1 Title

2 A generalized framework of AMOVA with any number of hierarchies and any level of
3 ploidies

4 Authors

5 Kang Huang^{1,2}, Yuli Li¹, Derek W. Dunn¹, Pei Zhang¹, Baoguo Li^{1,3}

6 Addresses

7 ¹ Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences,
8 Northwest University, Xi'an 710069, China

9 ² Department of Forest and Conservation Sciences, University of British Columbia,
10 Vancouver, BC V6T1Z4 Canada.

11 ³ Center for Excellence in Animal Evolution and Genetics, Chinese Academy of
12 Sciences, Kunming 650223 China

13 Keywords

14 Analysis of molecular variance, polyploidy, hierarchy, method-of-moment estimation,
15 maximum-likelihood estimation

16 Corresponding author

17 Baoguo Li

18 Telephone: +8613572209390; Fax: +86 029 88303304; E-mail: baoguoli@nwu.edu.cn

19 Running title

20 A generalized framework of AMOVA

Abstract

The analysis of molecular variance (AMOVA) is a widely used statistical model in the studies of population genetics and molecular ecology. The classical framework of AMOVA only supports haploid and diploid data, in which the number of hierarchies ranges from two to four. In practice, natural populations can be classified into more hierarchies, and polyploidy is frequently observed in contemporary species. The ploidy level may even vary within the same species, even within the same individual. We generalized the framework of AMOVA such that it can be used for any number of hierarchies and any level of ploidy. Based on this framework, we present four methods to account for the multilocus genotypic and allelic phenotypic data. We use simulated datasets and an empirical dataset to evaluate the performance of our framework. We make freely available our methods in a software, POLYGENE, which is freely available at <https://github.com/huangkang1987/>.

Keywords: analysis of molecular variance, polyploidy, hierarchy, method-of-moment estimation, maximum-likelihood estimation

Introduction

The *analysis of molecular variance* (AMOVA) is a statistical model for the molecular variation in a single species. AMOVA was developed by EXCOFFIER *et al.* (1992) based on the previous work of decomposing the total variance of gene frequencies into the variance components in different subdivision levels (COCKERHAM 1969; COCKERHAM

1973).

This statistical model was initially implemented for DNA haplotypes, but it can be applied to any marker datum, e.g., the codominant marker data and the dominant marker data (PEAKALL *et al.* 1995; PEAKALL AND SMOUSE 2006). The classical framework of AMOVA supports haploids and diploids, and the number of hierarchies ranges from two to four (individual, population, group and total population) (EXCOFFIER AND LISCHER 2010). In practice, natural populations can be classified into more than four hierarchies, and the ploidy level may vary within the same species or within the same individual.

In many species, physical or ecological barriers prevent random mating (MARTIN AND WILLIS 2007). The resulting partial or total isolation of populations results in genetic differentiation due to the interacting processes of genetic drift, differential gene-flow and natural selection (LANDE 1976). Because the factors restricting gene flow, such as geographical distance (WRIGHT 1943), landscape features (e.g., mountain, river, desert) (GEFFEN *et al.* 2004; CHAMBERS AND GARANT 2010; LAIT AND BURG 2013), ecological factors (e.g., salt concentration, climatic gradients) (LUPPI *et al.* 2003; YANG *et al.* 2014) and behavioral differences (e.g., parental care) (RUSSELL *et al.* 2004), are not all the same among populations, the gene flow between populations is unevenly distributed. For example, in humans, the intra-city gene-flow is higher than the intra-province, intra-nation and inter-nation gene flows. The population structure, in some situations, can be classified into multilevel hierarchies.

Polyploids represent a significant portion of plant species, with anywhere between

30% and 80% of angiosperms showing polyploidy (BUROW *et al.* 2001) and most lineages showing the evidence of paleopolyploidy (OTTO 2007). Due to their significant roles in molecular ecology, evolutionary biology and agriculture studies, polyploids have increasingly become the focus of theoretical and experimental research (AVNI *et al.* 2017; LING *et al.* 2018). There are two major problems in the population-genetics analysis of polyploids: genotyping ambiguity and double reduction.

For the *polymerase chain reaction* (PCR)-based markers, because the dosage of alleles cannot be determined by electrophoresis bands, the true genotype cannot be identified from the electrophoresis. This phenomenon is called *genotyping ambiguity*. For example, if an autotetraploid genotype *AAAB* has the same electrophoresis band type as the genotype *AABB*, then these two genotypes cannot be distinguished by electrophoresis.

In polyploids, double reduction occurs when a pair of sister chromatids is segregated into the same chromosome, which will cause the corresponding genotypic frequency deviating from *Hardy-Weinberg equilibrium* (HWE), where we assume that each allele will randomly appear within various genotypes. For autotetraploids, the rate α of double-reduction is assumed to be 0 under HWE, $1/7$ under *random chromosome segregation* (RCS) (HALDANE 1930), and $1/6$ under *complete equational segregation* (CES) (MATHER 1935). In the *partial equational segregation* (PES) model, the distance between the target locus and the centromere is incorporated into CES (HUANG *et al.* 2019).

Some software for the polysomic inheritance model has been developed, e.g., POLYSAT (CLARK AND JASIENIUK 2011), SPAGEDI (HARDY AND VEKEMANS 2002),

POLYRELATEDNESS (HUANG *et al.* 2014), GENODIVE (MEIRMANS AND TIENDEREN 2004; MEIRMANS AND LIU 2018), and STRUCTURE (PRITCHARD *et al.* 2000). However, some of them cannot solve the genotyping ambiguity, and all of them are supposed that the genotypic frequencies accord with HWE.

In this paper, we generalize the framework of AMOVA such that any number of hierarchies and any level of ploidy are allowed. Four methods are developed to account for multilocus genotypic and phenotypic data, including three method-of-moment methods and one maximum-likelihood method. Our model has been implemented in a software named POLYGENE, and it is freely available at <https://github.com/huangkang1987/>. POLYGENE is designed for genotypic or phenotypic datasets, which only supports homoploids to include more population-genetics analyses (e.g., phenotypic/genotypic distribution test).

Theory and modeling

There are three purposes of typical AMOVA: (i) estimate the variance components in different subdivision levels; (ii) measure the population differentiation with F -statistics (F_{IS} , F_{IT} , F_{ST} , etc.); (iii) test the significance of differentiation. In the following sections, we will briefly describe the general procedures of classic framework of AMOVA, then extend them to the generalized situation.

Classic framework

The procedures of AMOVA are as follows: (i) calculate the genetic distance between

two alleles or two haplotypes; (ii) calculate the *sum of squares* (SS), the degree of freedom and the *mean square* (MS) in each source of variation; (iii) solve variance components; (iv) calculate *F*-statistics; (v) perform permutation tests.

The collection consisting of some populations is called a *group*, denoted by g . We stipulate that each population can only belong to one group, and the union of all groups is the total population. Because an allele or a haplotype (for simplicity, we use ‘allele’ to refer them hereafter) is usually neither a discrete nor a continuous random variable (except the allele size in microsatellites), the SS cannot be calculated by the equation $SS = \sum_i (X_i - \bar{X})^2$. Using the genetic distance between any two alleles as a proxy, an alternative method can be used to calculate the SS, whose formula for a group of n allele copies is as follows:

$$SS = \sum_{1 \leq i < j \leq n} \frac{d_{ij}^2}{n}, \quad (1)$$

where d_{ij} is the genetic distance between the i^{th} and the j^{th} alleles. Such genetic distance is one of the following distances: nucleotide difference distance for DNA sequences (DNA sequence, EXCOFFIER *et al.* 1992), Euclidean distance for dominant markers (dominant marker, PEAKALL *et al.* 1995), *infinity allele model* (IAM) distance for codominant marker (codominant marker, COCKERHAM 1973) and *stepwise mutation model* distance (SMM) for microsatellites (microsatellite, SLATKIN 1995).

In variance decomposition, the genetic variance is decomposed as two to four hierarchies, including σ_{WI}^2 (within individuals), $\sigma_{AI/WP}^2$ (among individuals within

populations), $\sigma_{AP/WG}^2$ (among populations within groups) and σ_{AG}^2 (among groups), where σ_{WI}^2 and/or σ_{AG}^2 are sometimes ignored. Using all of the four hierarchies as an example, the layout of AMOVA is shown in Table 1.

By equating the expected MS with the observed MS, the unbiased estimates of variance components can be solved. After that, the F -statistics can be calculated by the following formulas:

$$\begin{aligned} F_{SC} &= \frac{\sigma_{AP/WG}^2}{\sigma_{AP/WG}^2 + \sigma_{AI/WP}^2 + \sigma_{WI}^2}, \\ F_{CT} &= \frac{\sigma_{AG}^2}{\sigma_{AG}^2 + \sigma_{AP/WG}^2 + \sigma_{AI/WP}^2 + \sigma_{WI}^2}, \\ F_{IS} &= \frac{\sigma_{AI/WP}^2}{\sigma_{AI/WP}^2 + \sigma_{WI}^2}, \\ F_{IT} &= \frac{\sigma_{AG}^2 + \sigma_{AP/WG}^2 + \sigma_{AI/WP}^2}{\sigma_{AG}^2 + \sigma_{AP/WG}^2 + \sigma_{AI/WP}^2 + \sigma_{WI}^2}, \\ F_{ST} &= \frac{\sigma_{AP/WG}^2 + \sigma_{AG}^2}{\sigma_{AG}^2 + \sigma_{AP/WG}^2 + \sigma_{AI/WP}^2 + \sigma_{WI}^2}. \end{aligned}$$

A permutation test is used to test the significance of differences. The null hypothesis is that there is no differentiation among individuals, populations or groups, and the observed differences are due to the random sampling. This statement is equivalent to the variance σ_{WI}^2 occupying 100% of the total variance, and thus the variances $\sigma_{AI/WP}^2$, $\sigma_{AP/WG}^2$ and σ_{AG}^2 together with various F -statistics are all zero.

In each permutation, the allele copies are randomly permuted in the total population to generate a new dataset. The variance components and the F -statistics are calculated for each permuted dataset to obtain their distributions under null hypothesis. The probability that each permuted variance component or each F -statistic is greater than the original value is used as a single-tailed P -value.

Generalized framework

Here, we present a generalized framework to decompose the genetic variance. In our framework, we do not need to use the degrees of freedom or the MS. Instead, we directly use the variance components to express the expected SS of various hierarchies, whose expressions are as follows:

$$\begin{aligned} E(SS_{WI}) &= \sigma_{WI}^2(H - N), \\ E(SS_{WP}) &= \sigma_{WI}^2(H - P) + \sigma_{AI/WP}^2 \left(H - \sum_{\rho} \sum_{i \in \rho} \frac{v_i^2}{v_{\rho}} \right), \\ E(SS_{WG}) &= \sigma_{WI}^2(H - G) + \sigma_{AI/WP}^2 \left(H - \sum_g \sum_{i \in g} \frac{v_i^2}{v_g} \right) + \sigma_{AP/WG}^2 \left(H - \sum_g \sum_{\rho \in g} \frac{v_{\rho}^2}{v_g} \right), \\ E(SS_{TOT}) &= \sigma_{WI}^2(H - 1) + \sigma_{AI/WP}^2 \left(H - \sum_i \frac{v_i^2}{v_t} \right) + \sigma_{AP/WG}^2 \left(H - \sum_{\rho} \frac{v_{\rho}^2}{v_t} \right) + \sigma_{AG}^2 \left(H - \sum_g \frac{v_g^2}{v_t} \right), \end{aligned} \quad (2)$$

where H and v_t are, respectively, the total number of haplotypes and alleles, v_i (v_{ρ} or v_g) is the number of haplotypes/alleles within the individual i (the population ρ or the group g), and the mobile subscript in $\sum_{\rho}$ ($\sum_g$ or $\sum_i$) is taken from all populations (all groups or all individuals). The estimates of variance components are identical to those in the classic framework. The step-by-step detailed derivations for these formulas in Equation (2) are provided in the Supplementary materials. Compared with previous frameworks (Table 1), Equation (2) is more regular, making it possible to be generalized.

