Frequency Spectrum Neutrality Tests: One for All and All for One

Guillaume Achaz[1]

*Systématique, Adaptation et Evolution (UMR 7138) and Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris VI,
Centre National de la Recherche Scientifique, Museum National d'Histoire Naturelle, Institut de Recherche pour le Développment,
75005 Paris, France*

ABSTRACT

Neutrality tests based on the frequency spectrum (*e.g.*, Tajima's *D* or Fu and Li's *F*) are commonly used by population geneticists as routine tests to assess the goodness-of-fit of the standard neutral model on their data sets. Here, I show that these neutrality tests are specific instances of a general model that encompasses them all. I illustrate how this general framework can be taken advantage of to devise new more powerful tests that better detect deviations from the standard model. Finally, I exemplify the usefulness of the framework on SNP data by showing how it supports the selection hypothesis in the lactase human gene by overcoming the ascertainment bias. The framework presented here paves the way for constructing novel tests optimized for specific violations of the standard model that ultimately will help to unravel scenarios of evolution.

T HE standard models of population genetics (*i.e.*, the Wright–Fisher model and related ones) constitute null models for which an amazing amount of theory has been developed. Population geneticists have used some aspect of the theory (*e.g.*, summary statistics) to test the goodness-of-fit of the standard model on a given data set. Rejection of the standard model typically suggests that alternative hypotheses, such as selection or demographic history, have to be accounted for. Although they test for more than neutrality, tests that compute the goodness-of-fit of the standard model have been referred to as "neutrality tests." Since different neutrality tests have varying sensitivity to different violations of the standard model, one typically uses a plethora of tests on the data set of interest. One then hopes that the evolutionary processes that generated the data set will be, at least partially, uncovered by the tests. Although neutrality tests based on population samples exhibit important diversity, they can be assigned to families such as "haplotype tests" (*e.g.*, Fu 1997; Depaulis and Veuille 1998) that use the distribution of haplotypes, "tree shape tests" that try to capture specific tree deformations (*e.g.*, Ramos-Onsins and Rozas 2002), and "frequency spectrum tests" that are based on the frequency spectrum (*e.g.*, Tajima 1989; Fu and Li 1993b; Fay and Wu 2000; Achaz 2008).

In this study, I investigate neutrality tests based on the frequency spectrum (hereafter referred to simply as neutrality tests) and show that they are all specific instances of a general framework. Neutrality tests compare two estimators of the population mutation parameter θ that characterizes the mutation–drift equilibrium. It is defined as $\theta = 2pN_e\mu$, where p is the ploidy (1 for haploids and 2 for diploids), N_e is the effective population size, and μ is the locus neutral mutation rate. When the standard model is true, the expectations of the several unbiased estimators of θ are equal.

Typical estimators of θ, in a sample of *n* sequences, are $\hat{\theta}_S = S/a_n$, where S is the number of polymorphic sites and $a_n = \sum_{i=1}^{n-1}(1/i)$ (Watterson 1975), and $\hat{\theta}_\pi = \pi$, where π is the average pairwise difference between all sequences in the sample (Tajima 1983). If an outgroup is available, mutations at frequency i/n can be distinguished from mutations at frequency $1 - i/n$. Following Fu (1995)'s notations, **ξ** is a vector that represents the unfolded frequency spectrum composed of ξ_i, the number of polymorphic sites at frequency i/n in the sample ($i \in [1, n - 1]$). When no outgroup is available, the frequency spectrum is folded and is given by a vector **η**, composed of η_i, the number of polymorphic sites at both frequencies i/n and $1 - i/n$. Accordingly, it has been shown that θ can be estimated from $\hat{\theta}_{\xi_1} = \xi_1$, with ξ_1 the number of derived singletons (Fu and Li 1993b), from $\hat{\theta}_{\eta_1} = ((n-1)/n)\eta_1$, with η_1 the total number of singletons (derived and ancestral) (Fu and Li 1993b), and from $\hat{\theta}_H = \sum_{i=1}^{n-1}(2i^2/(n(n-1)))\xi_i$ (Fay and Wu 2000). Recently, it has been suggested that singletons should be ignored when θ is estimated in samples with sequencing errors; this leads to estimators such as $\hat{\theta}_{\pi-\xi_1}, \hat{\theta}_{S-\xi_1}, \hat{\theta}_{\pi-\eta_1}$, and $\hat{\theta}_{S-\eta_1}$ (Achaz 2008). Other estimators of θ, such as $\hat{\theta}_\xi$ and $\hat{\theta}_\eta$, were designed to minimize their variance (Fu 1994b), although they can be computed using recursions only for a given value of θ.

Neutrality tests compute the goodness-of-fit of a statistic T, which is the difference between two estimators of θ, normalized by its standard deviation:

$$T = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\mathrm{Var}[\hat{\theta}_1 - \hat{\theta}_2]}} = \frac{t}{\sqrt{\alpha_n \theta + \beta_n \theta^2}}. \qquad (1)$$

For a given θ, under the standard model, T has a mean of $E[T] = 0$ and a variance of $\mathrm{Var}[T] = 1$. Lowercase letters (*e.g.*, t) denote the absolute difference (*i.e.*, the numerator only) and uppercase letters (*e.g.*, T) denote the normalized difference (Equation 1) throughout this work. Interestingly, the variance in the denominator is a function of both θ and θ^2. Because θ is unknown, the denominator cannot be computed as such. In practice, unbiased estimators of θ and θ^2 must be used instead. Because the variance of $\hat{\theta}_S$ vanishes asymptotically in a very large sample ($\lim_{n \to \infty} \mathrm{Var}[\hat{\theta}_S] = 0$), θ and θ^2 are, in practice, substituted by estimators based on S (TAJIMA 1989), which changes the mean and the variance of T to $E[T] \approx 0$ and $\mathrm{Var}[T] \approx 1$.

Tajima's D (TAJIMA 1989) is defined by $d = \hat{\theta}_\pi - \hat{\theta}_S$; the statistics proposed by FU and LI (1993b) are $f = \hat{\theta}_\pi - \hat{\theta}_{\xi_1}$, $f^* = \hat{\theta}_\pi - \hat{\theta}_{\eta_1}$, $d_2 = \hat{\theta}_S - \hat{\theta}_{\xi_1}$, and $d_2^* = \hat{\theta}_S - \hat{\theta}_{\eta_1}$. Another classical statistic is $h = \hat{\theta}_\pi - \hat{\theta}_H$ (FAY and WU 2000), even though its variance was not given by the authors. Finally, two other related neutrality tests that are, *a priori*, immune to sequencing errors were proposed: $y = \hat{\theta}_{\pi_{-\xi_1}} - \hat{\theta}_{S_{-\xi_1}}$ and $y^* = \hat{\theta}_{\pi_{-\eta_1}} - \hat{\theta}_{S_{-\eta_1}}$ (ACHAZ 2008). Other tests based on θ_ξ and θ_η (which are optimized for a given θ-value) as well as the difference between the observed and the expected values of the frequency spectrum were also proposed (FU 1996).

