

Cuauhtemoc Cervantes-Martinez · J. Steven Brown ·
Raymond Schnell · Juan C. Motamayor ·
Alan W. Meerow · Dapeng Zhang

A computer simulation study on the number of loci and trees required to estimate genetic variability in cacao (*Theobroma cacao* L.)

Received: 9 September 2005 / Revised: 21 November 2005 / Accepted: 30 January 2006
© Springer-Verlag 2006

Abstract Current methods for measures of genetic diversity of populations and germplasm collections are often based on statistics calculated from molecular markers. The objective of this study was to investigate the precision and accuracy of the most common estimators of genetic variability and population structure, as calculated from simple sequence repeat (SSR) marker data from cacao (*Theobroma cacao* L.). Computer simulated genomes of replicate populations were generated from initial allele frequencies estimated using SSR data from cacao accessions in a collection. The simulated genomes consisted of ten linkage groups of 100 cM in length each. Heterozygosity, gene diversity and the *F* statistics were studied as a function of number of loci and trees sampled. The results showed that relatively small random samples of trees were needed to achieve consistency in the observed estimations. In contrast, very large random samples of loci per linkage group were required to enable reliable inferences on the whole genome. Precision of estimates was increased by more than 50% with an increase in sample size from one to five loci per linkage group or 50 per genome, and up to 70% with ten loci per linkage group, or equivalently, 100 loci per genome. The use of fewer,

highly polymorphic loci to analyze genetic variability led to estimates with substantially smaller variance but with an upward bias. Nevertheless, the relative differences of estimates among populations were generally consistent for the different levels of polymorphism considered.

Introduction

Genetic diversity of plant species is a major concern among geneticists and plant breeders involved in preserving genetic variation. Current methods for measuring genetic diversity of populations and germplasm collections are often statistics calculated from molecular marker data (Mohammadi and Prasanna 2003). Economic and practical restrictions on the study of large portions of the genome have commonly led researchers to characterize the genetic variation of populations with relatively few molecular markers. Numerous studies have been reported in plant species using different-sized samples of individuals and marker density (Ni et al. 2002; Liu et al. 2003; Reif et al. 2004; Ahmad et al. 2003; Gao et al. 2004). Often, marker loci are selected on the number and frequency of alleles observed in preliminary screenings. The specific effect of the number of individuals and markers on the sample size on several measures of genetic distance has been recently studied by Kalinowski (2002a,b, 2005), using computer simulation methods. These studies yielded valuable information on the optimal sample size of individuals and loci in linkage equilibrium at differing levels of polymorphism, with several different divergence times, mutation rates, and models. However, his studies did not cover optimal marker density in situations in which loci are in linkage disequilibrium caused by physical linkage or reproduction of few individuals in a population for several generations. Also, the simulated population sizes and number of generations of random mating were considerably large to be representative for perennial plant species such as fruit trees; and finally, the numbers of markers studied were substantially small to consider the patterns of genetic distance estimates as representative of genomic values.

C. Cervantes-Martinez (✉) · J. S. Brown · R. Schnell ·
A. W. Meerow
United States Department of Agriculture-Agriculture
Research Service (USDA-ARS),
The Subtropical Horticulture Research Station (SHRS),
Miami, FL 33158, USA
e-mail: ccervantes@saa.ars.usda.gov

J. C. Motamayor
Mars Inc.,
c/o United States Department of Agriculture-
Agriculture Research Service (USDA-ARS),
The Subtropical Horticulture Research Station (SHRS),
Miami, FL 33158, USA

D. Zhang
United States Department of Agriculture-Agriculture
Research Service (USDA-ARS),
Beltsville, MD 20705, USA

In practice, genetic diversity estimates vary widely among loci, given the differential polymorphism patterns existing in the genome (Fuerst et al. 1977; Nei 1987; Weir 1996). Thus, it has been proposed that the mean and variance of any diversity measure can only be estimated accurately by considering the entire genome (Nei 1987). The distributional and statistical properties of the diversity estimators are often not well-known, but theoretical development has shown that their variances are highly dependent on the number of loci sampled, linkage disequilibrium, and the number of individuals sampled (Weir 1996). However, the effect of these factors on the precision of estimates for specific populations and sub-populations of plant species is yet unknown.

Cacao (*Theobroma cacao* L.) is a perennial fruit tree whose beans are the basis for chocolate production. Cacao has a small chromosome number ($n=x=10$), and a relatively small genome, approximately 2.8 times the size of the genome of *Arabidopsis thaliana* (L.) Heynh. (Lanaud et al. 1992; Couch et al. 1993; Brown et al. 2005). The use of molecular markers has been successful in delimiting genetic types (breeding groups), identifying incorrectly labeled samples, have given insight into relative genetic variability of groups, and in verifying biological hypotheses. The genetic structure and relationships of recognized breeding types, Forastero, Criollo, Trinitario, and Nacional were studied with molecular markers by Laurent et al. (1994) and later by Lerceteau et al. (1997), and their studies generally confirmed traditional breeding knowledge. The origin of the Trinitario type was specified to have been a cross between one sole Forastero type and Ancient Criollos. The genetic variability of the Criollo type is reported to be very low, while a larger amount of genetic variability is reported for the Forastero type (Laurent et al. 1994, Motamayor et al. 2003; Clement et al. 2003). It has since been recognized that there are serious problems in the mislabeling of clones in germplasm collections and breeding material. Estimates of between 15 to 44% mislabeling have been observed in international germplasm collections (Motilal and Butler 2003; Turnbull et al. 2004). Although previous studies have given substantial insight into cacao germplasm and have confirmed traditional breeding groups, more accurate estimators would help to determine genetic relationships among groups of clones from breeding programs or wild populations originating from small groups of founders and that are closely related. Precise and accurate estimates of genetic diversity are critical for international cacao breeding projects; therefore, sufficiently high marker density and tree sample sizes pose key questions. A high level of precision is necessary to guide marker and germplasm sampling, especially when introgression of incoming germplasm is involved, and also when variability within populations is larger than that existing among populations, as can arise after several generations of reproduction by seedlings and farmer selection.

The objectives of this research were to study the accuracy and precision of (a) observed heterozygosity, (b) gene diversity, and (c) the F statistics (Nei 1978, 1987;

Weir and Cockerham 1984; Weir 1996), by varying marker density and number of trees sampled, using computer simulation. The genome structure, effective population size, and number of generations of random mating were established in the simulated populations as previous research has described the genetic structure of cacao (*Theobroma cacao* L.) (Lanaud et al. 1995; Motamayor et al. 2002, 2003). We discuss optimal numbers of markers and trees in this context considering relevant principles of statistical and genetic drift theory.

Materials and methods

General description of methods

Cacao accessions, in reality, consist of clonal cultivars and clones selected from crosses, cultivated populations, or from the wild, that are not generally in Hardy–Weinberg equilibrium (HWE). Most of these populations (breeding groups) are relatively small, started from a reduced number of trees, and have few generations of intermating. Estimates from molecular mapping data have demonstrated as much as 25% of the genome being affected by linkage disequilibrium (Brown et al. 2005) and evidence of founder effect has been shown for the Trinitario group (Motamayor et al. 2003). Therefore, we followed a forward-in-time simulation approach in a classical population genetics scheme, as this is considered an appropriate method for studies of genetic systems in which the middle and long term effects are dependent on initial conditions (Rosenberg and Nordborg 2002).

