Analysis of Population Structure in Autotetraploid Species

Joëlle Ronfort,* Eric Jenczewski,* Thomas Bataillon* and François Rousset†

**Laboratoire de Génétique et d'Amélioration des Plantes, Institut National de la Recherche Agronomique, Domaine de Melgueil, 34130 Mauguio, France and †Laboratoire Génétique et Environnement, Institut des Sciences de l'Evolution, Université des Sciences et Techniques du Languedoc, 34095 Montpellier, France*

Manuscript received March 4, 1998
Accepted for publication June 23, 1998

ABSTRACT

Population structure parameters commonly used for diploid species are reexamined for the particular case of tetrasomic inheritance (autotetraploid species). Recurrence equations that describe the evolution of identity probabilities for neutral genes in an "island model" of population structure are derived assuming tetrasomic inheritance. The expected equilibrium value of F_{ST} is computed. In contrast to diploids, the correlation of genes between individuals within populations with respect to genes between populations (F_{ST}) may vary among loci due to the particular segregation patterns expected under tetrasomic inheritance and is consequently inappropriate for estimating demographic parameters in such populations. We thus define a new parameter (ρ) and derive its relationship with *Nm*. This relationship is shown to be independent from both the selfing rate and the proportion of double reduction. Finally, the statistical procedure required to evaluate these parameters using data on gene frequencies distribution among autotetraploid populations is developed.

DUE to its frequent occurrence among angiosperm species (from 30 to 50%; Stebbins 1971; Grant 1981), polyploidy is now recognized as an important step in the evolutionary diversification of flowering plants (Lewis 1980; Levin 1983; Stebbins 1985; Thompson and Lumaret 1992; Soltis and Soltis 1993; Bretagnolle and Thompson 1995, 1996; Petit *et al.* 1996, 1997). Polyploid species are commonly classified in two major types according to their presumed origin: allopolyploids are thought to result from hybridization between different taxa and subsequent chromosome doubling, while autopolyploids presumably stem from the chromosome doubling of the same genome, primarily by fusion of unreduced gametes (Bever and Felber 1992; Bretagnolle and Thompson 1995). Autotetraploidy was originally thought to be rare and maladaptive as compared to allopolyploidy. However, a growing number of studies using genetic information in addition to cytological and morphological traits confirm that autopolyploids are more common and of greater evolutionary importance than originally appreciated (Levin 1983; Crawford 1985; Rieseberg and Doyle 1989; Soltis and Soltis 1989).

Due to the addition of divergent genomes, inheritance in allopolyploids is disomic; *i.e.*, pairing behavior during meiosis is similar to that of nonhomologous pairs of chromosomes in diploids. In contrast, segregation patterns in autopolyploids are much more complex because more than two homologous chromosomes can pair during meiosis. Multivalents leading to polysomic inheritance are formed. This does not necessarily lead to random assortments of homologous chromosomes into gametes; two sister chromatids may also segregate into the same gamete (Figure 1). This phenomenon, known as "double reduction," is specific to autopolyploids. It increases the production of homozygous gametes as compared to what is expected under random chromosome segregation and is thus likely to alter many basic expectations of population genetics (Bever and Felber 1992). Because the frequency of double reduction depends on the occurrence of crossovers between the centromere and the locus under consideration (Figure 1), segregation patterns are expected to vary among loci, obscuring predictions regarding genetic aspects of autopolyploids.

Probably because of the agronomic significance of polyploid species, the consequences of polysomic inheritance and double reduction have been investigated, especially for self-fertilization and regular systems of inbreeding (Haldane 1930; Demarly 1963; Bennett 1968; Gallais 1990). In contrast, few investigations have dealt with the amount and patterns of genetic variation among naturally occurring autopolyploid populations, and theoretical models incorporating population structure and estimation procedures are still lacking for tetrasomic inheritance (Glendinning 1989; Bever and Felber 1992 for review; Moody *et al.* 1993).

For diploids, the distribution of genetic diversity within and among natural populations is commonly analyzed using theoretical models of population structure, for instance, the island model or the stepping stone model. Functions of probabilities of gene identity within and between units (populations, subpopulations), such

Corresponding author: Joëlle Ronfort, Laboratoire de Génétique et d'Amélioration des Plantes, Institut National de la Recherche Agronomique, Domaine de Melgueil, 34130 Mauguio, France.
E-mail: ronfort@ensam.inra.fr

