DOI: 10.1534/genetics.110.117275

Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended

Maxime Bonhomme,* Claude Chevalet,* Bertrand Servin,* Simon Boitard,* Jihad Abdallah,*,† Sarah Blott‡ and Magali SanCristobal‡,1

**Unité Mixte de Recherche 444 Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique Toulouse, F-31326 Castanet Tolosan Cedex, France, †Department of Animal Production, Faculty of Agriculture, An-Najah National University, Nablus, Palestine and ‡Centre for Preventive Medicine, Animal Health Trust, Kentford, Newmarket, Suffolk CB8 7UU, United Kingdom*

Manuscript received April 2, 2010
Accepted for publication May 22, 2010

ABSTRACT

Detecting genetic signatures of selection is of great interest for many research issues. Common approaches to separate selective from neutral processes focus on the variance of F_{ST} across loci, as does the original Lewontin and Krakauer (LK) test. Modern developments aim to minimize the false positive rate and to increase the power, by accounting for complex demographic structures. Another stimulating goal is to develop straightforward parametric and computationally tractable tests to deal with massive SNP data sets. Here, we propose an extension of the original LK statistic (T_{LK}), named $T_{F\text{–}\mathrm{LK}}$, that uses a phylogenetic estimation of the population's kinship ($\mathcal{F}$) matrix, thus accounting for historical branching and heterogeneity of genetic drift. Using forward simulations of single-nucleotide polymorphisms (SNPs) data under neutrality and selection, we confirm the relative robustness of the LK statistic (T_{LK}) to complex demographic history but we show that $T_{F\text{–}\mathrm{LK}}$ is more powerful in most cases. This new statistic outperforms also a multinomial-Dirichlet-based model [estimation with Markov chain Monte Carlo (MCMC)], when historical branching occurs. Overall, $T_{F\text{–}\mathrm{LK}}$ detects 15–35% more selected SNPs than T_{LK} for low type I errors ($P < 0.001$). Also, simulations show that T_{LK} and $T_{F\text{–}\mathrm{LK}}$ follow a chi-square distribution provided the ancestral allele frequencies are not too extreme, suggesting the possible use of the chi-square distribution for evaluating significance. The empirical distribution of $T_{F\text{–}\mathrm{LK}}$ can be derived using simulations conditioned on the estimated $\mathcal{F}$ matrix. We apply this new test to pig breeds SNP data and pinpoint outliers using $T_{F\text{–}\mathrm{LK}}$, otherwise undetected using the less powerful T_{LK} statistic. This new test represents one solution for compromise between advanced SNP genetic data acquisition and outlier analyses.

THE development of methods aiming at detecting molecular signatures of selection is one of the major concerns of modern population genetics. Broadly, such methods can be classified into four groups: methods focusing on (i) the interspecific comparison of gene substitution patterns, (ii) the frequency spectrum and models of selective sweeps, (iii) linkage disequilibrium (LD) and haplotype structure, and (iv) patterns of genetic differentiation among populations (for a review see Nielsen 2005). Tests based on the comparison of polymorphism and divergence at the species level inform on mostly ancient selective processes. Population-based approaches, however, are designed to pinpoint modern processes of local adaptation and speciation occurring among populations within a species. Such approaches also become crucial in the fields of agronomical and biomedical sciences, for instance, to pinpoint possible interesting (QTL) regions and disease susceptibility genes. Especially, human, livestock, and cultivated plants genetics may benefit from such methods while whole-genome single-nucleotide polymorphisms (SNPs) genotyping technologies are becoming routinely available (*e.g.*, Barreiro *et al.* 2008; Flori *et al.* 2009).

In the population genomic era (Luikart *et al.* 2003), identifying genes under selection or neutral markers influenced by nearby selected genes is a task in itself for quantifying the role of selection in the evolutionary history of species. Conversely, the accurate inference of demographic parameters such as effective population sizes, migration rates, and divergence times between populations relies on the use of neutral marker data sets. One approach of detecting loci under selection (outliers) with population genetic data is based on the genetic differentiation between loci influenced only by neutral processes (genetic drift, mutation, migration) and loci influenced by selection.

Lewontin and Krakauer's (LK) test for the heterogeneity of the inbreeding coefficient (F) across loci was the

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.117275/DC1.

[1] *Corresponding author:* UMR444 Laboratoire de Génétique Cellulaire, INRA Toulouse, BP52627, F-31326 Castanet Tolosan Cedex, France. E-mail: magali.san-cristobal@toulouse.inra.fr

first to be developed with regard to this concept (Lewontin and Krakauer 1973). The LK test was immediately subject to criticisms (Nei and Maruyama 1975; Lewontin and Krakauer 1975; Robertson, 1975a,b; Tsakas and Krimbas 1976; Nei and Chakravarti 1977; Nei *et al.* 1977). Indeed, its assumptions are likely to be violated due to loci with high mutation rate, variation of F due to unequal effective population size (N_e) among demes, and correlation of allele frequencies among demes due to historical branching. The robustness of the LK test to the effects of demography was tested through coalescent simulations by Beaumont and Nichols (1996). They tested the influence of different models of population structure on the joint distribution of F_{ST} (*i.e.*, the inbreeding coefficient F) and heterozygosity (H_e). The F_{ST} distribution under an infinite-island model is inflated for *low* H_e values under both the infinite-allele model (IAM) and the stepwise mutation model (SMM) (Beaumont and Nichols 1996). This tendency becomes, however, more marked when strong differences in effective size N_e and gene flow among demes occur, that is, when allele frequencies are correlated among local demes. This suggests an excess of false significant loci when one assumes an infinite-island model as a null hypothesis, while correlations of gene frequencies substantially occur. However, the F_{ST} distribution shows robustness properties for *high* H_e values (typical from microsatellite markers). Therefore, Beaumont and Nichols (1996) suggested the possibility of detecting outliers by using the distribution of neutral F_{ST} conditionally on H_e under the infinite-island model of symmetric migration, with mutation.

The problem of accounting for correlations of allele frequencies among subpopulations was discussed by Robertson (1975a), who showed how these correlations inflated the variance of the LK test. Different approaches were taken to cope with the problem. It was, for instance, proposed to restrict the analysis to pairwise comparisons (Tsakas and Krimbas 1976; Vitalis *et al.* 2001). However, as pointed out by Beaumont (2005), reducing the number of populations to be compared to many pairwise comparisons raises the problem of nonindependence in multiple testing and may reduce the power to detect outliers. Another way was to assume that subpopulation allele frequencies are correlated through a common migrant gene pool, that is, the ancestral population in a star-like population divergence. In this case, subpopulations evolve with an unequal number of migrants coming from the migrant pool and/or to different amounts of genetic drift. This demographic scenario can be explicitly modeled using the multinomial-Dirichlet likelihood approach (Balding 2003). This multinomial-Dirichlet likelihood (or Beta-binomial for biallelic markers such as SNPs) was implemented by Beaumont and Balding (2004) and subsequently by Foll and Gaggiotti (2008), Gautier *et al.* (2009), Guo *et al.* (2009), and Riebler *et al.* (2010), in a Bayesian hierarchical model in which the F_{ST} is decomposed into two components: a locus-specific (α) effect and a population-specific (β) effect. This Bayesian statistical model together with prior assumptions on α and β was implemented in a Markov chain Monte Carlo (MCMC) algorithm. A substantial improvement made by Foll and Gaggiotti (2008) was to use a reverse-jumping (RJ)-MCMC to simultaneously estimate the posterior distribution of a model with selection (with α and β) and of a model without selection (with β only). More recently, Excoffier *et al.* (2009) addressed the issue of accounting for "heterogeneous affinities between sampled populations"—in other words, accounting for migrant genes that do not necessarily originate from the same pool—by using a hierarchically structured population model. They showed by simulations that the false positive rate is lower under a hierarchically structured population model than under a simple island model, for the IAM and the SMM applicable to microsatellite markers and for a SNP mutation model. Excoffier *et al.*(2009) thus proposed to extend the Beaumont and Nichols (1996) method to a hierarchically structured population model.

Nowadays, a computational challenge is to analyze data sets with increasing numbers of markers and populations, under complex demographic histories, in a reasonable amount of time. This is especially the case in agronomical and biomedical sciences with the increasingly used biallelic SNP markers. A question arises as to whether F_{ST}-based methods would be sufficiently powerful to detect outliers with SNP markers. Indeed, for low H_e values, the inflation of the F_{ST} distribution under the infinite-island model accentuates dramatically when assuming a mutation model typical for SNPs (simulations of Eveno *et al.* 2008). Excoffier *et al.* (2009) corroborated these results and also indicated that the F_{ST} distribution is generally broader under a model of hierarchically structured populations when using SNP markers. In addition, as the authors pinpoint, although the hierarchical island model is more conservative than the island model, an excess of false positives can be obtained "if the underlying genetic structure is more complex . . . , for instance in case of complex demographic histories, involving population splits, range expansion, bottleneck or admixture events" (Excoffier *et al.* 2009, p. 12). The Bayesian hierarchical models developed by Beaumont and Balding (2004) and Foll and Gaggiotti (2008) effectively account for strong effective size and migration rate variation among subpopulations, but they still impose a star-like demographic model in which the current populations share a common migrant pool and are not supposed to have undergone historical branching. More practically, MCMC-based methods might suffer from a computational time requirement when analyzing large marker data sets such as SNP chips data sets. Therefore, the development of simple parametric tests potentially dealing with a summary

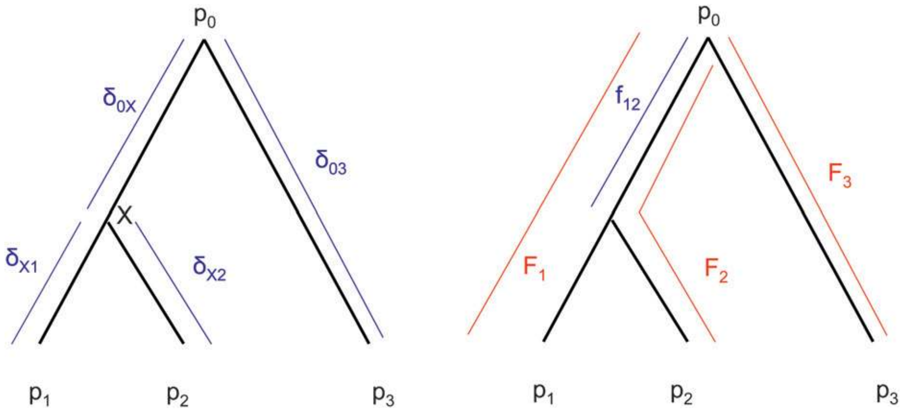


FIGURE 1.—Example of tree-like evolution: construction of the kinship matrix.

of the population tree, including historical branching as well as population size variation, remains an alternative solution to achieve a good compromise between advanced genetic data acquisition and outlier analyses.

In this article, we describe an extension of the original parametric LK test for biallelic markers that deals with complex population trees through a statistic that takes into account the kinship (or coancestry) matrix $\mathcal{F}$ between populations, under pure drift with no migration. The statistics of the classical test (T_{LK}) and its extension ($T_{\mathcal{F}\text{–LK}}$) are expected to follow a chi-square distribution with ($n - 1$) d.f., where n is the number of populations studied. Through forward simulations of neutral SNPs data under increasingly complex demographic histories, we obtained the empirical distribution of both statistics and showed that they follow a chi-square distribution provided the ancestral allele frequencies are not too extreme. These results also emphasize the robustness of these statistics to variation in demographic histories. Forward simulations of the same demographic models but including selection in one population allowed us to evaluate the power of both statistics to detect selection. We show that the extension of the LK test is more powerful at detecting outliers than the classical LK test for complex demographic histories. A comparison with one of the MCMC methods for multinomial-Dirichlet models (FOLL and GAGGIOTTI 2008) also revealed substantial additional power. We apply this new statistical test to a data set of SNP markers in known genes of the pig genome, taking advantage of the availability of microsatellite markers for the estimation of the kinship matrix. This new parametric test can help to screen large marker data sets and large numbers of populations for outliers in a reasonable amount of time, although we recommend to simulate the empirical distribution of the $T_{\mathcal{F}\text{–LK}}$ statistics conditionally on the estimated kinship matrix.

POPULATION MODEL AND NOTATIONS

We consider a set of n populations derived from a common ancestor and the frequencies $(p_1, p_2, \ldots, p_n)$ of one allele at a neutral biallelic locus. We assume their phylogeny is described by a tree (Figure 1), in which each branch is characterized by some amount of drift.

The kinship matrix: Due to drift and coancestries, frequencies p_i's are correlated, so that

$$\mathrm{Cov}(p_i, p_j) = f_{ij} p_0 (1 - p_0) \tag{1}$$

$$\mathrm{Var}(p_i) = f_{ii} p_0 (1 - p_0), \tag{2}$$

where p_0 is the frequency of the allele in the ancestor population, f_{ii} is the mean expected inbreeding coefficient of the ith population, and f_{ij} the kinship coefficient between populations i and j equal to the inbreeding coefficient of the most recent ancestor population common to i and j.

In Figure 1, for example, the calculations proceed as follows. Let δ_{UV} be the fixation index corresponding to the branch from U (an internal node or the root of the tree) to V (an internal node or a leaf of the tree, *i.e.*, one of the n populations). If the branch UV corresponds to t generations in a population of effective size N, $\delta_{UV} \simeq 1 - \exp(-t/2N)$ provided mutations are ignored. The tree of Figure 1 includes the root (O), the internal node (X), and the three populations 1, 2, and 3. Setting $f_{00} = 0$, we have

$$f_{11} = F_1 = 1 - (1 - \delta_{X1})(1 - \delta_{0X}) \tag{3}$$

$$f_{22} = F_2 = 1 - (1 - \delta_{X2})(1 - \delta_{0X}) \tag{4}$$

$$f_{33} = F_3 = \delta_{03} \tag{5}$$

$$f_{12} = \delta_{0X} \tag{6}$$

$$f_{13} = 0 \tag{7}$$

$$f_{23} = 0. \tag{8}$$

In the following, $\mathcal{F}$ stands for the matrix of the f_{ij}. For simplicity, diagonal elements f_{ii} are simply denoted as F_i. Under pure drift (without mutation) it can be demonstrated that $\mathcal{F}$ is invertible and positive definite.

Estimation: Let us consider L biallelic loci indexed by ℓ, whose first allele frequency in population i is $p_{i,\ell}$. A sample of genotyped individuals in each population provides an empirical estimate $\hat{p}_{i,\ell}$ of this allele frequency by simple counting.

We propose to make use of the neighbor-joining (NJ) tree (Saitou and Nei 1987) built from the Reynolds' genetic distances between pairs of populations (Reynolds *et al.* 1983), adding an outgroup so that the tree linking the n populations can be rooted. Then branch lengths of the NJ tree are estimates of the δ's and provide estimates of the elements of the $\mathcal{F}$ matrix. Since we assume in the following that frequency distributions are approximately Gaussian, an alternative approach could be to estimate δ-values by a likelihood approach as suggested by Weir and Hill (2002). However, these authors considered only the case where $\mathcal{F}$ is diagonal. Accounting for a general tree structure would make their approach more complicated and probably not needed since we did not find any strong difference between results obtained using true or estimated values.

Loci used to estimate $\mathcal{F}$ must be neutral. When genome-wide genotyping is available, one can consider that only a small fraction of genomic regions and hence of genotyped markers is or has been a target of selection, so that averaging over all loci will provide a good estimate of $\mathcal{F}$. We used this approach in our simulation-based study, where $\mathcal{F}$ was estimated from the simulated SNPs to be tested. Another possibility is to make use of a subset of markers (supposed neutral) to estimate $\mathcal{F}$ and then use it for testing departures from neutrality of another subset of markers. We used this approach to test for signature of selection in a real data set from pig populations. We took advantage of the availability of microsatellite markers for estimating $\mathcal{F}$, to test SNP markers in candidate genes.

TESTS OF SELECTION: LEWONTIN AND KRAKAUER AND EXTENSIONS

Distribution of the LK test: Consider L biallelic loci genotyped for a large set of individuals structured in n populations. Lewontin and Krakauer (1973) focused on the distribution of the F_{ST} statistic per locus and proposed a test statistic denoted here by T_{LK}. To simplify notations, the subscript ℓ per locus is omitted in the following. Note that the allele frequencies and the corresponding statistics depend on the current locus, while the kinship matrix $\mathcal{F}$ does not. Let $\mathbf{p} = (p_1, \ldots, p_j, \ldots, p_n)'$ be the n-vector of allelic frequencies of the first allele (say) in the n populations. The quantity F_{ST} is defined as

$$F_{\mathrm{ST}} = \frac{s_p^2}{\bar{p}(1-\bar{p})} = \frac{\left((1/(n-1))\sum_{i=1}^{n}(p_i-\bar{p})^2\right.}{\bar{p}(1-\bar{p})}, \qquad (9)$$

where $\bar{p}$ and s_P^2 are the sampling estimates of the mean and variance, respectively, of the vector p. The test statistic is equal to

$$T_{\mathrm{LK}} = \frac{n-1}{\bar{F}_{\mathrm{ST}}} F_{\mathrm{ST}}, \qquad (10)$$

where $\bar{F}_{\mathrm{ST}}$ is the average of F_{ST} in (9) over the L loci. Under the reference conditions considered by Lewontin and Krakauer (equal branch lengths, $F_i = f_{ii} = F$, and no correlations, $f_{ij} = 0$ for $i \neq j$), this test was shown to follow approximately a χ^2-distribution with $n - 1$ d.f.

In the following, we propose a new calculation of the first two moments of the F_{ST} statistic, in the case of a tree-like history of the n populations. Under genetic drift, the first two moments of $\mathbf{p}$ are

$$\mathbb{E}(\boldsymbol{p}) = p_0 \mathbf{1}_n \qquad (11)$$

$$\mathbb{V}(\boldsymbol{p}) = \mathcal{F} p_0(1-p_0), \qquad (12)$$

where p_0 is the founder allele frequency, $\mathbf{1}_n$ is the n-vector of 1's, and $\mathcal{F}$ is the kinship (or coancestry) $(n \times n)$ matrix linking the n populations.