According to the concept that each member in a hierarchy is a ‘vessel’ of genes, we can use the vessels V_0, V_1, V_2, V_3, V_4 and the corresponding expected SS_i ($i = 1, 2, 3, 4$) to describe each formula in Equation (2). In other words, we can use these vessels to rewrite Equation (2):

$$\begin{aligned}
 E(SS_1) &= \sigma_1^2 \left(|V_4| - \sum_{V_1} \sum_{V_0 \in V_1} \frac{|V_0|^2}{|V_1|} \right), \\
 E(SS_2) &= \sigma_1^2 \left(|V_4| - \sum_{V_2} \sum_{V_0 \in V_2} \frac{|V_0|^2}{|V_2|} \right) + \sigma_2^2 \left(|V_4| - \sum_{V_2} \sum_{V_1 \in V_2} \frac{|V_1|^2}{|V_2|} \right), \\
 E(SS_3) &= \sigma_1^2 \left(|V_4| - \sum_{V_3} \sum_{V_0 \in V_3} \frac{|V_0|^2}{|V_3|} \right) + \sigma_2^2 \left(|V_4| - \sum_{V_3} \sum_{V_1 \in V_3} \frac{|V_1|^2}{|V_3|} \right) + \sigma_3^2 \left(|V_4| - \sum_{V_3} \sum_{V_2 \in V_3} \frac{|V_2|^2}{|V_3|} \right), \quad (3) \\
 E(SS_4) &= \sigma_1^2 \left(|V_4| - \sum_{V_4} \sum_{V_0 \in V_4} \frac{|V_0|^2}{|V_4|} \right) + \sigma_2^2 \left(|V_4| - \sum_{V_4} \sum_{V_1 \in V_4} \frac{|V_1|^2}{|V_4|} \right) + \sigma_3^2 \left(|V_4| - \sum_{V_4} \sum_{V_2 \in V_4} \frac{|V_2|^2}{|V_4|} \right) \\
 &\quad + \sigma_4^2 \left(|V_4| - \sum_{V_4} \sum_{V_3 \in V_4} \frac{|V_3|^2}{|V_4|} \right),
 \end{aligned}$$

154 where V_i is a vessel at the level i , $|V_i|$ is the number of allele copies in V_i , SS_i is the SS
 155 within all vessels at the level i , σ_i^2 is the variance component among all V_{i-1} within V_i ,
 156 and the mobile subscript in $\sum_{V_i}$ is taken from all vessels at the level i . The subscript i
 157 ranges from 0 to 4, and the corresponding vessels represent, in turn, alleles, individuals,
 158 populations, groups and the total population. Equation (3) is in apple-pie order, which
 159 can be expressed as the forms of summation signs:

$$E(SS_i) = \sum_{j=1}^i \sigma_j^2 \left(|V_4| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} \right), \quad i = 1, 2, 3, 4.$$

160 If the hierarchy of individuals is ignored, then the vessel V_1 (V_2 or V_3) will
 161 represent a population (a group or the total population). In this situation, Equation (3)
 162 becomes

$$E(SS_i) = \sum_{j=1}^i \sigma_j^2 \left(|V_3| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} \right), \quad i = 1, 2, 3.$$

163 Generally, if there are $M + 1$ kinds of vessels at the levels ranging from 0 to M , then

Equation (3) can be generalized as follows:

$$E(SS_i) = \sum_{j=1}^i \sigma_j^2 \left(|V_M| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} \right), \quad i = 1, 2, \dots, M, \quad (4)$$

where V_M denotes the vessel of highest hierarchy (i.e., the total population). Equation (4) is the ultimate generalized form of AMOVA, which is extremely simple and can be applied to any number of hierarchies and any level of ploidy. We can also use matrices to express Equation (4):

$$\mathbf{S} = \mathbf{C}\mathbf{\Sigma},$$

where $\mathbf{S} = [E(SS_1), E(SS_2), \dots, E(SS_M)]^T$, $\mathbf{\Sigma} = [\sigma_1^2, \sigma_2^2, \dots, \sigma_M^2]^T$ and the coefficient matrix $\mathbf{C}$ is lower-triangular with type $M \times M$, whose ij^{th} element is

$$C_{ij} = \begin{cases} |V_M| - \sum_{V_i} \sum_{V_{j-1} \in V_i} \frac{|V_{j-1}|^2}{|V_i|} & \text{if } i \geq j, \\ 0 & \text{if } i < j, \end{cases} \quad i, j = 1, 2, \dots, M.$$

Then, a method-of-moment estimation of variance components can be given by $\hat{\mathbf{\Sigma}} = \mathbf{C}^{-1}\mathbf{\hat{S}}$, and the F -statistics can be solved by

$$\hat{F}_{ij} = 1 - \frac{\sum_{k=1}^i \hat{\sigma}_k^2}{\sum_{k=1}^j \hat{\sigma}_k^2}, \quad 1 \leq i < j \leq M. \quad (5)$$

Method-of-moment methods

For convenience, we call a method-of-moment estimator a *moment method*. In practice, the multilocus data are used to increase the accuracy of estimation. Based on the moment estimator described above, we develop three methods (called the *homoploid*, *anisoploid*

and *weighting genotypic methods*) to account for the multilocus genotypic or phenotypic data.

The **homoploid method** is only applicable to homoploids. In this method, all loci are treated as one dummy locus, and the dummy haplotypes are extracted from phenotypes. Meanwhile, the genetic distance between any two dummy haplotypes is calculated, and these dummy haplotypes are permuted to test the significance of differentiation. In diploids, this method is used in GENALEX (PEAKALL AND SMOUSE 2006).

To solve the genotyping ambiguity, we will use the posterior probabilities to weight the possible genotypes hidden behind a phenotype. The multiset consisting of alleles within an individual and at a locus is defined as a *genotype*, denoted by $\mathcal{G}$, and the set obtained by deleting the duplicated alleles in $\mathcal{G}$ is defined as the *phenotype* determined by $\mathcal{G}$, denoted by $\mathcal{P}$. In our previous paper (HUANG *et al.* 2019), the genotypic frequency $\Pr(\mathcal{G})$ and the phenotypic frequency $\Pr(\mathcal{P})$ under a double-reduction model (HWE, RCS, CES or PES) were calculated. On this basis, we are able to calculate the posterior probability $\Pr(\mathcal{G}|\mathcal{P})$ of a genotype $\mathcal{G}$ determining $\mathcal{P}$, whose formula is as follows:

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

After that, the probability $\Pr(A_{hl} = A_{lj})$ (or p_{hlj} for short) can be calculated by

$$p_{hlj} = \Pr(A_{hl} = A_{lj}) = \sum_{\mathcal{G}} \Pr(\mathcal{G}|\mathcal{P}) \Pr(A_{hl} = A_{lj}|\mathcal{G}), \quad (6)$$

where A_{hl} is the allele in the h^{th} dummy haplotype and at the l^{th} locus, A_{lj} is the j^{th} allele at the l^{th} locus, and $\mathcal{G}$ is taken from all possible genotypes determining $\mathcal{P}$.

In the homoploid method, because all loci are treated as one dummy locus, the

square of genetic distance $d_{hh'}$ between the h^{th} and the h'^{th} haplotype is the sum of squares of the distances in the form $d_{A_{lj}A_{lj'}}$ over all L loci, namely

$$d_{hh'}^2 = \sum_{l=1}^L d_{A_{hl}A_{h'l}}^2 = \sum_{l=1}^L \sum_{j=1}^{J_l} \sum_{j'=1}^{J_l} p_{hlj} p_{h'lj'} d_{A_{lj}A_{lj'}}^2,$$

where $d_{A_{lj}A_{lj'}}$ is the distance between the j^{th} and the j'^{th} allele at the l^{th} locus, and J_l is the number of alleles at the l^{th} locus. For an allele with missing data, its frequency refers the frequency in the corresponding population. In this method, because there is only one dummy locus, no additional weighting procedure is required for multilocus data.

The **anisoploid method** can be applied for both homoploids and anisoploids. In this method, the dummy alleles are extracted at each locus, and the missing data are ignored. Meanwhile, the genetic distance between two dummy alleles needs to be calculated locus by locus, and the dummy alleles are randomly permuted locus by locus during the permutation test.

For this method, the probability $\Pr(A_h = A_j)$ (or p_{hj} for short) in a phenotype at a target locus can also be expressed by Equation (6). Then the square of genetic distance $d_{A_hA_{h'}}$ between two allele copies A_h and $A_{h'}$ at this target locus is given by

$$d_{A_hA_{h'}}^2 = \sum_j^J \sum_{j'}^J p_{hj} p_{h'j'} d_{A_jA_{j'}}^2,$$

where J is the number of alleles at this target locus.

The untyped individuals (populations or groups) due to missing data should be directly skipped to avoid a denominator of zero. The global variance components for

214 multilocus data can be solved by using the formula $\mathbf{S} = \mathbf{C}\mathbf{\Sigma}$. There are two solving
 215 strategies: (i) find the sum of the whole $\mathbf{S}$ and the sum of the whole $\mathbf{C}$ over all loci, and
 216 then solve the global variance components, denoted by $\mathbf{\Sigma}_{g_1}$; (ii) solve $\mathbf{\Sigma}$ for each locus,
 217 and then find the sum of whole $\mathbf{\Sigma}$ over all loci, denoted by $\mathbf{\Sigma}_{g_2}$. Generally, $\mathbf{\Sigma}_{g_1}$ and $\mathbf{\Sigma}_{g_2}$
 218 are different, but they are approximately proportional to each other.

219 We adopt the first strategy because the global SS, the d.f. and the MS can also be
 220 obtained. This strategy has the same output style as the classic framework.