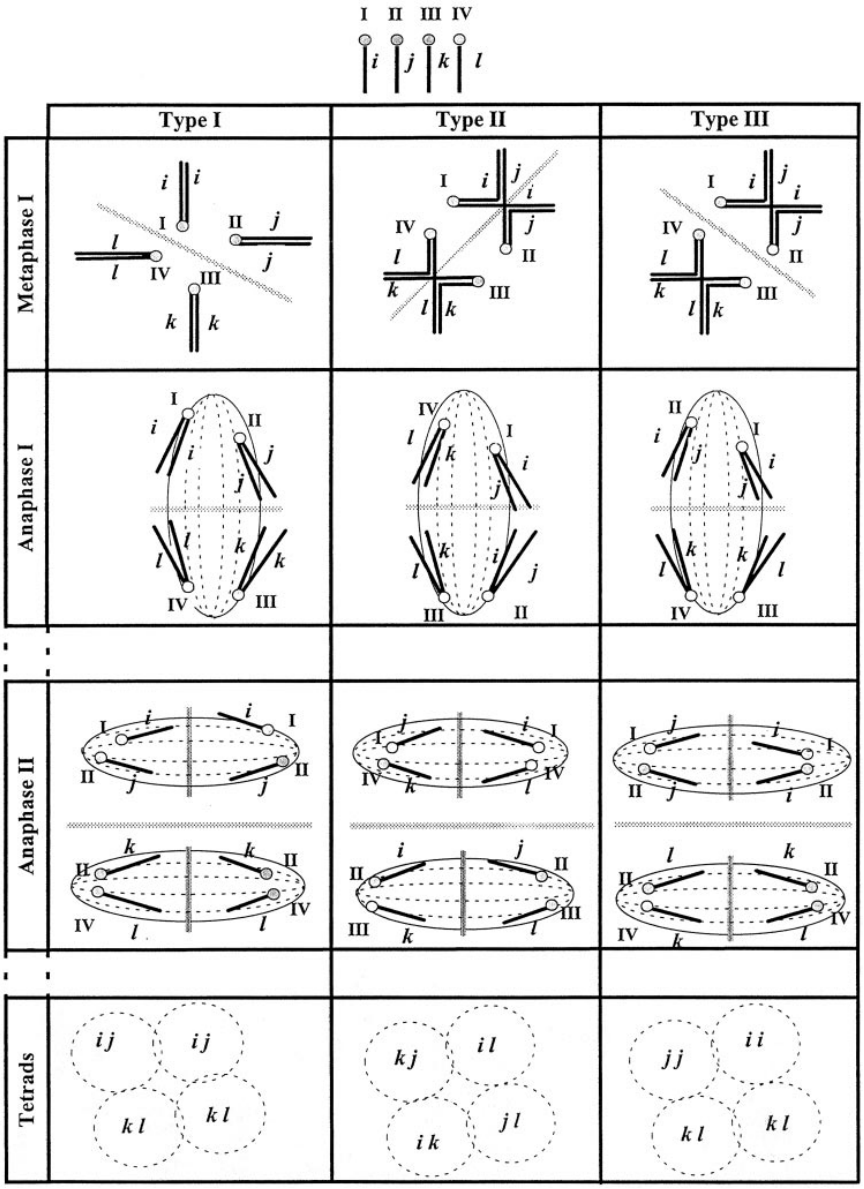

Figure 1.—Possible segregation patterns of a locus in an autotetraploid individual following the formation of a quadrivalent. Type I describes the segregation patterns expected when there is no crossover between the centromere and the locus. The first division is then reductional. When a crossover occurs between the centromere and the locus (Types II and III), the first division can be either equational (Type II) or reductional (Type III). Under Type III, the second division may then lead to double reduction. In the present case, gametes *ii* and *jj* have undergone double reduction.

as F_{ST} (Wright 1951), can be estimated using isozyme or DNA-based marker diversity and can be compared to expectations under specific models such as Wright's island or isolation by distance models (Slatkin and Barton 1989; Rousset 1997). The relationships between estimates and expectations can then be used to quantify gene flow between the studied units or even to understand how ecological and life history traits may influence the distribution of genetic variation within and among populations (see, for example, Loveless and Hamrick 1984; Hamrick and Godt 1990). However, models of population structure as well as estimation procedures have been almost exclusively devoted to diploid populations (see, however, Wright 1938).

The aim of this article is to develop a theoretical framework for the analysis of population structure in autotetraploid species. Recurrence equations that describe probabilities of gene identity under the island or isolation by distance models may be generalized for the case of tetrasomic inheritance; the case of the island model is given here as an illustration. Equilibrium values for traditional *F*-statistics parameters are derived. Because the proportion of double reduction may vary over loci, we define an additional function of probabilities of gene identity. This parameter seems appropriate to analyze population structure in autotetraploids, because its relationship with the migration rate and the population size is shown to be independent from both the selfing rate and the proportion of double reduction. Finally, following Weir and Cockerham (1984), we define estimators for the different parameters using the analysis of variance framework.

HIERARCHICAL GENIC STRUCTURE AND DEFINITION OF PARAMETERS

Let Q stand for the probability of identity, Q_0 for pairs of genes within individuals, Q_1 for pairs of gene between

individuals within subpopulations, and Q_2 for pairs of genes between subpopulations. Throughout this article, the notation Q_j will refer to probabilities of identity in state (IIS) and the j indices ($j = 0$ to 2) to the same pairs of genes. The addition of a dot on the top of a parameter will denote probabilities of identity by descent (IBD) (*i.e.*, $\dot{Q}_j$), and the standard notation $\equiv$ will be used to distinguish the definition of parameters from their values under particular models of population structure.

Under tetrasomic inheritance, four genes are available at a given locus. Then, a random pair of genes within individuals (Q_0) can be issued either from the same gamete (probability 1/3) or from two different gametes (probability 2/3). If Q_A and Q_B denote the probability of IIS associated, respectively, with these two categories of pairs of genes, then $Q_0 = (Q_A + 2Q_B)/3$.

Following Cockerham and Weir (1987, 1993; see also Rousset 1996), *F*-statistics parameters can be defined as

$$F_{IT} \equiv \frac{Q_0 - Q_2}{1 - Q_2} \tag{1}$$

$$F_{IS} \equiv \frac{Q_0 - Q_1}{1 - Q_1} = \frac{(Q_A + 2Q_B)/3 - Q_1}{1 - Q_1} \tag{2}$$

$$F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2}. \tag{3}$$

Another parameter we will consider is

$$\dot{\rho} \equiv \frac{\dot{Q}_1 - \dot{Q}_2}{(1 + \dot{Q}_A + 2\dot{Q}_B)/4 - \dot{Q}_2}. \tag{4}$$

This parameter is analogous to the "correlation between truly outcrossed mates" in diploids (Waller and Knight 1989; Tachida and Yoshimaru 1996). For diploids, interest in this correlation has come from the fact that the relationship between this parameter, the migration rate, and the population size is independent from the selfing rate (see Nagylaki 1983; Tachida and Yoshimaru 1996). We will show that, for $\dot{\rho}$ in autotetraploids, this relationship is moreover independent of the proportion of double reduction and therefore identical for all loci independently of their distance to the centromere.

Let us define $\dot{Q}_r$ as the IBD probability for two genes in different individuals located either in two different subpopulations ($r = 2$) or in the same subpopulation ($r = 1$) and use the relationship between coalescence of genes and identity probabilities (Malécot 1975; Tachida 1985; Slatkin and Voelm 1991): the probability of IBD for a pair of genes is the probability that neither gene has mutated between the present time and the time of first common ancestry, that is, their coalescence time (Malécot 1975; Slatkin 1991). This yields the expression

$$\dot{Q}_r = \sum_{t=1}^{\infty} (1 - \mu)^{2t} P(t) = E[(1 - \mu)^{2T}], \tag{5}$$

where μ is the mutation rate per generation, $P(t)$ the probability that two genes coalesce at generation t in the past, and T a random variable that describes the coalescence time for these two genes. As shown by Slatkin and Voelm (1991; see also Tachida and Yoshimaru 1996), if we think of the process as going backward in time, then T can be divided in two phases: T_1, the waiting time for two genes to be found in the same individual, and T_2, the time for the two genes in the same individual to coalesce (Figure 2). As a result, and since T_1 and T_2 are independent,

$$\dot{Q}_r = E[(1 - \mu)^{2(T_1 + T_2)}] = E[(1 - \mu)^{2T_1}] \cdot E[(1 - \mu)^{2T_2}]. \tag{6}$$

Because T_2 corresponds to a coalescence time, then using the relationships between the coalescence of genes and identity probabilities, $E[(1 - \mu)^{2T_2}]$ represents the IBD probability for a pair of genes when both are sampled in the same individual (Figure 2):

$$E[(1 - \mu)^{2T_2}] = \frac{1 + \dot{Q}_A + 2\dot{Q}_B}{4}. \tag{7}$$

Unlike T_2, T_1 in this instance is not a coalescence time but rather the "waiting time" for two genes initially at distance r to migrate within the same individual (Figure 2). To define T_1, we do not make any reference to identity between the two genes under consideration, and this waiting time will depend only on the initial distance between the two genes ($r = 1$ or 2) and on the way genes migrate within and between subpopulations. Hence, $E[(1 - \mu)^{2T_1}]$, which we will denote $\dot{h}_r$ in what follows, is not an IBD probability but simply denotes the probability that neither gene has mutated during T_1. Since double reduction affects only transition probabilities for genes within individuals, it does not affect T_1 nor $\dot{h}_r$. These two parameters are consequently independent from the proportion of double reduction.