It can be shown (see appendix a) that

$$\mathbb{E}(F_{\mathrm{ST}}) \simeq \bar{F} - \bar{f}, \qquad (13)$$

provided the number of populations is large enough, that

$$\mathbb{E}(T_{\mathrm{LK}}) \simeq (n-1), \qquad (14)$$

and that, approximating frequency distributions by the normal if F values are small,

$$\mathbb{V}(T_{\mathrm{LK}}) \simeq 2\frac{\sum_i \sum_j f_{ij}^2 - (2/n)\sum_i\left(\sum_j f_{ij}\right)^2 + (1/n^2)\left(\sum_i \sum_j f_{ij}\right)^2}{(\bar{F}-\bar{f})^2}, \qquad (15)$$

with

$$\bar{F} = \frac{1}{n}\sum_i F_i = \frac{1}{n}\sum_i f_{ii} \qquad (16)$$

and

$$\bar{f} = \frac{1}{n(n-1)}\sum_i \sum_{j \neq i} f_{ij}. \qquad (17)$$

With a star-like evolution (the nondiagonal elements in $\mathcal{F} = 0, \bar{f} = 0$) and with equal branch lengths ($\bar{F}_i = F$ for all i as in Lewontin and Krakauer 1973), the p_i's are assumed to be independent, identically distributed, and normal, so that T_{LK} follows the distribution of a chi square with $(n-1)$ d.f. This is the basic version of the test. In other cases, the test can be adapted, either recalculating its moments or defining another statistic to test the fit of data with the null hypothesis.

As shown in appendix a, the general expression (15) takes simpler forms in special cases of departure from the basic situation:

The phylogenetic tree of populations is structured but branch lengths are equal ($F_i = f_{ii} = F$ for all i). Then ROBERTSON (1975b) showed that

$$\mathbb{V}(T_{\mathrm{LK}}) \simeq 2(n-1)(1 + nV_{r'}), \qquad (18)$$

where $V_{r'}$ stands for the variance of correlation coefficients between gene frequencies (see APPENDIX A for the correspondence with the present notations). This result suggests that such correlations may imply a strong increase of the expected variance of the test.

Populations are independent (*i.e.*, the tree representing the phylogeny of populations has the structure of a star) but F values are heterogeneous. In that case one has

$$\mathbb{V}(T_{\mathrm{LK}}) \simeq 2(n-1)\left(1 + \left(1 - \frac{2}{n}\right)\frac{\mathbb{V}(F)}{\bar{F}^2}\right), \qquad (19)$$

where $\mathbb{V}(F)$ is the variance of F_i values.

Provided the departure from normality is not too strong, we propose an extension of the LK test to take account of any structuration on the moments of allele frequency distributions.

An extension of the LK test when the populations are structured—use of the $\mathcal{F}$ matrix: The previous calculation allows one to obtain the correct variance of the test. However, the chi-square distribution of the test is anyway only approximate, even assuming normality, because (i) the F_i's are heterogeneous, which implies that T_{LK} is a sum of squared random variables with different variances, and (ii) the denominator in (9) depends on the allele frequencies.

Assuming normality the joint distribution of allele frequencies is fully characterized by the initial frequencies p_0 and by the $\mathcal{F}$ matrix.

Let

$$\hat{p}_0 = \frac{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{p}}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n} \qquad (20)$$

be the unbiased linear estimate of p_0 with minimum variance, with $\mathbf{1}_n$ denoting the n-vector made of 1's. It may be noted that this estimate of p_0 is *not* the maximum-likelihood estimate, even under the normal assumption. When the n populations diverge from the founder in a star-like manner, but with different coancestry coefficients, then $\hat{p}_0 = \left(\sum_{j=1}^n p_j/F_j\right)/\left(\sum 1/F_j\right)$. Further, when the populations have the same size, as in the Lewontin and Krakauer test, then this estimator is the sample mean $\left(\hat{p}_0 = \bar{p}\right)$.

Let us note $\hat{p}_0 = \mathbf{w}'\mathbf{p}$, with w the n-vector

$$\mathbf{w} = \frac{\mathcal{F}^{-1}\mathbf{1}_n}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}. \qquad (21)$$

Then the first two moments of the estimator $\hat{p}_0$ of p_0 can be calculated:

$$\begin{aligned}\mathbb{E}(\hat{p}_0) &= \mathbf{w}'\mathbb{E}(\mathbf{p}) = p_0\\ \mathbb{V}(\hat{p}_0) &= \mathbf{w}'\mathbb{V}(\mathbf{p})\mathbf{w}\\ &= \frac{p_0(1-p_0)}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}.\end{aligned}$$

It follows that

$$\mathbb{E}(\hat{p}_0(1-\hat{p}_0)) = p_0(1-p_0)\left(1 - \frac{1}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}\right). \qquad (22)$$

If the ancestral allele frequencies p_0 were known, then the most interesting quadratic form in $\mathbf{p}$ would be

$$T_{F-\mathrm{LK}}(p_0) = (\mathbf{p} - p_0\mathbf{1}_n)'\mathbb{V}(\mathbf{p})^{-1}(\mathbf{p} - p_0\mathbf{1}_n), \qquad (23)$$

which follows a chi-square distribution with n d.f. However, since p_0 is unknown, it is replaced by its estimator $\hat{p}_0$, suggesting to define the test as

$$T_{F-\mathrm{LK}} = (\mathbf{p} - \hat{p}_0\mathbf{1}_n)'\mathbb{V}(\mathbf{p})^{-1}(\mathbf{p} - \hat{p}_0\mathbf{1}_n) = \frac{Q}{\hat{p}_0(1-\hat{p}_0)}. \qquad (24)$$

In practice, the above expression of $T_{F\text{-}\mathrm{LK}}$ is multiplied by the bias correction term $(1 - 1/(\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n))$ (see Equation 22), which is omitted in the following, for the sake of simplicity. When $\mathcal{F} = F\mathbf{I}_n$, the only difference between T_{LK} and $T_{F\text{-}\mathrm{LK}}$, apart from the bias correction term, resides in the estimation of F, either with $\bar{F}_{\mathrm{ST}}$ or with the estimation method proposed in this article (*Estimation* section).

The quadratic form

$$Q = (\mathbf{p} - \hat{p}_0\mathbf{1}_n)'\mathcal{F}^{-1}(\mathbf{p} - \hat{p}_0\mathbf{1}_n) \qquad (25)$$

can be written as $\mathbf{p}'\mathbf{M}\,\mathbf{p}$, where

$$\mathbf{M} = \mathcal{F}^{-1} - \frac{\mathcal{F}^{-1}\mathbf{1}_n\mathbf{1}_n'\mathcal{F}^{-1}}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}. \qquad (26)$$

Its first moment can be calculated as

$$\begin{aligned}\mathbb{E}(Q) &= \mathbb{E}(\mathbf{p})'\mathbf{M}\mathbb{E}(\mathbf{p}) + \mathrm{tr}(\mathbf{M}\mathbb{V}(\mathbf{p}))\\ &= p_0^2 1_n'\mathbf{M}\mathbf{1}_n + p_0(1-p_0)\mathrm{tr}(\mathbf{M}\mathcal{F})\\ &= (n-1)p_0(1-p_0).\end{aligned}$$

The second moment of Q is

$$\begin{aligned}\mathbb{V}(Q) &= 4\mathbb{E}(\mathbf{p})'\mathbf{M}\mathbb{V}(\mathbf{p})\mathbf{M}\mathbb{E}(\mathbf{p}) + 2\mathrm{tr}\big(\mathbf{M}\mathbb{V}(\mathbf{p})\mathbf{M}\mathbb{V}(\mathbf{p})\big)\\ &= 2(n-1)p_0^2(1-p_0)^2.\end{aligned}$$

Then $T_{F\text{-}\mathrm{LK}}$ has approximate expectation

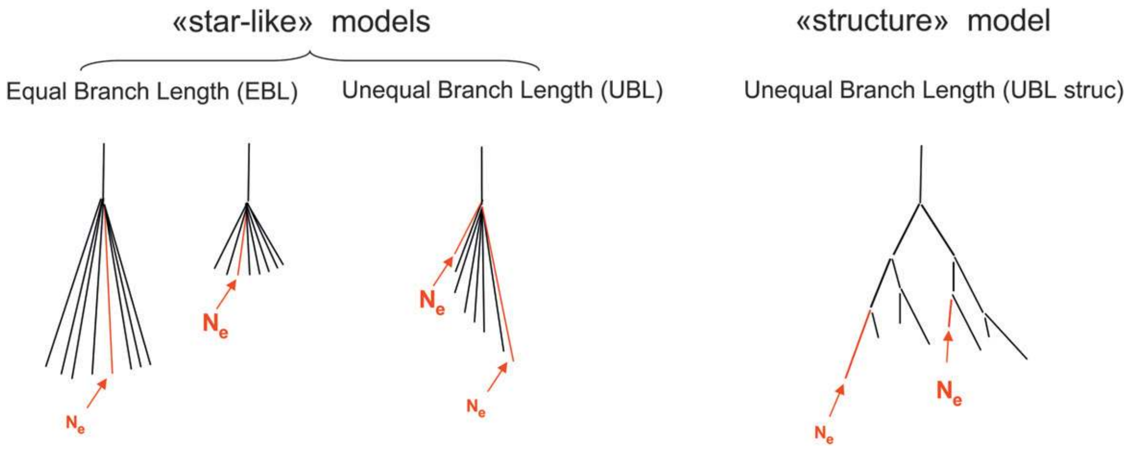

«star-like» models
«structure» model
Equal Branch Length (EBL)
Unequal Branch Length (UBL)
Unequal Branch Length (UBL struc)
N$_e$
N$_e$
N$_e$
N$_e$
N$_e$
N$_e$

FIGURE 2.—Models of population divergence simulated in this study. This schematic illustrates three sorts of demographic models simulated in this study: EBL, UBL, and UBL struc. Populations highlighted in red are those in which directional selection occurs. For the UBL models, we simulated selection in a large (big N_e) and a small (little N_e) population, in separate simulations. For the EBL model we simulated two scenarios, one with large and one with small populations, in which one population is selected.

$$\mathbb{E}(T_{F-\mathrm{LK}}) \approx \frac{\mathbb{E}(Q)}{\mathbb{E}[\hat{p}_0(1-\hat{p}_0)]} = n-1 \qquad (27)$$

and approximate variance

$$\mathbb{V}(T_{F-\mathrm{LK}}) \approx \frac{\mathbb{V}(Q)}{\mathbb{E}^2[\hat{p}_0(1-\hat{p}_0)]} = 2(n-1), \qquad (28)$$

so that $T_{F\text{–}\mathrm{LK}}$ follows approximately a χ^2_{n-1}-distribution under genetic drift. The case of a multiallelic locus is derived in APPENDIX B, but is not investigated further in this article.

SIMULATIONS

Simulation settings: We simulated haplotype samples of partially linked loci, under neutrality (H_0) and directional selection on one locus in one population (H_1). The choice of simulating partially linked loci was technically relevant because most SNPs data sets nowadays come from dense whole-genome scans. In all simulated scenarios of population divergence, the populations originate from an equilibrium ancestral population of constant size.

Neutral haplotype samples from this ancestral population were obtained by coalescent simulations using the MS software (HUDSON 2002). The generated haplotypes consisted of 1000 SNPs (or biallelic segregating sites) randomly distributed along a 100-Mb chromosome, resulting in a 100-kb distance between two SNPs, on average. Assuming a recombination rate of 1 cM/Mb, the recombination rate between two SNPs was fixed at 0.1 cM.

To simulate the evolution of the populations from the ancestral one in the same way for both neutrality and selection, we used forward simulations of the Wright–Fisher diploid model, further assuming stepwise changes in population size, population dichotomy, no mutations, and a uniform recombination rate. Different sorts of demographic models were simulated to explore the influence of demographic history on the statistical properties of both the classical LK statistic and the extension we propose. The first demographic model is a model of star-like population divergence with equal branch lengths (EBL) among populations, in which all populations evolve spontaneously from a common ancestor, independently from each other with the same inbreeding coefficient *F.* The second model is also a star-like divergence scenario but with unequal branch lengths (UBL) among populations. The third model is a model of populations structured by common ancestries with variation of branch length (UBL struc) (see Figure 2 and Table 1 for population schemes and the demographic parameters used).

Selection was modeled as follows: (i) selection occurs on a single locus (SNP) of the haplotype, (ii) selection occurs on the less frequent allele of the SNP (“0” and “1” are the ancestral and derived states, respectively), (iii) the allelic fitness k is linked to the selection coefficient s by $k = 1 + s$, leading to the genotypic fitness scheme

00	01	11
1	$1 + s$	$(1 + s)^2$

Note that in this case it is the derived allele that is under selection. Hence, the probability of drawing a given parental genotype to generate the next generation depends on the genotype frequency, which changes at each generation according to this selection scheme. In UBL models, we chose to simulate separately selection on “large” and “small” populations to better account for the heterogeneity of F among populations

TABLE 1

Demographic parameters of the different models simulated in this study

	Model										
	EBL ($F = 0.05$)[a] (8 pop)			UBL[a] (8 pop)			UBL struc (8 pop)				
	Groups										
	Outgroup	Ancestor	Population	Outgroup	Ancestor	Population	Outgroup	Ancestor	Ancestor	Ancestor	Population
	Population size										
	$N_e = 900$	$N_e = 900$	$N_e = 900^{b,c}$	$N_e = 500$	$N_e = 500$	$N_e = 100^{b,d}$	$N_e = 500$				$N_e = 100^{b,d}$
			$N_e = 900$			$N_e = 200$					$N_e = 200$
			$N_e = 900$			$N_e = 300$				$N_e = 100$	
			$N_e = 900$			$N_e = 350$				$N_e = 500$	
			$N_e = 900$			$N_e = 400$			$N_e = 500$		$N_e = 300$
			$N_e = 900$			$N_e = 600$					$N_e = 100$
			$N_e = 900$			$N_e = 800$		$N_e = 500$			
			$N_e = 900$			$N_e = 1000^{b,c}$					$N_e = 400$
									$N_e = 500$		$N_e = 600$
	EBL ($F = 0.4$)[a]									$N_e = 100$	
										$N_e = 500$	
	$N_e = 100$	$N_e = 100$	$N_e = 100^{b,d}$								$N_e = 800$
			$N_e = 100$								$N_e = 1000^{b,c}$
			$N_e = 100$								
			$N_e = 100$								
			$N_e = 100$								
			$N_e = 100$								
			$N_e = 100$								
			$N_e = 100$								
Generation nos.	$t = 200$	$t = 100$	$t = 100$	$t = 200$	$t = 100$	$t = 100$	$t = 200$	$t = 100$	$t = 25$	$t = 25$	$t = 50$

[a] In "star-like" models, the inbreeding coefficient $F = 1 - (1 - t/2N)^t = F_{ST}$. 8 pop, eight populations.

[b] Population in which directional selection occurs in the simulations (only one population is under selection in each simulation type).

[c] Large population.

[d] Small population.

when selection acts (Table 1). Selection was simulated for two intensities, $s = 0.05$ and $s = 0.20$.

We performed 10,000 simulations in each demographic scenario to cover small type I error. For each simulation, a matrix of unbiased Reynolds' genetics distances was computed from frequency data of the 1000 partially linked SNPs simulated. The $\mathcal{F}$ matrix was then estimated from branch lengths of a neighbor-joining tree (see *Estimation* section). The ancestral allele frequency p_0 was estimated using $\bar{p}$ for T_{LK} and using $\hat{p}_0$ (Equation 20) for $T_{F\text{–LK}}$. Then the T_{LK} and $T_{F\text{–LK}}$ statistics were calculated for each SNP, excluding the cases of complete fixation of any of the two alleles in the whole population set. To construct the H_1 distribution of both statistics, we recorded the T_{LK} and $T_{F\text{–LK}}$ values for the SNP under selection for each simulation. To construct the H_0 distribution, we drew at random one SNP position and recorded its associated T_{LK} and $T_{F\text{–LK}}$ values for each simulation under neutrality.

To allow an unbiased comparison of the empirical distributions to the theoretical distribution, we considered the ideal situation in which the true $\mathcal{F}$ matrix and ancestral allele frequency p_0 are known. In each simulation, we recorded the value of the ancestral allele frequency p_0 for each SNP, and we calculated $T_{\mathrm{LK}}(p_0)$ and $T_{F\text{–LK}}(p_0)$ accordingly (refer to Equations 9 and 10, where $\bar{p}$ is replaced by p_0, and Equation 23). The calculation of $T_{F\text{–LK}}(p_0)$ included the true $\mathcal{F}$ matrix.

The different empirical H_0 distributions of T_{LK} and $T_{F\text{–LK}}$ were compared to their theoretical expectations (*i.e.*, chi-square distribution with n or $n - 1$ d.f., depending on whether parameters had to be estimated or not). The power of each statistic to detect selected SNPs was evaluated as follows: first, we determined the 0.9, 0.95, 0.98, 0.99, and 0.999 quantiles of the empirical null distribution of each test from the simulations under neutrality. Then, the power of the tests was determined as the proportion of simulations for which the statistic was greater than a given quantile of the null. This allows power to be recorded as a function of the empirical type I error.

To compare the LK-based tests to the method of Foll and Gaggiotti (2008), we used their Bayes factor for selection of the selected SNP as a test statistic. As an implementation of the Foll and Gaggiotti (2008) method, we used the BAYESCAN software run with the default parameters. As this method requires a rather long computation time, comparisons were performed on 1000 simulations only, under UBL and UBL struc models for two selection intensities (0.05, 0.20). The power of this method and of the LK-based tests was evaluated as explained above.

Simulation results: *The empirical distributions of T_{LK} and $T_{F\text{–}LK}$ under neutrality, and the chi-square distribution:* The empirical distributions of $T_{\mathrm{LK}}(p_0)$ and $T_{F\text{–LK}}(p_0)$ have similar shapes in each demographic model (EBL, UBL, and UBL struc), with the same number of populations (*i.e.*, eight populations were simulated). We illustrate this under the more complex UBL struc model, with *Q–Q* plots that compare the empirical distribution of $T_{\mathrm{LK}}(p_0)$ and $T_{F\text{–LK}}(p_0)$ with the theoretical chi-square distribution (Figure 3). For each statistic, however, the right tail of the distribution varies slightly depending on the demographic model (Figure 3 for UBL struc and supporting information, Figure S1 and Figure S2 for EBL and UBL models). Overall, the empirical distributions of $T_{\mathrm{LK}}(p_0)$ and $T_{F\text{–LK}}(p_0)$ under neutrality appear relatively robust to increasingly complex demographies, whatever the range of ancestral allele frequencies (Figure 3, Figure S1, and Figure S2). In addition, we observed that the shape of the empirical distributions of $T_{\mathrm{LK}}(p_0)$ and $T_{F\text{–LK}}(p_0)$ appears to depend on p_0. When all simulated ancestral frequencies are included ($0 < p_0 < 1$), they do not fit the right tail of the chi-square distribution (Figure 3). Extreme p_0 values represented a high proportion of the simulations (Figure 3a). When accounting for less extreme p_0 values (*i.e.*, $0.2 < p_0 < 0.8$), the empirical distribution fit the chi-square distribution (Figure 3, b and c).

