Supplementary materials of ‘Performing Parentage Analysis for Polysomic Inheritances Based on Allelic Phenotypes’

Kang Huang, Gwendolyn Huber, Kermit Ritland, Derek W. Dunn, Baoguo Li

Appendices

A Double-reduction models

In the presence of double-reduction, a gamete will carry some *identical-by-double-reduction* (IBDR) alleles. For tetrasomic and hexasomic inheritances, there are only two and three allele copies within a gamete, respectively. Hence, there is at most one pair of IBDR alleles within a gamete. Therefore, we only need to use a single parameter to measure the degree of double-reduction.

For polysomic inheritance with a high ploidy level v, there may be more than one pair of IBDR alleles within a gamete. Therefore, it is necessary to add some additional parameters to measure the degree of double-reduction. Let α_i be the probability that a gamete carries i pairs of IBDR alleles. Then $\sum_{i=0}^{\lfloor v/4 \rfloor} \alpha_i = 1$, where $\lfloor v/4 \rfloor$ is the greatest integer not more than $v/4$. We call each α_i a *double-reduction rate*.

Geneticists have developed several simplified models to simulate double-reduction. In the *random chromosome segregation* (RCS) model, the crossing over between the target locus and the corresponding centromere is ignored. Therefore, there cannot be any IBDR allele in a gamete, and the genotypic frequencies accord with the HWE (Figure S1(A), Muller, 1914).

The *pure random chromatid segregation* (PRCS) model accounts for such crossings over, and assumes that the chromatids behave independently in the meiotic anaphase, and are randomly segregated into some gametes (Figure S1(B), Haldane, 1930). When a pair of sister chromatids are segregated into the same gamete, the double-reduction occurs.

In the *complete equational segregation* (CES) model, the whole arms of two pairing chromatids are supposed to be exchanged between the pairing chromosomes (Figure S1(C), Mather, 1935). Subsequently, the chromosomes are randomly segregated into the secondary oocytes in Metaphase I. If the pairing chromosomes are segregated into the same secondary oocyte, the duplicated alleles may be further segregated into a single gamete.

The probability that an allele within a chromatid is exchanged with a pairing chromatid is called the *single chromatid recombination rate*, denoted by r_s. In the CES model, the rate r_s is assumed to be one. This is an ideal assumption. In fact, the maximum value of r_s is 50% whenever the locus is located far from the centromere. Huang *et al.* (2019) presented a model by incorporating r_s into CES, called the *partial equational segregation* (PES) model. Let d be the distance (in centimorgans) from the target locus

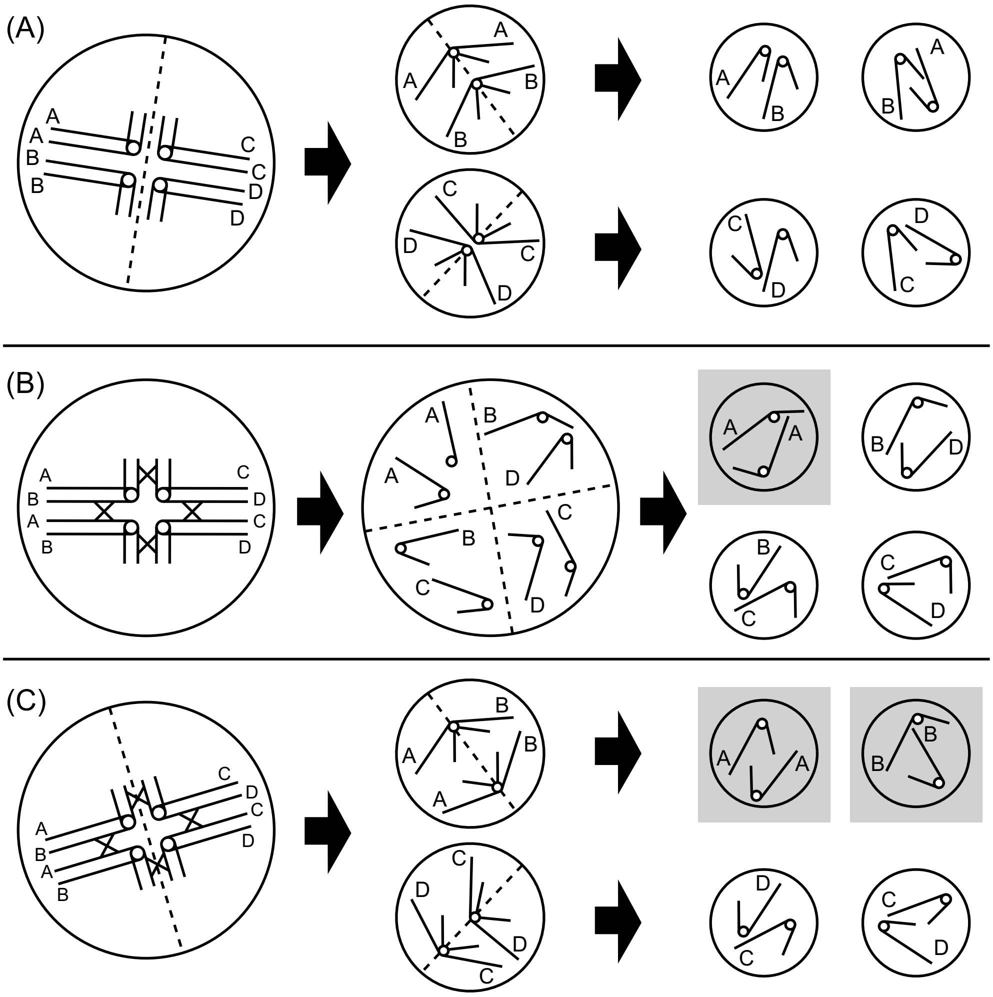

Figure S1: Diagram of double-reduction models under tetrasomic inheritance. The left column shows three primary oocytes, the middle column shows two secondary oocytes (in the rows marked (A) and (C)) or one tetrad (in the row marked (B)), and the right column shows three gametes. The gametes with a gray background carry IBDR alleles. We denote the cellular fissions by dashed lines, the arms of chromosomes by solid lines, and the centromeres by circles connecting solid lines. Each locus is located in a long arm of chromosomes and the identical-by-descent allele is denoted by the same letter as the corresponding locus. The row marked (A) is the sketch of RCS model. In this model, the crossing over between the target locus and its corresponding centromere is ignored (Muller, 1914). In the absence of crossing over, gametes may originate from any combination of homologous chromosomes, and two sister chromatids are never sorted into the same gamete (Parisod *et al.*, 2010). The row marked (B) is the sketch of PRCS model. This model accounts for the crossing over between the target locus and its corresponding centromere, and assumes that the chromatids behave independently in the meiotic anaphase, and are randomly segregated into the gametes (Haldane, 1930). When a pair of sister chromatids are segregated into the same gamete, the double-reduction occurs. The probability that two chromatids within the same gamete are a pair of sister chromatids is $4/\binom{8}{2}$, i.e. 1/7, where 4 is the number of pairs of sister chromatids, and $\binom{8}{2}$ is the number of ways to sample two chromatids from eight chromatids. The row marked (C) is the sketch of CES model. In this model, the pairs of homologous chromosomes are exchanged with the chromatids via recombination (Mather, 1935). The whole arms of sister chromatids are exchanged into different chromosomes. The probability that two homologous chromosomes within a single secondary oocyte are previously paired at a locus in Prophase I is 1/3. In this case, the fragments of these sister chromatids will be segregated into a single gamete at the ratio of 1/2, so the double-reduction rate is 1/6 for tetrasomic inheritance.

to its corresponding centromere. According to the Haldane's mapping function, the relational expression between r_s and d is as follows:

$$r_s = \frac{1}{2}\left[1 - \exp(-2d/100)\right].$$

In summary, different models are required to satisfy different conditions and their dimensions are also not the same. For example, there is an additional parameter r_s (or d) in the PES model, and thus the number of degrees of freedom in PES is higher. It is noteworthy that all of the four models mentioned above can be incorporated into a generalized framework (i.e. the double-reduction rates are used as the parameters to express the phenotypic probabilities for some models). Comparing with the RCS, PRCS and CES models, the number of parameters for such generalized model increases by $\lfloor v/4 \rfloor$. The double-reduction rates in four models are shown in Table S1.