We simulated populations that diverged from HWE due to drift by first simulating equilibrium populations and then randomly sampling founders of finite populations from two larger reference populations. In this case, each of three populations developed from both sets of founders is considered as a replicate population from the same reference population (Weir 1996). Next, the simulated finite subpopulations were random-mated, subject to drift and mutation for 50 generations; this would constitute an upper limit of random mating for most cacao populations. Finally, samples of trees were drawn from the resulting subpopulations for estimation of genetic diversity. Thus, variation was introduced into the simulations at several points. First, two reference populations were generated using different genetic parameter combinations. Second, three different founder populations were sampled from each reference population. Third, individual trees were sampled from subpopulations developed from each founder population. Lastly, loci were sampled from the entire genome to estimate genetic diversity.

Microsatellite data

The clones from which initial molecular marker data originated, consisted of 55 cacao accessions chosen from a random sample of 96 cacao clones from the germplasm

collection of the Tropical Agricultural Research and Training Center (CATIE), Turrialba, Costa Rica. Accessions were classified into two groups based on information about their genetic origin: group 1 consisted of ten accessions classified as Trinitario type, and group 2 consisted of 45 accessions classified as Trinitario type introgressed with Upper Amazon type.

Molecular marker analyses were performed at the United States Department of Agriculture-Agricultural Research Service (USDA-ARS) Crop Systems and Global Change Laboratory, Beltsville, MD, using 15 simple sequence repeat (SSR) markers from those described previously by Lanaud et al. (1999) and Risterucci et al. (2000). SSR amplification products were separated and scored using a CEQ8000 eight channel capillary DNA genetic analysis system (Beckman Coulter, Fullerton, CA) as described in Haymes et al. (2004). Principal Coordinate Analysis (Gower 1966) and Average Linkage Clustering were performed on the SSR data of the 55 samples from the two groups based on modified Roger's distance (Wright 1978). All members of the two groups clustered accurately (data not presented).

Data simulation—generation of reference populations

Two large base populations in HWE and linkage equilibrium were created for this study. The allele configuration across the genome was generated using data obtained from observed SSR loci in the sample of genotyped cacao accessions. The first base population was created using molecular marker information from accessions in group 1, known to contain a relatively small amount of genetic variability. The second base population was created from marker data from accessions corresponding to group 2, with a larger amount of genetic variability.

The simulated genome of each tree consisted of ten linkage groups ($2n=20$), corresponding to the map of Lanaud et al. (1995), each linkage group being set to 100 cM in length for computational simplicity in the simulation. A locus of neutral effect was placed each 1 cM, with a 50% probability of being polymorphic. The number of alleles per polymorphic locus was assigned randomly, varying from two to five for the first base population and from six to 13 for the second base population, similar to the molecular marker data used in the creation of each population.

Data simulation—generation of founders and reproduction of subpopulations

Alleles were assigned to loci on homologues by randomly sampling from a multinomial distribution. The allelic frequencies used corresponded to observed allele frequency distributions in the marker data. Each multilocus homologue then constituted a haplotype. Founder genotypes were created by randomly pairing haplotypes.

Each set of founders from the identical reference population were simulated using the corresponding polymorphic loci and allele frequency distributions. Each set of founders was random-mated by computer simulation to form progeny populations. Subsequent progeny were reproduced by forward simulation for 50 nonoverlapping generations, using a scheme similar to that described by Pálsson and Pamilo (1999) and MacLeod et al. (2005), described below.

Recombination probabilities between homologues were derived using a Poisson process, with the assumptions of no interference among loci (Haldane 1919) and recombination sites occurring at random. Homologues of each linkage group were randomly assigned to parental haploid genotypes, which were represented from one generation to next by their reproductive values obtained from a binomial distribution, with parameters $2N_t$ =number of haplotypes in generation t , and $p = (2N_t)^{-1}$. Haploid genomes were randomly paired to form diploids. Two sets were produced, each with three replicated populations. The two sets, sets 1 and 2, being derived from the first and second reference populations, corresponded to groups 1 and 2, respectively. The effective population size of trees (N_e) varied randomly from generation to generation, with a long-term effective size across generations (or harmonic mean across generations) of 201, 215, and 196 trees for populations 1, 2, and 3 of set 1, and 196, 174, and 217 trees for populations 1, 2, and 3 of set 2. The harmonic mean of the effective population size over generations (N_e^*) was 204 for set 1 and 194 for set 2. Neutral mutations were simulated in this research with the infinite-allele model (IAM; Kimura and Crow 1964). The number of mutations accumulated was obtained as a random draw from a Poisson distribution with parameter, $50\lambda\nu$, where the number 50 corresponds to the number of generations, λ is the chromosome length (cM) and ν is the mutation rate (1×10^{-3}). The position in the linkage group at which each mutation took place was assigned at random.

Data simulation—sampling trees and loci

Trees were randomly sampled from each of the three simulated replicate populations from both genetic groups, and loci were randomly sampled from each linkage group (Table 1). Fifty replications were drawn at each level of sampling of trees and loci in each population, giving a total of 2,500 observations for each level of sampling and a total of 7,500 observations over the three replications. Genetic variability estimates were calculated for each level of sampling of loci and trees. These estimates were then used to calculate empirical summary statistics of genetic variability estimators. An additional factor was investigated, in that, loci were sampled under two conditions; with and without restriction on the polymorphism level, i.e., selecting loci (markers) for those with higher polymorphic levels, or using no selection and including monomorphic markers. The two types of locus sampling were then

investigated for their potential effects on diversity estimates. Selection of loci with restriction consisted of sorting loci by level of polymorphism and selecting only those above a certain level. Summary statistics of diversity estimators were then calculated from both locus sampling methods.

Estimation of genetic variability in samples

Gene diversity, observed heterozygosity, and the F statistics were estimated by the linear model approach (Weir 1996).

Heterozygosity The linear model associated with the heterozygous state is given by:

$$x_{ijl} = \mu + \alpha_i + \beta_{ij} + \gamma_l + \delta_{il} + \varepsilon_{ijl}; \quad (1)$$

for $i = 1, 2, \dots, I; j = 1, 2, \dots, J_i; l = 1, 2, \dots, L$.

Here, x_{ijl} is an indicator variable that takes the value of 1 if the individual, j , of population, i , is heterozygous for locus, l , and 0 otherwise; α_i is the effect of population i ; β_{ij} is the effect of individual, j , of population, i ; γ_l is the effect of locus, l ; δ_{il} is the effect of the interaction of population, i , with locus, l ; and ε_{ijl} is the effect of the interaction of individual, j , of population, i , with locus, l . This model was analyzed first by incorporating the uncertainty due to sampling individual genomes with a specific set of loci, rather than all genomes of the population, and considering populations as a fixed factor in the linear model 1. Heterozygosity was estimated as the least squares mean for each locus, $\hat{H}_{il} = \hat{\mu} + \hat{\alpha}_i + \hat{\gamma}_l + \hat{\delta}_{il}$. The values $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\gamma}_l$ and $\hat{\delta}_{il}$ represent the solutions to the normal equations (Searle 1971).

The uncertainty introduced by sampling subpopulations from a larger population or species, as opposed to studying all subpopulations of the entire population or species, was considered in the second statistical analysis treating populations as a random factor. In this case, each population was considered as a random replication of an evolutionary process from a single ancestral population (Weir 1996). The observed heterozygosity was estimated as $\hat{H}_l = \hat{\mu} + \hat{\gamma}_l$. Here, $\hat{\mu}$ and $\hat{\gamma}_l$ are the solutions to the

normal equations (Searle 1971). Heterozygosity was also expressed as an average over loci within subpopulations as $\hat{H}_i = \hat{\mu} + \hat{\alpha}_i + \frac{1}{L} \sum_{l=1}^L (\hat{\gamma}_l + \hat{\delta}_{il})$ and over all subpopulations $\hat{H} = \hat{\mu} + \frac{1}{L} \sum_{l=1}^L \hat{\gamma}_l$, for the fixed and random models, respectively.