In the real situation of parameter estimation (see *Estimation* in POPULATION MODEL AND NOTATIONS for the estimation of p_0 and the $\mathcal{F}$ matrix), both estimators of p_0 ($\bar{p}$ and $\hat{p}_0$ in Equation 20) approximate well the true p_0 values (Figure S3). Moreover, the empirical distribution of $T_{F\text{–LK}}$ values based on various $\mathcal{F}$-matrix estimates is highly similar to the one calculated with the true $\mathcal{F}$ matrix (not shown). These results indicate that for both statistics the departure from the theoretical chi-square distribution under neutrality is mainly due to extreme p_0 values rather than problems related to parameter estimations.

Power comparison of the T_{LK} and $T_{F\text{–}LK}$ statistics: Power was calculated using the empirical distributions of the statistics, on the basis of simulations under neutrality and selection (see *Simulation settings* section). Some power properties common to both T_{LK} and $T_{F\text{–LK}}$ arise from this simulation study. First, the population size of the selected population has a major impact on the power to detect selected loci. For a given selection coefficient and whatever the type I error, we found that the power to detect selection is higher in a large population than in a small population (Figure 4), for both T_{LK} and $T_{F\text{–LK}}$. This was expected because the strength of a selection event is mainly determined by the product $N_e s$. The explanation is, however, more complex, since the population sizes also intervene in the weights $T_{F\text{–LK}}$ puts on each population. Second, the selection coefficient has a differential impact on the power, depending on the underlying demographic model. A larger selection coefficient does not result in a higher level of power in EBL and UBL models. However, a larger selection coefficient has a positive impact for detecting selected SNPs in UBL struc models. This can be explained by the fact that complete fixation was reached in some models but not in all of them.

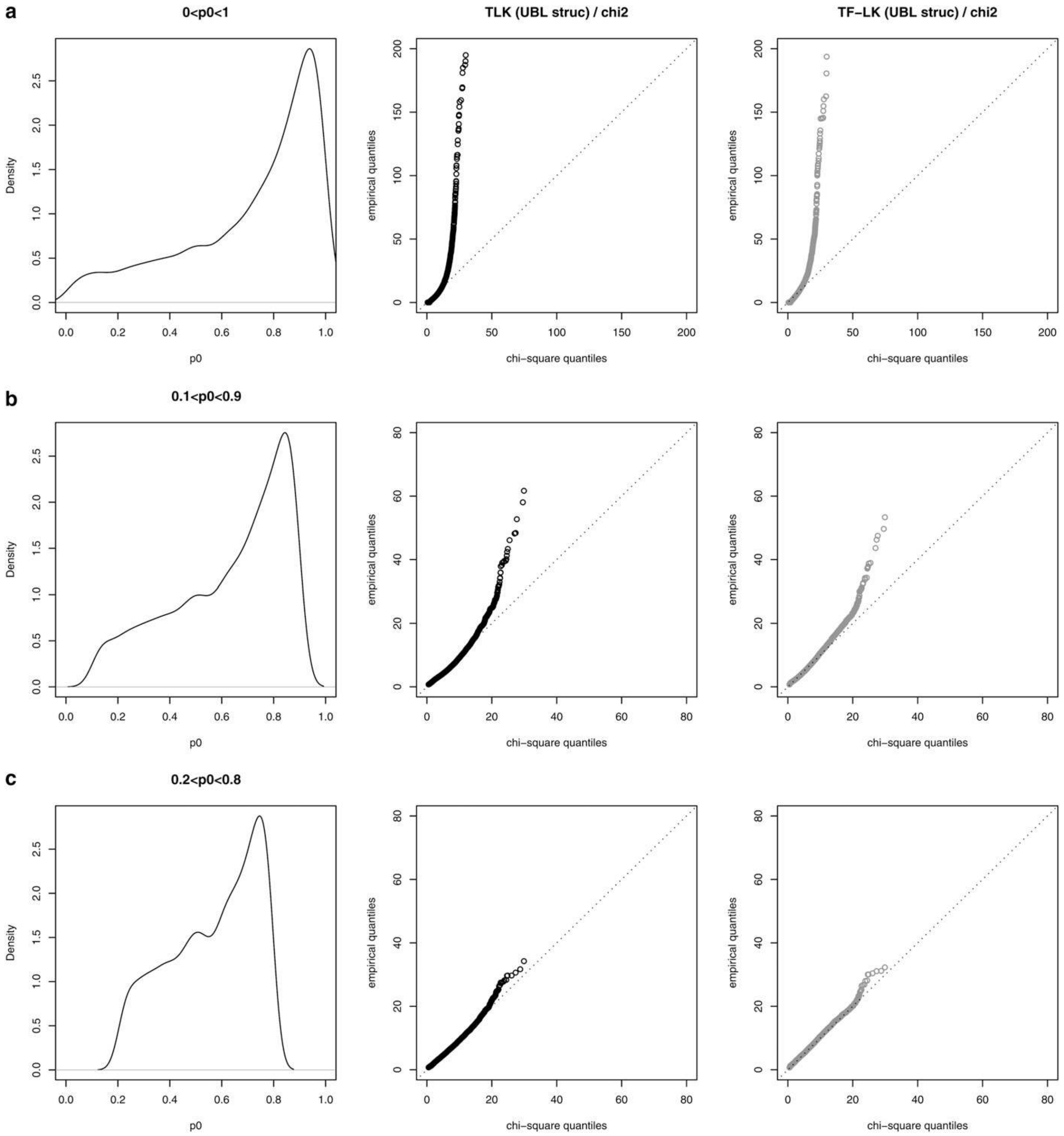


FIGURE 3.—Fit of T_{LK} and $T_{\mathcal{F}\text{–LK}}$ empirical distributions to a χ^2-distribution under the UBL struc scenario with eight populations, and dependency on p_0. (a–c) Left column, distribution of ancestral allele frequencies; center (resp. right) column, Q–Q plots of the empirical distribution of T_{LK} (resp. $T_{\mathcal{F}\text{–LK}}$) under neutrality ($H_0$) against the $\chi^2(8)$ distribution. For unbiased comparison of the empirical and theoretical distributions, we illustrate the ideal case in which p_0 and $\mathcal{F}$ are known.

Substantial differences in power occur between T_{LK} and $T_{\mathcal{F}\text{–LK}}$. We first consider the case in which selection acts on a large population relative to other populations. In UBL and UBL struc models, the detection power of $T_{\mathcal{F}\text{–LK}}$ is >20% greater than that of T_{LK} (Figure 4). In an EBL model, $T_{\mathcal{F}\text{–LK}}$ and T_{LK} have similar detection power, from 60 to 85% for $0.001 < \alpha < 0.1$. If selection acts on a small population relative to other populations, however, T_{LK} is more powerful than $T_{\mathcal{F}\text{–LK}}$ but it should be noted that the absolute power of both statistics is small in that case, especially at low type I errors. Restricting the window of possible p_0 values, for instance to $0.2 < p_0 < 0.8$, has a general negative effect on the power of the T_{LK} statistic, whatever the size of the population under

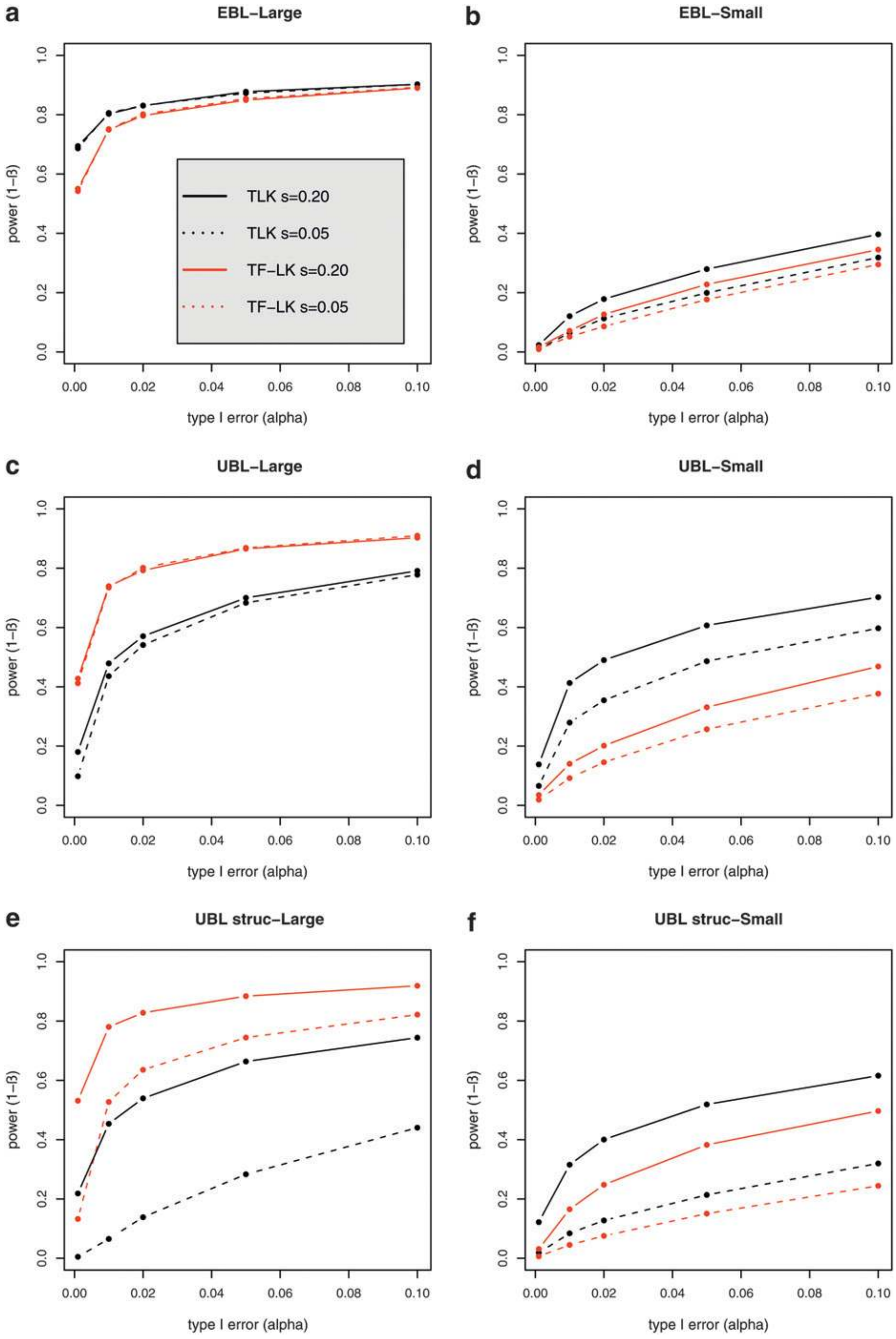


FIGURE 4.—(a–f) Power of T_{LK} and $T_{\mathcal{F}\text{–LK}}$ to detect selection in a large (a, c, and e) or small (b, d, and f) population, for different type I error values. Results are shown for different demographic models with eight populations and two selection coefficients (s = 0.05 and 0.20). The $\mathcal{F}$ matrix and p_0 are set to their estimated values.

selection (not shown). However, in complex UBL models when selection acts on a large population, the power of $T_{\mathcal{F}\text{–LK}}$ seems to benefit from intermediate ancestral frequencies ($0.2 < p_0 < 0.8$) for low type I error ($\alpha <$ 0.001). We also investigated the impact of the population sampling on power properties. For a given population tree, the power to detect selected SNPs with $T_{\mathcal{F}\text{–LK}}$ is increased by sampling more populations (Figure 5). This is not the case with T_{LK} for which the signal of selection seems masked by an increasing number of populations sampled.

We investigated the effect of estimating the $\mathcal{F}$ matrix on power. Selection may introduce a bias in the estimation of the $\mathcal{F}$ matrix, resulting in a loss of power for the tests based on $T_{\mathcal{F}\text{–LK}}$. Indeed, in EBL, UBL, and UBL struc models, the detection power obtained when estimating $\mathcal{F}$ (Figure 4) was reduced compared to that obtained when $\mathcal{F}$ is known (Figure S4), especially for small type I errors, *i.e.*, $0.001 < \alpha < 0.01$. In addition, for tests based on the $T_{\mathcal{F}\text{–LK}}$ statistic, the phylogenetic reconstruction may lead to the emergence of small internal branches and hence to small extradiagonal (f) values in the estimated $\mathcal{F}$ matrix. In the UBL models simulated, cutting small branch lengths had a positive effect on the power of $T_{\mathcal{F}\text{–LK}}$ (Figure 6a, cutoff values = 0.005). Indeed, the branch-cutting procedure trans-

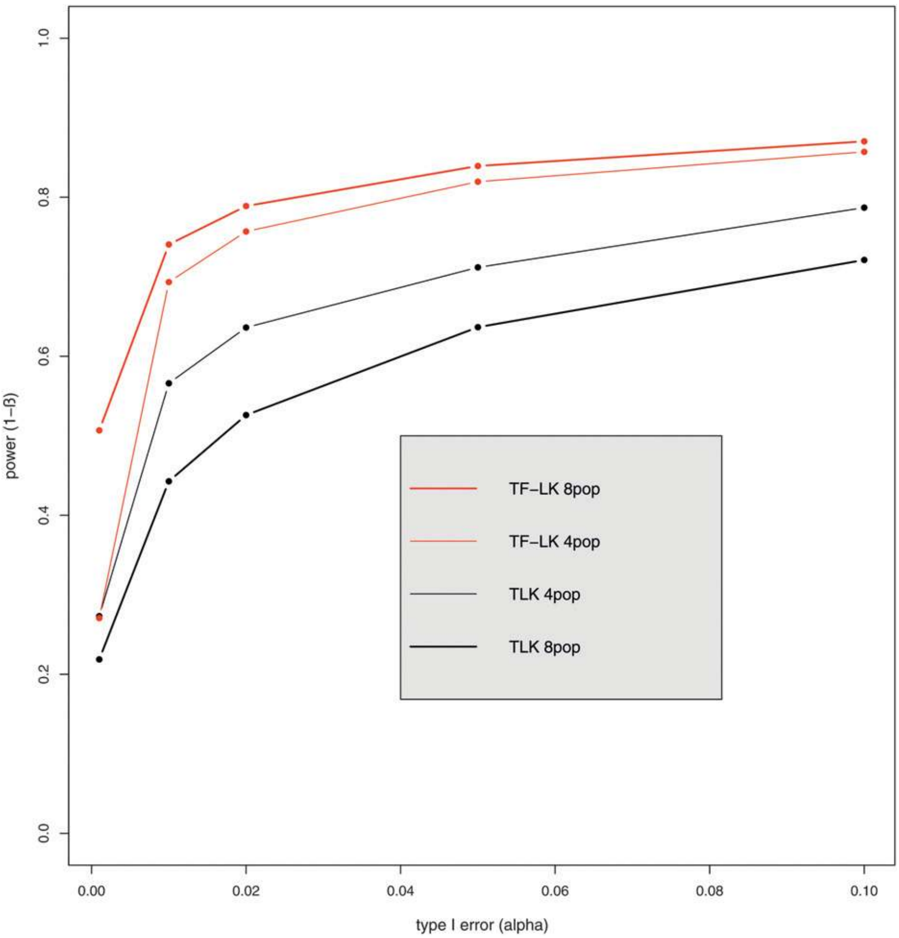


Figure 5.—Influence of population sampling on the power of $T_{\mathcal{F}\text{-LK}}$ when eight populations are simulated under a UBL struc model and the power is calculated on the basis of samples of four or eight populations. Each population sampling contains the selected population. The $\mathcal{F}$ matrices calculated on the basis of both kinds of population sampling do not have the same dimension but reflect similar amounts of genetic drift.

formed some trees inferred as (falsely) "structured" into "star-like" trees closer to the population trees simulated. In UBL struc models, however, cutting small branch lengths had a slightly negative effect on the power of $T_{\mathcal{F}\text{-LK}}$ (Figure 6b, cutoff value = 0.001). In some simulations, indeed, small branch lengths were neglected whereas they truly described the population tree and hence led to a decrease of power.

Finally, we compared the T_{LK} and $T_{\mathcal{F}\text{-LK}}$ tests with the MCMC method of Foll and Gaggiotti (2008) under UBL and UBL struc scenarios. We found that under a UBL scenario, the method of Foll and Gaggiotti (2008) had more detection power than T_{LK}, but not as much as $T_{\mathcal{F}\text{-LK}}$ whether one assumes the number of simulations was not enough for low type I errors (<0.001) (Figure 7, left). Under a UBL struc scenario,

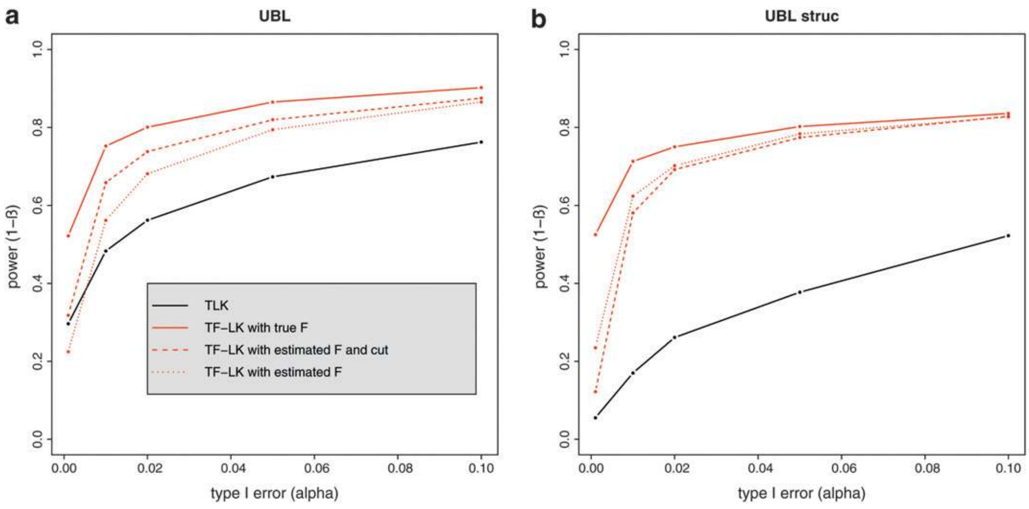


Figure 6.—Influence of a branch-cutting procedure on the power of $T_{\mathcal{F}\text{-LK}}$. This illustrates the effect on the power of $T_{\mathcal{F}\text{-LK}}$ of estimating the $\mathcal{F}$ matrix and of cutting small branch lengths in the phylogenetic tree. Branch lengths are cut as they correspond to f (extra diagonal) values <0.005 and <0.001, in (a) UBL and (b) UBL struc models, respectively.