221 In the **weighting genotypic method**, no dummy haplotypes are extracted. Instead,
 222 for any hierarchy i , the SS_i for each genotype hidden behind a phenotype is calculated,
 223 and then all sums of squares in the hierarchy i are weighted according to the
 224 corresponding posterior probabilities. We also find the sum of those SS_i over all loci in
 225 this method. For each locus, the SS_i is calculated by

$$SS_i = \sum_{V_i} \frac{1}{|V_i|} \left[\sum_{j=1}^{N_{V_i}} d^2(\mathcal{P}_j) + \sum_{1 \leq j < k \leq N_{V_i}} d^2(\mathcal{P}_j, \mathcal{P}_k) \right],$$

226 where $SS_1 = SS_{WI}$ (i.e., when $i = 1$), V_i is taken from all vessels in the hierarchy i , N_{V_i}
 227 is the number of phenotypes determined by the individuals within V_i , and at this locus,
 228 $d^2(\mathcal{P}_j)$ (or $d^2(\mathcal{P}_j, \mathcal{P}_k)$) is the weighted sum of squares of the distances within the
 229 phenotype $\mathcal{P}_j$ (or between the phenotypes $\mathcal{P}_j$ and $\mathcal{P}_k$), which can be calculated by the
 230 following formulas:

$$d^2(\mathcal{P}) = \frac{1}{2} \sum_{\mathcal{G}} \sum_{A, B \in \mathcal{G}} \Pr(\mathcal{G}|\mathcal{P}) d_{AB}^2,$$

$$d^2(\mathcal{P}_1, \mathcal{P}_2) = \sum_{\mathcal{G}_1} \sum_{\mathcal{G}_2} \sum_{A \in \mathcal{G}_1} \sum_{B \in \mathcal{G}_2} \Pr(\mathcal{G}_1|\mathcal{P}_1) \Pr(\mathcal{G}_2|\mathcal{P}_2) d_{AB}^2,$$

231 where $\mathcal{G}$ ($\mathcal{G}_1$ or $\mathcal{G}_2$) is taken from all candidate genotypes determining $\mathcal{P}$ ($\mathcal{P}_1$ or $\mathcal{P}_2$),
 232 and d_{AB} is the distance between the alleles A and B .

233 **Maximum-likelihood method**

234 We will develop a maximum-likelihood estimator to estimate the F -statistics and
 235 solve the variance components. For convenience, we call this method the *likelihood method*.
 236 In this method, a reversed procedure is used, such that the F -statistics are first estimated,
 237 and next the variance components and other statistics are solved.

238 To derive the expression of the likelihood for individuals, we first model some
 239 equations. A random distribution can be used to simulate the differentiation among
 240 individuals within a vessel V_i . We will choose some Dirichlet distribution for each
 241 hierarchy. That is because the standardized variance of each allele frequency accords
 242 with the corresponding F -statistic in that distribution. Therefore, no additional weighting
 243 procedure for the variance components is required.

244 Given a vessel V_i ($2 \leq i \leq M$), the allele frequencies $p_{11}, p_{12}, \dots, p_{1J}$ within an
 245 individual (i.e., within one of those V_1) are drawn from the Dirichlet distribution
 246 $\mathcal{D}(\alpha_{1i1}, \alpha_{1i2}, \dots, \alpha_{1iJ})$, where J is the number of alleles within this individual, and

$$\alpha_{1ij} = (1/F_{1i} - 1)p_{ij}, \quad j = 1, 2, \dots, J,$$

247 in which F_{1i} is the F -statistic among all individuals within V_i , and p_{ij} is the frequency
 248 of j^{th} allele in V_i . Then, the expectation and the variance of j^{th} allele frequency p_{1j} as
 249 a random variable are, respectively, p_{ij} and $F_{1i}p_{ij}(1 - p_{ij})$, and the standardized
 250 variance of p_{ij} as a random variable is exactly $F_{i,i+1}$, which is identical to Wright's

251 definition of F -statistics.

252 For simplicity, we let $\mathbf{p}_i$ be the vector consisting of the frequencies of all alleles in V_i ,
 253 i.e.

$$\mathbf{p}_i = [p_{i1}, p_{i2}, \dots, p_{ij}], \quad i = 1, 2, \dots, M.$$

254 Then, for each i with $2 \leq i \leq M$, the probability density function of $\mathbf{p}_1$ is as follows:

$$f(\mathbf{p}_1 | \mathbf{p}_i, F_{1i}) = \frac{\prod_{j=1}^J \Gamma(\alpha_{ij})}{\Gamma(\sum_{j=1}^J \alpha_{ij})} \prod_{j=1}^J p_{1j}^{\alpha_{ij}-1},$$

255 where $\Gamma(\cdot)$ is the gamma function, and $\alpha_{ij} = 1/F_{ij} - 1$. Assume that the alleles within
 256 V_1 are independently drawn according to the frequencies in $\mathbf{p}_1$. Then, the allele copy
 257 numbers of V_1 obey a multinomial distribution, and so the frequency $\Pr(\mathcal{G} | \mathbf{p}_1)$ of a
 258 genotype $\mathcal{G}$ conditional on $\mathbf{p}_1$ is

$$\Pr(\mathcal{G} | \mathbf{p}_1) = \binom{|V_1|}{c_1, c_2, \dots, c_J} p_{11}^{c_1} p_{12}^{c_2} \dots p_{1J}^{c_J},$$

259 where c_j is the number of the j^{th} allele copies in $\mathcal{G}$, $j = 1, 2, \dots, J$.

260 Now, the frequency $\Pr(\mathcal{G} | \mathbf{p}_i, F_{1i})$ of $\mathcal{G}$ conditional on both $\mathbf{p}_i$ and F_{1i} can be
 261 obtained from the weighted average of $\Pr(\mathcal{G} | \mathbf{p}_1)$ with $f(\mathbf{p}_1 | \mathbf{p}_i, F_{1i}) d\mathbf{p}_1$ as the weight,
 262 that is,

$$\Pr(\mathcal{G} | \mathbf{p}_i, F_{1i}) = \int_{\Omega} \Pr(\mathcal{G} | \mathbf{p}_1) f(\mathbf{p}_1 | \mathbf{p}_i, F_{1i}) d\mathbf{p}_1, \quad i = 2, 3, \dots, M,$$

263 where the integral domain Ω can be expressed as

$$\Omega = \{(p_{11}, p_{12}, \dots, p_{1J}) \mid p_{11} + p_{12} + \dots + p_{1J} = 1, p_{1j} \geq 0, j = 1, 2, \dots, J\}.$$

264 The integral can be converted into the following repeated integral with the multiplicity
 265 $J - 1$:

$$\begin{aligned}
 \Pr(\mathcal{G}|\mathbf{p}_i, F_{1i}) &= \int_0^1 \int_0^{1-p_{11}} \dots \int_0^{1-p_{11}-p_{12}-\dots-p_{1,J-2}} \Pr(\mathcal{G}|\mathbf{p}_1) f(\mathbf{p}_1|\mathbf{p}_i, F_{1i}) dp_{11} dp_{12} \dots dp_{1,J-1} \\
 &= \binom{|V_1|}{c_1, c_2, \dots, c_J} \frac{\Gamma(\alpha_{1i})}{\Gamma(|V_1| + \alpha_{1i})} \prod_{j=1}^J \frac{\Gamma(\alpha_{1ij} + c_j)}{\Gamma(\alpha_{1ij})} \\
 &= \binom{|V_1|}{c_1, c_2, \dots, c_J} \prod_{j=1}^J \prod_{j=0}^{c_j-1} (\alpha_{1ij} + j) / \prod_{j=0}^{|V_1|-1} (\alpha_{1i} + j), \tag{7}
 \end{aligned}$$

266 where $\alpha_{1i} = 1/F_{1i} - 1$. Importantly, $\alpha_{1i} \rightarrow +\infty$ if $F_{1i} \rightarrow 0^+$, thus the variance $\sigma_{1i}^2 \rightarrow 0$ if

267 $F_{1i} \rightarrow 0^+$. Since $\mathbf{p}_i$ is unavailable, the estimate $\hat{\mathbf{p}}_i$ is used as $\mathbf{p}_i$ in our calculation.

268 The frequency $\Pr(\mathcal{P}|\mathbf{p}_i, F_{1i})$ of a phenotype $\mathcal{P}$ conditional on both $\mathbf{p}_i$ and F_{1i} is

269 the sum of frequencies in the form $\Pr(\mathcal{G}|\mathbf{p}_i, F_{1i})$, where $\mathcal{G}$ is taken from all candidate

270 genotypes determining $\mathcal{P}$, in other words,

$$\Pr(\mathcal{P}|\mathbf{p}_i, F_{1i}) = \sum_{\mathcal{G}} \Pr(\mathcal{G}|\mathbf{p}_i, F_{1i}), \quad i = 2, 3, \dots, M.$$

271 Now, the global likelihood for individuals at a hierarchy i can be obtained, which is

272 the product of frequencies in the form $\Pr(\mathcal{P}_{lj}|\mathbf{p}_i, F_{1i})$ over all individuals and at all loci,

273 symbolically

$$\mathcal{L}_i = \prod_{l=1}^L \prod_{j=1}^N \Pr(\mathcal{P}_{lj}|\mathbf{p}_i, F_{1i}), \quad i = 2, 3, \dots, M.$$

274 Because the allele frequencies are already estimated, a downhill simplex algorithm

275 (NELDER AND MEAD 1965) can be used to find the optimal F_{1i} under the IAM model.

276 After that, the variance components can be solved from the F -statistics with an additional

277 constraint as follows:

$$E(SS_M) = \sum_{j=1}^M \sigma_j^2 \left(|V_M| - \sum_{V_{j-1} \in V_M} \frac{|V_{j-1}|^2}{|V_M|} \right),$$

278 where SS_M can be obtained from the allele frequencies of the total population under the

279 IAM model, that is,

$$SS_M = |V_M| \sum_{1 \leq i < j \leq J} p_{Mi} p_{Mj} d_{A_i A_j}^2,$$

280 where $d_{A_i A_j}$ is the IAM distance between the alleles A_i and A_j .

281 **Differentiation test**

282 In the homoploid/anisoploid method, the dummy haplotypes/alleles are extracted.
 283 Then, the differentiation test can be performed by permuting the dummy
 284 haplotypes/alleles. However, for the weighting genotypic and the likelihood methods,
 285 this cannot be done because there are neither dummy haplotypes nor dummy alleles
 286 being extracted in these two methods.

287 To solve this problem, we develop an alternative method to test the differences. In
 288 this method, the datasets of the same structure as the original datasets are randomly
 289 generated, where ‘the same structure’ means that there are the same individuals,
 290 populations and groups as well as the same missing data. More specifically, the
 291 genotypes of each individual are generated conditional on $\mathbf{p}_M$ and F_{1M} according to
 292 Equation (7) under the null hypothesis that there are no differences (i.e., $F_{1M} \rightarrow 0^+$).
 293 Moreover, the phenotypes can be obtained by removing the duplicated alleles in the
 294 generated genotypes.

295 For each generated dataset, the variance components and the F -statistics are
 296 estimated by the same procedures as above to obtain their empirical distributions.
 297 Similarly, the probability that each permuted variance component or each F -statistic is
 298 greater than the original value is used as a single-tailed P -value.

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables.

Evaluations

Simulated data

A Monte-Carlo simulation is used to assess the accuracy of the four methods mentioned above (three moment methods and one likelihood method) under different conditions: ploidy level, number of hierarchies and population differentiations. We choose three types of hierarchies: $M = 3, 4$ or 5 . If we denote by n_i the number of vessels in the form V_{i-1} in V_i , then the ploidy level n_1 of each individual (i.e., the number of allele copies in each individual at a locus) is set as $n_1 = 2, 4$ or 6 , and the number n_2 of individuals sampled in each population ranges from 5 to 50 at intervals of 5 . For those higher-hierarchy vessels, we set $n_3 = 4$ and $n_4 = n_5 = 2$. In the following discussion, we will use N to replace the symbol n_2 . Meanwhile, we set the number of loci per population as 10 and set the number of alleles per locus $J = 6$. We simulate these three types of hierarchies at each of the three ploidy levels in turn.

