






Inference of individual ploidy level using codominant markers

Kang Huang¹  | Derek W. Dunn¹  | Zhonghu Li¹  | Pei Zhang¹  | Yu Dai¹ |
Baoguo Li^{1,2} ¹Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an, China²Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

Correspondence

Baoguo Li, Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an 710,069, China.
Email: baoguoli@nwnu.edu.cn

Funding information

Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDB31020302; National Natural Science Foundation of China, Grant/Award Number: 31770411, 31730104, 31572278 and 31770425; Young Elite Scientists Sponsorship Program by CAST, Grant/Award Number: 2017QNRC001; National Key Programme of Research and Development, Ministry of Science and Technology, Grant/Award Number: 2016YFC0503200; Natural Science Basic Research Plan in Shaanxi Province of China, Grant/Award Number: 2018JM3024 and 2019JM258

Abstract

A significant portion of plant species are polyploids, with ploidy levels sometimes varying among individuals and/or populations. Current techniques to determine the individual ploidy, e.g., flow cytometry, chromosome counting or genotyping-by-sequencing, are often cumbersome. Based on the genotypic probabilities for polysomic inheritance under double-reduction, we developed a model to estimate allele frequency and infer the ploidy status of individuals from the allelic phenotypes of codominant genetic markers. The allele frequencies are estimated by an expectation-maximization algorithm in the presence of null alleles, false alleles, negative amplifications and self-fertilization, and the posterior probabilities are used to assign individuals into different levels of ploidy. The accuracy of this method under different conditions is evaluated. Our methods are freely available in a new software package, PLOIDYINFER, for use by other researchers which can be downloaded from <http://github.com/huangkang1987/ploidyinfer>.

KEYWORDS

allele frequency estimation, expectation-maximization algorithm, maximum-likelihood, null alleles, ploidy assignment

1 | INTRODUCTION

A significant portion of plant species are polyploids, with 24% of all plant taxa being polyploid (Barker, Arrigo, Baniaga, Li, & Levin, 2016), and 47% to 100% of angiosperm species traced to a polyploidy event in the past (Wood et al., 2009). Ploidy levels sometimes vary among individuals or populations (Gompert & Mock, 2017), with mixed-ploidy populations also occurring (e.g., *Spartina pectinata* [Kim, Rayburn, Boe, & Lee, 2012], *Chamerion angustifolium* [Martin & Husband, 2013], *Paspalum simplex* [Brugnoli, Urbani, Quarin, Martínez, & Acuña, 2013], *Aster amellus* [Castro, Loureiro, Procházka, & Münzbergová, 2012], *Lycoris radiata* [Liu, Xia, Zheng, Zhi, & Zhou, 2015]).

When such variation exists, determination of individual ploidy level can inform studies of ecology or trait variation and is required

for subsequent population genetic analyses (Gompert & Mock, 2017). Although the ploidy status can be determined by cytotype screening with flow cytometry or chromosome counting using microscopy (Peirson, Reznicek, & Semple, 2012), flow cytometry usually requires fresh samples, while chromosome counting requires root-tip squashes. Both methods are thus cumbersome and may be impractical under certain conditions. An R package, POLYSAT (Clark & Jasieniuk, 2011), provides a method to predict individual ploidy from the maximum number of unique identical-by-state (IBS) alleles within an individual. Similar methods have been described by Bisognin et al. (2009) and Robertson, Rich, and Allen (2010). However, the use of such methods may underestimate the true degree of ploidy when only a few alleles are present or if low polymorphic loci are used, e.g., biallelic markers. Gompert and Mock (2017) developed a method to detect individual ploidy levels with

genotyping-by-sequencing (GBS) analysis, which cannot be applied to multiallelic codominant markers.

This paper focuses on autopolyploids that display polysomic inheritance. In polysomic inheritance, more than two homologous chromosomes can pair at meiosis, resulting in the formation of multivalents and polysomic inheritance (Rieger, Michaelis, & Green, 1968). A peculiarity of polysomic inheritance is the possibility that a gamete inherits a single gene copy twice, termed double-reduction (Butruille & Boiteux, 2000). For example, an autotetraploid individual ABCD produces a gamete AA. Double-reduction results in gametes carrying identical-by-descent (IBD) alleles and increased homozygosity (Hardy, 2016). Double-reduction has a similar effect as inbreeding and can cause a deviation in genotypic probabilities from the Hardy-Weinberg equilibrium (HWE) (Luo et al., 2006). A brief description of double-reduction models, including random chromosome segregation (RCS), pure random chromatid segregation (PRCS), complete equational segregation (CES) and partial equational segregation (PES), can be found in Appendix S1.

In plants, mating systems often concur with the 'mixed mating model', in which a fraction of progeny are derived from self-fertilization and the remainder are derived from outcrossing (Ritland, 2002). Self-fertilization is the fusion of male and female gametes from a single genetic individual or colony. This can occur in many plant species and in some hermaphrodite animal species (Jarne & Charlesworth, 1993). Self-fertilization is an extreme case of inbreeding and results in the loss of genetic variation within individuals. When the selfing rate is high, many genetic loci become homozygous, which can lead to underestimation of ploidy level.

When the ploidy level is unknown, the genotyping can only obtain a set of alleles (or electrophoresis band type), defined here as a phenotype. As a result of reproductive isolation, there is usually a genetic differentiation among different levels of ploidy. Moreover, different ploidies have different inheritance models (e.g., disomic, tetrasomic), and the distribution of phenotypes will also differ among different ploidy levels. Both mechanisms can result in differences in the distributions of phenotypes among different ploidy levels, which can be used to infer the ploidy level.

Although SNPs are increasingly used, microsatellites are still important markers in population genetics. They are codominant and highly polymorphic. However, two major problems, null alleles and false alleles, may be present and result in either an under- or over-estimation of ploidy level. Null alleles are alleles that cannot be detected because of mutation, often within one of the primer binding sites (Brookfield, 1996). Null alleles are pervasive in microsatellite markers (Kalinowski, Wagner, & Taper, 2006; Ravinet et al., 2016). A false allele is the amplification of a polymerase chain reaction (PCR) artifact (Taberlet et al., 1996) and can be caused by slippage mutations (Schlötterer & Tautz, 1992) and primers annealing off-target (Yang, Chen, Chamberlin, & Benner, 2010).

In this study, we developed a method to estimate allele frequency and to infer the individual ploidy level based on phenotypes under polysomic inheritance, double-reduction, null alleles, negative amplification, false alleles and selfing. A Windows-based, user-friendly

software package, PLOIDYINFER, that incorporates this new method and supports a maximum ploidy level of 10 is available at <http://github.com/huangkang1987/ploidyinfer>.

2 | MATERIALS AND METHODS

2.1 | Rationale

In this study, we assumed that our model inferring the ploidy assignment of an individual from its phenotypes satisfied the following five conditions: (a) the population size of each ploidy level is large enough; (b) mating is random within each ploidy level, and a proportion of the offspring are produced by self-fertilization; (c) there is no subdivision within each ploidy level, and the distribution of genotypes is at the equilibrium state; (d) the genetic markers used are autosomal, codominant and unlinked; and (e) the candidate ploidy levels are known.

The set consisting of the alleles detected in an individual at a locus is called a phenotype of this individual, denoted by \mathcal{P} . The symbol \mathcal{P} is used to denote the collection consisting of all the phenotypes of this individual at L loci. A phenotype may contain no alleles due to null allele homozygotes or amplification failure. In such cases the phenotype is termed negative, and the symbol \emptyset is used for this.

The posteriori probability that an individual is at the ploidy level v is the assignment rate of this individual at the ploidy level v , or a ploidy assignment rate for abbreviation, written as $\Pr(v|\mathcal{P})$, which can be expressed as the following Bayes formula:

$$\Pr(v|\mathcal{P}) \propto \Pr(\mathcal{P}|v) \Pr(v), \quad (1)$$

where $\Pr(v)$ is the prior probability of the ploidy level v , and $\Pr(\mathcal{P}|v)$ is the multi-locus phenotypic probability, whose expression is.

$$\Pr(\mathcal{P}|v) = \prod_{l=1}^L \Pr(\mathcal{P}_l|v), \quad (2)$$

where \mathcal{P}_l is the l^{th} component in \mathcal{P} , i.e. the phenotype at the l^{th} locus, and $\Pr(\mathcal{P}_l|v)$ is the probability of \mathcal{P}_l at ploidy level v .