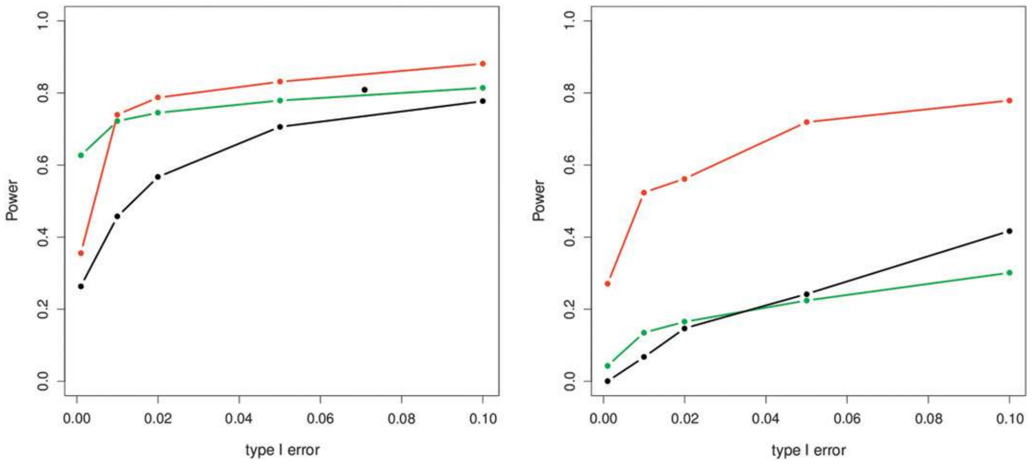


FIGURE 7.—Detection power obtained with T_{LK} (black), $T_{\mathcal{F}\text{-LK}}$ (red), and the method of FOLL and GAGGIOTTI (2008) (green) under a UBL (left) and a UBL struc scenario (right), for a selection coefficient of 0.05.

however, $T_{\mathcal{F}\text{-LK}}$ clearly outperformed the MCMC method for a wide range of type I errors (Figure 7, right). Indeed, $T_{\mathcal{F}\text{-LK}}$ detected 20–50% more selected SNPs than the MCMC method for type I errors ranging from 0.001 to 0.05. Similar results are obtained for $s = 0.20$ under both demographic scenarios (Figure S5). This difference in power under UBL struc scenarios may stem from the fact that the method of FOLL and GAGGIOTTI (2008) does not account for the hierarchical structure of populations, while $T_{\mathcal{F}\text{-LK}}$ does.

APPLICATION TO PIG SNPS DATA

One SNP data set was tested as an illustrative example for signature of selection: 34 SNPs located in candidate genes (BLOTT *et al.* 2003; A. DAY, G. EVANS, and S. BLOTT, unpublished data). The associated commercially important phenotypes concern reproductive performance, growth and fatness, meat quality, and disease resistance. Samples of four major European pig breeds were genotyped: the Landrace (LR) (LR01), the Large White (LW) (LW05), the Piétrain (PI) (PI03), and the Duroc (DU) (DU02). To estimate the $\mathcal{F}$ matrix for calculating the $T_{\mathcal{F}\text{-LK}}$ statistic, we made use of 50 genome-wide distributed microsatellite markers previously studied on the same samples in a previous project (PigBioDiv, see http://www.projects.roslin.ac.uk/pigbiodiv/ and SANCRISTOBAL *et al.* 2006). We used an Asian breed, the Meishan (MS01), as outgroup. We first explored the fit of the empirical distributions of T_{LK} and $T_{\mathcal{F}\text{-LK}}$ to the chi-square distribution. The empirical distributions were generated by simulating population history conditional on the previously estimated $\mathcal{F}$ matrix. To do so, we used forward simulations with parameterizations of N_e and split times that led to the estimated $\mathcal{F}$ matrix. Then, we simulated selection on one SNP in one population under the same conditions, to assess the power to detect selection in a real case. The empirical H_0 distribution of T_{LK} and $T_{\mathcal{F}\text{-LK}}$ in this case has a slightly shorter right tail than the chi-square distribution, (Figure 8). Moreover, $T_{\mathcal{F}\text{-LK}}$ was more powerful than T_{LK} (Figure 8). We performed single tests on the basis of the empirical distribution of T_{LK} and $T_{\mathcal{F}\text{-LK}}$, on each SNP, and we accounted for multiple testing using the Benjamini–Hochberg (BH) correction, which controls the false discovery rate (BENJAMINI and HOCHBERG 1995). The threshold for significance was set at 0.05. We also performed tests on the basis of the chi-square distribution (as in TESTS OF SELECTION section).

Single tests performed using T_{LK}, with either its empirical distribution or the chi-square distribution, pinpointed three outliers, *ESR*, *MQ30*, and *GHRHR* (Table 2). After correction for multiple tests (BH), there was no significant outlier. Single tests performed using $T_{\mathcal{F}\text{-LK}}$ with its empirical distribution pinpointed seven outliers (*NRAMP*, *HAL*, *ESR*, *REN*, *MQ30*, *MX1*, and *GHRHR*). Using the chi-square distribution, four outliers were detected (*HAL*, *ESR*, *REN*, and *MQ30*). After correction for multiple tests, only *ESR* and *MQ30* were significant. Overall, after correction for multiple tests, results of the chi-square test were similar to those obtained using the empirical distributions, but *P*-values were higher (Table 2), as expected since the chi-square distribution was more conservative in this case (Figure 8). Population SNP allele frequencies allowed us to identify the population(s) in which selection occurred. In our case, directional selection seems to have occurred in the Large White breed for a gene involved in reproductive performance (*ESR*) and for another gene *MQ30* (Figure 9). In addition, we confirmed that directional selection had occurred at the Halothane gene (*HAL*) in the Piétrain breed (Figure 9).

Figure 10 shows the neutral distribution of T_{LK} and $T_{\mathcal{F}\text{-LK}}$ conditional on heterozygosity (following the work of BEAUMONT and NICHOLS 1996), for the four pig breeds studied. T_{LK} and $T_{\mathcal{F}\text{-LK}}$ have similar shapes although T_{LK} has a slightly broader distribution for heterozygosity values >0.2. The SNPs *ESR*, *HAL*, and *MQ30* lie beyond the 0.999 quantile of the $T_{\mathcal{F}\text{-LK}}$ neutral envelope, with similarity to the single-test *P*-values we obtained (Table 2).

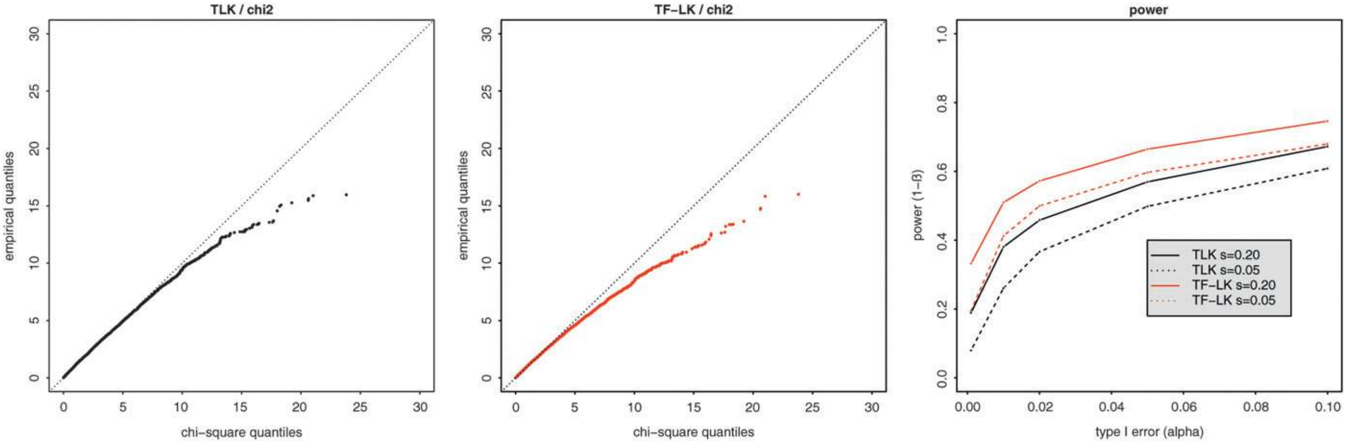


FIGURE 8.—Fit to the χ^2-distribution and power analysis of T_{LK} and $T_{F\text{–LK}}$ for a scenario mimicking the pig data set. The $\mathcal{F}$ matrix was estimated using data on 50 microsatellites. Forward simulations were parameterized conditionally on the $\mathcal{F}$ matrix.

DISCUSSION

We proposed an extension of LEWONTIN and KRAKAUER's (1973) method to detect signatures of selection in species with complex population trees, under pure genetic drift. We focused here on SNP data, but the method can also be applied to multiallelic loci. Using simulations of various population trees with or without selection, we compared the robustness and power of the original LK test, based on the T_{LK} statistic, and of the extension we proposed, based on the $T_{F\text{–LK}}$ statistic. In some simulation scenarios, comparisons with a model-based MCMC method (FOLL and GAGGIOTTI 2008) were also performed.

Empirical distributions of T_{LK} and $T_{F\text{–LK}}$ under neutrality: Simulations under neutrality indicate that the empirical distributions of T_{LK} and $T_{F\text{–LK}}$ are similar. They both do not fit the right-tail side of the chi-square distribution when including extreme p_0 values ($0 < p_0 < 1$), while they fit the chi-square distribution when considering only intermediate p_0 values (*i.e.*, $0.2 < p_0 < 0.8$). These observations hold whatever the demographic history of the populations (EBL, UBL, or UBL struc) and whether the parameters p_0 and $\mathcal{F}$ are estimated or not. The long right tail of the test distributions in the presence of extreme p_0 values results in an excess of false positives if the chi-square distribution is used as the null distribution for the test. Therefore, it is recommended rather to use the empirical distribution of the tests, which we did when evaluating the power of the methods. Alternatively, p_0 estimates at each tested SNP could be used as a proxy for choosing which distribution (empirical or theoretical) should be preferred to perform tests based on T_{LK} and $T_{F\text{–LK}}$.

The lack of fit of the T_{LK} and $T_{F\text{–LK}}$ distributions to the chi-square distribution in the case of extreme p_0 values can be explained as follows. First, these statistics are ratios (see Equations 9 and 24) and our derivations of their expected values and variances imply a first-order approximation of these ratios. When p_0 tends to zero or one, the denominators of the statistics become very large and this approximation is less accurate. Second, our derivations assume that the allele frequencies are normally distributed, which is also violated for extreme p_0 values.

Focusing on intermediate allele frequencies makes our derivations more accurate, and the good fit of the $T_{F\text{–LK}}$ distribution with the chi-square distribution is thus natural. More surprising is the equally good fit for UBL scenarios of the T_{LK} distribution with the chi-square distribution in this case. We note, however, that this result is consistent with the ones obtained by BEAUMONT and NICHOLS (1996), who showed that the F_{ST} distribution is robust to variations in the population structure for intermediate heterozygosity values. In the case of the UBL models, one likely explanation for the robustness of T_{LK} is that restricting to intermediate p_0 values effectively conditions on allele frequency trajectories that are compatible with the EBL hypothesis, therefore reducing the effect of population size differences. In the case of more complex structured models, this explanation alone may not be sufficient. But, as pointed out by BEAUMONT (2005), we can advocate the separation-of-timescales approximation (NORDBORG 1997; WAKELEY 1999, 2001; WAKELEY and ALIACAR 2001), which implies that in a wide range of structured population models, the allele frequencies can be approximated by the ones of a UBL model where several populations evolve independently from a common ancestral pool.

Another interesting issue is that the use of SNP data satisfies in principle one assumption underlying LK tests, *i.e.*, that mutations occur only in the ancestral population (the collecting phase of the separation-of-timescales approximation). Indeed, one criterion of the SNP ascertainment phase is that both alleles at a SNP marker must segregate in several of the populations studied, implying that the mutated allele is relatively ancient. Therefore, LK tests with SNP data can be applied to recently bifurcating populations (*i.e.*, live-

TABLE 2

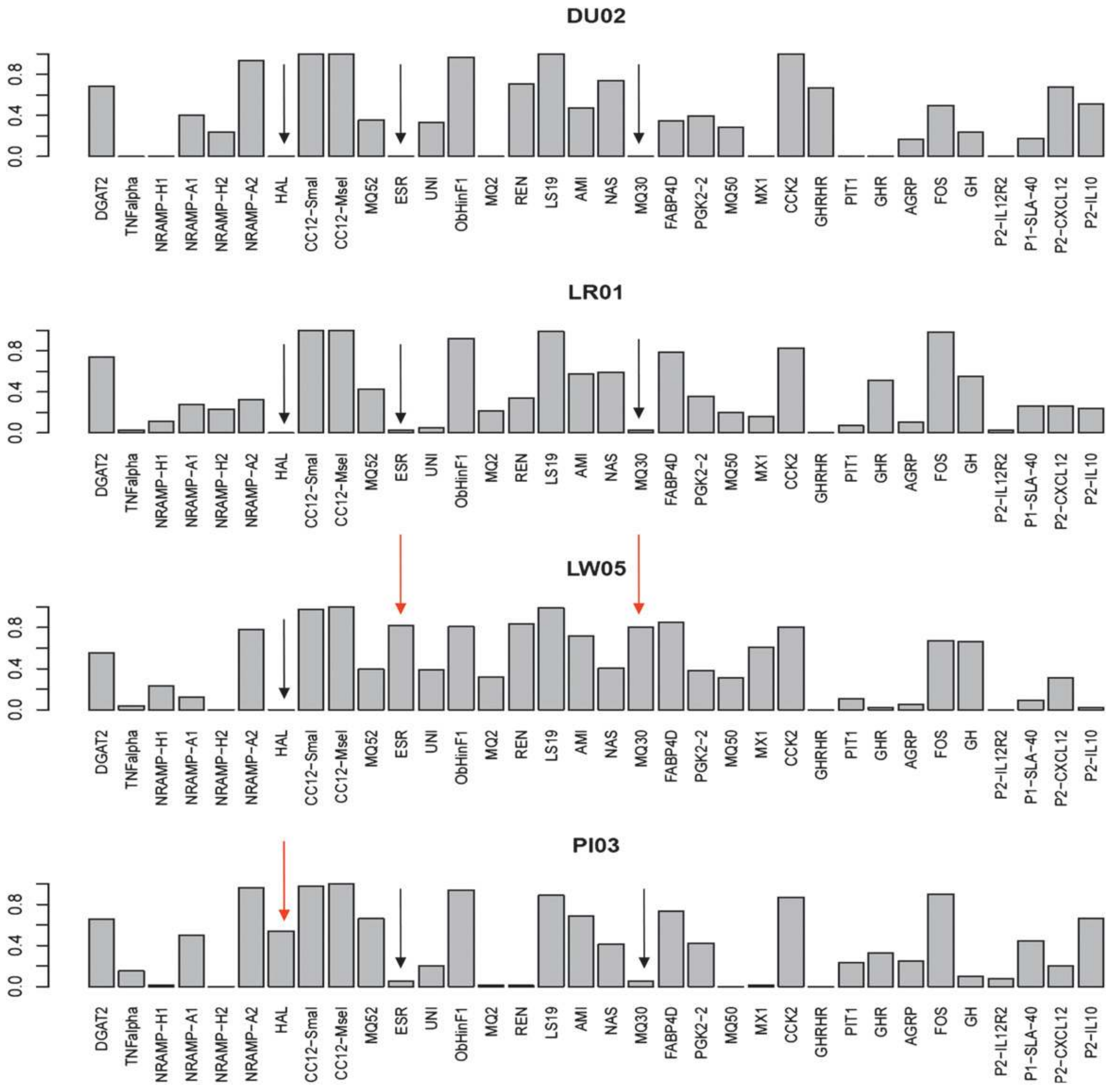
Nominal and corrected *P*-values on a 34-SNPs data set from PigBioDiv2, based on the empirical distribution of *T* and $T_{F\text{–}LK}$ and on the theoretical χ^2-distribution

	Empirical test				Chi-square test			
	T_{LK}		$T_{F\text{–}LK}$		T_{LK}		$T_{F\text{–}LK}$	
SNP name	*P*-value	BH	*P*-value	BH	*P*-value	BH	*P*-value	BH
9CP-DGAT2	0.9584	1.0000	0.9339	1.0000	0.9523	1.0000	0.9314	1.0000
23CP-TNFα	0.8150	1.0000	0.7344	1.0000	0.7934	1.0000	0.7323	1.0000
32CP-NRAMP-H1	0.6540	1.0000	0.5565	1.0000	0.6337	1.0000	0.5652	1.0000
41CP-NRAMP-A1	0.7213	1.0000	0.6128	1.0000	0.6967	1.0000	0.6185	1.0000
42CP-NRAMP-H2	0.5562	1.0000	0.4985	1.0000	0.5424	1.0000	0.5094	1.0000
43CP-NRAMP-A2	0.1313	0.6377	**0.0333**	0.1887	0.1361	0.6612	0.5090	0.3345
66CP-HAL	0.0521	0.4428	**0.0048**	0.0544	0.0583	0.4954	**0.0168**	0.1907
68CP-CC12-Smal	0.9833	1.0000	0.9751	1.0000	0.9803	1.0000	0.9707	1.0000
67CP-CC12-Msel	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
69CP-MQ52	0.8420	1.0000	0.7756	1.0000	0.8236	1.0000	0.7717	1.0000
105CP-ESR	**0.0074**	0.1513	**0.0005**	**0.0085**	**0.0129**	0.2465	**0.0022**	**0.0447**
106CP-UNI	0.7148	1.0000	0.5689	1.0000	0.6907	1.0000	0.5760	1.0000
107CP-ObHinF1	0.8841	1.0000	0.8544	1.0000	0.8681	1.0000	0.8454	1.0000
108CP-MQ2	0.4934	1.0000	0.3838	1.0000	0.4775	1.0000	0.4040	1.0000
109CP-REN	0.0792	0.5385	**0.0204**	0.1734	0.0846	0.5754	**0.0427**	0.3345
110CP-LS19	0.8117	1.0000	0.7115	1.0000	0.7901	1.0000	0.7086	1.0000
111CP-AMI	0.8985	1.0000	0.9049	1.0000	0.8839	1.0000	0.8996	1.0000
112CP-NAS	0.7641	1.0000	0.7703	1.0000	0.7409	1.0000	0.7661	1.0000
113CP-MQ30	**0.0089**	0.1513	**0.0005**	**0.0085**	**0.0145**	0.2465	**0.0026**	**0.0447**
8CP-FABP4D	0.4267	1.0000	0.5327	1.0000	0.4105	1.0000	0.5415	1.0000
100CP-PGK2-2	0.9993	1.0000	0.9985	1.0000	0.9983	1.0000	0.9969	1.0000
104CP-MQ50	0.7092	1.0000	0.5852	1.0000	0.6862	1.0000	0.5913	1.0000
219CP-MX1	0.0991	0.5615	**0.0291**	0.1887	0.1054	0.5970	0.0541	0.3345
220CP-CCK2	0.8598	1.0000	0.8877	1.0000	0.8426	1.0000	0.8824	1.0000
228CP-GHRHR	**0.0171**	0.1938	**0.0468**	0.2273	**0.0225**	0.2555	0.0731	0.3551
229CP-PITI	0.7637	1.0000	0.7041	1.0000	0.7402	1.0000	0.7006	1.0000
230CP-GHR	0.2283	0.8998	0.1555	0.5874	0.2260	0.8888	0.1860	0.7025
231CP-AGRP	0.8873	1.0000	0.8219	1.0000	0.8713	1.0000	0.8160	1.0000
232CP-FOS	0.3783	1.0000	0.3247	1.0000	0.3637	1.0000	0.3439	1.0000
233CP-GH	0.3426	1.0000	0.1914	0.6507	0.3307	1.0000	0.2182	0.7419
234CP-P2-IL12R2	0.8972	1.0000	0.8613	1.0000	0.8821	1.0000	0.8525	1.0000
235CP-P1-SLA-40	0.7084	1.0000	0.5956	1.0000	0.6856	1.0000	0.6025	1.0000
236CP-P2-CXCL12	0.5138	1.0000	0.5993	1.0000	0.4987	1.0000	0.6053	1.0000
237CP-P2-IL10	0.2382	0.8998	0.1211	0.5146	0.2353	0.8888	0.1521	0.6463