Here, I show that when using a general weighted linear combination of $\hat{\theta}_i = i\xi_i$ (or $\hat{\theta}_i^*$ when no outgroup is available), any estimators of θ [*i.e.*, $\hat{\theta}_\omega = (1/\sum \omega_i) \sum_i \omega_i i\xi_i$] and consequently any neutrality tests can be derived. NAWA and TAJIMA (2008) recently advocated the use of the $\hat{\theta}_i^*$ spectrum, which is expected to be uniform under the standard model, as a visual test for neutrality instead of the classical frequency spectrum. This last proposal is in complete agreement with the current work. Importantly, it has been previously reported that some θ-estimators and neutrality tests could be expressed as specific linear combinations of ξ_i or η_i (TAJIMA 1997; WAKELEY 2009). Furthermore, FU (1997) shows that several θ-estimators can be expressed as specific linear combinations of $\hat{\theta}_i$ ($\hat{\theta}_{L(x)} = (1/\sum i^{-x}) \sum_i i^{-x} i\xi_i$) or in a related framework that uses $\hat{\theta}_i^*$ instead of $\hat{\theta}_i$. $\hat{\theta}_H$ was subsequently designed as $\hat{\theta}_{L(-1)}$ (FAY and WU 2000). However, some estimators (like $\hat{\theta}_\pi$, $\hat{\theta}_{\pi_{-\xi_1}}$, or $\hat{\theta}_{S_{-\xi_1}}$) cannot be expressed using the FU (1997) framework. To the best of my knowledge, no previous study has explicitly derived the framework presented here. No work has yet highlighted the striking simplicity of θ-estimators and related tests, when expressed in this framework. I further show how the use of such a simple framework greatly facilitates the study of previous θ-estimators and their related neutrality tests and how it opens the door for constructing yet undiscovered interesting θ-estimators and neutrality tests with enhanced power.

MODEL

With an outgroup: According to FU (1995), we know that

$$E[\xi_i] = \theta/i \qquad (2)$$

$$\mathrm{Var}[\xi_i] = \theta/i + \sigma_{ii}\theta^2 \qquad (3)$$

$$\mathrm{Cov}[\xi_i, \xi_j] = \sigma_{ij}\theta^2, \qquad (4)$$

where σ_{ii} and σ_{ij} depend only on n and are given in Equation 2 of FU (1995). This shows that $E[i\xi_i] = \theta$ and therefore that any ξ_i can be used to construct an unbiased estimator of θ:

$$\hat{\theta}_i = i\xi_i. \qquad (5)$$

Consequently, a linear combination $\hat{\theta}_\omega$ of the $\hat{\theta}_i$'s (in which the weights sum to 1) is also an unbiased estimator of θ. Mathematically, it is expressed as

$$\hat{\theta}_\omega = \frac{1}{\sum_i \omega_i} \sum_{i=1}^{n-1} \omega_i\, i\xi_i, \qquad (6)$$

where ω_i is the weight of each $\hat{\theta}_i$ in the combined estimator. Therefore, any estimator based on the frequency spectrum can be solely described by an **ω**-vector. Importantly, it should be mentioned that FU (1997) also proposed a linear combination of $i\xi_i$, but in which only a subset of the weight vectors was used. Namely, the proposed weight vectors were restricted to $\omega_i = i^{-x}$.

Using Equations 3 and 4 the variance of $\hat{\theta}_\omega$ can be shown to be

$$\mathrm{Var}[\hat{\theta}_\omega]$$

$$= \left(\sum_i \omega_i\right)^{-2} \left(\sum_{i=1}^{n-1} \omega_i^2 i^2 \mathrm{Var}[\xi_i] + 2\sum_i \sum_{j>i} ij\omega_i\omega_j \mathrm{Cov}[\xi_i, \xi_j]\right)$$

$$= \left(\sum_i \omega_i\right)^{-2}$$

$$\times \left(\theta\left(\sum_i \omega_i^2 i\right) + \theta^2\left(\sum_i \omega_i^2 i^2 \sigma_{ii} + 2\sum_i \sum_{j>i} ij\omega_i\omega_j\sigma_{ij}\right)\right).$$

$$(7)$$

Following TAJIMA (1989), using Equation 1, one can compute a normalized statistic that is, in the general framework,

$$T_\Omega = \frac{\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}}{\mathrm{Var}^{1/2}[\hat{\theta}_{\omega_1} - \hat{\theta}_{\omega_2}]}, \tag{8}$$

which can be expressed as a function of an Ω-vector,

$$T_\Omega = \frac{\sum_i \Omega_i i\xi_i}{\sqrt{\alpha_n \theta + \beta_n \theta^2}} \tag{9}$$

with

$$\Omega_i = \frac{\omega_{1i}}{\sum_j \omega_{1j}} - \frac{\omega_{2i}}{\sum_j \omega_{2j}}$$

$$\alpha_n = \sum_i i\Omega_i^2$$

$$\beta_n = \sum_i i^2\Omega_i^2\sigma_{ii} + 2\sum_i\sum_{j>i} ij\Omega_i\Omega_j\sigma_{ij}.$$

The Ω-vector results from the difference between two weight vectors normalized to 1. As a consequence, (1) all elements of the Ω-vector sum to 0 and (2) the sum of all positive values cannot be >1 and the sum of all negative values cannot be < -1. Any vector that fits these two constraints can be considered, along with Equation 9, as a neutrality test.

Without an outgroup: If no adequate outgroup is available, the unfolded frequency spectrum and consequently the $\hat{\theta}_i$ spectrum, cannot be computed. This implies that one has to use the η folded frequency spectrum. Following FU (1995), we define $\eta_i = (\xi_i + \xi_{n-i})/(1 + \delta_{i,n-i})$ and therefore we have

$$E[\eta_i] = \phi_i\theta \tag{10}$$

$$\mathrm{Var}[\eta_i] = \phi_i\theta + \rho_{ii}\theta^2 \tag{11}$$

$$\mathrm{Cov}[\eta_i, \eta_j] = \rho_{ij}\theta^2, \tag{12}$$

where $\delta_{i,n-i}$ is a Kronecker delta (1 if $i = j$, and 0 otherwise) and where

$$\phi_i = \frac{n}{(1 + \delta_{i,n-i})i(n-i)}$$

$$\rho_{ii} = \frac{\sigma_{ii} + \sigma_{(n-i)(n-i)} + 2\sigma_{i(n-i)}}{(1 + \delta_{i,n-i})^2}$$

$$\rho_{ij} = \frac{\sigma_{ij} + \sigma_{i(n-j)} + \sigma_{(n-i)j} + \sigma_{(n-i)(n-j)}}{(1 + \delta_{i,n-i})(1 + \delta_{j,n-j})}.$$