Gene diversity Gene diversity and variance components calculated from allele frequencies were considered in this study as follows: let a_{lu} denote allele u ($u=1, 2, \dots, U_l$) of locus l , and y_{ijlur} , an indicator variable that takes the value of 1 if allele, r , at locus, l , of individual, j , of population, i , is type a_{lu} and zero otherwise. The linear model associated with the frequency of the allele a_{ijlur} is

$$y_{ijlur} = \mu_{lu} + \alpha_{ilu} + \beta_{ijlu} + \delta_{ijlur};$$

for $i = 1, 2, \dots, I; j = 1, 2, \dots, J_i; l = 1, 2, \dots, L; r = 1, 2$. (2)

Where μ_{lu} is the general mean of the allele frequency; α_{ilu} is the effect of population, i ; β_{ijlu} is the effect of individual (tree), j , nested in population, i , and δ_{ijlur} is the effect of homologue, r (Cockerham 1969). The frequency of allele, a_{lu} , was estimated as the least squares means $\hat{p}_{ilu} = \hat{\mu}_{lu} + \hat{\alpha}_{ilu}$ (within subpopulation i) and $\hat{p}_{lu} = \hat{\mu}_{lu}$ (across all subpopulations). Gene diversity estimators for fixed and random models were calculated as $\hat{D}_i = \frac{1}{L} \sum_{l=1}^L \left(1 - \frac{1}{2J_i}\right)^{-1} \left(1 - \sum_{u=1}^{U_l} \hat{p}_{ilu}^2\right)$ and $\hat{D} = \frac{1}{L} \sum_{l=1}^L \left(1 - \frac{1}{2J_l}\right)^{-1} \left(1 - \sum_{u=1}^{U_l} \hat{p}_{lu}^2\right)$, with $0 < \hat{p}_{ilu} < 1$ and $0 < \hat{p}_{lu} < 1$, respectively, and equal to 0 for $\hat{p}_{ilu} = 1$ and $\hat{p}_{lu} = 1$. The estimators \hat{D}_i and \hat{D} are unbiased when alleles are not correlated within and among populations (Weir 1996).

F statistics The F statistics consist of three parameters: F_{IS} , F_{IT} and F_{ST} , that indicate the degree by which the subdivision of a population in random drift decreases the

Table 1 Simulated sample sizes of trees and loci

Sampled trees per population	Sampled loci per linkage group	Polymorphic loci per linkage group ^a (%)
20, 40, 60, 80, 100	1, 2, 5, 10, 20, 40, 60, 80, 100	NA
100	1, 2, 5	5
100	1, 2, 5, 10	10
100	1, 2, 5, 10, 20	20
100	1, 2, 5, 10, 20, 40	40
100	1, 2, 5, 10, 20, 40, 60	60
100	1, 2, 5, 10, 20, 40, 60, 80	80
100	1, 2, 5, 10, 20, 40, 60, 80, 100	100

^aLoci were sorted in descending order by gene diversity, and only the indicated percentage of loci with the largest values were considered to sample for estimation of descriptive statistics

NA Not applicable

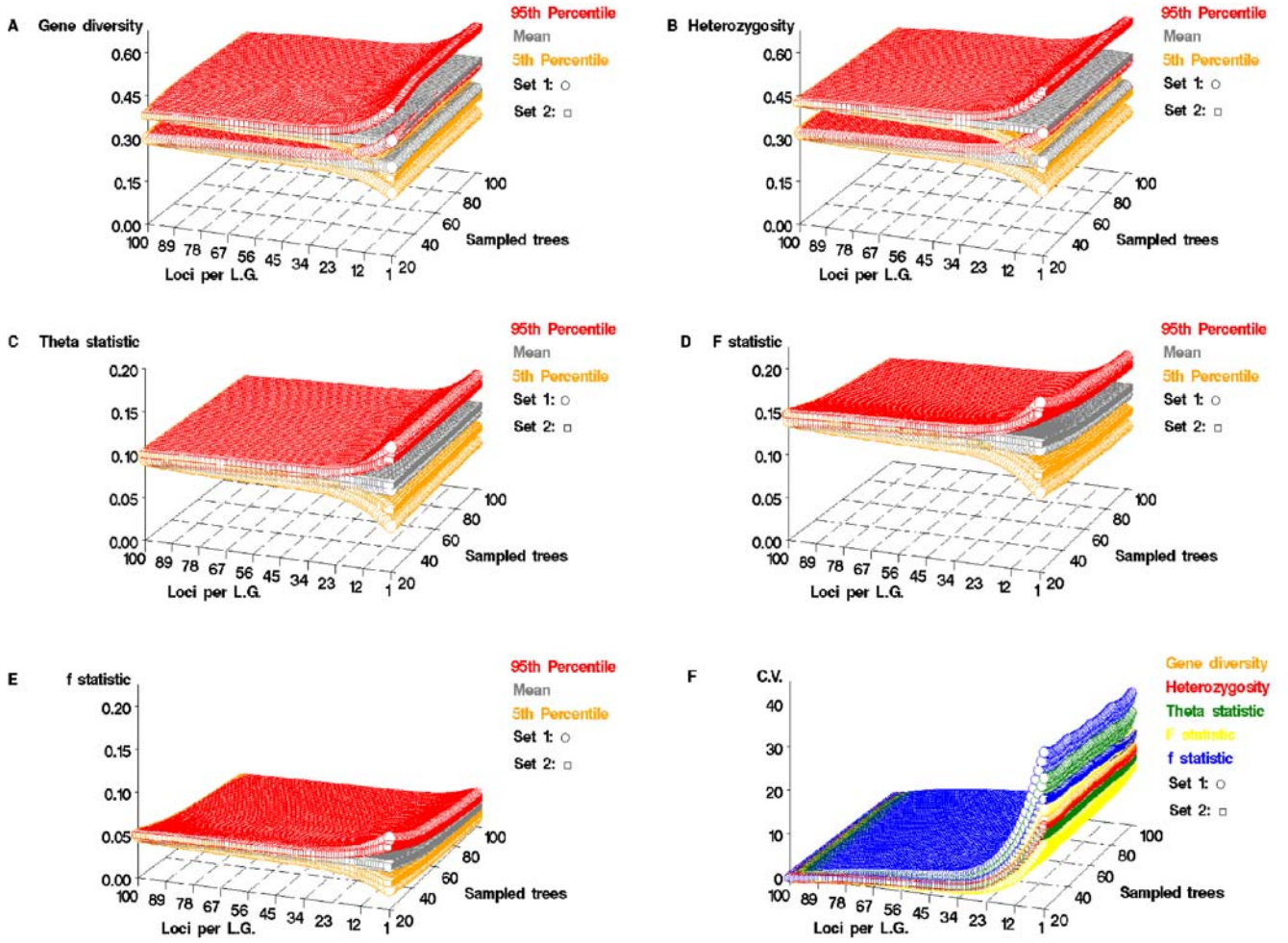


Fig. 1 Tridimensional scatter plots of genetic variability estimators plotted against sample sizes of loci and trees. Calculations were performed with the random model, considering trees and popula-

tions of each set as random. Interpolation among data points was made with a spline smoothing method (Meinguet 1979)