Table S1: The double-reduction rates in four models

Model	Alpha	Ploidy level				
		4	6	8	10	12
	α_1	0	0	0	0	0
RCS	α_2			0	0	0
	α_3					0
	α_1	1/7	3/11	24/65	140/323	1440/3059
PRCS	α_2			1/65	15/323	270/3059
	α_3					5/3059
	α_1	1/6	3/10	27/70	55/126	285/616
CES	α_2			3/140	5/84	65/616
	α_3					5/1848
	α_1	$r_s/6$	$3r_s/10$	$\frac{3}{70}r_s(10 - r_s)$	$\frac{5}{126}r_s(14 - 3r_s)$	$\frac{5}{616}r_s(84 - 28r_s + r_s^2)$
PES	α_2			$\frac{3}{140}r_s^2$	$\frac{5}{84}r_s^2$	$\frac{5}{616}r_s^2(14 - r_s)$
	α_3					$\frac{5}{1848}r_s^3$

B Likelihoods for genotypic data

The likelihood formulas stated in this section are applicable to the genotypic data of both diploids and autopolyploids.

We will first give the likelihood formulas in the absence of self-fertilization, and these formulas are identical to those in Kalinowski *et al.* (2007). For the first category in a parentage analysis (i.e. identifying the father when the mother is unknown), the likelihoods can be expressed as

$$\begin{aligned} \mathcal{L}(H_1) &= \Pr(\mathcal{G}_A)\big[(1-e)^2 T(\mathcal{G}_O \mid \mathcal{G}_A) + 2e(1-e)\Pr(\mathcal{G}_O) + e^2\Pr(\mathcal{G}_O)\big], \\ \mathcal{L}(H_2) &= \Pr(\mathcal{G}_A)\big[(1-e)^2 \Pr(\mathcal{G}_O) + 2e(1-e)\Pr(\mathcal{G}_O) + e^2\Pr(\mathcal{G}_O)\big]. \end{aligned} \tag{A1}$$

These two formulas are already listed in Equation (2), in which the second formula can be rewritten as $\mathcal{L}(H_2) = \Pr(\mathcal{G}_A)\Pr(\mathcal{G}_O)$ by merging similar terms.

For the second category (i.e. identifying the father when the mother is known), the likelihoods can be expressed as

$$\begin{aligned}\mathcal{L}(H_1) &= \Pr(\mathcal{G}_M)\Pr(\mathcal{G}_A)\big\{(1-e)^3 T(\mathcal{G}_O\,|\,\mathcal{G}_A,\mathcal{G}_M)\\ &\quad +e(1-e)^2\big[T(\mathcal{G}_O\,|\,\mathcal{G}_M)+T(\mathcal{G}_O\,|\,\mathcal{G}_A)+\Pr(\mathcal{G}_O)\big]\\ &\quad +3e^2(1-e)\Pr(\mathcal{G}_O)+e^3\Pr(\mathcal{G}_O)\big\}, \qquad\qquad \text{(A2)}\\ \mathcal{L}(H_2) &= \Pr(\mathcal{G}_M)\Pr(\mathcal{G}_A)\big\{(1-e)^3 T(\mathcal{G}_O\,|\,\mathcal{G}_M)+e(1-e)^2\big[T(\mathcal{G}_O\,|\,\mathcal{G}_M)+2\Pr(\mathcal{G}_O)\big]\\ &\quad +3e^2(1-e)\Pr(\mathcal{G}_O)+e^3\Pr(\mathcal{G}_O)\big\},\end{aligned}$$

where $\mathcal{G}_M$ is the observed genotype of the true mother.

For the third category (i.e. identifying the father and the mother jointly), the likelihoods can be expressed as

$$\begin{aligned}\mathcal{L}(H_1) &= \Pr(\mathcal{G}_{AM})\Pr(\mathcal{G}_A)\big\{(1-e)^3 T(\mathcal{G}_O\,|\,\mathcal{G}_A,\mathcal{G}_{AM})\\ &\quad +e(1-e)^2\big[T(\mathcal{G}_O\,|\,\mathcal{G}_{AM})+T(\mathcal{G}_O\,|\,\mathcal{G}_A)+\Pr(\mathcal{G}_O)\big]\\ &\quad +3e^2(1-e)\Pr(\mathcal{G}_O)+e^3\Pr(\mathcal{G}_O)\big\},\\ \mathcal{L}(H_2) &= \Pr(\mathcal{G}_{AM})\Pr(\mathcal{G}_A)\Pr(\mathcal{G}_O),\end{aligned} \qquad \text{(A3)}$$

where $\mathcal{G}_{AM}$ is the observed genotype of the alleged mother.

We will now give the likelihood formulas in the presence of self-fertilization. For the first category, the offspring is produced by selfing at a probability of s and by outcrossing at a probability of $1-s$. So, if we denote T_{s1} for $(1-s)T(\mathcal{G}_O\,|\,\mathcal{G}_A)+sT(\mathcal{G}_O\,|\,\mathcal{G}_A,\mathcal{G}_A)$, then the likelihood formulas can be obtained by replacing $T(\mathcal{G}_O\,|\,\mathcal{G}_A)$ with T_{s1} in the first formula in Equation (A1), whose expressions are as follows:

$$\begin{aligned}\mathcal{L}(H_1) &= \Pr(\mathcal{G}_A)\big[(1-e)^2 T_{s1}+2e(1-e)\Pr(\mathcal{G}_O)+e^2\Pr(\mathcal{G}_O)\big],\\ \mathcal{L}(H_2) &= \Pr(\mathcal{G}_A)\Pr(\mathcal{G}_O).\end{aligned}$$

For the second category, if the alleged father is not the same individual as the true mother, selfing cannot occur in H_1 but may occur in H_2. Thus, if we denote T_{s2} for $(1-s)T(\mathcal{G}_O\,|\,\mathcal{G}_M)+sT(\mathcal{G}_O\,|\,\mathcal{G}_M,\mathcal{G}_M)$, then the likelihood formulas can be obtained by replacing $T(\mathcal{G}_O\,|\,\mathcal{G}_M)$ with T_{s2} in the second formula in Equation (A2), whose expressions are as follows:

$$\begin{aligned}\mathcal{L}(H_1) &= \Pr(\mathcal{G}_M)\Pr(\mathcal{G}_A)\big\{(1-e)^3 T(\mathcal{G}_O\,|\,\mathcal{G}_A,\mathcal{G}_M)\\ &\quad +e(1-e)^2\big[T(\mathcal{G}_O\,|\,\mathcal{G}_M)+T(\mathcal{G}_O\,|\,\mathcal{G}_A)+\Pr(\mathcal{G}_O)\big]\\ &\quad +3e^2(1-e)\Pr(\mathcal{G}_O)+e^3\Pr(\mathcal{G}_O)\big\},\\ \mathcal{L}(H_2) &= \Pr(\mathcal{G}_M)\Pr(\mathcal{G}_A)\big\{(1-e)^3 T_{s2}+e(1-e)^2\big[T_{s2}+2\Pr(\mathcal{G}_O)\big]\\ &\quad +3e^2(1-e)\Pr(\mathcal{G}_O)+e^3\Pr(\mathcal{G}_O)\big\}.\end{aligned}$$

Moreover, if the alleged father is the same individual as the true mother, selfing must have occurred in H_1 and could not have occurred in H_2. Therefore, the likelihood formulas can be obtained by replacing $(1-e)^2 T(\mathcal{G}_O \,|\, \mathcal{G}_A)$ with $(1-e)^2 T(\mathcal{G}_O \,|\, \mathcal{G}_A, \mathcal{G}_A)$ and $(1-e)^2 \Pr(\mathcal{G}_O)$ with $(1-e)^2 T(\mathcal{G}_O \,|\, \mathcal{G}_A)$ in Equation (A1), whose expressions are as follows:

$$\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{G}_A)\big\{(1-e)^2 T(\mathcal{G}_O \,|\, \mathcal{G}_A, \mathcal{G}_A) + 2e(1-e)\Pr(\mathcal{G}_O) + e^2 \Pr(\mathcal{G}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{G}_A)\big\{(1-e)^2 T(\mathcal{G}_O \,|\, \mathcal{G}_A) + 2e(1-e)\Pr(\mathcal{G}_O) + e^2 \Pr(\mathcal{G}_O)\big\}.
\end{aligned}$$

For the third category, if the alleged father is not the same individual as the alleged mother, selfing cannot happen in H_1 but may happen in H_2. In this situation, the likelihood formulas are the same as those in Equation (A3). Moreover, if the alleged father is the same individual as the alleged mother, selfing must have occurred in H_1 but could not have occurred in H_2. Therefore, the likelihood formulas can be obtained by replacing $T(\mathcal{G}_O \,|\, \mathcal{G}_A)$ with $T(\mathcal{G}_O \,|\, \mathcal{G}_A, \mathcal{G}_A)$ in the first formula in Equation (A1), whose expressions are as follows:

$$\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{G}_A)\big\{(1-e)^2 T(\mathcal{G}_O \,|\, \mathcal{G}_A, \mathcal{G}_A) + 2e(1-e)\Pr(\mathcal{G}_O) + e^2 \Pr(\mathcal{G}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{G}_A)\Pr(\mathcal{G}_O).
\end{aligned}$$

