Interpretation of Variation Across Marker Loci as Evidence of Selection

Renaud Vitalis, Kevin Dawson, Pierre Boursot

Interpretation of Variation Across Marker Loci as Evidence of Selection

Renaud Vitalis[⋆†‡], Kevin Dawson[⋆§], and Pierre Boursot[⋆]

[⋆] *Laboratoire Génome, Populations et Interactions, C.C. 063, Université Montpellier II, 34095 Montpellier Cedex 05, France*

[†] *Laboratoire Génétique et Environnement, C.C. 065, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier II, 34095 Montpellier Cedex 05, France*

[‡] *Station Biologique de la Tour du Valat, Le Sambuc, 13200 Arles, France*

[§] *I.A.C.R. Long Ashton Research Station, Department of Agricultural Science, University of Bristol, Bristol BS41 9AF, United Kingdom*

Running head:

Evidence of selection from marker loci

Keywords:

Population divergence, genetic drift, coalescent theory, neutrality tests, IBD probabilities

Corresponding author:

Renaud VITALIS

Laboratoire Génétique et Environnement, C.C. 065

Institut des Sciences de l'Évolution de Montpellier

Université Montpellier II

Place Eugène Bataillon, 34095 Montpellier Cedex 05, France

Tel.: +33 4 67 14 32 50

Fax.: +33 4 67 14 36 22

E-mail: `vitalis@isem.univ-montp2.fr`

ABSTRACT

Population structure and history have similar effects on the genetic diversity at all neutral loci. However, some marker loci may also have been strongly influenced by natural selection. Selection shapes genetic diversity in a locus-specific manner. If we could identify those loci that have responded to selection during the divergence of populations, then we may obtain better estimates of the parameters of population history, by excluding these loci. Previous attempts have been made to identify outlier loci from the distribution of sample statistics under neutral models of population structure and history. Unfortunately these methods depend on assumptions about population structure and history, and these are usually unknown. In this paper, we define new population-specific parameters of population divergence, and construct sample statistics which are estimators of these parameters. We then use the joint distribution of these estimators to identify outlier loci that may be subject to selection. We found that outlier loci are easier to recognize when this joint distribution is conditioned on the total number of allelic states in the pooled sample, at each locus. This is because the conditional distribution is less sensitive to the values of nuisance parameters.

INTRODUCTION

PRESUMED neutral polymorphic loci are commonly used in making inferences about patterns of differentiation within or among populations of the same or closely related species. For this purpose, genetic distances (see, *e.g.*, NEI, 1972) or WRIGHT's (1951) F-statistics are estimated from allele-frequency data. Under particular models of population structure, these parameters are related to demographic or historical parameters, such as the effective population size, the rate of migration between populations or the time since the populations diverge from their common ancestral population.

However, misinterpretations can occur, if one is not able to clearly distinguish between the patterns generated by random genetic drift or by natural selection. The problem is that selective processes can also affect neutral loci. A locus which is neutral will respond to selection whenever it is in linkage disequilibrium (statistical association among allelic states at different loci) with other loci which are subject to selection. Such associations may arise by chance in small populations (HILL and ROBERTSON, 1966, 1968; OHTA and KIMURA, 1969). For example, stabilizing or balancing selection operating at a locus tends to maintain an elevated level of variation at closely linked neutral loci (HUDSON and KAPLAN, 1988; STROBECK, 1983). Selection acting on any locus has an effect on loosely linked loci, which resembles a reduction of effective population size (BARTON, 1995, 1998; ROBERTSON, 1961). Local adaptation tends to increase population differentiation at loci where selection acts, and very high F_{ST} values may be found at closely linked neutral loci (CHARLESWORTH *et al.*, 1997). The substitution of advantageous mutations at a locus may also reduce neutral variation at linked loci

(KAPLAN *et al.*, 1989; MAYNARD SMITH and HAIGH, 1974). Similarly, « background selection », caused by the selection against deleterious mutations (CHARLESWORTH *et al.*, 1993) results in a reduced effective population size for neutral genes in the region of the chromosome where this selection is acting. Background selection may also increase the apparent population differentiation (CHARLESWORTH *et al.*, 1997).

Therefore, it is of prime interest to identify loci that are responding to selection in order to exclude them from the genetic analysis of population structure or history. It was recognized early on by CAVALLI-SFORZA (1966) that any form of selection will affect some regions of the genome more than others, whereas population history, demography, migration and the mating system will affect the whole genome in the same way. Accordingly, LEWONTIN and KRAKAUER (1973) proposed two tests of selective neutrality. Both tests are based on the sampling distribution of a statistic $\widehat{F}$, the standardized variance of gene frequency, which is an estimator of the parameter F_{ST}. Their first test is a goodness of fit test comparing the observed distribution of $\widehat{F}$ estimates (one estimate from each locus) to a χ^2 distribution with $(n-1)$ degrees of freedom, where n is the number of populations sampled. The second test is based on the comparison of the observed variance of $\widehat{F}$ (across loci) noted s_F^2, with the theoretical variance approximated as

$$\sigma^2 = \frac{k\overline{F^2}}{n-1} \tag{1}$$

where $\overline{F}$ is the mean value of $\widehat{F}$ averaged across loci, and k is a constant which, according to LEWONTIN and KRAKAUER (1973), should not exceed 2 whatever the underlying distribution of allelic frequency. The ratio s_F^2/σ^2,

should be distributed approximately as a $\chi^2/d.f.$, the number of degrees of freedom $d.f.$ being determined by the number of bi-allelic loci.

However, since populations of the same species share, to a certain extent, a common history and since populations are connected through the dispersal of individuals, $\widehat{F}$ values will be correlated across loci. For example, the geographic and historical relationships between populations may have a hierarchical structure if populations have derived from a common ancestral population as a sequence of successive splits. This is the pattern expected when the fragmentation of a species range occurred as a sequence of population subdivisions. The effect of such a population history is always to increase the expected variance of $\widehat{F}$ (ROBERTSON, 1975a,b). Moreover, even simple models of divergence by drift (NEI and CHAKRAVARTI, 1977), island models (NEI *et al.*, 1977), or stepping stone models of dispersal (NEI and MARYUYAMA, 1975) inflate the expected variance, making LEWONTIN and KRAKAUER's (1973) test unreliable in most cases (LEWONTIN and KRAKAUER, 1975).

More recently, BOWCOCK *et al.* (1991) studied the world-wide human genetic differentiation based on DNA polymorphism. Simulating a reasonably well supported evolutionary scenario of divergence, they evaluated the theoretical distribution of F_{ST} conditional on initial gene frequencies. Among 100 nuclear RFLP markers a number of genes exhibited lower or, more often, higher variation than expected under neutrality. In an important paper, BEAUMONT and NICHOLS (1996) proposed a method based on the analysis of the expected distribution of F_{ST} conditional on heterozygosity rather than allele frequency. The conditional distribution, built under an island model

of population structure, is remarkably robust to a wide range of alternative models (colonisation, stepping-stone). Interestingly, departures from equilibrium do not alter much the expected distribution, whenever F_{ST} is less than 0.5. Yet, unequal numbers of immigrants per generation over the whole population generated some discrepancies with the symmetric island model for heterozygosities in the range [0.1, 0.5] (see Figure 3d in BEAUMONT and NICHOLS, 1996).

Thus, their approach might be flawed whenever the true population history consists of repeated branching events, or when the connectivity of populations is uneven. However, we can not infer patterns of migration or historical branching, and test for the homogeneity of the markers with the same data. This is what FELSENSTEIN (1982) described as the « infinitely many parameters » problem. A solution to this problem is to restrict attention to simple but realistic scenarios which may apply to any *pair* of populations (ROBERTSON, 1975b; TSAKAS and KRIMBAS, 1976). This reduces the number of parameters in the model. Here, we develop a model of population divergence. We define population-specific parameters, as functions of probabilities of identity for pairs of genes taken within or among populations. These parameters are simply related to the ratio of divergence time over effective population size. We construct simple estimators of these population-specific parameters. We then examine the expected joint distribution of these estimators, under a wide range of neutral scenarios of divergence. This suggests a new method to assess the homogeneity of response of genetic markers from empirical data. Finally, we apply our new method to a data set of allozyme loci from *Drosophila simulans* populations, and compare our results

to those obtained by using BEAUMONT and NICHOLS's (1996) method.

THE MODEL

We consider two haploid populations of constant sizes N_1 and N_2, which completely separated τ generations ago, from a single population of stationary size N_0. By complete separation, we mean that the populations did not exchange any migrants between the time of the split and the present. We do not assume that the common ancestral population was at equilibrium when it split. Instead, we allow the ancestral population to have gone through a bottleneck, τ_0 generations before present (with $\tau_0 > \tau$). Before this, the ancestral population was at mutation-drift equilibrium, with constant size N_e. Generations do not overlap. New mutations arise at a rate μ, and follow the infinite allele model (IAM). This model of population divergence is illustrated in Figure 1.

[FIGURE 1 about here.]

Let $Q_{w,i}$ be the probability that two genes sampled at random within population i are identical by descent (IBD) and Q_a, the probability that a gene sampled at random from population 1 is IBD to a gene sampled at random from population 2. IBD probabilities are defined as the probabilities that two genes have not mutated since their most recent common ancestor (MALÉCOT, 1975). The probability that a pair of genes are IBD is equal to the probability that these genes are identical in state (IIS), whenever the mutation process follows the IAM.