expected heterozygosity. F_{IS} and F_{IT} measure the correlation between uniting gametes respective to subpopulations and the entire population, respectively, and F_{ST} is the correlation between two random gametes from different subpopulations. Cockerham (1969) presented analogous parameters to the F statistics: F , θ , and f , that can be expressed as functions of the components of variance of frequencies of neutral genes. The relationship between both sets of parameters is given as $F=F_{IT}$, $\theta=F_{ST}$, and $f=F_{IS}$ (Weir and Cockerham 1984). Then, using Cockerham's notation, the F statistics were estimated as

$$\hat{f}_{il} = \frac{\sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{ilu}}^2}{\sum_{u=1}^{U_l} (\hat{\sigma}_{\beta_{ilu}}^2 + \hat{\sigma}_{\delta_{ilu}}^2)} \quad \text{and} \quad \hat{f}_i = \frac{\sum_{l=1}^L \sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{ilu}}^2}{\sum_{l=1}^L \sum_{u=1}^{U_l} (\hat{\sigma}_{\beta_{ilu}}^2 + \hat{\sigma}_{\delta_{ilu}}^2)} \quad \text{for the}$$

$$\text{fixed model, and } \hat{f}_l = \frac{\sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{lu}}^2}{\sum_{u=1}^{U_l} (\hat{\sigma}_{\beta_{lu}}^2 + \hat{\sigma}_{\delta_{lu}}^2)}, \quad \hat{f} = \frac{\sum_{l=1}^L \sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{lu}}^2}{\sum_{l=1}^L \sum_{u=1}^{U_l} (\hat{\sigma}_{\beta_{lu}}^2 + \hat{\sigma}_{\delta_{lu}}^2)},$$

$$\hat{F}_l = \frac{\sum_{u=1}^{U_l} (\hat{\sigma}_{\alpha_{lu}}^2 + \hat{\sigma}_{\beta_{lu}}^2)}{\sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{lu}}^2}, \quad \hat{F} = \frac{\sum_{l=1}^L \sum_{u=1}^{U_l} (\hat{\sigma}_{\alpha_{lu}}^2 + \hat{\sigma}_{\beta_{lu}}^2)}{\sum_{l=1}^L \sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{lu}}^2}, \quad \hat{\theta}_l = \frac{\sum_{u=1}^{U_l} \hat{\sigma}_{\alpha_{lu}}^2}{\sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{lu}}^2},$$

$$\text{and } \hat{\theta} = \frac{\sum_{l=1}^L \sum_{u=1}^{U_l} \hat{\sigma}_{\alpha_{lu}}^2}{\sum_{l=1}^L \sum_{u=1}^{U_l} \hat{\sigma}_{\beta_{lu}}^2} \quad \text{for the random model, with } \hat{\sigma}_{\beta_{lu}}^2 = \hat{\sigma}_{\alpha_{lu}}^2$$

+ $\hat{\sigma}_{\beta_{lu}}^2 + \hat{\sigma}_{\delta_{lu}}^2$. The values $\hat{\sigma}_{\beta_{ilu}}^2$ and $\hat{\sigma}_{\delta_{ilu}}^2$ are the moment estimators of the variance components for the fixed model, and $\sigma_{\beta_{lu}}^2$, $\sigma_{\delta_{lu}}^2$, and $\sigma_{\alpha_{lu}}^2$ are the components for the random model, both corresponding to the linear representation in Eq. 2.

The estimators \hat{D}_i , \hat{D} , \hat{H}_i , and \hat{H} were calculated considering both polymorphic and monomorphic loci. The F statistics, however, were calculated only for polymorphic loci, as these estimators are undefined for monomorphic loci. Fixed and random models were analyzed for each of the two sets of populations.

Table 2 Lengths of the percentile intervals containing the central 90% of estimates, presented as percentages with respect to the interval length calculated from the smallest locus sample size (bold)^a

Locus or loci per linkage group		Sample size																								
		Trees per population				40				60				80				100								
		Genetic variability estimator				Genetic variability estimator				Genetic variability estimator				Genetic variability estimator				Genetic variability estimator								
		GD	H	θ	F	f	GD	H	θ	F	f	GD	H	θ	F	f	GD	H	θ	F	f	GD	H	θ	F	f
1	1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		0.33	0.25	0.11	0.13	0.07	0.32	0.25	0.10	0.11	0.05	0.32	0.25	0.10	0.11	0.04	0.30	0.24	0.10	0.10	0.04	0.30	0.24	0.10	0.10	0.03
	2	68	70	69	69	72	69	67	70	71	71	67	67	68	67	72	70	69	70	69	69	72	69	70	69	71
	5	43	44	43	42	44	44	43	44	44	43	42	43	43	42	45	45	45	44	44	43	45	44	44	44	44
	10	30	30	30	29	30	29	29	30	30	30	29	29	29	29	31	30	31	29	29	29	30	30	28	29	31
	20	20	20	19	19	20	20	19	19	19	20	20	20	20	20	20	20	20	20	20	18	20	20	19	19	21
	40	12	12	12	12	12	12	12	12	12	13	12	12	12	12	12	12	13	13	12	12	12	12	12	12	12
	60	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	8	8	8	8	8	8	8	8	8	8
	80	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		0.35	0.33	0.08	0.09	0.05	0.37	0.34	0.07	0.08	0.04	0.36	0.33	0.07	0.08	0.03	0.36	0.32	0.07	0.07	0.03	0.37	0.34	0.07	0.07	0.02
	2	71	71	70	69	71	70	70	68	70	71	72	72	67	67	67	71	71	71	71	71	69	69	71	69	68
	5	45	46	42	42	43	44	44	43	44	45	45	46	41	42	43	46	46	43	43	44	44	44	42	42	43
	10	31	31	29	28	28	30	30	29	31	30	30	30	29	29	28	33	33	30	30	31	30	30	29	28	29
	20	20	20	19	19	20	20	20	20	20	20	21	21	19	19	19	21	21	20	20	20	19	19	20	19	19
	40	12	12	12	12	12	12	13	13	12	13	12	12	12	12	12	13	13	13	12	12	12	12	12	12	12
	60	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
	80	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

^aThe percentages were calculated relative to the actual length of the intervals (bold) obtained with the smallest locus sample size (one locus per L.G.). The length of the intervals can be calculated by multiplying the cell percentage by the corresponding initial length (bold) and dividing by 100

^bSet 1 and 2 correspond to the populations derived from the reference population 1 and 2, respectively

GD gene diversity, H heterozygosity

The analyses correspond to the random model

Summary statistics of simulations Summary statistics for the genetic variability estimators were calculated to determine optimal sample sizes and to compare theoretical expected values to verify the reliability of the simulation. The empirical mean, coefficient of variation, the fifth and 95th percentiles were obtained from all replications of locus sampling levels. Their expected values were approximated as the average over replicates of tree sample sizes (Efron and Tibshirani 1993; Gilks et al. 1996). Percentile intervals were calculated as the difference between the expected values of the fifth and 95th percentiles.

Expected values of diversity measures To determine the reliability of the simulation process, expected values of the F statistics were compared to the expected theoretical values. The F statistics are related to one another by the

equality $\frac{(1-F)}{(1-\theta)} = (1-f)$, with particular values of $F=\theta$ and $f=0$ for monoecious populations in random mating (Weir 1996). Thus, the degree of inbreeding within populations, f , was compared to $\frac{F-\theta}{1-\theta}$, F was compared to θ , and the harmonic mean of F and θ was compared to the expected accumulated inbreeding, $F_t = [\frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1}](1 - \nu)^2$ (Kimura and Crow 1964), where $t=50$, $\nu=10^{-3}$, and N_e =the harmonic mean of the effective population size across years. The expected inbreeding from random drift and mutation at equilibrium was calculated as $\frac{1-2\nu}{4N_e\nu-2\nu+1}$ (Kimura and Crow 1964).

