Testing Differentiation in Diploid Populations

Jérôme Goudet,* Michel Raymond,†,‡ Thierry de Meeüs§ and François Rousset†

**Institut de Zoologie et d'Ecologie Animale, Université de Lausanne, CH-1015 Lausanne, Switzerland,* †*Institut des Sciences de l'Evolution, UMR CNRS 5554, Laboratoire de Génétique et Environnement, 34095 Montpellier cedex 05, France,* ‡*Department of Genetics, University of Uppsala, S-75007 Uppsala, Sweden and* §*Laboratoire de Parasitologie Comparée, UMR CNRS 5555, Université de Montpellier II, 34095 Montpellier cedex 05, France*

Manuscript received February 8, 1996
Accepted for publication August 28, 1996

ABSTRACT

We examine the power of different exact tests of differentiation for diploid populations. Since there is not necessarily random mating within populations, the appropriate hypothesis to construct exact tests is that of independent sampling of genotypes. There are two categories of tests, F_{ST}-estimator tests and goodness of fit tests. In this latter category, we distinguish "allelic statistics", which account for the nature of alleles within genotypes, from "genotypic statistics" that do not. We show that the power of F_{ST}-estimator tests and of allelic goodness of fit tests are similar when sampling is balanced, and higher than the power of genotypic goodness of fit tests. When sampling is unbalanced, the most powerful tests are shown to belong to the allelic goodness of fit group.

SEVERAL procedures have been recently proposed in the literature to test for population differentiation. The development of fast microcomputers and of efficient algorithms has allowed the use of various exact tests (HUDSON *et al.* 1992; ROFF and BENTZEN 1992; RAYMOND and ROUSSET 1995b), which are preferable to traditionally used asymptotic tests, particularly when many infrequent alleles are considered, a situation now common after the recent development of DNA-based polymorphism.

The exact-X^2 (HUDSON *et al.* 1992) or the probability test (WEIR 1990; RAYMOND and ROUSSET 1995b) are based on the hypothesis of random sampling of genes. These tests are therefore appropriate for haploid individuals, where sampling individuals corresponds to the same process as sampling genes. For higher ploidy levels, these tests are also valid if genes within individuals are independent. For diploid organisms, this condition is equivalent to random mating at least during the latest generation, in each population. When this assumption is not valid, exact tests of population differentiation should be based on the hypothesis of independent sampling of genotypes. Here we present several exact tests assuming random sampling of diploid genotypes. Possible nonindependence of genes within individuals is therefore taken into account and could not affect their validity. Power comparison of these tests has been carried out to evaluate their performance in presence of distinct alternative hypotheses.

DEFINITION OF TESTS

Exact sampling distribution: Under the null hypothesis of absence of population differentiation, the conditional probability of the observed sample given the marginal counts is

$$\Pr(S) = \frac{\prod_{i=1}^{r} N_{i.}! \prod_{j=1}^{N_g} N_{.j}!}{N_{..}! \prod_{i=1}^{r} \prod_{j=1}^{N_g} N_{ij}!}, \qquad (1)$$

where N_{ij} is the number of individuals of genotype j in population i, $N_{i.}$ is the sample size of population i, $N_{.j}$ is the total number of individuals with genotype j, $N_{..}$ the total number of individuals sampled, r the number of samples and N_g the number of genotypes (an example is given in Table 1A). Exact tests are constructed by ranking all possible tables given the marginal counts according to the value of a particular statistic, and summing up the probabilities (as defined in Equation 1) of the tables with more extreme ranks. The resulting sum is the P value of the test.

Different classes of statistics to rank the tables will be considered here. A practical distinction is between estimators of F_{ST} and "goodness of fit" statistics such as the X^2, likelihood ratio, or (by a slight abuse of language) the sample probability $\Pr(S)$. A more important distinction is between statistics computed from allele counts ("allelic statistics", Table 1B) or from genotype counts ("genotypic statistics", Table 1A). For example, an exact-X^2 can be an allelic or genotypic goodness of fit statistic whether it is computed on allele or genotype counts. In both cases however, the distribution of the statistics under the null hypothesis is obtained using the sampling distribution under random sampling of genotypes, not of genes (*i.e.*, of Table 1A, not Table 1B). Allelic statistics take into account whether two genotypes share one allele or not, an information that is ignored by genotypic statistics that only consider whether genotypes are identical or different. In this

Corresponding author: Jérôme Goudet, Institut de Zoologie et D'Ecologie Animale, Bâtiment de Biologie, Université de Lausanne, CH-1015 Lausanne, Switzerland. E-mail: jerome.goudet@izea.unil.ch

