

ARTICLE



Linkage disequilibrium under polysomic inheritance

Kang Huang ^{1,2}, Derek W. Dunn¹, Wenkai Li¹, Dan Wang¹ and Baoguo Li^{1,3}✉

© The Author(s), under exclusive licence to The Genetics Society 2022

Linkage disequilibrium (LD) is the non-random association of alleles at different loci. Squared LD coefficients r^2 (for phased genotypes) and r_{Δ}^2 (for unphased genotypes) will converge to constants that are determined by the sample size, the recombination frequency, the effective population size and the mating system. LD can therefore be used for gene mapping and the estimation of effective population size. However, current methods work only with diploids. To resolve this problem, we here extend the linkage disequilibrium measures to include polysomic inheritance. We derive the values of r^2 and r_{Δ}^2 at equilibrium state for various mating systems and different ploidy levels. For unlinked loci, $E(r_{\Delta}^2) \approx \frac{1}{3(N_e - \eta)}$ for monoecious and dioecious (with random pairing) mating systems or $\frac{3+f}{3(1+f)(N_e - \eta)}$ for dioecious mating systems (with lifetime pairing), where f is the number of females in a half-sib family and η is a constant related to the ploidy level. We simulate the application of estimating N_e using unphased genotypes. We find that estimating N_e in polyploids requires similar sample sizes and numbers of loci as in diploids, with the main source of bias due to using 0.5 as the recombination frequency.

Keywords: Linkage disequilibrium; Polysomic inheritance; Recombination frequency; Effective population size; Mating system

Heredity (2022) 128:11–20; <https://doi.org/10.1038/s41437-021-00482-1>

INTRODUCTION

Linkage disequilibrium (LD) is the non-random association of alleles at different loci within individuals in a given population (Slatkin 2008), and can be influenced by many factors, such as selection, mutation, recombination, genetic drift, and the mating system (Nei 1987). Linkage disequilibrium can be measured by several parameters, such as the correlation coefficient r , Lewontin's (1964) D' , Hill's (1975) Q , Maruyama's (1982) D^* , Ohta's (1980) F' , and Brown et al.'s (1980) χ . The most frequently used measure of LD is the squared correlation coefficient r^2 (Hill and Weir 1994), which is the weighted sum of the squared correlation coefficient between alleles at two loci.

The influence of genetic drift on linkage disequilibrium in finite populations has been extensively studied in diploids (Ohta and Kimura 1969; Hill and Robertson 1968; Weir 1979; Weir and Cockerham 1979; Weir and Hill 1980; Sved and Feldman 1973; Hill 1974). In general, previous work has shown that the squared correlation coefficient r^2 (for phased genotypes) or r_{Δ}^2 (for unphased genotypes) will converge to a constant after several generations of random mating for unlinked loci, whereas more generations are required to converge for linked loci. This constant is determined by the sample size n , recombination frequency c , effective population size N_e and the mating system. Based on these four factors, LD has been incorporated into two major applications: (i) gene mapping (Hill and Weir 1994; Devlin and Risch 1995; Jorde 1995; Hosking et al. 2002; Hästbacka et al. 1992) and (ii) the estimation of effective population size (England et al. 2006; Hill 1981; Waples et al. 2014; Hayes et al. 2003; Sved et al. 2013), which enable either c or N_e to be solved when the other

three factors are known, respectively. However, current methods work only with organisms that are diploid.

Many plant species are polyploid, with 30–80% of angiosperm species being at least partially polyploid (Burow et al. 2001), with evidence for paleo-polyploidy in most plant lineages (Otto 2007). Although rare, polyploidy is also present in animals, such as in some salamanders, flatworms, leeches, brine shrimps, frogs and fishes. Polyploidy is also important in the evolution of both wild and cultivated plants, and plays a key role in plant breeding (Sattler et al. 2016; Udall and Wendel 2006). However, to date the effects of ploidy on LD has not been extensively studied.

Polysomic inheritance is expected in autopolyploids but not in allopolyploids, although complex mechanisms can lead to a mixture of disomic and polysomic inheritance in the same genome (segmental allopolyploids, Stift et al. 2008). There are at least three typical features in polysomic inheritances: (i) multivalents may be formed during meiosis (Rieger et al. 1968), resulting in a particular phenomenon in polysomic inheritance, termed the double-reduction (Butruille and Boiteux 2000), in which a gamete may inherit a single gene copy twice; (ii) the chromosomes are randomly paired and exchange their chromatid segments during meiosis, in which the recombination frequency c is $1-1/\nu$ if the corresponding loci are located on different chromosomes (ν is the ploidy level), ≤ 0.5 (in bivalent pairing) or 0.75 (in multivalent pairing) if the corresponding loci are located on the same chromosome (Fisher 1947; Sved 1964); (iii) the decay coefficient of heterozygosity (i.e., the ratio of single non-identity coefficients in the next and the current generations in the absence

¹Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an 710069, China. ²Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T1Z4, Canada. ³Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China. Associate editor Olivier Hardy. ✉email: baoguo@nwu.edu.cn

of mutation and migration) is $1 - \frac{1}{vN_e}$ in polyploids (N_e is the effective population size).

Here, we extend both the linkage disequilibrium measure D and Burrow's Δ statistic to account for polysomic inheritance, and calculate their corresponding squared correlation coefficients r^2 and r_{Δ}^2 . We also extend Weir and Hill's (1980) double non-identity framework to account for polysomic inheritance, and derive the expressions of these double non-identity coefficients under five mating systems. On this basis, we are able to derive $E(\hat{r}^2)$ and $E(\hat{r}_{\Delta}^2)$ at equilibrium state, and these two expectations are approximated by d^2 or δ^2 , respectively. Both approximations are closely related to the mating system together with the effective population size N_e and the recombination frequency c . We study the behavior of the squared correlation coefficient estimators \hat{r}^2 and \hat{r}_{Δ}^2 during genetic drift, investigate the influence of recombination frequency c on d^2 or δ^2 , simulate the application for reestimating effective population size N_e , and evaluate the statistical performance of estimating \hat{N}_e . We discuss the relationship between r^2 and c (or between r_{Δ}^2 and c), and that between r^2 and v (or between r_{Δ}^2 and v). We enable the estimation of Burrow's Δ , the testing of linkage disequilibrium based on Burrow's Δ , and the estimation of effective population size using our software package POLYGENE V1.3 (Huang et al. 2020), which is freely available via <http://github.com/huangkang1987/polygene>.

THEORY AND MODELING

LD measurements

We denote A and B for two alleles each from a different locus. The generalized LD measurement D between A and B is defined as the difference between the observed and the expected frequencies of the haplotype AB , where a haplotype is defined as a combination of alleles at multiple loci from a single set of chromosomes. We slightly revise the notations of both Weir and Cockerham (1979) and Weir and Hill (1980) and define five specific variants of D : (i) D_s^{AB} (for the same haplotype), (ii) D_d^{AB} (for two different haplotypes within the same individual), (iii) D_w^{AB} (for the within-individual component), (iv) D_b^{AB} (for the between-individual component) and (v) D_{AB} (for the usual LD measurement). These measurements can be defined by symbols as follows:

$$D_s^{AB} \stackrel{\text{def}}{=} P_s^{AB} - p_A p_B,$$

$$D_d^{AB} \stackrel{\text{def}}{=} P_d^{AB} - p_A p_B,$$

$$D_w^{AB} \stackrel{\text{def}}{=} P_w^{AB} - P_d^{AB},$$

$$D_b^{AB} \stackrel{\text{def}}{=} P_b^{AB} - p_A p_B,$$

$$D_{AB} \stackrel{\text{def}}{=} D_w^{AB} + D_b^{AB},$$

where P_s^{AB} is the probability that the alleles in the same haplotype are A and B , P_d^{AB} is the probability that alleles in different haplotypes within the same individual are A and B , and p_A and p_B are respectively the probabilities of A and B .

According to these definitions, the following expressions hold:

$$D_w^{AB} = D_s^{AB} - D_d^{AB}, D_b^{AB} = D_d^{AB} \text{ and } D_{AB} = D_s^{AB}.$$

The usual LD measurement D_{AB} is the covariance between A and B in the same haplotype, i.e., $D_{AB} = \text{Cov}(\mathcal{B}_A, \mathcal{B}_B)$, where $\mathcal{B}_A = 1$ if the first allele in the haplotype is A , otherwise $\mathcal{B}_A = 0$, and the meaning of \mathcal{B}_B is analogous.

The values of D_{AB} may be negative, and its range is influenced by the probabilities of A and B . It is therefore more intuitive to use Pearson's correlation coefficient r_{AB} to measure LD to convert the

range to $[-1, 1]$:

$$r_{AB} = \frac{D_{AB}}{\sqrt{Q_{AB}}} = \frac{\text{Cov}(\mathcal{B}_A, \mathcal{B}_B)}{\sqrt{\text{Var}(\mathcal{B}_A)\text{Var}(\mathcal{B}_B)}}.$$

where $Q_{AB} = \text{Var}(\mathcal{B}_A)\text{Var}(\mathcal{B}_B) = p_A p_X q_B q_X$ (X represents any allele distinct from both A and B , and thus $p_X = 1 - p_A$ and $q_X = 1 - q_B$).