Because all possible ploidy levels are assumed to be known, the sum $\sum_{\lambda=1}^{\Lambda} \Pr(v_{\lambda}|\mathcal{P})$ of all ploidy assignment rates is equal to one,

where Λ is the number of candidate ploidy levels. From Equations (1) and (2), to obtain the value of $\Pr(v|\mathcal{P})$, we need only consider how to calculate the values of $\Pr(\mathcal{P}_1|v)$, $\Pr(\mathcal{P}_2|v)$, ..., s . This problem is closely related to the allele frequencies, and are discussed below.

We adopted an iterative algorithm to estimate the ploidy assignment rates. In this process of estimation, an expectation-maximization (EM) algorithm was inserted to estimate the allele frequencies. The EM algorithm is explained in the section Allele frequency estimator. Here, we give an outline of the procedure of our iterative algorithm.

- (i) Let $\hat{\mathbf{Q}}_j (1 \leq j \leq N)$ be the vector consisting of the current ploidy assignment rates of the j^{th} individual, i.e., $\hat{\mathbf{Q}}_j = [\hat{\Pr}(v_1 | \mathcal{P}_j), \hat{\Pr}(v_2 | \mathcal{P}_j), \dots, \hat{\Pr}(v_\Lambda | \mathcal{P}_j)]$, whose initial vector is assumed as $[\Pr(v_1), \Pr(v_2), \dots, \Pr(v_\Lambda)]$. N is the total number of individuals.
- (ii) Estimate the allele frequencies at each ploidy level and at each locus by an EM algorithm based on $\hat{\mathbf{Q}}_j$ to calculate the probabilities of each \mathcal{P}_{ji} in \mathcal{P}_j for each candidate ploidy level, i.e., $\Pr(\mathcal{P}_{ji} | v_\Lambda)$.
- (iii) Calculate the updated ploidy assignment rates in $\hat{\mathbf{Q}}'_j$ by Equations (1) and (2), where $\hat{\mathbf{Q}}'_j = [\hat{\Pr}'(v_1 | \mathcal{P}_j), \hat{\Pr}'(v_2 | \mathcal{P}_j), \dots, \hat{\Pr}'(v_\Lambda | \mathcal{P}_j)]$.
- (iv) If the absolute value $\max_{1 \leq j \leq N} (|\hat{\mathbf{Q}}_j - \hat{\mathbf{Q}}'_j|)$ is greater than or equal to a predefined threshold (e.g., 10^{-5}), then substitute $\hat{\mathbf{Q}}_j$ with $\hat{\mathbf{Q}}'_j$ and then repeat steps (ii) and (iii). Otherwise, terminate this process.

In the above process, the final values of $\hat{\mathbf{Q}}_j$ are the estimates of the ploidy assignment rates of the j^{th} individual at all ploidy levels, and the allele frequencies can also be estimated. The schematic diagram of the two algorithms of our model is shown in (Figure 1).

2.2 | Genotypic and phenotypic probabilities

In polyploids, some genotypes share a common electrophoretic band type for codominant markers (e.g., microsatellites). For a PCR-based marker, the dosage of alleles cannot be determined, and the true genotype is unavailable, so a heterozygous phenotype represents multiple classes of genotypes. In the presence of null alleles, a homozygous phenotype also represents multiple candidate genotypes. In addition to null alleles, the occurrence of a negative phenotype may be due to other factors, such as a low DNA quality or an experimental error (Kalinowski et al., 2006). The probability that a phenotype becomes negative due to these other factors is called the negative amplification rate, denoted by β .

We used the multiset consisting of the alleles within a genotype G to denote G itself. We used the symbol \mathcal{P}^* to denote the set consisting of the visible alleles within G , which is the set concerning the effect of null alleles. For example, if the ploidy level is two, then,

$$\mathcal{P}^* = \begin{cases} AB & \text{if } G = AB \\ A & \text{if } G = AA \text{ or } AY \\ \emptyset & \text{if } G = YY \end{cases}$$

where A and B are distinct visible IBS alleles, Y is the null allele, and AA , AB , AY and YY are the genotype pattern. To further account for false alleles, we assumed each visible allele (e.g., A) outside \mathcal{P}^* could erroneously be amplified at a probability of e or can be correctly unamplified at a probability of $1-e$, where e is defined as the false allele rate. The set of observed alleles without considering any negative amplification (e.g., \mathcal{P}') will be a superset of or equal to \mathcal{P}^* . The transitional probability from \mathcal{P}^* to \mathcal{P}' is,

$$T(\mathcal{P}' | \mathcal{P}^*) = \begin{cases} \prod_{A \in \mathcal{P}'} (\delta_{A \in \mathcal{P}^*} + \delta_{A \notin \mathcal{P}^*} e) \prod_{A \notin \mathcal{P}'} (1-e) & \text{if } \mathcal{P}' \supseteq \mathcal{P}^*, \\ 0 & \text{if } \mathcal{P}' \not\supseteq \mathcal{P}^*, \end{cases}$$

where δ_X is a Boolean variable, which equals one if the expression X is true and zero otherwise.

To model negative amplification, we assumed the observed phenotype \mathcal{P} will randomly become \emptyset at a probability of β . The probability of a phenotype \mathcal{P} , conditional on the true genotype G , can be expressed as.

$$\Pr(\mathcal{P} | G) = \beta \delta_{\mathcal{P}=\emptyset} + (1-\beta) T(\mathcal{P}' = \mathcal{P} | \mathcal{P}^*), \quad (3)$$

In the presences of null alleles and negative amplification, the probability of a phenotype \mathcal{P} at a ploidy level v can be expressed as.

$$\Pr(\mathcal{P} | v) = \sum_G \Pr(G) \Pr(\mathcal{P} | G), \quad (4)$$

where G is taken from all possible genotypes at v , $\Pr(G)$ is the probability of G , and the value of $\Pr(\mathcal{P} | G)$ can be obtained from Equation (3).

With the Bayes formula, the posterior probability of G given by observing a phenotype \mathcal{P} is as follows:

$$\Pr(G | \mathcal{P}) = \frac{\Pr(\mathcal{P} | G) \Pr(G)}{\Pr(\mathcal{P} | v)}. \quad (5)$$

Under the condition that the allele frequencies are known, the probability $\Pr(G)$ can be calculated for a certain double-reduction model at the equilibrium state (Appendix S2).

2.3 | Selfing model

Hardy (2016) developed two method-of-moment estimators to estimate the selfing rate for polyploids. These estimators are based on single-locus heterozygosity and allele frequencies. For phenotypic data, both parameters are estimated from the phenotypes assuming each allele dosage configuration and each allele are of the same weight. Because the phenotypic probabilities and the posterior probabilities of hidden genotypes are influenced by self-fertilization, the estimation of both heterozygosity and allele frequencies will be inaccurate without the inclusion of self-fertilization.

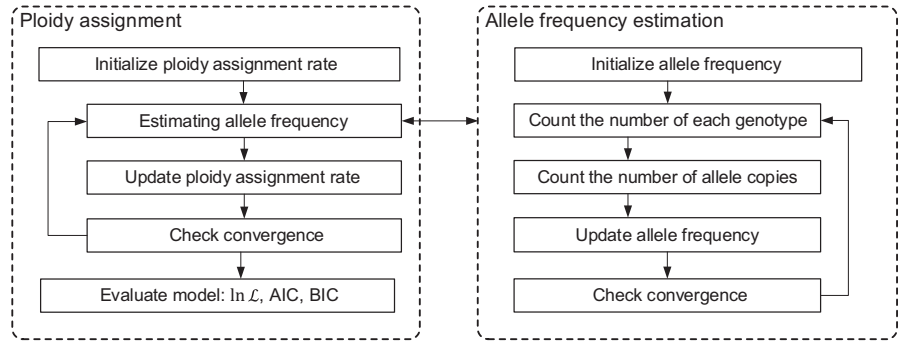
In the following, we incorporated self-fertilization into our model to simultaneously estimate the allele frequencies and selfing rate.

In the presence of selfing, the solving of genotypic probabilities under double-reduction is computationally difficult (Appendix S3). We therefore used an alternative method to obtain an approximate solution, using the inbreeding coefficient F as an intermediate variable and by assuming that inbreeding is only caused by selfing. The value of F at the equilibrium state under double-reduction and self-fertilization has been derived previously (Huang et al., 2019):

$$F = \frac{8\alpha + sv}{8\alpha + v(s + v - sv)},$$

where $\alpha = \sum_i i \alpha_i$, and α_i denotes the probability that a gamete carries i pairs of identical-by-double-reduction (IBDR) alleles.