More generally, let Q_h denote the IBD probability of any pair of genes: $h = (w, i)$ when two genes are sampled within population i, or $h = a$ when

one gene is sampled from each population. It is possible to give an expression for Q_h, as a function of the coalescence time (SLATKIN, 1991). Under a continuous time approximation (HUDSON, 1990)

$$Q_h = \int_0^\infty \gamma^t \mathrm{c_h}(t)\mathrm{d}t \tag{2}$$

where $\mathrm{c_h}(t)$ is the probability of coalescence at t for a pair of genes of type h, and $\gamma = (1-\mu)^2$. The waiting time for a coalescent event in a population of size N_i has an exponential distribution with mean N_i. The IBD probability for a pair of genes in population i reduces to

$$Q_{w,i} = \int_0^\tau \frac{\gamma^t}{N_i}\mathrm{e}^{-t/N_i}\mathrm{d}t + (1-C_i)Q_0 \tag{3}$$

where Q_0 is the IBD probability for two genes sampled at random from the common ancestral population at time τ (just before the split), and $(1-C_i) = \gamma^\tau \cdot \mathrm{e}^{-\tau/N_i}$ is the probability that the two genes neither coalesce nor mutate in the i^{th} population, in the time-interval $0 < t \leqslant \tau$. The first term on the right-hand side of equation (3) is the probability that the two genes coalesce in the time-period $0 < t \leqslant \tau$, and are IBD. Following equation (2), the IBD probability for a pair of genes sample at random from the common ancestral population just before the split at time τ is given by

$$Q_0 = \int_\tau^{\tau_0} \frac{\gamma^{t-\tau}}{N_0}\mathrm{e}^{-(t-\tau)/N_0}\mathrm{d}t + (1-C_0)\int_{\tau_0}^\infty \frac{\gamma^{t-\tau_0}}{N_e}\mathrm{e}^{-(t-\tau_0)/N_e}\mathrm{d}t \tag{4}$$

where $(1-C_0) = \gamma^{\tau_0-\tau} \cdot \mathrm{e}^{-(\tau_0-\tau)/N_0}$ is the probability that the two genes neither coalesce nor mutate in the time-interval $\tau < t \leqslant \tau_0$. The first term

on the right-hand side of equation (4) averages over the coalescent events occurring during the population bottleneck. During this time-interval ($\tau < t \leqslant \tau_0$) the waiting time for a coalescent event is exponentially distributed with mean N_0. The last term in equation (4) averages over coalescent events occurring in the ancestral population, at mutation-drift equilibrium. This last term represents the IBD probability for two randomly sampled genes in a stationary population of size N_e, which is $1/(1+\theta)$, with $\theta = 2N_e\mu$. Solving the integrals in the low-mutation limit (where $\gamma^t \approx \mathrm{e}^{-2\mu t}$), we find that the solution of equation (3) is

$$Q_{w,i} \approx \frac{1}{\theta_i + 1}\left[1 - \mathrm{e}^{-T_i(\theta_i+1)}\right] + \mathrm{e}^{-T_i(\theta_i+1)} \cdot Q_0 \tag{5}$$

where $\theta_i = 2N_i\mu$ and $T_i = \tau/N_i$. The value of Q_0 is given by the solution of equation (4)

$$Q_0 \approx \frac{1}{\theta_0 + 1}\left[1 - \mathrm{e}^{-T_0(\theta_0+1)}\right] + \mathrm{e}^{-T_0(\theta_0+1)}\left(\frac{1}{\theta + 1}\right) \tag{6}$$

where $\theta_0 = 2N_0\mu$ and $T_0 = (\tau_0 - \tau)/N_0$. The probability for a gene in population 1 to be IBD with a gene in population 2 is just given by

$$Q_a = \gamma^\tau Q_0 \tag{7}$$

Obviously, two such genes can not coalesce during the τ generations between the moment of divergence and the present. They are IBD only if their respective ancestors are IBD when populations 1 and 2 diverge, and furthermore, if they do not undergo mutation during the divergence. Now, it is useful to consider the parameter

$$F_i = \frac{Q_{w,i} - Q_a}{1 - Q_a} \tag{8}$$

It is worth noting that the weighted sum of F_i over the two populations gives the intraclass correlation for the probability of identity by descent for genes within populations, relatively to genes between populations. This is of particular interest, because the properties of the intraclass correlations for the probability of identity in state (« IIS correlations ») (COCKERHAM and WEIR, 1987) can be deduced from the properties of the corresponding intraclass IBD correlations, in the low-mutation limit (ROUSSET, 1996). Indeed, such ratios of identity probabilities of the form of equation (8) give the same low-mutation limit, whether one considers the infinite allele model or other mutation models (ROUSSET, 1996, 1997).

If we neglect new mutations arising during the divergence process, Q_a reduces to Q_0 and $Q_{w,i} = C_i(1 - Q_0) + Q_0$. Thus

$$F_i \approx 1 - \mathrm{e}^{-T_i} \tag{9}$$

Note that equation (9) gives a well known result when both daughter populations are assumed to have the same size N, so that $F_1 = F_2 = F \approx 1 - \mathrm{e}^{-\tau/N}$ (see, *e.g.*, REYNOLDS *et al.*, 1983). Hereafter, the parameter T_i will be referred to as a the « branch length » of population i. An important result is that, in the low-mutation limit, the new parameters F_1 and F_2 do not depend on the « nuisance parameters » θ or T_0. This suggests that a simple moment-based estimator $\widehat{T}_i$ of branch length can be derived as

$$\widehat{T}_i = -\ln(1 - \widehat{F}_i) \tag{10}$$

where $\widehat{F}_i$ is an estimator of F_i (see Appendix for details).

PROPERTIES

Simulation procedure: For each set of parameter values, a sequence of artificial data sets was generated using standard coalescent simulations, as described by, *e.g.*, HUDSON (1990). The simulations were performed as follows (see Figure 1 for an illustrated example of one simulated genealogy). For each population, the genealogy of a sample of n_i genes is generated for a period of time ranging from present to τ generations in the past. During this period, all the coalescent events are separated by exponentially distributed time-intervals, with means $N_1/\binom{n_1}{2}$ in population 1 and $N_2/\binom{n_2}{2}$ in population 2 (See Equation 3). At time τ, the number n_0 of lineages that remain represents the ancestors of all the genes sampled in populations 1 and 2. The genealogy of these lineages is generated for the time-period $[\tau, \tau_0]$, and all the coalescence events are separated by exponentially distributed time-intervals, with mean $N_0/\binom{n_0}{2}$ (see the first term in the right-hand side of equation 4). At time τ_0, the lineages that remain are the ancestors of all the genes sampled in populations 1 and 2. The genealogy of these n_e genes is generated for the period $[\tau_0, +\infty]$, with all coalescent events separated by exponentially distributed time-intervals with mean $N_e/\binom{n_e}{2}$ (see the second term in the right-hand side of equation 4). Once the complete genealogy is obtained, the mutation events are superimposed on the coalescent tree of lineages. In the results which follow, each artificial data set consisted of two

(haploid) samples of size $n = 100$, one from population 1 and the other from population 2.

Simulation results: By calculating the estimators $\widehat{F}_1$ and $\widehat{F}_2$ for each of these artificial data sets, it was possible to obtain a close approximation to the expected distribution of these estimators (see Appendix for details). Figure 2 shows this expected joint distribution of $\widehat{F}_1$ and $\widehat{F}_2$, for various combinations of the nuisance parameters θ and T_0. In this case, the « true » branch lengths were $T_1 = T_2 = 0.1$ (hence $F_1 = F_2 \approx 0.0953$). The expected value of the estimator $\widehat{F}_1$ (resp. $\widehat{F}_2$) was always close to the value of the parameter F_1 (resp. F_2). One can show that, by construction, the points $\left(\widehat{F}_1, \widehat{F}_2\right)$ lie within the upper-right triangle with vertices (1,1), (-1,1) and (1,-1). The joint distribution of these two statistics has a negative correlation. Most importantly, it is clear from this figure that the joint distribution of $\widehat{F}_1$ and $\widehat{F}_2$ depends strongly on the nuisance parameters, even though their expectations remain close to the true values of F_1 and F_2.

[FIGURE 2 about here.]

It can be seen that, for smaller values of T_0, the joint distribution becomes tighter as θ increases. On the other hand, for larger values of θ, the distribution is found to widen as T_0 increases. In both cases, it is the level of variation that remains before divergence which is crucial in shaping the joint distribution. With small θ and large T_0, the lineages coalesce rapidly before the divergence, and the number of distinct mutations (allelic states) that can be maintained is small. In this case, the variance of the estimates of populations branch lengths is large, as illustrated by the wide joint distribu-

tion of $\widehat{F}_1$ and $\widehat{F}_2$. Therefore, the joint distribution of $\widehat{F}_1$ and $\widehat{F}_2$ is not ideal for investigating the homogeneity of results of a set of molecular markers. Indeed, other factors such as heterogeneous mutation rates across loci may be invoked to explain disparities of branch length estimates among markers. Fortunately, this problem can be overcome by considering the joint distribution of $\widehat{F}_1$ and $\widehat{F}_2$, conditional upon the total number k of allelic states in the pooled sample at each locus. Figure 3 shows the estimated joint distribution for $T_1 = T_2 = 0.1$ (hence $F_1 = F_2 \approx 0.0953$), conditioned on $k = 4$. The combinations of nuisance parameter values are the same as in Figure 2.

[FIGURE 3 about here.]

The expected joint conditional distribution appears to be almost independent on the nuisance parameters. So, given the observed values for the parameters F_1 and F_2, and given the number of alleles in the sample, one can obtain the conditional joint distribution, and then a high probability region, that should contain 95% of the observed measures of pairwise $\widehat{F}_i$'s values. This result provides the justification for using the conditional distributions to analyze the homogeneity in the patterns of genetic differentiation revealed by a (large) set of markers.

APPLICATIONS

In this section, we present a methodology for identifying outlier loci by a pairwise analysis of populations. For each pair of populations (i, j), we suggest the following protocol:

1. For all loci, the statistics $\widehat{F}_i$ and $\widehat{F}_j$ are computed (see Appendix).

2. The parameters F_i and F_j are estimated as the averages among loci weighted by the heterozygosities $(1 - \widehat{Q}_i)$ and $(1 - \widehat{Q}_j)$, respectively (see Appendix). This corresponds to the weighting of loci suggested by WEIR and COCKERHAM (1984) for the multilocus estimator of F_{ST}.