For the total population (i.e., V_M), the allele frequencies $p_{M1}, p_{M2}, \dots, p_{MJ}$ ($M = 3, 4$ or 5 , $J = 6$) are randomly drawn from the Dirichlet distribution $\mathcal{D}(1, 1, \dots, 1)$ with all concentration parameters being equal to one. The F -statistic $F_{i,i+1}$ among all V_i within V_{i+1} is set as 0.05 . To simulate the differentiation, the allele frequencies in $\mathbf{p}_i$ for each V_i are independently generated according to both $\mathbf{p}_{i+1}$ and $F_{i,i+1}$. More specifically, $p_{i1}, p_{i2}, \dots, p_{iJ}$ are randomly drawn from the Dirichlet distribution $\mathcal{D}(\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{iJ})$, $i =$

320 $1, 2, \dots, M - 1$, where

$$\lambda_{ij} = (1/F_{i,i+1} - 1)p_{i+1,j}, \quad j = 1, 2, \dots, J,$$

321 in which $p_{i+1,j}$ is the frequency of j^{th} allele in the upper vessel V_{i+1} . Obviously, each

322 λ_{ij} is proportional to $p_{i+1,j}$.

323 The alleles in each individual are randomly drawn according to the allele
324 frequencies $p_{11}, p_{12}, \dots, p_{1J}$ for this individual (i.e., one of the vessels in the form V_1). For
325 polyploids, the duplicated alleles within a genotype $\mathcal{G}$ will be removed to convert $\mathcal{G}$
326 into a phenotype $\mathcal{P}$. The genotype $\mathcal{G}$ or the phenotype $\mathcal{P}$ is randomly set as $\emptyset$ at a
327 probability of 0.05 to simulate the negative amplification.

328 For any combination of simulation parameters, 5,000 datasets are generated, and
329 then the AMOVA for every generated dataset is performed by using each of these four
330 methods. The allele frequencies for each population at each locus are independently
331 estimated by using the double-reduction model under RCS with inbreeding. An
332 expectation-maximization algorithm modified from KALINOWSKI AND TAPER (2006) is
333 used to estimate the frequencies of alleles. In this algorithm, the initial value of each allele
334 frequency at a target locus is assigned as $1/J$, and then each frequency is iteratively
335 updated until the sequence consisting of those updated values is convergent. The
336 updated frequency $\hat{p}'_j$ is calculated by

$$\hat{p}'_j = \frac{\sum_{\mathcal{P}} \sum_{\mathcal{G}} \Pr(\mathcal{G}|\mathcal{P}) \Pr(A_j|\mathcal{G})}{\sum_{\mathcal{P}} \sum_{\mathcal{G}} \Pr(\mathcal{G}|\mathcal{P})}, \quad j = 1, 2, \dots, J,$$

337 where $\mathcal{P}$ is taken from all phenotypes at this target locus, $\mathcal{G}$ is taken from all possible
338 genotypes determining $\mathcal{P}$, $\Pr(\mathcal{G}|\mathcal{P})$ is the posterior probability of $\mathcal{G}$ determining $\mathcal{P}$,

and $\Pr(A_j|G)$ is the frequency of the j^{th} allele in G and at this target locus.

Because the dimension of variance components depends on the allele frequencies and the number of loci, to estimate the F -statistics, what we truly need is the proportion of $\sum_{k=1}^i \hat{\sigma}_k^2$ to $\sum_{k=1}^j \hat{\sigma}_k^2$ according Equation (5). Therefore, we use the bias and the RMSE of the F -statistics to evaluate the accuracy of estimates of F -statistics, where RMSE is the abbreviation of *root-mean-square error*.

Simulated results

The bias and the RMSE of $\hat{F}_{i,i+1}$ ($1 \leq i < M$) for diploids and under different conditions are shown in Figures 1 and 2, respectively. It can be found from Figure 1 that each bias is generally reduced as the sample size N increases, and its variation trend does not change as the number M of hierarchies increases. The bias of $\hat{F}_{M-1,M}$ is smaller than that of $\hat{F}_{i,i+1}$ ($i \leq M - 2$). For the anisoploid and the weighting genotypic methods, the estimates of F -statistics are unbiased. However, for the homoploid method, they are slightly biased due to the weighting for missing data. For the likelihood method, the bias is largest, reaching 0.05, but it drops to below 0.015 at $N = 50$. As M increases, the bias of $\hat{F}_{M-1,M}$ is within the range of the other F -statistics.

It can be seen from Figure 2 that for the three moment methods, the RMSEs of $\hat{F}_{i,i+1}$ are similar: if M is lower (e.g., $M = 3$), the variance decreases quickly as N increases. However, if M is larger, the RMSE of $\hat{F}_{i,i+1}$ is less sensitive to the changes in N , and the RMSE of $\hat{F}_{M-1,M}$ becomes more and more inaccurate as M increases. In contrast, for the likelihood method, the RMSE of $\hat{F}_{i,i+1}$ is less affected by M , and the RMSE of $\hat{F}_{M-1,M}$

lies among those of $\hat{F}_{i,i+1}$ ($i \leq M - 2$).

The bias and the RMSE of $\hat{F}_{i,i+1}$ for tetraploids and under different conditions are shown in Figures 3 and 4, respectively. It can be seen from Figure 3 that for the three moment methods, the estimates of F -statistics become biased for the polyploid phenotypic data. The bias of $F_{1,2}$ is largest, reaching -0.01 at $N = 50$. For the weighting genotypic method, the estimates of F -statistics are also biased, but their biases drop to 0.003 at $N = 50$. For the likelihood method, the bias is larger than that of the weighting genotypic method, reaching 0.02 at $N = 50$. As M increases, the bias of $\hat{F}_{M-1,M}$ is also around those of the other F -statistics.

Compared with the situation of diploids, the RMSE in Figure 4 is reduced in scale, while the patterns are similar to those in diploids. For the weighting genotypic method, the RMSE of $\hat{F}_{i,i+1}$ becomes less sensitive to N as M increases, and the RMSE of $\hat{F}_{M-1,M}$ is largest. For the likelihood method, the sensitivity of RMSE of $\hat{F}_{i,i+1}$ does not vary significantly as M increases.

Empirical data

We will use the human dataset of PEMBERTON *et al.* (2013) to evaluate our generalized framework of AMOVA. This dataset consists of 5795 individuals sampled from 267 worldwide populations (e.g., ethnic groups). These populations are genotyped at 645 autosomal microsatellite loci. The average genotyping rate is 97.02%. In this dataset, the notion of groups needs to be divided into two levels, called *groups I* and *groups II*, to generate a nested structure with five levels (individual, population, group I,

group II, total population).

The collection consisting of several populations in some countries or areas is defined as a *group I*, and the collection consisting of several populations in some region (e.g., East Asia or Middle East) is defined as a *group II*. For example, the populations of all Chinese nations are assigned to East Asia, whereas the population of the Uygur ethnic group is originally in Central South Asia. We still stipulate that each population (or each group I) can only belong to one group I (or one group II), and the union of all groups with the same level is the total population.

Empirical results

Because PEMBERTON *et al.* (2013) dataset is genotypic and because 2.98% of genotypes are missing, the weighting genotypic method is equivalent to the anisoploid method, and the homoploid method is biased. We only use the anisoploid and the likelihood methods for this dataset, whose results are shown in Table 2. Moreover, the results of the corresponding F -statistics are shown in Table 3.

According to Tables 2 and 3, the two kinds of results obtained by using these two methods are generally similar. The variance components within individuals contribute to the majority in these two kinds of results and the F -statistics are generally small (below 0.08), implying a medium difference among populations ($F_{ST} \approx 0.06$). For the anisoploid method, the value of inbreeding coefficients is small ($F_{IS} = 0.0119$), but it is significantly greater than zero, while it is exactly equal to zero for the likelihood method.

Discussion

RMSE

In this paper, we generalize the framework of AMOVA and propose four methods to solve the variance components and the F -statistics.

It can be found from the comparison of Figures 2 and 4 that the RMSE in diploids is smaller than that in tetraploids, implying that the estimations of variance components and F -statistics are more accurate for diploids, although there are some biases for the polyploid phenotypic data.

We also see from Figures 2 and 4 that for the three moment methods, the estimated F -statistic $\hat{F}_{M-1,M}$ becomes increasingly inaccurate as M increases. However, for the likelihood method, as M increases, the accuracy of $\hat{F}_{M-1,M}$ is not only unaffected but also the same as that of $\hat{F}_{i,i+1}$ ($i \leq M-2$). Therefore, the likelihood method can be used in the datasets with higher value of M .

Biasedness

For the homoploid method, the estimated F -statistic $F_{i,i+1}$ is biased for the genotypic dataset with missing data. This bias is caused by the weighting for missing data. The allele frequency of the missing genotypes refers to the allele frequency in the corresponding population, and we assume that $F_{IS} = 0$. Therefore, $F_{1,2}$ is underestimated in Figure 1. For the anisoploid and the weighting genotypic methods, because the missing genotype data are ignored, such a bias is avoided.

For the phenotypic data, all three moment methods become biased. There are two sources of these biases: (i) the extraction of dummy haplotypes; (ii) the estimation of allele frequencies.

The extraction of dummy haplotypes breaks the correlation between alleles within the same individual, which can bias the estimation of SS_{WI} . Therefore, the bias of $F_{1,2}$ in Figure 3 is largest. This bias can be reduced by increasing the sample size N (e.g., the bias is -0.1 at N increasing to 50, see Figure 3). For the weighting genotypic method, this bias can be eliminated.

For the genotypic data, the allele frequencies are estimated by counting the alleles within the corresponding genotypes, so this estimation is unbiased. For the phenotypic data, the allele frequencies are estimated by using an expectation-maximization algorithm modified from KALINOWSKI AND TAPER (2006). Because this algorithm is also a kind of maximum-likelihood method, such estimation is biased, and the bias is passed to the subsequent steps. However, it can be reduced to a negligible level if N is large enough (e.g., the bias is 0.003 at $N = 50$, see Figure 3).

Unbiasedness

Due to the unbiasedness of moment methods, some negative estimates of variance components and F -statistics may present when the level of differentiation is low or the sample size is small. We select three datasets to illustrate this phenomenon, where each dataset consists of two populations with identical diploids which are genotyped at only one biallelic locus. Specifically, in Dataset 1, each population contains four genotypes (1

AA, 1 BB and 2 AB), which are drawn from HWE; in Dataset 2, each population is heterozygote-deficient, only containing two homozygotes (1 AA and 1 BB); in Dataset 3, each population is heterozygote-excessive, not containing homozygotes (2 AB). Because there is only one locus and no missing data, the three moment methods are equivalent. We use the homoploid method as an example, whose results with 9999 permutations are shown in Table 4. The results by using the likelihood method are also shown in this table as a comparison.

It can be found from Table 4 that for the moment methods, some estimates of variance components and F -statistics are negative or greatly deviate from the true values. For the likelihood method, such negative values can be avoided, and the values of estimates can be ensured to lie in the range of biological meaning.

Empirical results

There are some differences in the results of AMOVA on PEMBERTON *et al.* (2013) dataset between the moment and the likelihood methods (see Tables 2 and 3). For example, the value of F_{IS} is significantly positive for the moment method, but it is exactly equal to zero for the likelihood method.

The differences come from the dissimilarity between the schemes of these two kinds of methods. For the moment method, the SS within each hierarchy and at each locus is calculated, and the occurrence of some rare genotypes/phenotypes can only slightly change the values of SS. Therefore, for the loci with a similar polymorphism, their influences on the values of SS, MS, Var etc. are also similar.