The values of r_{AB} may also be negative. However, the squared correlation coefficient r_{AB}^2 ranges from 0 to 1. We will adopt the average value of r_{AB}^2 across all allele pairs to evaluate the LD between two loci for the situation of phased genotypes. For diallelic loci, the averaged r_{AB}^2 across all allele pairs is equal to that of any allele pair.

The above LD measurements are applicable for phased genotypes although unphased genotypes are more common. For unphased genotypes, Burrows's Δ statistic (Cockerham and Weir 1977) can be used, and we will extend this to account for polysomic inheritance. By using D_w^{AB} and D_b^{AB} , Burrows's Δ statistic between A and B can be defined as $\Delta_{AB} \stackrel{\text{def}}{=} D_w^{AB} + v D_b^{AB}$, which is also equal to $D_s^{AB} + (v - 1) D_b^{AB}$. Moreover, for two-locus unphased genotypes, Burrow's Δ statistic can be expanded to:

$$\Delta_{AB} = \left(\sum_{i=1}^v \sum_{j=1}^v ij G_{B_j X^{v-i}}^{A_i X^{v-i}} \right) - v p_A p_B, \quad (1)$$

where X is an arbitrary allele distinct from both A and B , with each $G_{B_j X^{v-i}}^{A_i X^{v-i}}$ denoting a two-locus unphased genotypic frequency whose superscript (or subscript) is an unphased genotype containing exactly i copies of A (or j copies of B). In Supplementary Appendix A, we use triploids to illustrate how Δ_{AB} is expanded. Substituting the observed values of p_A , p_B and $G_{B_j X^{v-i}}^{A_i X^{v-i}}$ into Eq. (1), Δ_{AB} can be estimated.

Burrows's Δ is also $1/v$ times the covariance between the allele dosages of A and B within individuals, i.e., $\Delta_{AB} = \text{Cov}(C_A, C_B)/v$, where C_A and C_B are the allele dosages of A and B , respectively (Gao et al. 2008). In other words, $C_A = \sum_{i=1}^v \mathcal{B}_{A_i}$ and $C_B = \sum_{i=1}^v \mathcal{B}_{B_i}$, where i enumerates haplotypes within individuals. Similarly, it is more intuitive to use Pearson's correlation coefficient $r_{\Delta AB}$ to measure LD for unphased data, which is also equal to the correlation coefficient between C_A and C_B :

$$r_{\Delta AB} = \frac{\Delta_{AB}}{\sqrt{R_{AB}}} = \frac{\text{Cov}(C_A, C_B)/v}{\sqrt{\text{Var}(C_A)\text{Var}(C_B)}/v}.$$

where $\text{Cov}(C_A, C_B)$ and $\text{Var}(C_A)$ can be derived by

$$\begin{aligned} \text{Cov}(C_A, C_B) &= E(C_A C_B) - E(C_A)E(C_B) \\ &= \left(\sum_{i=1}^v \sum_{j=1}^v ij G_{B_j X^{v-i}}^{A_i X^{v-i}} \right) - v^2 p_A p_B, \end{aligned}$$

$$\begin{aligned} \text{Var}(C_A) &= E(C_A^2) - E^2(C_A) \\ &= \sum_{i=1}^v \sum_{j=1}^v E(\mathcal{B}_{A_i} \mathcal{B}_{A_j}) - v^2 p_A^2 \\ &= \sum_{i=1}^v E(\mathcal{B}_{A_i}) + \sum_{i \neq j} E(\mathcal{B}_{A_i} \mathcal{B}_{A_j}) - v^2 p_A^2 \\ &= v p_A + v(v-1)[\mathcal{F} p_A + (1 - \mathcal{F}) p_A^2] - v^2 p_A^2. \end{aligned}$$

In the expression of $\text{Var}(C_A)$, \mathcal{F} is the inbreeding coefficient and can be solved from the relation $P_{AA} = \mathcal{F} p_A + (1 - \mathcal{F}) p_A^2$, where P_{AA} is the probability of sampling two copies of A within the same individual without replacement. \mathcal{F} can be obtained by

$$\mathcal{F} = \frac{P_{AA} - p_A^2}{p_A p_X}.$$

Substituting the expression of \mathcal{F} into $r_{\Delta AB}$, a simplified expression of $\sqrt{R_{AB}}$ can be obtained

$$\begin{aligned}\sqrt{R_{AB}} &= \sqrt{\text{Var}(C_A)\text{Var}(C_B)}/v \\ &= \sqrt{[p_A p_X + (v-1)(P_{AA} - p_A^2)][q_B q_X + (v-1)(P_{BB} - q_B^2)]}.\end{aligned}\quad (2)$$

Likewise, $r_{\Delta AB}$ may be negative, but the squared correlation coefficient $r_{\Delta AB}^2$ ranges from 0 to 1, which can also be used to evaluate the LD between two loci for unphased genotypes.

In the following text, for simplicity, we will use $D_w, D_b, D, \Delta, Q, R, r$ and r_{Δ} to replace $D_w^{AB}, D_b^{AB}, D_{AB}, \Delta_{AB}, Q_{AB}, R_{AB}, r_{AB}$ and $r_{\Delta AB}$ in turn. Due to genetic drift, D^2 and Q (or Δ^2 and R) converge to zero after an infinite number of generations. However, the ratio r^2 of D^2 to Q (or the ratio r_{Δ}^2 of Δ^2 to R) converges to a constant, whose value is determined by the mating system together with the recombination frequency c and the effective population size N_e (Weir and Hill 1980). Therefore, the effective population size can be estimated from \hat{r}^2 (or \hat{r}_{Δ}^2) if the relationship between $E(\hat{r}^2)$ (or $E(\hat{r}_{\Delta}^2)$), mating system, c and N_e can be derived.

The values of r^2 and r_{Δ}^2 can be calculated by

$$r^2 = \frac{\hat{D}^2}{\hat{Q}} \quad \text{and} \quad r_{\Delta}^2 = \frac{\hat{\Delta}^2}{\hat{R}},$$

where $\hat{D}, \hat{\Delta}, \hat{Q}$, and \hat{R} can be calculated from the samples. However, these statistics are correlated, such that $E(\hat{r}^2)$ and $E(\hat{r}_{\Delta}^2)$ is hard to derive. If such correlations can be reduced or even eliminated (this can be done by some weighting scheme when multiple loci are used), then $E(\hat{r}^2)$ and $E(\hat{r}_{\Delta}^2)$ can be approximated by the ratio of two expectations, we denoted these ratios by d^2 and δ^2 .

$$E(\hat{r}^2) \approx \frac{E(\hat{D}^2)}{E(\hat{Q})} = d^2 \quad \text{and} \quad E(\hat{r}_{\Delta}^2) \approx \frac{E(\hat{\Delta}^2)}{E(\hat{R})} = \delta^2.\quad (3)$$

In the following sections, we extend Weir and Hill's (1980) double non-identity framework, to obtain the expressions of d^2 and δ^2 .

Double non-identity coefficients

The double non-identity coefficients can be used to derive the moments of various LD measurements. The term *identity* means identical-by-descent (IBD), i.e., two alleles are identical because they are inherited from a common ancestor. Based on Weir and Hill (1980), we establish 22 two-locus allele configurations for polysomic inheritances (Table 1) The observed and expected frequencies of these 22 configurations are denoted by P_i and E_i , respectively; and E_i is derived by the non-identity coefficients assuming no initial LD (Table 1). The descriptions of the non-identity coefficients, and the derivations of E_i are provided in Supplementary Appendix B. The moments of LD measurements can be expressed by E_i (Supplementary Appendix C), and can be further expanded as linear combinations of the double non-identity coefficients (Table 2).

The expressions of various moments can now be expressed uniformly by matrices. Let \mathbf{M} be the row vector consisting of the 7 moments (header row of Table 2), and let Φ be the column vector consisting of the 13 double non-identity coefficients (header column of Table 2). Denote \mathbf{A} as a 13×7 matrix, whose i^{th} column consists of the i^{th} column divided by the last column of Table 2. Then

$$\mathbf{M} = \Phi^T \mathbf{A}.\quad (4)$$

We call \mathbf{M} the moment vector, and Φ the double non-identity vector.

Transition matrix of double non-identity coefficients

The transition matrix of double non-identity coefficients can be used to describe the behavior of double non-identity coefficients due to genetic drift.

Let Φ be the double non-identity column vector in the current generation, and let Φ' be that in the next generation and Φ' can be expressed as $\Phi' = \Omega \Phi$. We call Ω the *transition matrix* from Φ to Φ' .