FIGURE 1 The schematic diagram of the two algorithms used in our model



We used a Dirichlet distribution to approximate the genotypic probability under inbreeding and double-reduction. Assume the allele frequencies within an individual $\mathbf{q} = [q_1, q_2, \dots, q_K]$ are drawn from a Dirichlet distribution $D(\gamma_1, \gamma_2, \dots, \gamma_K)$ (Pritchard, Stephens, & Donnelly, 2000), where $\gamma_k = (1/F - 1)p_k$ and p_k is the frequency of the k^{th} allele in the reference population. The probability density function of \mathbf{q} is.

$$f(\mathbf{q}|\mathbf{p}, F) = \Gamma(\gamma) \prod_{k=1}^K \frac{q_k^{\gamma_k-1}}{\Gamma(\gamma_k)}.$$

Here, $\gamma = 1/F - 1$, and \mathbf{p} denotes the vector of allele frequencies in the reference population, where $\mathbf{p} = [p_1, p_2, \dots, p_K]$, so that the expected value and variance of each q_k is p_k and $Fp_k(1-p_k)$, respectively.

Because the correlation between alleles within the same individuals relative to the reference population have been explained by the divergence between \mathbf{p} and \mathbf{q} , the alleles are independent of the background of \mathbf{q} . Therefore, the genotypic probability, conditional on \mathbf{q} , is given by.

$$\Pr(G|\mathbf{q}) = \binom{v}{n_1, n_2, \dots, n_K} \prod_{k=1}^K q_k^{n_k}$$

where n_k is the number of copies of the k^{th} allele in G , and $\binom{v}{n_1, n_2, \dots, n_K}$

is the multinomial coefficient. The genotypic probability conditional on \mathbf{q} and F can be derived by taking the weighted average of $\Pr(G|\mathbf{q})$ using $f(\mathbf{q}|\mathbf{p}, F)$ as the weight:

$$\Pr(G|\mathbf{p}, F) = \int_{\Omega} \Pr(G|\mathbf{q}) f(\mathbf{q}|\mathbf{p}, F) d\mathbf{q},$$

where the integral domain Ω can be expressed as $\Omega = \{(q_1, q_2, \dots, q_K) | q_1 + q_2 + \dots + q_K = 1, q_k \geq 0, k = 1, 2, \dots, K\}$. Such an integral can be converted into the following repeated integral with the multiplicity $K-1$:

$$\begin{aligned} \Pr(G|\mathbf{p}, F) &= \int_0^1 \int_0^{1-q_1} \dots \int_0^{1-q_1-q_2-\dots-q_{K-2}} \Pr(G|\mathbf{q}) f(\mathbf{q}|\mathbf{p}, F) dq_1 dq_2 \dots dq_{K-1} \\ &= \binom{v}{n_1, n_2, \dots, n_K} \frac{\Gamma(\gamma)}{\Gamma(\gamma+v)} \prod_{k=1}^K \frac{\Gamma(\gamma_k+n_k)}{\Gamma(\gamma_k)} \\ &= \binom{v}{n_1, n_2, \dots, n_K} \left[\prod_{k=1}^K \prod_{j=0}^{n_k-1} (\gamma_k+j) \right] / \prod_{j=0}^{v-1} (\gamma+j). \end{aligned} \quad (6)$$

2.4 | Allele frequency estimator

In this section, based on the current ploidy assignment rates of each individual mentioned in step (i) of our iterative algorithm, we adopted an EM algorithm (Dempster, Laird, & Rubin, 1977) to estimate the allele frequencies. Conversely, the results of the estimation were also used for our iterative algorithm to estimate the ploidy assignment rates. Our EM algorithm followed the method developed by Kalinowski and Taper (2006), which is also an iterative algorithm and can be used to maximize the genotypic likelihood as described below.

The product of genotypic probabilities of all individuals for a given ploidy level v and at a target locus is called the genotypic likelihood at v , denoted by $\mathcal{L}_{\text{geno}}(v)$, whose logarithmic expression is as follows:

$$\mathcal{L}_{\text{geno}}(v) = \sum_i \ln[\Pr(G_i)] \sum_{j=1}^N \Pr(v|\mathcal{P}_j) \Pr(G_i|\mathcal{P}_j),$$

where G_i is taken from all possible genotypes of \mathcal{P}_j , \mathcal{P}_j is the phenotypic vector of the j^{th} individual, and \mathcal{P}_j is the phenotype of the j^{th} individual. The sum $\sum_{j=1}^N \Pr(v|\mathcal{P}_j) \Pr(G_i|\mathcal{P}_j)$ represents the number of genotypes G_i in the population conditional on the current allele frequencies.

In our EM algorithm, we began with a trivial initial allele frequency vector at the target locus, then updated this vector by calculating the weighted average of allele copies. The genotypic likelihood is maximized after the allele frequencies vector converges (Dempster et al., 1977).

The updated frequency \hat{p}'_k of k^{th} allele (denoted as A_k) is the weighted average of allele copies within all genotypes; these values for the genotypes, $\sum_{j=1}^N \Pr(v|\mathcal{P}_j) \Pr(G_i|\mathcal{P}_j)$, are included as the weights, which are expressed as follows:

$$\hat{p}'_k = \frac{\sum_i \Pr(A_k | G_i) \sum_{j=1}^N \Pr(v|\mathcal{P}_j) \Pr(G_i|\mathcal{P}_j)}{\sum_i \sum_{j=1}^N \Pr(v|\mathcal{P}_j) \Pr(G_i|\mathcal{P}_j)}, \quad (7)$$

where $\Pr(A_k | G_i)$ is the frequency of A_k in G_i (which is known), and the meanings of other symbols are as in the above logarithmic expression of genotypic likelihood.

Similarly, the updated negative amplification rate $\hat{\beta}'$ can be expressed as

$$\hat{\beta}' = \frac{N_{\emptyset} \hat{\beta}}{N \Pr(P = \emptyset | v)}, \quad (8)$$

where N_{\emptyset} is the number of negative phenotypes \emptyset of all individuals at the target locus, which is a constant, and $\Pr(P = \emptyset | v)$ is the probability of the negative phenotype \emptyset at the ploidy level v , which can be calculated from Equation (4).

In our EM algorithm, we estimated the allele frequencies and the negative amplification rate simultaneously. The initial frequencies of visible alleles were assumed equal (i.e., $1/K$, where K is the number of alleles at the target locus, including the null allele). The initial value of $\hat{\beta}$ was arbitrarily set (e.g., 0.05, because the final estimate is independent of the initial value). Each $\Pr(v | \mathcal{P}_j)$ was taken from the current ploidy assignment rates of the j^{th} individual in step (i) in our iterative algorithm, so this was also known. The procedure of our EM algorithm was as follows:

- I Let $\hat{\mathbf{P}} = [\hat{p}_1, \hat{p}_2, \dots, \hat{p}_K, \hat{\beta}]$ be the vector consisting of current allele frequencies (including the null allele frequency) and the current negative amplification rate at the target locus, whose initial vector is $[1/K, 1/K, \dots, 1/K, 0.05]$.
- II Use $\hat{\mathbf{P}}$ to calculate various genotype probabilities, and then use Equations (4) and (5) to calculate $\Pr(P = \emptyset | v)$ and $\Pr(G_i | \mathcal{P}_j)$, $j = 1, 2, \dots, N$.
- III Substitute the values obtained in step (ii) into Equations (7) and (8) to calculate the updated components in the vector $\hat{\mathbf{P}}' = [\hat{p}'_1, \hat{p}'_2, \dots, \hat{p}'_K, \hat{\beta}']$.
- IV If $\max(|\hat{\mathbf{P}} - \hat{\mathbf{P}}'|)$ is greater than or equal to a predefined threshold (e.g., 10^{-5}), then replace $\hat{\mathbf{P}}$ with $\hat{\mathbf{P}}'$ and repeat steps (ii) and (iii). Otherwise, terminate this process.

Remark 1: In the above EM algorithm, the estimation is restricted to one ploidy level and to one target locus. Therefore, to obtain the estimates of all allele frequencies and all negative amplification rates, the algorithm needs to be run for $\Lambda \times L$ times.

Remark 2: The presence of both null alleles and negative amplification can be freely incorporated into our model. If the effects of null alleles and/or negative amplification are not considered, the initial value of \hat{p}_y or $\hat{\beta}$ is set to zero. If both factors are not considered, the negative phenotypes cannot be explained, and so they are not used for the estimation of allele frequencies. Because we do not develop an false allele estimator, the true false allele rate is used in the following simulation.

