienna, Faculty of Life Sciences, Rennweg 14, Vienna A-1030, Austria.
E-mail: david.field@univie.ac.at
Received 22 February 2017; revised 7 July 2017; accepted 24 July 2017; published online 4 October 2017

hexaploid individual, the presence of the two alleles (*a*, *b*) at a locus could reflect five possible genotypes (*aaaaab*, *aaaabb*, *aaabbb*, *aabbbb*, *abbbbb*). Therefore, many genotypes are indistinguishable and require the use of phenotypes (that is, the unique alleles present, Table 1). There is a long history of theory developed for understanding the population genetics of autopolyploids that incorporate some of the complexities of polysomic inheritance, double reduction and genotype ambiguity (for example, Haldane, 1930; Mather, 1935; Geiringer, 1949; Moody *et al.*, 1993; Ronfort *et al.*, 1998; Wu *et al.*, 2001; Luo *et al.*, 2006; Stift *et al.*, 2008; Meirmans and Tienderen, 2013). However, approaches that explicitly incorporate polysomic inheritance and double reduction to examine contemporary patterns of gene dispersal are currently unavailable for autopolyploids.

For diploids, population assignment is commonly achieved through frequency-based likelihood or full Bayesian approaches. Frequency-based methods such as *GeneClass* (for example, Cornuet *et al.*, 1999) use a sample of reference genotypes that provides information on the allele frequencies from each of the known (fixed) candidate populations. Individuals of unknown origin are then assigned probabilistically to their most likely population of origin. In contrast, the Bayesian method implemented in *Structure* (Falush *et al.*, 2007) uses an iterative algorithm (Markov Chain Monte Carlo) that randomly assigns individuals into a number of groups (predefined clusters) and converges when the assumption of Hardy–Weinberg and linkage equilibrium is fulfilled. Thus *Structure* simultaneously identifies the set of populations, their allele frequencies and the population membership coefficient of each individual, and these are updated until the best fit for the data is found. Currently, the only assignment approaches for polyploid data include *Genodive* (Meirmans and Van Tienderen, 2004) and *Structure* (Falush *et al.*, 2007), although both programs do not account for double reduction. Only *Structure* allows for phenotype markers; however, it remains unclear how accurate the method is for performing population assignment compared with a method that allows for polysomic inheritance with double reduction for autopolyploids. In addition, information on the maternal relationship for individual offspring (if known) is not utilized in existing assignment methods. However, in many cases, for example, seed collected from individual plants, including information on the genotype of the known maternal parent could increase the power of population assignment as only the population origin of the paternal parent requires evaluation.

Here we develop novel methods of population assignment for autotetraploid and autohexaploid species that explicitly account for polysomic inheritance with double reduction and ambiguous genotypes (implemented in the software *AutoPoly*). The main goal is to use allele frequency information from predefined reference genotypes sampled from a set of candidate populations and then assign a set of genotyped individuals of unknown origin (that is, offspring) to their most likely: (i) joint maternal and paternal source population, when both maternal and paternal origins are unknown (for example, seed dispersal), or (ii) paternal population of origin (for example, pollen dispersal), given the maternal parent is known (genotype and population of origin). For each of these approaches, we present methods for genotype (allele dosage known) and phenotype markers (allele dosage unknown). To assess the accuracy of these assignment methods in relation to *Structure*, we conducted a power analysis using simulated microsatellite (SSR) data and examined the effects of the number of loci, degree of population differentiation (F_{ST}), genotype ambiguity, maternal information, error rates and double reduction. From this, we address the following questions: (i) what is the difference in the accuracy of population assignment between genotype and phenotype data? (ii) does the inclusion of maternal information improve population assignment? (iii) how accurate is *AutoPoly* for providing point estimates of migration rates? and (iv) how does the accuracy of *AutoPoly* compare to *Structure*? Lastly, we test these methods using an empirical data set for the autohexaploid plant *Eremophila glabra*.

Table 1 Genotype classes, phenotypes and general formulas for the number of possible genotypes given *k* codominant alleles

Class	*Genotype*	*Phenotype*	*No. of genotypes*	k = *6*	
				Genotype	*Phenotype*
Monoallele	*aaaaaa*	*a*	k	6	6
Biallele I	*aaaaab*	*ab*	$k(k-1)$	30	15
Biallele II	*aaaabb*	*ab*	$k(k-1)$	30	
Biallele III	*aaabbb*	*ab*	$k(k-1)/2$	15	
Triallele I	*aaaabc*	*abc*	$k(k-1)(k-2)/2$	60	20
Triallele II	*aaabbc*	*abc*	$k(k-1)(k-2)$	120	
Triallele III	*aabbcc*	*abc*	$k(k-1)(k-2)/6$	20	
Quadriallele I	*aaabcd*	*abcd*	$k(k-1)(k-2)(k-3)/6$	60	15
Quadriallele II	*aabbcd*	*abcd*	$k(k-1)(k-2)(k-3)/4$	90	
Pentallele	*aabcde*	*abcde*	$k(k-1)(k-2)(k-3)(k-4)/24$	30	6
Hexallele	*abcdef*	*abcdef*	$k(k-1)(k-2)(k-3)(k-4)(k-5)/720$	1	1
Total				462	63

An example for the number of possible genotypes and phenotypes when $k=6$.

METHODS

Likelihood model for polyploid population assignment

In our model, individuals are autopolyploid with either four (Y_4) or six (Y_6) sets of chromosomes (that is, $2n=4x=$ tetraploid; $2n=6x=$ hexaploid), but all populations must have the same ploidy level for any given analysis. Random mating is assumed within each reference population (both in terms of zygote and gamete dispersal) and loci are assumed to be unlinked and in linkage equilibrium. We allow segregation patterns at a given locus to follow expectations for polysomic inheritance with multivalent formation under random chromatid segregation (RCdS). To allow for any double reduction rate (DRR), we use general formulas for DRR anywhere within the theoretical bounds (for RCdS, tetraploids, $0<\alpha<(1)/(7)$; hexaploids, $0<\beta<(3)/(11)$) (Mather, 1936; Geiringer, 1949). Here we assume the maximum double reduction follows that expected for RCdS rather than maximum equational segregation (tetraploids, $0<\alpha<(1)/(6)$; hexaploids, $0<\beta<(3)/(10)$). We assumed RCdS as this was more tractable for calculating general formulas for segregation ratios and the specific requirements for maximum equational segregation (that is, only one crossover event between locus and centromere) is rather restrictive. Moreover, for most empirical data sets, DRR at a given locus remains unknown (but see Stift *et al.*, 2008) but probably lies somewhere between the theoretical minimum and maximum. By always using general formulas for α and β, we circumvent the problem of other methods that do not allow for multivalent chromosome formation and double reduction (that is, *Structure* and *Genodive*) or assume that double reduction is fixed at either the theoretical minimum or maximum (for example, Buteler *et al.*, 1997).

We consider a set of I discrete populations that exchange zygotes (for example, seed) or gametes (for example, pollen). In each population, a representative sample of n individuals are either genotyped (allele dosage known) or phenotyped (allele dosage unknown). We let G_{ijm} and P_{ijm} denote the genotype and phenotype at the jth locus ($j=1, 2, \ldots, J$) for the mth individual ($m=1, 2, \ldots, M$) located in the ith population ($i=1, 2, \ldots, I$). For

example, an individual genotype (G_{ijm}) lists the alleles detected where the total alleles recorded must equal the ploidy level (for example, tetraploid, *aabc*; hexaploid, *aabcde*), whereas a phenotype (P_{ijm}) lists only the unique alleles present (for example, tetraploid, *abc*; hexaploid, *abcde*). We let $\boldsymbol{G}=\{G_{ijm}\}$ and $\boldsymbol{P}=\{P_{ijm}\}$ represent the matrix of genotypes or phenotypes of individuals in the sampled population. The model only allows for resolved genotypes or phenotypes for a given analysis (that is, cannot include both phenotypes and genotypes). Although the majority of empirical data sets will consist of phenotypes, we describe both approaches because beginning with unambiguous genotypes is an easier starting point.

Our assignment method builds on techniques designed for diploids for individual based population assignment using multilocus genotypes (Rannala and Mountain, 1997; Cornuet *et al.*, 1999) and consists of five main steps that calculate: (1) allele frequencies in each candidate (reference) population, (2) expected gamete frequencies at random mating equilibrium (RME) in each population, (3) expected genotype frequencies at RME, (4) assignment probabilities, and (5) simulations to determine the confidence intervals (CIs) for assignment. For each step, we describe the methods for autotetraploids followed by autohexaploids and, when required, derivations for both genotype and phenotype data. For the assignment probabilities (step 4) we describe separately the methods for: Model I = joint maternal and paternal population assignment with a single population origin (for example, seed dispersal), Model II = joint maternal and paternal population assignment with an admixed population origin, and Model III = paternal population assignment given the known maternal genotype of each offspring (for example, pollen dispersal).