Let Φ_0 be the double non-identity vector in the founder generation and let Φ_t be that in the t^{th} generation. This gives $\Phi_t = \Omega^t \Phi_0$. If a population is allowed to reproduce for several generations, the vector sequence is: $\Phi_0, \Phi_1, \Phi_2, \dots, \Phi_t, \dots$ and will reach a steady state as t increases. In other words, this sequence will converge to a constant vector, denoted by Φ_{∞} . This limit vector Φ_{∞} is independent to the initial vector Φ_0 if $\Phi_0 \neq \mathbf{0}$.

To simplify the model for polysomic inheritance, we established a virtual mating system, named the *haplotype sampling* (HS) mating system. In this mating system, it is assumed that each individual is reproduced by randomly sampling v haplotypes with replacement from the previous generation. The genes in an offspring therefore come from a maximum of v parents. Because the haplotypes within (or among) individuals are randomly sampled, there is no difference among dihaplotypic, trihaplotypic and quadhaplotypic double non-identity coefficients, symbolically $\Theta_1 = \Theta_2, \Gamma_1 = \Gamma_2 = \Gamma_3 = \Gamma_4$ and $\Delta_1 = \Delta_2 = \dots = \Delta_7$. Therefore, the transition matrix Ω in the HS mating system can be simplified as a 3×3 matrix, which is derived in Supplementary Appendix D. The full and simplified Ω are listed in Supplementary Table S3 and Table 3, respectively.

It is noteworthy that the sum of the combination coefficients of 1 in each column in Table 3 is exactly one, but the sum of each row of Ω is less than one. This indicates that the transition (i.e., a generation of random mating) will gradually reduce the double non-identity coefficients, and their values will eventually converge to zero, i.e., $\Omega^{\infty} = \mathbf{0}$. This also holds for the other mating systems and demonstrates the loss of heterozygosity and the fixation of alleles.

Although Φ_{∞} will eventually converge to zero, the ratio of the moments $E(\hat{D}^2)$ to $E(\hat{Q})$, and of the moments $E(\hat{\Delta}^2)$ to $E(\hat{R})$ will converge to some constants. This can be considered as the double non-identity vector Φ reaches a relatively stable state so the direction of Φ is constant during reproduction, symbolically $\Phi' = \nu \Phi$. The direction of Φ (say ω) and the scale factor ν can be solved by performing eigen-value decomposition for Ω , i.e., solving $\Omega \omega = \nu \omega$. It is also noteworthy that there are multiple eigenvalues, with the highest eigenvalue be of our interest. Therefore, d^2 and δ^2 can be calculated from Eq. (4) by substituting Ω with ω , i.e., $\mathbf{M}_{\omega} = \omega^T \mathbf{A}$. We denote the elements in \mathbf{M}_{ω} as $E_{\omega}(\cdot)$, e.g., $E_{\omega}(\hat{D}^2)$, then the exact d^2 and δ^2 are as follows:

$$d^2 = \frac{E_{\omega}(\hat{D}^2)}{E_{\omega}(\hat{Q})} \quad \text{and} \quad \delta^2 = \frac{E_{\omega}(\hat{\Delta}^2)}{E_{\omega}(\hat{R})}.\quad (5)$$

Approximations

Weir and Hill (1980) adopted a matrix decomposition technique to approximate ν and ω for disomic inheritance and also to approximate d^2 and δ^2 . We follow this approach to derive the approximate expressions of d^2 and δ^2 for the HS mating system and four additional mating systems.

Let Ω be the simplified transition matrix for the HS mating system, as detailed in Table 3. If N is large enough, the values of the terms with N^{-2} and N^{-3} in Table 3 will be small, then Ω can be decomposed to:

$$\Omega = \mathbf{T} + N^{-1} \mathbf{S} + \mathcal{O}(N^{-2}).$$

Table 1. Allele configurations and their expected frequencies.

Allele configuration	Freq. notation	Double non-identity	Expectation			Expectation notation
			$P_A q_B$	$P_A q_B (p_X + q_X)$	$P_A p_X q_B q_X$	
Digenic						
$Z_{B...}^A...$	P_1		1			E_1
$Z_{B...}^A...$	P_2		1			E_2
$Z_{B...}^A...$	P_3		1			E_3
Trigenic						
$Z_{B...}^{AA...}, Z_{BB...}^A...$	P_4		2	$-P$		E_4
$Z_{B...}^{AA...}, Z_{BB...}^A...$	P_5		2	$-P$		E_5
$Z_{B...}^A..., Z_{B...}^A...$	P_6		2	$-\Pi$		E_6
$Z_{B...}^A..., Z_{B...}^A...$	P_7		2	$-\Pi$		E_7
$Z_{B...}^A..., Z_{B...}^A...$	P_8		2	$-\Pi$		E_8
$Z_{B...}^{AA...}, Z_{BB...}^A...$	P_9		2	$-P$		E_9
Quadgenic						
Dihaplotypic						
$Z_{BB...}^{AA...}$	P_{10}	Θ_1	1	$-P$	Θ_1	E_{10}
$Z_{B...}^A...$	P_{11}	Θ_2	1	$-\Pi$	Θ_2	E_{11}
Trihaplotypic						
$Z_{B...}^A...$	P_{12}	Γ_1	1	$-\Pi$	Γ_1	E_{12}
$Z_{B...}^{AA...}, Z_{BB...}^A...$	P_{13}	Γ_2	2	$-P - \Pi$	$2\Gamma_2$	E_{13}
$Z_{B...}^A...$	P_{14}	Γ_3	1	$-\Pi$	Γ_3	E_{14}
$Z_{B...}^{AA...}$	P_{15}	Γ_4	1	$-P$	Γ_4	E_{15}
Quadhaplotypic						
$Z_{B...}^A...$	P_{16}	Δ_1	1	$-\Pi$	Δ_1	E_{16}
$Z_{B...}^{AA...}$	P_{17}	Δ_2	1	$-P$	Δ_2	E_{17}
$Z_{B...}^A...$	P_{18}	Δ_3	1	$-\Pi$	Δ_3	E_{18}
$Z_{B...}^{AA...}, Z_{BB...}^A...$	P_{19}	Δ_4	2	$-P - \Pi$	$2\Delta_4$	E_{19}
$Z_{B...}^A...$	P_{20}	Δ_5	1	$-\Pi$	Δ_5	E_{20}
$Z_{B...}^{AA...}$	P_{21}	Δ_6	1	$-P$	Δ_6	E_{21}
$Z_{B...}^{AA...}, Z_{BB...}^A...$	P_{22}	Δ_7	2	$-P - \Pi$	$2\Delta_7$	E_{22}

Z denotes an allele configuration, P and Π denote the single non-identity coefficient within and between individuals. Superscripts and subscripts denote the genotype patterns at two target loci, respectively. The vertical bars separate individuals, · denotes an allele copy in any form, and... denotes the remaining allele copies. The allele copies in the same position are in the same haplotype. For example, $Z_{B...}^A...$, $Z_{BB...}^A...$ and $Z_{B...}^A...$ denote the two target alleles are in the same haplotype, different haplotypes in the same individual, and different individuals, respectively.

For the matrices **T** and **S** in the principal part of **Ω**, with **Ω** given in Table 3 we obtain

$$\mathbf{T} = \begin{bmatrix} c_1^2 & -2c_1c & c^2 \\ 0 & -c_1 & c \\ 0 & 0 & 1 \end{bmatrix} \text{ and }$$

$$\mathbf{S} = \begin{bmatrix} \frac{c^2}{v_1} - \frac{1+2c_1c}{v} & \frac{4c(2c-1)}{v} - \frac{2c^2}{v_1} & \frac{2c^2(3-2v)}{v_1v} \\ -\frac{c_1}{v} & \frac{6c-3}{v} & -\frac{5c}{v} \\ 0 & \frac{4}{v} & -\frac{6}{v} \end{bmatrix},$$

where $c_i = c - i$ and $v_i = v - i$. Similarly, ν and ω can be decomposed to

$$\nu = 1 + N^{-1}r + \mathcal{O}(N^{-2}),$$

$$\omega = 1 + N^{-1}\mathbf{x} + \mathcal{O}(N^{-2}),$$

where $\mathbf{1} = [1, 1, 1]^T$ and $\mathbf{x} = [x_1, x_2, x_3]^T$. According to $\mathbf{\Omega}\omega = \nu\omega$, we obtain a matrix equation as follows:

$$\mathbf{T}\mathbf{1} + N^{-1}\mathbf{T}\mathbf{x} + N^{-1}\mathbf{S}\mathbf{1} = \mathbf{1} + N^{-1}\mathbf{x} + N^{-1}r\mathbf{1} + \mathcal{O}(N^{-2}).$$

Because $\mathbf{T}\mathbf{1} = \mathbf{1}$, if the term $\mathcal{O}(N^{-2})$ is omitted, we obtain $(\mathbf{S} - r\mathbf{1})\mathbf{1} = (\mathbf{I} - \mathbf{T})\mathbf{x}$.