TABLE 1

Layout for the two types of contingency tables described in text

	11	12	13	22	23	33	All
				A			
1	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}	N_{16}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	N_{24}	N_{25}	N_{26}	$N_{2.}$
3	N_{31}	N_{32}	N_{33}	N_{34}	N_{35}	N_{36}	$N_{3.}$
All	$N_{.1}$	$N_{.2}$	$N_{.3}$	$N_{.4}$	$N_{.5}$	$N_{.6}$	$N_{..}$
				B			
1		n_{11}		n_{12}		n_{13}	$n_{1.}$
2		n_{21}		n_{22}		n_{33}	$n_{2.}$
3		n_{31}		n_{32}		n_{33}	$n_{3.}$
All		$n_{.1}$		$n_{.2}$		$n_{.3}$	$n_{..}$

A. Layout of a R × C genotypic table. Sample made of three subsamples (rows). In this example, three alleles (six genotypes) were found at the locus analyzed. B. Layout of a R × C genic (or allelic) table derived from the genotypic one.

respect the usual estimators of F_{ST} are allelic statistics since they take into account the nature of alleles within genotypes.

***F_{ST}* estimators:** The parameter traditionally used to measure population differentiation is F_{ST} (WRIGHT 1951; COCKERHAM and WEIR 1987). Several estimators of F_{ST} were used to construct a rejection zone.

$\hat{\theta}_{WC}$. This is the estimator of F_{ST}, defined by WEIR and COCKERHAM (1984) as

$$\hat{\theta}_{WC} = \frac{\sum_{u=1}^{k} a_u}{\sum_{u=1}^{k}(a_u + b_u + c_u)},$$

where k is the number of alleles at the locus, and a_u, b_u and c_u are the among samples, among individuals within samples and within individual estimates of components of variance, respectively, of a nested analysis of variance on allele frequencies (COCKERHAM 1969, 1973; WEIR and COCKERHAM 1984). The P value of the exact $\hat{\theta}_{WC}$-test is computed as $P_{\hat{\theta}_{WC}} = \Sigma_{\hat{\theta}_{WC}(s_i) \geq \hat{\theta}_{WC}(S)} \Pr(s_i)$, where the summation is over all possible contingency tables s_i with a larger $\hat{\theta}_{WC}$ than the observed.

$\hat{\theta}_U$. This is an "unweighted" estimator of F_{ST}, defined by WEIR and COCKERHAM (1984) as

$$\hat{\theta}_U = 1/k \sum_{u=1}^{k} \frac{a_u}{a_u + b_u + c_u}.$$

The P value of the exact $\hat{\theta}_U$-test is computed as $P_{\hat{\theta}_U} = \Sigma_{\hat{\theta}_U(s_i) \geq \hat{\theta}_U(S)} \Pr(s_i)$.

$\hat{\theta}_{RH}$. This statistic is another estimate of F_{ST} (ROBERTSON and HILL 1984). An explicit formula is found in WEIR and COCKERHAM (1984):

$$\hat{\theta}_{RH} = 1/(k-1) \sum_{u=1}^{k} \frac{(1 - \tilde{p}_u) a_u}{a_u + b_u + c_u},$$

TABLE 2

Parameters values of the different alternative hypotheses

Sampling scheme	m	m'	Selfing rate	H_t	F_{ST}
4 × 64	0.09	0.12	0.0	0.83	0.0262
16 × 16	0.30	0.32	0.0	0.83	0.0262
64 × 64	0.12	0.12	0.0	0.94	0.0262
4 × 64	0.20	0.27	0.0	0.82	0.0090
	0.30	0.40	0.0	0.82	0.0044
	0.40	0.53	0.0	0.82	0.0022
	0.50	0.67	0.0	0.82	0.0010
	0.60	0.80	0.0	0.82	0.0003
	0.75	1.00	0.0	0.82	0.0000
16 × 16	0.40	0.43	0.0	0.82	0.0151
	0.50	0.53	0.0	0.82	0.0086
	0.60	0.64	0.0	0.82	0.0046
	0.75	0.80	0.0	0.82	0.0013
	0.85	0.91	0.0	0.82	0.0003
	0.94	1.00	0.0	0.82	0.0000
4 × 64	0.20	0.27	0.7	0.74	0.0138
	0.30	0.40	0.7	0.74	0.0067
	0.40	0.53	0.7	0.74	0.0033
	0.50	0.67	0.7	0.74	0.0015
	0.60	0.80	0.7	0.74	0.0005
	0.70	0.93	0.7	0.74	0.0001
	0.75	1.00	0.7	0.74	0.0000
16 × 16	0.30	0.32	0.7	0.75	0.0397
	0.40	0.43	0.7	0.74	0.0230
	0.50	0.53	0.7	0.74	0.0132
	0.60	0.64	0.7	0.74	0.0071
	0.75	0.80	0.7	0.74	0.0020
	0.85	0.91	0.7	0.74	0.0004
	0.94	1.00	0.7	0.74	0.0000

m, migration rate; m', migration rate corrected for the number of populations [$m' = Dm/(D\text{-}1)$]; H_t, expected value of the genetic diversity, F_{ST}, expected value of the parameter F_{ST}.