3. The expected joint distribution of $\widehat{F}_i$ and $\widehat{F}_j$ is generated by performing 10000 coalescent simulations for a given set of nuisance parameters values. This is repeated using a wide range of values for the nuisance parameters. In the *Drosophila simulans* data set discussed below, all the pairwise combinations for θ and T_0 where performed, with $\theta = 1$, 5 or 10, and $T_0 = 0.01$, 0.1 or 1. Thus, a total of 90000 coalescent simulations were performed in this example. The simulated sample size are chosen to be representative of those actually realized in the real data set.

4. For each expected joint distribution of $\widehat{F}_i$ and $\widehat{F}_j$, we construct all the distributions, conditional on the number of allelic states k in the pooled sample, for $k = 2, 3, \ldots$ (The pooled sample is the sample obtained by pooling the samples from populations i and j). Remember, there is one expected distribution for each set of nuisance parameters values. For each conditional distribution, we identify the « high probability » or « high density » region, in the range of the points $\widehat{F}_i$ and $\widehat{F}_j$, where 95% of the data is expected to lie (see Appendix for the construction of this high probability region).

5. For each value of the number of allelic states in the pooled sample, we superimpose a scatter plot of the observed data points (pairs of $\widehat{F}_1$ and $\widehat{F}_2$ values) over an outline of the 95% high probability region, in order to identify outlier loci.

***Drosophila simulans* data set:** We applied this method to a *Drosophila simulans* data set, described in SINGH *et al.* (1987) and CHOUDHARY *et al.* (1992). The raw data set was kindly provided by R. S. Singh and R. A. Morton. Among 111 allozyme loci, 43 were found to be polymorphic in the 5 populations studied in Europe and Africa. The samples consisted in isofemale lines maintained in the laboratory. The haploid sample sizes ranged from $n = 26$ to $n = 55$. Figure 4 shows the analysis performed on a particular pair of populations (France and Tunisia). The multilocus estimates of the parameters F_1 (French population) and F_2 (Tunisian population) were 0.0064 and 0.0617, respectively. The expected distributions with these averaged values, conditioned on the number of alleles in the pooled sample, are plotted with the actual monolocus pairwise $(\widehat{F}_1, \widehat{F}_2)$ estimates.

[FIGURE 4 about here.]

In the great majority of cases, the points fall within the 95% confidence region. With 43 loci we would expect two ($0.05 \times 43 \approx 2$) to lie outside the region by chance. But considering the joint distributions for loci with 3 or more alleles, we found 4 loci that clearly lie outside. Caution is required in the case of loci which lie on the borders of the possible range (Figure 4B). These correspond to loci that have an allele fixed in one population. Slight variations in the nuisance parameters can increase or decrease the relative proportion of loci that may fix one allele in a population. Indeed, we found some conditions under which the 95% envelope contained these two loci. This problem can remain even when we condition on the observed number of alleles. On the other hand, two other loci (coding for Glutamate Pyruvate Transaminase and Carbonic Anhydrase-3) are clear outliers of the expected

distributions (Figures 4C and 4D). In all pairwise comparisons which included the French population, these two loci fell either ourside, or on the edges of the 95% high probability region.

[FIGURE 5 about here.]

In all the pairs which included the population from Congo, two loci coding respectively for the Larval Protein-10 (Pt-10) and the Phosphoglucomutase (PGM) were found to lie outside or on the limit of the 95% high probability region (Figure 5). The locus coding for the Larval Protein-10 systematically gives a longer estimated branch length for this African population than do all other loci, while it gives similar branch lengths to other loci for the other populations. This suggests that genetic variation has been severely reduced by a factor other than genetic drift in this African population. The locus coding for Phosphoglucomutase gives a longer branch length estimate than the other loci in three cases (Figures 5A-C), and a shorter one in one case (Figure 5D). The locus coding for Phosphoglucomutase was also found to lie outside the limit of the 95% high probability region, in all the pairs which included the population from Seychelle Island (Figure 6). In order to strengthen our presumption that these loci were outside the limit allowed by a neutral model, we checked whether these loci also lie outside the limit of the 99% high probability region. The same results were obtained. For these loci, we did not find any plausible neutral scenario of divergence by drift which could provide such a scatter of points. We thus conclude that natural selection may have acted on these loci, or on closely linked regions within the genome.

[FIGURE 6 about here.]

We are more cautious about claiming that the loci coding for Glutamate Pyruvate Transaminase and Carbonic Anhydrase-3 have been or are subject to selection. These loci are clear outliers in some pairwise comparisons involving the French population, but only fall in the limits of the confidence region in other comparisons. Moreover, when considering 99% confidence regions instead of 95% confidence regions, some loci were no longer detected as outliers, but rather as lying on the edges of the confidence limit. The locus coding for isocitrate dehydrogenase-1 was found to be an outlier in three (out of four) pairs which included the population from Seychelle Island. Overall, six more loci were detected as outliers, in single pairwise comparisons. Therefore, we should be very careful in considering those latter loci as being under selection. Indeed, if a locus has responded to selection in one particular contemporary population since it became isolated, then we expect this locus to show up as an outlier in all (or most) comparisons involving this population. This pattern is exactly what we found for the two loci coding for Larval Protein-10 and Phosphoglucomutase in the Congo and Seychelle Island populations.

Evaluating the robustness of this method to the assumptions of the model: In the data set discussed above, it is likely that the populations of *D. simulans* have exchanged migrants after divergence. More generally, one can wonder whether complete isolation and divergence by random drift accurately describes natural situations. An alternative approach would be to develop a new model of population divergence, that allows subsequent migration after separation. But if we want to make inferences about a more realistic (and hence a more complex) model of divergence, then we need to

distinguish between the pattern of genetic differentiation which results from (*i*) recent separation followed by very little migration or (*ii*) ancient separation followed by a moderate amount of migration. This is a difficult task, that would require more powerful methods for inferring parameter values (*e.g.*, maximum likelihood; see NIELSEN and SLATKIN, 2000) that would be much more time consuming. Further note that NIELSEN and SLATKIN (2000) assume that the mutation rate is zero.

So, we are interested in testing if our method (which assumes evolution in complete isolation after divergence) is undermined when applied to pairs of populations that still exchange genes after divergence. It should be borne in mind that gene flow, like genetic drift, affects the whole genome in the same way. We generated artificial datasets under neutral models of population divergence, including high mutation rates and moderate levels of migration between populations. We used a modified version of the algorithm described by HUDSON (1990), that accounts for symmetric migration between populations. Considering populations 1 and 2 altogether, all events (coalescence and migration) are exponentially distributed with mean $N_1N_2/\left[N_2\binom{n_1}{2}\cdot N_1\binom{n_2}{2}+m(n_1+n_2)N_1N_2\right]$, where m is the backward migration rate (NORDBORG, 2001). Conditionally on the occurrence of one event, two genes coalesce in population 1 (resp. population 2) with probability $N_2\binom{n_1}{2}/\left[N_2\binom{n_1}{2}\cdot N_1\binom{n_2}{2}+m(n_1+n_2)N_1N_2\right]$ (resp. $N_1\binom{n_2}{2}/\left[N_2\binom{n_1}{2}\cdot N_1\binom{n_2}{2}+m(n_1+n_2)N_1N_2\right]$) or one gene migrate from population 2 to population 1 (resp. from population 1 to population 2) with probability $m\cdot n_1/\left[N_2\binom{n_1}{2}\cdot N_1\binom{n_2}{2}+m(n_1+n_2)N_1N_2\right]$ (resp. $m\cdot n_2/\left[N_2\binom{n_1}{2}\cdot N_1\binom{n_2}{2}+m(n_1+n_2)N_1N_2\right]$) (see NORDBORG, 2001; STROBECK,

1987; TAKAHATA, 1988).

For each set of parameters, we generated 20 datasets composed of two samples ($n_1 = n_2 = 50$) of 50 loci each. The parameter values are given in Table 1. For each dataset, we applied our method as described above. We generated joint distributions, conditional on the number of alleles, according to the actual numbers of alleles in each sample. For all sets of parameters, we grouped loci with 8 alleles and more in a single class. The number of joint conditional distributions generated per artificial dataset (*i.e.*, the number of classes for different numbers of alleles) ranged from 3 to 7. For each dataset, over all the joint conditional distributions taken together, we expected to detect $0.05 \times 50 = 2.5$ outlier loci, just by chance. We performed Wilcoxon's signed-rank tests (see, *e.g.*, MENDENHALL *et al.*, 1990) to determine if the distribution of the number of detected outlier loci was shifted to the right of 2.5 (one-tailed test).

[**TABLE 1** about here.]

Table 1 shows the total observed number of outlier loci (mean and median over 20 independent simulated datasets) detected for a range of nuisance parameter values (low and high mutation rates, short or long divergence by random drift, with or without migration). In no case could we reject the null hypothesis that the number of detected outlier loci was equal to 2.5 (against the alternative hypothesis that the number of detected outlier was greater than 2.5). Thus, our approach is conservative in the sense that the 95% confidence region contains at least 95% of the loci generated by a truly neutral model. At the level of 5% we do not (falsely) detect outlier loci in a sample of neutral markers (type I error).

Comparison with BEAUMONT and NICHOLS**'s (1996) method:** We also applied BEAUMONT and NICHOLS's (1996) procedure to the *D. simulans* data set. Based on a preliminary examination of the data, 3 loci (coding for α-Fucosidase, Dipeptidase-1 and Mannose Phosphatase Isomerase) were found to lie outside the 95% confidence region of the conditional joint distribution of $\widehat{F}_{ST}$ and mean heterozygosity. The percentiles were determined as described in BEAUMONT and NICHOLS (1996). Surprisingly, none of these 3 loci were detected as outliers using our method. There may be several reasons for this:

We suspect that, in the present case, the inclusion of a very distant insular population (Seychelle Island) may bias their analysis. Indeed, populations heterogeneous with respect to their demographic parameters (effective population sizes and migration rates) have been shown to strongly affect their method (BEAUMONT and NICHOLS, 1996). Isolation (low migration rates) together with population bottlenecks, can introduce a further bias. Consider as an extreme case, the fixation of a private allele at some locus in one population. This may be unexpected for a polymorphic locus in a mutation-migration-drift equilibrium model, unless there is a strong asymmetry, with some populations being smaller and receiving less immigrants than others. However, this is not unexpected for a model of separation and isolation, where there has been population bottlenecks. This may boost the F_{ST} estimate at some locus, and thus exclude it from the 95% high probability region. So, isolated populations should probably be excluded from BEAUMONT and NICHOLS's (1996) analysis.

