DOI: 10.1534/genetics.107.082198

Testing for Neutrality in Samples With Sequencing Errors

Guillaume Achaz[1]

Systématique, Adaptation et Evolution (UMR 7138) and Atelier de Bioinformatique, Université Pierre et Marie Curie-Paris VI, 75005 Paris, France

Manuscript received September 21, 2007
Accepted for publication April 18, 2008

ABSTRACT

Many data sets one could use for population genetics contain artifactual sites, *i.e.*, sequencing errors. Here, we first explore the impact of such errors on several common summary statistics, assuming that sequencing errors are mostly singletons. We thus show that in the presence of those errors, estimators of θ can be strongly biased. We further show that even with a moderate number of sequencing errors, neutrality tests based on the frequency spectrum reject neutrality. This implies that analyses of data sets with such errors will systematically lead to wrong inferences of evolutionary scenarios. To avoid to these errors, we propose two new estimators of θ that ignore singletons as well as two new tests $Y$and Y^* that can be used to test neutrality despite sequencing errors. All in all, we show that even though singletons are ignored, these new tests show some power to detect deviations from a standard neutral model. We therefore advise the use of these new tests to strengthen conclusions in suspicious data sets.

THE mode of evolution of a population sculpts the polymorphisms that segregate in all of its homologous sequences. Under a standard neutral model (*i.e.*, a Wright–Fisher model or any other related models), an incredible amount of theory has been developed to characterize those polymorphisms. Using what is expected about the polymorphisms, population geneticists have proposed several summary statistics measuring the departure from the standard model. These statistics are used to test for the likelihood of the standard model.

Most, if not all, of the neutrality tests based on the frequency spectrum compare different estimators of $\theta = 2pN_e\mu$, where p is the ploidy (1 for haploids and 2 for diploids), N_e is the effective population size, and μ the whole-locus mutation rate. Among a sample of n homologous sequences, several estimators of θ have been derived using a coalescent framework. This includes (i) $\hat{\theta}_S = S/a_n$ that uses S, the total number of polymorphic sites, along with the correcting factor $a_n = \sum_{i=1}^{n-1} 1/i$ (Watterson 1975); (ii) $\hat{\theta}_\pi = \pi$, where π is the average pairwise difference in the sample (Tajima 1983); (iii) $\hat{\theta}_{\xi_1} = \xi_1$ that uses ξ_1, the number of derived singletons (sites at frequency $1/n$) (Fu and Li 1993); and (iv) $\hat{\theta}_{\eta_1} = \eta_1 \times (n-1)/n$ based on η_1, the total number of singletons [sites at frequency $1/n$ or $(n-1)/n$] (Fu and Li 1993).

Statistics used for neutrality tests are typically based on the difference between two of these estimators, normalized by its standard deviation. Hence, using the estimators, several statistics were proposed: $D = (\theta_\pi - \theta_S)/\sqrt{\mathrm{Var}[\theta_\pi - \theta_S]}$ (Tajima 1989) as well as four other tests that compute differences between other estimators D^*_{FL}, F^*, D_{FL}, and F (Fu and Li 1993). Fu (1996) later proposed two other statistics, G_η and G_ξ, that compute the sum of the differences between the observed and the expected numbers of polymorphic sites at each frequency. Fu (1997) defined a general framework of tests that encompasses all of these tests, among others (like the H test of Fay and Wu 2000). A comparison of the main tests (Simonsen *et al.* 1995; Fu 1997) reveals that all these statistics usually behave similarly, although some violation of the standard model induces notable differences. For example, all statistics have almost the same power to detect population growth (Fu 1997) but show differences in detecting selective sweeps (Simonsen *et al.* 1995; Fu 1997; Fay and Wu 2000). The original Tajima's D is usually one of the most powerful tests (Simonsen *et al.* 1995; Fu 1997), although it is not systematically the case (see, *e.g.*, Teshima *et al.* 2006).