where $\tilde{p}_u$ is the observed frequency of allele u. The P value of the exact $\hat{\theta}_{RH}$-test is computed as $P_{\hat{\theta}_{RH}} = \Sigma_{\hat{\theta}_{RH}(s_i) \geq \hat{\theta}_{RH}(S)} \Pr(s_i)$.

Genotypic goodness of fit statistics: Three statistics have been used in this category.

X^2. The X^2 is the well-known chi-square, *i.e.*,

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{N_g} \frac{(N_{ij} - N_{i.}\overline{p}_{.j})^2}{N_{i.}\overline{p}_{.j}},$$

where $\overline{p}_{.j} = N_{.j}/N_{..}$. The proportion of tables with a higher or equal X^2, $P_{X^2} = \Sigma_{X^2(s_i) \geq X^2(S)} \Pr(s_i)$, is the P value of an exact genotypic-X^2 test.

G. The G value is the traditional log likelihood ratio (SOKAL and ROHLF, 1981):

$$G = -2 \sum_{i} \sum_{j} N_{ij} \ln \left(\frac{N_{ij}}{N_{i.}\overline{p}_{.j}} \right).$$

An exact genotypic G-test is constructed by computing the proportion of tables with a higher or equal G value, *i.e.*, $P_G = \Sigma_{G(s_i) \geq G(S)} P(s_i)$.

TABLE 3

Power of the different exact tests described in the text

		Goodness of fit tests							
		F_{ST} estimator tests			Genotypic tests			Allelic tests	
Sampling scheme	F_{ST}	$\hat{\theta}_{RH}$	$\hat{\theta}_{U}$	$\hat{\theta}_{WC}$	Pr	X^2	G	$\Pr_a$	G_a
4 × 64	0.0262	**999**	**999**	990	965	966	965	**999**	**999**
	0.0090	**875**	873	731	484	489	497	872	873
	0.0044	**535**	531	392	226	227	227	521	523
	0.0022	**250**	244	172	109	114	111	243	232
	0.0010	**136**	130	109	71	85	75	129	129
	0.0003	64	**67**	61	66	64	66	65	65
16 × 16	0.0262	**994**	**994**	980	782	798	792	**994**	**994**
	0.0151	908	**913**	799	413	429	422	904	899
	0.0086	**610**	606	471	241	268	248	578	577
	0.0046	287	**291**	242	137	137	139	260	269
	0.0013	95	**97**	**97**	74	87	76	89	95
	0.0003	55	55	62	**81**	73	70	67	70

4 × 64 stands for four populations of 64 individuals. 16 × 16 stands for 16 populations of 16 individuals. Numbers shown are the number of replicates (out of 1000) that gave significant ($P \leq 0.05$) results. Bold characters indicate the highest observed power for a given alternative hypothesis. All alternative hypotheses are generated with random mating within demes.

Probability. The probability Pr of the observed sample (Equation 1) can be used to define the rejection zone. This procedure gives a test generally known as the Fisher exact test on contingency table or Probability test. The latter designation will be used here. The P value is computed as $P_{\Pr} = \Sigma_{\Pr(s_i) \leq \Pr(S)} \Pr(s_i)$.

Allelic goodness of fit statistics: The same goodness of fit statistics can be computed on the genic tables derived from the genotypic ones (that is, under the hypothesis of independent sampling of genotypes): the X_a^2, the G_a value, and the probability $\Pr_a$ of the genic table under the null hypothesis of nongenic differentiation. The definitions are the same as above with the interpretation that n_{ij} is the count of allele j in sample i, and n_g the number of distinct alleles:

$$\Pr(S) = \frac{\prod_{i=1}^{r} n_{i.}! \prod_{j=1}^{n_g} n_{.j}!}{n_{..}! \prod_{i=1}^{r} \prod_{j=1}^{n_g} n_{ij}!}. \quad (2)$$

Test computation: The computation of each exact test requires the evaluation of all tables given the marginal values, which is practically impossible for most cases (GAIL and MANTEL 1977). Instead of computing this exact probability, an unbiased estimate can be obtained using an appropriate sampling procedure under