Allele frequencies

The first step in population assignment requires that allele frequencies are estimated in each of the reference populations to be evaluated as potential source populations. For genotype data, the frequency of each allele in a given population can be directly counted from information in the genotype matrix at each locus ($\boldsymbol{G}_{ij}$), as in diploids. In contrast, for phenotype data ($\boldsymbol{P}_{ij}$), only the distinct alleles that are carried by an individual are known. We use two alternative approaches: (i) Expectation-Maximization (EM)-based estimation, and (ii) marginal (weighted) allele frequency. The EM method follows the approach outlined by De Silva *et al.* (2005) and implemented in *Polysat* (Clark and Jasieniuk, 2011) and we run this approach assuming no selfing (for example, self-incompatible plants). One limitation of this method is that it assumes only RCeS occurs, meaning that double reduction under RCdS is not incorporated. To avoid the problem of using unknown priors or restricted assumptions on the nature of polysomic inheritance, we also use an alternative estimate based on the marginal allele frequency. This approach is equivalent to summing the allele counts over the set of possible genotypes for each given phenotype, which can be approximated by determining the number of individuals in each phenotypic class and weighting these proportionally to the number of alternative alleles. Here we let the vector $\boldsymbol{p}_{ij}=\{p_{1ij},\ldots,p_{kij}\}$, where p_{kij} is the frequency of the kth allele at the jth locus in the ith population. Given the vector of phenotypes P_{ij} and Y_4 (tetraploid), we find the frequency of the kth allele as:

$$\Pr\left(p_{kji}|P_{ij},Y_4\right)=\frac{n_{k4}+(4/3)n_{k3}+2n_{k2}+4n_{k1}}{4N_i} \quad (1)$$

where we denote N_i as the total number of individuals in the ith population and n_{k4}, n_{k3}, n_{k2} and n_{k1} are the number of quadriallele, triallele, biallele and monoallele individuals carrying the kth allele, respectively. In the case of quadriallele (first term nominator; n_{k4}), the allele counts are unambiguous as the genotype is known. For triallele phenotypes (the second term on the nominator; $(4/3)n_{k3}$), there can be four total copies of kth allele across three alternative genotypes. Similarly, for biallele phenotypes (third term; $2n_{k2}$), summing across possible genotypes there are a total of six copies for three genotypes. Lastly, for monoallele phenotypes (last term nominator; $4n_{k1}$), these are unambiguous as genotype is known. Following this same procedure, for Y_6 (hexaploid):

$$\Pr\left(p_{kji}|P_{ij},Y_6\right)=\frac{n_{k6}+1.2n_{k5}+1.5n_{k4}+2n_{k3}+3n_{k2}+6n_{k1}}{6N_i} \quad (2)$$

where n_{k6}, n_{k5}, n_{k4}, n_{k3}, n_{k2} and n_{k1} are the number of hexallele, pentallele, quadriallele, triallele, biallele and monoallele individuals carrying the kth allele, respectively. Compared with the EM-based estimation, the marginal allele frequency method may result in a bias towards more uniform allele frequencies, particularly when the population sample is small.

Population gametic probabilities

The next step requires the expected frequency of each gamete in each reference population under the assumption of random mating (RME), given the allele frequencies are known and a given DRR. For autopolyploids with RCdS, we must first calculate the expected frequencies of all possible gametes at equilibrium from the allele frequencies. In contrast to diploids, autopolyploids do not reach equilibrium after one generation of random mating but approach this asymptotically (Haldane, 1930; Geiringer, 1949; Bever and Felber, 1992). However, this can be approximated with general limit formulas for RME under segregation patterns intermediate between RCeS and RCdS (Geiringer, 1949).

We denote the vector $\boldsymbol{g}_{ij}=\{g_{1ij},\ldots,g_{mij}\}$, where g_{mij} is the expected frequency of the mth gamete at the jth locus in the ith population at RME. Tetraploids can transmit two allele copies and thus two classes of gametes are possible, either a monoallele or a biallele which occur with the frequencies x_{kk} and $2y_{kk'}$, respectively (where allele $k\neq k'$). For example, in a tetraploid population with $k=2$ unique alleles, there are three possible gametes. For clarity, we use notation x_{11} in place of x_{kk}, this gives x_{11}, $2y_{12}$, and x_{22}. To explicitly allow for polysomic inheritance and any DRR, we use the general limit formulas derived by Geiringer (1949) to calculate the equilibrium gamete frequencies for any given probability of double reduction ($0<\alpha<0.1428$) (also see Wricke and Weber, 1986) as:

$$\begin{aligned}\Pr\left(x_{11}|\boldsymbol{p}_{ij},Y_4\right)&=p_{kij}{}^2+\tfrac{3\alpha}{2+\alpha}p_{kij}\left(1-p_{kij}\right)\\ \Pr\left(2y_{12}|\boldsymbol{p}_{ij},Y_4\right)&=p_{kij}q_{k'ij}\tfrac{4-4\alpha}{2+\alpha}\end{aligned} \quad (3)$$

where p_{kij} and $q_{k'ij}$ is the frequency of alleles k_1 and k_2 at the jth locus in the ith population. When $\alpha=0$, the general limit formulas reduce to the binomial expansion of $(p+q,\ldots,k_i)^{Y/2}$, where $x_{11}=p^2$, $2y_{12}=2pq$ and $x_{22}=q^2$.

Hexaploid gametes transmit three alleles and can be classified into three classes in a hexaploid depending on the number of unique alleles they carry. These include monoallele gametes (x_{kkk}), biallele ($3y_{kkk'}$ and $3y_{kk'k'}$) and triallele gametes ($6z_{kk'k''}$). For example, a hexaploid population with a total of $k=3$ alleles, there are 10 possible gametes (x_{111}, x_{222}, x_{333}, $3y_{112}$, $3y_{122}$, $3y_{223}$, $3y_{233}$, $3y_{113}$, $3y_{133}$ and $6z_{123}$). We used the general limit formulas (Equation 28; Geiringer, 1949) to calculate the equilibrium gamete frequencies for any DRR ($0<\beta<0.2727$). The probability of the ith population producing monoallele, biallele and triallele gametes in a hexaploid population is:

$$\begin{aligned}\Pr\left(x_{111}|\boldsymbol{p}_{ij},Y_6\right)&=\tfrac{27(1-\beta)(3-\beta)}{(9+\beta)(9+2\beta)}p_{kij}^3+\tfrac{45\beta(3-\beta)}{(9+\beta)(9+2\beta)}p_{kij}^2+\tfrac{20\beta^2}{(9+\beta)(9+2\beta)}p_{kij}\\ \Pr\left(3y_{112}|\boldsymbol{p}_{ij},Y_6\right)&=\tfrac{27(1-\beta)(3-\beta)}{(9+\beta)(9+2\beta)}p_{kij}^2q_{k'ij}+\tfrac{15\beta(3-\beta)}{(9+\beta)(9+2\beta)}p_{kij}q_{k'ij}\\ \Pr\left(6z_{123}|\boldsymbol{p}_{ij},Y_6\right)&=\tfrac{27(1-\beta)(3-\beta)}{(9+\beta)(9+2\beta)}p_{kij}q_{k'ij}r_{k''ij}\end{aligned} \quad (4)$$

Where p_{kij}, $q_{k'ij}$, $r_{k''ij}$ represent the observed frequency of allele k_1, k_2 and k_3, respectively. When $\beta=0$, the general formulas reduce to the trinomial expansion of $(p+q+r,\ldots,k_i)^{Y/2}$ (for example, $x_{111}=p^3$, $3y_{112}=3p^2q$, $6z_{123}=6pqr$).

Genotype and phenotype probabilities

For tetraploids, when the mother is unknown, the probability of observing the genotype G_{ijm} for individual m at locus j, given it is from population i, is dependent on assuming the individual is solely from population i and the vector of expected gametic frequencies from population i (g_{ij}) and polyploidy = Y_4. This equates to the genotype probabilities at RME, $\Pr(G_{ijm}|i, g_{ij}Y_4)$. Henceforth, to distinguish genotypes we denote alleles with letters and their allele copy number with subscripts. The probability of observing each

genotype class follows Geiringer (1949) as

$$\begin{aligned}
&\Pr\left(a_4|i,\boldsymbol{g}_{ij},Y_4\right) = (x_{11})^2\\
&\Pr\left(a_3b|i,\boldsymbol{g}_{ij},Y_4\right) = 4x_{11}y_{12}\\
&\Pr\left(a_2b_2|i,\boldsymbol{g}_{ij},Y_4\right) = 4(y_{12})^2 + 2x_{11}x_{22}\\
&\Pr\left(a_2bc|i,\boldsymbol{g}_{ij},Y_4\right) = 4x_{11}y_{23} + 8y_{12}y_{13}\\
&\Pr\left(abcd|i,\boldsymbol{g}_{ij},Y_4\right) = 8y_{12}y_{34} + 8y_{13}y_{24} + 8y_{14}y_{23}
\end{aligned} \tag{5}$$

Unlike for tetraploids, as far as we are aware, there are no general formulas to calculate the expected genotype frequencies for hexaploids that take into account double reduction. Therefore, we derived the expected genotype frequencies for hexaploids at RME equilibrium that simply follows the random union of gametes within each population (Appendix A1). We follow the same notation for tetraploids. The probabilities of observing each genotype class given the individual from population i are:

$$\begin{aligned}
&\Pr(a_6|i,\boldsymbol{g}_{ij},Y_6) = (x_{111})^2\\
&\Pr(a_5b|i,\boldsymbol{g}_{ij},Y_6) = 6x_{111}y_{112}\\
&\Pr(a_4b_2|i,\boldsymbol{g}_{ij},Y_6) = 6x_{111}y_{122} + 9(y_{112})^2\\
&\Pr(a_3b_3|i,\boldsymbol{g}_{ij},Y_6) = 2x_{111}x_{222} + 18y_{112}y_{122}\\
&\Pr(a_4bc|i,\boldsymbol{g}_{ij},Y_6) = 12x_{111}z_{123} + 18y_{112}y_{113}\\
&\Pr(a_3b_2c|i,\boldsymbol{g}_{ij},Y_6) = 6x_{111}y_{223} + 36y_{112}z_{123} + 18y_{113}y_{122}\\
&\Pr(a_2b_2c_2|i,\boldsymbol{g}_{ij},Y_6) = 18y_{112}y_{233} + 18y_{133}y_{122} + 18y_{113}y_{223} + 36(z_{123})^2\\
&\Pr(a_3bcd|i,\boldsymbol{g}_{ij},Y_6) = 12x_{111}z_{234} + 36y_{112}z_{134} + 36y_{113}z_{124} + 36y_{114}z_{123}\\
&\Pr(a_2b_2cd|i,\boldsymbol{g}_{ij},Y_6) = 36y_{112}z_{234} + 18y_{114}y_{224} + 18y_{114}y_{223}\\
&\qquad\qquad + 72z_{123}z_{124} + 36z_{134}y_{122}\\
&\Pr(a_2bcde|i,\boldsymbol{g}_{ij},Y_6) = 36y_{112}z_{345} + 36y_{113}z_{245} + 36y_{114}z_{235} + 36y_{115}z_{234}\\
&\qquad\qquad + 72z_{123}z_{145} + 72z_{134}z_{125} + 72z_{124}z_{135}\\
&\Pr(abcdef|i,\boldsymbol{g}_{ij},Y_6) = 72(z_{123}z_{456} + z_{124}z_{356} + z_{125}z_{346} + z_{126}z_{345}\\
&\qquad\qquad + z_{134}z_{256} + z_{135}z_{246} + z_{136}z_{245} + z_{145}\\
&\qquad\qquad z_{236} + z_{146}z_{235} + z_{234}z_{156})
\end{aligned} \tag{6}$$

For hexaploids, this gives 11 distinct genotype classes (Table 1). Although the number of possible genotypes increases rapidly with the number of alleles (for example, when $k=6$, gives 462 genotypes), the expected genotype frequencies at RME for each can be calculated on the basis of their respective genotype class.