Now, following (2), and using (3), the IBD probability for two genes in different individuals at distance r reduces to

$$\dot{Q}_r = \dot{h}_r \cdot \frac{1 + \dot{Q}_A + 2\dot{Q}_B}{4} \tag{8}$$

and, putting this formula into (4), yields the following expression for $\dot{\rho}$:

$$\dot{\rho} = \frac{(1 + \dot{Q}_A + 2\dot{Q}_B)\dot{h}_1/4 - (1 + \dot{Q}_A + 2\dot{Q}_B)\dot{h}_2/4}{(1 + \dot{Q}_A + 2\dot{Q}_B)/4 - (1 + \dot{Q}_A + 2\dot{Q}_B)\dot{h}_2/4} = \frac{\dot{h}_1 - \dot{h}_2}{1 - \dot{h}_2}. \tag{9}$$

This parameter is of interest for two reasons: (1) Because the $\dot{h}_r$s are independent of the coefficient of double reduction, this equation shows that this is also true for $\dot{\rho}$; (2) as will be shown later, the expected value of $\dot{\rho}$ can be deduced with minimal effort from previous models of haploid populations.

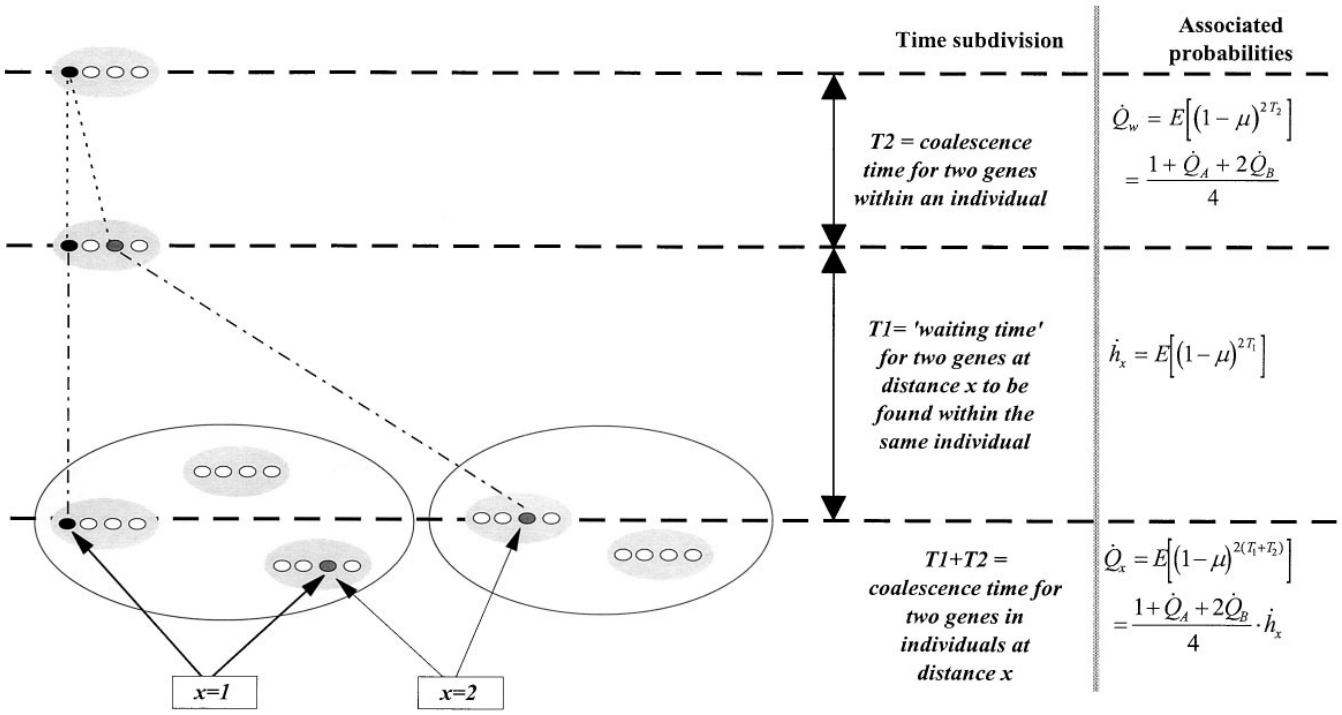

Figure 2.—Waiting times for coalescence of two genes located in two different individuals that are at distance x at the present time ($x = 1$ for individuals located in the same population; $x = 2$ for individuals in two different populations) and their associated probabilities. Populations are represented by large ellipses; small circles represent genes (four such circles denote an individual) shaded in black or gray for the genes studied, in white for genes not considered.

Consider now,

$$\frac{\dot{\rho}}{1-\dot{\rho}} = \frac{\dot{Q}_1 - \dot{Q}_2}{(1+\dot{Q}_A+2\dot{Q}_B)/4 - \dot{Q}_1}. \qquad (10)$$

Noting that

$$\frac{1+3\dot{F}_{IS}}{4} = \frac{(1+\dot{Q}_A+2\dot{Q}_B)/4 - \dot{Q}_1}{1-\dot{Q}_1},$$

we can always write

$$\frac{\dot{F}_{ST}}{1-\dot{F}_{ST}} = \frac{(1+\dot{Q}_A+2\dot{Q}_B)/4 - \dot{Q}_1}{1-\dot{Q}_1} \cdot \frac{\dot{Q}_1 - \dot{Q}_2}{(1+\dot{Q}_A+2\dot{Q}_B)/4 - \dot{Q}_1},$$

which reduces to

$$\frac{\dot{F}_{ST}}{1-\dot{F}_{ST}} = \frac{1+3\dot{F}_{IS}}{4} \cdot \frac{\dot{\rho}}{1-\dot{\rho}}. \qquad (11)$$

This may be compared to the result of the diploid model with selfing (Tachida and Yoshimaru 1996), in which one can write

$$\frac{\dot{F}_{ST}}{1-\dot{F}_{ST}} = \frac{1+\dot{F}_{IS}}{2} \cdot \frac{\dot{\rho}}{1-\dot{\rho}}. \qquad (12)$$

EQUILIBRIUM VALUES OF THE PARAMETERS IN AN ISLAND MODEL

We consider a finite island model (Wright 1951) of population structure: a set of n subpopulations, each consisting of N individuals, with nonoverlapping generations. Individuals are monoecious and subpopulations exchange migrant gametes at a rate m. Each migrant has an equal chance of coming from each of the other $n-1$ subpopulations. Genes are assumed to be neutral and the mutation rate μ is the same for all alleles. Following Nagylaki (1983) and Crow and Aoki (1984), two notations will be used. After migration, the proportion of pairs of genes that originate from one subpopulation in the previous generation is $a = (1-m)^2 + m^2/(n-1)$ for genes within a subpopulation, and $b = (1-a)/(n-1)$ for genes from different subpopulations. In each subpopulation, a proportion S of offspring is produced through selfing and the proportion of double reduction for the studied locus is denoted α.