TABLE 4

As Table 3 but alternative hypotheses generated with 70% selfing

		Goodness of fit tests							
		F_{ST} estimator tests			Genotypic tests			Allelic tests	
Sampling scheme	F_{ST}	$\hat{\theta}_{RH}$	$\hat{\theta}_{U}$	$\hat{\theta}_{WC}$	Pr	X^2	G	$\Pr_a$	G_a
4 × 64	0.0138	**748**	740	628	570	560	547	741	740
	0.0067	405	**411**	319	275	266	259	406	406
	0.0033	191	189	170	114	115	111	**201**	197
	0.0015	106	**111**	108	86	88	86	110	104
	0.0005	**85**	80	71	73	65	76	83	81
	0.0001	55	50	50	52	48	51	55	**58**
16 × 16	0.0397	967	968	925	877	871	873	967	**972**
	0.0230	**780**	**780**	658	579	535	548	749	753
	0.0132	**469**	**469**	397	312	304	304	465	461
	0.0071	232	**234**	195	168	151	170	212	217
	0.0020	**88**	83	61	71	73	63	81	85
	0.0004	57	56	60	54	**69**	**69**	58	57

Bold indicates the highest observed power for a given alternative hypothesis.

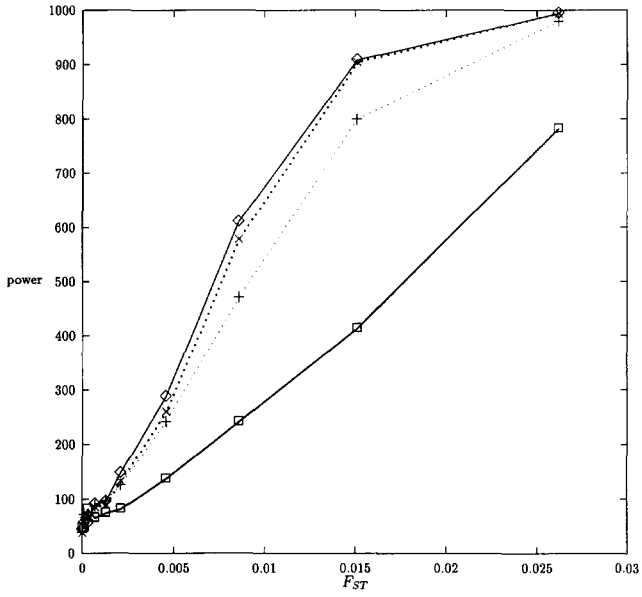

FIGURE 1.—Power of different tests as a function of alternative hypotheses. Population sampled made of 16 demes of 16 individuals. No selfing. ◇, $\hat{\theta}_{RH}$-test; +, $\hat{\theta}_{WC}$-test; □, Pr-test; ×, $\Pr_a$-test.

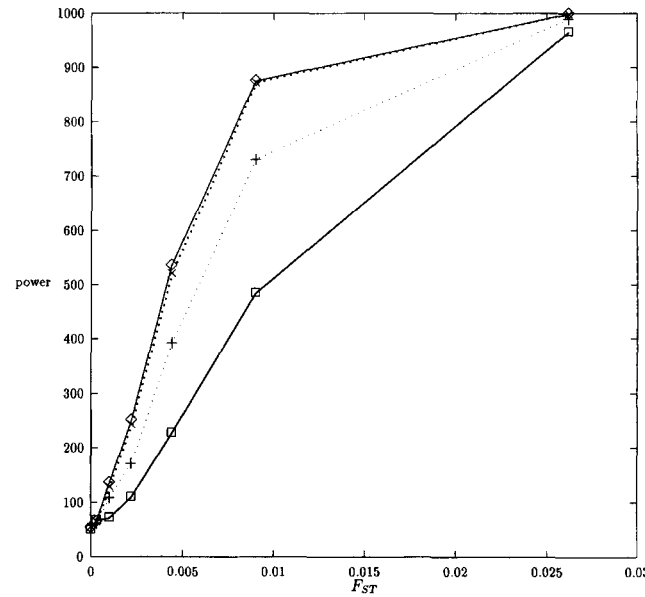

FIGURE 2.—As in Figure 1. Population sampled made of four demes of 64 individuals. No selfing. ◇, $\hat{\theta}_{RH}$-test; +, $\hat{\theta}_{WC}$-test; □, Pr-test; ×, $\Pr_a$-test.

the null hypothesis. We used either a permutation-based method such as the one developed by HUDSON *et al.* (1992) or the Markov chain method developed by RAYMOND and ROUSSET (1995b) (see also GUO and THOMPSON, 1992). All programs were tested by hand computation or by comparison with each other on various datasets. Type I error was fixed at 5%, as is usually done in biological studies.