Remark 3: For the PES model (Appendix S1), the value of the single chromatid recombination rate r_s is usually unknown. Because r_s is determined by the distance between the locus and the centromere (e.g., d), r_s is equal among different ploidies. We nested a downhill simplex algorithm (Nelder & Mead, 1965) outside of the allele frequency estimator to estimate the r_s . This algorithm is another method

to maximize the phenotypic likelihood. In this process, the vector $\hat{\mathbf{P}} = [\hat{p}_1, \hat{p}_2, \dots, \hat{p}_K, \hat{\beta}]$ for each ploidy level will be used, as will the formula $\arg \max_{r_s \in [0,1]} \sum_{\lambda=1}^{\Lambda} \ln \mathcal{L}(v_{\lambda})$.

Remark 4: In the estimation of the selfing ratio, the values s of different loci at the same ploidy level were assumed to be equal. Similarly, we nested a downhill simplex algorithm (Nelder & Mead, 1965) outside of the allele frequency estimator to maximize the phenotypic likelihood, the estimate $\text{hats} = \arg \max_{s \in [0,1]} \sum_{l=1}^L \ln \mathcal{L}(l)$. We also incorporated a method-of-moment selfing estimator based on the standardized identity disequilibrium coefficient (Hardy, 2016) in PLOIDYINFER.

It is noteworthy that when the selfing model is used, the genotypic and phenotypic probabilities are obtained by Equation (6) using the inbreeding coefficient F as an intermediate variable. However, both s and r_s will influence F , and the same value of F can be determined by multiple combinations of (s, r_s) . Therefore, the r_s and s cannot be simultaneously estimated using the maximum likelihood estimator.

Additionally, if three parameters ($r_s + p_y + \beta$ or $s + p_y + \beta$) are simultaneously estimated for allele frequencies, the locus should have at least three visible alleles. Because a biallelic locus has only four phenotypes (A, B, AB, and \emptyset), the degrees of freedom will only be three, which is insufficient to estimate more than three parameters.

2.5 | Criteria for model selection

If we consider the effects of null alleles and negative PCRs in a model or use the PES double-reduction model, then we may obtain a better estimate of an allele frequency (a lower root mean square error, RMSE) or a better ploidy assignment (a higher likelihood). However, such a model becomes complex because more parameters are used to explain various data, which may cause an overfitting to these data. This model will thus be penalized according to either the Akaike information criterion (AIC) (Akaike, 1974), the corrected Akaike information criterion (AICc) (McQuarrie & Tsai, 1998) or the Bayesian information criterion (BIC) (Schwarz, 1978). Therefore, we used the AIC or the BIC to evaluate a model. If a model has a smaller value of the AIC or the BIC, it is regarded as the preferred model. The values of AIC and BIC can be calculated by the following formulas:

$$\text{AIC} = N_{\text{par}} - 2 \ln \mathcal{L},$$

$$\text{AICc} = \text{AIC} + \frac{2N_{\text{par}}^2 + 2N_{\text{par}}}{N - N_{\text{par}} - 1},$$

$$\text{BIC} = N_{\text{par}} \ln N - 2 \ln \mathcal{L},$$

where N_{par} is the number of parameters used in a model, and \mathcal{L} is the phenotypic likelihood, whose logarithmic expression is.

$$\ln \mathcal{L} = \sum_{l=1}^L \sum_{j=1}^N \sum_{\lambda=1}^{\Lambda} \ln [\Pr(\mathcal{P}_j)] \Pr(v | \mathcal{P}_j).$$

If we use K'_l to denote the number of visible alleles in all individuals at the l^{th} locus, because the sum of these allele frequencies at this locus is equal to one, we can use K'_l parameters to express them. If the effect of either null alleles, negative amplification or a combined effect of both is considered, either one or two additional parameters will be required. In the PES double-reduction model, an extra parameter (the single chromatid exchange rate r_s) was added, whose values at all ploidy levels were assumed to be equal. In the selfing model, the selfing rate s was equal at different loci at the same ploidy level. In addition, the parameters (allele frequencies, negative amplification rate) at different loci at different ploidies were independently estimated. Hence, the total number N_{par} of parameters is given by the following formula:

$$N_{\text{par}} = \sum_{l=1}^L [\Lambda(K'_l - 1 + B_y + B_\beta) + B_{\text{PES}}] + \Lambda B_s + (\Lambda - 1)N, \quad (9)$$

where B_y , B_β , B_{PES} and B_s are four Boolean variables such that $B_y = 1$ if the null allele is considered, $B_\beta = 1$ if the negative amplification is involved, $B_{\text{PES}} = 1$ if the PES double-reduction model is chosen, and $B_s = 1$ when the selfing model is used; otherwise, these variables are all equal to zero.

3 | SIMULATIONS

For this application, we first evaluated the accuracy of the allele frequency estimator. Second, we evaluated the accuracy of individual ploidy assignment by computer simulation and the requirement of our model; finally, we validated our model with empirical data. The selfing rate was estimated with the maximum likelihood estimator during each simulation.

3.1 | Allele frequency estimation

Monte-Carlo simulations were used to evaluate the accuracy of the allele frequency estimator. Because the PES model was incompatible with the selfing model, two independent applications were performed to evaluate the accuracy of the allele frequency estimator. During each simulation, the prior probabilities of all ploidy levels were equal and were not updated during iteration.

Both applications generate the genotype at 20 loci according to the PES model, where the distance between this locus and the corresponding centromere is randomly drawn from a uniform distribution ranging from 0 to 100 cm. The single chromatid recombination rate at each locus was obtained from Haldane's mapping function, with the double-reduction rates obtained from Table S1. The selfing rate was set to zero in application (i) and to 0.1 or to 0.3 in application (ii).

Three levels for the initial number K' of visible alleles at each locus were set, which were $K' = 2, 4$ or 6 , to represent different levels of polymorphism. For the alleles at each locus, their initial number was $K'+1$ (including one null allele), and their initial frequencies were

drawn from the Dirichlet distribution with the concentration parameters equal to one:

$$p_1, p_2, \dots, p_{K'}, p_y \sim D(1, 1, \dots, 1).$$

Second, we created a monoecious population that contained individuals with disomic, tetrasomic and hexasomic inheritance. To simulate the phenotypes, we created a founder population of 1,000 individuals whose genotypes are in accordance with the HWE. The individuals of the population were then simulated to reproduce for 15 generations to let the genotypic probabilities reach equilibrium. Finally, we sampled 50 to 200 individuals at an interval of 10 from the final generation. For each offspring, the parents were two distinct individuals randomly chosen from the previous generation at a probability of $1-s$ or were the same randomly chosen individual at a probability of s . The following procedure was performed to simulate meiosis: (i) the chromosomes were randomly paired, and the alleles are exchanged between the paired chromosomes at a probability of r_s ; (ii) the chromosomes were randomly segregated into two secondary oocytes; and (iii) the alleles within a chromosome were randomly segregated into two gametes. Fertilization was then simulated by the merging of two gametes each from one parent.

The actual allele frequencies were calculated from the genotypes of the sampled individuals. These genotypes were then converted into phenotypes by the removal of both null and duplicated alleles and the insertion of false alleles at $e = 0.05$. The phenotypes were then randomly set to \emptyset at a probability of $\beta = 0.05$ to simulate negative amplification. Because of the insufficient degrees of freedom of the distribution of phenotypes at biallelic loci, the true value of β was set to zero for loci with two visible alleles and is not considered in the estimation. Furthermore, the maximum number of false alleles was two; this number and the false allele rate e were used as a priori information in the estimation.

Finally, we estimated the allele frequencies from the phenotypes by PES + r_s + p_y + β model in application (i). Because both r_s and s could not be estimated simultaneously, PRCS + s + p_y + β model was used for application (ii). Both applications were repeated 10,000 times, and the difference between the estimated and actual parameters (including visible allele frequency, p_y , β , r_s and s) compared. The RMSE and the bias for them were calculated to evaluate the accuracy of our allele frequency estimator:

$$\text{RMSE}(\hat{X}) = \sqrt{\text{Average}[(\hat{X} - X)^2]},$$

$$\text{Bias}(\hat{X}) = \text{Average}(\hat{X} - X).$$

The RMSE of the estimated visible allele frequencies is shown in Figure 2, where each subfigure shows the graph of the RMSE as a function of the sample size (at three levels for the initial number $K'+1$ of alleles at each locus). The bias of estimated visible allele frequencies is shown in Figure S2. The RMSE and bias of estimated null allele frequencies, negative amplification rate, single chromatid recombination rate and selfing rate are shown in Figures S2–S8, respectively.