Although, we cannot compute the $\hat{\theta}_i = i\xi_i$ spectrum (as defined above), we can compute a folded $\hat{\theta}_i^*$ spectrum defined as

$$\hat{\theta}_i^* = \phi_i^{-1}\eta_i. \tag{13}$$

This folded $\hat{\theta}_i^*$ spectrum is the visual neutrality test proposed by NAWA and TAJIMA (2008). Using a similar

reasoning to that above, a linear combination of $\hat{\theta}_i^*$ leads to a generic unbiased estimator of θ defined as

$$\hat{\theta}_\omega^* = \frac{1}{\sum_i \omega_i^*} \sum_{i=1}^{n/2} \omega_i^* \phi_i^{-1}\eta_i, \tag{14}$$

whose variance is given by

$$\mathrm{Var}[\hat{\theta}_\omega^*]$$
$$= \left(\sum_i \omega_i^*\right)^{-2}\left(\theta\left(\sum_i \omega_i^{*2}\phi_i^{-1}\right)\right.$$
$$\left. + \theta^2\left(\sum_i \omega_i^{*2}\phi_i^{-2}\rho_{ii} + 2\sum_i\sum_{j>i}\phi_i^{-1}\phi_j^{-1}\omega_i^*\omega_j^*\rho_{ij}\right)\right). \tag{15}$$

Consequently, the corresponding neutrality test T_Ω^* is

$$T_\Omega^* = \frac{\sum_i \Omega_i^*\phi_i^{-1}\eta_i}{\sqrt{\alpha_n^*\theta + \beta_n^*\theta^2}} \tag{16}$$

with

$$\Omega_i^* = \frac{\omega_{1i}^*}{\sum_j \omega_{1j}^*} - \frac{\omega_{2i}^*}{\sum_j \omega_{2j}^*}$$

$$\alpha_n^* = \sum_i \phi_i^{-1}\Omega_i^{*2}$$

$$\beta_n^* = \sum_i \phi_i^{-2}\Omega_i^{*2}\rho_{ii} + 2\sum_i\sum_{j>i}\phi_i^{-1}\phi_j^{-1}\Omega_i^*\Omega_j^*\rho_{ij}.$$

It is important to mention that TAJIMA (1997) previously showed that D, F^*, and D_2^* could be expressed as a linear combination of η_i. More precisely, the vectors used then correspond in the present framework to $\sum_i \Omega_i^*\phi_i^{-1}/\sqrt{\alpha_n^*\hat{\theta}_S + \beta_n^*\hat{\theta}_S^2}$. This vector definition emphasizes the weight on each η_i rather than on each $\hat{\theta}_i^*$.

With or without an outgroup: Using both definitions of $\hat{\theta}_i$ (Equation 5) and $\hat{\theta}_i^*$ (Equation 13), it is easy to show that we have

$$\hat{\theta}_i^* = \frac{1}{n}\left((n-i)\hat{\theta}_i + i\hat{\theta}_{n-i}\right). \tag{17}$$

As a consequence, the use of an ω^*-vector along with the η folded frequency spectrum is equivalent to the use of an ω-vector with the ξ unfolded frequency spectrum only when we have

$$\frac{1}{n}\left((n-i)\omega_i^*\hat{\theta}_i + i\omega_i^*\hat{\theta}_{n-i}\right) = \frac{1}{(1 + \delta_{i,n-i})}\left(\omega_i\hat{\theta}_i + \omega_{n-i}\hat{\theta}_{n-i}\right)$$

$$\omega_i^* = \frac{n}{(n-i)(1 + \delta_{i,n-i})}\omega_i$$
$$= \frac{n}{i(1 + \delta_{i,n-i})}\omega_{n-i}. \tag{18}$$

This makes clear that there is an equivalent ω^*-vector for any ω-vector that adheres to the following constraint:

TABLE 1

Basic characteristics of previous estimators of θ

Estimators	ω	ω^* (when defined)	Variance ($n = 30$)		
			$\theta = 1$	$\theta = 10$	$\theta = 100$
$\hat{\theta}_S$	$\omega_i = i^{-1}$	$\omega_i^* = \frac{n}{i(n-i)(1+\delta_{i,n-i})}$	0.36	12.8	1052
$\hat{\theta}_\pi$	$\omega_i = n-i$	$\omega_i^* = \frac{n}{(1+\delta_{i,n-i})}$	0.59	27.4	2419
$\hat{\theta}_{\eta_1}$	$\omega_1 = (n-1),\ \omega_{n-1} = 1,\ \omega_{1<i<n-1} = 0$	$\omega_1^* = n,\ \omega_{i>1}^* = 0$	1.22	35.1	2839
$\hat{\theta}_{S-\eta_1}$	$\omega_1 = \omega_{n-1} = 0,\ \omega_{1<i<n-1} = i^{-1}$	$\omega_1^* = 0,\ \omega_{i>1}^* = \frac{n}{i(n-i)(1+\delta_{i,n-i})}$	0.52	21.4	1833
$\hat{\theta}_{\pi-\eta_1}$	$\omega_1 = \omega_{n-1} = 0,\ \omega_{1<i<n-1} = (n-i)$	$\omega_1^* = 0,\ \omega_{i>1}^* = \frac{n}{(1+\delta_{i,n-i})}$	0.68	31.9	2825
$\hat{\theta}_{\xi_1}$	$\omega_1 = 1,\ \omega_{i>1} = 0$	—	1.15	25.0	1599
$\hat{\theta}_H$	$\omega_i = i$	—	1.55	65.0	5597
$\hat{\theta}_{S-\xi_1}$	$\omega_1 = 0,\ \omega_{i>1} = i^{-1}$	—	0.51	20.3	1730
$\hat{\theta}_{\pi-\xi_1}$	$\omega_1 = 0,\ \omega_{i>1} = (n-i)$	—	0.68	31.5	2790

$$i\omega_i = (n-i)\omega_{n-i}. \qquad (19)$$

To fold the frequency spectrum, the weight $i\omega_i$ associated with ξ_i (and not with $\hat{\theta}_i$) has to be the same as the weight $(n-i)\omega_{n-i}$ associated with ξ_{n-i}. This translates into an $i\omega_i$ vector that is symmetric around $n/2$. Furthermore, when the constraint (expressed in Equation 19) is fulfilled, we can write, for any $0 \leq f \leq 1$,

$$\omega_i^* = \frac{n}{(1+\delta_{i,n-i})}\left(f\frac{\omega_i}{(n-i)} + (1-f)\frac{\omega_{n-i}}{i}\right),$$

which leads interestingly for $f = (n-i)/n$ to

$$\omega_i^* = (\omega_i + \omega_{n-i})\frac{1}{(1+\delta_{i,n-i})}. \qquad (20)$$

The weights on $\hat{\theta}_i^*$ simply result from the sums of the weights on $\hat{\theta}_i$ and on $\hat{\theta}_{n-i}$ that are pooled when the spectrum is folded. In that respect, any ω-vector complying to Equation 19 can be used without the help of an outgroup. The ω^*-vectors are then a subset of all possible values of the ω-vectors. The former can be computed from the latter by using Equation 18 or 20.

Because $\mathbf{\Omega}$ is the difference between two normalized ω-vectors, all relationships between ω and ω^* expressed above also hold for $\mathbf{\Omega}$ and $\mathbf{\Omega}^*$.