In contrast, for the likelihood method, the SS is more sensitive to the distribution of genotypes/phenotypes, and the occurrence of some rare genotypes/phenotypes (e.g., homozygotes of rare alleles) can greatly affect the values of SS. The influence of a single rare genotype/phenotype may be equal to those of thousands of common genotypes/phenotypes. For the loci with a similar polymorphism, their influences on the values of SS, MS, Var etc. may be dramatic.

Applications

The calculating speed of the homoploid method is fastest during the permutation test. For this method, the genetic distance matrix only needs to be calculated one time, and it is permuted in a very fast way during the permutation test. More specifically, a permutation $k_1 k_2 \cdots k_H$ of the number codes $1, 2, \cdots, H$ is randomly generated, where H is the order of the genetic distance matrix (i.e., the number of alleles). Let d'_{ij} be equal to $d_{k_i k_j}$, where d'_{ij} is the ij^{th} element in the permuted genetic distance matrix, and $d_{k_i k_j}$ is the $k_i k_j^{\text{th}}$ element in the original distance matrix, $i, j = 1, 2, \cdots, H$. This technique can largely reduce the time expense, especially for a large dataset. For the other methods, the genetic distance should be calculated at each locus and in each iteration, so the calculating speeds of these methods are far slower than that of homoploid method. The drawback of the homoploid method is that the genetic distances are biased for the genotypic dataset with missing data or for the polyploid phenotypic data. Therefore, the homoploid method is suitable for a high-quality genotypic data (with a high genotyping rate) or a large dataset (e.g., next-generation sequencing data).

Although the calculating speed of the anisoploid method is slower than that of the homoploid method, it is still faster than those of the other two methods. That is because the whole dataset needs to be regenerated during the permutation test in the other two methods. For the anisoploid method, because the missing data are ignored during the calculation, the genetic distances are unbiased for genotypic data with a low genotyping rate. Therefore, this method is suitable for a low-quality genotypic data.

For the weighting genotypic method, because no dummy haplotypes are extracted, the genetic distances are less biased for the phenotypic data. In this method, instead of the use of permutation test, it randomly generates the dataset under the hypothesis that there is no differentiation. After that, it also calculates the probability that the variance components or the F -statistics at each locus and in each iteration are greater than the observed values. Therefore, this method is suitable for the polyploid phenotypic data.

For the three moment methods, there are two problems: (i) the RMSE of each $\hat{F}_{M-1,M}$ increases as the hierarchy number M increases, and (ii) some negative variance components or some negative estimates of F -statistics may present when the difference due to the unbiasedness is small. For the likelihood method, these two problems can be avoided, and the RMSEs of the estimated F -statistics are insensitive to M . In addition, various values of estimation are always in the biologically meaningful range. Therefore, this method is suitable for a larger M and/or for datasets for which a part of the results obtained by using these moment methods cannot be explained.

Acknowledgment

KH would like to thank Prof. Kermit Ritland for providing the space of visiting scholar in the University of British Columbia. This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020302), the National Natural Science Foundation of China (31770411, 31730104, 31572278 and 31770425), the Young Elite Scientists Sponsorship Program by CAST (2017QNRC001), and the National Key Programme of Research and Development, Ministry of Science and Technology (2016YFC0503200). DWD is supported by a Shaanxi Province Talents 100 Fellowship.

References

- Avni, R., M. Nave, O. Barad, K. Baruch, S. O. Twardziok *et al.*, 2017 Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357: 93-97.
- Burow, M. D., C. E. Simpson, J. L. Starr and A. H. Paterson, 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: 823-837.
- Chambers, J. L., and D. Garant, 2010 Determinants of population genetic structure in eastern chipmunks (*Tamias striatus*): the role of landscape barriers and sex-biased dispersal. *Journal of Heredity* 101: 413-422.
- Clark, L. V., and M. Jasieniuk, 2011 POLYSAT: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources* 11: 562-566.
- Cockerham, C. C., 1969 Variance of gene frequencies. *Evolution* 23: 72-84.
- Cockerham, C. C., 1973 Analyses of gene frequencies. *Genetics* 74: 679-700.
- Excoffier, L., and H. E. Lischer, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564-567.
- Excoffier, L., P. E. Smouse and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479-491.

533 Geffen, E. L. I., M. J. Anderson and R. K. Wayne, 2004 Climate and habitat barriers to
534 dispersal in the highly mobile grey wolf. *Molecular Ecology* 13: 2481-2490.

535 Haldane, J. B., 1930 Theoretical genetics of autopolyploids. *Journal of Genetics* 22: 359-372.

536 Hardy, O. J., and X. Vekemans, 2002 SPAGeDi: a versatile computer program to analyse
537 spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*
538 2: 618-620.

539 Huang, K., K. Ritland, S. Guo, M. Shattuck and B. Li, 2014 A pairwise relatedness estimator
540 for polyploids. *Molecular Ecology Resources* 14: 734-744.

541 Huang, K., T. C. Wang, D. W. Dunn, P. Zhang, R. C. Liu *et al.*, 2019 Genotypic frequencies at
542 equilibrium for polysomic inheritance under double-reduction. *G3: Genes, Genomes,*
543 *Genetics*: doi: 10.1534/g1533.1119.400132.

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

546 Lait, L. A., and T. M. Burg, 2013 When east meets west: population structure of a high-latitude
547 resident species, the boreal chickadee (*Poecile hudsonicus*). *Heredity* 111: 321-329.

548 Lande, R., 1976 Natural selection and random genetic drift in phenotypic evolution. *Evolution*:
549 314-334.

550 Ling, H. Q., B. Ma, X. L. Shi, H. Liu, L. L. Dong *et al.*, 2018 Genome sequence of the progenitor
551 of wheat A subgenome *Triticum urartu*. *Nature* 557: 424.

552 Luppi, T. A., E. D. Spivak and C. C. Bas, 2003 The effects of temperature and salinity on larval
553 development of *Armases rubripes* Rathbun, 1897 (Brachyura, Grapsoidea, Sesarmidae),
554 and the southern limit of its geographical distribution. *Estuarine, Coastal and Shelf*
555 *Science* 58: 575-585.

556 Martin, N. H., and J. H. Willis, 2007 Ecological divergence associated with mating system
557 causes nearly complete reproductive isolation between sympatric *Mimulus* species.
558 *Evolution* 61: 68-82.

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

561 Meirmans, P. G., and S. Liu, 2018 Analysis of Molecular Variance (AMOVA) for
562 autopolyploids. *Frontiers in Ecology and Evolution* 6: 66.

563 Meirmans, P. G., and P. H. V. Tienderen, 2004 GENOTYPE and GENODIVE : two programs
564 for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4:
565 792-794.

566 Nelder, J. A., and R. Mead, 1965 A simplex method for function minimization. *The computer*
567 *journal* 7: 308-313.

568 Otto, S. P., 2007 The evolutionary consequences of polyploidy. *Cell* 131: 452-462.

569 Peakall, R., and P. E. Smouse, 2006 GENALEX 6: genetic analysis in Excel. Population genetic
570 software for teaching and research. *Molecular Ecology Notes* 6: 288-295.

571 Peakall, R., P. E. Smouse and D. Huff, 1995 Evolutionary implications of allozyme and RAPD
572 variation in diploid populations of dioecious buffalograss *Buchloe dactyloides*.
573 *Molecular Ecology* 4: 135-148.

574 Pemberton, T. J., M. DeGiorgio and N. A. Rosenberg, 2013 Population structure in a
575 comprehensive genomic data set on human microsatellite variation. *G3: Genes, Genomes,*

Genetics 3: 891-907.

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

Russell, E. M., Y. Yom-Tov and E. Geffen, 2004 Extended parental care and delayed dispersal: northern, tropical, and southern passerines compared. *Behavioral Ecology* 15: 831-838.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-462.

Wright, S., 1943 Isolation by distance. *Genetics* 28: 114.

Yang, J. Y., S. A. Cushman, J. Yang, M. B. Yang and T. J. Bao, 2014 Effects of climatic gradients on genetic differentiation of *Caragana* on the Ordos Plateau, China. *Landscape Ecology* 28: 1729-1741.

Author contributions

KH and BGL designed the project, KH and YLL constructed the model and wrote the draft, KH designed the software, PZ performed the simulations and analyses, and DWD checked the model and helped to write the manuscript.

Tables

Table 1. The layout of AMOVA. The total SS is decomposed into the SS in different sources of variation. Each expected MS is expressed here as a function of variance components.

Source of variation	d.f.	SS	MS	Expected MS
Within individual	N	SS_{WI}	$\frac{SS_{WI}}{G - 1}$	σ_{WI}^2
Among individuals within populations	$N - P$	$SS_{WP} - SS_{WI}$	$\frac{SS_{WP} - SS_{WI}}{G - 1}$	$2\sigma_{AI/WP}^2 + \sigma_{WI}^2$
Among populations within groups	$P - G$	$SS_{WG} - SS_{WP}$	$\frac{SS_{WG} - SS_{WP}}{G - 1}$	$n\sigma_{AP/WG}^2 + 2\sigma_{AI/WP}^2 + \sigma_{WI}^2$

Among groups	$G - 1$	$SS_{TOT} - SS_{WG}$	$\frac{SS_{TOT} - SS_{WG}}{G - 1}$	$n''\sigma_{AG}^2 + n'\sigma_{AP/WG}^2 + 2\sigma_{AI/WP}^2 + \sigma_{WI}^2$
--------------	---------	----------------------	------------------------------------	---

597 Here, N , P or G denotes the number of individuals, populations or groups,
 598 respectively, and SS_{WI} , SS_{WP} , SS_{WG} or SS_{TOT} denotes the SS within individuals, within
 599 populations, within groups or in the total population, which can be obtained by Equation
 600 (1). The coefficients n, n' and n'' are, respectively, calculated by

$$601 \quad n = \frac{2N-2 \sum_g \sum_{\rho \in g} N_{\rho}^2 / N_g}{P-G}, \quad n' = \frac{2 \sum_g \sum_{\rho \in g} N_{\rho}^2 / N_g - 2 \sum_{\rho} N_{\rho}^2 / N}{G-1} \quad \text{and} \quad n'' = \frac{2N-2 \sum_g N_g^2 / N}{G-1},$$

602 in which N_{ρ} or N_g is the number of individuals in the population ρ or in the group g ,
 603 respectively, and the mobile subscript in $\sum_g$ (or in $\sum_{\rho}$) is taken from all groups (or all
 604 populations) in the total population.