P-value is for a single test. BH is *P*-value corrected for multiple testing, according to the Benjamini–Hochberg method (controlled for false discovery rate). *P*-values considered as significant at the 5% level are in boldface type, showing outlier SNPs most likely under directional selection. Four populations are studied: DU02, LW05, PI03, and LR01.

stock, recently colonizing or invasive populations), but also in principle to deeply divergent populations, provided the selected SNPs segregate in several of the populations studied. In contrast, the use of multiallelic loci (*i.e.*, microsatellites) should be handled with caution because they can potentially have mutated more recently (in the scattering phase of the separation-of-timescales approximation). This can affect the distribution of F_{ST} (Flint *et al.* 1999; Storz *et al.* 2004) and therefore the results of LK tests.

Power of T_{LK} and $T_{F\text{–}LK}$: If selection acts on a large population, $T_{F\text{–}LK}$ is more powerful than T_{LK}. This difference of power is remarkable at low type I errors. In UBL and UBL struc models $T_{F\text{–}LK}$ detects 20% and 15–35% more selected SNPs than T_{LK}, respectively. However, if selection acts on a small population, T_{LK} may be more powerful than $T_{F\text{–}LK}$ for UBL models, although this trend disappears for low type I errors. To interpret these observations, let us consider the simpler case of a UBL model where the $\mathcal{F}$ matrix is known. In this case, $T_{F\text{–}LK}$ is proportional to $\sum_{i=1}^{n}(1/F_i)(p_i - p_0)^2$, so that populations with a large F_i (*i.e.*, a small population size) have little influence on the distribution of the statistic. Thus, the relative size of the population where selection occurs has a strong impact on the power of the test. On the other hand, T_{LK} is proportional to $\sum_{i=1}^{n}(p_i - p_0)^2$, so that all populations have the same weight and the size of the population where selection


DU02

LR01

LW05

PI03

Figure 9.—Allelic frequencies of 34 SNPs in four major pig breeds of the PigBioDiv2 project. Arrows pinpoint the outlier SNPs, black in populations where no selection occurs, red in populations where selection occurs.

occurs does not directly matter. In fact, the disadvantage of T_{LK} compared to $T_{\mathcal{F}\text{–LK}}$ is its larger variance, due to the fact that it does not account for the F_i's. For large type I errors, the power of the test is essentially determined by the difference between the expected value of the statistic under selection and under neutrality, so the larger variance of T_{LK} is not an important problem. However, for very small type I errors, this problem of variance has a clear negative impact on T_{LK}'s power. It is important to note here that the small type I errors are the most relevant in practical applications, because genomic scans for selection have to deal with an important multiple-testing issue.

In practice, the $\mathcal{F}$ matrix is unknown and the power of $T_{\mathcal{F}\text{–LK}}$ will depend on how well it can be estimated. In our simulations, only a small percentage of SNPs were influenced by selection due to hitchhiking. Consequently, $\mathcal{F}$ was in general well estimated and the power of $T_{\mathcal{F}\text{–LK}}$ with an estimated $\mathcal{F}$ was almost as good as with a known $\mathcal{F}$. However, it is advisable to be cautious when testing dense SNP genotyping data in only a few genomic regions. In our application to pig SNP data, we avoided this bias by estimating the $\mathcal{F}$ matrix with an independent data set of microsatellite loci. Remarkably, the power of $T_{\mathcal{F}\text{–LK}}$ depends on a comprehensive population sampling in a given population tree, because the estimation of the $\mathcal{F}$ matrix is less biased when the population in which selection occurs is "diluted" among a high number of populations.

When lots of populations are tested and nearly neutral multilocus genotypes are available, the phylogenetic framework is perhaps the most convenient way of estimating the $\mathcal{F}$ matrix, as was proposed in this work. However, when the population number is not too large, alternative methods such as approximate Bayesian computation (ABC) methods (Beaumont *et al.* 2002; Marjoram *et al.* 2003) could be considered, as they potentially deal with more summary statistics than only

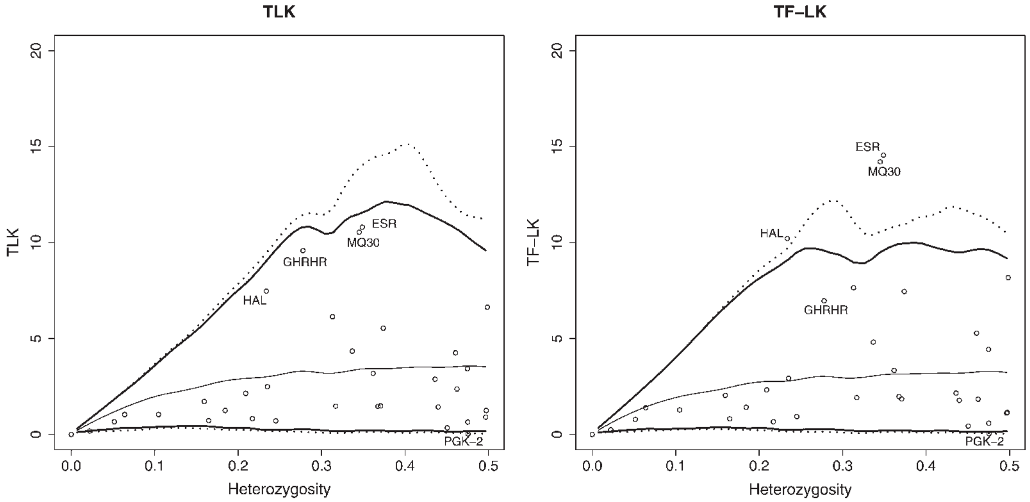


FIGURE 10.—Distributions of T_{LK} and $T_{\mathcal{F}-\mathrm{LK}}$ conditional on heterozygosity and test for outliers with 34 SNPs (candidate genes) of the PigBioDiv2 project. Top, middle, and bottom solid lines delineate the neutral envelope containing 98% of the values, with the mean values. Top and bottom dotted lines delineate the 99.8% envelope.

one distance to infer the population tree necessary to calculate the $\mathcal{F}$ matrix.

The ancestral allele frequency p_0 of the selected allele has a complex influence on the detection power of T_{LK} and $T_{\mathcal{F}-\mathrm{LK}}$. On one hand, extreme p_0 values induce a long right tail of the statistic distributions under neutrality, which reduces the power. On the other hand, the evidence of selection is stronger if the selected allele was initially at low frequency (saying it differently, the difference between the expected value of the statistics under selection and under neutrality is larger for small p_0 values). The combination of these two antagonistic effects implies that conditioning on intermediate p_0 values may lead to either an increase or a decrease of power, depending on the evolution scenario and on the test. As already outlined above, the size of the population where selection occurs will have more effect on the results obtained with $T_{\mathcal{F}-\mathrm{LK}}$ than on those obtained with T_{LK}. Indeed, conditioning on intermediate p_0 values will increase the power of $T_{\mathcal{F}-\mathrm{LK}}$ if selection acts in a large population, but decrease it if selection acts in a small population. These observations may be important to understand the influence of SNP ascertainment, which typically favors alleles with intermediate ancestral frequencies, on the detection power of the tests.

Software: A general workflow for the application of the test to a real data set is presented in Figure 11. We implemented R and python codes that (i) compute the matrix of Reynolds' genetic distances (LAVAL *et al.* 2002) between populations from a matrix of SNP genotype frequencies, (ii) compute a NJ tree from this Reynolds' matrix (or another Reynolds' matrix if provided), (iii) build an estimate of the $\mathcal{F}$ matrix from the output of the NJ tree, (iv) compute the test statistics, and (v) compute the χ^2-approximated P-values, the empirical distribution of the test statistics under the null (conditioned on $\mathcal{F}$), and the null envelope conditioned on heterozygosity. The codes and the pig data files are available at http://qgp.jouy.inra.fr/flk or as File S1 and File S1 cont.

Methodological perspectives: Some methodological issues arise from these observations. First, the F_{ST} distribution (analogous to the T_{LK} statistic) was shown to be sensitive to complex patterns of migration and sharp differences in the migration rate among populations [island models, hierarchically structured models (BEAUMONT and NICHOLS 1996; EXCOFFIER *et al.* 2009)]. The sensitivity of $T_{\mathcal{F}-\mathrm{LK}}$ to correlations of allele frequencies among populations due to migration events should also be considered with regard to robustness and power. Although gene flow among closely related populations should not in principle bias the estimation of the population tree—the bias would concern only branch lengths after the split—gene flow among distantly related populations is expected to mask the true population tree. Second, a simulation study of the robustness and power of $T_{\mathcal{F}-\mathrm{LK}}$ when testing multiallelic loci with a high mutation rate, such as microsatellite loci or haplotypes, would be interesting.

CONCLUSION

A practical motivation for the development of an extension of the LK test was to provide a powerful and rapid parametric statistical test for detecting the signature of selection in somewhat complex population trees with large marker data sets in many populations. BEAUMONT and BALDING (2004) and FOLL and GAGGIOTTI (2008) developed Bayesian hierarchical models on the basis of a multinomial-Dirichlet likelihood that arises naturally under the separation-of-timescales approximation. These methods explicitly model population-specific (β-) effects that actually correspond to variation of the inbreeding coefficient F (or F_{ST}) among populations. The fact that these methods implement robust statistical modeling, including likelihood expression and estimation using MCMC, makes them computationally prohibitive for large marker data sets and large numbers of populations. On the other hand, methods

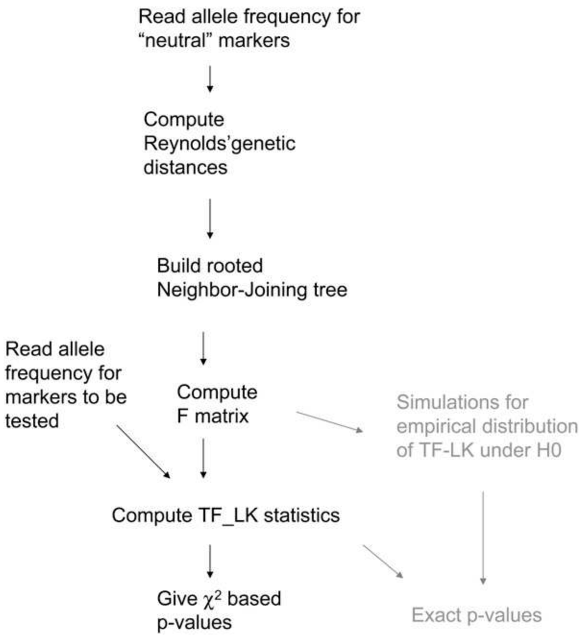

Read allele frequency for "neutral" markers
Compute Reynolds'genetic distances
Build rooted Neighbor-Joining tree
Read allele frequency for markers to be tested
Compute F matrix
Simulations for empirical distribution of TF-LK under H0
Compute TF_LK statistics
Give χ^2 based p-values
Exact p-values

Figure 11.—Workflow for the *F*–LK test. Shaded terms illustrate optional items.

based on an island model (Beaumont and Nichols 1996) or a hierarchically structured model (Excoffier *et al.* 2009) are computationally convenient and quite conservative, but may tend to omit more complex demographic histories involving N_e variation among populations and historical branching. To help in screening large marker data sets for outliers in relatively complex population trees, we propose an additional method that accounts for N_e variation among populations and historical branching, assuming pure genetic drift and no migration in its current state. The statistical test based on either the empirical distribution of the $T_{F\text{–LK}}$ statistic or the theoretical chi-square distribution is generally more powerful than a classical LK test based on T_{LK}. In scenarios where the populations are hierarchically structured, it is also more powerful than the MCMC method of Foll and Gaggiotti (2008). This extended LK test thus represents a quick and powerful tool in the context of genomic scans for selection using population data.

We thank David Robelin for the Quantitative Genetic Platform web site and Juliette Riquet, Hélène Gilbert, and Maria Inès Fariello for helpful discussions. The authors thank the DNA providers of the European project PigBioDiv (Bio4-CT-0188). J.A. acknowledges financial support from PigBioDiv2 (QLK5-CT-2002-01059) and M.B. acknowledges financial support from DéLiSus (ANR-07-GANI-001).

LITERATURE CITED

Balding, D., 2003 Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. **63**(3): 221–230.

Barreiro, L., G. Laval, H. Quach, E. Patin and L. Quintana-Murci, 2008 Natural selection has driven population differentation in modern humans. Nat. Genet. **40**(3): 340–345.

Beaumont, M., 2005 Adaptation and speciation: What can f-st tell us? Trends Ecol. Evol. **20**(8): 435–440.

Beaumont, M., and D. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. **13**(4): 969–980.

Beaumont, M., and R. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. Ser. B Biol. Sci. **263**(1377): 1619–1626.

Beaumont, M., W. Zhang and D. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate—a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. **57**(1): 289–300.

Blott, S., L. Andersson, M. Groenen, M. San Cristobal, C. Chevalet *et al.*, 2003 Characterisation of genetic variation in the pig breeds of China and Europe—the pigbiodiv2 project. Arch. Zootecnia **52**(198): 207–217.

Eveno, E., C. Collada, M. Guevara, V. Léger, A. Soto *et al.*, 2008 Contrasting patterns of selection at pinus pinaster ait. drought stress candidate genes as revealed by genetic differentiation analyses. Mol. Biol. Evol. **25**(2): 417–437.

Excoffier, L., T. Hofer and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity **103**(4): 285–298.

Flint, J., J. Bond, D. Rees, A. Boyce, J. Roberts-Thomson *et al.*, 1999 Minisatellite mutational processes reduce f(st) estimates. Hum. Genet. **105**(6): 567–576.

Flori, L., S. Fritz, F. Jaffrézic, M. Boussaha, I. Gut *et al.*, 2009 The genome response to artificial selection: a case study in dairy cattle. PLoS ONE **4**(8): e6595.

Foll, M., and O. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics **180:** 977–993.

Gautier, M., L. Flori, A. Riebler, F. Jaffrezic, D. Laloe *et al.*, 2009 A whole genome Bayesian scan for adaptive genetic divergence in west African cattle. BMC Genomics **10:** 550.

Guo, F., D. Dey and K. Holsinger, 2009 A Bayesian hierarchical model for analysis of snp diversity in multilocus, multipopulation samples. J. Am. Stat. Assoc. **104**(485): 142–154.

Hudson, R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**(2): 337–338.

Laval, G., M. SanCristobal and C. Chevalet, 2002 Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet. Sel. Evol. **34**(4): 481–507.

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

Lewontin, R. C., and J. Krakauer, 1975 Testing heterogeneity of f-values. Genetics **80:** 397–398.

Luikart, G., P. England, D. Tallmon, S. Jordan and P. Taberlet, 2003 The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. **4**(12): 981–994.

Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100**(26): 15324–15328.

Nei, M., and A. Chakravarti, 1977 Drift variances of fst and gst statistics obtained from a finite number of isolated populations. Theor. Popul. Biol. **11**(3): 307–325.

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes—comment. Genetics **80:** 395.

Nei, M., A. Chakravarti and Y. Tateno, 1977 Mean and variance of fst in a finite number of incompletely isolated populations. Theor. Popul. Biol. **11**(3): 291–306.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197–218.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

Reynolds, J., B. Weir and C. Cockerham, 1983 Estimation of the co-ancestry coefficient—basis for a short-term genetic-distance. Genetics **105:** 767–779.

Riebler, A., L. Held and W. Stephan, 2010 Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics **178:** 1817–1829.

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

ROBERTSON, A., 1975b Lewontin-Krakauer test for neutral genes—comment. Genetics **80:** 396.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method—a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4): 406–425.

SANCRISTOBAL, M., C. CHEVALET, C. S. HALEY, R. JOOSTEN, A. P. RATTINK *et al.*, 2006 Genetic diversity within and between European pig breeds using microsatellite markers. Anim. Genet. **37:** 189–198.

STORZ, J., B. PAYSEUR and M. NACHMAN, 2004 Genome scans of dna variability in humans reveal evidence for selective sweeps outside of africa. Mol. Biol. Evol. **21**(9): 1800–1811.

TANABE, K., and M. SAGAE, 1992 An exact Cholesky decomposition and the generalized inverse of the variance–covariance matrix of the multinomial distribution, with applications. J. R. Stat. Soc. Series B Stat. Methodol. **54:** 211–219.

TSAKAS, S., and C. KRIMBAS, 1976 Testing the heterogeneity of f values: a suggestion and a correction. Genetics **84:** 399–401.

VITALIS, R., K. DAWSON and P. BOURSOT, 2001 Interpretation of variation across marker loci as evidence of selection. Genetics **158:** 1811–1823.