RESULTS

The model described above shows that all estimators of θ based on the frequency spectrum are linear combinations of $\hat{\theta}_i = i\xi_i$, weighted by a specific vector ω. When no outgroup is available, one can use a linear combination of $\hat{\theta}_i^* = \phi_1^{-1}\eta_i$, weighted by a vector ω^*. Consequently, neutrality tests can be expressed as a linear combination of $\hat{\theta}_i$ (or $\hat{\theta}_i^*$) weighted by a vector $\mathbf{\Omega}$ (or $\mathbf{\Omega}^*$), for which a variance can be computed easily. Three applications of the model are developed below. First, I reinvestigate the previous estimators of θ and their corresponding neutrality tests and frame their intrinsic properties in terms of the $\hat{\theta}_i$ ($\hat{\theta}_i^*$) spectrum. Then, since previous tests are only specific instances of the framework, I show how the model can be used to build new tests that are more powerful than previous ones. Finally, I exemplify the benefit of the framework on real data that are known to be subject to an ascertainment bias.

Previous θ-estimators and neutrality tests: Using Equation 6, all previously reported θ-estimators are given by an ω-vector (Table 1). When defined, the corresponding ω^*-vectors are also provided (Table 1). A graphical representation of four estimators of θ is shown in Figure 1. Figure 1 highlights that both $\hat{\theta}_S$ and $\hat{\theta}_\pi$ emphasize the low-frequency polymorphic sites in their estimation of θ (although not as much as $\hat{\theta}_{\xi_1}$, which is solely based on derived singletons) and that, on the contrary, $\hat{\theta}_H$ gives more weight to ancestral polymorphisms. Framed in the folded spectrum, $\hat{\theta}_S$ still weights more low plus high frequencies whereas $\hat{\theta}_\pi$ has a uniform weight. Potentially, using other weight vectors, one could express any undiscovered estimator of θ based on the frequency spectrum.

The numerical variances of the previous estimators of θ are reported in Table 1 (for $n = 30$ and $\theta = 1, 10, 100$). They can be computed either by their original derivations or by Equation 7. This clearly shows that, among previous estimators of θ, the variance of $\hat{\theta}_S$ is the smallest and that of $\hat{\theta}_H$ is the largest. This can be explained by the fact that the variance of $\hat{\theta}_i$ increases with i. As a consequence $\hat{\theta}_H$, which puts more weight on ancestral alleles, shows a larger variance. Interestingly, estimators without singletons have relatively small variances.

Previous neutrality tests are given in Table 2. A graphical representation of the $\mathbf{\Omega}$-vectors (and $\mathbf{\Omega}^*$ when defined) used in four previous tests is reported
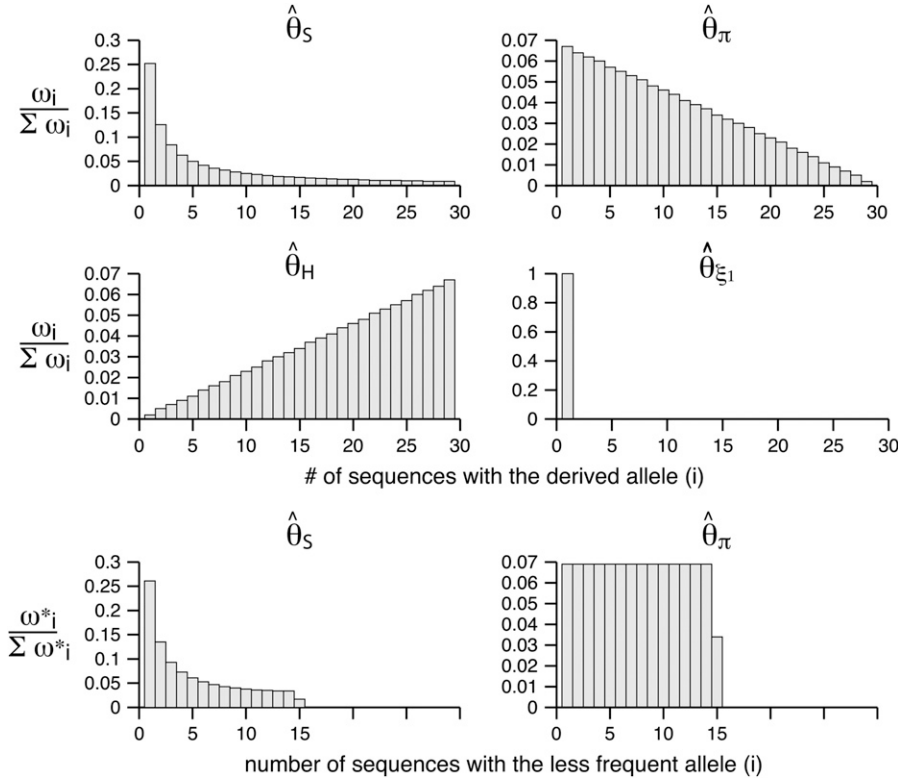
FIGURE 1.—Estimators of θ. A graphical view of the weight vectors of four typical estimators of θ (for $n = 30$). All values of the normalized vector sum to 1. In the top four panels, the **ω**-vectors that are defined for the unfolded frequency spectrum (**ξ**) are given, whereas the two bottom ones are the **ω***-vectors that are defined for the folded frequency spectrum (**η**). For estimators that can be defined in terms of both **ω** and **ω*** (here $\hat{\theta}_\pi$ and $\hat{\theta}_S$), the latter can be computed from the former with $\omega_i^* = \omega_i + \omega_{n-i}$ (when $i \neq n - i$) or $\omega_i^* = \omega_i$ (when $i = n - i$).

in Figure 2. Figure 2 shows that the sensitivity of the different tests differs although they share some common features. For example, D and F^* both are negatively sensitive to both low and high frequencies (although more sensitive to low frequencies). D shows opposite sensitivity between medium frequencies and low/high frequency, whereas F^* shows poor sensitivity to medium-frequency polymorphisms. F and F^* have opposite effects on doubletons and singletons. Thus, deviations that enhance both will have opposite effects. Finally, H is oppositely skewed by low and high frequencies.

One crucial aspect of neutrality tests is their important variance under the neutral model. This variance induces a large confidence interval and therefore

decreases their power to detect a deviation. It has been argued that this variance is a consequence of the tree shape variance and that neutrality tests based on the frequency spectrum are doomed to exhibit low power (FELSENSTEIN 1992b).