When individuals are autotetraploid, there are $4N$ genes in each subpopulation. Then, provided neither gene has mutated [with probability $\gamma = (1-\mu)^2$], genes originating from the same subpopulation are identical by descent with probability $(1+3\dot{Q}_0)/4N + (1-1/N)\dot{Q}_1$, while genes from different subpopulations are identical by descent with probability $\dot{Q}_2$. The recurrence relations for $\dot{Q}_1$ and $\dot{Q}_2$ are as follows (t denoting time in generation):

$$\dot{Q}_{1,t+1} = \gamma \cdot \left\{ a \cdot \left[\frac{1 + 3\dot{Q}_{0,t}}{4N} + \left(1 - \frac{1}{N}\right)\dot{Q}_{1,t} \right] + (1 - a)\dot{Q}_{2,t} \right\} \quad (13)$$

$$\dot{Q}_{2,t+1} = \gamma \cdot \left\{ b \cdot \left[\frac{1 + 3\dot{Q}_{0,t}}{4N} + \left(1 - \frac{1}{N}\right)\dot{Q}_{1,t} \right] + (1 - b)\dot{Q}_{2,t} \right\}. \quad (14)$$

Combining these two relationships, we obtain

$$\dot{Q}_{1,t+1} - \dot{Q}_{2,t+1} = \gamma \cdot \left\{ (a - b) \cdot \left[\frac{1 + 3\dot{Q}_{0,t}}{4N} + \left(1 - \frac{1}{N}\right)\dot{Q}_{1,t} \right] + (b - a)\dot{Q}_{2,t} \right\}. \quad (15)$$

At equilibrium, the Q_i's do not change, hence

$$(\dot{Q}_1 - \dot{Q}_2) \cdot \left[1 - \gamma \cdot (a - b) \cdot \left(1 - \frac{1}{N}\right) \right] = \gamma \cdot (a - b) \cdot \frac{1 + 3\dot{Q}_0 - 4\dot{Q}_2}{4N}. \quad (16)$$

Using $d = a - b$, this equation can be expressed as

$$C = \frac{\dot{Q}_1 - \dot{Q}_2}{1 + 3\dot{Q}_0 - 4\dot{Q}_2} = \frac{\gamma d}{4N(1 - \gamma d) + 4\gamma d}. \quad (17)$$

Noting that $(1 + \dot{Q}_A + 2\dot{Q}_B)/4 - \dot{Q}_1 = (1 + 3\dot{Q}_0 - 4\dot{Q}_1)/4$, then substituting this into (10) and using (17), yields

$$\frac{\dot{\rho}}{1 - \dot{\rho}} = \frac{4(\dot{Q}_1 - \dot{Q}_2)}{1 - 3\dot{Q}_0 - 4\dot{Q}_1} = \frac{4C}{1 - 4C} = \frac{\gamma d}{N(1 - \gamma d)}, \quad (18)$$

which is the same result as in the diploid (or haploid) model. Using $\theta = n/(n - 1)$, Equation 18 becomes

$$\frac{\dot{\rho}}{1 - \dot{\rho}} = \frac{1}{2N(m\theta + \mu)} \cdot (1 + O(m) + O(\mu)), \quad (19)$$

i.e.,

$$\frac{\dot{\rho}}{1 - \dot{\rho}} \approx \frac{1}{2N(m\theta + \mu)} \quad (20)$$

and, using (11),

$$\frac{\dot{F}_{ST}}{1 - \dot{F}_{ST}} \approx \frac{1 + 3\dot{F}_{IS}}{4} \cdot \frac{1}{2N(m\theta + \mu)}. \quad (21)$$

As one may note, we do not need to know identity probabilities within subpopulations (Q_0 and Q_1) to derive these results. For diploids, the expected equilibrium value of F_{IS} depends on the selfing rate (S), and the population size (N). For autotetraploids, it also depends on the proportion of double reduction that increases the proportion of homozygous gametes produced [see, for example, Bennett (1968) for the case of a (single) large autotetraploid population]. When neither selfing nor double reduction are occurring in the population (*i.e.*, $S = 0$ and $\alpha = 0$), $F_{IS} = 0$. Equation 21 can then be further simplified into

$$\dot{F}_{ST} \approx \frac{1}{1 + 8Nm\theta + 8N\mu}. \quad (22)$$

Expected values of $\dot{\rho}$ can be computed for other mutation models as previously described (*e.g.*, Crow and Aoki 1984; Rousset 1996), as well as for other geographical models. Under isolation by distance models, it can be shown that $\rho_r/(1 - \rho_r) \approx r/(2D\sigma^2)$ + Constant, for a pair of populations at distance r in a one-dimensional model, and $\rho_r/(1 - \rho_r) \approx \ln(r)/(2D\pi\sigma^2)$ + Constant, in a two-dimensional model, where D is the population density and σ^2 is a measure of dispersal (Rousset 1997).

POPULATION PARAMETERS ESTIMATION

Consider a dataset describing the genotypic constitution of autotetraploid individuals sampled (at random) from a set of r subpopulations. Each subpopulation is represented by n_i individuals (sample size), where i refers to the ith subpopulation. To build estimators for the level of population differentiation, we use the linear model with hierarchical effects (subpopulations, individuals within subpopulations, and genes within individuals) developed by Cockerham (1969, 1973) for the analysis of diploid population structure. Now x_{ijk} is an indicator variable describing the state of the kth gene ($1 \le k \le 4$, instead of $1 \le k \le 2$ for diploids) in the jth sampled individual ($1 \le j \le n_i$) of the ith subpopulation ($1 \le i \le r$). For a particular allele u, $x_{ijk:u} = 1$, if the gene is u, $x_{ijk:u} = 0$ otherwise, and the ANOVA setup is as follows:

$$\begin{aligned} \sum_{\text{Subpop}}^{r} \sum_{\text{Indiv}}^{n_i} \sum_{\text{Genes}}^{4} (x_{ijk:u} - x_{\cdots:u})^2 &= \sum_i \sum_j \sum_k (x_{ijk:u} - x_{ij:u})^2 \\ &\quad + \sum_i \sum_j \sum_k (x_{ij:u} - x_{i\cdot:u})^2 \\ &\quad + \sum_i \sum_j \sum_k (x_{i\cdot:u} - x_{\cdots:u})^2 \\ &= SS_{\text{g[enes]:}u} + SS_{\text{i[ndividuals]:}u} \\ &\quad + SS_{\text{s[ubpopulation]:}u}. \end{aligned}$$

Using the same developments as for diploids (Weir 1996), the following sum of squares expectations can be derived (details are given in the appendix): for genes within individuals

$$\varepsilon(SS_{\text{g[enes]}}) = 3S_1(1 - Q_0); \quad (23a)$$

for genes between individuals within subpopulations

$$\varepsilon(SS_{\text{i[ndivis]}}) = W_d \cdot (4(Q_0 - Q_1) + (1 - Q_0)); \quad (23b)$$

and for genes between individuals from different subpopulations

$$\varepsilon(SS_{s[ubpops]}) = 4W_a \cdot (Q_1 - Q_2) + W_w \cdot [4(Q_0 - Q_1) + (1 - Q_0)], \quad (23c)$$

where $S_1 = \Sigma_i n_i$, $S_2 = \Sigma_i n_i^2$, $W_d \equiv S_1 - r$, $W_a \equiv S_1 - S_2/S_1$, and $W_w \equiv r - 1$.