All computer simulations and statistical analyses were performed with a set of SAS macros for Windows Version 9.1.2 (SAS Institute, Cary, NC). The parameters of the linear models 1 and 2 were estimated with the MIXED

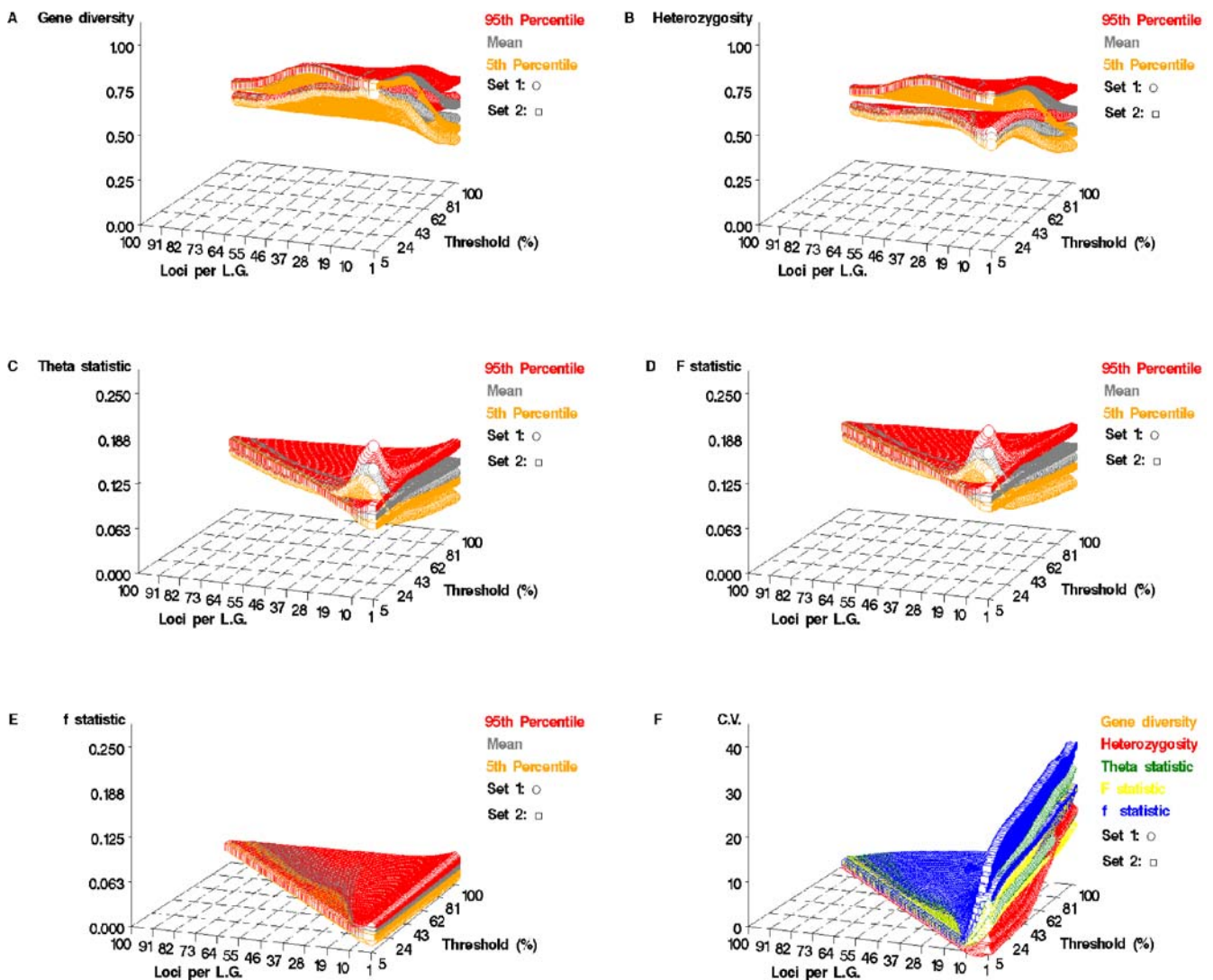


Fig. 2 Tridimensional scatter plots of genetic variability estimators plotted against sample sizes of loci and trees. Calculations were performed with the random model, considering trees and populations of each set as random. The locus sample space was bounded by

the indicated percentage (threshold) of loci with the largest gene diversity. Interpolation among data points plot were made with the spline smoothing method (Meinguet 1979)

Table 3 Lengths of the percentile intervals containing the central 90% of estimates, presented as percentages with respect to the interval length calculated from the smallest locus sample size (bold)^a

<i>Set</i> Locus or loci per linkage group		<i>Locus number</i> ^b (sample space)																			
		5					10					20					40				
		Estimator					Estimator					Estimator					Estimator				
		GD	H	θ	<i>F</i>	<i>f</i>	GD	H	θ	<i>F</i>	<i>f</i>	GD	H	θ	<i>F</i>	<i>f</i>	GD	H	θ	<i>F</i>	<i>f</i>
1	1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		.06	.09	.08	.08	.02	.06	.09	.08	.08	.02	.06	.08	.07	.08	.02	.11	.10	.07	.08	.02
		61	62	61	62	62	66	67	66	65	66	74	72	70	71	67	68	71	71	71	70
		0	0	0	0	0	33	33	33	33	33	41	40	40	40	41	43	42	43	43	42
							0	0	0	0	0	24	23	23	23	23	28	27	28	28	28
												0	0	0	0	0	16	16	16	16	16
																	0	0	0	0	0
2	1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		.01	.03	.03	.03	.01	.02	.04	.04	.04	.01	.03	.04	.04	.04	.02	.04	.06	.05	.05	.02
		60	62	62	62	60	69	69	66	67	69	69	71	72	73	69	72	70	73	69	67
		0	0	0	0	0	35	35	34	34	34	39	40	41	41	37	44	42	43	42	43
							0	0	0	0	0	23	23	23	24	23	30	28	28	27	28
												0	0	0	0	0	16	16	16	16	15
																	0	0	0	0	0
Set	Loci per linkage group	<i>Locus number</i> ^c																			
		60					80					100									
		Estimator					Estimator					Estimator									
		GD	H	θ	<i>F</i>	<i>f</i>	GD	H	θ	<i>F</i>	<i>f</i>	GD	H	θ	<i>F</i>	<i>f</i>					
		100	100	100	100	100	100	100	100	100	100	100	100	100	100	100					
		.24	.19	.08	.08	.03	.29	.23	.09	.09	.03	.31	.25	.10	.10	.03					
		67	67	69	69	70	71	70	72	70	73	69	70	68	70	69					
42	42	44	44	44	43	43	43	43	44	43	44	42	42	44							
28	27	29	29	28	31	30	30	30	30	29	30	29	30	29							
18	18	18	18	18	20	20	19	19	19	19	19	19	19	21							
9	9	9	9	9	12	12	11	11	11	12	12	12	12	12							
0	0	0	0	0	7	7	6	6	7	8	8	8	8	8							
					0	0	0	0	0	5	5	5	5	5							
										0	0	0	0	0							
2	1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		.24	.22	.05	.06	.02	.34	.31	.06	.07	.02	.37	.33	.07	.07	.02					
		70	70	70	70	74	72	71	69	69	71	69	69	69	70	70					
		43	42	42	42	45	45	45	43	43	44	44	44	43	42	43					
		29	29	28	28	30	31	31	29	29	30	30	30	29	30	30					
		18	18	18	18	19	20	20	19	19	20	20	20	20	20	20					
		9	9	9	9	9	7	7	11	11	11	12	12	12	12	12					
0	0	0	0	0	6	6	6	6	6	9	9	8	8	8							
					0	0	0	0	0	5	5	5	5	5							
										0	0	0	0	0							