Moreover, in general, the loci which were outliers in our analysis gave small values of (global) F_{ST}. But from the shape of the joint distribution

of F_{ST} and heterozygosity, it seems that BEAUMONT and NICHOLS's (1996) analysis is likely to detect outlier loci which exhibit unusually large F_{ST} values. However, a process which would cause an apparent decrease of genetic variation at one locus in a single local population, without leading to a decrease of the variation over all populations, would not be detected BEAUMONT and NICHOLS's (1996) procedure. In other words, if selection acts on one locus at a local scale, pairwise comparisons of populations is more likely to be efficient for detecting outlier loci.

DISCUSSION

Using population-specific estimators of branch lengths: Conventional pairwise genetic distances or pairwise measures of population differentiation are based on the assumption that the sizes of populations are equal and constant through time or that dispersal, if any, is symmetric. For example, the pairwise F_{ST} parameter is defined as a ratio of identity probabilities within and among populations. But the within-population term is taken as an average over the pair of populations. Thus, the definition of the parameter implicitly assumes that both populations share the same demographic parameters. WEIR and COCKERHAM's (1984) estimator θ of F_{ST} is constructed to have low bias and variance, assuming that the populations are independent replicates of the same stochastic process. This means that populations are supposed to have the same size, and that they do not exchange migrants. Without these assumptions, θ would be a complex function of unequal (within-population) identity probabilities.

In contrast, the F_i parameters defined here make sense even when the

populations are of unequal size. The only assumption we make is that when the two populations have separated, they remain completely isolated. From the estimation of F_i's for a pair of populations, we can infer the branch lengths. The ratio of these branch length estimates is inversely proportional to the ratio of effective population sizes. Thus, these estimates may be seen as measures of the intensity of genetic drift that has occurred since population divergence. The main drawback to this approach is that when estimates of IIS probabilities are smaller within populations than among (*i.e.*, $\widehat{Q}_{w,i} < \widehat{Q}_a$), $\widehat{F}_i$ becomes negative, and the moment-based estimator of branch length fails. Although this can arise just by chance for some loci, averaging $\widehat{Q}$ estimates over loci reduces the problem.

Provided that we obtain good estimates of branch lengths for a pair of populations (which requires the pooling of information from many independent loci) we may be able to evaluate the consistency of locus-specific estimates. Indeed, the joint distribution of branch length estimates, conditioned on the number of alleles in the pooled sample, depends only weakly on nuisance parameters of the simple model of divergence by drift. In particular, this conditional distribution is not sensitive to departures from mutation-drift equilibrium before isolation, or to differences in mutation rates.

Detection of selection acting on genetic markers: We saw from the analysis of the *D. simulans* data set that the great majority of loci always fall in the confidence region of the conditional pairwise distributions of branch length estimates, while some loci do not. Overall, we identified two loci that were probably subject to selection in the population from Congo. We concluded that the distribution of variability at these loci may be shaped by

other forces than mutation and drift. Furthermore, we identified two other loci that either lie on the edges, or fall just outside the high probability region of the expected conditional distribution in the French population, although we should be cautious about these latter loci. It is noteworthy that our estimation of the density of F_i parameters (see Appendix) is discontinuous, because of the discrete nature of the data (the allele counts). This is particularly true when the number of alleles on which the distribution is conditioned is small (for a given set of parameters, the lower the number of allelic states, the more discontinuous the null distribution: see Figure 4). Using discrete distributions is clearly preferable to using some (unnecessary) continuous approximations to it. Moreover, whenever the null distribution is based on the same number of allelic states and the same number of genes as in the sample, there is no tendency for loci to show up as outlier just because of the discrete nature of the distribution (*i.e.*, a locus can not, by construction, shows up *between* arc-shaped areas, located at the edge of some distributions). Yet, when an apparent outlier lie very close to the 95% high probability region, it is highly advisable to check whether this locus also lie outside the 99% high probability region.

The main criticisms of LEWONTIN and KRAKAUER's (1973) attempts to interpret across-loci heterogeneity of F_{ST} values arose from their failure to consider allele frequencies as random variables, whose distribution depends on the underlying model of population structure and history. Indeed, uneven patterns of dispersal among populations (NEI and MARYUYAMA, 1975) or sequences of population splits within the species (ROBERTSON, 1975a,b) may strongly undermine the approach. LEWONTIN and KRAKAUER (1975)

acknowledged that their tests might be limited to situations were the true population structure did not depart too much from the island model.

However, conditioning the distribution of F_{ST} on the heterozygosity (BEAUMONT and NICHOLS, 1996) or on gene frequency for biallelic loci (BOWCOCK *et al.*, 1991) has been shown to give surprisingly robust results, in the sense that strong departures from the model assumptions do not alter very much the distribution. The strongest effect on the joint expected distribution of F_{ST} and heterozygosity occurs when populations are heterogeneous with respect to their demographic parameters (BEAUMONT and NICHOLS, 1996), for example when populations are founded by very different numbers of individuals, or when populations are arranged in an irregular stepping-stone lattice. However, BEAUMONT and NICHOLS (1996) considered a large number d of subpopulations in the metapopulation ($d = 100$) and this parameter strongly influences the expected heterozygosity [$H_e \approx 4Nd\mu/(1 + 4Nd\mu)$, for diploids]. In addition, at a local scale, F_{ST} is only weakly influenced by the total population size Nd (ROUSSET, 2001). The number of populations has a stronger role than acknowledged by BEAUMONT and NICHOLS (1996) in determining whether mutation has an effect on F_{ST} or not. It has been shown that, considering smaller numbers of populations, F_{ST} estimates may be reduced by mutation, especially with a stepwise mutation model (see FLINT *et al.*, 1999). With $d = 100$ islands, the sets of parameters used in BEAUMONT and NICHOLS (1996) did not account for any case where mutation may depress F_{ST}.

As already suggested by TSAKAS and KRIMBAS (1976), restricting LEWONTIN and KRAKAUER's (1973) tests to pairs of populations removes all kinds

of dependence on the unknown population structure. Indeed, whatever their history, two populations ultimately descend from a single ancestral one in the past. Still, nuisance parameters may broaden the joint distribution of pairwise F_i's (Figure 2). However, conditioning on the number of alleles (Figure 3) also gives distributions that are robust enough to variations in the values of nuisance parameters. It is obvious that for each analysis of a pair of populations, we deliberately discard the information brought by other populations, which may decrease the power of the method (TSAKAS and KRIMBAS, 1976). But we believe that this enables us to explain a wider range of patterns than any symmetrical model, such as the island model. In this respect, our approach is conservative. Moreover, we found that low or moderate gene flow did not undermine our approach, in the sense that we did not falsely detect outlier loci, when they were neutral (Table 1). We compared and discussed the performance of our method to that of BEAUMONT and NICHOLS's 1996 using the empirical data from SINGH *et al.* (1987) and CHOUDHARY *et al.* (1992). We further tested whether our method would falsely rejected neutral loci (type I error), under a wide range of nuisance parameter values (see Table 1). In particular, since the method assumes that the mutations arising after divergence can be neglected, we checked that high mutation rates do not weaken the approach.

We have found that patterns such as those identified in the Tunisia *vs* Congo data set as evidence of selection, can be produced by « neutral models » where the coalescent process occurs independently at each locus. Indeed, similar scatters of points could be obtained whenever the parameters F_1 and F_2 vary across loci, having particularly high values at certain loci (results

not shown). Models of this type provide a rough approximation to models of unlinked neutral loci, some of which having been strongly influenced by selection (remind that the effect of selection resembles a reduction in the effective population size experienced by these loci, as described by BARTON, 1995, 1998; ROBERTSON, 1961). So, it is certainly plausible that the patterns which we have identified in the Tunisia *vs* Congo data set were produced by selection. A thorough investigation of the conditions under which our method fails to identify selected loci (type II error) would be desirable. However, this is not feasible, as the range of models which incorporate selection is very large.

An important task for the future is to consider a more general neutral model of the divergence of two populations, where gene flow may continue after the moment of « separation ». It is also desirable to extend this approach to more elaborate neutral models, incorporating recombination. More sophisticated estimators of the divergence parameters (branch lengths) would then be required. We assumed that the mutation process follows the IAM and we allowed a wide range of possible mutation rates. In the IAM, genes that are identical in state are also identical by descent. This may not be the case with other mutation models such as with the K allele or stepwise mutation processes, which can produce IIS genes that are not IBD (homoplasy). The IAM is probably an adequate model for allozyme data. It is certainly not so appropriate for potentially more variable markers, such as microsatellites. Recent studies reveal that the processes of mutation of microsatellite markers may be more complex than previously thought and may vary greatly among loci (ESTOUP and ANGERS, 1998). Furthermore, the effect of homoplasy on measures of population subdivisions is not simple (ROUSSET, 1996).

Therefore, further studies should be conducted to test the application of our method across different classes of nuclear markers that differ in processes of mutation. Clearly, if a whole class of marker loci, which are known to have a very distinct mutation process, are identified as outliers by our analysis, then this class of markers should be interpreted with caution.

If we could identify those marker loci that have responded to selection during the process of divergence, then we may be able to obtain improved estimates of the parameters of population structure and history, by excluding these loci (Ross *et al.*, 1999). Our method differs from previous ones in allowing selection to be detected in particular populations, and in some pairwise comparisons but not others. This opens up the possibility that markers may be discarded only in the analysis of those populations where there is evidence that they have responded to selection. It is also of interest to use this approach to screen the genome for regions that have responded to strong selection in the recent past. If populations have diverged phenotypically and if this has been caused by selection, then it may even be possible to identify candidate regions for the Quantitative Trait Loci (QTL) underlying this adaptive divergence.