This matrix equation is a linear equation set with 3 equations and 4 unknowns, the solutions of which are as follows:

$$r = -2/v, x_1 = \frac{c^2v + (1 - 2c)v_1}{(2 - c)cv_1v} + \zeta, x_2 = \zeta, x_3 = \zeta \ (\zeta \text{ is any number}).$$

If we let $\zeta = 0$, we obtain a special solution: $r = -2/v$ and $\mathbf{x} = \left[\frac{c^2v + (1 - 2c)v_1}{(2 - c)cv_1v}, 0, 0 \right]^T$. Replacing this solution into the expressions of

Table 2. Essential factors of moment expressions.

	$E(\hat{D}_w^2)$	$E(\hat{D}_b^2)$	$E(\hat{D}_w\hat{D}_b)$	$E(\hat{D}^2)$	$E(\hat{\Delta}^2)$	$E(\hat{Q})$	$E(\hat{R})$	Divisor
$\Phi_1\Theta_1$	$n^2v^2 \times (2 + vv_2)$	λ_7	$-nv \times (vv_1 + \lambda_1)$	$v_1^2\lambda_5$	$2v^2n_1^2v_1^2$	$2v_1^2$	$2v^2n_1^2v_1^2$	$n^3v^3v_1/Q$
$\Phi_2\Theta_2$	n^2v^2	2	$-nv$	λ_5	$v^2\lambda_2$	2	$2v^2$	n^3v^2/n_1Q
Γ_1	$-2n^2v^2$	$-2\lambda_1$	$nv(1 + \lambda_4)$	$-2v_1\lambda_3$	$2v^2v_1\lambda_2$	$4v_1$	$4v^2v_1$	n^3v^2/n_1Q
Γ_2	0	$-2\lambda_1$	$-nvv_2$	$-2v_1\lambda_3$	$-4v^2n_1v_1$	$4v_1$	$-4v^2n_1v_1$	$n^3v^2/2n_1Q$
Γ_3	0	4	$-nv$	$-2\lambda_3$	$-2v^2n_2$	4	$4v^2$	n^3v/n_1n_2Q
Γ_4	$-2n^2v^2v_2$	$2\lambda_1\lambda_4$	λ_6	$-2v_1^2\lambda_3$	$4v^2n_1^2v_1^2$	$4v_1^2$	$4v^2n_1^2v_1^2$	$n^3v^3v_1/v_2Q$
Δ_1	n^2v^2	λ_7	$-nv\lambda_4$	$2v_1^2$	$v^2v_1^2\lambda_2$	$2v_1^2$	$2v^2v_1^2$	n^3v^2/n_1Q
Δ_2	0	1	0	1	v^2	1	$v^2n_1^2$	$n^3v^2/n_1v_1^2Q$
Δ_3	0	$-2\lambda_1$	nv	$4v_1$	$-2v^2n_2v_1$	$4v_1$	$4v^2v_1$	n^3v/n_1n_2Q
Δ_4	0	v_1	0	v_1	v^2v_1	v_1	$-v^2n_1v_1$	$n^3v/2n_1n_2Q$
Δ_5	0	1	0	1	v^2	1	v^2	$n^3/n_1n_2n_3Q$
Δ_6	n^2v^2	λ_4^2	$-nv\lambda_4$	v_1^2	$v^2n_1^2v_1^2$	v_1^2	$v^2n_1^2v_1^2$	$n^3v^3v_1/v_2v_3Q$
Δ_7	0	$-2\lambda_4$	nv	$2v_1$	$-2v^2n_1v_1$	$2v_1$	$-2v^2n_1v_1$	$n^3v^2/2n_1v_2Q$

For brevity, we denote $n - i$ by n_i and $v - i$ by v_i , and let $\lambda_1 = n_2v + 2$, $\lambda_2 = nn_2 + 2$, $\lambda_3 = nv - 2$, $\lambda_4 = vn_1 + 1$, $\lambda_5 = 2 + nv\lambda_3$, $\lambda_6 = nv[2 + (\lambda_3 - \lambda_4)(1 + \lambda_4)]$, $\lambda_7 = 2 + 2n_2v + \lambda_2v^2$.

Table 3. Simplified Ω^T for HS mating system.

1	N^{-1}	1	N^{-1}	N^{-2}	1	N^{-1}	N^{-2}	N^{-3}
$(1-c)^2$	$\frac{c^2}{v-1} - \frac{1+2c(c-1)}{v}$	0	$\frac{1-c}{v}$	$\frac{2c-1}{v^2}$	0	0	$\frac{2}{v^2}$	$-\frac{2}{v^3}$
$2c(1-c)$	$\frac{4c(2c-1)}{v} - \frac{2c^2}{v-1}$	$1-c$	$\frac{6c-3}{v}$	$\frac{2-8c}{v^2}$	0	$\frac{4}{v}$	$-\frac{12}{v^2}$	$\frac{8}{v^3}$
c^2	$\frac{2c^2(3-2v)}{v(v-1)}$	c	$-\frac{5c}{v}$	$\frac{6c}{v^2}$	1	$-\frac{6}{v}$	$\frac{11}{v^2}$	$-\frac{6}{v^3}$

Each element of Ω^T is a combination of $1, N^{-1}, N^{-2}$ and N^{-3} with the combination coefficients in the corresponding cell. The combination coefficients are zero for unrepresented terms.

ν and ω yields

$$\nu \approx \frac{Nv - 2}{Nv} \text{ and } \omega \approx \left[1 + \frac{c^2v + (1 - 2c)v_1}{(2 - c)cN_e v_1 v}, 1, 1 \right]^T.$$

Now, by substituting Φ with ω and A with $A_1 = \lim_{n \rightarrow \infty} A$ in Eq. (4), it can be calculated that

$$E_\omega(\hat{D}^2) = E_\omega(\hat{\Delta}^2) \approx \frac{c^2v + (1 - 2c)v_1}{(2 - c)cN_e v_1 v} \text{ and } E_\omega(\hat{Q}) = E_\omega(\hat{R}) \approx 1.$$

Therefore, the approximated d^2 and δ^2 are as follows:

$$d_{HS}^2 \approx \frac{c^2v + (1 - 2c)v_1}{(2 - c)cN_e v_1 v} \text{ and } \delta_{HS}^2 \approx \frac{c^2v + (1 - 2c)v_1}{(2 - c)cN_e v_1 v}.$$

To include the effect of finite sample size, higher order terms in A should be included. We derive the approximations of d_{HS}^2 and δ_{HS}^2 by ignoring higher order terms of A , and find that d_{HS}^2 and δ_{HS}^2 converge to

$$d_{HS}^2 \approx \frac{c^2v + (1 - 2c)v_1}{(2 - c)cN_e v_1 v} + \frac{1}{vn - 1}, \tag{6a}$$

$$\delta_{HS}^2 \approx \frac{c^2v + (1 - 2c)v_1}{(2 - c)cN_e v_1 v} + \frac{1}{n - 1}, \tag{6b}$$

where N_e and N are equivalent under the HS mating system, n is the sample size. The additional terms $1/(vn - 1)$ and $1/(n - 1)$ are corrections for finite sample size (see Supplementary Appendix E for details). The results from Eqs. (6a) and (6b) accord with those of

Ohta and Kimura (1969) and Weir and Hill (1980) for the monoecious selfing mating system in diploids.

The transition of single non-identity coefficients satisfies the relations: $P' = \frac{Nv-1}{Nv}P$ and $\pi' = \frac{Nv-1}{Nv}\pi$. Moreover, if two loci are located at the two extremities on the same chromosome under bivalent pairing, and the thirteen double non-identity coefficients are all equal to P^2 and $\Phi' = (\frac{Nv-1}{Nv})^2\Phi$, and thus also the corresponding eigenvalue $\nu = (\frac{Nv-1}{Nv})^2 \approx \frac{Nv-2}{Nv}$. By comparing with the previous conclusion of $\nu \approx \frac{Nv-2}{Nv}$ by substituting $\zeta = 0$, we see that $r = -2/v$ is a good approximation to the rate of loss of heterozygosity at the pairs of independent loci.

We follow Weir and Hill (1980) to establish four additional mating systems. Two are monocious mating systems: (i) selfing being allowed (termed MS), and (ii) selfing being excluded (termed ME). In both of these mating systems, the effective population size N_e is the same as the population size N . The other two mating systems we use are both dioecious systems: (i) dioecious with random pairing (termed DR), and dioecious with lifetime pairing (termed DH). In DR, each offspring is produced from a new pairing. In DH, each individual remains in a single reproductive unit for its entire lifetime. Moreover, in both DR and the DH, there are M males and F females in the population for each generation and $F = fM$, the effective population size is calculated by $N_e = \frac{4MF}{M+F}$.