From Equations 23a–23c, we obtain

$$Q_1 - Q_2 = \frac{W_d\varepsilon(SS_s) - W_w\varepsilon(SS_i)}{4W_aW_d} \quad (24)$$

$$1 - Q_2 = \frac{1}{4W_aW_d}\left[\frac{W_aW_d}{S_1}\varepsilon(SS_g) + (W_a - W_w)\varepsilon(SS_i) + W_d\varepsilon(SS_s)\right], \quad (25)$$

which yield an estimator of F_{ST}:

$$\hat{F}_{ST} = \frac{W_dSS_s - W_wSS_i}{[W_dSS_p + (W_a - W_w)SS_i + (W_aW_d/S_1)SS_g]}. \quad (26)$$

Now, noting that $1 + 3Q_0 - 4Q_1 = 1 - Q_0 + 4(Q_0 - Q_1) = \varepsilon(SS_I)/W_d$, we have

$$\frac{\hat{\rho}}{1 - \hat{\rho}} = \frac{4 \cdot (\hat{Q}_1 - \hat{Q}_2)}{1 + 3\hat{Q}_0 - 4\hat{Q}_1} = \frac{W_dSS_s - W_wSS_i}{W_aSS_i}. \quad (27)$$

An estimator of $\hat{F}_{IT} \equiv 1 - (1 - \hat{Q}_1)/(1 - \hat{Q}_2)$ is

$$\hat{F}_{IT} = \frac{4(W_aW_d/3S_1)SS_g}{[W_dSS_p + (W_a - W_w)SS_i + (W_aW_d/S_1)SS_g]}, \quad (28)$$

and

$$\hat{F}_{IS} = 1 - \frac{1 - \hat{F}_{IT}}{1 - \hat{F}_{ST}}. \quad (29)$$

For all these parameters, multilocus estimates (*i.e.*, combining the information from all alleles and all loci) are defined as the sum of locus-specific numerators divided by the sum of locus-specific denominators (see also Reynolds *et al.* 1983; Weir 1996). For example,

$$\frac{\hat{\rho}}{1 - \hat{\rho}} = \frac{\sum_{l=1}^{n_l}\sum_{u=1}^{u_l}(W_dSS_S - W_wSS_i)_{lu}}{\sum_{l=1}^{n_l}\sum_{u=1}^{u_l}(W_aSS_i)_{lu}}, \quad (30)$$

where l refers to the lth loci and u to the uth allele (with n_l, the number of locus and u_l, the number of alleles at locus l). Given the dependency of *F*-statistics on the proportion of double reduction (see above), multilocus estimates of these parameters will be appropriate to make inferences about the balance between migration (and/or mutation) and drift only if $\alpha = 0$ for all the studied loci. As soon as $\alpha \neq 0$ for at least one locus, only the estimate of ρ will have this property.

DISCUSSION

The aim of this study was to adapt the use of Wright's F_{ST} to estimate population structure and gene flow in autotetraploid species. In contrast to diploids, F_{ST} estimates in autotetraploids are expected to vary across the loci as a consequence of different amounts of double reduction during meiosis (Figure 1 and Introduction). This problem is illustrated in Equation (11) because F_{IS} will vary depending on both the selfing rate and the proportion of double reduction (α). Since the proportion of double reduction for a given locus is difficult to assess empirically and because population structure estimates should be based on several loci, we defined a new function of identity probabilities, ρ, which is an analogue to the "correlation between truly outcrossed mates" previously defined for diploids (Waller and Knight 1989; Tachida and Yoshimaru 1996). For both diploids and tetraploids, the relationship between this correlation and the product Nm is independent from the selfing rate (except when selfing affects migration). For autotetraploids, interest in ρ comes mainly from the fact that this relationship is also independent of the proportion of double reduction and therefore identical for all loci independently of their distance to the centromere. The parameter ρ can consequently be used to assess population structure over many loci, without any prior knowledge concerning the proportion of double reduction.

Inspection of the relationship between ρ and F_{ST} (11) shows that F_{ST} is increased by a factor $(1 + 3F_{IS})/4$ when self-fertilization or double reduction occurs within subpopulations. This means that like self-fertilization, double reduction reduces the effective subpopulation size and hence promotes differentiation among subpopulations (for the studied locus). The complication due to partial selfing or double reduction can be absorbed in the single parameter F_{IS} and by defining the effective population size as $N_Z = N/(1 + 3F_{IS})$. Equation (21) can then be used with N_Z replacing N, *i.e.*, $F_{ST}/(1 - F_{ST}) = 1/(8N_Zm\theta + 8N_Z\mu)$, while ρ is still equal to $\rho/(1 - \rho) \approx 1/(2Nm\theta + 2N\mu)$, which depends only on the migration rate, mutation rate, and the demographic population size (*i.e.*, N, not N_Z). Comparison of Equation 11 with the results of the diploid model (12) further shows that self-fertilization has a greater influence on differentiation in autotetraploids as compared to diploids.

When ignoring selfing and double reduction, the expected effect of drift under the island model of population structure is halved at equilibrium as compared to expectations for diploids, *i.e.*, $F_{ST} \approx 1/(1 + 4Nm\theta + 4N\mu)$ (Crow and Aoki 1984; Cockerham and Weir 1987). This can be interpreted as the decrease in the rate of coalescence of genes within subpopulations and is due to the fact that the probability of drawing the same gene within an individual is reduced to 1/4 in an autotetraploid species instead of 1/2 in diploids. In other words, this means that, for a same demographic population size, the effective population size is doubled in an autotetraploid population as compared to a diploid one. This result is in accordance with earlier work