As a consequence, an ideal neutrality test should minimize its variance under the standard model. The variances of the denominator of previous neutrality tests are given in Table 2 (for $n = 30$ and $\theta = 1, 10, 100$). It is also important to mention that previous derivations of f, f^*, y, and y^* variances give different values. Simulations show that the new derivations are the correct ones (supporting information, Table S1). First, it should be noted that the original D test has a very low variance

TABLE 2

Basic characteristics of neutrality tests

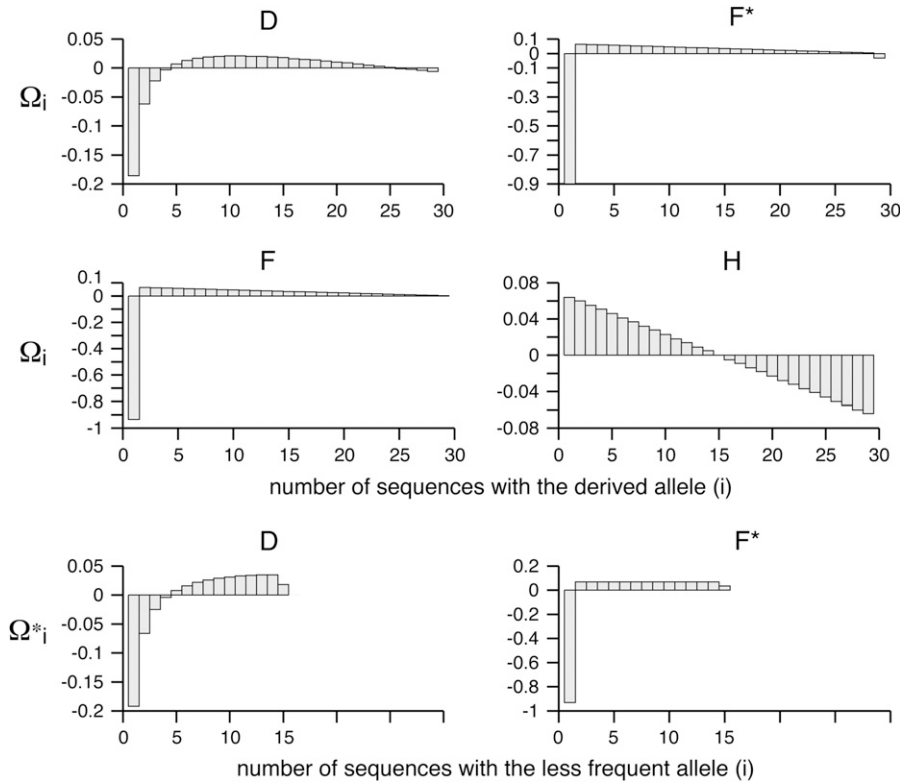| Test	$\hat{\theta}_1$	$\hat{\theta}_2$	Mandatory outgroup	Variance ($n = 30$)		
				$\theta = 1$	$\theta = 10$	$\theta = 100$
d	$\hat{\theta}_\pi$	$\hat{\theta}_S$		0.18	8.2	728
f	$\hat{\theta}_\pi$	$\hat{\theta}_{\xi_1}$	No	1.62	51.9	4,084
d_2	$\hat{\theta}_S$	$\hat{\theta}_{\xi_1}$		0.93	25.8	1,910
y	$\hat{\theta}_{\pi-\xi_1}$	$\hat{\theta}_{\pi-\xi_1}$		0.12	6.2	558
h	$\hat{\theta}_\pi$	$\hat{\theta}_H$		0.98	40.0	3,417
f^*	$\hat{\theta}_\pi$	$\hat{\theta}_{\eta_1}$	Yes	1.71	63.8	5,314
d_2^*	$\hat{\theta}_S$	$\hat{\theta}_{\eta_1}$		0.99	34.5	2,805
y^*	$\hat{\theta}_{\pi-\eta_1}$	$\hat{\theta}_{S-\eta_1}$		0.12	5.8	524
T_Ω	$\omega_{1i} = e^{-0.9i}$	$\omega_{2i} = 1$		1.19	37.1	2,895
	$\omega_{1i} = \binom{30}{i}0.5^{30}$	$\omega_{2i} = 1$	Yes	2.48	151.4	14,167

FIGURE 2.—Neutrality tests. A graphical view of the weight vectors of four typical neutrality tests (for $n = 30$). Because the $\boldsymbol{\Omega}$-vectors used for neutrality tests are computed as a difference between two normalized vectors, all values of $\boldsymbol{\Omega}$ sum to 0. In the top four panels, the $\boldsymbol{\Omega}$-vectors that are defined for the unfolded frequency spectrum ($\boldsymbol{\xi}$) are given, whereas the two bottom ones are the $\boldsymbol{\Omega}$*-vectors that are defined for the folded frequency spectrum ($\boldsymbol{\eta}$). For estimators that can be defined in terms of both $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$* (here D and F *), the latter can be computed from the former in the way that ω_i^* can be deduced from ω_i.

when compared to all other tests. This is connected to the low variance of both $\hat{\theta}_S$ and $\hat{\theta}_\pi$. Second, Y and Y* tests have also a small variance, although they ignore an important fraction of the data (*i.e.*, singletons). All other tests have a similar variance.

This predicts that D typically will be sensitive to low, medium, and high frequencies and should be more powerful because it has a relatively low variance under neutrality. Therefore, it has the potential to be an excellent neutrality test and it appears that it is often one of the most powerful tests (SIMONSEN *et al.* 1995; FU 1997). H is sensitive either low or high frequencies; however, its larger variance predicts that it will be useful only when the distortion in the θ-spectrum is very strong. In practice, it is powerful only when there is a large excess of high-frequency polymorphisms. The singleton tests appear to be good candidates to capture an excess of singletons, although they neglect other deviations in the spectrum. The Y and Y* tests have low variance, although ignoring singletons can lead to low power especially when they are in excess (ACHAZ 2008).

Building new tests: To design new neutrality tests using this framework I started by analyzing the deviation of the average $\hat{\theta}_i$ spectrum, which is expected to be uniform under the standard models. Furthermore, because FU (1995) showed that the covariance between ξ_i's is weak when compared to their variance, visual inspection of the variance of $\hat{\theta}_i$ provides a first approximation to the expected variance of $\hat{\theta}_\omega$ and therefore of their related T_ω tests. I studied two deviations from the

standard model: a severe bottleneck and isolated populations with migration.

The severe bottleneck was simulated as a sudden change of size from N chromosomes to $N/100$ that lasts for a time $T_1 = 0.1$ (in N generations). Accordingly, the coalescent rates within the bottleneck are accelerated by 0.01 and the simulations were performed as in SIMONSEN *et al.* (1995). Sampling was performed after a time T_b has elapsed after the bottleneck. The mean and the standard deviation of $\hat{\theta}_i$ are given in Figure 3a for two times, $T_b = 0.03$ and $T_b = 0.3$. Figure 3 shows that most of the deviation comes from the sites with low frequency. Therefore, I designed a new test that captures the deviations within low frequencies. In this test, I used a first vector of $\omega_{1i} = e^{-\alpha i}$, with $\alpha = 0.9$ and a second uniform vector $\omega_{2i} = 1$. This results in an exponentially decreasing weight for low-frequency mutations (Figure 3) that is positive for frequency $i/n \leq 0.13$. The choice of $\alpha = 0.9$ was mostly empirical, although using $\alpha = 0.8$ or $\alpha = 1$ leads to similar results (data not shown). As stressed in the DISCUSSION, this study aims at illustrating how easy it is to create new tests with enhanced power; power optimization deserves an entire new study. A graphical view of the $\boldsymbol{\Omega}$-vector associated with this new T_Ω test is given in Figure 3 and its variance is reported in Table 2. Most of the weight of this test is given to low frequencies and its variance is comparable to those of other neutrality tests. The power of this new test and of D, F, and H is reported in Figure 3. Results show that the new test outperforms the previous tests