For phenotypes, the expected frequency of full homozygotes and heterozygotes (for example, monoallele and hexallele for a hexaploid, respectively) are equal to genotype frequencies at RME. To evaluate genotype probabilities for partial heterozygotes, we must account for the lack of allele copy number. To address this problem, we take the sum of the probabilities of obtaining each of the possible genotypes. For example, for a phenotype with three unique alleles detected, *abc*, the set of possible genotypes is $G_{ijm}\in\{P_{ijm}\}=\{aabc, abbc, abcc\}$. Therefore, the probability of obtaining each of the alternative genotypes is proportional to their frequencies at RME, which depends on the equilibrium gametic frequencies in the ith population (g_{ij} and ploidy (Y_x), given the population from which the individual was sampled,

$$\Pr\left(P_{ijm}|i,g_{ij},Y_x\right) = \sum_{G_{ijm}\in\{Ph\}}\Pr(G_{ijm}|i,g_{ij},Y_x) \tag{7}$$

Assignment probabilities

Model I. Population assignment (single candidate, mother unknown). Here we assume individual m is from two unknown parents that belong to a single candidate population. This probability comes directly from the probability of observing the genotype at each of the candidate populations. Assuming the alleles at the J loci are independent (no linkage), we calculate the probability of observing the multilocus genotype $\boldsymbol{G}_{im}$, or phenotype $\boldsymbol{P}_{im}$, at a given candidate population i as the product of the probabilities at each locus:

$$\begin{aligned}
&\Pr(\mathbf{G}_{im}|i,g_i,Y_x) = \prod_{j=1}^{J}\Pr\left(G_{ijm}|i,g_{ij},Y_x\right)\\
&\Pr(\mathbf{P}_{im}|i,g_i,Y_x) = \prod_{j=1}^{J}\Pr\left(P_{ijm}|i,g_{ij},Y_x\right)
\end{aligned} \tag{8}$$

when an allele is absent from population i, this results in a zero probability of gametes and genotypes that carry that allele. This reduces the probability of the multilocus genotype/phenotype to zero, although the particular allele may be rare in the population or missing among the reference individuals (Cornuet *et al.*, 1999). To account for this problem, we follow Rannala and Mountain (1997) and let the frequency of the absent allele be proportional to the inverse of the number of gene copies at the locus, adjusted by the number of observed alleles as, $p_{kji^*}=(1/K_j)/N_{ij}Y$, where K_j is the total number of alleles detected across all populations for the jth locus, N_i are the total number of individuals sampled in the ith candidate population and Y is the ploidy level (for example, $Y=6$ for hexaploid). Given this allele frequency, we re-calculate allele frequencies proportionally so that the sum of allele frequencies in each population sums to one and then re-calculate the probability of all possible gametes and phenotypes/genotypes.

Model II. Population assignment (admixed individuals). We now consider the situation in which we assume one parent is a resident of population i and assume the other parent belongs to a different candidate population i'. We denote this first-generation admixed genotype as $G_{[i,i']jm}$, where one gamete is from the resident population i (that is, $g_{[i]j}\{..\}$) and the other gamete is from population i' (that is, $g_{[i,i']j}\{..\}$). For tetraploids, the probability of observing individual m which is a mixed (F_1) genotype $G_{[i,i']jm}$ at locus j depends on the gametic frequencies in each of the two populations (g_{ij}, $g_{i'j}$) and can be written as $\Pr\left(G_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_4\right)$. We replace the term $G_{[i,i']}$ for each specific genotypic class as follows:

$$\begin{aligned}
&\Pr\left(a_4[i,i']jm|i,i',g_{ij},g_{i'j},Y_4\right) = x_{11[i]}x_{11[i']}\\
&\Pr\left(a_3b_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_4\right) = 2x_{11[i]}y_{12[i']} + 2x_{11[i']}y_{12[i]}\\
&\Pr\left(a_2b_2[i,i']jm|i,i',g_{ij},g_{i'j},Y_4\right) = x_{11[i]}x_{22[i']} + x_{11[i']}x_{22[i]} + 2y_{12[i]}y_{12[i']} + 2y_{12[i']}y_{12[i]}\\
&\Pr\left(a_2bc_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_4\right) = 2x_{11[i]}y_{23[i']} + 2x_{11[i']}y_{23[i]} + 4y_{12[i]}y_{13[i']} + 4y_{12[i']}y_{12[i]}\\
&\Pr\left(abcd_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_4\right) = 4y_{12[i]}y_{34[i']} + 4y_{12[i']}y_{34[i]} + 4y_{13[i]}y_{24[i']} + 4y_{13[i']}y_{24[i]}\\
&\qquad\qquad + 4y_{14[i]}y_{14[i']} + 4y_{14[i']}y_{14[i]}
\end{aligned} \tag{9}$$

Similarly, the probability of mixed genotypes for a hexaploid individual is $\Pr\left(G_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right)$ and are described in Appendix A2.

Phenotype data are treated as outlined for the case of non-admixed individuals (Equation 7). Similarly, the probability of observing the multilocus genotype $\boldsymbol{G}_{[i,i']m}$, or phenotype $\boldsymbol{P}_{[i,i']m}$, assuming it is an F_1 between population i and i', follows that of Equation 8.

Model III. Population assignment (paternal origin given mother known). Now we consider a situation where individual offspring are sampled and the identity of the female (mother) and her population of origin and genotype/phenotype are known, but the location and identity of the male (father) is unknown. The location of the unknown father may be in any of the candidate populations, including that of the known female. The intent is to determine for a given offspring (o), the most likely source of the male gamete (x_m), given that the female parent (f) is known. This can be expressed in a similar framework used for paternity analysis (for example, Meagher, 1986). We evaluate the probability of obtaining the offspring genotype (G_{oj}) given the following relationship: f is a parent of o and the male parent, m_i, is located in the ith population. We make the assumption that the female and male parents are not F1 or recent immigrants to the population in which they were sampled. The probability depends on the gamete frequencies in the ith population (g_{ij}), ploidy (Y_x), and the DRR (α_j or β_j) at the jth locus and can be written as:

$$\Pr\left(G_{oj}|G_{fj},g_{ij},Y_x,\alpha_j;\beta_j\right) = \sum_{x_f}\Pr(G_{oj}|x_f)\Pr(x_f|G_{fi})\Pr(x_m|g_{ij}) \tag{10}$$

where x_f and x_m are the female and males gametes, respectively, $P(x_f|G_{fj})$ is the gamete segregation probability from a given female genotype and $P(x_m|g_{ij})$ is the probability of the male gamete given the expected gamete frequencies in a candidate population (Equations 3 and 4). For polyploids, there can be many alternative gametes that two parents could have contributed towards the offspring, hence we sum over all the possible gametes segregating from the known female parent. Although gametic segregation ratios have been described previously for autohexaploids for fixed RCeS or RCdS, most loci probably exhibit intermediate DRR between the two extremes. Therefore, for hexaploids we derived generalized segregation probabilities for any value of β (Appendix A3).

Likelihood of population assignment

For each individual, we evaluate the likelihood of population assignment to each of the candidate populations and, being a first generation, between all pair-wise candidate populations. Following Cornuet *et al.* (1999), for Model I we take the logarithms of the genotype probabilities in each candidate population ($i = 1, 2,\ldots, I$). For Model II, we take the logarithms of each hth population pair ($h = 1, 2,\ldots, H$), where the number of admixed population genotype/phenotype possibilities is $H = I(I-1)$. In the case of two candidate populations i and i', for example, the log-likelihoods of the observing individual m, assuming it is solely from population i, or assumed m is an admixed genotype originating from two populations [i, i'] would be:

$$\begin{matrix}\ln[\Pr(\boldsymbol{G}_{im}|g_i, Y_x)]\\ \ln[\Pr(\boldsymbol{G}_{i'm}|g_{i'}, Y_6)]\\ \ln[\Pr(\boldsymbol{G}_{[i,i']m}|g_i, g_{i'}, Y_6)]\end{matrix} \quad (11)$$

Each individual is assigned to the population or admixed population pair in which the likelihood of observing the individual's genotype/phenotype is the highest. Using this method, a candidate population or mixed population pair is always assigned from among the set of reference populations. In order to discriminate between the most likely candidates, we use the statistic, ln Δ, as the difference in the log-likelihood of the most likely candidate ($\ln_1$) and the second most likely ($\ln_2$):

$$\ln \Delta = \ln_1 - \ln_2 \quad (12)$$

Confidence intervals

We used simulations to assess the accuracy of assignment procedures and identify the critical values of ln Δ. The aim here is to provide a measure of confidence that an individual belongs to the assigned candidate population or jointly to two populations in the case of admixed genotypes/phenotypes. We generated new sets of multilocus genotypes/phenotypes for each population by drawing gametes according to their expected frequencies in the reference samples. Similarly, for population assignment when the mother is unknown, we generated new sets of mixed multilocus genotypes/phenotypes for each of the hth population pairs ($h = 1, 2,\ldots, H$). Next we compare ln Δ between groups of simulated individuals that were assigned correctly and incorrectly. Critical values for population assignment were approximated from the distribution of ln Δ values of the simulated data (typically $n = 10\,000$; see Supplementary Information S1 and Supplementary Figure S1 for more details).