WAKELEY, J., 1999 Nonequilibrium migration in human history. Genetics **153:** 1863–1871.

WAKELEY, J., 2001 The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. **59**(2): 133–144.

WAKELEY, J., and N. ALIACAR, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

WEIR, B., and W. HILL, 2002 Estimating f-statistics. Annu. Rev. Genet. **36:** 721–750.

Communicating editor: M. W. FELDMAN

APPENDIX A: DISTRIBUTION OF LEWONTIN AND KRAKAUER'S TEST IN A STRUCTURED POPULATION

In the following, we derive the first two moments of the basic test (Equation 10).

We first write the sum of the numerator in Equation 9, in matrix product,

$$\sum_i (p_i - \bar{p})^2 = \mathbf{p}' \cdot \mathbf{M}_{\mathrm{LK}} \cdot \mathbf{p}, \tag{A1}$$

where **p** is the n-vector of p_i's, $\mathbf{M}_{\mathrm{LK}}$ is the $n \times n$ matrix equal to $\mathbf{I}-(1/\mathrm{n})\mathbf{E}$, **I** is the $n \times n$ identity matrix, and **E** is the $n \times n$ matrix made up of 1's. The expectation can be written as

$$\mathbb{E}\Big(\sum_i (p_i - \bar{p})^2\Big) = \mathbb{E}(\mathbf{p}') \cdot \mathbf{M}_{\mathrm{LK}} \cdot \mathbb{E}(\mathbf{p}) + \mathrm{trace}(\mathbf{M}_{\mathrm{LK}} \cdot \boldsymbol{G}), \tag{A2}$$

where $\mathbf{G} = p_0(1-p_0)\boldsymbol{\mathcal{F}}$ is the variance–covariance matrix of frequencies. The first term is 0 since all p_i's have the same expectation p_0 (hence $\mathbf{M}_{\mathrm{LK}} \cdot \mathrm{E}(\mathbf{p})$=0). Further,

$$\mathrm{trace}(\mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G}) = p_0(1-p_0)\Big(\mathrm{trace}(\boldsymbol{\mathcal{F}}) - \frac{1}{n}\mathrm{trace}(\mathbf{E} \cdot \boldsymbol{\mathcal{F}})\Big) \tag{A3}$$

$$= p_0(1-p_0)\left(\sum_i f_{ii} - \frac{1}{n}\sum_i \sum_j f_{ij}\right). \tag{A4}$$

Denoting by $\bar{F}$ and $\bar{f}$ the mean value of fixation indexes F_i and the mean value of the fixation indexes f_{ij} attached to ancestral populations common to all pairs of observed populations (Equations 16 and 17), one gets

$$\mathbb{E}\Big(\sum_i (p_i - \bar{p})^2\Big) = (n-1)(\bar{F} - \bar{f})p_0(1-p_0); \tag{A5}$$

hence

$$\mathbb{E}(s_p^2) = (\bar{F} - \bar{f})p_0(1-p_0). \tag{A6}$$

Similarly, we have

$$\mathbb{E}(\bar{p}(1-\bar{p})) = p_0(1-p_0) - p_0(1-p_0)\left(\bar{f} + \frac{1}{n}(\bar{F} - \bar{f})\right), \tag{A7}$$

where the second term is equal to minus the variance of $\bar{p}$. In fact the expression $\bar{f} + (1/n)(\bar{F} - \bar{f})$ can be shown to be small, in general smaller than the reciprocal of the number n of populations.

Turning to the expectation of F_{ST}, the error made when replacing the expectation of the ratio by the ratio of expectations is of the same order of magnitude ($<$1:n), so that we can write

$$\mathbb{E}(F_{\mathrm{ST}}) \simeq \bar{F} - \bar{f}. \tag{A8}$$

Assuming normality, the sum of squares (Equation A1) has a variance equal to

$$\mathbb{V}\left(\sum_i (p_i - \bar{p})^2\right) = 2\,\mathrm{trace}(\mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G} \cdot \mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G}). \tag{A9}$$

We have

$$\mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G} \cdot \mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G} = p_0^2(1-p_0)^2\left(\mathbf{I} - \frac{1}{n}\mathbf{E}\right) \cdot \boldsymbol{\mathcal{F}} \cdot \left(\mathbf{I} - \frac{1}{n}\mathbf{E}\right) \cdot \boldsymbol{\mathcal{F}}; \tag{A10}$$

hence

$$\mathrm{trace}(\mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G} \cdot \mathbf{M}_{\mathrm{LK}} \cdot \mathbf{G}) = p_0^2(1-p_0)^2\left(\mathrm{trace}(\boldsymbol{\mathcal{F}}^2) - \frac{2}{n}\mathrm{trace}(\boldsymbol{\mathcal{F}} \cdot \mathbf{E} \cdot \boldsymbol{\mathcal{F}}) + \frac{1}{n^2}\mathrm{trace}(\mathbf{E} \cdot \boldsymbol{\mathcal{F}} \cdot \mathbf{E} \cdot \boldsymbol{\mathcal{F}})\right) \tag{A11}$$

since the *trace* operator is commutative. Denoting by a dot the summation over indexes $\left(\sum_i f_{ij} = f_{.j}, \sum_i \sum_j f_{ij} = f_{..}\right)$, we have

$$\mathrm{trace}(\boldsymbol{\mathcal{F}}^2) = \sum_i \sum_j f_{ij}^2 \tag{A12}$$

$$\mathrm{trace}(\boldsymbol{\mathcal{F}} \cdot \mathbf{E} \cdot \boldsymbol{\mathcal{F}}) = \sum_i f_{i.}^2 \tag{A13}$$

$$\mathrm{trace}(\mathbf{E} \cdot \boldsymbol{\mathcal{F}} \cdot \mathbf{E} \cdot \boldsymbol{\mathcal{F}}) = f_{..}^2. \tag{A14}$$

As before, we assume that the number of populations is large enough for the variance of F_{ST} to be approximated by the ratio of the variance of the numerator, as calculated above, to the square of the expectation of $\bar{p}(1-\bar{p})$ (Equation A7).

Assuming that the number of loci is large enough for the variance of $\bar{F}_{\mathrm{ST}}$ (Equation 10) to be neglected, the previous expressions allow the first two moments of the test to be derived for any coancestry structure (matrix $\boldsymbol{\mathcal{F}}$) of the populations, Equations 13 and 15.

Robertson (1975a) considered the case of a structured history causing correlations between allele frequencies (nonzero f_{ij} values, with equal branch lengths ($F_i = f_{ii} = F$). With these conditions expressions (A12), (A13), and (A14) become

$$\begin{aligned}
\mathrm{trace}(\boldsymbol{\mathcal{F}}^2) &= nF^2 + \sum_i \sum_{j \neq i} f_{ij}^2 \\
&= nF^2 + n(n-1)\bar{f}^2 + \sum_i \sum_{j \neq i} (f_{ij} - \bar{f})^2 \\
\mathrm{trace}(\boldsymbol{\mathcal{F}} \cdot E \cdot \boldsymbol{\mathcal{F}}) &= \sum_i (F + \sum_{j \neq i} f_{ij})^2 \\
&= nF^2 + 2n(n-1)F\bar{f} + \sum_i \left(\sum_{j \neq i} f_{ij}\right)^2 \\
&= nF^2 + 2n(n-1)F\bar{f} + \sum_i \left(\sum_{j \neq i} (f_{ij} - \bar{f}) + (n-1)\bar{f}\right)^2 \\
&= nF^2 + 2n(n-1)F\bar{f} + n(n-1)^2\bar{f}^2 + \sum_i \left(\sum_{j \neq i} (f_{ij} - \bar{f})\right)^2 \\
\mathrm{trace}(E \cdot \boldsymbol{\mathcal{F}} \cdot E \cdot \boldsymbol{\mathcal{F}}) &= (nF + n(n-1)\bar{f})^2 \\
&= n^2(F^2 + 2(n-1)F\bar{f} + (n-1)^2\bar{f}^2).
\end{aligned}$$

Setting

$$v_1 = \frac{\sum_i \sum_{j \neq i} (f_{ij} - \bar{f})^2}{n(n-1)}$$

and

$$v_2 = \frac{\sum_i \left(\sum_{j \neq i} (f_{ij} - \bar{f}) \right)^2}{n(n-1)^2},$$

the sum

$$\mathrm{trace}(\boldsymbol{\mathcal{F}}^2) - \frac{2}{n}\mathrm{trace}(\boldsymbol{\mathcal{F}} \cdot E \cdot \boldsymbol{\mathcal{F}}) + \frac{1}{n^2}\mathrm{trace}(E \cdot \boldsymbol{\mathcal{F}} \cdot E \cdot \boldsymbol{\mathcal{F}})$$

in Equation A11 becomes equal to

$$(n-1)((F - \bar{f})^2 + nv_1 - 2(n-1)v_2).$$

Comparing with Robertson's notations (Robertson 1975a, p. 785), his δ_{ij} is

$$\delta_{ij} = \frac{f_{ij} - \bar{f}}{F};$$

$v_1 = (F - \bar{f})^2$ times his $V_{r'}$ term, which is the variance of "internal" correlation coefficients between gene frequencies in different populations defined as

$$r'_{ij} = \frac{f_{ij} - \bar{f}}{\bar{F} - \bar{f}};$$

and v_2 corresponds to a second term he found small with respect to the first one, to get Equation 18.

In the case of independence between populations (the tree has a star structure), but heterogeneous F_i values (the populations show different heterozygosities), we get another simplified expression. Assuming no correlation between allele frequencies ($\bar{f} = 0$), the expectation is not changed,

$$\mathbb{E}(F_{\mathrm{ST}}) = \bar{F}, \tag{A15}$$

and the previous expressions for the variance become

$$\mathrm{trace}(\boldsymbol{\mathcal{F}}^2) = \sum_i F_i^2 \tag{A16}$$

$$\mathrm{trace}(\boldsymbol{\mathcal{F}} \cdot E \cdot \boldsymbol{\mathcal{F}}) = \sum_i F_i^2 \tag{A17}$$

$$\mathrm{trace}(E \cdot \boldsymbol{\mathcal{F}} \cdot E \cdot \boldsymbol{\mathcal{F}}) = n^2 \bar{F}^2 \tag{A18}$$

so that we get

$$\mathbb{V}\Big(\sum_i (p_i - \bar{p})^2 \Big) = 2p_0^2(1-p_0)^2 \Big(\Big(1 - \frac{2}{n}\Big) \sum_i F_i^2 + \bar{F}^2 \Big) \tag{A19}$$

$$= 2p_0^2(1-p_0)^2(n-1)(\bar{F}^2 + \Big(1 - \frac{2}{n}\Big)\mathbb{V}(F)) \tag{A20}$$

if we set

$$\mathbb{V}(F) = \frac{1}{n-1} \sum_i (F_i - \bar{F})^2. \tag{A21}$$

Then, the variance of T_{LK} is changed from $2(n-1)$, the value corresponding to a chi-square distribution, to

$$\mathbb{V}(T_{\mathrm{LK}}) \simeq 2(n-1)\Big(1+\Big(1-\frac{2}{n}\Big)\frac{\mathbb{V}(F)}{\bar{F}^2}\Big). \tag{A22}$$

Evaluating the variance of F_i values can be obtained from the variance of Reynolds' distances R_{ij}, which estimate the mean F values of populations i and j from their proximal common ancestor population. Indeed, with no correlation $(\bar{f}=0)$, $R_{ij}=\frac{1}{2}(F_i+F_j)$, so that

$$\mathbb{V}(F) = 2\mathbb{V}(R). \tag{A23}$$

APPENDIX B: MULTIALLELIC VERSION OF BASIC AND EXTENDED LK TESTS

In the following we extend the test to the case of multiallelic markers.

Consider a locus with A alleles. Let $\mathbf{P}=(\mathbf{p}'_1,\ldots,\mathbf{p}'_a,\ldots,\mathbf{p}'_A)'$ denote the nA-vector of allele frequencies sorted by population within allele number: $\mathbf{p}_a$ denotes the n-vector of frequencies of allele a in the n populations. Under drift,

$$\mathbb{E}(\mathbf{P}) = (p_{01}\mathbf{1}'_n,\ \ldots\ , p_{0a}\mathbf{1}'_n,\ \ldots\ , p_{0A}\mathbf{1}'_n)' = \mathbf{p}_0\otimes\mathbf{1}_n, \tag{B1}$$

where ⊗ denotes the Kronecker product and $\mathbf{p}_0$ is now the A-vector of founder allele frequencies. The variance of $\mathbf{P}$,

$$\mathbb{V}(\mathbf{P}) = \mathbf{B}_0\otimes\boldsymbol{\mathcal{F}}, \tag{B2}$$

involves the $(n\times n)$-matrix $\boldsymbol{\mathcal{F}}$ and the $(A\times A)$-matrix $\mathbf{B}_0=\mathrm{diag}(\mathbf{p}_0)-\mathbf{p}_0\mathbf{p}'{}_0$. The estimator of founder frequencies is now $\hat{\mathbf{P}}_0=(\mathbf{1}'_n\mathbf{w}'\mathbf{p}$ and can be written as $(\mathbf{I}_A\otimes\mathbf{1}_n\mathbf{w}')\mathbf{P}$, with $\mathbf{w}$ as in Equation 21. The multiallelic equivalent of $T_{F-\mathrm{LK}}(p_0)$ in (23) is

$$\tilde{T}_{F-\mathrm{LK}}(\mathbf{P}_0) = (\mathbf{P}-\mathbf{P}_0)'(\mathbf{B}_0\otimes\boldsymbol{\mathcal{F}})^{-1}(\mathbf{P}-\mathbf{P}_0) \tag{B3}$$

$$= (\mathbf{P}-\mathbf{P}_0)'(\mathbf{B}_0^-\otimes\boldsymbol{\mathcal{F}}^{-1})(\mathbf{P}-\mathbf{P}_0), \tag{B4}$$

where $\mathbf{B}_0^-$ is the Moore–Penrose generalized inverse of $\mathbf{B}_0$. It can be explicitly written as (Tanabe and Sagae 1992)

$$\mathbf{B}_0^- = (\mathbf{I}_A-\mathbf{1}_A\mathbf{1}'_A)\mathrm{diag}^{-1}(\mathbf{p}_0)(\mathbf{I}_A-\mathbf{1}_A\mathbf{1}'_A). \tag{B5}$$

Replacing $\mathbf{p}_0$ with $\hat{\mathbf{p}}_0$ in $\mathbf{P}_0$ and $\mathbf{B}_0$ leads to the quadratic form

$$\begin{aligned}\tilde{T}_{F-\mathrm{LK}} &= (\mathbf{P}-\hat{\mathbf{P}}_0)'(\hat{\mathbf{B}}_0\otimes\boldsymbol{\mathcal{F}})^{-1}(\mathbf{P}-\hat{\mathbf{P}}_0)\\ &= \mathbf{P}'(\mathbf{I}_A-\mathbf{I}_{nA}\otimes\mathbf{1}_n\mathbf{w})'(\hat{\mathbf{B}}_0^-\otimes\boldsymbol{\mathcal{F}}^{-1})(\mathbf{I}_A\otimes\mathbf{I}_{nA}-\mathbf{1}_n\mathbf{w}')\mathbf{P}\\ &= P(\hat{\mathbf{B}}_0^-\otimes\mathbf{M})\mathbf{P},\end{aligned}$$

where $\mathbf{M}$ is the $(n\times n)$ matrix in Equation 26.

In the particular case when the number of alleles is two, $\tilde{T}_{F-\mathrm{LK}}$ reduces to $T_{F\text{-}\mathrm{LK}}$ in (24), so that considering one of the two alleles or both alleles is equivalent.

From the calculation of the moments of $\tilde{T}_{F-\mathrm{LK}}$ (see below), we get

$$\mathbb{E}(\tilde{T}_{F-\mathrm{LK}}) \approx (n-1)(A-1) \tag{B6}$$

$$\mathbb{V}(\tilde{T}_{F-\mathrm{LK}}) \approx 2(n-1)(A-1) \tag{B7}$$

so that $\tilde{T}_{F-\mathrm{LK}}$ has approximately a $\chi^2_{(n-1)(A-1)}$-distribution under the null hypothesis of genetic drift.

Moment calculations: The same type of demonstration as in appendix a is used for the extension of the LK test, so that only main results are presented.

When $\boldsymbol{P_0}$ is known: The expectation of the statistic test is

$$\begin{aligned}\mathbb{E}(\tilde{T}_{F-\mathrm{LK}}(P_0)) &= \mathrm{trace}[(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}})^{-}\mathbb{V}(\mathbf{P})]\\ &= \mathrm{trace}[(\mathbf{B}_0^{-} \otimes \boldsymbol{\mathcal{F}}^{-1})(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}})]\\ &= \mathrm{trace}[(\mathbf{B}_0^{-}\mathbf{B}_0) \otimes (\boldsymbol{\mathcal{F}}^{-1}\boldsymbol{\mathcal{F}})]\\ &= \mathrm{trace}(\mathbf{I}_n)\mathrm{trace}(\mathbf{B}_0^{-}\mathbf{B}_0)\\ &= n(A-1).\end{aligned}$$

Similarly, assuming approximate normality,

$$\begin{aligned}\mathbb{V}(\tilde{T}_{F-\mathrm{LK}}(\mathbf{P}_0)) &= 2\,\mathrm{trace}[(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}})^{-}\mathbb{V}(\mathbf{P})(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}})^{-}\mathbb{V}(\mathbf{P})]\\ &= 2n(A-1).\end{aligned}$$

When $\boldsymbol{P}_0$ is unknown: First, note that $\mathbf{P} - \hat{\mathbf{P}}_0 = (\oplus_{a=1}^{A}(\mathbf{I}_n - \mathbf{W}))\mathbf{P}$, where $\mathbf{W}$ is the $(n \times n)$ matrix built with identical lines equal to $\mathbf{w}$ (Equation 21). It can also be shown that the ath diagonal element of $\mathbf{B}_0$ is $\sum_{b\neq a} 1/p_{0,b}$ and the (a, b) element is equal to $\sum_{c\neq a,b} 1/p_{0,c}$.