Singletons are the class of polymorphisms that have the highest impact on several statistics including D. An excess or a deficit of singletons strongly skews the statistics; this deviation can lead to the rejection of the standard model. As we discuss below, sequencing errors are mostly singletons. In that respect, these errors have a strong harmful potential for population geneticists who want to infer evolutionary scenarios from data sets. Even when substantial effort is made to correct sequencing errors, singletons are typically not taken into account (Innan *et al.* 2003).

"Sequencing errors" are used here in their broadest sense. They encompass any errors that are introduced

[1]*Address for correspondence:* Atelier de Bioinformatique, Université Pierre et Marie Curie 4, place Jussieu, Boîte courrier 1202, 75005 Paris, France. E-mail: achaz@abi.snv.jussieu.fr

during the experimental procedure. Sequencing is done directly either from amplified genomic regions or from a cloned fragment.

In the former case, the sequencing is done on a pool of amplified fragments and some of them typically contain PCR errors. Actually, the errors rate of the replication enzymes used in PCR are typically high: regular taq polymerases and reverse transcriptases make $\sim 10^{-4}$ errors/bp and Pfu polymerases, which proofread the newly synthesized fragments, make $\sim 10^{-6}$ errors/bp (see, *e.g.*, the Invitrogen website). Because direct sequencing is a consensus of all amplifications, it does not cause many errors. However, it appears that new sequencing techniques (*i.e.*, pyrosequencing) exhibit an error rate that can be as high as 10^{-3} errors/bp (Wang *et al.* 2007). Whatever is the rate of errors, only increasing the coverage (the number of independent replicates) reduces the amount of sequencing errors.

Alternatively, one can sequence a clone that is a genomic fragment embedded in a plasmid. If the cloned fragment is a PCR product, it typically contains one or more sequencing errors. Indeed, even though the consensus of all amplifications does not have errors, individual sequences from amplified fragments typically contain errors. Any error that is carried by the clone itself cannot be corrected by increasing the coverage. Estimates of the error rate for cloned amplified fragments from the maize genome give 7×10^{-4} errors/bp (Eyre-Walker *et al.* 1998; Tiffin and Gaut 2001). Here also, increasing the number of independent clones of the same genomic region can be used to reduce the number of errors.

There are several reasons why only a single clone is sequenced. A first obvious reason is cost. Typically, exploratory genome projects [*e.g.*, the génolevures project (Souciet *et al.* 2000)] invest more in the number of sequenced genomes than in the coverage of each sequence. This type of analysis culminates with the metagenomic projects [*e.g.*, Sargasso Sea project (Venter *et al.* 2004)], where random chunks of DNA are cloned from the samples of the environment and subsequently sequenced. Another interesting example is when sequences from a given organism cannot be independently cloned twice. This happens either when no clonal culture is available or when the organism cannot be grown. This is the case of several microorganisms and almost all viruses. We emphasize the case of individual cloning of retroviruses (see, for example, the method develop by Palmer *et al.* 2003), where, for each individual, only a single clone is obtained by reverse PCR.

Finally, ancient DNA also contains many errors due to DNA degradation (Green *et al.* 2006). In this last case, chemical damage that accumulates in the molecule through the years makes it impossible to retrieve the original sequence at some sites.

Recently, a very interesting approach explicitly incorporated sequencing errors (among other experimental biases) into coalescent models (Knudsen and Miyamoto 2007). This new model can be used to estimate, through a full maximum-likelihood framework, population parameters as well as sequencing error rate. Although this is a very powerful approach, it remains extremely computationally intensive and will be strongly affected by recombination events. We therefore think that summary statistics estimations (moment methods) are complementary to these types of methods. In that regard, the impact of sequencing errors on $\hat{\theta}_S$ and $\hat{\theta}_\pi$ has been studied very recently by Johnson and Slatkin (2008). The authors developed a finite-site model to account for these errors on the basis of their probability of occurrence per site in a finite sequence. Complementarily, we developed here an infinite-site model (*i.e.*, all sequencing errors are new singletons) to characterize their impact in detail and avoid them without any prior knowledge of their likelihood.