Genotyping errors

To examine the effects of genotype/phenotype errors on critical values, we modified the simulated individuals according to two sources of error, e_1 and e_2, where e_1 is the probability of allelic dropout (removing an allele from a phenotype or genotype) and e_2 is the probability of an allele being mis-scored. For the latter, e_2, an allele is replaced with an alternative proportional to the allele frequencies in the sampled population. These simulations did not explicitly model null alleles; however, increasing rates of allelic dropout will generate similar effects on critical values.

Power analysis with simulated data sets

In order to compare the performance of these different polyploid assignment methods, we simulated populations using an individual-based model with polysomic inheritance and double reduction. Briefly, this simulation considered a finite island model with migration between 10 populations of constant size, each containing 1000 hermaphrodite individuals (with no selfing) with 24 microsatellite loci and a mutational rate ($\mu = 2 \times 10^{-4}$). By running separate simulations with different migration rates, we obtained replicate data sets with five different levels of average population differentiation (F_{ST}: 0.03, 0.06, 0.09, 0.13 and 0.20; see Supplementary Information S2 for more details).

We ran each combination of F_{ST} (0.03, 0.06, 0.09, 0.13 and 0.20), with three different number of loci (6, 12 and 24), two marker types (genotype, phenotype) and two model approaches (mother unknown (Models I and II) and mother known (Model III)) for two ploidy levels (tetraploid and hexaploid). Once the model was at mutation–migration equilibrium, we randomly sampled a set of reference samples ($n = 60$ from each population) and generated a set of offspring ($n = 10\,000$). Here, among the final set of offspring, the migration rate was increased to $m = 0.5$, so that ~5000 were generated from random mating within populations and the remaining from interpopulation mating. Although this represents an atypically high migration rate, this facilitated comparisons of accuracy on the equal sample size between different metapopulations with different F_{ST}. Owing to computational constraints for forward-time simulation of polyploid populations, we obtained 10 replicates for each combination of the above parameters (total $n = 1200$ simulations). Low s.e. for accuracy among replicates, particularly for higher F_{ST} 0.13 and 0.20 and >12 loci (s.e. in accuracy <1%) suggests that this number was sufficient to demonstrate differences in the various assignment methods. Given that the incorrect choice of α and β had little impact on accuracy (Supplementary Information S2), we drew a random DRR at each locus between the theoretical minimum and maximum values for RCdS for α (0–0.14) and β (0–0.2727). With *AutoPoly*, individuals were assigned to their most likely paternal population of origin at 80% confidence calculated from simulating $n = 10\,000$ individuals with error rates ($e_1 = 0.005$ and $e_2 = 0.005$).

Migration rate point estimates

We also examined the performance of *AutoPoly* to provide estimates of interpopulation migration rates. Although population assignment is not designed to explicitly estimate migration rates, point estimates can be obtained by dividing the number of detected immigrants by the total sample size (Manel *et al.*, 2003). Here we simulated autohexaploid populations with different migration rates for populations with different degrees of population differentiation (F_{ST}) and number of loci that resemble data typically available for studies of natural populations. Following the same procedure described in the power simulations, once the simulation reached drift-mutation equilibrium, we generated a final set of offspring ($n = 10\,000$ from each population). For the progeny, the migration rate, m, is the probability that individual offspring were generated through interpopulation mating, while $1 - m$ were generated from random mating within populations. We ran 10 replicates for each of the following combination of parameters: migration rate m (0.02, 0.1, 0.2), F_{ST} (0.03, 0.06, 0.09), and number of loci (6, 12). Here we examined phenotype markers, mother known (Model III) and a fixed intermediate DRR ($\beta = 0.136$) (that is, $n = 180$ simulations), which was assumed known in the assignment test. Individuals were assigned with the same conditions in the previous power simulations.

AutoPoly and Structure

In order to compare the performance of *AutoPoly* with *Structure*, we use a subset of the same simulated data detailed above for the power simulations. Here we focus on tetraploid populations with lower levels of population differentiation (F_{ST}: 0.03, 0.06, 0.09). Initial simulations identified little difference in accuracy between the programs with higher levels of F_{ST} (both methods >98% accuracy).

With *AutoPoly*, individuals were assigned as in the above simulations. We used *Structure* version 2.3.4 (Falush *et al.*, 2007) to assign individuals to their most likely candidate population or admixed population pair. Here we used the admixture model, updated allele frequencies only for the reference individuals, set the number of genetic clusters to $K = 10$ and used sampling location as prior information (LOCPRIOR). Therefore, unlike studies that aim to search for the most likely number of clusters, here we assume K is known and equal to the

number of demes in the simulated metapopulation. All data sets were run for a burn-in period of 20 000 and 200 000 iterations of the Markov Chain Monte Carlo (see Supplementary Information S2 for more details). The POPINFO parameter is not available for ploidy >2. Therefore, we assigned each individual using the membership coefficients (that is, ancestry proportion) Q_k, which represents the posterior probability of membership to each of the $K=10$ clusters (here Q_1, Q_{k+1},…,Q_{10}). We assigned an individual to belong solely to the kth population if $Q_k > A_Q$, where A_Q is the threshold Q_k value for assignment. We tested a range of A_Q thresholds (0.9, 0.8, 0.7, 0.6). If all coefficients were in the bounds of $(1-A_Q) < Q < A_Q$, an individual was instead assigned as admixed, with the most likely population pair involved being the two populations with the first and second highest Q values.

Autohexaploid empirical example

As an empirical test of this method, we used microsatellite (SSR) data available for the autohexaploid bird-pollinated shrub *Eremophila glabra* ssp. *glabra* from central Australia. At the study site, *E. glabra* is found in a series of discrete populations that are separated by agriculture. In this $\sim 15 \times 15$ km^2 area, reference landscape grids were delineated and intensively surveyed for *E. glabra* plants, identifying 15 discrete populations. To obtain reference allele frequencies, 32–62 individuals at each population were phenotyped at six highly polymorphic microsatellite markers (allelic dosage cannot be resolved in *E. glabra*). For four populations, 7–11 seed were phenotyped from up to 11 known plants. DNA extraction methods and SSR protocols follow those of Elliott (2009). Diversity and divergence measures were calculated using the adult samples from each of the 15 populations (see Supplementary Information S3 for more details).

We performed population assignment for each of the offspring phenotypes using *AutoPoly* when the mother is known (Model III). Therefore, we are assessing the most likely origin of the paternal (pollen) parent that may be the same population as the mother or any of the other 14 candidate populations sampled. Given *E. glabra* is bird pollinated, it is feasible that pollen could be dispersed among any of the candidate populations within the 15×15 km^2 area. To calculate CIs, we simulated five replicates of $n=10\,000$ individuals at each of the two different total error rates $E=0.01$ and 0.03 (where $E=e_1+e_2$ and $e_1=0.005$ and 0.015 and $e_2=0.005$ and 0.015). To examine the effect of using the incorrect DRR, we used randomly drawn values between the theoretical minimum and maximum in the simulations 'true DRR' but drew another set of randomly drawn 'assumed DRR' to use in the assignment calculations of the individual data and compare to when these values are the same. We used random values because real loci are more likely to vary somewhere between the theoretical minimum and maximum due simply to variation in marker position on the chromosome and distance from the centromere. As we discovered in the power simulations for data sets with low F_{ST} (0.03) and marker number (that is, 6 loci), *Structure* would not give consistent results with the *E. glabra* data due to a lack of model convergence.

RESULTS

Power analysis

The accuracy of the population assignment for both tetraploids (Figure 1) and hexaploids (Figure 2) increased with more loci and higher population differentiation. The accuracy was generally similar between ploidy levels, with 0.5–3% greater accuracy for hexaploids compared with tetraploids. Accuracy can be improved by increasing the confidence threshold, although this comes at the expense of the number of individuals that can be assigned for data with low information content. For example, for only 6 loci, phenotypes and mother known, a mean of ~84% could be assigned at 80% confidence, compared with mean 35% assigned at 95% confidence. However, the accuracy improves rapidly (exceeds 90%) under moderate levels of population differentiation ($F_{ST}=0.09$) even with only six loci regardless of marker type or model approach.

The difference in assignment accuracy between genotype and phenotype markers was relatively low for most simulation parameters but was most evident at lower population differentiation. For example, with 6 loci, genotypes had between 3% ($F_{ST}=0.03$) and 0.5% ($F_{ST}=0.20$) greater accuracy for tetraploid data when using Model III (mother known; Figure 1). Greater differences in accuracy between genotypes and phenotypes were observed with increasing error rates and under mother unknown models. For example, with a total error rate $=0.03$ and assuming that the mother was unknown, genotypes had 14.1% ($F_{ST}=0.03$) to 9.9% ($F_{ST}=0.20$) greater accuracy than phenotypes.

The inclusion of maternal information (Model III—mother known) resulted in a large improvement in the accuracy of population assignment compared with Models I and II (mother unknown). For example, with phenotypes, the mother known model had 19.1% ($F_{ST}=0.03$) to 4.4% ($F_{ST}=0.20$) greater accuracy than mother unknown (Figures 1 and 2). The difference between these model types was most evident at low levels of population differentiation and when only 6 or 12 loci are available. With 24 loci and $F_{ST} \geqslant 0.13$, there was no difference in assignment accuracy.

Migration rates

The power simulations suggest that false positives can substantially inflate the estimated number of immigrant genotypes (Figure 3). Substantial inflation of immigrant genotypes of ~25% above the true value was generated when population differentiation was low ($F_{ST}=0.03$) and few loci are simulated ($n=6$) (Figure 3a). Genotyping individuals for more loci ($n=12$) reduced this bias to ~10%, while scenarios with higher population differentiation of $F_{ST}=0.06$ and $F_{ST}=0.09$ reduced this bias further to ~2% (Figure 3b) and $<1\%$ (Figure 3c), respectively (assuming loci $=12$). Although the degree of bias is considerable for data sets with low power, the bias observed in the estimated migration rate is relatively consistent across different values of the true migration rate.