^aThe percentages were calculated relative to the actual length of the intervals (bold) obtained with the smallest locus sample size (1 locus per L.G.). The length of the intervals can be calculated by multiplying the cell percentage by the corresponding initial length (bold) and dividing by one hundred

^bLoci considered with the largest gene diversity from which the random samples of loci were obtained

^cSet 1 and 2 correspond to the replicate populations derived from the reference population 1 and 2, respectively

GD gene diversity, *H* heterozygosity

The sampling of loci was performed among the most polymorphic. The analyses correspond to the random model

procedure using the method of moments (Littell et al. 1996). As a final step for ease of interpretation of the results, tridimensional scatter plots were produced using the nonparametric spline-smoothing method (Meinguet 1979), defining horizontal axes as sample sizes of loci and trees, and the vertical axis as genetic variability estimate.

Results

Random model

Unrestricted locus sample space Summary statistics of gene diversity, observed heterozygosity, and the F statistics for different combinations of tree and locus sample sizes are shown in Fig. 1. These estimates correspond to the case of no restriction on locus sampling. The means of gene diversity and heterozygosity showed little variation across levels of tree sampling for a given number of loci in both sets of populations. Larger mean values were observed for set 2 due to greater allelic diversity of their reference population.

The percentile intervals containing 90% of the estimates of gene diversity and heterozygosity varied markedly more with the number of loci sampled than with the number of trees (Fig. 1). These intervals are also presented in Table 2 as percentages obtained from the interval of the smallest locus sample. One locus per linkage group produced gene diversity and heterozygosity estimates contained in the smallest percentile intervals of length 0.24 for set 1 and 0.32 for set 2. Samples of five loci per linkage group increased the precision of estimates by reducing the interval by more than 50%, to 0.11 and 0.15, for sets 1 and 2, respectively, whereas samples of ten loci per linkage group (10% per genome) caused a reduction in the length of approximately 70%, to 0.07 and 0.11, respectively. Increasing the sample size, however, from ten to 20 loci per linkage group only reduced the interval lengths by an additional 10%. Further reduction of the percentile intervals of both statistics tended asymptotically towards zero, when approximately 25 or more loci per linkage group (250 loci per genome) were included in the analyses.

Regarding the F statistics, the means of F and f had a slight tendency to decrease and of θ to increase with larger samples of trees (Fig. 1), and these trends were mostly consistent over sizes of locus samples. The estimated mean values of F , θ , and f using all loci information and the largest samples of trees, were 0.11, 0.09, and 0.02 for set 1 and 0.12, 0.10, and 0.02 for set 2. The f values were approximately equal to the ratio, $\frac{F-\theta}{1-\theta}$, all tending towards zero, or equivalently, F and θ tending to be equal. This result was expected, as the correlation of alleles was the same within and among individuals in random mating, monoecious populations (Weir 1996). Also, the harmonic means between F and θ were 0.10 and 0.11 for sets 1 and 2, respectively, closely approximating their corresponding expected accumulated inbreeding of 0.11 and 0.11 for set 1

and set 2, respectively, as described by the formula of Kimura and Crow (1964). This agreement between empirical mean values of the F statistics and theoretical expectations shows that the approach employed constitutes a reliable method for studying optimal sample sizes of markers and trees for more precise estimates of genetic diversity.

The percentile intervals for the F statistics were also reduced by increasing the locus sample size in a manner similar to that observed for gene diversity and heterozygosity (Fig. 1, Table 2). These trends were also in agreement with the trends observed for the coefficients of variation (%C.V.) (Fig. 1). Kalinowski (2002a) found similar patterns in the reduction of the %C.V. of four measurements of genetic distance, including F_{ST} .

Restricted locus sampling space The summary statistics for diversity estimators were calculated for loci obtained by randomly subsampling only the loci with the largest gene diversity (Fig. 2). This proportion of loci is presented as the percentage of the total loci per linkage group and is equivalent to a threshold of locus selection; a threshold of 50% refers to a sample space including only 50 of the 100 loci per linkage group. A 100% threshold corresponds to a nonrestricted space relative to gene diversity and includes both polymorphic and monomorphic loci (Table 1).

Surfaces formed by means of gene diversity and heterozygosity on the sampling plane over loci and trees showed a decreasing slope across increasing sizes of locus samples (Fig. 2). Mean gene diversity and heterozygosity estimates based only on highly polymorphic loci would be upwardly biased compared to estimates based on all loci, polymorphic and monomorphic. In this research, the extreme bias of mean gene diversity values were 131 and 95% for sets 1 and 2, and mean values of heterozygosity were biased by 90 and 93% for sets 1 and 2, respectively. The F and θ statistics were also overestimated by a maximum of approximately 50% for set 1, and 10% for set 2, respectively. These percentages correspond to the comparison of estimates calculated from the five loci with the largest gene diversity and estimates calculated from all loci. The bias decreased gradually in all cases as the threshold of the sample space was increased, or equivalently, as more loci with lower gene diversity were considered in the sample space.

In addition to the observed bias in the mean values, the percentile intervals of gene diversity, heterozygosity, and the F and θ statistics were shorter when the loci were randomly sampled from spaces with the largest gene diversity (Fig. 2c and d, Table 3). These interval lengths became progressively longer when more loci were included in the sample space. For example, increasing the sample size from one locus to two loci per linkage group reduced the interval lengths by approximately 40%, for sample spaces of five loci with the largest gene diversity. However, the reduction in the interval lengths was approximately 30% less when all loci were included in the sample space. These trends were similar and consistent for the C.V. of the different estimators.

As gene diversity is a function that increases with the number of alleles and whose maximum is reached when all alleles are equally frequent, then selecting the loci with the largest gene diversity is equivalent to selecting the loci with the highest number of alleles with approximately equal frequency. Therefore, these results show that analyzing variability with only a few highly polymorphic loci, each with many alleles, leads to an upward bias of gene diversity, heterozygosity, and θ when the inference is genome-wise. Thus, sampling restrictions on the polymorphic state of loci should be avoided when the inference is to be made to the entire genome (Nei 1987), otherwise the resulting bias must be accepted.

Fixed model

Unrestricted locus sample space The summary statistics for gene diversity, heterozygosity, and the f statistic for the fixed model showed very similar trends to those observed for the random model. This was expected, as differences of allele frequencies between the two sets of populations were small ($F \approx 0.110$ and 0.115 for sets 1 and 2), derived by only 50 generations of random genetic drift. The percentile intervals of gene diversity and heterozygosity estimators of populations corresponding to different sets did not overlap for sample sizes of 20 or more loci per linkage group. In fact, nearly 100 loci per linkage group were required to separate the percentile intervals of populations belonging to the same set.