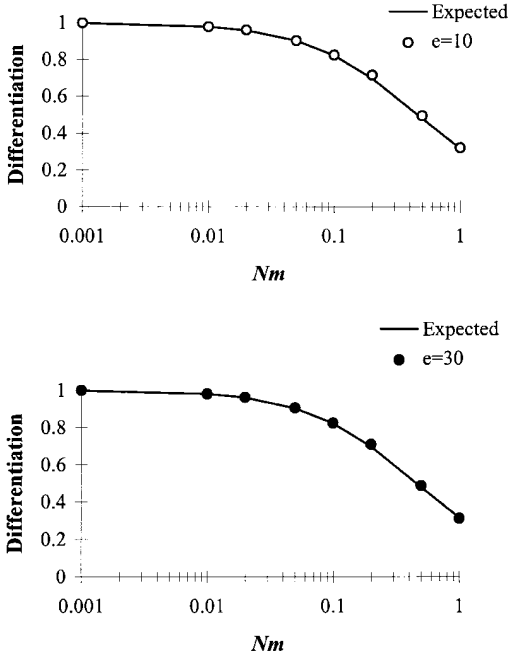

Figure 3.—Comparison of average values of the estimator ($\hat{\rho}$) with expected equilibrium values of ρ. An island model was simulated assuming $K = 9$ allelic states for each of 10 loci, $n = 12$ subpopulations and $\mu = 10^{-5}$ (see text for a complete description of the simulation procedure). To determine the number of generations required for the population to reach its equilibrium, sampling was performed in generations 200, 400, 1000, 2000, and 5000. For $N = 50$, the estimates were stable after 2000 generations. Each symbol gives the value of $\hat{\rho}$ computed after 2000 generations and averaged over 10 replicated simulations (standard errors were always less than 10^{-3} and are therefore not shown) for $N = 50$, no selfing, $\alpha = 0$. ○ is for sample size $e = 10$, ● for $e = 30$. Lines were computed using the expected equilibrium value of ρ, *i.e.*, $\rho \approx 1/(1 + 2Nm\theta + 2N\mu)$ for small *Nm*. Simulations were performed for two other parameter sets: $N = 50$, $S = 0.2$, $\alpha = 0$; $N = 50$, no selfing, $\alpha = 2/7$. ρ is identical for these different parameter sets and the average values of the estimates ($\hat{\rho}$) are too close to be distinguished on the figure.

on autotetraploids (Moody *et al.* 1993) that assumed that mutation alone was opposed to genetic drift (*i.e.*, $m = 0$, $F_{IS} = 0$, $\alpha = 0$): $F_{ST} = 1/(1 + 8N\mu)$.

Following the linear model derived in Cockerham (1969, 1973) for diploid data, estimators for *F*-statistics and ρ can be computed through hierarchical analyses of variance of gene frequencies. Simulations were performed to assess possible bias in the estimation of ρ due to small sample sizes. We simulated a finite island model composed of *n* monoecious subpopulations of size *N*. In each subpopulation, 10 neutral, independent loci (recombination rate = 0.5), each with *K* possible allelic states (*K*-allele model), and all segregating according to the same proportion of double reduction (α) were modeled. Initial frequencies of the different allelic states were made equal in all the subpopulations (initial frequency = $1/K$). Each subpopulation had the same mating system: complete outcrossing ($S = 0$) or partial selfing ($S \neq 0$). We assumed discrete and nonoverlapping generations. Mutation occurs at a rate μ per locus per generation, each allele having an equal chance to mutate toward one of the *K*-1 other allelic states. Migration occurs through male gametes only: to produce the next generation in a given subpopulation, each pollen grain was sampled independently, and with probability *m* it was chosen among gametes from the remaining $n - 1$ subpopulations. As shown in Figure 3, the discrepancies between the average value of the estimator and the expected value of ρ are very small even for small sample sizes, with either $S = 0$ or $S \neq 0$ and $\alpha \neq 0$.

We wrote a computer program estimating *F*-statistics and the parameter ρ according to the ANOVA setup developed above (details of the computations are given in the appendix). The program provides estimations for ρ, F_{ST}, F_{IS}, and F_{IT} for each allele as well as estimates combining data over alleles and over loci. To test for a departure from $F_{ST} = 0$, the program allows for Fisher's exact test on (population × genotypes) contingency tables [for each locus separately, see Raymond and Rousset (1995) for the diploid model]. Exact tests on contingency tables in which cell counts are tetraploid genotypes are valid even if there is double reduction. As for diploid datasets (Raymond and Rousset 1995), the software allows for analysis either over the whole set of populations or for pairs of populations. The program containing both estimations and exact tests procedures is available upon request.

We thank D. Couvet and P. Jarne for discussions, M. Raymond for advice concerning the computer program, and J. M. Prosperi for comments on the manuscript. This work was supported by a grant from the French "Bureau des Ressources Génétiques" to E.J. and J.R. This is contribution number 98-085 of the Institut des Sciences de l'Evolution.

LITERATURE CITED

Bennett, J. H., 1968 Mixed self- and cross-fertilization in a tetrasomic species. Biometrics **24:** 485–500.

Bever, J. D., and F. Felber, 1992 The theoretical population genetics of autopolyploidy. Oxford Surv. Evol. Biol. **8:** 185–217.

Bretagnolle, F., and J. D. Thompson, 1995 Gametes with somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. New Phytol. **129:** 1–22.

Bretagnolle, F., and J. D. Thompson, 1996 An experimental study of ecological differences in winter growth between sympatric diploid and autotetraploid *Dactylis glomerata*. J. Ecol. **84:** 343–351.

Cockerham, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

Cockerham, C. C., 1973 Analysis of gene frequencies. Genetics **74:** 679–700.

Cockerham, C. C., and B. S. Weir, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA **84:** 8512–8514.

Cockerham, C. C., and B. S. Weir, 1993 Estimation of gene flow from *F*-Statistics. Evolution **47:** 855–863.

Crawford, D. J., 1985 Electrophoretic data and plant speciation. Syst. Bot. **10:** 405–416.

Crow, J. F., and K. Aoki, 1984 Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA **81:** 6073–6077.