ACKNOWLEDGEMENTS

We are very grateful to R.S. Singh and R.A. Morton for providing the *Drosophila simulans* data set. We thank I. Olivieri for helpful comments on a previous draft of this manuscript and S. Billiard for valuable discussions about the structured coalescent. We are grateful to two anonymous reviewers who constructively commented on and criticized the manuscript. This work

was funded by the contract number BIO4-CT96-1189 of the Commission of the European Communities (DG XII) to P.B., and R.V. was also partially funded by the Fondation Sansouire. This is publication number 2001-XXX of the Institut des Sciences de l'Évolution de Montpellier.

LITERATURE CITED

BARTON, N. H., 1995 Linkage and the limits to natural selection. Genetics, **140**: 821–841.

BARTON, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res., **72**: 123–133.

BEAUMONT, M. A., and R. A. NICHOLS, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. B, **263**: 1619–1626.

BOWCOCK, A. M., J. R. KIDD, J. L. MOUNTAIN, J. M. HEBERT, L. CAROTENUTO, K. K. KIDD and L. L. CAVALLI-SFORZA, 1991 Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. Genetics, **88**: 839–843.

CAVALLI-SFORZA, L. L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. B, **164**: 362–379.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics, **134**: 1289–1303.

CHARLESWORTH, B., M. NORDBORG and D. CHARLESWORTH, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res., **70**: 155–174.

CHOUDHARY, M., M. B. COULTHART and R. S. SINGH, 1992 A comprehensive study of genic variation in natural populations of *Drosophila*

melanogaster. VI. patterns and processes of genic divergence between *D. melanogaster* and its sibling species, *D. simulans*. Genetics, **130**: 843–853.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics, **74**: 697–700.

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

ESTOUP, A., and B. ANGERS, 1998 Microsatellites and minisatellites for molecular ecology: Theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology*, edited by G. R. CARVALHO. IOS Press, Amsterdam.

FELSENSTEIN, J., 1982 How can we infer geography and history from gene frequencies ? J. Theor. Biol., **96**: 9–20.

FLINT, J., J. BOND, D. C. REES, A. J. BOYCE, J. M. ROBERTS-THOMSON, L. EXCOFFIER, J. B. CLEGG, M. A. BEAUMONT, R. A. NICHOLS and R. M. HARDING, 1999 Minisatellite mutational processes reduce F_{ST} estimates. Hum. Genet., **105**: 567–576.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet. Res., **8**: 269–294.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet., **38**: 226–231.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol., **7**: 1–44.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics, **120**: 831–840.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The hitchhiking effect revisited. Genetics, **123**: 887–899.

LEWONTIN, R. C., and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. Genetics, **74**: 175–195.

LEWONTIN, R. C., and J. KRAKAUER, 1975 Testing the heterogeneity of F values. Genetics, **80**: 397–398.

MALÉCOT, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol., **8**: 212–241.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res., **23**: 23–35.

MENDENHALL, WM, DD WACKERLY and RL SCHEAFFER, 1990 *Mathematical statistics with applications.* PWS-KENT Publishing Company, Boston.

NEI, M., 1972 Genetic distance between populations. Am. Nat., **106**: 283–292.

NEI, M., and A. CHAKRAVARTI, 1977 Drift variance of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. Theor. Popul. Biol., **11**: 307–325.

NEI, M., A. CHAKRAVARTI and Y. TATENO, 1977 Mean and variance of F_{ST} in a finite number of incompletely isolated populations. Theor. Popul. Biol., **11**: 291–306.

NEI, M., and T. MARYUYAMA, 1975 Lewontin-Krakauer test for neutral genes. Genetics, **80**: 395.

NIELSEN, R., and M. SLATKIN, 2000 Likelihood analysis of ongoing gene flow and historical association. Evolution, **54**: 44–50.

NORDBORG, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of statistical genetics*, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. John Wiley & Sons, Ltd, Chichester.

OHTA, T., and M. KIMURA, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res., **13**: 47–55.

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.

ROBERTSON, A., 1961 Inbreeding in artificial selection programmes. Genetical Research, **2**: 189–194.

ROBERTSON, A., 1975a Gene frequency distribution as a test of selective neutrality. Genetics, **81**: 775–785.

ROBERTSON, A., 1975b Remarks on the Lewontin-Krakauer test. Genetics, **80**: 396.

Ross, K. G., D. D. Shoemaker, M. J. B. Krieger, J. DeHeer and L. Keller, 1999 Assessing genetic structure with multiple classes of molecular markers: A case study involving the introduced fire ant *Solenopsis invicta*. Mol. Biol. Evol., **16**: 525–543.

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.

Rousset, F., 2001 Inferences from spatial population genetics, pp. 179–212 in *Handbook of statistical genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Ltd, Chichester.

Singh, R. S., M. Choudhary and J. R. David, 1987 Contrasting patterns of geographic variation in the cosmopolitan sibling species *Drosophila melanogaster* and *D. simulans*. Biochem. Genet., **25**: 27–40.

Slatkin, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res., **58**: 167–175.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics, **103**: 545–555.

Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. Genetics, **117**: 149–153.

Takahata, N., 1988 The coalescent in two partially isolated diffusion populations. Genet. Res., **52**: 213–222.

Tsakas, S., and C. B. Krimbas, 1976 Testing the heterogeneity of F values: A suggestion and a correction. Genetics, **84**: 399–401.

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. Eugen., **15**: 323–354.

APPENDIX

Parameters estimation: For any given allele u, we use the indicator variable x_{iju} for describing the state of the j^{th} gene in the i^{th} population, with $i = (1, 2)$. $x_{iju} = 1$ if the allelic type is u, $x_{iju} = 0$, otherwise. Let p_{iu} be the frequency of allele u in the i^{th} population. Then $p_{iu} = \mathcal{E}(x_{iju}|\ \mathbf{p})$, where $\mathcal{E}(\quad|\ \mathbf{p})$ denotes the expectation, conditional on the array $\mathbf{p}$ of all the allele frequencies. Considering the second moments of the random variable x_{iju}, it follows that $\mathcal{E}(x_{iju}^2|\ \mathbf{p}) = p_{iu}$ and, since individuals are sampled independently from the i^{th} population, $\mathcal{E}(x_{iju}x_{ij'u}|\ \mathbf{p}) = p_{iu}^2$ for $j' \neq j$. Then, summing over all alleles gives the probability for two genes in population i to be identical in state (IIS)

$$Q_{w,i} = \mathcal{E}\left(\sum_{u=1}^{k} p_{iu}^2\right) \tag{A 1}$$

where $\mathcal{E}$ denotes now the expectation over the distribution of allele frequencies $\mathbf{p}$ and k is the number of alleles in the population. The IIS probability for two genes respectively taken in population 1 and 2 is given by

$$Q_a = \mathcal{E}\left[\sum_{u=1}^{k}(p_{1u}p_{2u})\right] \tag{A 2}$$

An unbiased estimator of the frequency of allele u among n_i sampled individuals from the i^{th} population is simply given by $\widehat{p}_{iu} = \sum_{j=1}^{n} x_{iju}/n_i$. Expanding the square of this expression, and then taking expectation, gives $\mathcal{E}(\widehat{p}_{iu}^2|\ \mathbf{p}) = \left[p_{iu} + n_i(n_i - 1)p_{iu}^2\right]/n_i$. Therefore,

$$\widehat{Q}_{w,i} = \sum_{u=1}^{k} \left[\widehat{p}_{iu}\left(n_i\widehat{p}_{iu} - 1\right)\right] / (n_i - 1) \qquad \text{(A 3)}$$

is an unbiased estimator of the probability for two genes in population j to be identical in state, with k being the number of alleles in the sample. Similarly

$$\widehat{Q}_a = \sum_{u=1}^{k} \left(\widehat{p}_{1u}\widehat{p}_{2u}\right) \qquad \text{(A 4)}$$

is an unbiased estimator of the IIS probability of two genes taken in the ancestral population, before divergence. Approximating the expectation of a ratio by the ratio of expectations, an estimator of F_i is given by

$$\widehat{F}_i = \frac{\sum_{u=1}^{k} \left[\widehat{p}_{iu}\left(n_i\widehat{p}_{iu} - 1\right) / (n_i - 1) - \widehat{p}_{1u}\widehat{p}_{2u}\right]}{1 - \sum_{u=1}^{k}\left(\widehat{p}_{1u}\widehat{p}_{2u}\right)} \qquad \text{(A 5)}$$

When combining the information brought by all alleles at more than one locus, a multilocus estimator is defined as the ratio of the sum of locus-specific numerators over the sum of locus-specific denominators (see, *e.g.*, WEIR and COCKERHAM, 1984). It is worth noting that, when daughter population sizes are equal, this simple way to estimate parameters (*i.e.*, equating Qs to $\widehat{Q}$s in equation (8) to get $\widehat{F}$) directly yields Cockerham's estimators (COCKERHAM, 1973; WEIR and COCKERHAM, 1984) developed with the methods of analysis of variance (see ROUSSET, 2001, for a thorough demonstration of the equivalence between estimator formulas based on analyses of variance and expressions in terms of frequency of identical genes). Our estimator differs from previous ones (*e.g.*, REYNOLDS *et al.*, 1983) in allowing separate parameters F_is for each population.