Restricted locus sample space The expected inbreeding at equilibrium, when the loss of alleles by random drift balances the gain of new alleles by mutation, is approximately $F \approx 0.55$ and 0.56 for sets 1 and 2, respectively (Kimura and Crow 1964). As differences of allele frequencies among populations would be greatest at equilibrium, relatively smaller locus sample sizes at that point would be required to differentiate gene diversities and heterozygosity. However, many more generations of genetic drift are required to reach the equilibrium condition. Alternatively, estimates of gene diversity and heterozygosity based on highly polymorphic loci would show the relative differences among populations of the same set with fewer loci without increasing the number of generations of random drift. As in the case of sampling loci from an unrestricted sample space, the genetic diversity estimators of each population for the fixed model showed very similar trends to those of the random model for combinations giving optimal sample size. The relative differences of gene diversity and heterozygosity among population means of the same set were approximately constant across locus sample sizes. Nevertheless, using the most polymorphic loci alone would upwardly bias the absolute values of the estimates for genome-wise inference.

Discussion

We focused this research on two classes of statistics commonly used to analyze genetic variability in plant populations as follows: (a) gene diversity and heterozygosity as absolute estimates of the amount of genetic variability, useful for identifying populations with substantial variability and potential for agronomic evaluation as germplasm donors; (b) the F statistics as relative measurements of genetic variability, providing information about genetic differentiation among populations and insight about genetic change over time. Our forward-in-time simulation approach was designed to emulate the development process of small cacao populations, reproducing at random in isolation over a limited number of generations. The estimation process of the two classes of statistics also provided insight about properties of the estimators, as they result from calculation from SSR data of cacao populations. Summary statistics of the estimates showed that relatively small samples of trees are sufficient to achieve consistency in the observed estimations. In fact, little increase in precision was found by increasing the sample size beyond 20 trees per population. This is in agreement with Kalinowski (2005), who, using standard coalescent simulation techniques, found that fewer than 20 individuals per population are sufficient to estimate F_{ST} values greater than 0.05.

Reductions of approximately 75% for genetic distances were also reported by Kalinowski (2002a) with an increment of the sample size from two to 32 independent loci with at least two alleles per locus. In contrast however, very large samples of loci were required in this research to enable reliable inferences due to variation across the genome. Precision of the estimates was increased by more than 50% with an increase in sample size from one to five loci per linkage group (50 loci per genome) and up to 70% with an increase to ten loci per linkage group (100 per genome). The large discrepancy between both types of estimates is because loci in linkage equilibrium (independent) as simulated by Kalinowski are more informative than loci in linkage disequilibrium. However, collecting information from many independent loci is difficult to accomplish, as allele frequencies of loci become dependent due to physical linkage and shared histories. In this study, the expected distance between adjacent loci of the same linkage group for unrestricted sampling was 34, 9, 5, 3, 2, 1, and 1 cM for sample sizes of 2, 10, 20, 40, 60, 80, and 100 loci per linkage group, respectively.

The random variation of the F statistics across neutral loci with no mutation, selection, or migration, is generated by the sampling error of allele frequencies. For example, let p_{ilu} be the frequency of the allele, a_{lu} , of locus, l , in population, i , with expected value over populations of μ_{lu} . The expected value of Cockerham's parameters over a large number of random mating populations is written as

$F = \theta = \frac{\sigma_{pL}^2}{D_i}$. The numerator measures the variation of the allele frequencies among populations, and the denominator is the gene diversity over all populations. Random variations in the components of this equality can occur among loci just by sampling error in a limited number of small populations.

The simplest case to observe the random variation of F statistics is in the loci with two alleles in the first generation of random drift, as the allele frequency in each population belongs to a binomial distribution (Nei 1987). For a specific set of populations, the allele frequencies represent a random sample of the allele distribution, and so, θ becomes a random variable. The behavior of the variability of θ attributed to genetic sampling is shown in Fig. 3. The dispersion of θ values is reduced by increasing the number of populations considered to obtain the estimates, converging to the expected value for a very large number of populations. Each independent set of populations represents an independent estimate of θ . Alternatively, independent loci of the same set of populations can be used to provide estimates of θ equivalent to one locus over different sets of populations (Weir and Cockerham 1984). Graphically, the information of each independent locus of the same set of populations would also represent a point in space of random dispersion (Fig. 3). Therefore, a more precise estimate of θ would be obtained by increasing

either the number of loci or the number of small populations, or both.

Molecular data sets containing five or more SSR markers per linkage group, or equivalently 50 SSRs or more per genome, are still somewhat problematic due to the expense of materials and the amount of labor required using available commercial technology. Placing restrictions on the basis of polymorphism has been an alternative to reduce the locus sample size and still estimate genetic diversity (Kalinowski 2002a). However, genome-wise estimates would be upwardly biased if markers are selected with criteria that restrict the sampling of loci to those with higher diversity (Fig. 2a). Moreover, the gene diversity and heterozygosity for each individual population would be also upwardly biased, in a magnitude inversely proportional to the values of θ . Thus, upward bias would be larger for populations in early generations of random drift, with small values of θ and a larger proportion of polymorphic loci. Kalinowski (2002b) also showed that larger samples of loci are required to reduce the %C.V. of estimates for populations with relatively small genetic differentiation, or equivalently, low values of F_{ST} .

In contrast, the overestimation of gene diversity and heterozygosity, originated by considering only highly polymorphic loci, would be substantially lower for populations with a large proportion of fixed loci. Several populations and breeding groups have shown this genetic structure for the cacao types: Nacional (Lerceteau et al.