Demarly, Y., 1963 Génétique des tétraploïdes et amélioration des plantes. Ann. Amelior. Plantes **13:** 307–400.
Gallais, A., 1990 *Théorie de la Selection en Amélioration des Plantes.* Masson, Paris.
Glendinning, D. R., 1989 Some aspects of autotetraploid population dynamics. Theor. Appl. Genet. **78:** 233–242.
Grant, V., 1981 *Plant Speciation,* Ed. 2. Columbia University Press, New York.
Haldane, J. B. S., 1930 Theoretical genetics of autotetraploids. J. Genet. **22:** 359–372.
Hamrick, J. L., and J. W. Godt, 1990 Allozyme diversity in plant species, pp. 43–63 in *Plant Population Genetics, Breeding and Genetic Resources,* edited by A. H. D. Brown, M. T. Clegg, A. L. Kahler and B. S. Weir. Sinauer Associates Inc., Sunderland, MA.
Levin, D. A., 1983 Polyploidy and novelty in flowering plants. Am. Nat. **122:** 1–25.
Lewis, W. H., 1980 Polyploidy in angiosperms: dicotyledons, pp. 241–268 in *Polyploidy, Biological Relevance,* edited by W. H. Lewis. Plenum Press, New York.
Loveless, M. D., and J. L. Hamrick, 1984 Ecological determinants of genetic structure in plant populations. Annu. Rev. Ecol. Syst. **15:** 65–95.
Malécot, G., 1948 *Les Mathématiques de l'Hérédité.* Masson, Paris.
Malécot, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. **8:** 212–241.
Moody, M. E., L. D. Mueller and D. E. Soltis, 1993 Genetic variation and random drift in autotetraploid populations. Genetics **134:** 649–657.
Nagylaki, T., 1983 The robustness of neutral modes of geographical variation. Theor. Popul. Biol. **24:** 268–294.
Petit, C., J. D. Thompson and F. Bretagnolle, 1996 Phenotypic plasticity in relation to ploidy level and corm production in the perennial grass *Arrhenatherum elatius.* Can. J. Bot. **74:** 1964–1973.
Petit, C., P. Lesbros, X. Ge and J. D. Thompson, 1997 Variation in flowering phenology and selfing rate across a contact zone between diploid and tetraploid *Arrhenaterum elatius* (Poaceae). Heredity **79:** 31–40.
Raymond, M., and F. Rousset, 1995 An exact test for population differentiation. Evolution **49:** 1280–1283.
Reynolds, J., B. S. Weir and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105:** 767–779.
Rieseberg, L. H., and M. F. Doyle, 1989 Tetrasomic segregation in the naturally occurring autotetraploid *Allium nevii* (Alliaceae). Hereditas **111:** 31–36.
Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357–1362.
Rousset, F., 1997 Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145:** 1219–1228.
Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.
Slatkin, M., and N. H. Barton, 1989 A comparison of three indirect methods for estimating average levels of gene flow. Evolution **43:** 1349–1368.
Slatkin, M., and L. Voelm, 1991 F_{ST} in a hierarchical island model. Genetics **127:** 627–629.
Sokal, R. R., and F. J. Rohlf, 1995 *Biometry,* Ed. 3. Freeman and Company, New York.
Soltis, D. E., and P. S. Soltis, 1989 Genetic consequences of autopolyploidy in *Tolmiea* (Saxifragaceae). Evolution **43:** 586–594.
Soltis, D. E., and P. S. Soltis, 1993 Molecular data and the dynamic nature of polyploidy. Crit. Rev. Plant Sci. **12:** 243–273.
Stebbins, G. L., 1971 *Chromosomal Evolution in Higher Plants.* Addison-Wesley, Reading, MA.
Stebbins, G. L., 1985 Polyploidy, hybridization and the invasion of new habitats. Ann. MO Bot. Gard. **72:** 824–832.
Tachida, H., 1985 Joint frequencies of alleles determined by separate formulation for the mating and mutation systems. Genetics **111:** 963–974.
Tachida, H., and H. Yoshimaru, 1996 Genetic diversity in partially selfing populations with the stepping-stone structure. Heredity **77:** 469–475.
Thompson, J. D., and R. Lumaret, 1992 The evolutionary dynamics of polyploid plants: origins, establishment and persistence. Trends Ecol. Evol. **7:** 302–307.
Waller, D. M., and S. E. Knight, 1989 Genetic consequences of outcrossing in the cleistogamous annual, *Impatiens capensis.* II. Outcrossing rates and genotypic correlations. Evolution **43:** 860–869.
Weir, B. S., 1996 *Genetic Data Analysis II: Methods for Discrete Population Genetic Data.* Sinauer Associates, Sunderland, MA.
Weir, B. S., and C. C. Cockerham, 1984 Estimating *F*-Statistics for the analysis of population structure. Evolution **38:** 1358–1370.
Wright, S., 1938 The distribution of gene frequencies in populations of polyploids. Proc. Natl. Acad. Sci. USA **24:** 372–377.
Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Communicating editor: M. Slatkin

APPENDIX

Computation of expected sum of squares of gene frequencies involved in estimating ρ and *F*-statistics: Let E_u denote the expectation of $x_{ijk:u}$, and π_u, the expected frequency of the allele u. Then $\varepsilon[(x_{ijk:u} - E_u)^2] = \pi_u - \pi_u^2$, where ε denotes expectation. Then, summing over all alleles, we obtain

$$\varepsilon_{ijk} = \varepsilon\left[\sum_u (x_{ijk:u} - E_u)^2\right] = 1 - \sum_u \pi_u^2 = 1 - Q_3, \qquad \text{(A1)}$$

where Q_3 denotes the identity probability for genes from different independent replicate populations, then, in the following, we write $[x_{ijk} - E]$ for the sum over alleles. Using the relationship $(x_k - x_{k'})^2 = (x_k - E)^2 + (x_{k'} - E)^2 - 2 \cdot (x_k - E) \cdot (x_{k'} - E)$, we obtain a useful equation for the covariance of two genes, *i.e.*,

$$\varepsilon[(x_k - E)(x_{k'} - E)] = \varepsilon[(x_k - E)^2] - \varepsilon[(x_k - x_{k'})^2]/2. \qquad \text{(A2)}$$

This, derived for different pairs of genes, yields the covariances

$$\varepsilon[(x_{ijk} - E)(x_{ijk'} - E)] = (1 - Q_3) - (1 - Q_0) = Q_0 - Q_3 \qquad \text{(A3)}$$

for genes within individuals,

$$\varepsilon[(x_{ijk} - E)(x_{ij'k'} - E)] = (1 - Q_3) - (1 - Q_1) = Q_1 - Q_3 \qquad \text{(A4)}$$

for genes between individuals within subpopulations, and

$$\varepsilon[(x_{ijk} - E)(x_{i'j'k'} - E)] = (1 - Q_3) - (1 - Q_2) = Q_2 - Q_3 \qquad \text{(A5)}$$

for genes between individuals in different subpopulations. These relationships can be used to derive the expectations

$$\varepsilon_{ij.} = \varepsilon[(x_{ij.} - E)^2] = \frac{1}{16}\varepsilon\left[\sum_{k=1}^{4}(x_{ijk} = E)^2 + \sum_{k=1}^{4}\sum_{k \neq k'}(x_{ijk} - E)(x_{ijk'} - E)\right],$$

i.e.,

$$\varepsilon_{ij.} = (1 - Q_3) - \frac{3}{4} \cdot (1 - Q_0)$$

$$\varepsilon_{i..} = \varepsilon[(x_{i..} - E)^2]$$
$$= \frac{1}{16n_i}\varepsilon\Bigg[\sum_j^{n_i}\sum_k^4 (x_{ijk} - E)^2 + \sum_j^{n_i}\sum_k^4\sum_{k\neq k'} (x_{ijk} - E)\cdot(x_{ijk'} - E) + \sum_j^{n_i}\sum_{j\neq j'}\sum_k^4\sum_{k\neq k'} (x_{ijk} - E)\cdot(x_{ij'k'} - E)\Bigg], \quad \text{(A6)}$$

i.e.,

$$\varepsilon_{i..} = (1 - Q_3) - \frac{n_i - 1}{n_i}(1 - Q_1) - \frac{3}{4n_i}(1 - Q_0)$$

$$\varepsilon_{...} = \varepsilon[(x_{...} - E)^2] = \varepsilon\left[\left(\frac{1}{4S_1}\sum_i\sum_j\sum_k x_{ijk} - E\right)^2\right], \quad \text{(A7)}$$

i.e.,

$$\varepsilon_{...} = (1 - Q_3) - (1 - \frac{S_2}{S_1^2})\cdot(1 - Q_2) - \left(\frac{S_2 - S_1}{S_1^2}\right) \cdot (1 - Q_1) - \frac{3}{4S_1}(1 - Q_0), \quad \text{(A8)}$$

where $S_1 = \Sigma_i n_i$ and $S_2 = \Sigma_i n_i^2$.