The quadratic form $\tilde{T}_{F-\mathrm{LK}}$ can be written as $\mathbf{P}'\tilde{\mathbf{M}}\mathbf{P}$, with

$$\begin{aligned}\tilde{\mathbf{M}} &= (\oplus_{a=1}^{A}(\mathbf{I}_n - \mathbf{W}))'(\hat{\mathbf{B}}_0^{-} \otimes \boldsymbol{\mathcal{F}}^{-1})(\oplus_{a=1}^{A}(\mathbf{I}_n - \mathbf{W}))\\ &= \hat{\mathbf{B}}_0^{-} \otimes [(\mathbf{I}_n - \mathbf{W})'\boldsymbol{\mathcal{F}}^{-1}(\mathbf{I}_n - \mathbf{W})]\\ &= \hat{\mathbf{B}}_0^{-} \otimes \mathbf{M}\end{aligned}$$

with $\mathbf{M}$ defined in Equation 26. Then, incidently, $\tilde{T}_{F-\mathrm{LK}}$ can be written as

$$\tilde{T}_{F-\mathrm{LK}} = \sum_{a,b} p_a^{'}\mathbf{M}\mathbf{p}_b \cdot \left(\sum_{c\neq a,b}\frac{1}{\hat{p}_{0,c}}\right) \tag{B8}$$

with $\sum_a \mathbf{p}_a = \mathbf{1}_n$.

Coming back to the matrix notations, and following calculation lines of the biallelic case (Equations 22, 27, and 28), but neglecting the bias term in (22), $\hat{\mathbf{B}}_0$ is replaced by its expectation $\mathbf{B}_0$, and

$$\begin{aligned}\mathbb{E}(\tilde{T}_{F-\mathrm{LK}}) &= \mathbb{E}(\mathbf{P})'\tilde{\mathbf{M}}\mathbb{E}(\mathbf{P}) + \mathrm{trace}(\tilde{\mathbf{M}}\mathbb{V}(\mathbf{P}))\\ &= \mathbf{P}'{}_0(\mathbf{B}_0^{-} \otimes \mathbf{M})\mathbf{P}_0 + \mathrm{trace}((\mathbf{B}_0^{-} \otimes \mathbf{M})(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}}))\\ &= \mathrm{trace}((\mathbf{B}_0^{-}\mathbf{B}_0) \otimes (\mathbf{M}\boldsymbol{\mathcal{F}}))\\ &= \mathrm{trace}(\mathbf{M}\boldsymbol{\mathcal{F}})\mathrm{trace}(\mathbf{B}_0^{-}\mathbf{B}_0)\\ &= (n-1)(A-1),\end{aligned}$$

since

$$\mathrm{trace}(\mathbf{M}\boldsymbol{\mathcal{F}}) = \mathrm{trace}\left(\mathbf{I}_n - \frac{\boldsymbol{\mathcal{F}}^{-1}\mathbf{1}\mathbf{1}'}{\mathbf{1}'\boldsymbol{\mathcal{F}}^{-1}\mathbf{1}}\right) = n-1, \tag{B9}$$

and $\mathbf{B}_0$ has rank $(A-1)$.

Similarly, assuming approximate normality,

$$\begin{aligned}\mathbb{V}(\tilde{T}_{F-\mathrm{LK}}) &= 4\mathbb{E}(\mathbf{P})'\tilde{\mathbf{M}}\mathbb{V}(\mathbf{P})\tilde{\mathbf{M}}\mathbb{E}(\mathbf{P}) + 2\,\mathrm{trace}(\tilde{\mathbf{M}}\mathbb{V}(\mathbf{P})\tilde{\mathbf{M}}\mathbb{V}(\mathbf{P}))\\ &= 4\mathbf{P}'{}_0(\mathbf{B}_0^{-} \otimes \mathbf{M})(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}})(\mathbf{B}_0^{-} \otimes \mathbf{M})\mathbf{P}_0\\ &\quad +\ 2\,\mathrm{trace}((\mathbf{B}_0^{-} \otimes \mathbf{M})(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}})(\mathbf{B}_0^{-} \otimes \mathbf{M})(\mathbf{B}_0 \otimes \boldsymbol{\mathcal{F}}))\\ &= 2\,\mathrm{trace}((\mathbf{B}_0^{-}\mathbf{B}_0) \otimes (\mathbf{M}\boldsymbol{\mathcal{F}}) \cdot (\mathbf{B}_0^{-}\mathbf{B}_0) \otimes (\mathbf{M}\boldsymbol{\mathcal{F}}))\\ &= 2\,\mathrm{trace}(\mathbf{M}\boldsymbol{\mathcal{F}}\mathbf{M}\boldsymbol{\mathcal{F}})\mathrm{trace}(\mathbf{B}_0^{-}\mathbf{B}_0\mathbf{B}_0^{-}\mathbf{B}_0)\\ &= 2(n-1)(A-1),\end{aligned}$$

since $\mathbf{M}\boldsymbol{\mathcal{F}}\mathbf{M}\boldsymbol{\mathcal{F}} = \mathbf{M}\boldsymbol{\mathcal{F}}$ and $\mathbf{B}_0\mathbf{B}_0^{-}\mathbf{B}_0 = \mathbf{B}_0$ by definition of the generalized inverse.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.117275/DC1

Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended

Maxime Bonhomme, Claude Chevalet, Bertrand Servin, Simon Boitard, Jihad Abdallah, Sarah Blott and Magali SanCristobal

DOI: 10.1534/genetics.110.117275

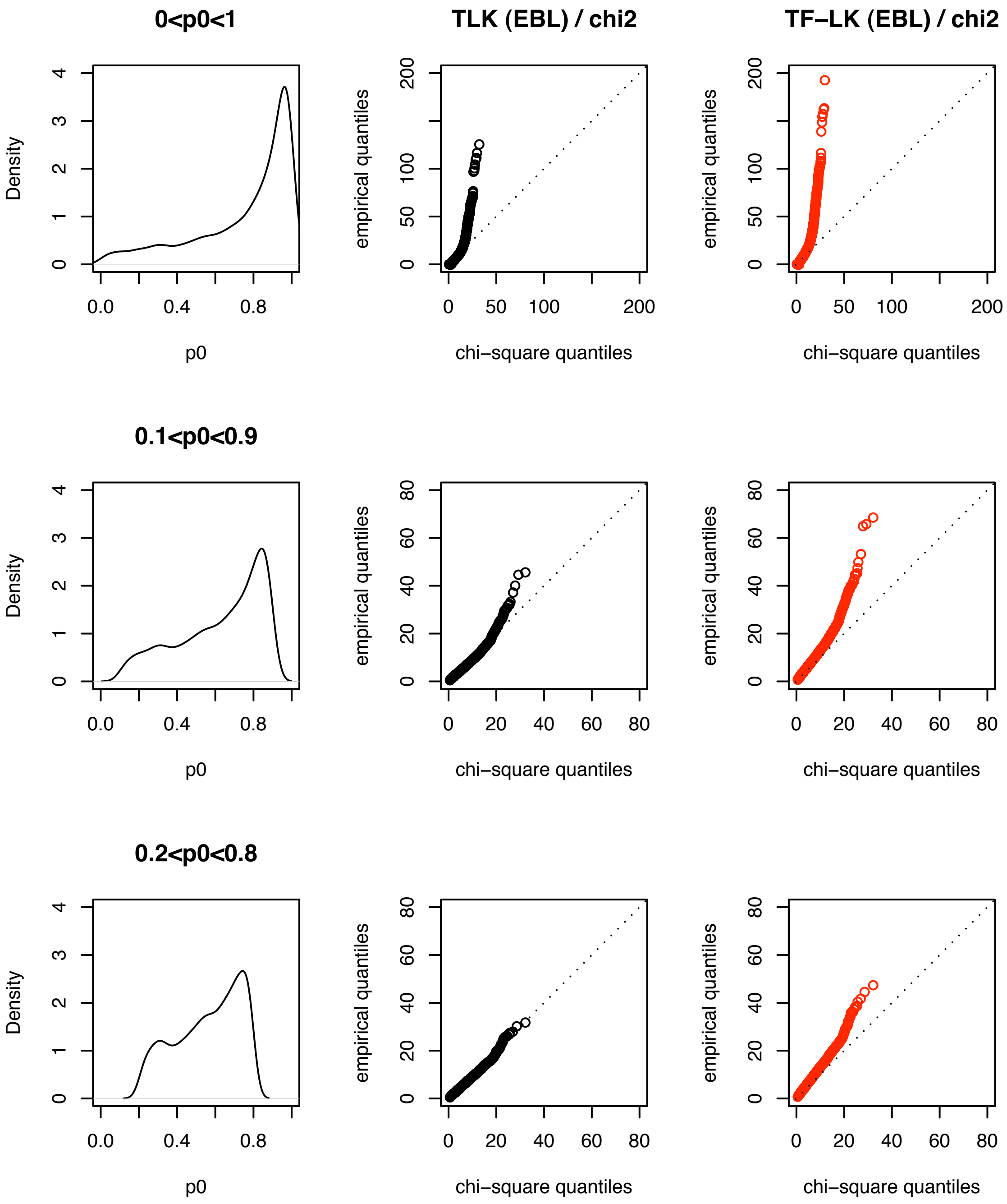


FIGURE S1.—This figure reports the same description as Figure 3, but for an EBL model.

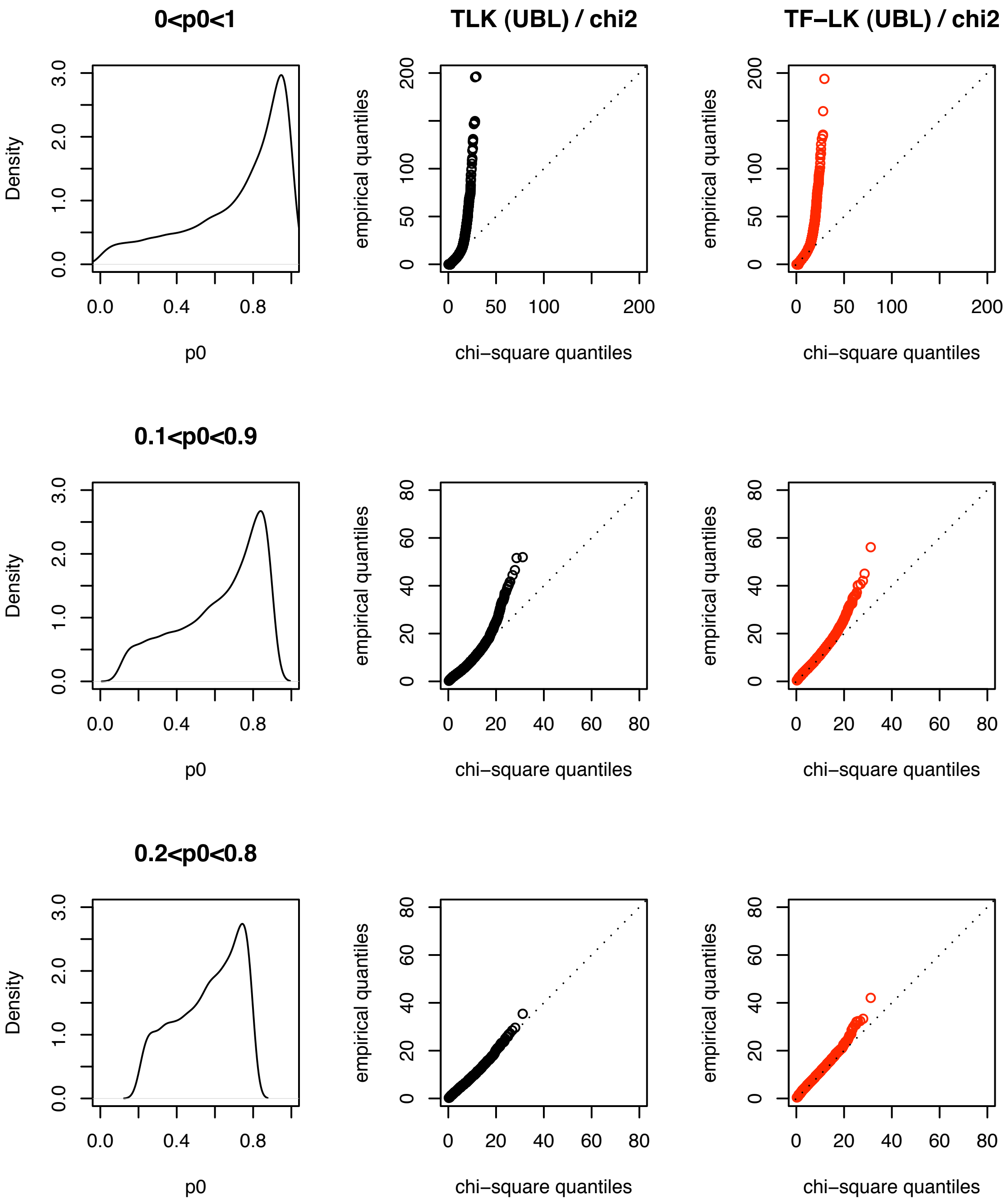


FIGURE S2.—This figure reports the same description as Figure 3, but for an UBL model.

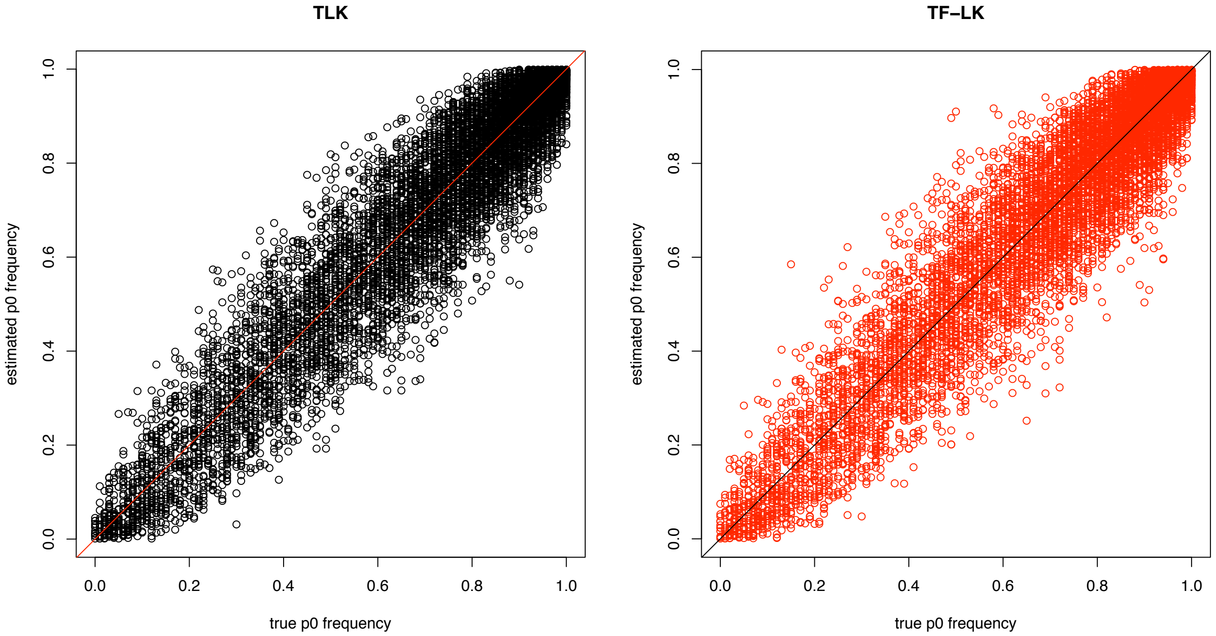


FIGURE S3.—This figure illustrates the correlation between p_0 estimates and true p_0 values recorded in the simulations. The T_{LK} and $T_{F\text{-}LK}$ estimators of p_0 can be considered as relevant estimators, with minimum variance.