Estimation of the density of F_i parameters: For each set of parameter values, coalescent simulations were performed, thus generating « artificial data sets ». Each artificial data set yields a pair of estimates $\widehat{F}_1$ and $\widehat{F}_2$. An approximation to the expected joint distribution was obtained as follows. First, a 2-dimensional histogram was constructed. Recall that the points $\left(\widehat{F}_1, \widehat{F}_2\right)$ are constrained to lie within the upper-right triangle of a square with vertices (-1,-1), (1,-1), (-1,1) and (1,1). The whole square region was covered by a 2-Dimensional array (or mesh) of 100×100 square cells. Each cell has thus sides of length 0.02. Each observation $\left(\widehat{F}_1, \widehat{F}_2\right)$ was binned in the appropriate cell. The cell counts were divided by the total number of observations, to obtain a discrete probability distribution over the 2-dimensional array. This discrete distribution is a close approximation to the expected joint distribution of the estimators $\left(\widehat{F}_1, \widehat{F}_2\right)$. The q-level « high probability region » ($q = 95\%$, or any other value) is constructed as follows. The cells are sorted in order of decreasing probability. Finally, starting from the cells with the highest associated probabilities, cells are sequentially added to the confidence region, until the cumulative probability of the whole set of cells obtained is equal to (or just exceeds) the chosen q-value.

From this procedure, we obtain for each simulation a region within which a proportion q of the data lies. Notice that this confidence region is not necessarily continuous. Constructing the high probability region using the discrete distribution is clearly preferable to using some (unnecessary) continuous approximation to it.

TABLE 1

Results from applications to various divergence scenarios

			Detected outliers		
μ	θ	T_0	*mean*	*median*	p value
No migration: $m = 0$					
10^{-5}	1	1	1.85	2.0	0.98
10^{-5}	10	10^{-2}	1.15	1.0	1.00
10^{-3}	1	1	2.60	3.0	0.28
10^{-3}	10	10^{-2}	1.75	2.0	0.76
Low migration: $m = 0.01$					
10^{-5}	1	1	2.30	2.5	0.79
10^{-5}	10	10^{-2}	2.25	2.0	0.77
10^{-3}	1	1	2.00	2.0	0.99
10^{-3}	10	10^{-2}	1.20	1.0	1.00
Moderate migration: $m = 0.1$					
10^{-5}	1	1	2.30	2.0	0.87
10^{-5}	10	10^{-2}	2.05	2.0	0.96
10^{-3}	1	1	2.25	2.0	0.89
10^{-3}	10	10^{-2}	1.85	2.0	0.98

For all sets of parameters, 50 loci were scored among 100 haploid sampled individuals (50 in each population). The mean number of detected outlier loci is given, as well as the median of the distribution of that number. We provide the p value of Wilcoxon's signed-rank tests, performed on the distributions of detected outliers, to determine whether this distribution was shifted to the right of 2.5 (one-tailed test).

FIGURE CAPTIONS

Figure 1. A gene genealogy under our model, for $n = 10$ genes sampled in each population. In this example, the parameters values are $N_1 = N_2 = 100$, $N_0 = 500$, $N_e = 1000$, $\tau = 50$, $\tau_0 = 150$ and $\mu = 10^{-3}$.

Figure 2. Expected distribution of pairs of $\widehat{F}_1$ and $\widehat{F}_2$ estimates, for wide ranges of values of the nuisance parameters $\theta = 2N_e\mu$ and T_0. $T_i = \tau/N_i$ is 0.10 for both daughter populations (with $\tau = 50$ and $N_1 = N_2 = 500$), giving an expected value $F_i \approx 0.0953$, as indicated by the dotted lines. For all parameter sets, $\mu = 10^{-4}$ and $N_0 = 1000$. One hundred individuals are sampled in each daughter population. The light gray area defines a region in which 95% of the simulated points are expected to lie (see Appendix for details).

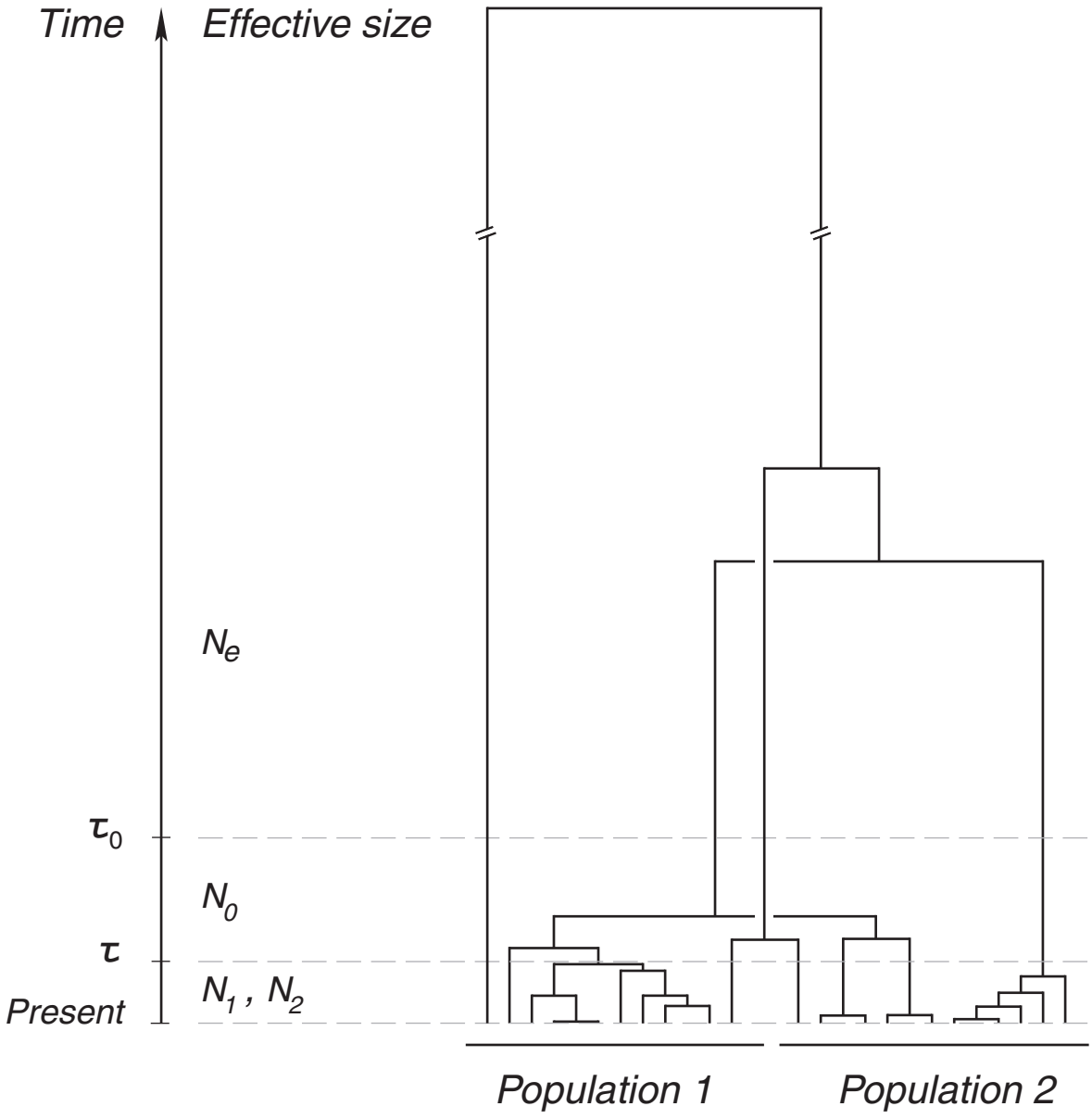
Figure 3. Expected distribution of pairs of $\widehat{F}_1$ and $\widehat{F}_2$ estimates conditioned on a number of alleles in the sample equal to 4. As in Figure 1, wide ranges of values have been used for the nuisance parameters. The dotted lines indicate the expected values for F_1 and F_2.

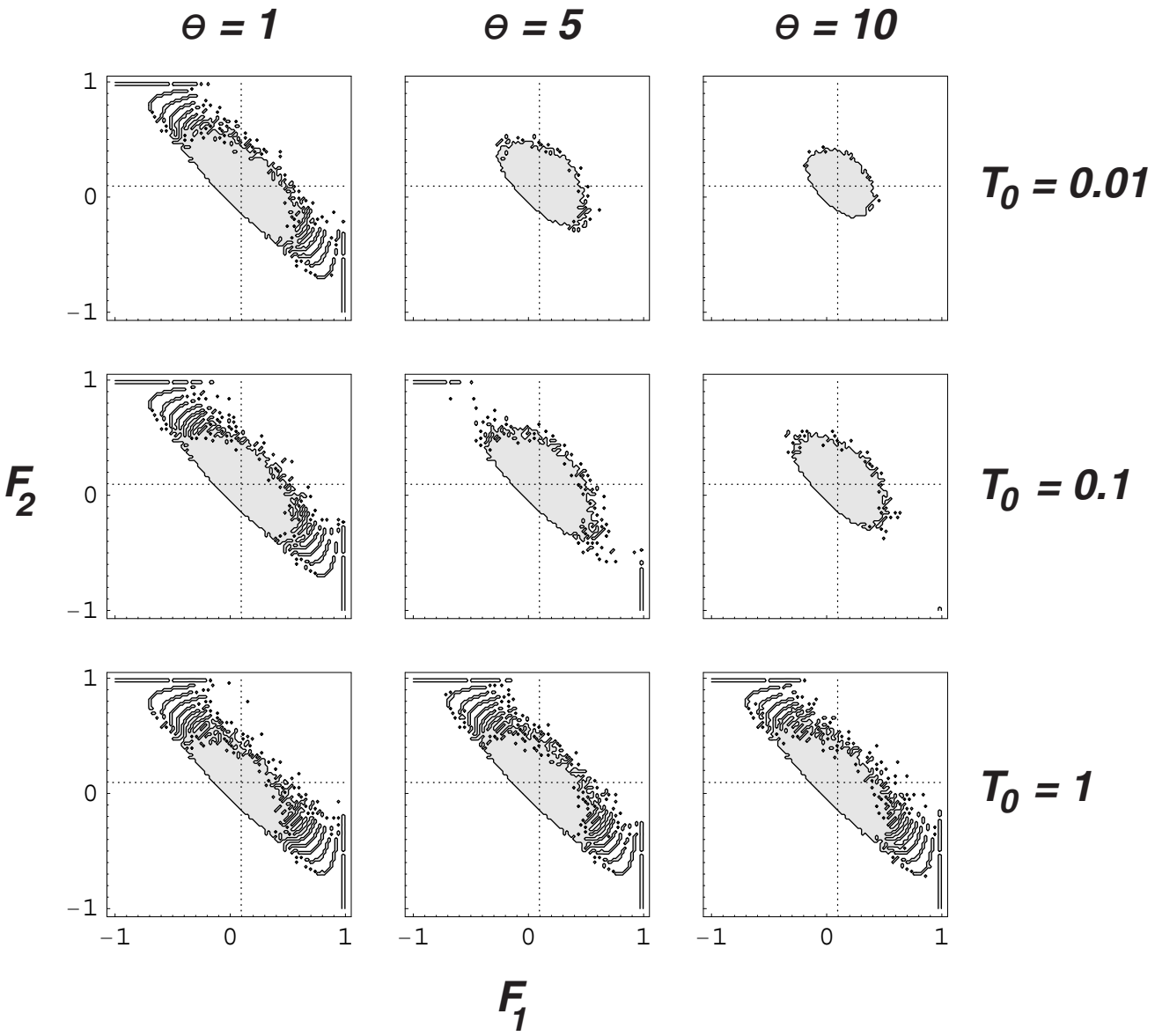
Figure 4. $\widehat{F}_1$ and $\widehat{F}_2$ values estimated from 43 loci in *Drosophila simulans* for the pairwise comparison of the populations from France ($n = 55$) and Tunisia ($n = 52$). n is the number of isofemale lines typed for each enzymatic system (haploid sample size). Each locus is represented with a black dot. The averaged values are $\widehat{F}_1 = 0.0064$ and $\widehat{F}_2 = 0.0617$ as indicated by the dotted lines. Thin lines enclose a region in which 95% of the simulated data points are expected to lie. Four distributions are shown, conditioned on the number of allelic states in the whole sample. A. Expected distribution of pairwise F_i