605

606

607 Table 2. The degrees of freedom, sum of squares (SS), mean squares (MS), estimated
608 variance components (Var) and the variance percentage of (microsatellite, SLATKIN 1995)
609 dataset

Method	Source	d.f.	SS	MS	Var	%	P	Perm. Mean	Perm. Var
Anisoploid	WI	3626266	1293746.5	0.357	0.357	92.68	1.000	0.381	0.000
	AI/WP	3454743	1262341.6	0.365	0.004	1.12	0.000	0.000	0.000
	AP/WC1	133491	87578.6	0.656	0.007	1.82	0.000	0.000	0.000
	AC1/WC2	32227	36005.9	1.117	0.003	0.77	0.000	0.000	0.000
	AC2	5160	86911.1	16.843	0.014	3.62	0.000	0.000	0.000
	Total	7251887	2766583.6	0.381	0.385	100.00	-	-	-
Likelihood	WI	3626266	1310274.0	0.361	0.361	94.00	1.000	0.381	0.000
	AI/WP	3454743	1248297.8	0.361	0.000	0.00	0.624	0.000	0.000
	AP/WC1	133491	88346.1	0.662	0.007	1.88	0.000	0.000	0.000
	AC1/WC2	32227	45688.1	1.418	0.005	1.26	0.000	0.000	0.000
	AC2	5160	73977.7	14.337	0.011	2.86	0.000	0.000	0.000
	Total	7251887	2766583.6	0.381	0.384	100.00	-	-	-

610 *WI: within individuals; AI/WP: among individuals within populations; AP/WC1:
611 among populations within groups I; AC1/WC2: among groups I within groups II;
612 AC2: among groups II.

613

614

615 Table 3. The value, significance, permuted mean and permuted variance of F -statistics

F -statistics	Anisoploid				likelihood			
	Value	P	Mean $\times 10^{-6}$	Var $\times 10^{-8}$	Value	P	Mean $\times 10^{-4}$	Var $\times 10^{-6}$
F_{IS}	0.0119	<0.0001	0.0425	7.3783	0.0000	0.1649	0.0000	0.0000
F_{IC1}	0.0307	<0.0001	1.1174	7.2052	0.0196	<0.0001	1.7837	0.3049
F_{IC2}	0.0384	<0.0001	1.2765	7.1643	0.0323	<0.0001	3.9708	0.5911
F_{IT}	0.0732	<0.0001	1.3279	7.1507	0.0600	<0.0001	6.1550	1.0658
F_{SC1}	0.0190	<0.0001	1.0729	0.2298	0.0196	<0.0001	1.7837	0.3049
F_{SC2}	0.0268	<0.0001	1.2320	0.1936	0.0323	<0.0001	3.9708	0.5911
F_{ST}	0.0620	<0.0001	1.2835	0.1751	0.0600	<0.0001	6.1550	1.0658
F_{C1C2}	0.0080	<0.0001	0.1584	0.0812	0.0130	<0.0001	2.1884	0.0948
F_{C1T}	0.0438	<0.0001	0.2100	0.0516	0.0412	<0.0001	4.3740	0.3788
F_{C2T}	0.0362	<0.0001	0.0513	0.0156	0.0286	<0.0001	2.1870	0.0947

616

Table 4. Results of AMOVA by using the moment and likelihood methods

Estimator	Dataset	Source	d.f.	SS	MS	Var	%	Perm. Mean	Perm. Var	<i>P</i>	<i>F</i> -statistics	Value	Perm. Mean	Perm. Var	<i>P</i>
MOM	1	WI	8	2.000	0.250	0.250	100.0	0.267	0.008	0.302	F_{IS}	0.143	-0.023	0.143	0.169
		AI/WP	6	2.000	0.333	0.042	16.7	-0.001	0.010	0.168	F_{IT}	0.000	-0.016	0.140	0.507
		AP	1	0.000	0.000	-0.042	-16.7	0.000	0.003	0.763	F_{ST}	-0.167	-0.012	0.028	0.763
		Total	15	4.000	0.267	0.250	100.0	-	-	-					
	2	WI	4	0.000	0.000	0.000	0.0	0.287	0.019	0.913	F_{IS}	1.000	-0.074	0.341	0.000
		AI/WP	2	2.000	1.000	0.500	200.0	-0.003	0.032	0.000	F_{IT}	1.000	-0.058	0.337	0.000
		AP	1	0.000	0.000	-0.250	-100.0	0.001	0.016	0.943	F_{ST}	-1.000	-0.048	0.158	0.943
		Total	7	2.000	0.286	0.250	100.0	-	-	-					
	3	WI	4	2.000	0.500	0.500	200.0	0.286	0.019	0.000	F_{IS}	-1.000	-0.067	0.345	0.766
		AI/WP	2	0.000	0.000	-0.250	-100.0	0.001	0.033	0.766	F_{IT}	-1.000	-0.054	0.339	0.766
		AP	1	0.000	0.000	0.000	0.0	-0.001	0.016	0.476	F_{ST}	0.000	-0.057	0.162	0.476
		Total	7	2.000	0.286	0.250	100.0	-	-	-					
ML	1	WI	8	2.133	0.267	0.267	100.0	0.239	0.002	0.644	F_{IS}	0.000	0.091	0.032	0.304
		AI/WP	6	1.600	0.267	0.000	0.0	0.024	0.002	0.304	F_{IT}	0.000	0.116	0.038	0.358
		AP	1	0.267	0.267	0.000	0.0	0.009	0.001	0.511	F_{ST}	0.000	0.031	0.005	0.376
		Total	15	4.000	0.267	0.267	100.0	-	-	-					
	2	WI	4	0.000	0.000	0.000	0.0	0.247	0.006	0.949	F_{IS}	1.000	0.113	0.068	0.000
		AI/WP	2	1.333	0.667	0.333	100.0	0.033	0.007	0.000	F_{IT}	1.000	0.151	0.078	0.000
		AP	1	0.667	0.667	0.000	0.0	0.018	0.003	0.569	F_{ST}	0.000	0.049	0.019	0.321
		Total	7	2.000	0.286	0.333	100.0	-	-	-					
	3	WI	4	1.143	0.286	0.286	100.0	0.247	0.006	0.000	F_{IS}	0.000	0.113	0.068	0.256
		AI/WP	2	0.571	0.286	0.000	0.0	0.033	0.007	0.256	F_{IT}	0.000	0.151	0.078	0.444

	AP	1	0.286	0.286	0.000	0.0	0.018	0.003	0.569	F_{ST}	0.000	0.049	0.019	0.321
	Total	7	2.000	0.286	0.286	100.0	-	-	-					
618	* Dataset 1: pop1 = pop2 = {AA, BB, AB, AB}; Dataset 2: pop1 = pop2 = {AA, BB}; Dataset 3: pop1 = pop2 = {AB, AB}.													

619

Figure legends

Figure 1. The bias of estimated F -statistic $\hat{F}_{i,i+1}$ for diploids as a function of sample size N sampled from each population under different conditions ($1 \leq i < M$). For each of the four methods listed at the top, the results are shown in the column where this method is located. Each row shows the results of a structure, where the structure '2xNx4x2' means that $M = 4$, and there are two groups, each group containing four populations and each population consisting of N diploids. The meanings of other two structures can be analogized. Each solid, dashed, dash-dotted or dotted line denotes the bias of $\hat{F}_{1,2}$, $\hat{F}_{2,3}$, $\hat{F}_{3,4}$ or $\hat{F}_{4,5}$, respectively, corresponding to the value of N .

Figure 2. The RMSE of estimated F -statistic $\hat{F}_{i,i+1}$ for diploids and under different conditions ($1 \leq i < M$). The meanings of columns and rows are as indicated in Figure 1. Each solid, dashed, dash-dotted or dotted line denotes the RMSE of $\hat{F}_{1,2}$, $\hat{F}_{2,3}$, $\hat{F}_{3,4}$ or $\hat{F}_{4,5}$, respectively, corresponding to the value of N .

Figure 3. The bias of estimated F -statistic $\hat{F}_{i,i+1}$ for tetraploids and under different conditions. The meanings of columns, rows and lines are as indicated in Figure 1.

Figure 4. The RMSE of estimated F -statistic $\hat{F}_{i,i+1}$ for tetraploids and under different conditions. The meanings of columns, rows and lines are as indicated in Figure 2.

Figures

Figure 1

Figure 2

Figure 3

Figure 4

Supplementary materials

Linear model

Let A be an allele randomly taken from the total population, and let p be the mean frequency of A in the total population. We will focus on the biases of the frequency p related with an allele a to carry out our discussion. Our linear model is developed from COCKERHAM (1969; 1973), which is described by the following function:

$$y_{g\rho ia} = p + d_g + d_{g\rho} + d_{g\rho i} + d_{g\rho ia},$$

where a is arbitrary, and the relations among g, ρ, i, a are nested, that is, $a \in i \subseteq \rho \subseteq g$; d_g is the bias of the frequency p in the group g relative to the total population, $d_{g\rho}$ is the bias of p in the population ρ relative to the group g , $d_{g\rho i}$ is the bias of p in the individual i relative to the population ρ , and $d_{g\rho ia}$ is the bias of p in the allele a relative to the individual i . It is worth pointing out that because of the nested relation, g and ρ are uniquely determined so long as i is given. We stipulate that the condition $E(y_{g\rho ia}) = p$ should be satisfied in this model.

Because $E(y_{g\rho ia}) = p$ and the allele frequencies obey a binomial distribution, we have $\text{var}(y_{g\rho ia}) = p(1 - p)$, that is, $\sigma_{\text{TOT}}^2 = p(1 - p)$.

According to COCKERHAM (1969; 1973), the formulas for $\text{cov}(y_{g\rho ia}, y_{g'\rho'i'a'})$ under various situations are

$$\text{cov}(y_{g\rho ia}, y_{g'\rho' i' a'}) = \begin{cases} \text{var}(y_{g\rho ia}) = p(1-p) & \text{if } g = g', \rho = \rho', i = i', a = a'; \\ \text{cov}_{AA/WI} = F_{IT}p(1-p) & \text{if } g = g', \rho = \rho', i = i', a \neq a'; \\ \text{cov}_{AI/WP} = F_{ST}p(1-p) & \text{if } g = g', \rho = \rho', i \neq i'; \\ \text{cov}_{AP/WG} = F_{CT}p(1-p) & \text{if } g = g', \rho \neq \rho'; \\ \text{cov}_{AG} = 0 & \text{if } g \neq g'. \end{cases}$$

664 For the final situation, because the alleles among groups are assumed to be independent,
 665 the value of corresponding F -statistic is zero, and so $\text{cov}_{AG} = 0p(1-p) = 0$. Moreover,
 666 the formulas for $E(y_{g\rho ia}y_{g'\rho' i' a'})$ under various situations are

$$E(y_{g\rho ia}y_{g'\rho' i' a'}) = \begin{cases} p(1-p) + p^2 & \text{if } g = g', \rho = \rho', i = i', a = a'; \\ F_{IT}p(1-p) + p^2 & \text{if } g = g', \rho = \rho', i = i', a \neq a'; \\ F_{ST}p(1-p) + p^2 & \text{if } g = g', \rho = \rho', i \neq i'; \\ F_{CT}p(1-p) + p^2 & \text{if } g = g', \rho \neq \rho'; \\ 0 + p^2 & \text{if } g \neq g'. \end{cases}$$

667 In fact, for the first situation, $E(y_{g\rho ia}y_{g'\rho' i' a'})$ becomes $E(y_{g\rho ia}^2)$, so

$$E(y_{g\rho ia}y_{g'\rho' i' a'}) = \text{var}(y_{g\rho ia}) + [E(y_{g\rho ia})]^2 = p(1-p) + p^2.$$