AutoPoly and Structure

We found that the two programs exhibited very similar results when more marker data was available (number of loci $=12$ or 24) and population differentiation was high ($F_{ST}=0.09$) (Figure 4). In contrast, the performance of *Structure* (with a threshold Q value of $A_Q > 0.7$) was substantially lower than *AutoPoly* for all parameters combinations with six loci and $F_{ST}=0.06$ and with $\leqslant 12$ loci for the lowest differentiation of $F_{ST}=0.03$. In some cases, this prevented *Structure* from running with data sets of six loci for either genotype or phenotype (that is, missing data points; Figures 4a and b).

The impact of marker type had drastically different effects on the accuracy of the two methods. In the case of *AutoPoly*, the accuracy of genotypes was only slightly higher than phenotypes (see above). In contrast, *Structure* exhibited much higher accuracy (up to ~40%) with genotypes than with phenotypes for the exact same parameters, except when 24 loci were available or population differentiation was high ($F_{ST}=0.09$).

Population assignment in *E. glabra*

In the *E. glabra* populations, we detected an average of 14.5–23 alleles across loci in each population (A, Supplementary Table S1). Mean pair-wise population differentiation was low ($Rho=0.082 \pm 0.031$ s.d.; Supplementary Table S2) and similar to the lowest differentiation examined in the power testing simulations (that is, $F_{ST}=0.03$; $Rho=0.11 \pm 0.01$ s.d., Supplementary Tables S3–S7).

Using simulations of mating events within and among populations of *E. glabra*, the number of individuals that could be assigned at a given confidence threshold was generally lower than suggested from power simulations. Simulations of *E. glabra* data indicate that between

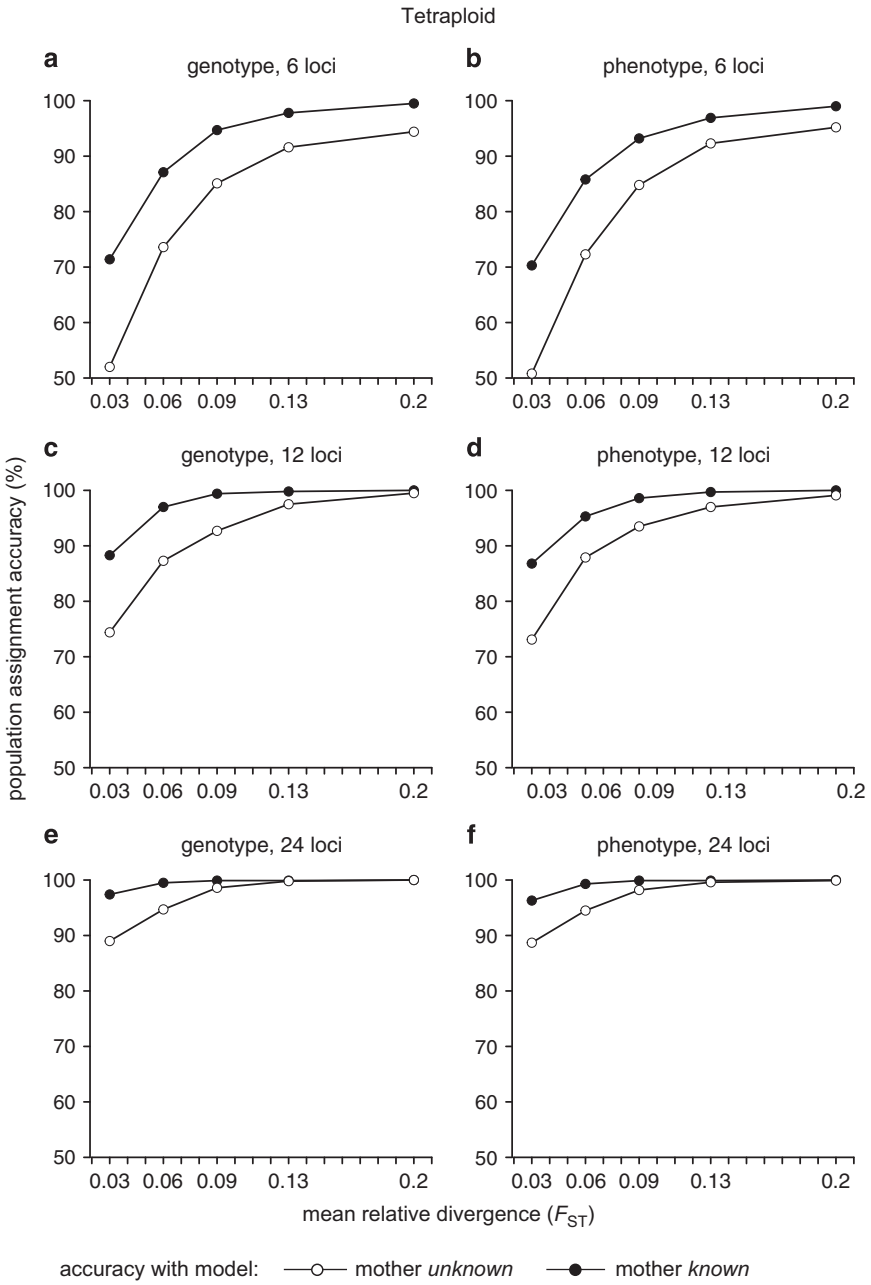

Figure 1 Simulations of autotetraploid populations showing the percentage of individuals assigned to the correct population for different levels of population differentiation (F_{ST}), number of loci (6, 12 and 24), marker types (genotype and phenotype) and model conditions (mother known and mother unknown). Indicated is the mean accuracy (percentage of individuals correctly assigned) ($n = 10$ replicates) for 10 000 simulated progeny when the mother information of each progeny is included in the analysis and when the mother information is not included. Individuals assigned at 95% confidence levels. The actual proportion of immigrants among the progeny was fixed to 50% for all replicates (see Methods section for more details).

71.7% and 63.1 % of individuals ($n = 10\,000$) could be assigned at 95% confidence using total error rates 0.01 and 0.03, respectively (Table 2). As found in the power testing simulations (Supplementary Information S2), using the correct DRR versus randomly drawing a new set of DRR for each locus (that is, DRR known versus unknown) had little impact on assignment (71.7% and 71.5% assigned, at 95% CI, respectively; Table 2).

With the actual *E. glabra* individuals, 57% and 77% of 467 individuals could be assigned at 95% and 90% confidence, respectively (with simulated error = 0.01). Using a 90% confidence threshold resulted in 2–8% higher immigration rates, suggesting that the strict threshold may reduce false positives at the expense of having fewer assigned individuals.

DISCUSSION

Ever since the pioneering theory on autotetraploids by Haldane (1930), investigating the population genetics of natural polyploid populations has remained an ongoing challenge for biologists. We provide a new framework for population assignment for autopolyploids that complements existing methods implemented in *Structure* (Falush *et al.*, 2007). The performance simulations imply that these new methods fill an important gap, enabling population assignment

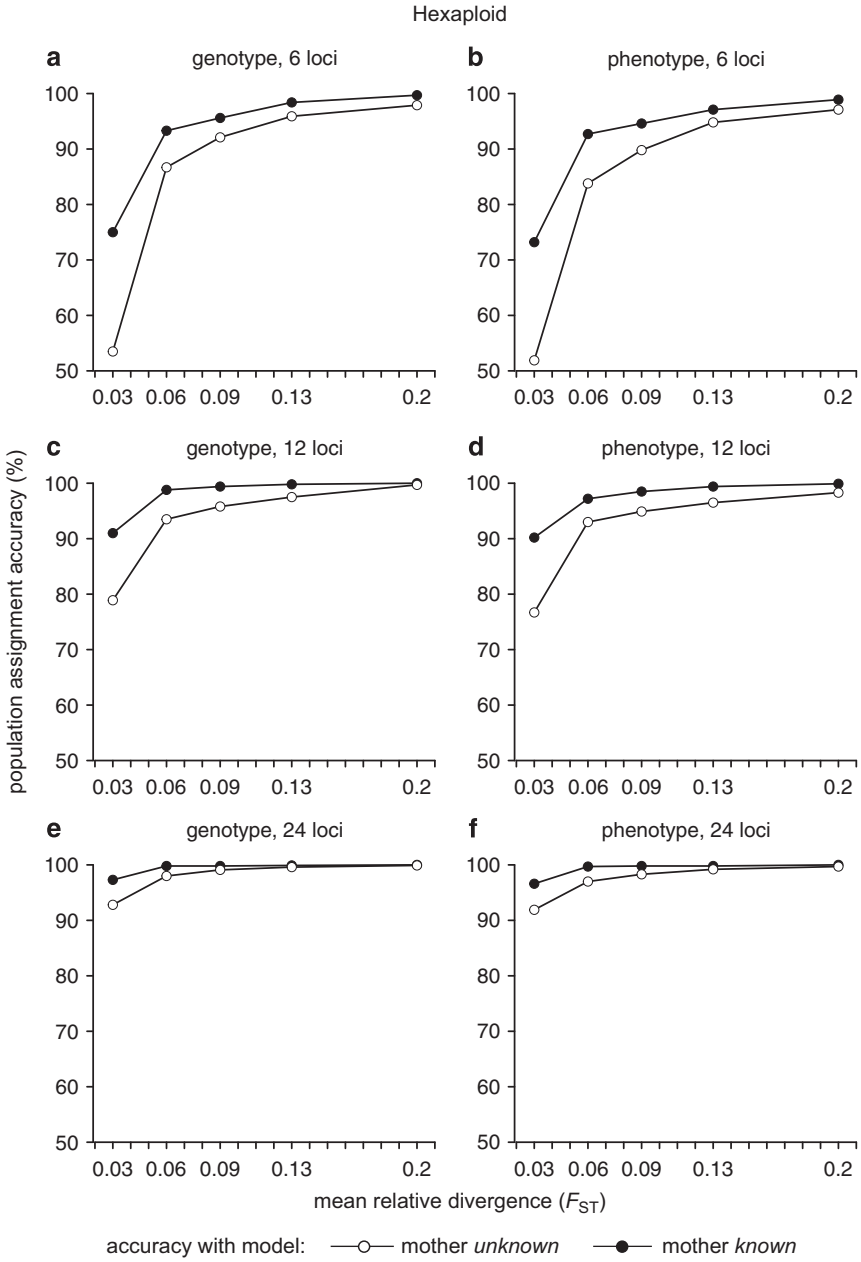

Figure 2 Simulations of autohexaploid populations showing the percentage of individuals assigned to the correct population for different levels of population differentiation (F_{ST}), number of loci (6, 12 and 24), marker types (genotype and phenotype) and model conditions (see Figure 1 legend for more details).

when population differentiation is low and when few polymorphic markers are available. We discuss the main factors that influence the performance of these methods and its application to empirical data. We then conclude by considering current challenges and future directions for population assignment in polyploids.