a

Impact of a severe bottleneck on the θ spectrum
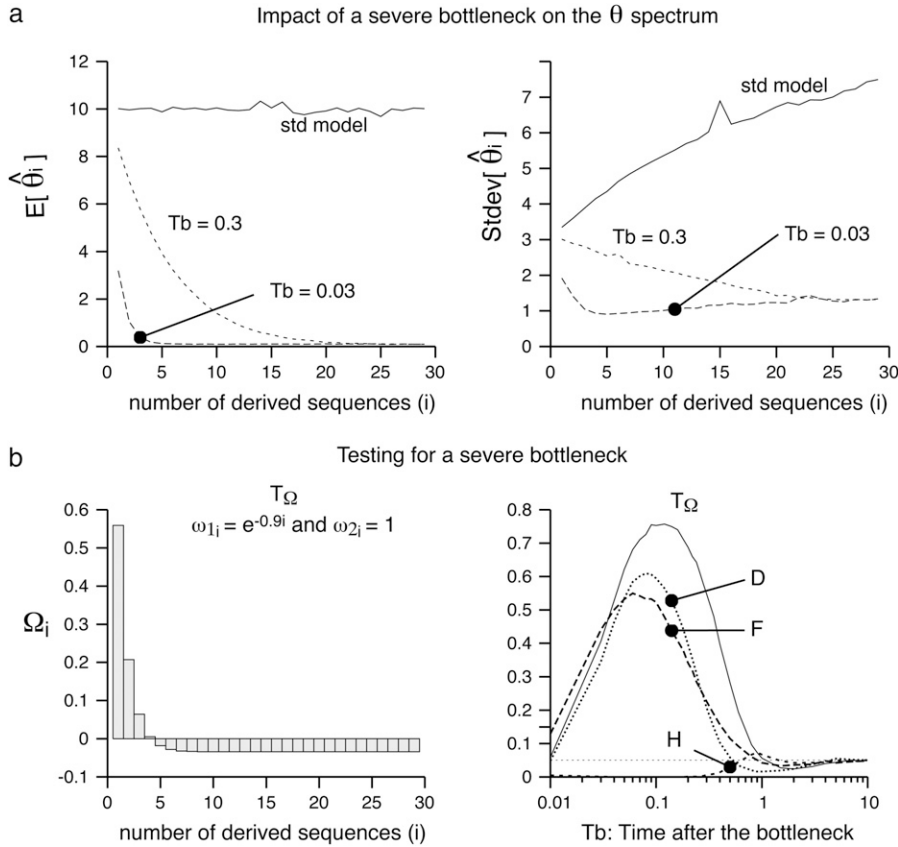

b

Testing for a severe bottleneck

FIGURE 3.—Example of a severe bottleneck. (a) The mean and the standard deviation of the $\hat{\theta}_i$ spectrum that is observed in simulations ($n = 30$, 10^4 replicates) of a standard model or of a recent severe bottleneck (reduction of $f = 1/100$ for a time $T_1 = 0.1$). In both times after the bottleneck ($T_b = 0.03$ and $T_b = 0.3$), the observed trend is similar: an excess in low frequency of $\hat{\theta}_i$, though stronger for $T_b = 0.03$. (b) Left, the weight vector of a new neutrality test (here T_Ω) is reported. It focuses its sensitivity on low frequencies: $\Omega_i = e^{-0.9i}/\sum_j e^{-0.9j} - 1/30$. (b) Right, the power of four neutrality tests is compared in detecting a severe bottleneck as a function of the time elapsed after the bottleneck. The new test shows enhanced power to detect the bottleneck (more power for a longer time).

by 20% and is able to detect the deviation for a longer time.

The 95% confidence intervals were built using coalescent simulations under the standard model, using a fixed number of segregating sites (HUDSON 1993; DEPAULIS and VEUILLE 1998). Although there has been much debate on how confidence intervals should be set (DEPAULIS et al. 2001; MARKOVTSOVA et al. 2001; WALL and HUDSON 2001), it has been clearly shown that the choice of a particular method does not alter the results in standard models (RAMOS-ONSINS et al. 2007) and therefore is not discussed here.

In the second scenario, I compared the power of neutrality tests in detecting a case of isolation with migration (e.g., NIELSEN and WAKELEY 2001). In the simulations, the isolation event happened at time $T_i = 3$ and both populations were sampled equally ($n_a = n_b = 15$). The migration rate between the two populations is variable. Similar to the analysis of the bottleneck, I first report the mean and the standard deviation of the $\hat{\theta}_i$ spectrum. Figure 4 shows that most of the deviation comes from the sites at frequency 15/30. Additionally, for a small enough migration rate ($M = 0.1$), there are almost no polymorphisms with frequency >0.5. Although the standard deviations are large, the coefficients of variations (variance/mean) are relatively small. To design a new test, I used for the first ω-vector the probabilities given by a binomial law, $\omega_{1i} = \binom{i}{n} p^i (1 -$

$p)^{n-i}$ with $p = 0.5$ and $n = 30$ and a uniform vector $\omega_{2i} = 1$ as a second vector.

This was motivated by the idea of designing a test that specifically captures an excess of medium-frequency polymorphisms. A graphical view of the resulting Ω-vector is given in Figure 4 and its variance is given in Table 2. Almost all the weight of this test is given to the $13 < i < 17$ sites. The variance of this new test is large, and this is related to the large variance of $\hat{\theta}_{n/2}$ in the sample with even n. Despite this large variance, the test clearly outperforms all previous tests (Figure 4).