For the transitional probability $T(\mathcal{G}_O \,|\, \mathcal{G}_A)$ or $T(\mathcal{G}_O \,|\, \mathcal{G}_A, \mathcal{G}_M)$ and so on in this section, it should be calculated by $T(G_O \,|\, G_F)$ or $T(G_O \,|\, G_F, G_M)$ because these genotypes are assumed correctly genotyped in calculating these transitional probabilities, i.e. $\mathcal{G}_O = G_O$, $\mathcal{G}_F = G_F$, $\mathcal{G}_M = G_M$. Similarly, for the genotypic frequency $\Pr(\mathcal{G}_A)$ or $\Pr(\mathcal{G}_O)$ and so on in some formula listed in this section, it should be calculated by $\Pr(G_A)$ or $\Pr(G_O)$ because the genotyping errors does not change the distribution of genotypes, i.e. $\Pr(\mathcal{G}) = \Pr(G = \mathcal{G})$.

For diploids without self-fertilization, the formulas of genotypic frequency and two transitional probabilities have been given in the section *Marshall et al.'s (1998) diploid model*.

For diploids with self-fertilization, the transitional probabilities do not change, but the genotypic frequency is related to the inbreeding coefficient F, denoted by $\Pr(G \,|\, \mathbf{p}, F)$, which can be calculated by

$$\Pr(G \,|\, \mathbf{p}, F) = \begin{cases} Fp_i + (1-F)p_i^2 & \text{if } G = A_iA_i, \\ 2(1-F)p_ip_j & \text{if } G = A_iA_j, \end{cases}$$

where F can be converted from the selfing rate s by the relational expression

$$F = \frac{s}{2-s}.$$

Above two formulas will be extended from disomic to polysomic inheritances in Appendix C.

For autopolyploids without self-fertilization, the genotypic frequency $\Pr(G)$ from tetrasomic to decasomic inheritances for each double-reduction model has been derived in Huang *et al.* (2019), and the transitional probabilities $T(G_O \,|\, G_F)$ and $T(G_O \,|\, G_F, G_M)$ are given in Appendix D.

For autopolyploids with self-fertilization, the transitional probabilities do not change, but the exact genotypic frequency is unavailable. As an alternative, we give its approximate solution, whose derivation is given in Appendix C.

C Genotypic and phenotypic frequencies

We have previously discussed the generalized genotypic frequencies from tetrasomic to decasomic inheritances under any double-reduction model (Huang *et al.*, 2019). We will further incorporate self-fertilization into these genotypic frequencies.

In the presence of self-fertilization, if the ploidy level is high, the calculation of the genotypic frequencies from their analytical expressions is problematic (see Appendix K for details). As an alternative, we give an approximate solution by using the inbreeding coefficient F as an intermediate variable under the assumption that the inbreeding is only caused by both self-fertilization and double-reduction. The analytical expression of F at an equilibrium state under both double-reduction and selfing was derived in Huang *et al.* (2019), which is

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

where s is the selfing rate, v is the ploidy level, and α is the expected number of pairs of IBDR alleles within a gamete. The value of α can be calculated by $\alpha = \sum_i i\alpha_i$, in which α_i is a double-reduction rate, whose value is listed in Table S1.

Let's now consider the genotypic frequencies incorporating both inbreeding and double-reduction. Let $p_1, p_2, \cdots, p_K$ be all allele frequencies in a population, and let γ_k be $(1/F - 1)p_k$, $k = 1, 2, \cdots, K$. Denote $\mathbf{p} = [p_1, p_2, \cdots, p_K]$ and $\gamma = \sum_{k=1}^{K} \gamma_k$. Assume that $q_1, q_2, \cdots, q_K$ are all allele frequencies of an individual, which are drawn from the Dirichlet distribution $\mathcal{D}(\gamma_1, \gamma_2, \cdots, \gamma_K)$ (Pritchard *et al.*, 2000). Denote $\mathbf{q} = [q_1, q_2, \cdots, q_K]$. Then the probability density function of $\mathbf{q}$ is

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

the expectation $\mathrm{E}(q_k)$ is p_k, and the variance $\mathrm{Var}(q_k)$ is $Fp_k(1 - p_k)$, $k = 1, 2, \cdots, K$. Moreover, for any q_k, its standardized variance is exactly F. From this, we see that these conditions accord with those of the definition of Wright's F-statistics. Hence the inbreeding coefficient F can be defined as the standardized variance of allele frequencies among individuals in the same population.

Because the correlation between alleles within the same individual relative to the population is explained by the divergence from $\mathbf{p}$ to $\mathbf{q}$, the alleles within the same genotype are independent relative to $\mathbf{q}$. Therefore, the frequency $\Pr(G \,|\, \mathbf{q})$ of a genotype G conditional on $\mathbf{q}$ is one of terms in the expansion of polynomial $(p_1 + p_2 + \cdots + p_K)^v$, i.e. the following term:

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

where n_k is the number of copies of the k^{th} allele in G, $k = 1, 2, \cdots, K$.

Next, the frequency $\Pr(G \,|\, \mathbf{p}, F)$ of G conditional on $\mathbf{q}$ and F is the weighted average of all frequencies in the form of $\Pr(G \,|\, \mathbf{q})$, with $f(\mathbf{q} \,|\, \mathbf{p}, F)\mathrm{d}\mathbf{q}$ as a weight, that is

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

where the integral domain Ω can be expressed as

$$\Omega = \{(q_1, q_2, \cdots, q_K) \,|\, q_1 + q_2 + \cdots + q_K = 1, q_k \geqslant 0, k = 1, 2, \cdots, K\}.$$

Such integral can be converted into the following repeated integral with the multiplicity $K - 1$:

$$\Pr(G \,|\, \mathbf{p}, F) = \int_0^1 \int_0^{1-q_1} \cdots \int_0^{1-q_1-q_2-\cdots-q_{K-2}} \Pr(G \,|\, \mathbf{q}) f(\mathbf{q} \,|\, \mathbf{p}, F)\mathrm{d}q_1 \mathrm{d}q_2 \cdots \mathrm{d}q_{K-1}.$$

It can now be calculated from the expressions of $\Pr(G \,|\, \mathbf{q})$ and $f(\mathbf{q} \,|\, \mathbf{p}, F)$ mentioned above that

$$\Pr(G \,|\, \mathbf{p}, F) = \binom{v}{n_1, n_2, \cdots, n_K} \prod_{k=1}^{K} \prod_{j=0}^{n_k-1} (\gamma_k + j) \Big/ \prod_{j'=0}^{v-1} (\gamma + j'). \tag{A4}$$

Equation (A4) is the approximate solution with F as an intermediate variable. Here, if self-fertilization is considered, the genotypic frequency $\Pr(\mathcal{G})$ should be calculated by Equation (A4), otherwise, the formula of $\Pr(\mathcal{G})$ under each double-reduction model is given in Huang *et al.* (2019).

Based on the derivation above, we are now able to express the phenotypic frequencies whilst considering the presence of negative amplifications. If β is the negative amplification rate, the frequency $\Pr(\mathcal{P})$ for each phenotype $\mathcal{P}$ is the weighted average of $\mathcal{B}_{\mathcal{P}=\varnothing}$ and $\sum_{\mathcal{G} \rhd \mathcal{P}} \Pr(\mathcal{G})$ with β and $1 - \beta$ as their weights, i.e.

$$\Pr(\mathcal{P}) = \beta\, \mathcal{B}_{\mathcal{P}=\varnothing} + (1 - \beta) \sum_{\mathcal{G} \rhd \mathcal{P}} \Pr(\mathcal{G}). \tag{A5}$$

Besides, if the negative amplifications are not considered, it only needs to set β as zero in Equation (A5).