The transition matrix Ω for each of the four additional mating systems (MS, ME, DR and DH) is a 13×13 matrix, whose element expressions are derived in Supplementary Appendices F–H. The matrices T and S in the principal part of Ω for all five mating systems are listed in Supplementary Appendix I. The approximate expressions of d^2 and δ^2 for additional

mating systems can be derived with the same method (details can be found in Supplementary Appendix J) and are shown as follows:

$$d_{MS/ME/DR}^2 \approx \frac{8c_2c^2 - 4c_2cv(5c - 1) + 2v^2(7c_2c^2 + c + 2) - 3c_1^2v^3(c + 1)}{c_2c(cv_2 + v)(3v - 4)v^2N_e} + \frac{1}{vn - 1},$$

$$\delta_{MS/ME/DR}^2 \approx \frac{v^2[4 - 3v + 8c^2 - 14c - cv(2c^2 + 4c - 13) + c_2cv^2(c + 1)]}{c_2c(cv_2 + v)(3v - 4)v^2(N_e - \eta)} + \frac{1}{n - 1};$$

$$d_{DH}^2 \approx \left\{ (1 + f)[cv(3v^2 + 2v - 8) - v^2(3v - 4)] + c^2(3v - 4)[v^2 - 10v + 4 + f(v^2 - 8v + 4)] - c^3v_2[3v^2 - 10v + 4 + f(3v^2 - 8v + 4)] \right\} / [c_2c(1 + f)(cv_2 + v)(3v - 4)v^2N_e] + \frac{1}{vn - 1},$$

$$\delta_{DH}^2 \approx v^2 \left\{ c^3(3 + f)v_2v - (1 + f)(3v - 4) - c^2[3v^2 - 8 + f(v^2 + 4v - 8)] - c[f(2v^2 - 13v + 14) + 3(2v^2 - 7v + 6)] \right\} / [c_2c(1 + f)(cv_2 + v)(3v - 4)v^2(N_e - \eta)] + \frac{1}{n - 1}.$$

The approximate expressions of d^2 and δ^2 from disomic to decasomic are presented in Supplementary Tables S5 and S6. They follow a general pattern:

$$d^2 = \frac{C}{N_e} + \frac{1}{vn - 1} \quad \text{and} \quad \delta^2 = \frac{C}{N_e - \eta} + \frac{1}{n - 1}. \tag{7}$$

where η is equal to 0 for the HS mating system, $\frac{2(v-2)(v-1)}{v^2}$ for the MS mating system, or $\frac{4(v-1)^2}{v^2}$ for the ME/DR/DH mating systems. The values of C for approximated d^2 and δ^2 between unlinked loci located on either the same chromosome ($c = 0.5$) or different chromosomes ($c = 1 - 1/v$) are presented in Table 4.

SIMULATIONS AND EVALUATIONS

Behaviors of \hat{r}^2 and \hat{r}_Δ^2

In this section, we discuss the behaviors of the squared correlation coefficient estimators \hat{r}^2 and \hat{r}_Δ^2 during reproduction and provide the exact and the approximate values of d^2 or δ^2 for reference.

Due to the correlation between \hat{D}^2 and \hat{Q} (or between $\hat{\Delta}^2$ and \hat{R}), $E(\hat{r}^2)$ (or $E(\hat{r}_\Delta^2)$) is not equal to d^2 (or δ^2), which introduces some

biases when few loci are used. To solve this problem, Waples (2006) used an empirical equation to adjust \hat{r}_Δ^2 for di-allelic loci, which can be extended to multi-allelic loci by collapsing alleles. We use an alternative method to eliminate such correlations and bias. Assuming all locus pairs share the same parameters (c, n, N_e, v and mating system), then their d^2 (or δ^2) are respectively the same, and their \hat{r}^2 (or \hat{r}_Δ^2) can be weighted to approximate d^2 (or δ^2). The multi-locus estimates of \hat{r}^2 and \hat{r}_Δ^2 are calculated by

$$\hat{r} = \frac{\sum_{(l_1, l_2)} \sum_{A \in l_1, B \in l_2} \hat{D}_{AB}^2}{\sum_{(l_1, l_2)} \sum_{A \in l_1, B \in l_2} \hat{Q}_{AB}} \quad \text{and} \quad \hat{r}_\Delta^2 = \frac{\sum_{(l_1, l_2)} \sum_{A \in l_1, B \in l_2} \hat{\Delta}_{AB}^2}{\sum_{(l_1, l_2)} \sum_{A \in l_1, B \in l_2} \hat{R}_{AB}}, \tag{8}$$

where (l_1, l_2) is taken from all locus pairs, the symbol $A \in l_1$ (or $B \in l_2$) represents A (or B) is taken from all alleles at the first (or the second) locus in (l_1, l_2) .

We adopt a Monte-Carlo method to simulate the behavior of \hat{r}^2 and \hat{r}_Δ^2 . During simulation, a population with the MS mating system is generated, which contains 40 or 80 individuals with a ploidy level of either 2 or 4. Next, the individuals generated are genotyped at 200 linked diallelic loci pairs, with a recombination frequency 0.1 for each locus pair. Although we generate 400 loci, only 200 loci pairs with $c = 0.1$ are used in calculating \hat{r}^2 and \hat{r}_Δ^2 . The population is then allowed to reproduce for 250 generations. For each generation, by using the data of genotypes of all individuals under various situations, \hat{r}^2 and \hat{r}_Δ^2 are calculated by Eq. (8), and the exact and the approximate d^2 and δ^2 are also calculated by Eqs. (5) and (6a, 6b), respectively. This process is performed 300,000 times in total. The results are shown in Fig. 1.

Figure 1 shows that the approximate d^2 or δ^2 are both slightly higher than their exact value, and both the exact and the approximate d^2 or δ^2 decrease as N_e or v increases. The values of \hat{r}^2 and \hat{r}_Δ^2 are both initially 1, and reduce respectively to exact d^2 and δ^2 values after about 40 generations. Henceforth, \hat{r}^2 and \hat{r}_Δ^2 both achieve a relatively stable state and remain around the exact values of d^2 and δ^2 for several generations. In particular, if the ploidy level is four, these values will both converge to the exact d^2 and δ^2 values as the number of generations increases.

Due to genetic drift, some loci become fixed and are excluded from the simulation, causing the number L of locus pairs used for genotyping to decline. The correlation between the numerator and the denominator in each of both formulas in Eq. (8) therefore increases, such that \hat{r}^2 and \hat{r}_Δ^2 correspondingly decrease. The duration of a stable state depends on three factors: (i) ploidy level v , (ii) effective population size N_e and (iii) the number L of locus pairs. As the value of each of these factors increases, the longer the duration of the stable state of both \hat{r}^2 and \hat{r}_Δ^2 .

Table 4. Coefficient C for approximated d^2 and δ^2 .

	v	HS	MS/ ME/DR		DH	
		d^2 & δ^2	d^2	δ^2	d^2	δ^2
Same chromosome $c = 0.5$	2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{3+2f}{6(1+f)}$	$\frac{3+f}{3(1+f)}$
	4	$\frac{1}{9}$	$\frac{61}{480}$	$\frac{1}{3}$	$\frac{75+61f}{480(1+f)}$	$\frac{3+f}{3(1+f)}$
	6	$\frac{1}{15}$	$\frac{19}{252}$	$\frac{1}{3}$	$\frac{22+19f}{252(1+f)}$	$\frac{3+f}{3(1+f)}$
	8	$\frac{1}{21}$	$\frac{559}{10560}$	$\frac{1}{3}$	$\frac{627+559f}{10560(1+f)}$	$\frac{3+f}{3(1+f)}$
	10	$\frac{1}{27}$	$\frac{277}{6825}$	$\frac{1}{3}$	$\frac{609+554f}{13650(1+f)}$	$\frac{3+f}{3(1+f)}$
Different chromosomes ($c = 1 - 1/v$)	2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{3+2f}{6(1+f)}$	$\frac{3+f}{3(1+f)}$
	4	$\frac{1}{15}$	$\frac{439}{5280}$	$\frac{109}{330}$	$\frac{673+439f}{5280(1+f)}$	$\frac{343+109f}{330(1+f)}$
	6	$\frac{1}{35}$	$\frac{1147}{30870}$	$\frac{579}{1715}$	$\frac{1747+1147f}{30870(1+f)}$	$\frac{1779+579f}{1715(1+f)}$
	8	$\frac{1}{63}$	$\frac{22357}{1068480}$	$\frac{5662}{16695}$	$\frac{33921+22357f}{1068480(1+f)}$	$\frac{17226+5662f}{16695(1+f)}$
	10	$\frac{1}{99}$	$\frac{18529}{1383525}$	$\frac{18775}{55341}$	$\frac{56093+37058f}{2767050(1+f)}$	$\frac{56845+18775f}{55341(1+f)}$