Overcoming the ascertainment bias: As an example of the power of designing new neutrality tests, I analyzed SNP data (from HapMap) around the Lactase gene (LCT), which has been shown to exhibit a footprint of a recent strong selective sweep in European populations (BERSAGLIERI et al. 2004) as well in eastern African populations (TISHKOFF et al. 2007). This pattern of recent selection is one of the strongest in the human genome (NIELSEN et al. 2005). Indeed, it has been advanced that the lactase-persistence phenotype (the ability to digest milk as an adult) has been advantageous in European populations of farmers (especially in Northern European ones). The SNPs that are tightly associated with the selective sweep in Europeans are located at 13–22 kb upstream of the gene start (BERSAGLIERI et al. 2004). From HapMap (release 27, February 2009) I gathered all SNPs in a window of
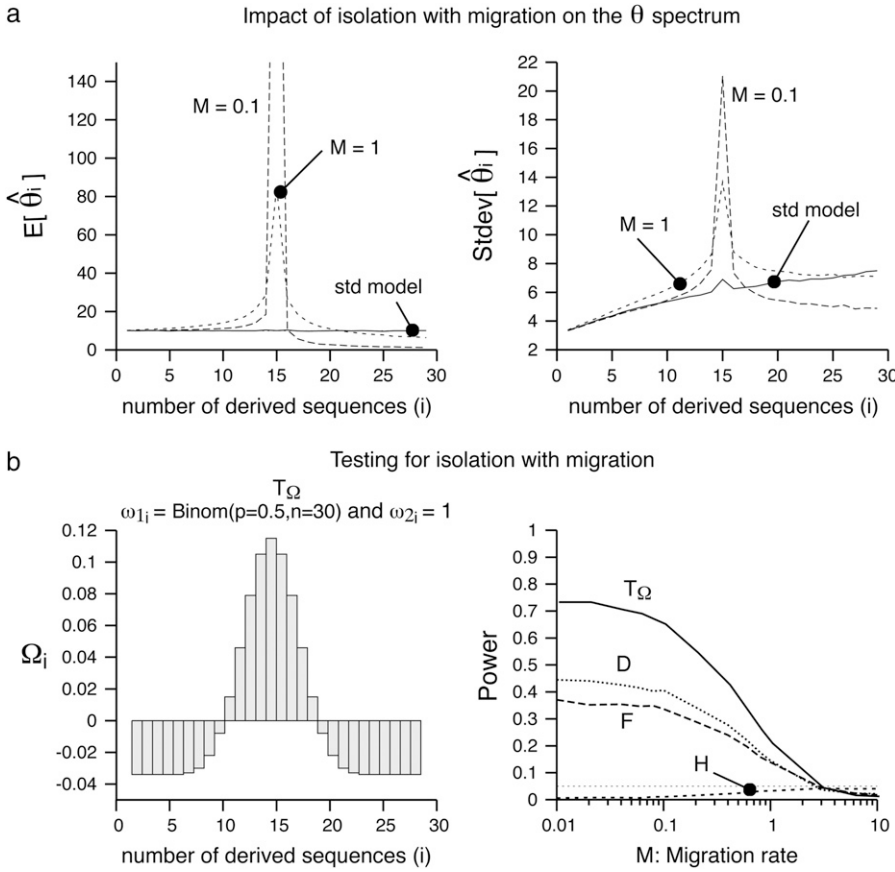
Figure 4.—Isolation with migration. (a) The mean and the standard deviation of the $\hat{\theta}_i$ spectrum that is observed in simulations ($n = 30$, 10^4 replicates) of a standard model or of an isolation with migration model (two populations equally sampled, $n_a = n_b = 15$ that were a single ancestral panmictic population at time $T_i = 3$). In both sampling migration rates between the two populations ($M = 0.1$ and $M = 1$), the observed trend is similar: an excess of $\hat{\theta}_{15}$, though much stronger for $M = 0.1$. (b) Left, the weight vector of a new neutrality test (T_ω) that focuses its sensitivity on $i = 15$. The weight vector used here is

$$\Omega_i = \binom{30}{i} 0.5^{30} / \sum_j \binom{30}{j} 0.5^{30} - 1/30,$$

where $\binom{30}{i} 0.5^{30}$ is obtained using a binomial with $p = 0.5$ and $n = 30$. (b) Right, the power of four neutrality tests is compared when detecting the population structure as a function of the migration rate. The new test displays much more power to detect the population structure.

100 kb centered at the start of the lactase gene. This includes 50 kb upstream and the entire gene. I considered only SNPs whose sample size was at least 85 chromosomes. Because the sample size of all SNPs was not identical, I used the observed frequencies to generate a folded frequency spectrum of 85 chromosomes for the following populations: Utah residents with northern and western European ancestry from the CEPH collection (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); and Yoruban in Ibadan, Nigeria (YRI).

According to the literature, one expects to find a trace of an ongoing selective event in the CEU population only. Without the help of an outgroup, this would translate into an excess of low-frequency polymorphism in the folded frequency spectrum (typically, negative D, F^*, and D_2^*). Computation of the standard neutrality tests shows a deficit of low-frequency polymorphism rather than an excess. This deficit is even often significant (Table 3). This is clearly caused by the ascertainment bias in the data set. Because the polymorphisms were first screened in a small group and further genotyped in larger groups, rare variants are underrepresented (*e.g.*, Kuhner *et al.* 2000; Clark *et al.* 2005). This ascertainment bias has been subject to various corrections (*e.g.*, Wakeley *et al.* 2001; Nielsen *et al.* 2004). To avoid any correction, I computed a T_Ω^* test where the weights of both ω_i^* and ω_2^* vectors were

set to 0 for $i < 8$. The remaining two vectors were computed using $\hat{\theta}_\pi$ and $\hat{\theta}_S$. As a consequence, this test is D-like in that it considers only polymorphisms with frequencies in the range $[0.09, 0.91]$. This is reminiscent of ignoring the singletons data set where sequencing errors are suspected (Achaz 2008). Results (Table 3) show that this test significantly deviates from the standard model for the CEU population. Ignoring fewer polymorphisms (*e.g.*, only the 5% that are of low frequency) or changing the minimum sample size leads to similar results (data not shown).

DISCUSSION

Here I developed a unifying framework for θ-estimators on the basis of the frequency spectrum. Namely, all known estimators of θ are linear combinations of $\hat{\theta}_i = i\xi_i$ (or $\hat{\theta}_i^* = \phi_i^{-1}\eta_i$). Because neutrality tests based on the frequency spectrum are simple functions of these θ-estimators, the framework can be used to derive them. All tests (of this family) proposed so far are embedded in the framework. Using the model, I have shown that estimators of θ based on a folded spectrum always have an unfolded equivalent. The reciprocal, however, is not true.

Besides its unifying appeal, the model developed here can be used in several ways. First, I showed how it can be used to compute the variance of all estimators of θ and consequently of statistics such as $t = \hat{\theta}_1 - \hat{\theta}_2$. All variances

TABLE 3

Neutrality tests in the lactase region

Population	D	D_2^*	F^*	T_{Ω^*}
CEU	−0.16	0.81	0.48	−1.79*
CHB	2.04*	1.97**	2.29*	0.25
JPT	0.92	2.22**	1.93*	0.28
YRI	1.00	2.18**	1.94*	−0.04

$*P < 0.05$, $**P < 0.01$.

of all estimators can be computed either using this framework or from their previous derived analytical formula. The same should be true for all t. Importantly, the computation of f, f^*, y, and y^* revealed differences between both methods. Simulations demonstrate that the previous formulas were not correct while the new ones are. Besides a minor error in the f and f^* variance (corrected in SIMONSEN et al. 1995), it appears that the $\text{Cov}[\pi, \xi_1]$ that was derived by FU and LI (1993b) is inexact. Therefore the variances of f and f^* (FU and LI 1993b) as well as the variances of y and y^* (ACHAZ 2008) that were using this covariance carried along the error. Framed within the model presented here, all variances are correct. Finally, it can be used to compute the variance of h that was not given by the authors (FAY and WU 2000).

One potentially interesting development is to find an ω-vector that minimizes variance of the associated estimator of θ. This problem was previously addressed thoroughly (FELSENSTEIN 1992a,b; FU and LI 1993a; FU 1994a,b). Indeed, it was shown that phylogenetic estimates have lower variance than estimators based on summary statistics (FELSENSTEIN 1992b; FU and LI 1993a; FU 1994b). Moreover, FU (1994a,b) proposed a general method to find weight vectors that minimize the variance of the estimators and showed that the best vector actually depends on the value of θ itself. Nonetheless, it remains true that some estimators have less variance than others (i.e., $\hat{\theta}_S$ vs. $\hat{\theta}_\pi$), whatever is the value of θ. This latter observation suggests that re-exploring this question of minimizing the variance may be of interest.