Generating alternative hypotheses: Our simulations assume that all alleles are different at initial time, and, to generate the alternative hypotheses, follow the progress of inbreeding with time in a neutral finite island model of population. Asymptotic values of fixation indices and of heterozygosity are known in this model. The validity of simulations were checked by comparing the observed values to their theoretical expectations. One thousand independent loci (replicates) were generated for each alternative hypothesis. The set of parameters we used is described in Table 2.

The simulations were run for a sufficient number of generations to insure that equilibrium between drift and migration has been reached. For each set of parameters we checked that asymptotic equilibrium was attained. Since we are interested in finding tests with high power, we focused our attention on high migration rates (small F_{ST}).

For most of our results, the entire population was sampled without replacement, therefore only the effect of the genetic sampling is observed. However, the effect of the statistical sampling may be much larger than that of the genetic sampling. For this reason, we simulated 64 populations made of 64 individuals, from which we sampled 16 individuals in 16 populations. These simulations were run for a large number of generations, much larger than the time necessary to reach equilibrium, and samples where taken every 500 generations until the 3000th generation to look at the effect of global genetic diversity on the power of the different tests. We also took a series of unbalanced samples from our populations, since the behavior of our tests may differ between balanced and unbalanced sampling. Pseudo-random numbers were generated according to MARSAGLIA *et al.* (1990) for the Markov Chain algorithm and according to L'ECUYER (1988) for the randomization algorithm.

RESULTS

In all cases, the power of the X_a^2 and $\hat{\theta}_{RH}$-test were the same, so the X_a^2 power is not shown in the tables. Results when the null hypotheses were true always gave power very close to 5% as expected and are therefore not shown either.

Exhaustive sampling, balanced samples: Results are shown in Tables 3 and 4 and Figures 1 and 2. In all cases, genotypic goodness of fit tests are less powerful than allelic tests that use the additional information of the identity of the alleles within genotypes. For each alternative hypothesis, the different allelic tests have similar power. The main exception is when $\hat{\theta}_{WC}$ is used to define the rejection zone, which lead to a decrease in power in most conditions (Tables 3 and 4 and Figures 1 and 2).

For comparable alternative hypotheses, our results indicate that it may be better to sample more individu-

TABLE 5

Power of goodness of fit statistics under sampling of genotypes or alleles

		Allelic tests			
		On genotypic tables		On genic tables	
Sampling scheme	F_{ST}	Pr_a	G_a	Pr_{aa}	G_{aa}
16 × 16	0.0151	904	899	**906**	905
	0.0086	578	577	593	**599**
	0.0013	89	95	98	**103**

Random mating within populations is assumed. Bold characters indicate the highest observed power for a given alternative hypothesis.

als from fewer populations than the reverse, at least when the alternative hypothesis is close to the null (*i.e.*, F_{ST} is small, Tables 3 and 4).

When there is random mating within populations, genes within individuals are independent. We may then compare the power of tests based on the probability distribution of alleles (which are obtained from considering the tables with the same marginal counts as the allelic table, Table 1B) and that of genotypes. Table 5 shows that the loss of power of the latter is small.

Partial sampling, balanced samples: Results are shown in Table 6. Ranking of the different tests is similar to what we obtain under exhaustive sampling. As the number of generations increases, genetic diversity decreases, so does the power of the tests (Table 6).

Unbalanced samples: A discrepancy appears when samples are unbalanced between the probability test (Pr_a) or the *G*-test (G_a), which performed better than the F_{ST}-tests. G_a seems overall slightly more powerful than Pr_a (Table 7). Among the F_{ST}-tests, $\hat{\theta}_{RH}$ and $\hat{\theta}_U$ performed equally well, with a tendency for the former to do slightly better when samples are balanced, and slightly worse when the samples are unbalanced (Tables 6 and 7).

TABLE 7

Effect of unequal sampling on the power of different exact tests

	Tests power				
F_{ST}	$\hat{\theta}_{RH}$	$\hat{\theta}_U$	$\hat{\theta}_{WC}$	Pr_a	G_a
		A			
0.0090	353	366	325	479	**496**
0.0044	165	169	159	**237**	236
		B			
0.0138	244	268	270	**372**	364
0.0067	146	145	151	189	**196**
		C			
0.0262	880	883	822	914	**917**
0.0086	320	324	272	341	**351**
		D			
0.0397	758	774	706	788	**804**
0.0132	243	**255**	210	250	251

Partial sampling was carried out on four populations of 64 individuals (two samples of five individuals, two of 64, A: no selfing; B: 70% selfing) and on 16 populations of 16 individuals (six samples of five individuals, five of 10, six of 16, C: no selfing and D: selfing). Bold characters indicate the highest observed power for a given alternative hypothesis.