Accuracy of population assignment methods

Using a likelihood method that utilizes phenotypes and accounts for polysomic inheritance, we show that population assignment with microsatellite markers can reliably detect the origin of individuals or their gametes (for example, pollen). Based on the power simulations, knowledge of the degree of population differentiation, either F_{ST} or *Rho* (Ronfort *et al*., 1998), can be used to predict the performance of polyploid population assignment. Despite some inherent differences in allelic diversity found in polyploids, these simulations showed similar levels of performance to methods reported for diploids (Rannala and Mountain, 1997; Cornuet *et al*., 1999). When using maternal parent information, assignment accuracy is high (~95%) at relatively low F_{ST} (0.06) and with a modest number of loci (6). For plant studies, this will assist in the improvement of assignment accuracy of pollen dispersal, although assignment of seed will remain more challenging and may require genotyping more loci. Considering that investigations of contemporary dispersal patterns in plants often involve the collection of open pollinated seed arrays from known maternal parents, these results highlight the benefits of including maternal information for population assignment.

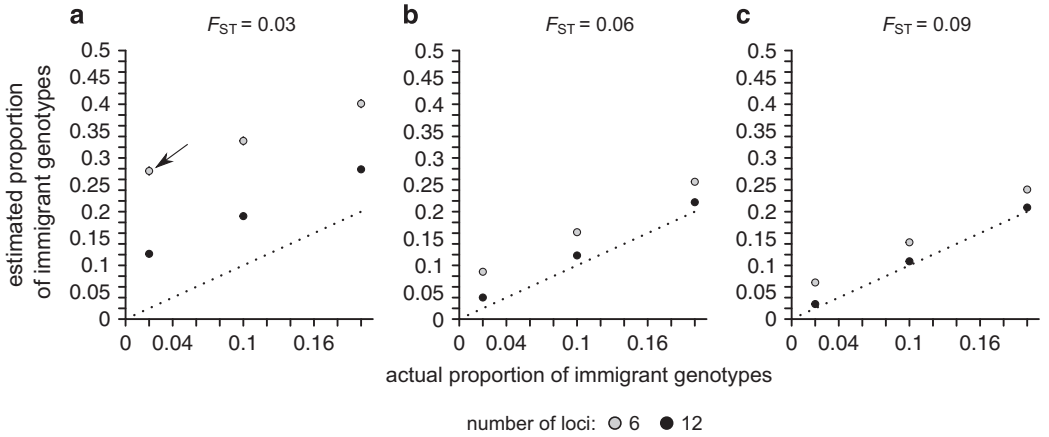

Figure 3 Simulations of autohexaploid populations showing the actual proportion of simulated immigrants between populations against the proportion estimated by *AutoPoly*. Shown are three different levels of population differentiation (F_{ST}) and two number of loci (6, 12) for phenotype data with mother known model. Arrow indicates the simulated population parameters most similar to *E. glabra* data. The dotted line indicates when simulated and estimated values are equal.

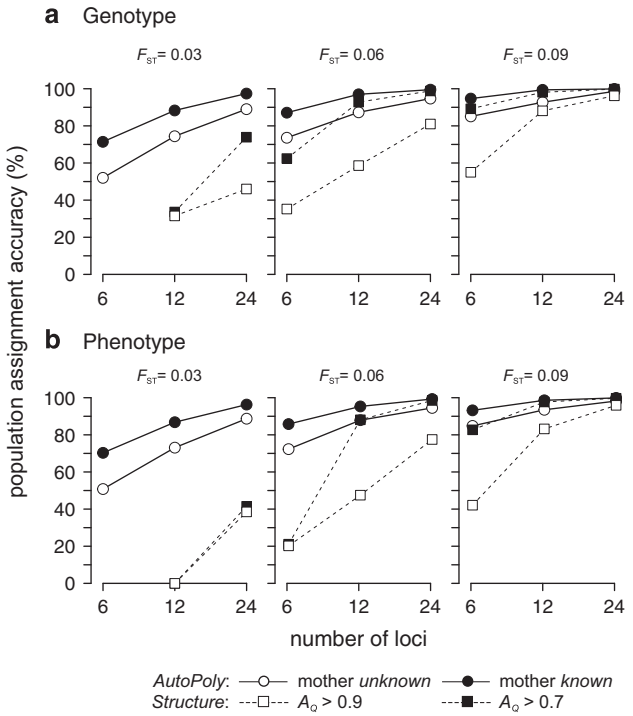

Figure 4 The percentage of simulated autotetraploid individuals assigned to the correct population using *AutoPoly* (circles) and *Structure* (squares) for (**a**) genotype (allele copy number known) and (**b**) phenotype markers (allele copy number unknown). Mean accuracy (percentage of individuals correctly assigned) ($n=10$ replicates) for 10 000 simulated progeny for each combination of parameters, including three levels of population differentiation (F_{ST}: 0.03, 0.06, 0.09) and number of loci (6, 12 and 24). For *AutoPoly*, individuals assigned at 95% confidence levels and the models tested include mother unknown (Models I and II; open circles) and mother known (Model III; filled circle). For *Structure*, we assigned each individual to their most likely source using thresholds of the membership coefficient Q of $A_Q > 0.9$ (open square) or $A_Q > 0.7$ (closed square) (see Methods section for details). The actual proportion of immigrants among the progeny was fixed to 20% for all replicates.

The methods we present for autopolyploids, like those of *Structure*, are not designed to specifically estimate migration rates. However, rough point estimates can be obtained by dividing the number of detected immigrants by the total sample size (Manel *et al.*, 2003). Although it may be tempting to obtain migration rates among polyploid populations with this method, our simulations predict that significant overestimates of migration rates will be obtained using point estimates under some scenarios (that is, 6 loci and $F_{ST}=0.03$). Nevertheless, genotyping more loci can substantially reduce the amount of bias. For example, at $F_{ST}=0.06$ and migration 2%, the ~8% overestimate at six loci reduces to 2% with 12 loci. Until further theory is developed that explicitly models migration rates for autopolyploids (for example, Wilson and Rannala, 2003), similar

Table 2 The percentage of simulated *Eremophila glabra* individuals assigned to each of the 15 populations at 95, 90 and 80% confidence

Error[a]	*DRR*[b]	*Strict (95%)*	*Moderate (90%)*	*Relaxed (80%)*
0.01	Correct	71.7	85.6	100
0.03	Correct	63.1	78.9	100
0.01	Random	71.5	85.4	100
0.03	Random	62.3	78.5	100

Abbreviation: DRR, double reduction rate.
[a]Total proportion of the simulated error (error $= e_1+e_2$) due to allelic drop out ($e_1 = 0.005$, 0.015) and mistyping ($e_2 = 0.005$, 0.015).
[b]DRR is either: (1) correct, we randomly chose a DRR for each locus from a uniform distribution ($0 \leqslant \beta \leqslant 0.2727$) and used the same values in calculation of assignment likelihoods, or (2) random, we randomly chose a DRR for each locus as in (1) but use a different random draw of DRR at each locus for the calculation of assignment likelihoods.

Table 3 An example of the generalized gametic segregation frequencies for the three possible biallelic hexaploid genotypes carrying two alleles ($k=2$)

Genotype	*Gametic frequencies*			
	aaa	*aab*	*abb*	*bbb*
a_5b	$\frac{3+\beta}{6}$	$\frac{3-2\beta}{6}$	$\frac{\beta}{6}$	—
a_4b_2	$\frac{3+3\beta}{15}$	$\frac{9-5\beta}{15}$	$\frac{3-\beta}{15}$	$\frac{\beta}{15}$
a_3b_3	$\frac{1+3\beta}{20}$	$\frac{9-3\beta}{20}$	$\frac{9-3\beta}{20}$	$\frac{1-3\beta}{20}$

The proportion of double reduced gametes $=\beta$. For the full set of gametic segregation frequencies, see Supplementary Table S7.

simulations will be required to assess the power of individual empirical data sets and the extent of the overestimation bias.

Polysomic inheritance, double reduction and unknown allele dosage

Contrary to expectations of substantially lower phenotype performance, we found only small differences between genotype and phenotype methods. Similarly, we found that the uncertainty in the DRR had little impact on the accuracy of population assignment. It is often assumed that these aspects are major limiting factors for population genetic analysis of polyploids. This has led to the development of methods to infer full genotypes by estimating the allele copy number from isozyme band intensity (Young and Brown, 1999) or electropherogram peak areas (Esselink *et al.*, 2004). Methods have also been developed to estimate allele frequencies from phenotype markers using maximum likelihood (De Silva *et al.*, 2005) or iterative-based procedures (Markwith *et al.*, 2006). These methods exhibit their own error and require prior information that may be unavailable (for example, selfing rate); this, together with the small difference we detected between phenotype and genotypes, suggests that our simple approach using marginal allele frequencies may be sufficient for likelihood-based population assignment.