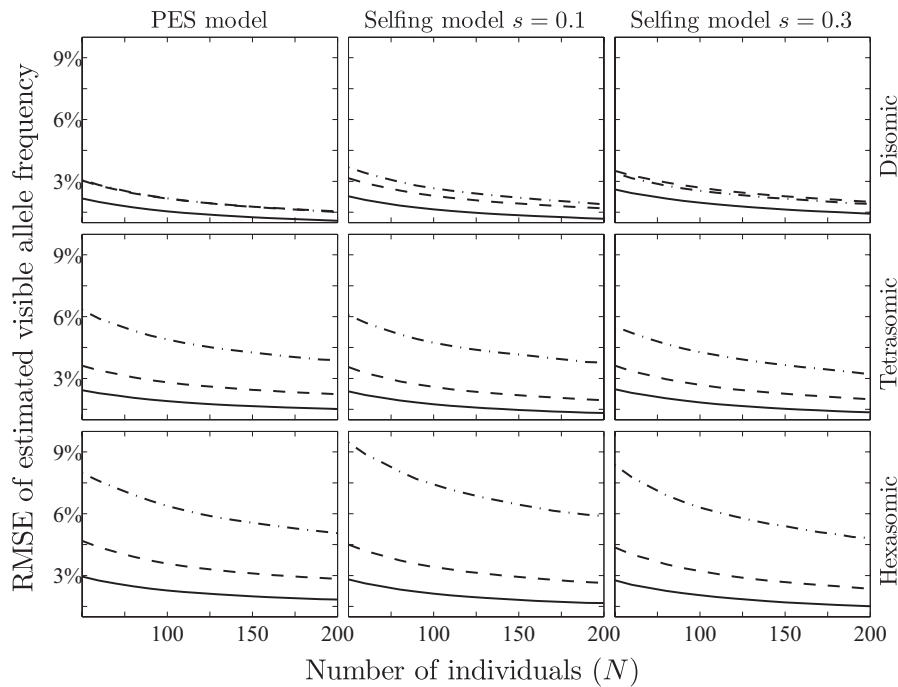


FIGURE 2 The RMSE of the estimated visible allele frequencies. We simulated three populations with the inherited types from disomic to hexasomic, and each result is shown on a separate row. The selfing rate was simulated at three levels (0, 0.1 and 0.3). For outcrossed populations, the allele frequencies were estimated by a $PES + r_s + p_y + \beta$ model, and for self-fertilized populations, the $PRCS + s + p_y + \beta$ model was used. The corresponding results obtained by 20 loci and 10,000 simulations are successively shown in the three columns from left to right. The number N of individuals sampled in the final generation ranges from 50 to 200. The initial number of visible alleles at each locus is 2, 4 or 6, and the graph of the corresponding RMSE as a function of sample size (N) is plotted by a dash-dotted, dashed or solid curve

In Figure 2, each RMSE decreases as the sample size increases, the value of which is approximately 0.01–0.06 at $N = 200$, with each curve near to an asymptote when the sample size is high. The RMSE decreases as K' increases and increases as v increases. The selfing rate only slightly changes the RMSE.

The pattern of the RMSE of the null allele frequencies is similar to that for visible allele frequencies (Figure S3), but the RMSE of \hat{p}_y for $K' = 2$ is at its lowest for disomic inheritance.

The pattern of the RMSE for a negative amplification rate is also similar to that for the visible allele frequencies (Figure S5), although this decreases as v increases.

Both r_s and s cannot be estimated simultaneously so the results of each are shown in the same figure. The selfing rate can be reliably estimated from the data, and the RMSE can reach 0.02 to 0.04 at $N = 200$ (Figure S7). However, the RMSE of r_s is very high, approximately 10 to 20 times that of s .

3.2 | Simulated data

Here, we simulated a mixed-ploidy population to test the correct assignment rate under different conditions. The correct assignment rate is defined as the probability that an individual is assigned to the true ploidy level.

The simulated population is monoecious, consisting of individuals with five levels of even ploidies, which range from diploid to decaploid. This population is assumed to be genotyped at L unlinked loci, and the distances between each of these loci and their corresponding centromeres are drawn from a uniform distribution from 0 to 100 cm. The number L of loci is obtained by the round function $L = \text{Round}(K'_{\text{tot}}/K')$, where K'_{tot} is the total number of visible alleles, and K' is the number of visible alleles per locus. There are three

levels for K' , i.e., $K' = 2, 4$ or 6 , and there are four levels for K'_{tot} , i.e., $K'_{\text{tot}} = 50, 80, 110$ or 140 .

The initial frequency \bar{p}_y of null alleles is set at two levels, i.e., $\bar{p}_y = 0.05$ or 0.1 , and the initial frequencies of visible alleles are all equal. To simulate the genetic drift, the allele frequencies for each ploidy level were drawn from a Dirichlet distribution (Pritchard et al., 2000):

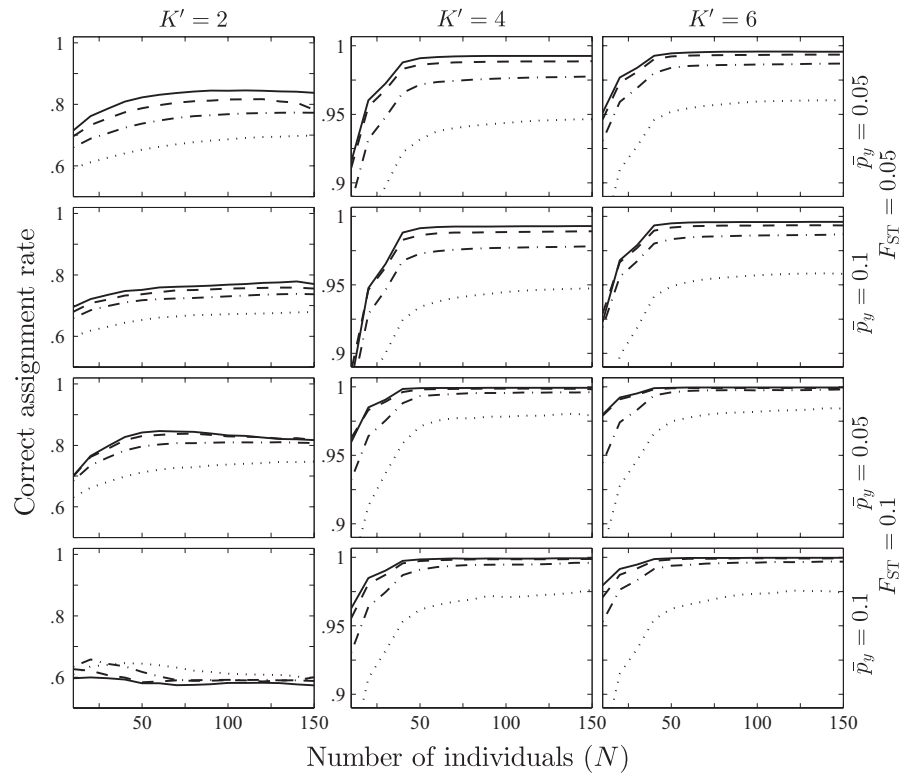
$$p_1, p_2, \dots, p_{K'}, p_y \sim D(\gamma_1, \gamma_2, \dots, \gamma_{K'}, \gamma_y).$$

This assumes that each γ_k is proportional to the corresponding initial allele frequency \bar{p}_k , such that $\gamma_k = (1/F_{ST} - 1)\bar{p}_k$, and the variance of each p_k can be expressed as $\text{Var}(p_k) = F_{ST}\bar{p}_k(1 - \bar{p}_k)$. This is consistent with the definition of Wright's F_{ST} (Templeton, 2006). F_{ST} is set at two levels, which are $F_{ST} = 0.05$ or 0.1 .

Thereafter, we generated 1,000 individuals for each ploidy level. Genotypes were randomly drawn according to their predicted probabilities under the RCS model to accelerate the simulation. This population was reproduced for 15 generations to let the genotypic probabilities reach equilibrium, with three levels of selfing rate simulated during reproduction, i.e., $s = 0, 0.1$ or 0.3 . We sampled 10 to 150 individuals (denoted as N) at an interval of 10 individuals in the final generation for each ploidy level. Each genotype was converted to a phenotype to be used to perform the ploidy assignment.