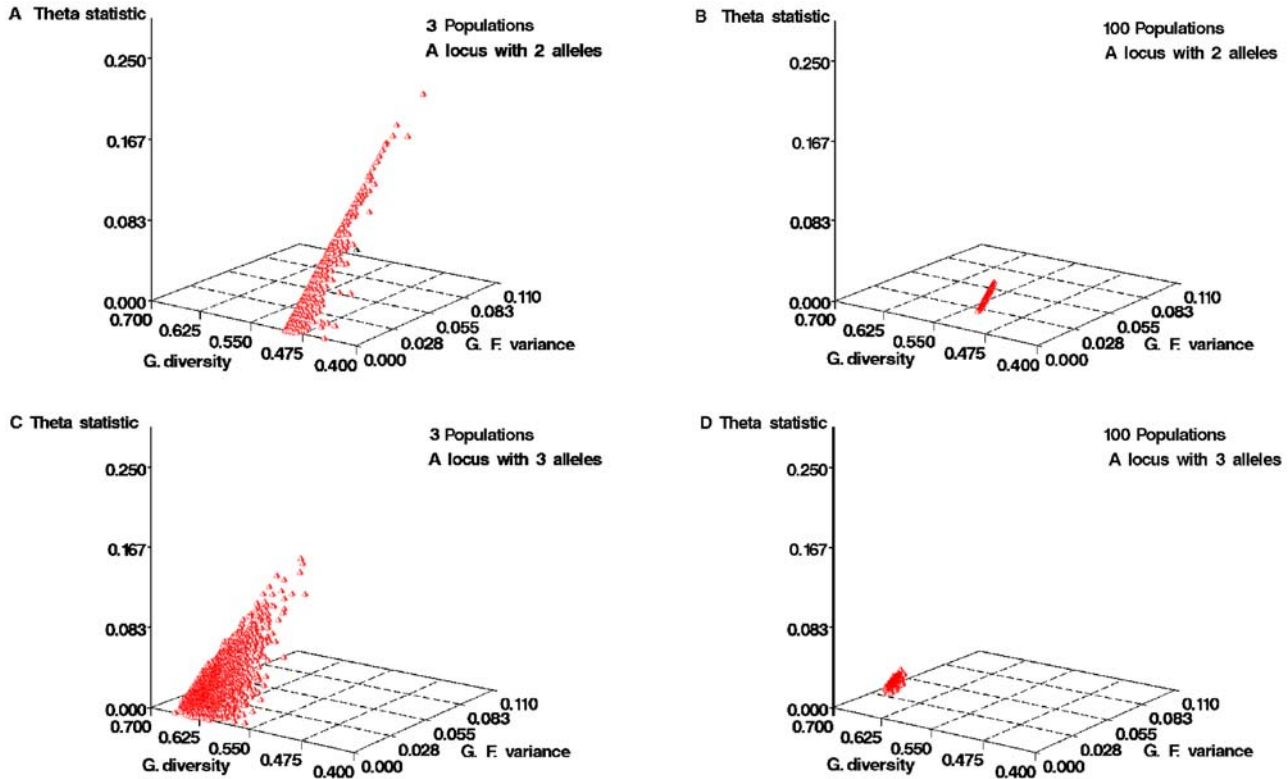


Fig. 3 Empirical distribution (obtained with 2,500 iterations) of gene diversity (*G. diversity*), gene frequency variance (*G.F. variance*), and the θ statistic. The estimates correspond to the first

generation of genetic drift of a locus with 2 (A,B) or 3 (C,D) alleles equally frequent in populations of size 20

1997), French Guiana (Laurent et al. 1994), Criollo, and Amelonado (Motamayor et al. 2002, 2003). Although current estimates of genetic diversity for these groups are expected to be reliable, given their genetic structure, we advocate increasing the marker density to at least five loci per linkage group to obtain at least a precision as we described here.

Like most of the simulation research, the results of this study answer specific questions related to the conditions of the simulation process, as large samples of loci distributed in linkage groups, and finite populations with small values of θ . These results can also be seen as the minimum sample sizes to estimate genetic variability for populations of fruit tree species with larger values of θ and a considerable number of independent loci. However, the generalization of these results to other species depends on parameters such as the number of trees per population, number of subpopulations, generations of random mating, number of linkage groups, linkage group lengths, and population differentiation. Other types of parameters such as the genome size and physical distance among loci were beyond the scope of this research. Selection of cacao trees in producing areas is currently being carried out in several countries of Asia, Africa, and Latin America, with selection for clones as cultivars, maintaining as much genetic diversity as possible. Cultivated cacao trees are often the result of a mix of hybrids produced by the national research institutes, and older, recognized cultivars. In many cases, the plant material is closely related with few generations of intermating, and small θ values. Therefore, marker densities at least as large as the ones described in the present study would be required to achieve the levels of precision discussed.

Acknowledgement We acknowledge Dr. James B. Holland for his critical review of this manuscript and suggestions.

References

Ahmad R, Potter D, Southwick SM (2003) Genotyping of peach and nectarine cultivars with SSR and SRAP markers. *J Am Soc Hortic Sci* 128:898–903

Brown JS, Schnell RJ, Motamayor JC, Lopes U, Kuhn DN, Borrone JW (2005) Resistance gene mapping for Witches' Broom disease in *Theobroma cacao* L. in a F_2 population using SSR markers and candidate genes. *J Amer Soc Hortic Sci* 130: 366–373

Clement D, Risterucci AM, Motamayor JC, N'Goran J, Lanaud C (2003) Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. *Genome* 46:103–111

Cockerham CC (1969) Variance of gene frequencies. *Evolution* 23:72–84

Couch J, Zintel HA, Fritz P (1993) The genome of the tropical tree *Theobroma cacao* L. *Mol Gen Genet* 237:123–128

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, NY

Fuerst PA, Chakraborty R, Nei M (1977) Statistical studies on protein polymorphism in natural populations: I. Distribution of single locus heterozygosity. *Genetics* 86:455–483

Gao Z-H, Shen ZJ, Han Z-H, Fang JG, Zhang Y-M, Zhang Z (2004) Microsatellite markers and genetic diversity in Japanese apricot (*Prunus mume*). *HortScience* 39:1571–1574

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov Chain Monte Carlo in practice. Chapman and Hall, London

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53:325–338

Haldane JBS (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. *J Genet* 8:299–309

Haymes KM, Ibrahim IA, Mischke S, Saunders J (2004) Rapid isolation of DNA from chocolate and date palm tree crops. *J Agric Food Chem* 52:5456–5462

Kalinowski ST (2002a) How many alleles per locus should be used to estimate genetic distances? *Heredity* 88:62–65

Kalinowski ST (2002b) Evolutionary and statistical properties of three genetic distances. *Mol Ecol* 11:1263–1273

Kalinowski ST (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity* 94:33–36

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738

Lanaud C, Hamon P, Duperray C (1992) Estimation of the nuclear DNA content of *Theobroma cacao* L. by flow cytometry. *Café Cacao* 36:3–8

Lanaud C, Risterucci AM, N'Goran AJK, Clement D, Flament MH, Laurent V, Falque M (1995) A genetic linkage map of *Theobroma cacao* L. *Theor Appl Genet* 91:987–993

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJJ (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Mol Ecol* 8:2141–2152

Laurent V, Risterucci AM, Lanaud C (1994) Genetic diversity in cocoa revealed by cDNA probes. *Theor Appl Genet* 88: 193–198

Lerceteau E, Robert T, Pétiard V, Crouzillat D (1997) Evaluation of the extent of genetic variability among *Theobroma cacao* using RAPD and RFLP markers. *Theor Appl Genet* 95:10–19

Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996) SAS System for mixed models. SAS Institute, Cary, NC

Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165:2117–2128

MacLeod AK, Haley CS, Woolliams JA, Stam P (2005) Marker densities and the mapping of ancestral junctions. *Genet Res Camb* 85:69–79

Meinguet J (1979) Multivariate interpolation at arbitrary points made simple. *J Appl Math Phys* 30:292–304

Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient tools and considerations. *Crop Sci* 43:1235–1248

Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* 89:380–386

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. *Heredity* 91:322–330

Motilal L, Butler D (2003) Verification in global cacao germplasm collections. *Genet Resour Crop Evol* 50:799–807

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583–590

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, NY

Ni J, Colowit PM, Mackill DJ (2002) Evaluation of genetic diversity in rice subspecies using microsatellite markers. *Crop Sci* 42:601–607

Pálsson S, Pamilo P (1999) The effects of deleterious mutations on linked, neutral variation in small populations. *Genetics* 153:475–483

- Reif JC, Xia XC, Melchinger AE, Warburton ML, Hoisington DA, Beck D, Bohn M, Frisch M (2004) Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR markers. *Crop Sci* 44:326–334
- Risterucci AM, Grivet L, N’Goran JAK, Pieretti I, Flament MH, Lanud C (2000) A high-density linkage map of *Theobroma cacao* L. *Theor Appl Genet* 101:948–955
- Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* 3:380–390
- Searle SR (1971) *Linear models*. Wiley, NY
- Turnbull CJ, Butler DR, Cryer NC, Zhang D, Lanaud C, Daymond AJ, Ford CS, Wilkinson MJ, Hadley P (2004) Tackling mislabelling in cocoa germplasm collections. *INGENIC Newsletter* 9:8–11
- Weir BS (1996) *Genetic data analysis II*. Sinauer, Sunderland, MA
- Weir BS, Cockerham C (1984) Estimating F -statistics for the analysis of population structure. *Evolution* 38:1358–1370
- Wright S (1978) *Evolution and genetics of populations*, vol IV. The University of Chicago