D Transitional probabilities

In our model with a ploidy level greater than two, we establish two formulas of transitional probabilities $T(G_O \,|\, G_F)$ and $T(G_O \,|\, G_F, G_M)$, whose expressions are as follows:

$$
\begin{aligned}
T(G_O \,|\, G_F) &= \sum_{g_F \subset G_F \uplus G_F} T(g_F \,|\, G_F) \Pr(G_O \setminus g_F), \\
T(G_O \,|\, G_F, G_M) &= \sum_{g_F \subset G_F \uplus G_F} T(g_F \,|\, G_F) T(G_O \setminus g_F \,|\, G_M),
\end{aligned}
\tag{A6}
$$

where the operations $\uplus$ and $\setminus$ are respectively the union and difference of multisets, G_O, G_F and G_M are in turn the genotypes of the offspring, the father and the mother at a locus, g_F and $G_O \setminus g_F$ are the genotypes of the sperm and the egg that form the offspring, $\Pr(G_O \setminus g_F)$ is gamete frequency of the egg, and $T(g_F \,|\, G_F)$ and $T(G_O \setminus g_F \,|\, G_M)$ are two transitional probabilities from a zygote to a gamete, which have been derived in Equation (A7).

It is noteworthy that there cannot be any double-reduction under the RCS model or the PES model with $r_s = 0$ (see Table S1), then the double-reduction should not be considered. In other words, the expression $g_F \subset G_F \uplus G_F$ in Equation (A6) has to be replaced by $g_F \subset G_F$ under these situations.

Huang *et al.* (2019) derived the generalized gamete frequency $\Pr(g)$ and zygote frequency $\Pr(G)$ (Huang *et al.*, 2019). They also derived the generalized transitional probability $T(g \,|\, G)$ from a zygote G to a gamete g, which can be used at any even ploidy level v and under any double-reduction model, whose expression is

$$
T(g \,|\, G) = \sum_{i=0}^{\lfloor v/4 \rfloor} \sum_{j_1 + j_2 + \ldots + j_K = i} \frac{\prod_{k=1}^{K} \delta_k \binom{n_k}{j_k} \binom{n_k - j_k}{m_k - 2j_k}}{\binom{v}{i} \binom{v-i}{v/2-2i}} \alpha_i,
\tag{A7}
$$

where n_k (or m_k) is the number of copies of the k^{th} allele in G (or in g), α_i is a double-reduction rate, and δ_k is a binary variable, which is used to exclude the values outside the variation range D of j_k, such that $\delta_k = 1$ if $j_k \in D$, or $\delta_k = 0$ if $j_k \notin D$. The variation range D of j_k can be expressed as

$$
\max(0, m_k - n_k) \leqslant j_k \leqslant \min(n_k, m_k/2).
$$

In fact, for the binomial coefficient $\binom{n_k}{j_k}$, n_k and j_k should satisfy the condition $0 \leqslant j_k \leqslant n_k$. Similarly, for $\binom{n_k - j_k}{m_k - 2j_k}$, we have $0 \leqslant m_k - 2j_k \leqslant n_k - j_k$, or equivalently $m_k - n_k \leqslant j_k \leqslant m_k/2$. Therefore, the expression of D holds.

E Likelihoods under phenotype method

Under the PHENOTYPE method, if self-fertilization is not considered, the likelihoods for the first category in a parentage analysis can be expressed as

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_A) \big[(1-e)^2 T(\mathcal{P}_O \,|\, \mathcal{P}_A) + 2e(1-e) \Pr(\mathcal{P}_O) + e^2 \Pr(\mathcal{P}_O) \big], \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_A) \Pr(\mathcal{P}_O).
\end{aligned}
$$

For the second category, the likelihoods can be expressed as

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\big\{(1-e)^3 T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_M) \\
&\quad + e(1-e)^2\big[T(\mathcal{P}_O \,|\, \mathcal{P}_M) + T(\mathcal{P}_O \,|\, \mathcal{P}_A) + \Pr(\mathcal{P}_O)\big] \\
&\quad + 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\big\{(1-e)^3 T(\mathcal{P}_O \,|\, \mathcal{P}_M) + e(1-e)^2\big[T(\mathcal{P}_O \,|\, \mathcal{P}_M) + 2\Pr(\mathcal{P}_O)\big] \\
&\quad + 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\big\}.
\end{aligned}
$$

For the third category, they can be expressed as

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\big\{(1-e)^3 T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_{AM}) \\
&\quad + e(1-e)^2\big[T(\mathcal{P}_O \,|\, \mathcal{P}_{AM}) + T(\mathcal{P}_O \,|\, \mathcal{P}_A) + \Pr(\mathcal{P}_O)\big] \\
&\quad + 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),
\end{aligned}
$$

where $\Pr(\mathcal{P}_A)$, $\Pr(\mathcal{P}_O)$, $\Pr(\mathcal{P}_M)$ and $\Pr(\mathcal{P}_{AM})$ are calculated by Equation (A5), $T(\mathcal{P}_O \,|\, \mathcal{P}_A)$, $T(\mathcal{P}_O \,|\, \mathcal{P}_M)$ and $T(\mathcal{P}_O \,|\, \mathcal{P}_{AM})$ by Equation (3), and $T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_M)$ and $T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_{AM})$ by Equation (4).

If self-fertilization is considered, like the situations of Appendix B, each pair of likelihood formulas can be obtained by modifying the existing formulas. For the first category, the likelihood formulas are

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_A)\big[(1-e)^2 T_{s1} + 2e(1-e)\Pr(\mathcal{P}_O) + e^2\Pr(\mathcal{P}_O)\big], \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O),
\end{aligned}
$$

where $T_{s1} = (1-s)T(\mathcal{P}_O \,|\, \mathcal{P}_A) + sT(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_A)$. For the second category, if $A \not\equiv M$, then

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\big\{(1-e)^3 T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_M) \\
&\quad + e(1-e)^2\big[T(\mathcal{P}_O \,|\, \mathcal{P}_M) + T(\mathcal{P}_O \,|\, \mathcal{P}_A) + \Pr(\mathcal{P}_O)\big] \\
&\quad + 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_M)\Pr(\mathcal{P}_A)\big\{(1-e)^3 T_{s2} + e(1-e)^2\big[T_{s2} + 2\Pr(\mathcal{P}_O)\big] \\
&\quad + 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\big\},
\end{aligned}
$$

where $T_{s2} = (1-s)T(\mathcal{P}_O \,|\, \mathcal{P}_M) + sT(\mathcal{P}_O \,|\, \mathcal{P}_M, \mathcal{P}_M)$; if $A \equiv M$, then

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_A)\big\{(1-e)^2 T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2\Pr(\mathcal{P}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_A)\big\{(1-e)^2 T(\mathcal{P}_O \,|\, \mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2\Pr(\mathcal{P}_O)\big\},
\end{aligned}
$$

where $T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_A)$ and $T(\mathcal{P}_O \,|\, \mathcal{P}_M, \mathcal{P}_M)$ are calculated by Equation (4). For the third category, if $A \not\equiv AM$, then

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\big\{(1-e)^3 T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_{AM}) \\
&\quad + e(1-e)^2\big[T(\mathcal{P}_O \,|\, \mathcal{P}_{AM}) + T(\mathcal{P}_O \,|\, \mathcal{P}_A) + \Pr(\mathcal{P}_O)\big] \\
&\quad + 3e^2(1-e)\Pr(\mathcal{P}_O) + e^3\Pr(\mathcal{P}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_{AM})\Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O);
\end{aligned}
$$

if $A \equiv AM$, then

$$
\begin{aligned}
\mathcal{L}(H_1) &= \Pr(\mathcal{P}_A)\big\{(1-e)^2 T(\mathcal{P}_O \,|\, \mathcal{P}_A, \mathcal{P}_A) + 2e(1-e)\Pr(\mathcal{P}_O) + e^2\Pr(\mathcal{P}_O)\big\}, \\
\mathcal{L}(H_2) &= \Pr(\mathcal{P}_A)\Pr(\mathcal{P}_O).
\end{aligned}
$$

F Estimation of genotyping error rate (continuous)

In this appendix, we will use the trio mismatches to describe how to estimate the genotyping error rate. The trio mismatch in a true parents-offspring trio may be caused by the genotyping errors in this offspring or in the parents. If the offspring or if both parents encounter a genotyping error, the probability of observing a trio mismatch is equal to the exclusion rate for the third category, denoted by δ_o. If only one parent encounters a genotyping error, the probability of observing a trio mismatch is equal to the exclusion rate for the second category, denoted by δ_p. Moreover, if each individual in a selfed trio encounters a genotyping error, the probability of observing a trio mismatch is denoted by δ_s. Therefore, the probability γ of observing a trio mismatch in a true parents-offspring trio can be expressed as

$$\gamma = e[(1-s_t)(\delta_o + 2\delta_p) + 2s_t\delta_s] + e^2[(1-s_t)(\delta_o - 4\delta_p) - s_t\delta_s] + e^3(1-s_t)(\delta_o - 2\delta_p), \qquad \text{(A8)}$$

where s_t is the frequency of selfing in the reference trios.