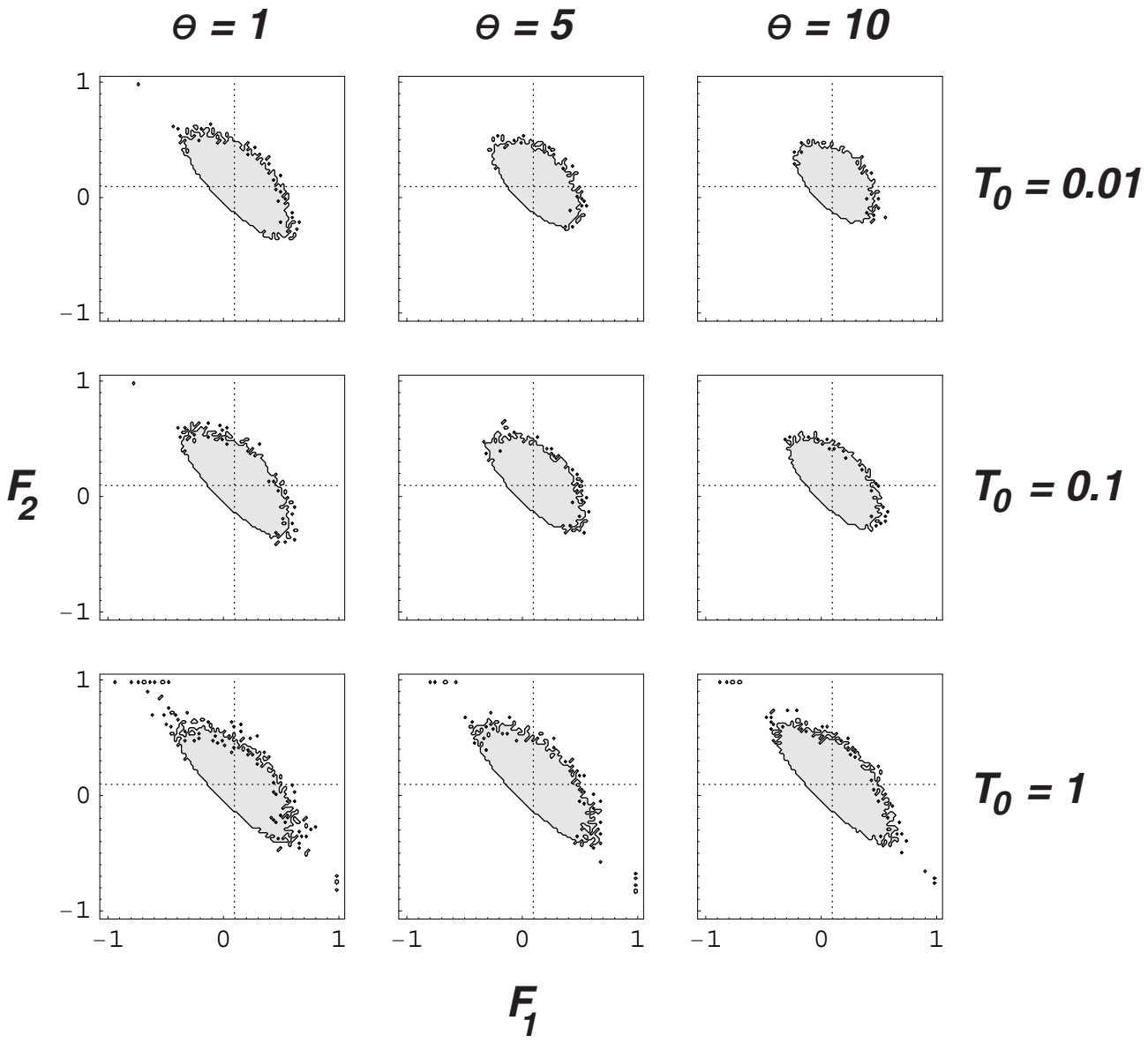
estimates conditioned on a number k of allelic states equal to 2. B. *idem* with $k = 3$. C. *idem* with $k = 4$. D. *idem* with $k = 5$. Black arrows indicate outlier loci. The loci coding for the Glutamate Pyruvate Transaminase (GPT) and Carbonic Anhydrase-3 (Ca-3) are shown respectively in C and D.

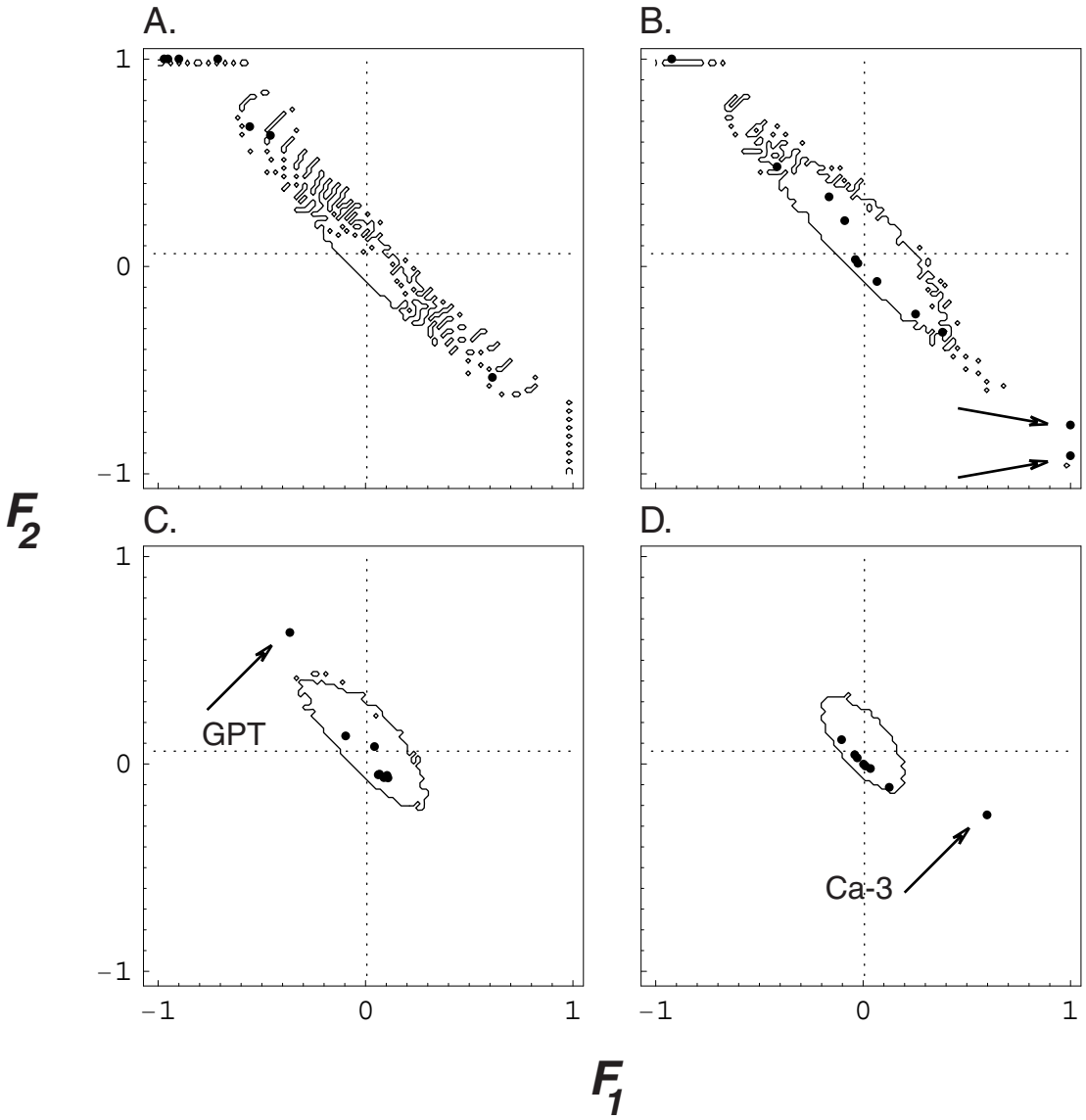
Figure 5. $\widehat{F}_1$ and $\widehat{F}_2$ values estimated from 43 loci in *Drosophila simulans* for all the pairwise comparisons involving the population from Congo ($n = 45$). A. Expected distribution for the populations from France ($n = 55$) and Congo. B. Tunisia ($n = 52$) *vs* Congo. C. Congo *vs* Cape Town, South Africa ($n = 32$). D. Congo *vs* Seychelle Island ($n = 26$). All distributions are conditioned on $k = 4$. Each locus is represented with a black dot. Dotted lines give the expected values for $\widehat{F}_1$ and $\widehat{F}_2$. For each expected conditional distribution, black arrows indicate the loci coding for the Larval Protein-10 (Pt-10) and Phosphoglucomutase (PGM).