NAWA and TAJIMA (2008) recently proposed to use the $\hat{\theta}_i^*$ spectrum instead of the classical frequency spectrum as a visual test for neutrality. This can be extended to the unfolded $\hat{\theta}_i$ spectrum if an outgroup is available. The study presented here fully supports this idea. The visual inspection of the $\hat{\theta}_i$ spectrum indicates why some tests will reject neutrality. Contrary to what intuition may suggest, when one is interested in θ-estimation, the appropriate representation for weight vectors is the ω-vector as defined above rather than weights on the ξ_i themselves (or on the η_i as in TAJIMA 1997).

When an outgroup is used to unfold the spectrum, the choice of the appropriate outgroup is of critical importance. If the outgroup is not adequate (too distant or too close), misoriented sites will have a disastrous effect on θ-estimations and therefore on related neutrality tests (BAUDRY and DEPAULIS 2003). This adds to the difficulty of using tests based on the full ξ-spectrum. However, when low and high frequencies can be sorted apart, much power is gained in terms of choosing the adequate evolutionary scenario. For example, no high frequencies are overrepresented under recent growth or severe bottlenecks.

Specific problems that concern only some area of the spectrum can be handled easily by setting to 0 all weights in the suspicious area. For example, the sequencing errors can be avoided when the singletons are ignored (ACHAZ 2008). With the current framework, by ignoring the low-frequency polymorphisms, the ascertainment bias can be overcome and the pattern expected from selection at the lactase gene appears. This strategy has endless extensions as long as we have some prior knowledge of the suspicious area.

Finally, I think that this framework opens the door for new estimations of θ and the related neutrality tests. Using simple examples, I show how the power of neutrality tests can easily be improved to detect deviations from the standard model. To optimize the power of the future new tests, one could (1) minimize their variance under the standard model, (2) select their area of sensitivity on the basis of prior knowledge of the impact of specific deviations, and (3) use recombination estimates to compute smaller confidence intervals (WALL 1999) (because recombination results in quasi-independent replicates that lower the variance of the θ-estimators). By building specific tests that will be sensitive to specific deviations, one could envision how several selected tests will be able to help the population geneticist to choose between different possible scenarios for a given data set. Another interesting alternative would be to use the different θ-estimators as summary statistics to infer the best parameters for a given evolutionary scenario (e.g., using ABC analysis).

The source code for this study was designed as a C++ library for the simulations and a C library for sequence analysis and is available upon request. A dedicated web version of the tests is available at http://wwwabi.snv. jussieu.fr/achaz/neutralitytest.html. Furthermore, the tests will be incorporated in a future release of DNAsp.

LITERATURE CITED

ACHAZ, G., 2008 Testing for neutrality in samples with sequencing errors. Genetics **179:** 1409–1424.

BAUDRY, E., and F. DEPAULIS, 2003 Effect of misoriented sites on neutrality tests with outgroup. Genetics **165:** 1619–1622.

BERSAGLIERI, T., P. C. SABETI, N. PATTERSON, T. VANDERPLOEG, S. F. SCHAFFNER et al., 2004 Genetic signatures of strong recent pos-

itive selection at the lactase gene. Am. J. Hum. Genet. **74:** 1111–1120.

CLARK, A. G., M. J. HUBISZ, C. D. BUSTAMANTE, S. H. WILLIAMSON and R. NIELSEN, 2005 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. **15:** 1496–1502.

DEPAULIS, F., and M. VEUILLE, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788–1790.

DEPAULIS, F., S. MOUSSET and M. VEUILLE, 2001 Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol. Biol. Evol. **18:** 1136–1138.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

FELSENSTEIN, J., 1992a Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet. Res. **60:** 209–220.

FELSENSTEIN, J., 1992b Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

FU, Y. X., 1994a Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

FU, Y. X., 1994b A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

FU, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

FU, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

FU, Y. X., and W. H. LI, 1993a Maximum likelihood estimation of population parameters. Genetics **134:** 1261–1270.

FU, Y. X., and W. H. LI, 1993b Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanism of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

KUHNER, M. K., P. BEERLI, J. YAMATO and J. FELSENSTEIN, 2000 Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics **156:** 439–447.

MARKOVTSOVA, L., P. MARJORAM and S. TAVARÉ, 2001 On a test of Depaulis and Veuille. Mol. Biol. Evol. **18:** 1132–1133.

NAWA, N., and F. TAJIMA, 2008 Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. Genes Genet. Syst. **83:** 353–360.

NIELSEN, R., and J. WAKELEY, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.

NIELSEN, R., M. J. HUBISZ and A. G. CLARK, 2004 Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics **168:** 2373–2382.

NIELSEN, R., S. WILLIAMSON, Y. KIM, M. J. HUBISZ, A. G. CLARK *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566–1575.

RAMOS-ONSINS, S. E., and J. ROZAS, 2002 Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. **19:** 2092–2100.

RAMOS-ONSINS, S. E., S. MOUSSET, T. MITCHELL-OLDS and W. STEPHAN, 2007 Population genetic inference using a fixed number of segregating sites: a reassessment. Genet. Res. **89:** 231–244.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TAJIMA, F., 1997 Estimation of the amount of DNA polymorphism and statistical tests of the neutral mutation hypothesis based on DNA polymorphism, pp. 149–164 in *Progess in Population Genetics and Human Evolution*. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

TISHKOFF, S. A., F. A. REED, A. RANCIARO, B. F. VOIGHT, C. C. BABBITT *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. **39:** 31–40.

WAKELEY, J., 2009 *Coalescent Theory, an Introduction*. Roberts and Company. Greenwood Village, Colorado.

WAKELEY, J., R. NIELSEN, S. N. LIU-CORDERO and K. ARDLIE, 2001 The discovery of single-nucleotide polymorphisms–and inferences about human demographic history. Am. J. Hum. Genet. **69:** 1332–1347.

WALL, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

WALL, J. D., and R. R. HUDSON, 2001 Coalescent simulations and statistical tests of neutrality. Mol. Biol. Evol. **18:** 1134–1135.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

GENETICS

Frequency Spectrum Neutrality Tests: One for All and All for One

Guillaume Achaz

TABLE S1

Variance of neutrality tests ($\theta = 10$, $n = 30$, 10^5 replicates)

Test	Publication	Original Derivation	Current Framework	Simulation
d	TAJIMA (1989)	8.21	8.21	8.24
d_2	FU and LI (1993b)	25.83	25.83	25.89
$d_2{}^*$	FU and LI (1993b)	34.47	34.47	34.58
h	FAY and WU (2000)	na	39.96	39.55
f^*	FU and LI (1993b)	61.44	63.84	63.98
f	FU and LI (1993b)	49.59	51.85	51.83
y^*	ACHAZ (2008)	7.72	5.81	5.80
y	ACHAZ (2008)	7.40	6.18	6.17

Please note that for F^*, the typo FU and LI (1993b) equation was corrected, as it is in SIMONSEN *et al.* (1995).