The values of s_t and γ can be estimated from the reference trios identified from a single application or from multiple applications based on the same dataset, and δ_o and δ_s can be estimated from a similar Monte-Carlo algorithm mentioned above. The procedures are broadly as follows: randomly sample three (or two) individuals, considering them as a trio (or a selfed trio), and next calculate the probability that the genotypes/phenotypes at a locus of this trio (or this selfed trio) are mismatched, which is used as $\hat{\delta}_o$ (or $\hat{\delta}_s$) at this locus.

Under the assumption of random mating, the joint distribution of parental genotypes/phenotypes is the product of two observed genotypic/phenotypic frequencies, such that we can randomly sample two individuals and assume they are parents in the estimation of δ_o. However, in the estimation of δ_p, the joint distribution of parent-offspring genotypes/phenotypes cannot be estimated via this method. That is because the parent-offspring genotypes are correlated. As an alternative, we use the empirical distribution of genotypes/phenotypes of reference pairs to approximate the joint distribution of parent-offspring genotypes/phenotypes. More specifically, we randomly sample a matched pair (as a mother-offspring pair) from the reference pairs and an individual (as an alleged father) from all samples, considering them as a trio, and calculate the probability that the genotypes/phenotypes at a locus of this trio are mismatched, which is used as $\hat{\delta}_p$ at this locus.

The single-locus estimate $\hat{e}_l$ at the l^{th} locus can be obtained by solving Equation (A8), whose variance $\mathrm{Var}(\hat{e}_l)$ can be approximately expressed as $\mathrm{Var}(\hat{e}_l) \approx e/(n_{rl}\hat{\delta}_l)$. Moreover, the multi-locus estimate $\hat{e}$ is the weighted average of single-locus estimates across all loci, that is $\hat{e} = \sum_l w_l \hat{e}_l$, where $w_l = n_{rl}\hat{\delta}_l / \big(\sum_{l'} n_{rl'}\hat{\delta}_{l'}\big)$. The variance $\mathrm{Var}(\hat{e})$ can be approximately expressed as $\mathrm{Var}(\hat{e}) \approx e/\big(\sum_l n_{rl}\hat{\delta}_l\big)$.

G Estimation of sample rate (continuous)

Assume that the assignment rates a_c and a_u as well as the selfing rate s_u can be reliably estimated under an application and a confidence level, and that n_c is the number of cases. Because the number of assigned cases n_a obeys the binomial distribution $\mathrm{B}(n_c; a)$, the assignment rate a can be estimated by $\hat{a} = n_a/n_c$. Therefore, the sample rate p_s can be estimated by Equations (5), (6) or (7), and the variance $\mathrm{Var}(\hat{p}_s)$ can be calculated by the formula $\mathrm{Var}(\hat{p}_s) = \mathrm{E}(\hat{p}_s^2) - [\mathrm{E}(\hat{p}_s)]^2$.

However, it is unfortunate that the true value of a is unknown, then we cannot directly apply the binomial distribution $\mathrm{B}(n_c; a)$ to perform various calculations. As an alternative, we select the uniform distribution $\mathrm{U}(0,1)$ as the prior distribution obeyed by a, and then give the posterior distribution obeyed by a according to the Bayes formula, where the expected value $\mathrm{E}(a)$ for the posterior distribution is

$$\mathrm{E}(a) = \frac{n_a + 1}{n_c + 2}.$$

Now, we can perform various calculations so long as we let the value of a in $\mathrm{B}(n_c; a)$ be equal to $\frac{n_a+1}{n_c+2}$.

In actual conditions, multiple applications and multiple confidence levels will be used jointly to increase the accuracy of sample rate estimation. For convenience, we denote $\hat{p}_{si}$ for the estimated value of p_s under an application and a confidence level. According to the previous derivations, $\hat{p}_{si}$ together with its variance can be calculated under the assumption that a_c, a_u and s_u can be reliably estimated. Like the estimation of genotyping error rate, the estimate $\hat{p}_s$ is the weighted average of the estimated values of p_s under all selected applications and all selected confidence levels, symbolically $\hat{p}_s = \big(\sum_i w_i \hat{p}_{si}\big)/\big(\sum_i w_i\big)$, where $w_i = 1/\mathrm{Var}(\hat{p}_{si})$.

Finally, let's consider the estimation of selfing rate s_u under multiple confidence levels. In actual conditions, the loci may be insufficient, causing that there are only few cases to assign the parent at a high confidence level (e.g. $\Delta > \Delta_{0.99}$). Besides, the genotyping error rate may be high, causing that the false parent may be assigned at a low confidence level (e.g. $\Delta > 0$) when the true parent is not sampled. To avoid these problems, we jointly use three confidence levels (80%, 95% and 99%) in POLYGENE for each application.

The estimated value $\hat{s}_u$ is the ratio of n_s to n_a, i.e. $\hat{s}_u = n_s/n_a$ under an application and a confidence level, where n_s is the number of selfing cases. If we select the three confidence levels 99%, 95% and 80%, then $\hat{s}_u$ is the weighted average of the corresponding ratio values of n_s to n_a, that is

$$\hat{s}_u = \frac{n_{s,0.99} + n_{s,0.95} + n_{s,0.80}}{n_{a,0.99} + n_{a,0.95} + n_{a,0.80}}.$$

H Pseudo-dominant approach

The pseudo-dominant approach was used in Rodzen *et al.* (2004) and Wang and Scribner (2014). In this approach, the codominant data are converted into the dominant data. More specifically, each visible allele is defined as a virtual dominant marker, whose observed phenotype is either present (denoted by $\{A\}$) if this allele is detected, or absent (denoted by $\varnothing$) if this allele is not detected. We denote $\mathcal{P}^D$ for the phenotype at a dominant marker. Moreover, the LOD scores are calculated by the diploid likelihood formulas listed below. These formulas are originally derived in Gerber *et al.* (2000) by using the transitional probability $T(\mathcal{G}\,|\,G)$ from a true genotype G to an observed genotype $\mathcal{G}$ based on an alternative genotyping error model, where

$$T(\mathcal{G}\,|\,G) = (1-e)\Pr(\mathcal{G})\mathcal{B}_{G=\mathcal{G}} + e\mathcal{B}_{G\neq\mathcal{G}}.$$

The above formula is different to that listed in Equation (1). Because the possible phenotypes at a dominant marker are $\{A\}$ and $\varnothing$, the degree-of-freedom is only one. Therefore, the null allele frequency, the selfing rate and the negative amplification rate cannot be estimated. Besides, we will use the formulas and the model given in Rodzen *et al.* (2004) to evaluate the efficiency of this approach.

Next, the transitional probability from one phenotype or a pair of phenotypes to another phenotype at a dominant marker is described in Tables 1 and 2 in Gerber *et al.* (2000).

The phenotypic frequency at a dominant marker in diploids is

$$\Pr(\mathcal{P}^D) = \begin{cases} (1-p)^2 & \text{if } \mathcal{P}^D = \varnothing, \\ 1-(1-p)^2 & \text{if } \mathcal{P}^D = \{A\}, \end{cases}$$

where p is the frequency of the dominant allele A at this dominant marker, and p is estimated from the observed phenotypic frequencies, whose estimated expression is $\hat{p} = 1 - \sqrt{\widehat{\Pr}(\mathcal{P}^D = \varnothing)}$.