AutoPoly versus Structure

Comparison of the two methods showed that, at low population differentiation, two to three times as many microsatellite loci may be required to use *Structure* compared with *AutoPoly*. Although they exhibited near identical levels of accuracy when population divergence was high, *AutoPoly* performed substantially better for most parameters at moderate ($F_{ST} = 0.06$ with < 12 loci) to low levels of divergence ($F_{ST} = 0.03$) and performed better with phenotypes. With phenotype markers and few loci, we observed greater variance in *Q* values and in some cases a lack of model convergence (that is, missing data points; Figures 4a and b). It seems unlikely that the lower accuracy of *Structure* with phenotypes and low information content is due to *AutoPoly* explicitly incorporating double reduction, as using the correct DRR had little impact on assignment accuracy. One possibility is that phenotype data with low F_{ST} makes it difficult for *Structure* to accurately estimate allele frequencies in each population while simultaneously assigning individuals to clusters. In contrast, for *AutoPoly* allele frequencies in each of the candidate populations are given, and the assignment directly comes from the probability of obtaining the phenotype given the individual is from each candidate population.

Empirical example

With the autohexaploid *E. glabra*, population assignment with six SSR markers using phenotypes with maternal information could identify the origin of about half of the offspring samples (at 95% confidence). This was less than the proportion predicted from simulating the actual *E. glabra* data, which suggested that 63% of the offspring could be assigned at 95% confidence. The simulations suggest that the level of population differentiation and number of loci currently available for the *E. glabra* data set contains too little power to estimate population assignment with high accuracy. This effect has been noted previously in diploid organisms (see Rannala and Mountain, 1997) and may be partially overcome by genotyping more loci. The decision on whether to generate more marker data or increase confidence thresholds (at the cost of fewer assignable individuals) will depend on the biological question and the importance of minimizing false positives versus false negatives.

Lower assignment success between the power simulations and the *Eremophila* data also suggests that some complexities encountered in natural populations may need to be incorporated into the theory. Contributing factors likely include higher marker error rates (including null alleles), the presence of close relatives and more variable population differentiation among natural populations. We also assume that all possible source populations have been sampled, populations are randomly mating and dispersal occurs randomly with respect to the surrounding populations. However, isolation by distance and recent bursts of dispersal may generate complex genetic compositions in the candidate reference data (that is, due to recent admixture) and the offspring pool. Incorrect assumptions on the demography and extent of relatedness among individuals can result in the mis-specification of the simulations and introduce bias into estimates of the CIs. Future efforts will be required to quantify which of these factors contribute most to the overall number of type I and type II errors and what level of migration we can expect to detect for a given set of population parameters and sample size.

CONCLUSIONS

Population assignment using genetic markers holds the promise of rapid estimation of contemporary dispersal patterns in natural populations. There are, however, significant challenges when applying these methods to natural polyploid populations. Further development of the theory may benefit from explicitly modelling error rates, double reduction and the probability of unsampled data (alleles and candidate populations). This could be achieved in a full Bayesian framework (for example, Hadfield *et al.*, 2006), although the computational burden of higher ploidy level and genotype uncertainty would make this a non-trivial task. As for diploids, the marker power as well as the sampling strategy will determine what level of accuracy can be achieved and how many individuals can be assigned with a high degree of confidence. Trade-offs exists between the number of individuals sampled for estimating allele frequencies and the number of individuals used to assess mating patterns (Meirmans, 2015). This will also depend on the ecology (for example, dispersal vector) of the organism and the demographic context in which each discrete population resides. When dispersal rates are low among populations, few immigrants will be generated, making it difficult to quantify migration rates without very large sample sizes (Manel *et al.*, 2003). We should therefore always remain cautious when inferring dispersal rates from genetic data and interpret these patterns using knowledge about the ecology and demography of the study organism.

DATA ARCHIVING

The *AutoPoly* R package that runs the population assignment methods described here (including documentation and example files) and empirical data sets are available at the Dryad Digital Repository (http://datadryad.org/), via http://dx.doi.org/10.5061/dryad.bc498. Updated versions of the program are also available from the Comprehensive R Archive Network, (https://cran.r-project.org), or https://github.com/dfield007/AutoPoly.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Alec Zwart for providing programming assistance; Lan Li for laboratory work and Tara Hopley, Freddie Loyman and Gina Leach for assistance in the field. We also thank Melinda Pickup, Tom Ellis and three anonymous reviewers for useful comments that improved the quality of the manuscript. This work was supported by a grant to AGY and LMB from Land and Water Australia.

Berry O, Tocher MD, Sarre SD (2004). Can assignment tests measure dispersal? *Mol Ecol* **13**: 551–561.
Bever J, Felber F (1992). The theoretical population genetics of autopolyploidy. In: Antonovics J, Futuyma D (eds). *Oxford Surveys in Evolutionary Biology*. Oxford University Press: Oxford. pp 185–217.
Buteler MI, Labonte DR, Macchiavelli RE (1997). Determining paternity in polyploids: hexaploid simulation studies. *Euphytica* **96**: 353–361.
Cain ML, Milligan BG, Strand AE (2000). Long-distance seed dispersal in plant populations. *Am J Bot* **87**: 1217–1227.
Clark LV, Jasieniuk M (2011). polysat: an R package for polyploid microsatellite analysis. *Mol Ecol Resour* **11**: 562–566.
Cornuet J-M, Piry S, Luikart G, Estoup A, Solignac M (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**: 1989–2000.
De Silva H, Hall A, Rikkerink E, McNeilage M, Fraser L (2005). Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity* **95**: 327–334.
Elliott CP (2009). Isolation and characterization of microsatellites in the bird-pollinated, autohexaploid, *Eremophila glabra* ssp. *glabra* (R.Br. (Ostenf.)) (Myoporaceae), an Australian endemic plant. *Mol Ecol Resour* **9**: 1242–1246.
Esselink GD, Nybom H, Vosman B (2004). Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting-peak ratios) method. *Theor Appl Genet* **109**: 402–408.
Falush D, Stephens M, Pritchard JK (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* **7**: 574–578.
Geiringer H (1949). Chromatid segregation of tetraploids and hexaploids. *Genetics* **34**: 665–684.
Hadfield JD, Richardson DS, Burke T (2006). Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol Ecol* **15**: 3715–3730.
Haldane JBS (1930). Theoretical genetics of autopolyploids. *J Genet* **22**: 359–372.
Luo ZW, Zhang Z, Zhang RM, Pandey M, Gailing O, Hattemer HH *et al.* (2006). Modeling population genetic data in autotetraploid species. *Genetics* **172**: 639–646.
Mable BK (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biol J Linn Soc* **82**: 453–466.
Manel S, Schwartz MK, Luikart G, Taberlet P (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol Evol* **18**: 189–197.
Markwith SH, Stewart DJ, Dyer JL (2006). TETRASAT: a program for the population analysis of allotetraploid microsatellite data. *Mol Ecol Notes* **6**: 586–589.
Mather K (1935). Reductional and equational separation of the chromosomes in bivalents and multivalents. *J Genet* **30**: 53–78.
Mather K (1936). Segregation and linkage in autotetraploids. *J Genet* **32**: 287–314.
Meagher TR (1986). Analysis of paternity within a natural population of *Chamaelirium luteum*. 1. Identification of most-likely male parents. *Am Nat* **128**: 199–215.
Meirmans PG (2015). Seven common mistakes in population genetics and how to avoid them. *Mol Ecol* **24**: 3223–3231.
Meirmans PG, Tienderen PHV (2013). The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* **110**: 131–137.
Meirmans PG, Van Tienderen P (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol* 792–794.
Moody ME, Mueller LD, Soltis DE (1993). Genetic variation and random drift in autotetraploid populations. *Genetics* **134**: 649–657.
Obbard D, Harris S, Pannell J (2006). Simple allelic-phenotype diversity and differentiation statistics for allopolyploids. *Heredity* **97**: 296–303.
Otto S, Whitton J (2000). Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401–437.
Paetkau D, Calvert W, Stirling I, Strobeck C (1995). Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* **4**: 347–354.
Paetkau D, Slade R, Burden M, Estoup A (2003). Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Mol Ecol* **13**: 55–65.
Ramsey J, Schemske DW (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rev Ecol Syst* **29**: 467–501.
Rannala B, Mountain JL (1997). Detecting immigration by using multilocus genotypes. *PNAS* **94**: 9197–9201.
Ronfort J, Jenczewski E, Bataillon T, Rousset F (1998). Analysis of population structure in autotetraploid species. *Genetics* **150**: 921–930.
Soltis DE, Soltis PS, Tate JA (2004). Advances in the study of polyploidy since Plant speciation. *New Phytol* **161**: 173–191.
Stift M, Berenos C, Kuperus P, van Tienderen PH (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to Rorippa (Yellow Cress) microsatellite data. *Genetics* **179**: 2113–2123.
Wilson GA, Rannala B (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**: 1177–1191.
Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH (2009). The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci* **106**: 13875–13879.
Wricke G, Weber W (1986). *Quantitative Genetics and Selection in Plant Breeding*. de Gruyter: Berlin, Germany.
Wu R, Gallo-Meagher M, Littell RC, Zeng ZB (2001). A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. *Genetics* **159**: 869–882.
Young A, Brown AH (1999). Genetic structure of fragmented populations of the endangered daisy *Rutidosis leptorrhynchoides*. *Conserv Biol* **13**: 256–265.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)