The effect of selfing and nonrandom mating: When there is departure from random mating, the power of the tests drops (Tables 3 and 4).

Relationships to estimation: $\hat{\theta}_{WC}$ provides the least biased estimator of F_{ST} for all sampling strategies, but has the largest variance when genetic diversity is high (Tables 8–10). As genetic diversity decreases, the variance of the three estimators increases (Figure 3). Noteworthy is the effect of unbalanced sampling on the variance of $\hat{\theta}_U$ and $\hat{\theta}_{RH}$: while $\hat{\theta}_{RH}$ has always a lower variance when all samples are of equal sizes (Tables 8 and 9), $\hat{\theta}_U$ has a lower variance when sample sizes are unequal (Table 10).

TABLE 6

Effects of the sampling scheme and of genetic diversity on the power of different exact tests

				Tests power				
Sampling scheme	Generation	F_{ST}	H_t	$\hat{\theta}_{RH}$	$\hat{\theta}_U$	$\hat{\theta}_{WC}$	Pr_a	G_a
Exhaustive sampling								
16 × 16		0.0262	0.83	**994**	**994**	980	**994**	**994**
4 × 64		0.0262	0.83	**999**	**999**	990	**999**	**999**
Partial sampling in 64 × 64								
16 × 16	500	0.0262	0.94	996	996	989	**997**	**997**
	1000	0.0262	0.88	947	949	907	949	**951**
	2000	0.0262	0.78	**734**	733	689	731	728
	3000	0.0262	0.69	629	627	601	627	**635**

Partial sampling was carried out on 64 populations made of 64 individuals. H_t, expected heterozygosity. Bold characters indicate the highest observed power for a given alternative hypothesis.

TABLE 8

Mean of the three estimators of F_{ST} with no selfing

Sampling scheme	Estimators of F_{ST}			
	F_{ST}	$\hat{\theta}_{RH}$	$\hat{\theta}_{U}$	$\hat{\theta}_{WCSampling\ scheme}$
4 × 64	0.0262	0.0226 (0.76)	0.0229 (0.78)	0.0264 (1.14)
	0.0090	0.0082 (0.40)	0.0083 (0.42)	0.0092 (0.60)
	0.0044	0.0040 (0.32)	0.0041 (0.32)	0.0044 (0.44)
	0.0022	0.0020 (0.25)	0.0020 (0.25)	0.0020 (0.32)
	0.0010	0.0010 (0.22)	0.0010 (0.23)	0.0011 (0.30)
	0.0003	0.0003 (0.22)	0.0003 (0.22)	0.0003 (0.29)
16 × 16	0.0262	0.0242 (0.74)	0.0243 (0.74)	0.0266 (0.94)
	0.0151	0.0140 (0.60)	0.0141 (0.60)	0.0151 (0.73)
	0.0086	0.0081 (0.51)	0.0082 (0.51)	0.0085 (0.62)
	0.0046	0.0043 (0.45)	0.0043 (0.45)	0.0047 (0.57)
	0.0013	0.0012 (0.41)	0.0012 (0.41)	0.0013 (0.53)
	0.0003	0.0002 (0.39)	0.0002 (0.39)	0.0002 (0.51)

SD × 100 in parentheses.

DISCUSSION

Overall, the most powerful tests are the allelic probability and the allelic G, carried out on the genotypic tables (respectively Pr_a-test and G_a-test). These conclusions stand for all sampling schemes. While power is similar between these two tests and the F_{ST}-tests for balanced sampling (*e.g.*, Table 6), the loss of power when samples are unbalanced is larger in F_{ST}-tests than in the Pr_a-test and the G_a-test (Table 7). For example, for an alternative hypothesis defined by $F_{ST} = 0.009$, we find no significant differences in power between the F_{ST}-tests and the Pr_a or G_a tests (Table 3), but the latter are at least 13% more powerful than F_{ST}-tests when samples are unbalanced (Table 7).

Sampling strategies: Power is much higher when samples are balanced than when they are not. We also find that when gene flow is high, it seems better to sample few populations with many individuals rather than more populations with less individuals. This may be because one gets a better estimation of local allele frequencies, thereby increasing the ability of the test to reject smaller allele frequency differences. When samples are balanced, most exact tests that use the identity of alleles information perform equally well.

Among the tests based on F_{ST} estimators, $\hat{\theta}_{WC}$, while being the least biased estimator, is also the least powerful. $\hat{\theta}_{RH}$ and $\hat{\theta}_{U}$ perform equally well, with the former doing slightly better under balanced sampling and slightly worse under unbalanced sampling.