Eight PRCS-based models were evaluated, where each model was a PRCS model with three parameters being considered or unconsidered: null alleles ($+p_y$), negative amplification ($+\beta$) or selfing ($+s$). The model with the lowest BIC was chosen. For each parameter combination of \bar{p}_y , F_{ST} , s , N , K' and K'_{tot} , we performed 1,000 simulations and used the correct assignment rate to evaluate the

FIGURE 3 The correct assignment rate at $s = 0.1$. The number of individuals sampled at each ploidy level is from 10 to 150. The number K' of visible alleles per locus is 2, 4 or 6, and the corresponding results are shown in different columns. The initial null allele frequency \bar{p}_y is 0.05 or 0.1, and Wright's F_{ST} is also 0.05 or 0.1. The result for each combination of \bar{p}_y and F_{ST} obtained by 1,000 simulations is shown in one row. The total number of visible alleles K'_{tot} is 50, 80, 110 or 140, the number of loci is determined by Round (K'_{tot}/K'), and the graph of the corresponding correct assignment rate as a function of sample size (N) is shown in a dotted, dash-dotted, dashed or solid curve



efficiency of our model. The results for the correct assignment rate as a function of sample size at $s = 0.1$ are shown in Figure 3, and the results for $s = 0$ and 0.3 are shown in Figures S9 and S10, respectively.

The results for $K' = 2$ performed the worst and were insufficient to perform further analyses. The correct assignment rate here can reach a maximum value of approximately 0.9, 0.8 and 0.7 at $s = 0, 0.1$ and 0.3, respectively.

The results for $K' > 2$ were similar, with the correct assignment rate generally increasing with increased sample size. The two curves increased steeply with small sample sizes and become flat at approximately $N = 50$. The results for $K' = 6$ are more accurate than that for $K' = 4$ at low rates of genetic differentiation (by 0.5%–2%).

The presence of self-fertilization will largely reduce the accuracy of the model when $K = 2$. When $K' > 2$, the correct assignment rate is reduced by 2%–5% when s is increased from 0 to 0.3 (Figures S9 and S10). This effect can be offset when rates of genetic differentiation are high ($F_{ST} = 0.1$).

The presence of null alleles can also reduce the correct assignment rate, but its effects are very limited when $K' > 2$ (<0.5%) and can be negligible. A high degree of genetic differentiation can improve the correct assignment rate for $K' > 2$ by 0%–5%, especially when the selfing rate is also high (Figure S10).

3.3 | Empirical data

We used real phenotypic data for the Bermuda buttercup (*Oxalis pes-caprae*) to validate our model (Ferrero et al., 2015). The Bermuda

buttercup is a polyploid, highly clonal geophyte that is native to South Africa (Born, Linder, & Desmet, 2007). This plant is considered a widespread and noxious invasive weed occurring in disturbed sites in all Mediterranean climatic regions (Ferrero et al., 2013). In its native habitats, both diploid ($2n = 2x = 14$ chromosomes) and tetraploids cytotypes ($2n = 4x = 28$ chromosomes) are present, with tetraploidy being the most common ploidy level (te Beest et al., 2011). The sterile pentaploid cytotype ($2n = 5x = 35$ chromosomes) is the most widespread form throughout the introduced range (Castro et al., 2013).

Ferrero et al. (2015) collected genetic samples from 22 populations from South Africa and the Western Mediterranean. The ploidy levels were determined by flow cytometry, and there were 42, 242 and 96 diploids, tetraploids and pentaploids. The genetic diversity was evaluated from the phenotypes at seven nuclear microsatellite loci. The total number of individuals was 380, the number of visible alleles per locus ranged from nine to 17, and 78 visible alleles were detected in total.

Because the sample size of each population was relatively small (ranging from six to 21), we did not perform ploidy assignment for each population. Instead, we assumed that all samples originated from the same mixed-ploidy population. We used different models to perform the allele frequency estimations and ploidy assignments for the total population. The number of parameters, likelihoods, BIC, correct assignment rates and number of misidentified individuals are shown in Table 1.

From Table 1, it can be seen that the correct assignment rate varied among models from 0.571 to 0.924. The correct assignment rate is negatively correlated with the BIC. The model with the optimal BIC is $RCS + p_y$, whose correct assignment rate is 0.913. Misidentifications mainly occurred between adjacent ploidies.

4 | DISCUSSION

4.1 | Allele frequency estimator

Although polyploids have more genetic information if the unambiguous genotypes are available (Huang, Ritland, Guo, Shattuck, & Li, 2014), due to the genotyping ambiguity, one cannot determine the dosage of alleles within a phenotype. This makes it difficult to accurately estimate allele frequencies for polyploids.

Our new allele frequency estimator is an extension of Kalinowski and Taper's (2006) maximum-likelihood estimator, which utilizes an EM algorithm (Dempster et al., 1977) to maximize likelihood locus by locus. There are also some method-of-moment null allele estimators (e.g., Brookfield, 1996; Chakraborty, Andrade, Daiger, & Budowle, 1992; van Oosterhout, Hutchinson, Wills, & Shipley, 2004; Summers & Amos, 1997). However, the estimates of null allele frequencies

by moment estimators may be negative because they are unbiased. Estimates by maximum-likelihood estimators are asymptotically unbiased and always lie within a biological meaningful range, which can be directly used for ploidy assignment without truncation. Moreover, the maximum-likelihood estimator yields a lower RMSE (Kalinowski & Taper, 2006).

The estimation of allele frequencies becomes decreasingly accurate with an increase in ploidy level, especially for low polymorphic markers (see Figure 2 and S3). The reasons for this are as follows. For high levels of ploidy, there are many allele copies within each individual; they are generally not all the same, so there will be few homozygous phenotypes. Additionally, the estimated allele frequencies depend largely on a small portion of homozygous phenotypes, which are sensitive to the number of homozygous phenotypes. A little variation in this number will cause a large change in the estimation.

TABLE 1 The number of parameters, likelihood, BIC, correct assignment rate and number of misidentified individuals in ploidy assignment for the data set of Ferrero et al. (2015) evaluated with different models

Model	Extra	N_{par}	$\ln \mathcal{L}$	BIC	Correct	4 → 2	4 → 5	5 → 2	5 → 4
RCS		979	-8140.2	22,095.8	0.57	127	5	2	29
RCS	+ p_y	1,000	-7527.6	20,995.4	0.91	14	8	0	11
RCS	+ β	1,000	-8308.6	22,557.4	0.57	125	6	2	29
RCS	+s	982	-7683.3	21,199.9	0.92	8	13	0	10
RCS	+ β + s	1,003	-7845.2	21,648.5	0.92	8	11	0	10
RCS	+ p_y + s	1,003	-7527.6	21,013.3	0.91	14	8	0	11
RCS	+ p_y + β	1,021	-7476.0	21,017.0	0.90	20	9	0	9
RCS	+ p_y + β + s	1,024	-7474.4	21,031.6	0.90	19	9	0	9
PRCS		979	-7983.0	21,781.4	0.65	97	4	1	30
PRCS	+ p_y	1,000	-7528.8	20,997.8	0.91	14	8	0	11
PRCS	+ β	1,000	-8151.5	22,243.3	0.65	97	5	1	30
PRCS	+s	982	-7683.3	21,199.9	0.92	8	13	0	10
PRCS	+ β + s	1,003	-7845.2	21,648.4	0.92	8	11	0	10
PRCS	+ p_y + s	1,003	-7528.6	21,015.1	0.91	14	8	0	11
PRCS	+ p_y + β	1,021	-7473.8	21,012.5	0.90	19	9	0	9
PRCS	+ p_y + β + s	1,024	-7474.4	21,031.6	0.90	19	9	0	9
CES		979	-7963.7	21,742.8	0.67	92	4	1	30
CES	+ p_y	1,000	-7529.4	20,998.9	0.91	14	8	0	11
CES	+ β	1,000	-8132.3	22,204.8	0.66	93	5	1	30
CES	+s	982	-7683.3	21,199.9	0.92	8	13	0	10
CES	+ β + s	1,003	-7845.2	21,648.5	0.92	8	11	0	10
CES	+ p_y + s	1,003	-7528.9	21,015.8	0.91	14	8	0	11
CES	+ p_y + β	1,021	-7473.5	21,011.8	0.90	19	9	0	9
CES	+ p_y + β + s	1,024	-7474.4	21,031.6	0.90	19	9	0	9
PES		986	-7959.6	21,776.2	0.66	95	4	1	30
PES	+ p_y	1,007	-7525.9	21,033.6	0.91	14	8	0	11
PES	+ β	1,007	-8127.9	22,237.5	0.66	94	6	1	30
PES	+ p_y + β	1,028	-7473.5	21,053.5	0.91	19	9	0	8

Note: The column headers of the form $x \rightarrow y$ denotes the number of individuals with true ploidy level x who were misidentified as ploidy level y . All diploids are correctly identified, i.e. $2 \rightarrow 4$ and $2 \rightarrow 5$ are all equal to zero, so these results are not shown.