Here we focus on the impact of sequencing errors on neutrality tests based on the frequency spectrum. First, we first explore how strongly sequencing errors can affect the estimators of θ. Then, we show that the D and F statistics (Tajima 1989; Fu and Li 1993) can be highly skewed by sequencing errors. Therefore, we propose new θ-estimators that will be insensitive to sequencing errors ($\hat{\theta}_{S_{-\xi_1}}$, $\hat{\theta}_{S_{-\eta_1}}$, $\hat{\theta}_{\pi_{-\eta_1}}$, and $\hat{\theta}_{\pi_{-\xi_1}}$) and two related statistics (Y^* and Y) that can be used to test neutrality despite the presence of sequencing errors. We then analyze the sensitivity of the tests on the basis of these two statistics to detect some violations of the standard model: variable population size (bottleneck), selection, and isolation (an extreme case of population structure) with and without sequencing errors.

RESULTS

Regardless of the experimental artifact leading to sequencing errors, the errors are very likely to be uniformly distributed along the sequences. As a consequence, if sequences are long enough, almost all sequencing errors will be singletons. If this is true, the singletons that we observe in a sample of sequences are a mixture of real singletons and sequencing errors. There are two types of mutations that are singletons: the ξ_1 ones at frequency $1/n$ and the ξ_{n-1} ones at frequency $(n-1)/n$. Without the help of an outgroup, these two classes will be considered a single class of mutation, η_1.

The number of sequencing errors, ε, depends on both the locus rate of errors, μ_{err} and the number of sequences, n. Actually, ε can be defined as a Poisson random variable with a parameter $n\mu_{\text{err}}$. It is interesting to see that ε and S, the number of real mutations, increase linearly with the sequence length (*i.e.*, $\mu_{\text{seq}} = L_{\text{seq}} \times \mu_{\text{site}}$). This implies that increasing the length of the sequence of interest will not alter the fraction of artifactual sites. On the contrary, S and ε do not exhibit

similar relationship with n, the number of sequences. ε increases linearly with n whereas S increases only logarithmically with n. In that respect, increasing the number of sequences in the sample will inflate the fraction of artifactual sites and then worsen the situation. It is even more dramatic when one considers only the singletons (ξ_1), whose average number does not depend on n ($E[\xi_1] = \theta$). As a result, adding new sequences will add only new artifactual singletons but no real ones.

Impact of sequencing errors on θ-estimators: We first study how sequencing errors will affect four common θ estimators: $\hat{\theta}_\pi$, $\hat{\theta}_S$, $\hat{\theta}_{\xi_1}$, and $\hat{\theta}_{\eta_1}$. Even though all of these estimators will be inflated by sequencing errors, we can expect that the errors will not equally affect all of them. In fact, each error adds a new singleton as well as a new segregating site but adds only $2/n$ to the average pairwise difference. Since both $S$and ε are independent, all covariances between real and artifactual sites are null. Using the expectations of π, S, η_1, and ξ_1 as well as their variances (Watterson 1975; Tajima 1983; Fu and Li 1993) (Equations B1–B13), we can express the mean and the variances of all biased estimators as

$$E[\hat{\theta}_\pi^{\text{err}}] = \theta + 2\mu_{\text{err}} \tag{1}$$

$$\text{Var}[\hat{\theta}_\pi^{\text{err}}] = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 + \frac{4}{n}\mu_{\text{err}} \tag{2}$$

$$E[\hat{\theta}_S^{\text{err}}] = \theta + \frac{n}{a_n}\mu_{\text{err}} \tag{3}$$

$$\text{Var}[\hat{\theta}_S^{\text{err}}] = \frac{\theta}{a_n} + \frac{b_n}{a_n^2}\theta^2 + \frac{n}{a_n^2}\mu_{\text{err}} \tag{4}$$

$$E[\hat{\theta}_{\xi_1}^{\text{err}}] = \theta + n\mu_{\text{err}} \tag{5}$$

$$\text{Var}[\hat{\theta}_{\xi_1}^{\text{err}}] = \theta + \left(a_n\frac{2n}{(n-1)(n-2)} - \frac{4}{(n-2)}\right)\theta^2 + n\mu_{\text{err}} \tag{6}$$

$$E[\hat{\theta}_{\eta_1}^{\text{err}}] = \theta + (n-1)\mu_{\text{err}} \tag{7}$$

$$\text{Var}[\hat{\theta}_{\eta_1}^{\text{err}}] = \frac{(n-1)}{n}\theta + \left(a_n\frac{2(n-1)}{n^2} - \frac{1}{n^2}\right)\theta^2 + \frac{(n-1)^2}{n}\mu_{\text{err}}. \tag{8}$$