Relationships to estimation: The different weightings given to alleles have an influence on the statistical outcome. Giving more weight to rare alleles provides estimators with lower variance under the null hypothesis and therefore more powerful test statistics. This finding stands for F_{ST} (ROBERTSON and HILL 1984; WEIR and COCKERHAM 1984; LONG 1986) and related quantities

TABLE 9

As Table 8 but with 70% selfing

Sampling scheme	Estimators of F_{ST}			
	F_{ST}	$\hat{\theta}_{RH}$	$\hat{\theta}_{U}$	$\hat{\theta}_{WC}$
4 × 64	0.0138	0.0125 (0.82)	0.0127 (0.85)	0.0142 (1.13)
	0.0067	0.0062 (0.61)	0.0063 (0.63)	0.0069 (0.82)
	0.0033	0.0031 (0.50)	0.0031 (0.52)	0.0035 (0.68)
	0.0015	0.0013 (0.46)	0.0013 (0.47)	0.0016 (0.64)
	0.0005	0.0006 (0.43)	0.0006 (0.44)	0.0006 (0.56)
	0.0001	0.0002 (0.40)	0.0002 (0.41)	0.0002 (0.53)
16 × 16	0.0397	0.0366 (1.41)	0.0372 (1.41)	0.0400 (1.67)
	0.0230	0.0211 (1.18)	0.0214 (1.19)	0.0232 (1.44)
	0.0132	0.0124 (0.94)	0.0127 (0.96)	0.0139 (1.22)
	0.0071	0.0065 (0.83)	0.0067 (0.85)	0.0075 (1.08)
	0.0020	0.0018 (0.77)	0.0018 (0.78)	0.0018 (0.96)
	0.0004	0.0004 (0.73)	0.0004 (0.74)	0.0004 (0.96)

SD × 100 in parentheses.

TABLE 10

As Table 8 but with unequal sample sizes

Estimators of F_{ST}			
F_{ST}	$\hat{\theta}_{RH}$	$\hat{\theta}_{U}$	$\hat{\theta}_{WC}$
		A	
0.0090	0.0085 (0.88)	0.0085 (0.87)	0.0090 (1.05)
0.0044	0.0042 (0.75)	0.0042 (0.72)	0.0044 (0.78)
		B	
0.0138	0.0123 (1.61)	0.0125 (1.56)	0.0138 (1.83)
0.0067	0.0062 (1.48)	0.0062 (1.43)	0.0066 (1.56)
		C	
0.0262	0.0238 (1.05)	0.0240 (1.04)	0.0262 (1.27)
0.0086	0.0080 (0.77)	0.0081 (0.76)	0.0087 (0.94)
		D	
0.0397	0.0362 (1.61)	0.0368 (1.56)	0.0405 (1.83)
0.0132	0.0120 (1.48)	0.0123 (1.43)	0.0137 (1.56)

A, B, C, D as in Table 7. SD × 100 in parentheses.

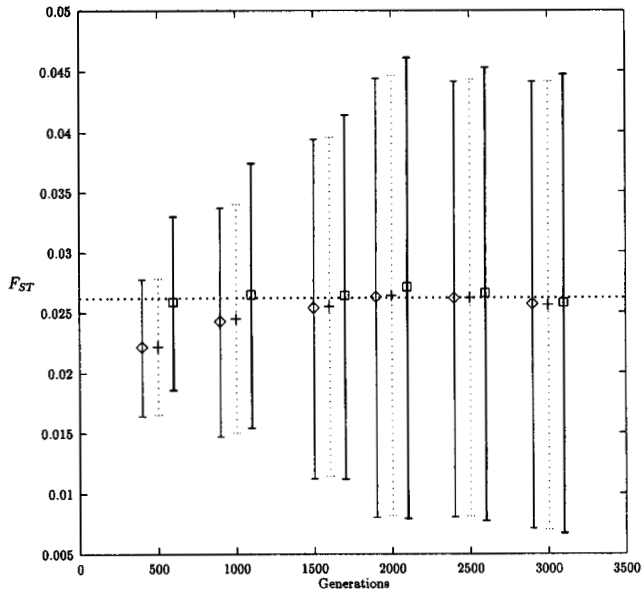

FIGURE 3.—Observed values of the three F_{ST} estimators ± 1 SD. ◇, $\hat{\theta}_{RH}$; +, $\hat{\theta}_{U}$; □, $\hat{\theta}_{WC}$; . . . , expected value of F_{ST}.