Now, the likelihood formulas listed below can be used for the actual calculation by using these transitional probabilities and phenotypic frequencies: for the first category in a parentage analysis,

$$\begin{aligned} \mathcal{L}(H_1) &= (1-e)^2 T(\mathcal{P}_O^D\,|\,\mathcal{P}_A^D)\Pr(\mathcal{P}_A^D) + e(1-e)\big[\Pr(\mathcal{P}_O^D) + \Pr(\mathcal{P}_A^D)\big] + e^2, \\ \mathcal{L}(H_2) &= (1-e)^2 \Pr(\mathcal{P}_O^D)\Pr(\mathcal{P}_A^D) + e(1-e)\big[\Pr(\mathcal{P}_O^D) + \Pr(\mathcal{P}_A^D)\big] + e^2; \end{aligned}$$

for the second category,

$$\begin{aligned}
\mathcal{L}(H_1) &= (1-e)^3 T(\mathcal{P}_O^D \,|\, \mathcal{P}_A^D, \mathcal{P}_M^D) \Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + \\
&\quad e(1-e)^2 \big[\Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + T(\mathcal{P}_O^D \,|\, \mathcal{P}_M^D) \Pr(\mathcal{P}_M^D) + T(\mathcal{P}_O^D \,|\, \mathcal{P}_A^D) \Pr(\mathcal{P}_A^D)\big] + \\
&\quad e^2(1-e)\big[\Pr(\mathcal{P}_O^D) + \Pr(\mathcal{P}_A^D) + \Pr(\mathcal{P}_M^D)\big] + e^3, \\
\mathcal{L}(H_2) &= (1-e)^3 T(\mathcal{P}_O^D \,|\, \mathcal{P}_M^D) \Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + \\
&\quad e(1-e)^2 \big[\Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + T(\mathcal{P}_O^D \,|\, \mathcal{P}_M^D) \Pr(\mathcal{P}_M^D) + \Pr(\mathcal{P}_O^D) \Pr(\mathcal{P}_A^D)\big] + \\
&\quad e^2(1-e)\big[\Pr(\mathcal{P}_O^D) + \Pr(\mathcal{P}_A^D) + \Pr(\mathcal{P}_M^D)\big] + e^3;
\end{aligned}$$

for the third category,

$$\begin{aligned}
\mathcal{L}(H_1) &= (1-e)^3 T(\mathcal{P}_O^D \,|\, \mathcal{P}_A^D, \mathcal{P}_M^D) \Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + \\
&\quad e(1-e)^2 \big[\Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + T(\mathcal{P}_O^D \,|\, \mathcal{P}_M^D) \Pr(\mathcal{P}_M^D) + T(\mathcal{P}_O^D \,|\, \mathcal{P}_A^D) \Pr(\mathcal{P}_A^D)\big] + \\
&\quad e^2(1-e)\big[\Pr(\mathcal{P}_O^D) + \Pr(\mathcal{P}_A^D) + \Pr(\mathcal{P}_M^D)\big] + e^3, \\
\mathcal{L}(H_2) &= (1-e)^3 \Pr(\mathcal{P}_O^D) \Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + \\
&\quad e(1-e)^2 \big[\Pr(\mathcal{P}_A^D) \Pr(\mathcal{P}_M^D) + \Pr(\mathcal{P}_O^D) \Pr(\mathcal{P}_M^D) + \Pr(\mathcal{P}_O^D) \Pr(\mathcal{P}_A^D)\big] + \\
&\quad e^2(1-e)\big[\Pr(\mathcal{P}_O^D) + \Pr(\mathcal{P}_A^D) + \Pr(\mathcal{P}_M^D)\big] + e^3.
\end{aligned}$$

I Exclusion approach

Although the exclusion approach is not as accurate as the likelihood approach, the number of mismatches can be used as a reference. Here, we extend the exclusion approach to polysomic inheritances, and this extended approach can be incorporated into our framework, such that the effects of double-reduction, null alleles, negative amplifications and self-fertilization can all be freely accommodated.

The logic of the exclusion approach is relatively simple: if the alleged parents are able to produce the offspring, they cannot be excluded. We will here give two extended definitions of matches by using the genotypic data.

Given an alleged parent-offspring pair, if there exists a gamete g_A produced by the alleged parent at a locus, such that g_A is a subset of the offspring genotype $\mathcal{G}_O$ at this locus, then such a pair is termed *matched* at this locus. The condition in this definition can be described by symbols as follows: $\exists g_A \subset \mathcal{G}_A \uplus \mathcal{G}_A$, such that $g_A \subset \mathcal{G}_O$; or equivalently, $\max\left\{\mathcal{B}_{g'_A \subset \mathcal{G}_O} \,|\, g'_A \subset \mathcal{G}_A \uplus \mathcal{G}_A\right\} = 1$, where $\mathcal{G}_A$ is the genotype of the alleged parent at this locus.

Given an alleged parents-offspring trio, if there exist two gametes g_F and g_M produced by the alleged father and the alleged mother at a locus, respectively, such that the fusion of g_F and g_M results in the offspring genotype $\mathcal{G}_O$ at this locus, then such a trio is termed *matched* at this locus. Similarly, the conditions in this definition can be described as follows: $\exists g_F \subset \mathcal{G}_{AF} \uplus \mathcal{G}_{AF}$, $\exists g_M \subset \mathcal{G}_{AM} \uplus \mathcal{G}_{AM}$, such that $g_F \uplus g_M = \mathcal{G}_O$; or equivalently,

$$\max\left\{\mathcal{B}_{g'_F \uplus g'_M = \mathcal{G}_O} \,|\, g'_F \subset \mathcal{G}_{AF} \uplus \mathcal{G}_{AF},\, g'_M \subset \mathcal{G}_{AM} \uplus \mathcal{G}_{AM}\right\} = 1,$$

where $\mathcal{G}_{AF}$ (or $\mathcal{G}_{AM}$) is the genotype of the alleged father (or the alleged mother) at this locus.

Finally, it is important to highlight that under the RCS model or the PES model with $r_s = 0$, the expressions, used to describe the two definitions and involved in the double-reduction, should be revised, i.e. we must replace $g_A \subset \mathcal{G}_A \uplus \mathcal{G}_A$ by $g_A \subset \mathcal{G}_A$, $g_F \subset \mathcal{G}_{AF} \uplus \mathcal{G}_{AF}$ by $g_F \subset \mathcal{G}_{AF}$ and $g_M \subset \mathcal{G}_{AM} \uplus \mathcal{G}_{AM}$ by $g_M \subset \mathcal{G}_{AM}$.

J Allele frequency estimation

We adopt an *expectation-maximization* (EM) algorithm (Dempster *et al.*, 1977) to estimate the allele frequencies for phenotypic data. This algorithm follows the methods of Kalinowski and Taper (2006), which is an iterative algorithm used to maximize the genotypic likelihood. The *genotypic likelihood* at a locus is defined as the product of genotypic frequencies of all individuals at this locus, denoted by $\mathcal{L}_{\text{geno}}$, whose logarithmic expression is

$$\ln \mathcal{L}_{\text{geno}} = \sum_{\mathcal{P}} \sum_{\mathcal{G} \rhd \mathcal{P}} \Pr(\mathcal{G} \,|\, \mathcal{P}) \ln[\Pr(\mathcal{G})],$$

in which $\mathcal{P}$ is taken from the phenotypes of all individuals at this locus, $\mathcal{G}$ is taken from all genotypes determining $\mathcal{P}$ at the same locus, $\Pr(\mathcal{G} \,|\, \mathcal{P})$ is the posterior probability of $\mathcal{G}$ determining $\mathcal{P}$, and $\Pr(\mathcal{G})$ is the frequency of $\mathcal{G}$.

The initial frequencies of amplifiable alleles are assumed to be equal to $1/K$, where K is the number of alleles, including the null allele A_y. The updated frequency $\hat{p}'_k$ of the k^{th} allele A_k is the weighted average of frequencies of A_k in all genotypes at a locus, with the posterior probabilities of these genotypes as their weights, whose expression is

$$\hat{p}'_k = \frac{\sum_{\mathcal{P}} \sum_{\mathcal{G} \rhd \mathcal{P}} \Pr(\mathcal{G} \,|\, \mathcal{P}) \Pr(A_k \,|\, \mathcal{G})}{\sum_{\mathcal{P}} \sum_{\mathcal{G} \rhd \mathcal{P}} \Pr(\mathcal{G} \,|\, \mathcal{P})}, \quad k = 1, 2, \cdots, K,$$

where $\Pr(A_k \,|\, \mathcal{G})$ is the frequency of A_k in $\mathcal{G}$.

Our algorithm also includes simultaneously the estimation of negative amplification rate β. Because the final estimated value of β is independent to the initial value, the initial value can be arbitrarily selected (e.g. 0.05). The updated negative amplification rate $\hat{\beta}'$ can be expressed as

$$\hat{\beta}' = \frac{N_{\varnothing} \hat{\beta} / \Pr(\mathcal{P} = \varnothing)}{N},$$

where $N_{\varnothing}$ is the number of negative phenotypes at this locus, N is the number of all individuals, $\hat{\beta}$ is the current negative amplification rate, and $\hat{\beta}/\Pr(\mathcal{P} = \varnothing)$ is the posterior probability that a negative phenotype is the result of negative amplification.

If $\max\{|\hat{p}_k - \hat{p}'_k| \,|\, k = 1, 2, \cdots, K\}$ and $|\hat{\beta} - \hat{\beta}'|$ are less than a predefined threshold (e.g. 10^{-5}) or if the iterative times reach 2000, the iteration is terminated, where $\hat{p}_k$ is the current frequency of A_k.