Now, the basic relationship $\varepsilon[\Sigma_i^r w_i(x_i - x)^2] = \varepsilon[\Sigma_i^r w_i (x_i - E)^2] - \varepsilon[\Sigma_i^r w_i (x - E)^2]$ can be used to write sum of squares expectations, for genes within individuals,

$$\varepsilon(SS_{g[enes]}) = \varepsilon\left[\sum_i^r\sum_j^{n_i}\sum_k^4 (x_{ijk} - x_{ij.})^2\right]$$
$$= \varepsilon\left[\sum_i\sum_j\sum_k (x_{ijk} - E)^2\right]$$
$$- \varepsilon\left[\sum_i\sum_j\sum_k (x_{ij.} - E)^2\right]$$

and using (A6) and (A7), we obtain

$$\varepsilon(SS_{g[enes]}) = 4S_1\varepsilon_{ijk} - 4S_1\varepsilon_{ij.} = 3S_1(1 - Q_0). \quad \text{(A9)}$$

Following the same procedure and denoting $W_d \equiv S_1 - r$, $W_a \equiv S_1 - S_2/S_1$, and $W_w \equiv r - 1$, we find the following sum of squares expectations:

for genes between individuals within subpopulations (using A6 and A7),

$$\varepsilon(SS_{i[ndivis]}) = 4S_1\varepsilon_{ij.} - 4\sum_i n_i\varepsilon_{i..}$$
$$= W_d \cdot (4(Q_0 - Q_1) + (1 - Q_0)); \quad \text{(A10)}$$

for genes between individuals from different subpopulations (using A7 and A8),

$$\varepsilon(SS_{s[ubpops]}) = 4\sum_i n_i\varepsilon_{i..} - 4S_1\varepsilon_{...}$$
$$= 4W_a \cdot (Q_1 - Q_2) + W_w \cdot [4(Q_0 - Q_1) + (1 - Q_0)]. \quad \text{(A11)}$$

As for diploids (Cockerham and Weir 1987), the components of variance of the nested ANOVA model ($x_{ijk:u} = \mu_u + \alpha_{i:u} + \beta_{j:u} + \varepsilon_{ijk:u}$) can be expressed as linear functions of identity probabilities, *i.e.*,

$$\sigma_a^2 \equiv Q_{1:u} - Q_{2:u} = (1 - Q_{2:u})\, F_{ST}, \quad \text{(A12)}$$

$$\sigma_b^2 \equiv Q_{0:u} - Q_{1:u} = (1 - Q_{2:u})(F_{IT} - F_{ST}), \quad \text{(A13)}$$

and

$$\sigma_e^2 \equiv 1 - Q_{0:u} = (1 - Q_{2:u})(1 - F_{IT}). \quad \text{(A14)}$$

ANOVA framework for the estimation of ρ and *F*-statistics: To compute sum of squares, straight way gene frequencies were used instead of indicator variables (x_{ijk}). This method is based on the following relationships between gene frequency estimates and the indicator variable [see Weir (1996) for the diploid case],

$$\bar{x}_{...} = \sum_i n_i\tilde{p}_{A_i}/\sum_i n_i = \tilde{p}_A$$

$$\bar{x}_{i..} = \sum_j\sum_k x_{ijk}/4n_i = \tilde{p}_{A_i}$$

$$\bar{x}_{ij.}^2 = \varepsilon(\sum_k x_{ijk})^2/16 = (\tilde{p}_{A_i} + 3\tilde{P}_{AA_i})/4,$$

where $\tilde{p}_{A_i} = \Sigma_j\Sigma_k x_{ijk}/4n_i$ and $\tilde{P}_{AA_i} = \tilde{P}_{0,i} + \tilde{P}_{1,i}/2 + \tilde{P}_{,2;i}/6$ with $P_{0,i}$, $\tilde{P}_{1,i}$, and $\tilde{P}_{2,i}$ standing, respectively, for the proportion of monogenic (AAAA), trigenic (AAAa), and digenic (AAab) individuals in the *i*th population (Malécot 1948). These relationships yield more convenient expressions for the variance components of the analysis, as shown in Table 1.

TABLE 1

Nested analysis of variance layout for estimation of the variance components of population structure in autotetraploid organisms, and corresponding gene frequency-based equations.

Source	d.f.	Sum of squares	Expected mean square[a]
Between populations	$r - 1$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{4}(\bar{x}_{i..} - \bar{x}_{...})^2 = 4\sum_{i=1}^{r} n_i(\tilde{p}_{Ai} - \tilde{p}_A)^2$	$p_A(1 - p_A)[(1 - F_{IT}) + 4(F_{IT} - F_{ST}) + 4n_c F_{ST}] = \sigma_e^2 + 4\sigma_b^2 + 4n_c\sigma_a^2$
Individuals in populations	$\sum_{i=1}^{r}(n_i - 1) = n. - r$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{4}(\bar{x}_{ij.} - \bar{x}_{xi..})^2 = \sum_{i=1}^{r} n_i(\tilde{p}_{Ai} + 3\tilde{P}_{AA_i} - 4\tilde{p}_{Ai}^2)$	$p_A(1 - p_A)[(1 - F_{IT}) + 4(F_{IT} - F_{ST})] = \sigma_e^2 + 4\sigma_b^2$
Genes in individuals	$\sum_{i=1}^{r} n_i(b - 1) = 3n.$	$\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{4}(\bar{x}_{ijk} - \bar{x}_{ij.})^2 = 3\sum_{i=1}^{r} n_i(\tilde{p}_{A_i} - \tilde{P}_{AA_i})$	$p_A(1 - p_A)(1 - F_{IT}) = \sigma_e^2$

This table directly follows a two-way nested ANOVA (see Sokal and Rohlf 1995). Corresponding data design: r populations ($1 \leq i \leq r$) of sample size n_i ($1 \leq j \leq n_i$), and b genes registered for each individual ($1 \leq k \leq 4$).

[a] $n_c = 1/(r - 1)\,[\Sigma_{i=1}^{r} n_i - \Sigma_{i=1}^{r} n_i^2 / \Sigma_{i=1}^{r} n_i]$.