668 For the second situation, $E(y_{g\rho ia}y_{g'\rho' i' a'})$ becomes $E(y_{g\rho ia}y_{g\rho ia'})$. Because F_{IT} is the
 669 probability $\Pr(y_{g\rho ia} \equiv y_{g\rho ia'})$ in the sense that two distinct alleles within a same
 670 individual in the total population are IBD, we obtain

$$\begin{aligned} E(y_{g\rho ia}y_{g'\rho' i' a'}) &= \Pr(y_{g\rho ia}y_{g\rho ia'} = 1) \\ &= \Pr(y_{g\rho ia} = 1, y_{g\rho ia'} = 1) \\ &= \Pr(y_{g\rho ia} \equiv y_{g\rho ia'})\Pr(y_{g\rho ia} = 1) \\ &\quad + [1 - \Pr(y_{g\rho ia} \equiv y_{g\rho ia'})]\Pr(y_{g\rho ia} = 1)\Pr(y_{g\rho ia'} = 1) \\ &= F_{IT}p + (1 - F_{IT})p^2 \\ &= F_{IT}p(1-p) + p^2. \end{aligned}$$

671 For the remaining situations, the derivations are similar, and omitted.

There are the following relations between the F -statistics and the variance components:

$$\begin{aligned} F_{IT} &= 1 - \frac{\sigma_{WI}^2}{\sigma_{TOT}^2}, \\ F_{ST} &= 1 - \frac{\sigma_{WI}^2 + \sigma_{AI/WP}^2}{\sigma_{TOT}^2}, \\ F_{CT} &= 1 - \frac{\sigma_{WI}^2 + \sigma_{AI/WP}^2 + \sigma_{AP/WG}^2}{\sigma_{TOT}^2}, \\ \sigma_{TOT}^2 &= \sigma_{AI/WP}^2 + \sigma_{AP/WG}^2 + \sigma_{AG}^2. \end{aligned}$$

Because $\sigma_{TOT}^2 = p(1 - p)$, we obtain

$$\begin{aligned} \sigma_{WI}^2 &= (1 - F_{IT})p(1 - p), \\ \sigma_{AI/WP}^2 &= (F_{IT} - F_{ST})p(1 - p), \\ \sigma_{AP/WG}^2 &= (F_{ST} - F_{CT})p(1 - p), \\ \sigma_{AG}^2 &= F_{CT}p(1 - p). \end{aligned} \tag{8}$$

We will use the symbol $\bar{y}_{g\rho i}$ ($\bar{y}_{g\rho}$, $\bar{y}_g$ or $\bar{y}_t$) to denote the average of values of the function $y_{g\rho ia}$ when a is taken from all alleles within the individual i (the population ρ , the group g or the total population). Then,

$$\begin{aligned} \bar{y}_{g\rho i} &= \frac{1}{v_i} \sum_{a \in i} y_{g\rho ia}, \quad \bar{y}_{g\rho} = \frac{1}{v_\rho} \sum_{a \in \rho} y_{g\rho ia} = \frac{1}{v_\rho} \sum_{i \in \rho} \sum_{a \in i} y_{g\rho ia}, \\ \bar{y}_g &= \frac{1}{v_g} \sum_{a \in g} y_{g\rho ia} = \frac{1}{v_g} \sum_{\rho \in g} \sum_{a \in \rho} y_{g\rho ia} = \frac{1}{v_g} \sum_{i \in g} \sum_{a \in i} y_{g\rho ia} = \frac{1}{v_g} \sum_{\rho \in g} \sum_{i \in \rho} \sum_{a \in i} y_{g\rho ia}, \\ \bar{y}_t &= \frac{1}{v_t} \sum_a y_{g\rho ia} = \frac{1}{v_t} \sum_g \sum_{\rho \in g} \sum_{i \in \rho} \sum_{a \in i} y_{g\rho ia}, \end{aligned}$$

where v_i , v_g , v_ρ or v_t is the number of alleles within the individual i , the population ρ , the group g or the total population, respectively.

Derivation for the formula of expected SS_{WI}

The expectations of $\bar{y}_{g\rho i}^2$ and $y_{g\rho ia}\bar{y}_{g\rho i}$ are calculated as follows:

$$\begin{aligned}
 E(\bar{y}_{g\rho i}^2) &= \text{var}(\bar{y}_{g\rho i}) + [E(\bar{y}_{g\rho i})]^2 = \text{var}\left(\frac{1}{v_i} \sum_{a \in i} y_{g\rho ia}\right) + p^2 \\
 &= \frac{1}{v_i^2} \left[\sum_{a \in i} \text{var}(y_{g\rho ia}) + \sum_{\substack{a, a' \in i \\ a \neq a'}} \text{cov}(y_{g\rho ia}, y_{g\rho ia'}) \right] + p^2 \\
 &= \frac{1}{v_i^2} [v_i p(1-p) + (v_i^2 - v_i) F_{IT} p(1-p)] + p^2 \\
 &= \frac{1}{v_i} p(1-p) + \frac{v_i - 1}{v_i} F_{IT} p(1-p) + p^2, \\
 E(y_{g\rho ia} \bar{y}_{g\rho i}) &= E\left(\frac{1}{v_i} \sum_{a' \in i} y_{g\rho ia} y_{g\rho ia'}\right) \\
 &= \frac{1}{v_i} \left[E(y_{g\rho ia}^2) + \sum_{\substack{a' \in i \\ a' \neq a}} E(y_{g\rho ia} y_{g\rho ia'}) \right] \\
 &= \frac{1}{v_i} [p(1-p) + p^2] + \frac{v_i - 1}{v_i} [F_{IT} p(1-p) + p^2] \\
 &= \frac{1}{v_i} p(1-p) + \frac{v_i - 1}{v_i} F_{IT} p(1-p) + p^2.
 \end{aligned}$$

682 Comparing the two calculated results, we have $E(\bar{y}_{g\rho i}^2) = E(y_{g\rho ia} \bar{y}_{g\rho i})$. In addition,

$$\sigma_{\text{WI}}^2 = (1 - F_{IT})p(1-p).$$

683 Using these facts, we can easily obtain the next result:

$$E[(y_{g\rho ia} - \bar{y}_{g\rho i})^2] = \left(1 - \frac{1}{v_i}\right) \sigma_{\text{WI}}^2.$$

684 Moreover, we have

$$\sum_a \left(1 - \frac{1}{v_i}\right) = H - \sum_i \sum_{a \in i} \frac{1}{v_i} = H - \sum_i \frac{v_i}{v_i} = H - N,$$

685 where H and N are the numbers of alleles and individuals in the total population,

686 respectively. Now, we can easily derive the first formula of expected SS in Equation (2):

$$\begin{aligned}
 E(\text{SS}_{\text{WI}}) &= E\left[\sum_a (y_{g\rho ia} - \bar{y}_{g\rho i})^2\right] = \sum_a E[(y_{g\rho ia} - \bar{y}_{g\rho i})^2] \\
 &= \sum_a \left(1 - \frac{1}{v_i}\right) \sigma_{\text{WI}}^2 \\
 &= \sigma_{\text{WI}}^2 (H - N).
 \end{aligned}$$

687 Derivation for the formula of expected SS_{WP}

688 The expectations of $\bar{y}_{g\rho}^2$ and $y_{g\rho ia} \bar{y}_{g\rho}$ are calculated as follows:

$$\begin{aligned}
 E(\bar{y}_{g\rho}^2) &= \text{var}(\bar{y}_{g\rho}) + [E(\bar{y}_{g\rho})]^2 = \text{var}\left(\frac{1}{v_\rho} \sum_{a \in \rho} y_{g\rho ia}\right) + p^2 \\
 &= \frac{1}{v_\rho^2} \left[\sum_{a \in \rho} \text{var}(y_{g\rho ia}) + \sum_{i \in \rho} \sum_{\substack{a, a' \in i \\ a \neq a'}} \text{cov}(y_{g\rho ia}, y_{g\rho ia'}) \right. \\
 &\quad \left. + \sum_{\substack{i, i' \in \rho \\ i \neq i'}} \sum_{\substack{a \in i \\ a' \in i'}} \text{cov}(y_{g\rho ia}, y_{g\rho i' a'}) \right] + p^2 \\
 &= \frac{1}{v_\rho} p(1-p) + \frac{\sum_{i' \in \rho} v_{i'}^2 - v_\rho}{v_\rho^2} F_{IT} p(1-p) + \frac{v_\rho^2 - \sum_{i' \in \rho} v_{i'}^2}{v_\rho^2} F_{ST} p(1-p) + p^2, \\
 E(y_{g\rho ia} \bar{y}_{g\rho}) &= E\left(\frac{1}{v_\rho} \sum_{i' \in \rho} \sum_{a' \in i'} y_{g\rho ia} y_{g\rho i' a'}\right) \\
 &= \frac{1}{v_\rho} \left[E(y_{g\rho ia}^2) + \sum_{\substack{a' \in i \\ a' \neq a}} E(y_{g\rho ia} y_{g\rho ia'}) + \sum_{\substack{i' \in \rho \\ i' \neq i}} \sum_{a' \in i'} E(y_{g\rho ia} y_{g\rho i' a'}) \right] \\
 &= \frac{1}{v_\rho} p(1-p) + \frac{v_i - 1}{v_\rho} F_{IT} p(1-p) + \frac{v_\rho - v_i}{v_\rho} F_{ST} p(1-p) + p^2.
 \end{aligned}$$

689 Now, by using Equation (8), it is not difficult to calculate that

$$E[(y_{g\rho ia} - \bar{y}_{g\rho})^2] = \sigma_{WI}^2 \left(1 - \frac{1}{v_\rho}\right) + \sigma_{AI/WP}^2 \left(1 + \frac{\sum_{i' \in \rho} v_{i'}^2}{v_\rho^2} - \frac{2v_i}{v_\rho}\right).$$

690 Moreover, because P is the number of populations in the total population, we have

$$\begin{aligned}
 \sum_a \left(1 - \frac{1}{v_\rho}\right) &= H - \sum_\rho \sum_{a \in \rho} \frac{1}{v_\rho} = H - \sum_\rho 1 = H - P, \\
 \sum_a \left(1 + \frac{\sum_{i' \in \rho} v_{i'}^2}{v_\rho^2} - \frac{2v_i}{v_\rho}\right) &= H + \sum_\rho \frac{v_\rho \sum_{i' \in \rho} v_{i'}^2}{v_\rho^2} - \sum_\rho \sum_{i \in \rho} \frac{2v_i^2}{v_\rho} = H - \sum_\rho \sum_{i \in \rho} \frac{v_i^2}{v_\rho},
 \end{aligned}$$

691 Now, the second formula of expected SS in Equation (2) can be derived as follows:

$$\begin{aligned}
 E(SS_{WP}) &= E\left[\sum_a (y_{g\rho ia} - \bar{y}_{g\rho})^2\right] = \sum_{a \in i} E[(y_{g\rho ia} - \bar{y}_{g\rho})^2] \\
 &= \sum_{a \in i} \left[\sigma_{WI}^2 \left(1 - \frac{1}{v_\rho}\right) + \sigma_{AI/WP}^2 \left(1 + \frac{\sum_{i' \in \rho} v_{i'}^2}{v_\rho^2} - \frac{2v_i}{v_\rho}\right) \right] \\
 &= \sigma_{WI}^2 (H - P) + \sigma_{AI/WP}^2 \left(H - \sum_\rho \sum_{i \in \rho} \frac{v_i^2}{v_\rho} \right).
 \end{aligned}$$