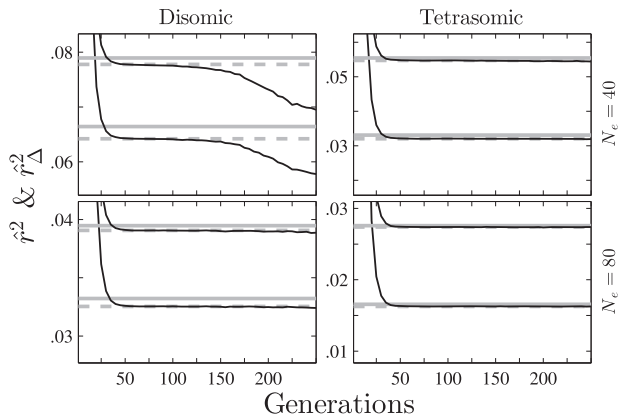


Fig. 1 The behaviors of \hat{r}^2 and \hat{r}^2_{Δ} during reproduction for the MS mating system (set $N_e = 40$ or 80 , $v = 2$ or 4 , $L = 200$ and $c = 0.1$). Each of the two columns shows the results of a different ploidy level, and each of the two rows shows the results of a different effective population size. Solid gray lines denote approximate d^2 or δ^2 , dotted gray lines denote exact d^2 or δ^2 , and solid lines denote \hat{r}^2 or \hat{r}^2_{Δ} , where the lines representing δ^2 (or \hat{r}^2_{Δ}) are above those representing d^2 (or \hat{r}^2) for each situation.

We also simulate the behaviors of \hat{r}^2 and \hat{r}^2_{Δ} during reproduction for five mating systems (including cases with f being set to either 2 or 5 for the DR and the DH mating systems). The simulation process is as follows. First, a population for each of the five mating systems is generated, which contains 40 individuals with a ploidy level of either 2, 4, 6 or 8. Next, these 40 individuals are genotyped as described for the previous simulation. Then, the population is allowed to reproduce for 50 generations. For each generation, by using data of the genotypes of all individuals under various situations, \hat{r}^2 and \hat{r}^2_{Δ} are calculated. The exact and approximate d^2 and δ^2 values are also calculated. The process is repeated 30,000 times. The results are shown in Supplementary Fig. S1, and are similar to those shown in Fig. 1. However, the approximate values of d^2 and δ^2 deviate more from their exact values for some mating systems.

Finally, we also simulate the behaviors of \hat{r}^2 and \hat{r}^2_{Δ} for the MS mating system under different recombination frequencies (set $N_e = 80$, $v = 2$ or 4 , $L = 200$ and $c = 0.001, 0.002, 0.004, 0.01, 0.02, 0.04, 1$ or 2). The simulation process is similar to the previous method and is performed 20,000 times. The population is allowed to reproduce for 100 generations. For each generation, \hat{r}^2 and \hat{r}^2_{Δ} are calculated, with the results shown in Supplementary Fig. S2. This shows that the convergent rates for \hat{r}^2 or \hat{r}^2_{Δ} among different ploidy levels differ little as the number of generations increase, but are strongly affected by the recombination frequency: the higher the recombination frequency, the faster the rate of convergence.

Recombination frequency

To investigate the influence of the recombination frequency c on d^2 and δ^2 , the exact and the approximate d^2 and δ^2 are calculated for each mating system under different recombination frequencies (set $N_e = 100$, $n = 100$, $v = 2, 4, 6$ or 8 , $f = 1$ for DR and $f = 2$ or 5 for DH). The recombination frequency c ranges from 0 to 1. The results for the MS mating system are shown in Fig. 2, and the results for all mating systems (including MS) are uniformly shown in Supplementary Fig. S3.

Figure 2 shows that d^2 or δ^2 are high at a low recombination frequency and decrease gradually to a relatively low value as c increases. The rate of decrease steepens as the ploidy level increases. However, after c reaches ~ 0.5 , d^2 (at $v = 2$) or δ^2 (at all ploidy levels) both begin to increase. The approximate values of d^2 are close to their exact values, whilst the difference between the approximate and the exact values of δ^2 are more obvious, especially when $c > 0.5$.

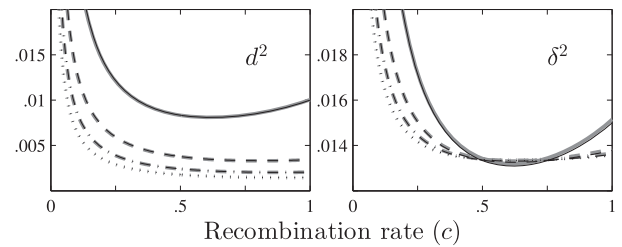


Fig. 2 The relationship between d^2 (or δ^2) and the recombination frequency c for the MS mating system (set $N_e = 100$, $n = 100$ and $v = 2, 4, 6$ or 8). The solid, dashed, dash-dotted and dotted lines denote the values for diploids, tetraploids, hexaploids and octoploids in turn, and the gray and the black lines denote the exact and the approximate values, respectively.

The exact values for d^2 and δ^2 for the unlinked loci located on the same or different chromosomes are calculated for all five mating systems (set $N_e = 100$, $n = 100$, $v = 2, 4, 6, 8$ or 10 , $c = 0.5$ or $1 - 1/v$ and $f = 1, 2$ or 5 for DR/DH). Moreover, the error rates for d^2 or δ^2 under different conditions are also calculated. The results are presented in Supplementary Table S7. It is clear that the difference between $\delta^2_{c=0.5}$ and $\delta^2_{c=1-1/v}$ is low under all conditions, but the difference between $d^2_{c=0.5}$ and $d^2_{c=1-1/v}$ is ~ 50 to 100 times higher. For example, for tetraploids, the error rate is about 13% for d^2 but only 0.13% for δ^2 .

Estimation of effective population size

In this section, we estimate the effective population size N_e from unphased genotypes. We derived the relationships among v , c , n , N_e and δ^2 in the *Theory and modeling* section, e.g., Eq. (6b), where v and n are known, δ^2 can be substituted by \hat{r}^2_{Δ} , \hat{N}_e can be solved if c is known.

Close-linked loci take a long time to reach a mutation-drift equilibrium (Supplementary Fig. S2) and provide past information regarding N_e . Some estimators use this feature to estimate the time series of N_e , but need a priori information about recombination frequency (e.g., Tenesa et al. 2007; Santiago et al. 2020; Hollenbeck et al. 2016). For contemporary N_e , some estimators (e.g., England et al. 2006) assume that all loci are unlinked, and they use a recombination frequency 0.5 for all loci pairs. In polysomic inheritances, the recombination frequency is $1 - 1/v$ between two loci located on different chromosomes. Because $\delta^2_{c=0.5}$ and $\delta^2_{c=1-1/v}$ are close, with the error rate at most 1.5% (Supplementary Table S7), we assume the recombination frequency $c = 0.5$ between any two loci.

We preliminarily solve N_e using the approximated δ^2 by Eq. (7):

$$\hat{N}_{e,\text{initial}} = \frac{c}{\hat{r}^2_{\Delta} - 1/(n-1)} + \eta, \quad (9)$$

where \hat{r}^2_{Δ} is calculated by Eq. (8).

We further optimize the solution using the exact δ^2 , i.e., Eq. (5). The exact δ^2 is related to the double non-identity coefficients and the effective population size N_e . Therefore, the exact δ^2 can be regarded as a function of N_e , denoted by $\delta^2(N_e)$ such that \hat{N}_e is the root of the following equation:

$$\delta^2(\hat{N}_e) - \hat{r}^2_{\Delta} = 0,$$

and we solve \hat{N}_e with Newton's method using $\hat{N}_{e,\text{initial}}$ as the initial solution. This approach is denoted as NEWTON'S approach. According to Eq. (8) and the central limit theorem, \hat{r}^2_{Δ} can be approximated with a normal distribution when there are many loci. Substituting δ^2 with \hat{r}^2_{Δ} and N_e with \hat{N}_e in Eq. (7) and assuming $\hat{r}^2_{\Delta} \sim \mathcal{N}(\mu, \sigma^2)$, it can be found that $[\hat{r}^2_{\Delta} - 1/(n-1)]/C$ is accord with $\mathcal{N}(\mu - 1/(n-1), \sigma^2/C^2)$ and is equal to $1/(\hat{N}_e - \eta)$. Therefore, $\hat{N}_e - \eta$ is in accordance with an inverse normal distribution whose expectation is undefined (Robert 1991). It is thus meaningless to evaluate the statistical performance of \hat{N}_e because its expected value is not defined. To avoid this problem, we instead evaluate

the statistical performance of $1/\hat{N}_e$, which is approximately unbiased according to Eq. (9).