Null alleles and negative amplifications can both be freely incorporated into our model. If the null alleles are not considered, the candidate genotypes extracted from a phenotype only need to be set as 'not containing A_y'. If the negative amplifications are not considered, the initial value of β only needs to be set as zero. If both factors are not considered, the negative phenotype cannot be explained, and so $\varnothing$ is discarded in the allele frequency estimation together with the subsequent analyses.

We also nest a downhill simplex algorithm (Nelder and Mead, 1965) outside the EM algorithm to estimate the selfing rate s. The estimated value $\hat{s}$ is obtained by maximizing the phenotypic likelihood $\mathcal{L}_{\text{pheno}}$, that is $\hat{s} = \underset{s \in [0,1]}{\arg\max} \, \mathcal{L}_{\text{pheno}}$, where $\mathcal{L}_{\text{pheno}} = \prod\limits_{\mathcal{P}} \Pr(\mathcal{P})$.

K Reasons for computational difficulty

In the absence of selfing, the generalized form of genotypic frequencies can be obtained by two methods (Huang *et al.*, 2019). The first method is the *non-linear method.* In this method, we establish a non-linear equation set with the frequencies $\Pr(G_1), \Pr(G_2), \cdots, \Pr(G_I), \Pr(g_1), \cdots, \Pr(g_J)$ as the unknowns and the frequencies $p_1, p_2, \cdots, p_K$ as the parameters, whose expression is as follows:

$$\begin{cases} \Pr(G_i) = \sum\limits_{\mu=1}^{J} \Pr(g_\mu) \Pr(G_i \setminus g_\mu), & i = 1,\, 2,\, \cdots,\, I, \\ \Pr(g_j) = \sum\limits_{\nu=1}^{I} \Pr(G_\nu) T(g_j \,|\, G_\nu), & j = 1,\, 2,\, \cdots,\, J, \\ p_k = \sum\limits_{\nu=1}^{I} \Pr(G_\nu) \Pr(A_k \,|\, G_\nu), & k = 1,\, 2,\, \cdots,\, K, \end{cases} \tag{A9}$$

where $I = \binom{2v}{v}$, $J = \binom{v/2+v}{v/2}$, $K = v + 1$ (I, J and K are the numbers of zygotes, gametes and alleles at a locus, respectively), $\Pr(G_i \setminus g_\mu) = \Pr(g = G_i \setminus g_\mu)$, $T(g_j \,|\, G_\nu)$ is the transitional probability from G_ν to g_j, and p_k and $\Pr(A_k \,|\, G_\nu)$ are the frequencies of A_k in the population and in G_ν, respectively. If the ploidy level v is equal to 4, 6, 8 or 10, the number of equations in Equation set (A9) is 90, 1015, 13374 or 187770, and the number of unknowns is 85, 1008, 12265 or 187759. We now see that these numbers will increase rapidly with an increase in ploidy level. Therefore, this will cause a computational difficulty for Equation set (A9) at a high ploidy level.

In order to overcome such a computational difficulty, we adopt another method, named the *linear method*, to obtain the zygote frequencies. For this method, briefly speaking, we will first use Equation set (A9) to calculate the gamete frequencies at a biallelic locus. Next, we split these alleles one by one at this locus until they are split into $v/2 + 1$ alleles so as to more expediently obtain the zygote frequencies at a multi-allelic locus. Finally, we use the former I equations in Equation set (A9), i.e.

$$\Pr(G_i) = \sum_{\mu=1}^{J} \Pr(g_\mu) \Pr(G_i \setminus g_\mu), \quad i = 1,\, 2,\, \cdots,\, I,$$

to calculate the zygote frequencies. This method can be described by a linear equation set $\mathbf{Ax} = \mathbf{b}$. Because there are no sufficient constraint conditions to obtain a unique solution for such linear equation set when $v \geqslant 12$, this method can only be applied from tetrasomic to decasomic inheritances (Huang *et al.*, 2019).

In the presence of selfing, for the linear method, although the gamete frequencies can be solved for $v < 12$, the zygote frequencies cannot be easily calculated from the gamete frequency. That is because for any $i \in I$, the i^{th} equation in Equation set (A9) should be modified as

$$\Pr(G_i) = (1-s) \sum_{\nu=1}^{J} \Pr(g_\nu) \Pr(G_i \setminus g_\nu) + s \sum_{\mu=1}^{I} \sum_{\nu=1}^{J} \Pr(G_\mu) T(g_\nu \,|\, G_\mu) T(G_i \setminus g_\nu \,|\, G_\mu).$$

For the non-linear method, the calculation is more difficult when the ploidy level is high.

L Supplementary figures

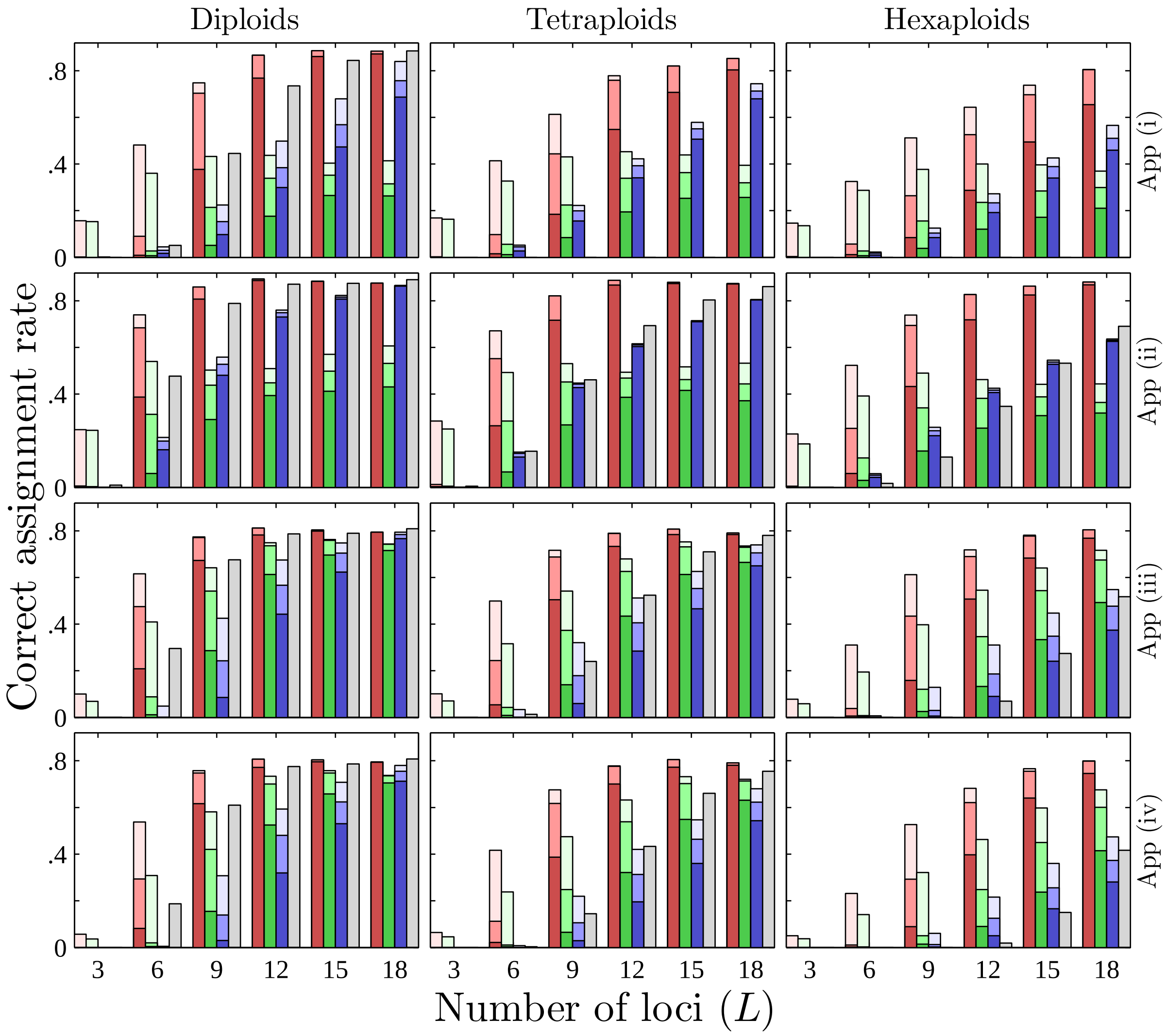

Figure S2: The correct assignment rate as a function of the number of loci L by using the phenotypic data at the selfing rate 0. The ploidy levels, applications, methods, confidence levels and the definitions of bars together with their shading are as for Figure 2.