692 Derivation for the formula of expected SS_{WG}

693 The expectations of $\bar{y}_g^2$ and $y_{g\rho ia}\bar{y}_g$ are calculated as follows:

$$\begin{aligned} E(\bar{y}_g^2) &= \text{var}(\bar{y}_g) + [E(\bar{y}_g)]^2 = \text{var}\left(\frac{1}{v_g} \sum_{a \in g} y_{g\rho ia}\right) + p^2 \\ &= \frac{1}{v_g^2} \left[\sum_{a \in g} \text{var}(y_{g\rho ia}) + \sum_{i \in g} \sum_{\substack{a, a' \in i \\ a \neq a'}} \text{cov}(y_{g\rho ia}, y_{g\rho ia'}) \right. \\ &\quad \left. + \sum_{\rho \in g} \sum_{\substack{i, i' \in \rho \\ i \neq i'}} \sum_{\substack{a \in i \\ a' \in i'}} \text{cov}(y_{g\rho ia}, y_{g\rho i' a'}) + \sum_{\substack{\rho, \rho' \in g \\ \rho \neq \rho'}} \sum_{\substack{a \in \rho \\ a' \in \rho'}} \text{cov}(y_{g\rho ia}, y_{g\rho' i' a'}) \right] + p^2 \\ &= \frac{1}{v_g} p(1-p) + \frac{\sum_{i' \in g} v_{i'}^2 - v_g}{v_g^2} F_{IT} p(1-p) \\ &\quad + \frac{\sum_{\rho' \in g} v_{\rho'}^2 - \sum_{i' \in g} v_{i'}^2}{v_g^2} F_{ST} p(1-p) + \frac{v_g^2 - \sum_{\rho' \in g} v_{\rho'}^2}{v_g^2} F_{CT} p(1-p) + p^2, \\ E(y_{g\rho ia}\bar{y}_g) &= E\left(\frac{1}{v_g} \sum_{\rho' \in g} \sum_{i' \in \rho'} \sum_{a' \in i'} y_{g\rho ia} y_{g\rho' i' a'}\right) \\ &= \frac{1}{v_g} \left[E(y_{g\rho ia}^2) + \sum_{\substack{a' \in i \\ a' \neq a}} E(y_{g\rho ia} y_{g\rho ia'}) + \sum_{\substack{i' \in \rho \\ i' \neq i}} \sum_{a' \in i'} E(y_{g\rho ia} y_{g\rho i' a'}) \right. \\ &\quad \left. + \sum_{\substack{\rho' \in g \\ \rho' \neq \rho}} \sum_{a' \in \rho'} E(y_{g\rho ia} y_{g\rho' i' a'}) \right] \\ &= \frac{1}{v_g} p(1-p) + \frac{v_i - 1}{v_g} F_{IT} p(1-p) + \frac{v_\rho - v_i}{v_g} F_{ST} p(1-p) \\ &\quad + \frac{v_g - v_\rho}{v_g} F_{CT} p(1-p) + p^2. \end{aligned}$$

694 Now, by using Equation (8), it is easy to calculate that

$$\begin{aligned} E[(y_{g\rho ia} - \bar{y}_g)^2] &= \sigma_{WI}^2 \left(1 - \frac{1}{v_g}\right) + \sigma_{AI/WP}^2 \left(1 + \frac{\sum_{i' \in g} v_{i'}^2}{v_g^2} - \frac{2v_i}{v_g}\right) \\ &\quad + \sigma_{AP/WG}^2 \left(1 + \frac{\sum_{\rho' \in g} v_{\rho'}^2}{v_g^2} - \frac{2v_\rho}{v_g}\right). \end{aligned}$$

695 Moreover, because G is the number of all groups, if we let $x = i$ or $x = \rho$, then

$$\sum_a \left(1 - \frac{1}{v_g}\right) = H - \sum_g \sum_{a \in g} \frac{1}{v_g} = H - \sum_g 1 = H - G,$$

$$\sum_a \left(1 + \frac{\sum_{x' \in g} v_{x'}^2}{v_g^2} - \frac{2v_x}{v_g} \right) = H + \sum_g \frac{v_g \sum_{x' \in g} v_{x'}^2}{v_g^2} - 2 \sum_g \sum_{x \in g} \frac{v_x^2}{v_g} = H - \sum_g \sum_{x \in g} \frac{v_x^2}{v_g}.$$

696 Now, the third formula of expected SS in Equation (2) can be derived as follows:

$$\begin{aligned} E(SS_{WG}) &= E \left[\sum_a (y_{g\rho ia} - \bar{y}_g)^2 \right] = \sum_a E \left[(y_{g\rho ia} - \bar{y}_g)^2 \right] \\ &= \sum_a \left[\sigma_{WI}^2 \left(1 - \frac{1}{v_\rho} \right) + \sigma_{AI/WP}^2 \left(1 + \frac{\sum_{i' \in g} v_{i'}^2}{v_g^2} - \frac{2v_i}{v_g} \right) \right. \\ &\quad \left. + \sigma_{AP/WG}^2 \left(1 + \frac{\sum_{\rho' \in g} v_{\rho'}^2}{v_g^2} - \frac{2v_\rho}{v_g} \right) \right] \\ &= \sigma_{WI}^2(H - G) + \sigma_{AI/WP}^2 \left(H - \sum_g \sum_{i \in g} \frac{v_i^2}{v_g} \right) + \sigma_{AP/WG}^2 \left(H - \sum_g \sum_{\rho \in g} \frac{v_\rho^2}{v_g} \right). \end{aligned}$$

697 Derivation for the formula of expected SS_{TOT}

698 The expectations of $\bar{y}_t^2$ and $y_{g\rho ia}\bar{y}_t$ are calculated as follows:

$$\begin{aligned} E(\bar{y}_t^2) &= \text{var}(\bar{y}_t) + [E(\bar{y}_t)]^2 = \text{var} \left(\frac{1}{v_t} \sum_a y_{g\rho ia} \right) + p^2 \\ &= \frac{1}{v_t^2} \left[\sum_a \text{var}(y_{g\rho ia}) + \sum_i \sum_{\substack{a, a' \in i \\ a \neq a'}} \text{cov}(y_{g\rho ia}, y_{g\rho ia'}) \right. \\ &\quad + \sum_\rho \sum_{\substack{i, i' \in \rho \\ i \neq i'}} \sum_{\substack{a \in i \\ a' \in i'}} \text{cov}(y_{g\rho ia}, y_{g\rho i' a'}) + \sum_g \sum_{\substack{\rho, \rho' \in g \\ \rho \neq \rho'}} \sum_{\substack{a \in \rho \\ a' \in \rho'}} \text{cov}(y_{g\rho ia}, y_{g\rho' i' a'}) \\ &\quad \left. + \sum_{\substack{g, g' \\ g \neq g'}} \sum_{\substack{a \in g \\ a' \in g'}} \text{cov}(y_{g\rho ia}, y_{g' \rho' i' a'}) \right] + p^2 \\ &= \frac{1}{v_t} p(1 - p) + \frac{\sum_{i'} v_{i'}^2 - v_t}{v_t^2} F_{IT} p(1 - p) + \frac{\sum_{\rho'} v_{\rho'}^2 - \sum_{i'} v_{i'}^2}{v_t^2} F_{ST} p(1 - p) \\ &\quad + \frac{\sum_{g'} v_{g'}^2 - \sum_{\rho'} v_{\rho'}^2}{v_t^2} F_{CT} p(1 - p) + \frac{v_t^2 - \sum_{g'} v_{g'}^2}{v_t^2} \times 0 + p^2, \end{aligned}$$

$$\begin{aligned}
 E(y_{g\rho ia}\bar{y}_t) &= E\left(\frac{1}{v_t} \sum_{g'} \sum_{\rho' \in g} \sum_{i' \in \rho} \sum_{a' \in i'} y_{g\rho ia} y_{g'\rho' i' a'}\right) \\
 &= \frac{1}{v_t} \left[E(y_{g\rho ia}^2) + \sum_{\substack{a' \in i \\ a' \neq a}} E(y_{g\rho ia} y_{g\rho ia'}) + \sum_{\substack{i' \in \rho \\ i' \neq i}} \sum_{a' \in i'} E(y_{g\rho ia} y_{g\rho i' a'}) \right. \\
 &\quad \left. + \sum_{\substack{\rho' \in g \\ \rho' \neq \rho}} \sum_{a' \in \rho'} E(y_{g\rho ia} y_{g\rho' i' a'}) + \sum_{\substack{g' \neq g}} \sum_{a' \in g} E(y_{g\rho ia} y_{g' \rho' i' a'}) \right] \\
 &= \frac{1}{v_t} p(1-p) + \frac{v_i - 1}{v_t} F_{IT} p(1-p) + \frac{v_\rho - v_i}{v_t} F_{ST} p(1-p) \\
 &\quad + \frac{v_g - v_\rho}{v_t} F_{CT} p(1-p) + \frac{v_t - v_g}{v_t} \times 0 + p^2.
 \end{aligned}$$

699 Now, according to Equation (8), we can easily calculate that

$$\begin{aligned}
 E[(y_{g\rho ia} - \bar{y}_t)^2] &= \sigma_{WI}^2 \left(1 - \frac{1}{v_t}\right) + \sigma_{AI/WP}^2 \left(1 + \frac{\sum_{i'} v_{i'}^2}{v_t^2} - \frac{2v_i}{v_t}\right) \\
 &\quad + \sigma_{AP/WG}^2 \left(1 + \frac{\sum_{\rho'} v_{\rho'}^2}{v_t^2} - \frac{2v_\rho}{v_t}\right) + \sigma_{AG}^2 \left(1 + \frac{\sum_{g'} v_{g'}^2}{v_t^2} - \frac{2v_g}{v_t}\right).
 \end{aligned}$$

700 Moreover, we have $\sum_a \left(1 - \frac{1}{v_t}\right) = H - 1$ and

$$\sum_a \left(1 + \frac{\sum_{x'} v_{x'}^2}{v_t^2} - \frac{2v_x}{v_t}\right) = H + \frac{\sum_{x'} v_{x'}^2}{v_t} - 2 \sum_x \frac{v_x^2}{v_t} = H - \sum_x \frac{v_x^2}{v_t}, \quad x = i, \rho, g.$$

701 Now, let us derive the final formula in Equation (2):

$$\begin{aligned}
 E(SS_{TOT}) &= E\left[\sum_a (y_{g\rho ia} - \bar{y}_t)^2\right] = \sum_a E[(y_{g\rho ia} - \bar{y}_t)^2] \\
 &= \sum_a \left[\sigma_{WI}^2 \left(1 - \frac{1}{v_t}\right) + \sigma_{AI/WP}^2 \left(1 + \frac{\sum_{i'} v_{i'}^2}{v_t^2} - \frac{2v_i}{v_t}\right) \right. \\
 &\quad \left. + \sigma_{AP/WG}^2 \left(1 + \frac{\sum_{\rho'} v_{\rho'}^2}{v_t^2} - \frac{2v_\rho}{v_t}\right) + \sigma_{AG}^2 \left(1 + \frac{\sum_{g'} v_{g'}^2}{v_t^2} - \frac{2v_g}{v_t}\right) \right] \\
 &= \sigma_{WI}^2 (H - 1) + \sigma_{AI/WP}^2 \left(H - \sum_i \frac{v_i^2}{v_t}\right) + \sigma_{AP/WG}^2 \left(H - \sum_\rho \frac{v_\rho^2}{v_t}\right) + \sigma_{AG}^2 \left(H - \sum_g \frac{v_g^2}{v_t}\right).
 \end{aligned}$$