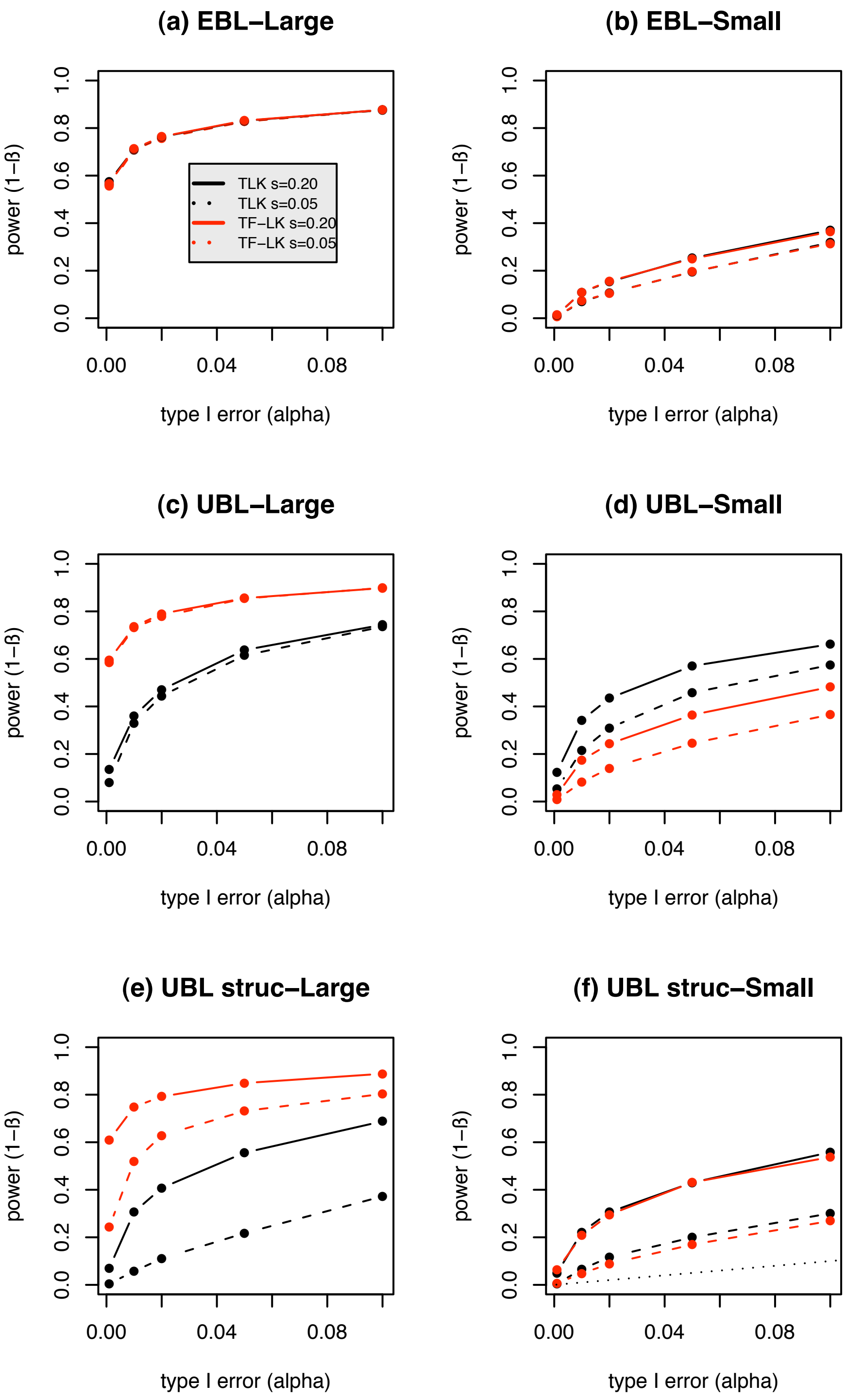
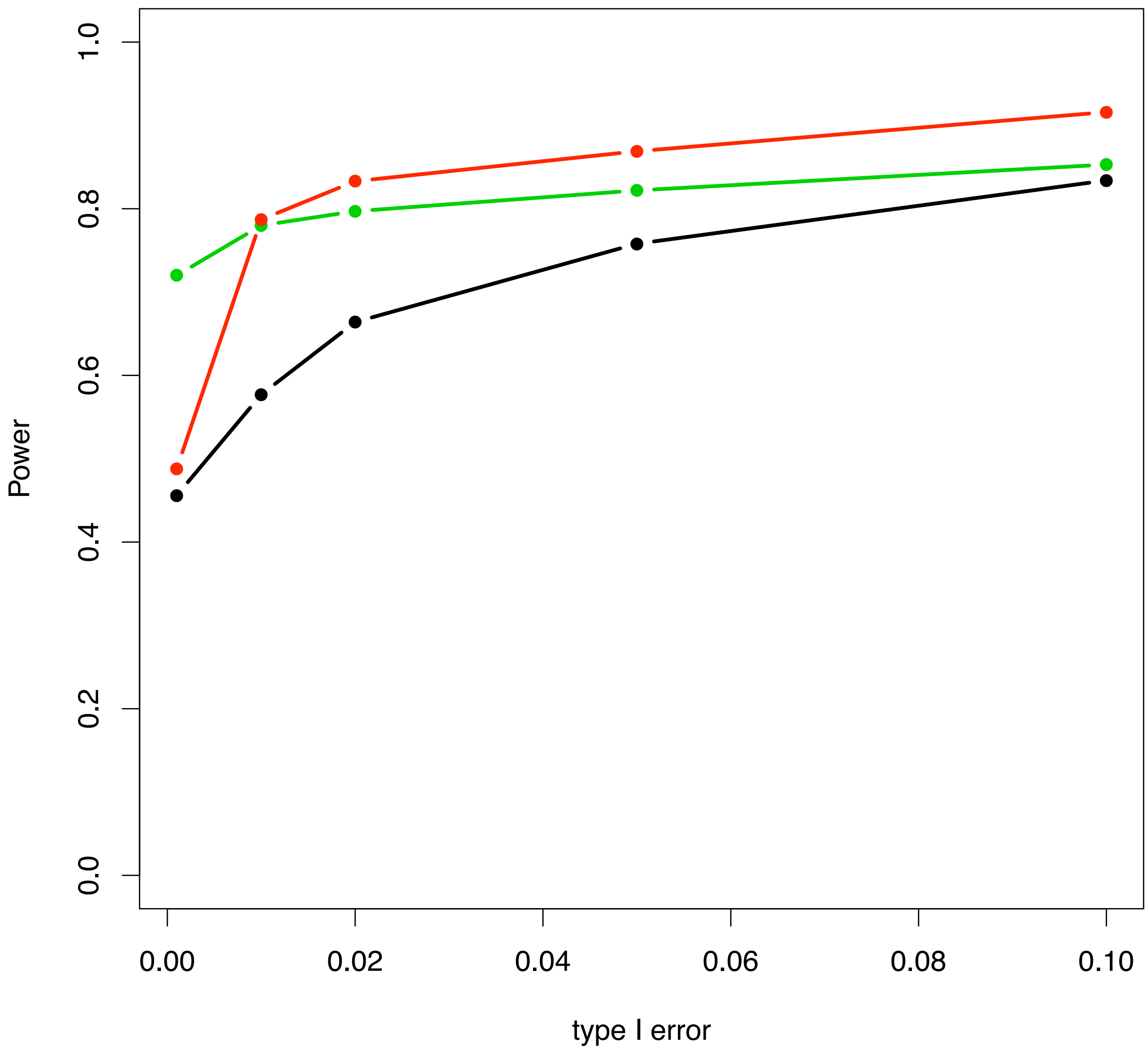


FIGURE S4.—This figure reports the same description as Figure 4 but for $p_0$and $\boldsymbol{F}$ equal to their true value.

A



B

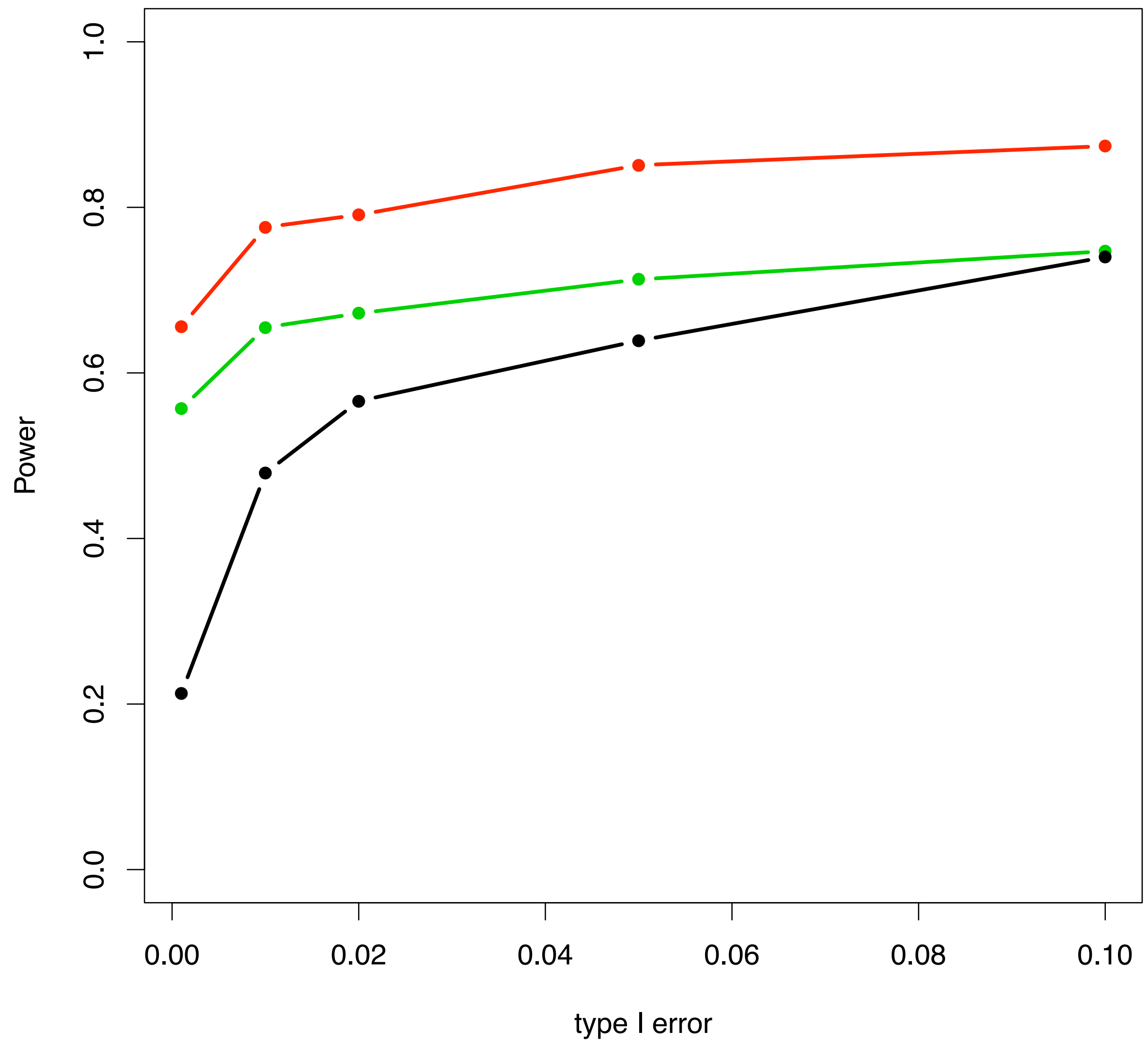


FIGURE S5.—This figure reports the same description as Figure 7 with a selection coefficient equal to *s*=0.20.

FILE S1

Detecting selection in population trees: the Lewontin and Krakauer test extended

This page is documentation on how to compute the extended LK test (named FLK) on a SNP (biallelic marker) dataset. Instructions are provided along with a set of input file examples containing the pig data analyzed in Bonhomme et al.

All files are available for download as FileS1.zip at http://www.genetics.org/cgi/content/full/genetics.110.117275/DC1. Additionally, this information is available at http://qgp.jouy.inra.fr/flk.

Principle
The principle of the test is to compare patterns of differences between **allele frequencies** in several populations to their expectation under a neutral evolution. The null hypothesis of neutral evolution assumes a **tree structure** with branch length corresponding to the amount of genetic drift in each population (**F**). This tree is estimated from the **matrix of Reynold's genetic distances** between populations, using the neighbor joining (NJ) algorithm.

Software requirements
All software needed are freely available on all common computer operating systems. Please install the following required packages to use the programs provided.

The test calculations are performed using R. The **ape** package is needed to estimate the NJ tree.
To derive the empirical distributions of the test under neutrality, a python program is provided (see below). It requires the **simuPOP** and **numpy** packages to run.

Input

Main input
In order to perform the test, the user needs to provide data on allele frequencies for several populations. To build the population tree, the program needs an **outgroup** population used to root the NJ tree. This file contains one line per population. Each line starts with the population name, followed by the list of allele frequencies for this population.

As it is assumed markers are biallelic (SNP), only one allele frequency is needed per marker. It doesn't matter the allele which frequency is reported in the file, as long as it is the same allele for all populations.

Excerpt of the input file for the pig dataset

```
    GBDU02 0.6875 0 0 0.40425532 0.23958333 ...
    FRLR01 0.73958333 0.03125 0.11 0.27659574 0.23469388 ...
    GBLW05 0.55 0.03703704 0.22916667 0.125 0 ...
    DEPI03 0.65306122 0.15306122 0.01 0.5 0 ...
    FRMS01 0.3125 0 0 0 0.375 ...
```

For this dataset, the outgroup population is FRMS01.

Additional input
Additionally, the user may provide a file with the Reynolds genetic distances already computed. This is convenient (and recommended) if the SNP data is small and restricted to a few regions of the genome. The format of the file is as follows:
Each line contains first the population name and then the corresponding row of the matrix of reynolds genetic distances. It is assumed that the population order is the same for the row and the columns. Population names in this file **must** match the ones in the main input file, although the order might be different.

Reynolds Genetic Distances for the pig dataset.

```
    GBDU02 0.0000 0.2422 0.2850 0.3916 0.2647
    FRLR01 0.2422 0.0000 0.1732 0.3396 0.1501
    GBLW05 0.2850 0.1732 0.0000 0.3572 0.1774
    FRMS01 0.3916 0.3396 0.3572 0.0000 0.3436
    DEPI03 0.2647 0.1501 0.1774 0.3436 0.0000
```

In the pig data analysed by Bonhomme et al., the Reynolds genetic distances were computed from microsatellite data.

Computing FLK test

In order to compute the FLK test, you will need the R code provided in the file FLK.R.
We provide instructions to use the code through an example R session on the pig data. An analysis of the input file must follow the same steps. The R statements are in **bold** and comments in *italic*:

```
## import the functions
source('FLK.R')
## Read the SNP frequency data
freq=read.table('pig.dat',row.names=1)
## Read the matrix of Reynolds Genetic Distances
DR=read.table('pig.dist',row.names=1)
## Estimate the population tree with provided Reynolds matrix
F=Fij(freq,outgroup='FRMS01',D=DR)
## Alternatively estimate the population tree using Reynolds distances
## computed on the SNP data (not recommended here)
Fsnp=Fij(freq,outgroup='FRMS01')
## Now compute the FLK and LK tests
tests=FLK(freq,F)
```

The **FLK** R function returns a data frame where each line corresponds to results for a SNP. The order of the SNPs in the data frame is the same as on input. For each SNP, the function returns the mean heterozygosity (Ht), the FLK statistic (F.LK), the associated asymptotic p-value (F.LK.p.val), the original LK statistic (LK) and associated asymptotic p-value (LK.p.val).
Excerpt of the data frame obtained on the pig data:

```
Ht F.LK F.LK.p.val LK LK.p.val
0.45036473 4.422247e-01 0.931388145 0.34041443 0.95225673
0.10454975 1.286550e+00 0.732329480 1.03240612 0.79341130
0.15934366 2.034670e+00 0.565242014 1.71449799 0.63371538
0.43976966 1.783754e+00 0.618476643 1.43776139 0.69670761
0.20902125 2.316534e+00 0.509360914 2.14718524 0.54242600
0.37367636 7.443688e+00 0.059023131 5.54232704 0.13612880
```

Additional output files are created by the Fij function. It returns the estimated F matrix in a file named fij.txt and the NJ tree in the file named tree.txt. These files are needed to derive the empirical null distribution of the FLK statistic (see below).

Empirical null distribution of FLK

We provide a program called **FLKnull** to derive the empirical null distribution of the FLK statistic. This program performs simulations conditional on the dataset analysed (that is the population tree estimated from the data). The program needs the **fij.txt** and **tree.txt** files created by the Fij R function (see above).

Running the program

To run the program, open a terminal and go the directory containing the results of the analysis. Then just run the program by typing **python FLKnull**. This will perform 10,000 simulations conditional on the estimated population tree. Optionnaly, more (or less although not recommended) simulations can be specified as an argument to the program. For example typing **python FLKnull 50000** will lead to performing 50,000 simulations. Note that the simulation process can take some time.

FLKnull returns the empirical quantiles of the null distribution of the tests for different heterozygocities. The results are provided in an output file named '**envelope.txt**'. Each line of the file is composed of:

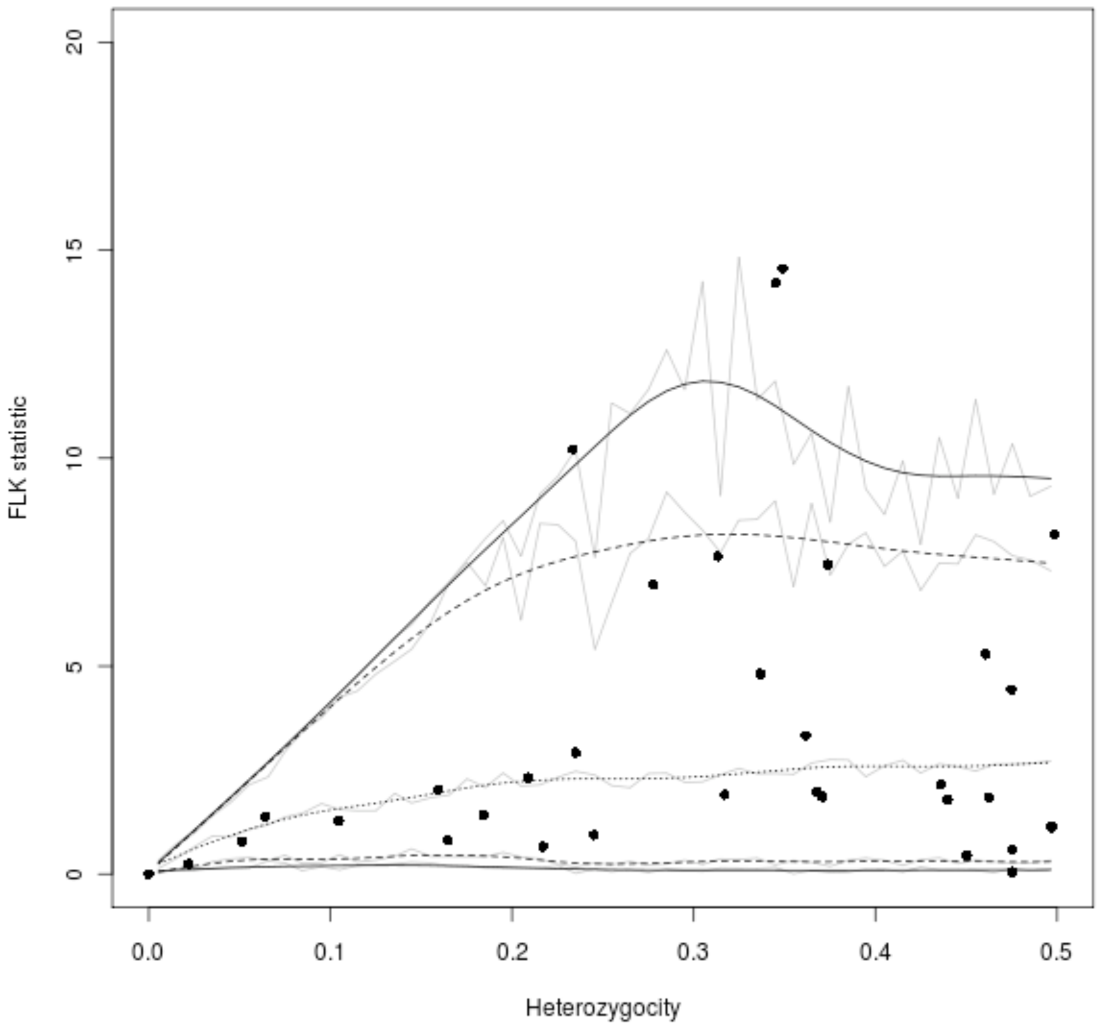
- Heterozygosity
- 0.005, 0.025, 0.5 (median), 0.975, 0.995 quantiles of the null distribution

The output file has a header as first line indicating the values for the different columns.

Plotting the distribution

We provide another R code to plot the null distribution envelope. This is done by calling **source('plotNull.R')** within your session (provided the output file 'envelope.txt' is in the current working directory). The actual estimated quantiles are plotted in gray. Because of the variance due to the simulation process, the envelope is better represented by fitting a spline on the actual quantiles. These are the lines represented in black: the solid lines correspond to the 0.005 and 0.995 quantiles, the dashed lines the 0.025 and 0.975 quantiles and the doted line to the median. If you find the variance around the spline to be too large, perform more simulations as explained above.

You can then add the observed value of your data by calling **points(tests$Ht,tests$F.LK,pch=16)**. On the pig data this results in the following figure:


FLK statistic
Heterozygocity