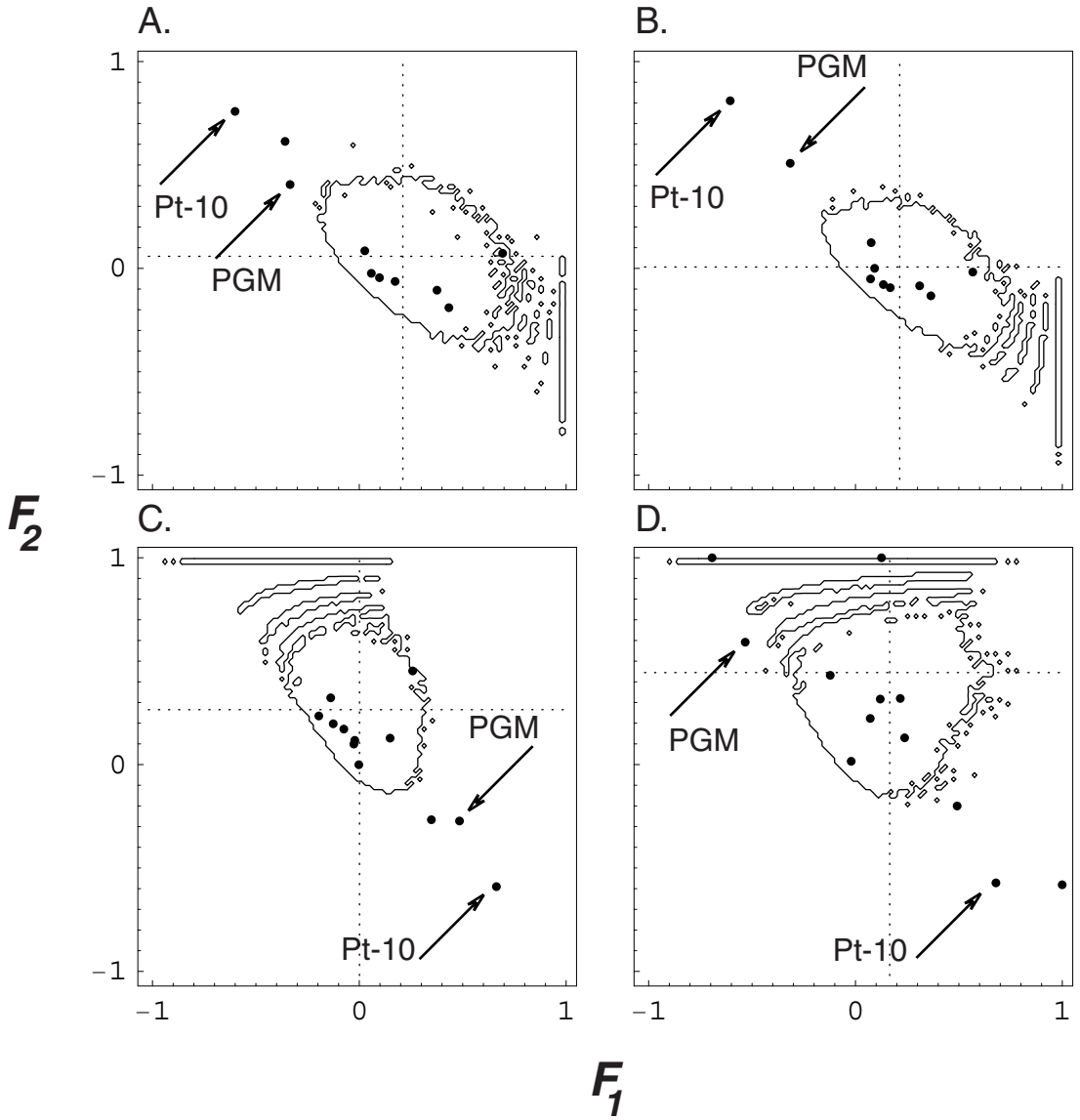
Figure 6. $\widehat{F}_1$ and $\widehat{F}_2$ values estimated from 43 loci in *Drosophila simulans* for all the pairwise comparisons involving the population from Seychelle Island ($n = 26$). A. Expected distribution for the populations from France ($n = 55$) and Seychelle Island. B. Tunisia ($n = 52$) *vs* Seychelle Island. C. Congo ($n = 45$) *vs* Seychelle Island. D. Cape Town, South Africa ($n = 32$) *vs* Seychelle Island. Distributions in A and C are conditioned on $k = 4$ and distributions in B and D are conditioned on $k = 3$. Each locus is represented with a black dot. Dotted lines give the expected values for $\widehat{F}_1$ and $\widehat{F}_2$. For each expected conditional distribution, black arrows indicate the locus coding for Phosphoglucomutase (PGM).

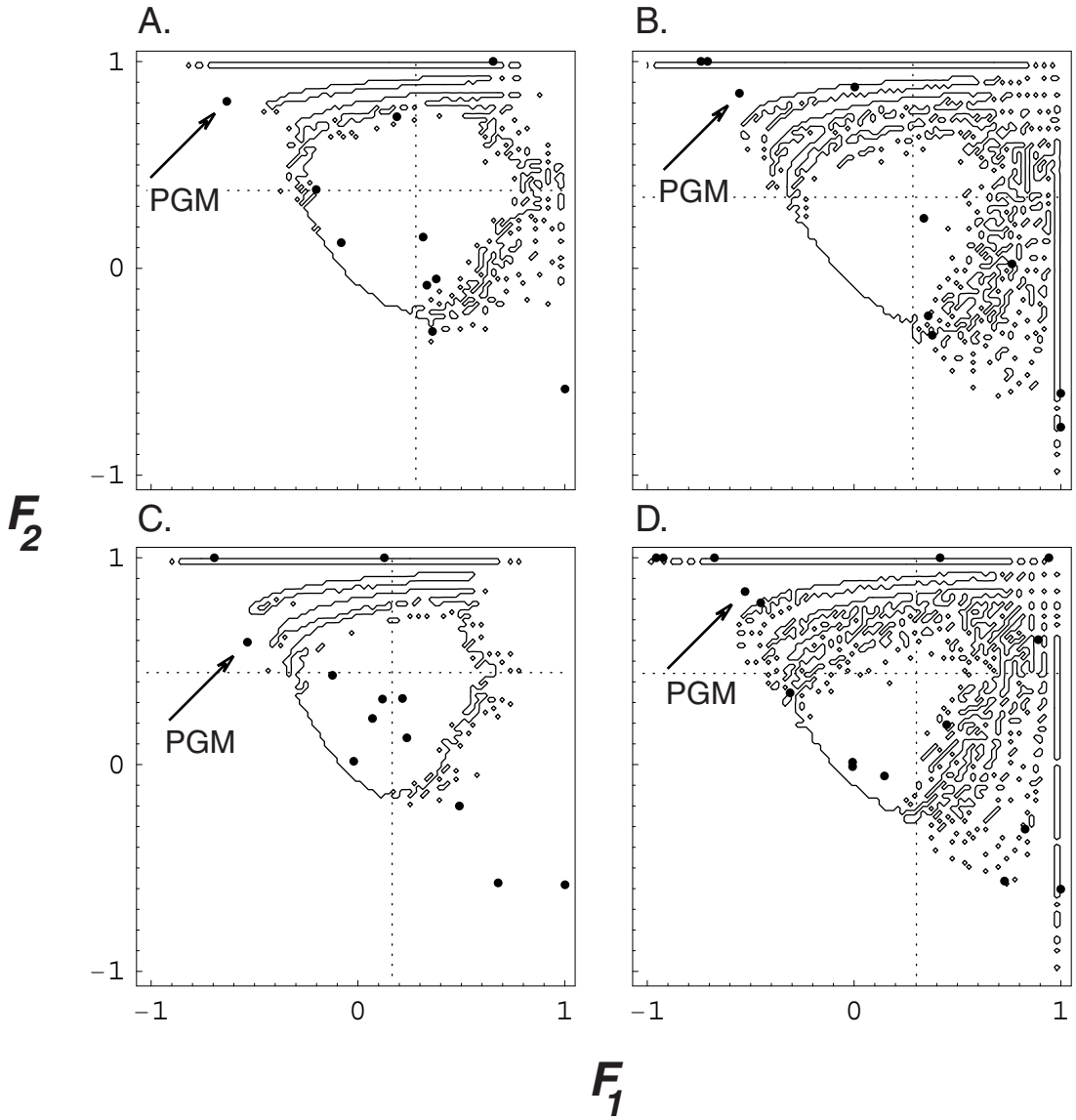

Time
Effective size
N_e
τ_0
N_0
τ
N_1 , N_2
Present
Population 1
Population 2


θ = 1
θ = 5
θ = 10
T0 = 0.01
T0 = 0.1
T0 = 1
F2
F1
1
0
−1
−1
0
1


θ = 1
θ = 5
θ = 10
T0 = 0.01
T0 = 0.1
T0 = 1
F2
F1
1
0
−1


A.
B.
C.
D.
GPT
Ca-3
1
0
−1
−1
0
1
F2
F1


A.
B.
C.
D.
Pt-10
PGM
PGM
Pt-10
PGM
Pt-10
PGM
Pt-10
1
0
−1
−1
0
1
F2
F1


A.
B.
C.
D.
PGM
PGM
PGM
PGM
1
0
−1
1
0
−1
−1
0
1
−1
0
1
F2
F1