We use a Monte-Carlo method to simulate the estimation of effective population size N_e from unphased genotypes, and then evaluate the statistical performance of NEWTON'S approach under different ploidy levels, numbers of loci, numbers of alleles and sample sizes. Two types of SSR markers are used during simulation: (i) SNP (diallelic) and (ii) SSR (hexa-allelic). For simulation, first a founder population with 200 individuals all with a ploidy level of either 2, 4, 6 or 8 is created. To avoid the fixation of alleles, each allele in the founder generation is set as being unique. Second, the 200 individuals are genotyped at 100 or 200 diallelic SNPs, or at 20 or 40 hexa-allelic SSRs. These loci are assumed to be isometrically distributed on 10 chromosomes, and the length of each chromosome is 100 cM. Third, the founder population is allowed to reproduce for a fixed number of generations to reach the linkage equilibrium; the number of generations is 44 or 86 for SNP, and 11 or 19 for SSR; during meiosis, it is assumed that the chromosomes form bivalents. Fourth, after the final generation has been attained, to reduce the number of alleles k , we repeat collapsing two randomly selected alleles until the value of k is less than 2 (for SNP) or 6 (for SSR). Fifth, for the final generation, 400 individuals are created in total, and n individuals are randomly sampled from this generation, where $n = 40, 80, \dots, 400$ (interval 40). Finally, using the data of unphased genotypes of the n individuals sampled ($n = 40, 80, \dots, 400$), \hat{N}_e can be estimated by using NEWTON'S approach. We use the MS mating system as an example and performed 2000 replicates for each configuration. If we subsequently let $\hat{V} = 1/\hat{N}_e$, the bias and the RMSE of \hat{V} can be calculated, the results being shown in Fig. 3 and Supplementary Fig. S4. The simulated bias and RMSE of \hat{N}_e are shown in Supplementary Fig. S4.

Figure 3 shows that the results for SNP are more biased than those for SSR, with \hat{V} slightly increasing as the number of loci L also increases. The bias of \hat{V} is small, and is generally less than 2×10^{-3} , especially less than 3×10^{-4} for the hexasomic and the octosomic inheritances, thus \hat{V} is nearly unbiased, as expected.

Supplementary Fig. S5 shows that the RMSEs of \hat{V} decrease as n increases, the values of which are similar among different ploidy levels. Moreover, the RMSEs for polyploids are slightly smaller than that for diploids. In general, the performances of SNPs and SSRs are similar.

DISCUSSION

LD test

We here follow the method proposed by Weir and Cockerham (1979) to extend two LD measures, D and the Burrow's Δ , to account for different levels of polysomic inheritance. These two

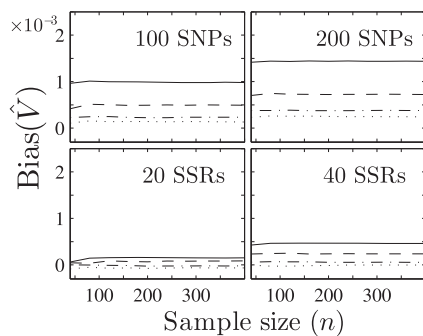


Fig. 3 The relationship between the bias of \hat{V} and the sample size n (set $N_e = 200$, $v = 2, 4, 6$ or 8 , $L = 100$ or 200 for SNP and $L = 20$ or 40 for SSR). The results are obtained from the unphased genotypes of 40–400 individuals (interval 40). The solid, dashed, dash-dotted and dotted lines denote results for disomic, tetrasomic, hexasomic and octosomic inheritances in turn.

measures can be used to perform the LD test. The null hypothesis of a LD test is that a pair of loci is under linkage equilibrium, which is equivalent to all D_{AB} (or all Δ_{AB}) values being equal to zero.

For a sample with n individuals, there are nv haplotypes. The observed and the expected occurrences of a haplotype AB are, respectively, nvP_s^{AB} and nvP_{AQ_B} . Because $D_{AB} = P_s^{AB}P_s^{AB} - P_{AQ_B}$, the χ^2 statistic for the LD measure D can be established as follows:

$$\chi_D^2 = nv \sum_{AB} \frac{\hat{D}_{AB}^2}{P_{AQ_B}} \text{ with d.f. } (k_1 - 1) \times (k_2 - 1),$$

where d.f. is the number of degrees of freedom, k_i is the number of alleles among the allele copies in those haplotypes at the i^{th} locus ($i = 1, 2$), A is taken from all k_1 alleles at the first locus, and B is taken from all k_2 alleles at the second locus.

Next, for a sample with n individuals, there are nv^2 allele pairs, the observed and the expected occurrences of an allele pair AB are respectively $nvP_s^{AB} + nv(v-1)P_d^{AB}$ and $nv^2P_{AQ_B}$. Because $\Delta_{AB} = P_s^{AB} + (v-1)P_d^{AB} - vP_{AQ_B}$, the χ^2 statistic for Burrow's Δ statistic can be established as follows:

$$\chi_{\Delta}^2 = n \sum_{AB} \frac{\hat{\Delta}_{AB}^2}{P_{AQ_B}} \text{ with d.f. } (k_1 - 1) \times (k_2 - 1).$$

d^2 and δ^2

In this study, various moments of LD measures are derived by extending Weir and Hill's (1980) double non-identity coefficients, and thus the exact d^2 can be obtained by using the moments $E(\hat{D}^2)$ and $E(\hat{Q})$ under various mating systems. The exact δ^2 can also be obtained by using the moments $E(\hat{\Delta}^2)$ and $E(\hat{R})$. Hence the value of \hat{r}^2 (or \hat{r}_{Δ}^2) can be approximately replaced by that of d^2 (or δ^2) under each mating system at the equilibrium state. Moreover, the approximate expressions of d^2 and δ^2 under various mating systems are derived by using the transitional matrix, and several relationships are discussed, such as the relationship between \hat{r}^2 (or \hat{r}_{Δ}^2) and the number of generations during reproduction, the relationship between d^2 (or δ^2) and the recombination frequency c , and so on.

Figure 1 shows that after the population has been allowed to reproduce for about 40 generations, \hat{r}^2 (or \hat{r}_{Δ}^2) reaches a relatively steady state, remaining close to the exact d^2 (or δ^2) for several generations. Then, \hat{r}^2 (or \hat{r}_{Δ}^2) begins to decrease again, due to both the fixation of alleles and the positive correlation between \hat{D}^2 and \hat{Q} (or between $\hat{\Delta}^2$ and \hat{R}). As the number of loci decreases, the number of terms in the numerator or the denominator in Eq. (8) is reduced, due to the weighted scheme in Eq. (8) being unable to effectively eliminate the correlation. The number of generations at which \hat{r}^2 (or \hat{r}_{Δ}^2) begins to decrease again depends on v , N_e , L and the initial heterozygosity.

Supplementary Fig. S2 shows that regardless of \hat{r}^2 or \hat{r}_{Δ}^2 , the smaller the recombination frequency, the slower the rate of convergence. Generally, \hat{r}^2 and \hat{r}_{Δ}^2 decrease to a relatively steady state after about $-4.21/\ln(1-c)$ generations. Moreover, under the same recombination frequency, the convergent rates of \hat{r}^2 (or \hat{r}_{Δ}^2) are similar for all levels of ploidy but differ markedly under different recombination frequencies.

Figure 2 (and Supplementary Fig. S3) shows that the relationship between d^2 (or δ^2) and the recombination frequency c has two main features: (i) if c is small (e.g., <0.25), both d^2 and δ^2 for polysomic inheritance decreases more rapidly than those for disomic inheritance and (ii), the difference between $d_{c=0.5}^2$ and $d_{c=1-1/v}^2$ under polysomic inheritance is considerable (the error rate ranges from 10% to 23%), whereas the difference between $\delta_{c=0.5}^2$ and $\delta_{c=1-1/v}^2$ is negligible (the error rate is less than 1.5% for non-HS mating systems).

For (i), this infers that a higher density genetic map is required to detect any linkage in polyploids. A rough estimate would be the

locus density in tetraploids (hexaploids or octoploids) to be 1.58 (2.16 or 2.67) times that for diploids (estimated by the threshold $\delta^2 = 0.2$, see Fig. 2). However, if the locus density is sufficient, the gene mapping in polyploids may be more accurate than that in diploids due to the steep slope of the curve at a low c .

For (ii) this indicates that it is unnecessary to distinguish whether two loci are located on the same chromosome or not if the effective population size N_e is estimated by \hat{r}_{Δ}^2 . From this reason, we can simply let the recombination frequency between any two loci be equal to 0.5, as is assumed in other methods (e.g., England et al. 2006). However, it is necessary to assume that two loci are located on different chromosomes if N_e is estimated by \hat{r}^2 using phased genotypes.

Effective population size

Among the parameters v , n , r^2 , r_{Δ}^2 , N_e , c and f , the first two v and n are known, the next two r^2 and r_{Δ}^2 can be estimated from the genotype data, and the mating system and the ratio f can be obtained from either a priori information, field observations or experiments. The remaining two parameters N_e and c are the parameters we usually need to estimate, and one can be estimated if the other is known.