(BARTON and SLATKIN 1986; SLATKIN and BARTON 1989) as well as for F_{IS} (ROBERTSON and HILL 1984; ROUSSET and RAYMOND 1995).

The weighting given to the different samples for the F_{ST}-estimators was proportional to their size. This is the weighting advocated by WEIR and COCKERHAM (1984). HUDSON *et al.* (1992) used an equivalent weighting. Nevertheless, this weighting scheme for sample sizes may not be the best for the present testing purposes. NEI and CHESSER (1983) suggested to weight equally all samples, on the ground that population size is usually unknown and may not be reflected in sample sizes. This point needs further investigations.

Genetic diversity: The levels of genetic diversity present in our simulated populations are closer to what is expected from molecular markers such as microsatellites than from isozymes (Tables 2 and 6). However, mutation was not considered in this investigation. Particular mutation processes could generate allele frequency distributions different from the one generated by drift/migration alone and could affect the power of the different tests.

Conclusions: The tests presented here are single locus, multi-allelic tests. F_{ST} and goodness of fit tests are easily generalized to multi-locus data if these can be considered independent.

For diploid sexual organisms, exact tests of subdivision should be based on the hypothesis of independent sampling of genotypes when the random sampling of alleles is inappropriate (alleles within individuals are not independent when there is nonrandom mating). Even when there is random mating, the power of allelic test based on genotypic tables is similar to that of tests based on genic tables when there is random mating. However, when samples are small compared to the level of variability (such that most sampled individuals have a unique genotype), allelic test on genic tables could be more powerful.

Tests ignoring the identity of alleles (genotypic goodness of fit statistics) are less powerful than tests acknowledging it. While a test more powerful than those described here could be found, it should emerge from the allelic statistics class rather than the genotypic one.

This work was financed in part by grant 31-43443.95 of the Swiss National Science Foundation (J.G.), by PICS 290 of Centre National de la Recherche Scientifique (T.D.) and GDR1105 (Programme Environnement, Vie et Société du CNRS). This is contribution 96-139 of the Institut des Sciences de l'Evolution (URA CNRS 327). M.R. thanks P. PAMILO for the opportunity to spend a year at the department of Genetics, Uppsala University, Sweden. The G_d-test will be included in future releases of the computer programs FSTAT (GOUDET 1995) and Genepop (RAYMOND and ROUSSET 1995a).

LITERATURE CITED

BARTON, N. H., and M. SLATKIN, 1986 A quasi equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity **56:** 409–415.

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.

GAIL, M., and N. MANTEL 1977 Counting the number of contingency tables with fixed margins. J. Am. Stat. Assoc. **72:** 859–862.

GOUDET, J., 1995 Fstat Version 1.2. A computer program to calculate F-statistics. J. Hered. **86:** 485–486.

GUO, S. W., and E. A. THOMPSON, 1992 Performing the exact test of Hardy-Weinberg proportions for multiple alleles. Biometrics **48:** 361–372.

HUDSON, R. R., D. D. BOOS and N. L. KAPLAN, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

L'ECUYER, P., 1988 Efficient and portable random number generators. Comm. ACM **31:** 147–157.

LONG, J. C., 1986 The allelic correlation of the Gainj and Kalam speaking people. I. The estimation of and interpretation of WRIGHT's F-statistics. Genetics **112:** 629–647.

MARSAGLIA, G., A. ZAMAN and W. W. TSANG, 1990 Toward a universal random number generator. Stat. Prob. Lett. **8:** 35–39.

NEI, M., and R. K. CHESSER, 1983 Estimation of fixation indices and gene diversities. Am. J. Hum. Genet. **47:** 253–259.

RAYMOND, M., and F. ROUSSET, 1995a GENEPOP (version 1.2): a population genetics software for exact test and ecumenicism. J. Hered. **86:** 248–249.

RAYMOND, M., and F. ROUSSET, 1995b An exact test of population differentiation. Evolution **49:** 1280–1283.

ROBERTSON A., and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

ROFF, D. A., and P. BENTZEN, 1992 Detecting geographical subdivision: a comment on a paper by HUDSON *et al.* Mol. Biol. Evol. **9:** 968.

ROUSSET, F., and M. RAYMOND, 1995 Testing heterozygote excess and deficiency. Genetics **140:** 1413–1419.

SLATKIN, M., and N. H. BARTON, 1989 A comparison of three indirect methods for estimating levels of gene flow. Evolution **43:** 1349–1368.

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry,* Ed. 2. Freeman and Co., New York.

WEIR, B. S., 1990 Genetic data analysis. Sinauer, 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., 1951 The genetical structure of populations. Ann. Eugenics **15:** 323–354.

Communicating editor: B. S. WEIR