APPENDIX A1

We use a similar approach that Wricke and Weber (1986) used for tetraploids, to derive general formulas for expected genotype frequencies in an autohexaploid at RME. First, we generated a pair-wise multiplicative matrix (Table A1) between each of the possible gametes using the general formulas above (Equation 4). Next, we obtain the sum of each of the off-diagonal elements. For example, the genotype a_4b_2 can arise through the union of either the gametes, (i) x_{111} (*aaa*) and $3y_{122}$ (*abb*) or (ii) $3y_{112}$ (*aab*) and $3y_{112}$ (*aab*). The sum of these frequencies gives us the expected genotype frequency of each (genotype) class given the gametic frequencies at RME. For this example, this equates to the sum of the these terms (see elements in bold, Table A1) as

$$\begin{aligned}\Pr\left(a_4b_2|i,\boldsymbol{g}_{ij},Y_6\right) &= 3x_{111}y_{122}+3x_{111}y_{112}+9(y_{112})^2\\ &= 6x_{111}y_{112}+9(y_{112})^2\end{aligned}$$

Table A1 Genotype frequencies at random mating equilibrium for a hexaploid population, an example of the approach with $k=2$ alleles

	x_{111} (aaa)	$3y_{112}$ (aab)	$3y_{122}$ (abb)	x_{222} (abb)
x_{111} (*aaa*)	$(x_{111})^2$ (*aaaaaa*)	$3x_{111}y_{112}$ (*aaaaab*)	**$3x_{111}y_{122}$ (*aaaabb*)**	$x_{111}x_{222}$ (*aaabbb*)
$3y_{112}$ (*aab*)	$3x_{111}y_{112}$ (*aaaaab*)	**$9(y_{112})^2$ (*aaaabb*)**	$9y_{112}y_{122}$ (*aaabbb*)	$3y_{112}x_{222}$ (*aaabbb*)
$3y_{122}$ (*abb*)	**$3x_{111}y_{122}$ (*aaaabb*)**	$9y_{112}y_{122}$ (*aaabbb*)	$9(y_{122})^2$ (*aabbbb*)	$3y_{112}x_{222}$ (*aabbbb*)
x_{222} (*abb*)	$x_{111}x_{222}$ (*aaabbb*)	$3x_{222}y_{112}$ (*aaabbb*)	$3x_{222}y_{122}$ (*aabbbb*)	$(x_{222})^2$ (*bbbbbb*)

All elements involved in generating genotype a_4b_2 are listed in bold.

APPENDIX A2

Using the same principles described for tetraploids, the probability of obtaining mixed genotypes for hexaploid individuals is $\Pr\left(G_{[i,i']jm}|g_{ij},g_{i'j},Y_6\right)$, where the probabilities for each genotype class are

$$\Pr\left(a_{6[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right)=\left(x_{111[i]}x_{111[i']}+x_{111[i']}x_{111[i]}\right)/2$$

$$\Pr\left(a_5b_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right)=3\left(x_{111[i]}y_{112[i']}+x_{111[i']}y_{112[i]}\right)$$

$$\Pr\left(a_4b_{2[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right)=3\left(x_{111[i]}y_{122[i']}+x_{111[i']}y_{122[i]}+3y_{112[i]}y_{112[i']}\right)$$

$$\begin{aligned}\Pr\left(a_3b_{3[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right) &= x_{111[i]}x_{222[i']}+x_{111[i']}x_{222[i]}\\ &\quad+9\left(y_{112[i]}y_{122[i']}+y_{112[i']}y_{122[i]}\right)\end{aligned}$$

$$\begin{aligned}\Pr\left(a_4bc_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right) &= 6\left(x_{111[i]}z_{123[i']}+x_{111[i']}z_{123[i]}\right)\\ &\quad+9\left(y_{112[i]}y_{113[i']}+y_{112[i']}y_{113[i]}\right)\end{aligned}$$

$$\begin{aligned}\Pr\left(a_3b_2c_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6\right) &= 3\left(x_{111[i]}y_{223[i']}+x_{111[i']}y_{223[i]}\right)\\ &\quad+18\left(y_{112[i]}z_{123[i']}+y_{112[i']}z_{123[i]}\right)\\ &\quad+9\left(y_{113[i]}y_{122[i']}+y_{113[i']}y_{122[i]}\right)\end{aligned}$$

$$\begin{aligned}\Pr(a_2b_2c_{2[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6) &= 9(y_{112[i]}y_{233[i']}+y_{112[i']}y_{233[i]}\\ &\quad+y_{133[i]}y_{122[i']}+y_{133[i']}y_{122[i]}\\ &\quad+y_{113[i]}y_{223[i']}+y_{113[i']}y_{223[i]}\\ &\quad+2z_{123[i]}z_{123[i']})\end{aligned}$$

$$\begin{aligned}\Pr(a_3bcd_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6) &= 6(x_{111[i]}z_{234[i']}+x_{111[i']}z_{234[i]})\\ &\quad+18(y_{112[i]}z_{134[i']}+y_{112[i']}z_{134[i]}\\ &\quad+y_{113[i]}z_{124[i']}+y_{113[i']}z_{124[i]}\\ &\quad+y_{114[i]}z_{123[i']}+y_{114[i']}z_{123[i]})\end{aligned}$$

$$\begin{aligned}\Pr(a_2b_2cd_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6) &= 18y_{112[i]}z_{234[i']}+18y_{112[i']}z_{234[i]}\\ &\quad+9y_{113[i]}y_{224[i']}+9y_{113[i']}y_{224[i]}\\ &\quad+9y_{114[i]}y_{223[i']}+9y_{114[i']}y_{223[i]}\\ &\quad+36z_{123[i]}z_{124[i']}+36z_{123[i']}z_{124[i]}\\ &\quad+18z_{134[i]}y_{122[i']}+18z_{134[i']}y_{122[i]}\end{aligned}$$

$$\begin{aligned}\Pr(a_2bcde_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6) &= 18(y_{112[i]}z_{345[i']}+y_{112[i']}z_{345[i]}\\ &\quad+y_{113[i]}z_{245[i']}+y_{113[i']}z_{245[i]}\\ &\quad+y_{114[i]}z_{235[i']}+y_{114[i']}z_{235[i]}\\ &\quad+y_{115[i]}z_{234[i']}+y_{115[i']}z_{234[i]})\\ &\quad+36(z_{123[i]}z_{145[i']}+z_{123[i']}z_{145[i]}\\ &\quad+z_{134[i]}z_{125[i']}+z_{134[i']}z_{125[i]}\\ &\quad+z_{124[i]}z_{135[i']}+z_{124[i']}z_{135[i]})\end{aligned}$$

$$\begin{aligned}&\Pr(abcdef_{[i,i']jm}|i,i',g_{ij},g_{i'j},Y_6)\\ &\quad=36(z_{123[i]}z_{456[i']}+z_{123[i']}z_{456[i]}+z_{124[i]}z_{356[i']}\\ &\qquad+z_{124[i']}z_{356[i]}+z_{125[i]}z_{346[i']}+z_{125[i']}z_{346[i]}\\ &\qquad+z_{126[i]}z_{345[i']}+z_{126[i']}z_{345[i]}+z_{134[i]}z_{256[i']}\\ &\qquad+z_{134[i']}z_{256[i]}+z_{135[i]}z_{246[i']}+z_{135[i']}z_{246[i]}\\ &\qquad+z_{136[i]}z_{245[i']}+z_{136[i']}z_{245[i]}+z_{145[i]}z_{236[i']}\\ &\qquad+z_{145[i']}z_{236[i]}+z_{146[i]}z_{235[i']}+z_{146[i']}z_{235[i]}\\ &\qquad+z_{234[i]}z_{156[i']}+z_{234[i']}z_{156[i]})\end{aligned}\tag{A1}$$

APPENDIX A3

For an autohexaploid, the number of possible gametes under RCeS that can originate at a given locus is given by drawing without replacement three alleles from a set of six by simply using the binomial coefficient $\binom{6}{3}=20$. However, for RCeD there are many possible gametes due to random chromatid assortment. In this case, we draw

without replacement 3 alleles from 12 to account for the independence of sister chromatids $\binom{12}{3} = 220$. The expected frequency of each gamete from a given genotype then follows the hypergeometric multinomial distribution as:

$$\frac{\binom{K_1}{k_1}\binom{K_2}{k_2}\binom{K_3}{k_3}\binom{K_4}{k_4}\binom{K_5}{k_5}\binom{K_6}{k_6}}{\binom{C}{n}} \quad \text{(A2)}$$

Where the number of copies of the ith allele in the parental genotype is K_i and the number of the ith allele present in the gamete is k_i. For RCeS: $k_i \leqslant K_i$, and $\sum K_i = 6$ and $\sum k_i = 3$, while for RCdS: $\sum K = 12$ and $\sum k = 3$. The values C and n are as described above for a binomial coefficient.

To demonstrate how to derive the general formulas for β, we use a hypothetical individual hexaploid at a single locus with the genotype, a_4bc. Assuming complete random chromosome segregation (RCeS), there are four possible unique gametes (*aaa*, *aab*, *aac*, *abc*), whereas for complete random chromatid segregation (RCdS), three unreduced gametes (*aaa*, *aab*, *aac*) and five double reduced gametes are possible (*abb*, *acc*, *bbc*, *cca*, *ccb*). We determine the frequency of each gamete under complete RCeS and the frequency under complete RCdS assuming only double reduced gametes were formed (Equation A2). Generalized formulas are the sum of these terms multiplied by $1-\beta$ and β, respectively, to obtain the segregation ratio for any given DRR. For example, for the phenotype, *AB*, (which could be either genotype a_5b, a_4b_2, a_3b_3, a_2b_4 or ab_5), let us assume the genotype is known to be a_3b_3. Using Equation 9, we find the expected frequency of the gamete a_3 under RCeS and RCdS are $(1-\beta)/20$, and $24/120(\beta)$, respectively. Generalized forms are equal to the sum of these terms (that is, $(1+3\beta)/20$). For this example, there are 22 possible gametic segregation ratios for 4 possible genotypes and gametes. We followed this procedure to determine the generalized formulas for all possible gametes from each of the possible (in Table 1) hexaploid genotypes (for the complete segregation matrix, see Supplementary Table S8). For autotetraploids, we used the general formulas for segregation ratios available elsewhere (Bever and Felber, 1992).