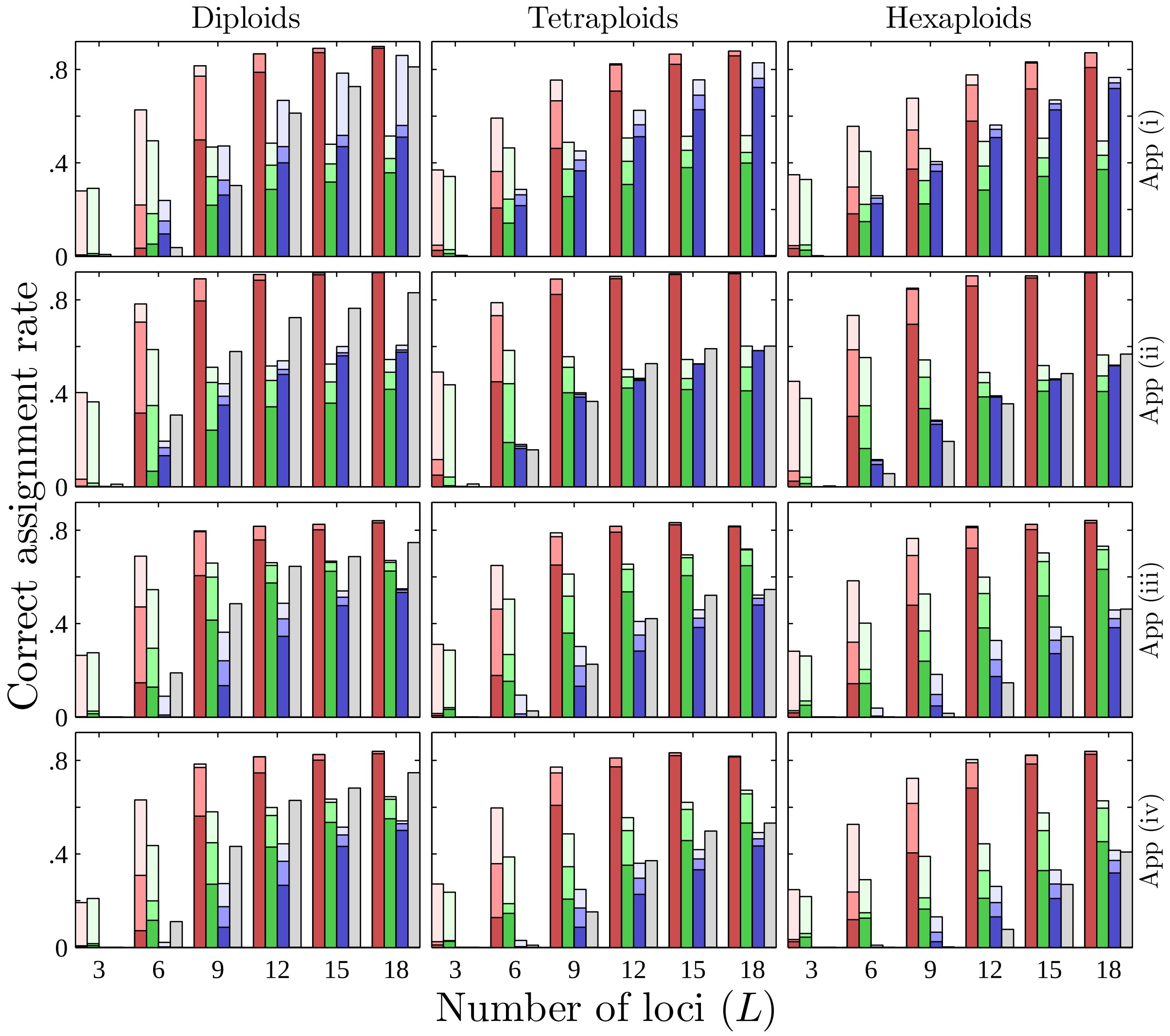

Figure S3: The correct assignment rate as a function of the number of loci L by using the phenotypic data at the selfing rate 0.3. The remaining are as for Figure **??**.

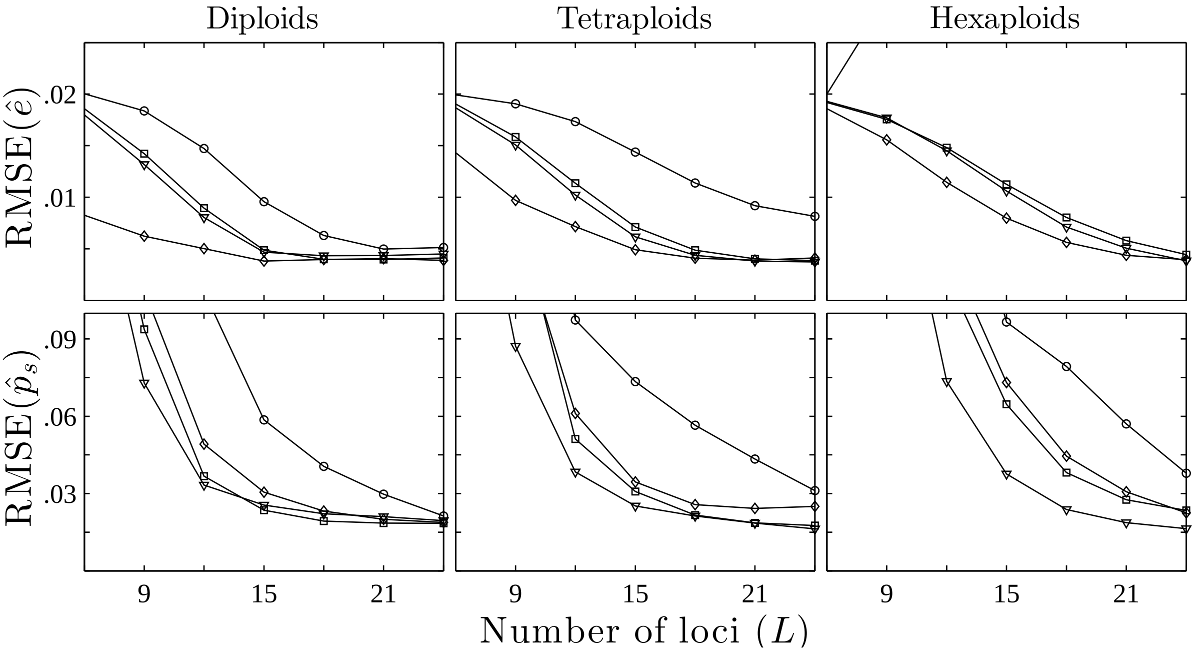

Figure S4: The RMSE of the estimated genotyping error rate $\hat{e}$ or the estimated sample rate $\hat{p}_s$ as a function of the number of loci L at $e = 0.02$ and $p_s = 0.8$. The remaining are as for Figure **??**.

LITERATURE CITED

Dempster, A. P., N. M. Laird, D. B. Rubin *et al.*, 1977 Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39: 1–38.

Gerber, S., S. Mariette, R. Streiff, C. Bodenes, and A. Kremer, 2000 Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Molecular Ecology 9: 1037–1048.

Haldane, J. B. S., 1930 Theoretical genetics of autopolyploids. Journal of Genetics 22: 359–372.

Huang, K., T. C. Wang, D. W. Dunn, P. Zhang, R. C. Liu *et al.*, 2019 Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. G3: Genes, Genomes, Genetics 9: 1693–1706.

Kalinowski, S. T., and M. L. Taper, 2006 Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. Conservation Genetics 7: 991–995.

Kalinowski, S. T., M. L. Taper, and T. C. Marshall, 2007 Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Molecular Ecology 16: 1099–1106.

Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7: 639–655.

Mather, K., 1935 Reductional and equational separation of the chromosomes in bivalents and multivalents. Journal of Genetics 30: 53–78.

Muller, H. J., 1914 A new mode of segregation in gregory's tetraploid *Primulas*. The American Naturalist 48: 508–512.

Nelder, J. A., and R. Mead, 1965 A simplex method for function minimization. The Computer Journal 7: 308–313.

Parisod, C., R. Holderegger, and C. Brochmann, 2010 Evolutionary consequences of autopolyploidy. New Phytologist 186: 5–17.

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

Rodzen, J. A., T. R. Famula, and B. May, 2004 Estimation of parentage and relatedness in the polyploid white sturgeon (*Acipenser transmontanus*) using a dominant marker approach for duplicated microsatellite loci. Aquaculture 232: 165–182.

Wang, J., and K. T. Scribner, 2014 Parentage and sibship inference from markers in polyploids. Molecular Ecology Resources 14: 541–553.