After simulation, we evaluate the RMSE and the bias of \hat{V} (i.e., $1/\hat{N}_e$). The curves of RMSE among different ploidy levels are similar, indicating that estimating N_e in polyploids requires similar numbers of samples and loci as in diploids. The performance of 100/200 diallelic SNPs is as good as that of 20/40 hexa-allelic SSRs (Supplementary Fig. S5), indicating that the RMSE is mainly determined by the number $\sum_i (k_i - 1)$ of independent alleles. The results for polyploids may be better than for diploids due to smaller biases (Fig. 3).

Some possible sources of this bias of \hat{V} are enumerated as follows. (i) According to Eq. (9), $\hat{r}_{\Delta}^2 - 1/(n - 1)$ is proportional to $1/(N_e - \eta)$, not $1/N_e$, indicating that the estimation of $1/(N_e - \eta)$ may be unbiased, but the estimation of $1/N_e$ is biased. (ii) The recombination frequency between two loci located on the same chromosome is less than 0.5, but it is assumed to be 0.5. (iii) The recombination frequency between two loci located on different chromosomes is $1 - 1/v$, but it is also assumed to be 0.5.

We suggest that (ii) is the main source of this bias. This is because the bias is largely influenced by both the number L of loci used and the ploidy level v (Fig. 3). Because the length of each chromosome is 100 cM, the loci become denser at higher levels of L . The value of δ^2 between two close loci (implying smaller c) therefore increases in the deviation from $\delta_{c=0.5}^2$ (Fig. 2). In addition, the simulation results for polyploids are less biased. This is because the curve of δ^2 at a higher ploidy level is flat for most situations (e.g., $c > 0.2$). To validate our prediction, we use unlinked loci to regenerate the results in Fig. 3, where the loci are on the same chromosome and the distance between two neighboring loci is long (10^{30} cM). The results show the bias is reduced to 10^{-5} (Supplementary Fig. S6).

The bias sources (ii) and (iii) can be reduced if the a priori information is available: (i) if the combination frequency between any two loci is known, the exact δ^2 can be calculated between all loci pairs and averaged. In this case, Eq. (8) should use the arithmetic mean of \hat{r}^2 and \hat{r}_{Δ}^2 ; (ii) if the lengths of chromosomes (in centimorgan) are known, assuming the loci are uniformly distributed on the chromosomes, then the exact δ^2 can be calculated; (iii) if the genome size and the number of chromosomes are both known, we can assume the length of the chromosomes accord with a particular distribution (e.g., triangular or uniform) and obtain the exact δ^2 (Waples et al. 2016); With NEWTON'S approach as we described, the exact δ^2 can be considered a function of the true N_e , then N_e can be estimated; (iv) if the genetic data are sufficient, it is possible to cluster the loci into some linkage groups, and the loci in different lineage groups will be used to perform the estimation of N_e . This can be achieved using a specific software package designed for diploid N_e estimation, i.e., NEESTIMATOR V2 (Do et al. 2014).

Non-independent samples

Non-independent samples can also be a potential bias source (Waples, personal communications). For non-independent samples due to random sampling, there is not extra bias. For non-independent samples due to non-random sampling, e.g., the relatives are more likely to be together sampled, extra bias is introduced.

We performed a simple simulation to show such bias, the results with different sampling strategies (random sampling, pair sampling of relatives) are compared. The bias of \hat{V} is increased under non-random sampling at a low sample size and approaches that under random sampling as n increases (Supplementary Fig. S6). Such bias is mainly due to the overestimation of \hat{r}_{Δ}^2 and $\hat{\Delta}^2$.

We derived the LD moments under pair sampling of clones in Supplementary Appendix K. The LD moments under non-random sampling are related to the sample size, the probability of non-random sampling, the types of relatives, the single and the double non-identity coefficients, the allele probability product pq , and the heterozygosities. Therefore, d^2 and δ^2 cannot be derived by the method used in this manuscript, i.e., Eq. (5), and the elimination of such bias can be a direction of future studies.

REFERENCES

- Brown AHD, Feldman MW, Nevo E (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96:523–536
- Burow MD, Simpson CE, Starr JL, Paterson AH (2001) Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.): broadening the gene pool of a monophyletic polyploid species. *Genetics* 159:823
- Butruille DV, Boiteux LS (2000) Selection–mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proc Natl Acad Sci USA* 97:6608–6613
- Cockerham CC, Weir BS (1977) Digenic descent measures for finite populations. *Genet Res* 30:121–147
- Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29:311–322
- Do C, Waples RS, Peel D, Macbeth G, Tillett BJ, Ovenden JR (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Mol Ecol Resour* 14:209–214
- England PR, Cornuet J-M, Berthier P, Tallmon DA, Luikart G (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv Genet* 7:303
- Fisher RA (1947) The theory of linkage in polysomic inheritance. *Philos Trans R Soc Lond Ser B Biol Sci* 233:55–87
- Gao XY, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* 32:361–369
- Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* 2:204–211
- Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* 13:635–643
- Hill WG (1974) Disequilibrium among several linked neutral genes in finite population I. Mean changes in disequilibrium. *Theor Popul Biol* 5:366–392
- Hill WG (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theor Popul Biol* 8:117–126
- Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genet Res* 38:209–216
- Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theor Appl Genet* 38:226–231
- Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* 54:705
- Hollenbeck C, Portnoy D, Gold J (2016) A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity* 117:207–216
- Hosking LK, Boyd PR, Xu CF, Nissim M, Cantone K, Purvis IJ, Khakhar R, Barnes MR, Liberwirth U, Hagen-Mann K (2002) Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *Pharmacogenomics J* 2:165
- Huang K, Dunn DW, Ritland K, Li BG (2020) polygene: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods Ecol Evol* 11:448–456

- Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* 56:11
- Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49:49
- Maruyama T (1982) Stochastic integrals and their application to population genetics. In: Kimura M (ed) *Molecular evolution, protein polymorphism and the neutral theory*. Japan Scientific Societies Press, Tokyo, p 151–166
- Nei M (1987) *Molecular evolutionary genetics*. Columbia university press, New York
- Ohta T (1980) Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. *Genet Res* 36:181–197
- Ohta T, Kimura M (1969) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* 63:229
- Otto SP (2007) The evolutionary consequences of polyploidy. *Cell* 131:452–462
- Rieger R, Michaelis A, Green MM (1968) *A glossary of genetics and cytogenetics: classical and molecular*. Springer-Verlag, New York, NY
- Robert C (1991) Generalized inverse normal distributions. *Stat Probabil Lett* 11:37–41
- Santiago E, Novo I, Pardiñas AF, Saura M, Wang J, Caballero A (2020) Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. *Mol Biol Evol* 37:3642–3653
- Sattler MC, Carvalho CR, Clarindo WR (2016) The polyploidy and its key role in plant breeding. *Planta* 243:281–296
- Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9:477
- Stift M, Berenos C, Kuperus P, van Tienderen PH (2008) Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cross) microsatellite data. *Genetics* 179:2113–2123
- Sved JA (1964) The relationship between diploid and tetraploid recombination frequencies. *Heredity* 19:585–596
- Sved JA, Cameron EC, Gilchrist AS (2013) Estimating effective population size from linkage disequilibrium between unlinked loci: theory and application to fruit fly outbreak populations. *PLoS ONE* 8:e69078
- Sved JA, Feldman MW (1973) Correlation and probability methods for one and two loci. *Theor Popul Biol* 4:129–132
- Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17:520–526
- Udall JA, Wendel JF (2006) Polyploidy and crop improvement. *Crop Sci* 46:5–3–5–14
- Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv Genet* 7:167–184. <https://doi.org/10.1007/s10592-005-9100-y>
- Waples RS, Antao T, Luikart G (2014) Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics* 197:769–780
- Waples RK, Larson WA, Waples RS (2016) Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity* 117:233–240. <https://doi.org/10.1038/hdy.2016.60>
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35:235–254
- Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 42:105
- Weir BS, Hill WG (1980) Effect of mating structure on variation in linkage disequilibrium. *Genetics* 95:477–488

ACKNOWLEDGEMENTS

We thank Dr. Robin Waples, two anonymous reviewers and the subject editor Prof. Olivier J. Hardy for their helpful suggestions and comments. KH thanks Prof. Kermit Ritland for providing a visiting professor position at UBC.

AUTHOR CONTRIBUTIONS

KH and BGL conceived the ideas, KH and WKL constructed the model, DW checked the model, KH and DWD wrote the draft and DWD edited the manuscript.

FUNDING INFORMATION

This study is funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020302), the National Natural Science Foundation of China (31730104, 32170515, 31770411, 32070453), and the Innovation Capability Support Program of Shaanxi (2021KJXX-027). DWD is supported by a Shaanxi Province Talents 100 Fellowship and KH is supported by a scholarship from China Scholarship Council.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41437-021-00482-1>.

Correspondence and requests for materials should be addressed to Baoguo Li.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.