The negative amplification rate, single chromatid recombination rate and selfing rate can all be estimated simultaneously with the allele frequencies.

As the ploidy level increases, the negative phenotypes caused by null allele homozygotes are reduced. Therefore, the negative amplification rate becomes more accurate at higher ploidy levels (Figure S5).

The single chromatid recombination rate can slightly influence the phenotypic probabilities, which can be easily concealed by sampling variance. Therefore, the estimate of r_s can be inaccurate (Figure S7). In contrast, the selfing rate is estimated based on data from all loci, the data volume is 20 times of that of r_s in our simulation. Therefore, the estimation of s is much more accurate than that of r_s (Figure S5). Fortunately, because the value of r_s is the same in different populations at the same locus, the accuracy of r_s can be increased by sampling more populations.

4.2 | Individual ploidy assignment

The assignment of individual ploidy depends on prior probabilities, individual phenotypes and allele frequencies. We used an iterative algorithm to alternately update the allele frequencies and ploidy assignment rate. The ploidy assignment rate is convergent during iteration. This iterative algorithm uses the difference in phenotypic distributions among various ploidies to distinguish individuals of different ploidy levels. This difference is determined by two major factors: (i) population genetic structure (including genetic differentiation and inbreeding/subdivision) and (ii) the polymorphism of the loci used.

Because of the resulting low reproductive success, gene-flow between ploidies is usually lower than intraploidy gene-flow, which generates differentiation among ploidies. A moderate differentiation will enhance ploidy assignment. Our results also indicated that our method will perform best at a higher differentiation ($F_{ST} = 0.1$, Figure 3). However, if the differentiation is extremely high ($F_{ST} > 0.5$), most phenotypes will be homozygous, and it becomes difficult to perform ploidy assignment.

The level of inbreeding/subdivision can affect the degree of homozygosity and make the phenotypic distributions deviate from double-reduction equilibrium expectations. This will also reduce the accuracy of ploidy assignment. The selfing model uses the inbreeding coefficient as an intermediate variable, which can help explain the phenotypic distribution under inbreeding/subdivision and to some extent increase the accuracy of the simulation.

Where there are only a few alleles at a locus, the phenotypes among individuals with high levels of ploidy (e.g., hexaploids) will be similar. It therefore becomes difficult to distinguish individuals with high levels of ploidy. For example, nearly all phenotypes are the heterozygote AB at a biallelic marker in such a case. During simulation, our algorithm is inaccurate when loci with two visible alleles are used (Figure 3, far left). Therefore, biallelic markers are unsuitable for ploidy assignment.

Furthermore, the rates of false alleles, null alleles and negative amplification can also influence the phenotypic distribution and result in the phenotypic distributions becoming more similar: false alleles increase the proportion of heterozygous phenotypes and phenotypes with more alleles, while null alleles and negative amplification increase the proportion of negative phenotypes. Therefore, all of these factors will reduce the accuracy of ploidy assignment.

4.3 | Conditions for successful ploidy inference

In our simulation, we found that to reach a correct assignment rate of 95%, under relatively good conditions (a high differentiation and a low selfing rate), our algorithm requires approximately 50 visible alleles and 40 individuals for each ploidy level. There should also be at least three visible alleles at each locus. Under relatively poor conditions (a low differentiation and a high selfing rate), approximately 70 (for higher polymorphic loci, $K' = 6$) or 100 (for lower polymorphic loci, $K' = 4$) visible alleles and 50 individuals at each ploidy level are required. Biallelic markers are unsuitable for performing ploidy assignment.

We emphasize that our simulations are conservative because we used five possible ploidy levels. In practice, the situation may be less restrictive due to only two or three ploidy levels being present in real populations (Ferrero et al., 2015; Grabowski, Morris, Casler, & Borevitz, 2014; McAllister & Miller, 2016).

When the numbers of loci and sample sizes are relatively low (5 to 10 individuals per ploidy level), we suggest that researchers use other methods to determine the true ploidy levels and to use these samples as reference material. If these ploidy levels are used as a priori information, it will greatly improve the ploidy assignment and will reduce the required number of loci and sample sizes by approximately 20% to 30% (five reference individuals) as a conservative estimation.

4.4 | Empirical data

The correct assignment rate of the data set of Ferrero et al. (2015) ranges from 0.571 to 0.924. There are some nonideal situations in this data set that preclude the optimal utilization of our model, and some imperfect results appear.

We assumed that all samples are from a single mixed-ploidy population and that there is no subdivision within each ploidy level. However, the data set of Ferrero et al. (2015) consists of 22 populations, in which 10 populations are collected from South Africa, and the others from the western Mediterranean. The differentiation among populations at the same ploidy level results in both genotypic and phenotypic probabilities deviating from the expectations of our model, and hence the correct assignment rate is reduced.

Although our model cannot perfectly assign individuals to their true ploidy, the correct assignment rate can be improved by the following three approaches. (a) Perform ploidy assignment separately

for each population or region. This eliminates any effects of population subdivision but may require increased sample sizes. (b) Increase the number of loci. Ferrero et al. (2015) only used seven microsatellites, which is relatively few for modern population genetics analysis. (c) Determine the true ploidy level for some individuals and use these as reference material.

When the ploidy levels are known, our model can also be used to estimate the allele frequency, null allele frequency, negative amplification rate, selfing rate, double-reduction rate, and genetic differentiation; calculate the likelihood, AIC, AICc and BIC; and test the distribution of the phenotype.

ACKNOWLEDGMENTS

We would like to thank the three anonymous reviewers for their suggestions and comments and thank Professor Olivier J. Hardy for the implementation of the selfing rate estimator. This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020302), the National Natural Science Foundation of China (31770411, 31730104, 31572278 and 31770425), the Young Elite Scientists Sponsorship Program by CAST (2017QNRC001), the National Key Programme of Research and Development, Ministry of Science and Technology (2016YFC0503200), and the Natural Science Basic Research Plan in Shaanxi Province of China (2018JM3024, 2019JM258). DWD is supported by a Shaanxi Province Talents 100 Fellowship, and KH is supported by a scholarship from China Scholarship Council.

AUTHOR CONTRIBUTIONS

K.H. and B.G.L. designed the project. K.H. established the model and wrote the draft. P.Z. and Y.D. performed simulations. Z.H.L. and D.W.D. checked the model and edited the manuscript.

DATA ACCESSIBILITY

The executable files, the source code, and the user manual of PLOIDYINFER V1.4 are available on GitHub (<http://github.com/huangkang1987/ploidyinfer>) and Zenodo (<https://doi.org/10.5281/zenodo.2466735>). The simulation program and its source code are given as supplementary information to this journal Ferrero V, Barrett SC, Castro S et al. (2014) Data from: Invasion genetics of the Bermuda buttercup (*Oxalis pes-caprae*): complex intercontinental patterns of genetic diversity, polyploidy and heterostyly characterize both native and introduced populations. Dryad Digital Repository. <https://doi.org/10.5061/dryad.r69g1>.

ORCID

Kang Huang  <https://orcid.org/0000-0002-8357-117X>

Derek W. Dunn  <https://orcid.org/0000-0001-5909-1224>

Zhonghu Li  <https://orcid.org/0000-0001-6763-5586>

Pei Zhang  <https://orcid.org/0000-0003-4433-382X>

Baoguo Li  <https://orcid.org/0000-0001-7430-3889>

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., & Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytologist*, 210, 391–398. <https://doi.org/10.1111/nph.13698>
- Bisognin, C., Seemüller, E., Citterio, S., Velasco, R., Grando, M. S., & Jarausch, W. (2009). Use of SSR markers to assess sexual vs. apomictic origin and ploidy level of breeding progeny derived from crosses of apple proliferation-resistant *Malus sieboldii* and its hybrids with *Malus domestica* cultivars. *Plant Breeding*, 128, 507–513.
- Born, J., Linder, H. P., & Desmet, P. (2007). The Greater Cape Floristic Region. *Journal of Biogeography*, 34, 147–162.
- Brookfield, J. F. Y. (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology*, 5, 453–455. <https://doi.org/10.1111/j.1365-294X.1996.tb00336.x>
- Brugnoli, E. A., Urbani, M. H., Quarín, C. L., Martínez, E. J., & Acuña, C. A. (2013). Diversity in diploid, tetraploid, and mixed diploid-tetraploid populations of *Paspalum simplex*. *Crop Science*, 53, 1509–1516. <https://doi.org/10.2135/cropsci2012.08.0497>
- Butruille, D. V., & Boiteux, L. S. (2000). Selection-mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proceedings of the National Academy of Sciences*, 97, 6608–6613. <https://doi.org/10.1073/pnas.100101097>
- Castro, S., Ferrero, V., Costa, J., Sousa, A. J., Castro, M., Navarro, L., & Loureiro, J. (2013). Reproductive strategy of the invasive *Oxalis pes-caprae*: Distribution patterns of floral morphs, ploidy levels and sexual reproduction. *Biological Invasions*, 15, 1863–1875. <https://doi.org/10.1007/s10530-013-0414-2>
- Castro, S., Loureiro, J., Procházka, T., & Münzbergová, Z. (2012). Cytotype distribution at a diploid-hexaploid contact zone in *Aster amellus* (Asteraceae). *Annals of Botany*, 110, 1047–1055. <https://doi.org/10.1093/aob/mcs177>
- Chakraborty, R., Andrade, M. D., Daiger, S. P., & Budowle, B. (1992). Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Annals of Human Genetics*, 56(1), 45–57. <https://doi.org/10.1111/j.1469-1809.1992.tb01128.x>
- Clark, L. V., & Jasieniuk, M. (2011). polysat: An R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, 11, 562–566. <https://doi.org/10.1111/j.1755-0998.2011.02985.x>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1–38.
- Ferrero, V., Barrett, S. C., Castro, S., Caldeirinha, P., Navarro, L., Loureiro, J., & Rodríguez-Echeverría, S. (2015). Invasion genetics of the Bermuda buttercup (*Oxalis pes-caprae*): Complex intercontinental patterns of genetic diversity, polyploidy and heterostyly characterize both native and introduced populations. *Molecular Ecology*, 24, 2143–2155.
- Ferrero, V., Castro, S., Costa, J., Acuña, P., Navarro, L., & Loureiro, J. (2013). Effect of invader removal: Pollinators stay but some native plants miss their new friend. *Biological Invasions*, 15, 2347–2358. <https://doi.org/10.1007/s10530-013-0457-4>
- Gompert, Z., & Mock, K. E. (2017). Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Molecular Ecology Resources*, 17, 1156–1167. <https://doi.org/10.1111/1755-0998.12657>

- Grabowski, P. P., Morris, G. P., Casler, M. D., & Borevitz, J. O. (2014). Population genomic variation reveals roles of history, adaptation and ploidy in switchgrass. *Molecular Ecology*, 23, 4059–4073. <https://doi.org/10.1111/mec.12845>
- Hardy, O. J. (2016). Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources*, 16, 103–117. <https://doi.org/10.1111/1755-0998.12431>
- Huang, K., Ritland, K., Guo, S. T., Shattuck, M., & Li, B. G. (2014). A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, 14, 734–744. <https://doi.org/10.1111/1755-0998.12217>
- Huang, K., Wang, T. C., Dunn, D. W., Zhang, P., Cao, X. X., Liu, R. C., & Li, B. G. (2019). Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. *G3: Genes, Genomes, Genetics*, 9, 1693–1706. <https://doi.org/10.1534/g3.119.400132>
- Jarne, P., & Charlesworth, D. (1993). The evolution of the selfing rate in functionally hermaphrodite plants and animals. *Annual Review of Ecology and Systematics*, 24, 441–466. <https://doi.org/10.1146/annurev.es.24.110193.002301>
- Kalinowski, S. T., & Taper, M. L. (2006). Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics*, 7, 991–995. <https://doi.org/10.1007/s10592-006-9134-9>
- Kalinowski, S. T., Wagner, A. P., & Taper, M. L. (2006). ml-relate: A computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes*, 6, 576–579. <https://doi.org/10.1111/j.1471-8286.2006.01256.x>
- Kim, S., Rayburn, A., Boe, A., & Lee, D. K. (2012). Neopolyploidy in *Spartina pectinata* Link: 1. Morphological analysis of tetraploid and hexaploid plants in a mixed natural population. *Plant Systematics and Evolution*, 298, 1073–1083. <https://doi.org/10.1007/s00606-012-0617-5>
- Liu, Y. X., Xia, T., Zheng, Y. H., Zhi, Y. Q., & Zhou, J. (2015). Genetic diversity and the population structure at two ploidy levels of *Lycoris radiata* as revealed by SCoT analysis. *Biochemical Systematics and Ecology*, 62, 106–114. <https://doi.org/10.1016/j.bse.2015.08.003>
- Luo, Z. W., Zhang, Z., Zhang, R. M., Pandey, M., Gailing, O., Hattmer, H. H., & Finkeldey, R. (2006). Modeling population genetic data in autotetraploid species. *Genetics*, 172, 639–646. <https://doi.org/10.1534/genetics.105.044974>
- Martin, S. L., & Husband, B. C. (2013). Adaptation of diploid and tetraploid *Chamerion angustifolium* to elevation but not local environment. *Evolution*, 67, 1780–1791.
- McAllister, C. A., & Miller, A. J. (2016). Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *American Journal of Botany*, 103, 1314–1325.
- McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection*. Singapore: World Scientific.
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7, 308–313. <https://doi.org/10.1093/comjnl/7.4.308>
- Peirson, J. A., Reznicek, A. A., & Semple, J. C. (2012). Polyploidy, infra-specific cytotype variation, and speciation in Goldenrods: The cytogeography of *Solidago* subsect. *Humiles* (Asteraceae) in North America. *Taxon*, 61, 197–210.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.
- Ravinet, J., Westram, A., Johannesson, K., Butlin, R., André, C., & Panova, M. (2016). Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Molecular Ecology*, 25, 287–305.
- Rieger, R., Michaelis, A., & Green, M. M. (1968). *A glossary of genetics and cytogenetics: Classical and molecular*. New York, NY: Springer-Verlag.
- Ritland, K. (2002). Extensions of models for the estimation of mating systems using *n* independent loci. *Heredity*, 88, 221–228. <https://doi.org/10.1038/sj.hdy.6800029>
- Robertson, A., Rich, T. C. G., Allen, A. M., et al. (2010). Hybridization and polyploidy as drivers of continuing evolution and speciation in *Sorbus*. *Molecular Ecology*, 19, 1675–1690.
- Schlötterer, C., & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Research*, 20, 211–215. <https://doi.org/10.1093/nar/20.2.211>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464. <https://doi.org/10.1214/aos/1176344136>
- Summers, K., & Amos, W. (1997). Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. *Behavioral Ecology*, 8, 260–267.
- Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., ... Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, 24, 3189–3194. <https://doi.org/10.1093/nar/24.16.3189>
- te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubesova, M., & Pysek, P. (2011). The more the better? The role of polyploidy in facilitating plant invasions. *Annals of Botany*, 109, 19–45. <https://doi.org/10.1093/aob/mcr277>
- Templeton, A. R. (2006). *Population genetics and microevolutionary theory*. Hoboken: John Wiley.
- van Oosterhout, C., Hutchinson, W. F., Wills, D. P., & Shipley, P. (2004). micro-checker: Software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, 4, 535–538. <https://doi.org/10.1111/j.1471-8286.2004.00684.x>
- Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., & Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, 106, 13875–13879. <https://doi.org/10.1073/pnas.0811575106>
- Yang, Z. Y., Chen, F., Chamberlin, S. G., & Benner, S. A. (2010). Expanded genetic alphabets in the polymerase chain reaction. *Angewandte Chemie International Edition*, 49, 177–180. <https://doi.org/10.1002/anie.200905173>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Huang K, Dunn DW, Li ZH, Zhang P, Dai Y, Li BG. Inference of individual ploidy level using codominant markers. *Mol Ecol Resour*. 2019;19:1218–1229. <https://doi.org/10.1111/1755-0998.13032>