The expectations are in good agreement with the intuition; the effect is the strongest on θ_{ξ_1} and the weakest on θ_π. It is noteworthy to mention that for $n > 4$, the bias is stronger for θ_S than for θ_π. It also shows that increasing the number of sequences will inflate the bias for all estimators except for θ_π. The examination of the variances also shows some interesting properties. As expected from intuition, adding sequencing errors inflates all variances. In the original variances (set μ_{err} to 0), only the variance of $\hat{\theta}_S$ vanishes as n increases, and this is the reason why θ is preferentially estimated from S. The other variances typically converge to a constant when n increases. Adding sequencing errors drastically changes this pattern. Indeed, the relationship between the variances and n becomes linear for $\hat{\theta}_{\xi_1}$ and $\hat{\theta}_{\eta_1}$, sublinear for $\hat{\theta}_S$, and converges to a constant for $\hat{\theta}_\pi$. With a moderate rate of sequencing error (*i.e.*, $\mu_{\text{err}} = 0.1$) and a typical sample size ($n \leq 100$), $\text{Var}[\hat{\theta}_S] > \text{Var}[\hat{\theta}_\pi]$ for small θ-values (*i.e.*, $\theta \leq 2$).

Therefore, in the presence of nonnegligible sequencing errors, estimations of θ should be carefully performed. Since both the mean and the variance of $\hat{\theta}_\pi$ are less affected by errors, their use is less inadequate than the use of $\hat{\theta}_S$, especially when the sample size gets large or when $\theta \leq 1$. This is in good agreement with the predictions of a finite-site model (Johnson and Slatkin 2008). However, all these estimators of θ are biased; this consequently motivates the derivation of new estimators that are immune to sequencing errors.

New θ-estimators: The simplest way to correct for sequencing errors would be to just ignore some of the observed singletons in the sequences. However, this assumes that we are able to estimate properly the number of artifactual singletons, which is unrealistic. Therefore, we derive new estimators that do not make use of the singletons to estimate θ. Although there is no way to compute a revised θ_{η_1} and θ_{ξ_1}, since they are solely based on singletons, one can compute new estimators derived from both the number of segregating sites and the average pairwise differences when singletons are ignored. Depending on the availability of an outgroup to orientate the mutations, these estimators are defined as $S_{-\xi_1}$ and $\pi_{-\xi_1}$ (with outgroup) or $S_{-\eta_1}$ and $\pi_{-\eta_1}$ (no outgroup). The means of these values were derived by removing the expected numbers of singletons in a sample. The expected number of singletons is given by either $E[\xi_1]$ or $E[\eta_1] = E[\xi_1] + E[\xi_{n-1}]$. These two values are respectively equal to $E[\xi_1] = \theta$ and $E[\eta_1] = (n/(n-1))\theta$ (Fu and Li 1993). It is important to point out that a singleton weighs $2/n$ on π. As a consequence, the expectations of these values are

$$E[S_{-\eta_1}] = \theta \times \left(a_n - \frac{n}{(n-1)}\right) \tag{9}$$

$$E[S_{-\xi_1}] = \theta \times (a_n - 1) \tag{10}$$

$$E[\pi_{-\eta_1}] = \theta \times \frac{(n-3)}{(n-1)} \tag{11}$$

$$E[\pi_{-\xi_1}] = \theta \times \frac{(n-2)}{n}. \tag{12}$$

Corresponding variances are presented in appendix b (see Equations B22, B23, B31, and B32). We used the

mean values to derive unbiased estimators of θ that should be insensitive to sequencing errors:

$$\hat{\theta}_{S_{-\eta_1}} = \frac{S_{-\eta_1}}{a_n - n/(n-1)} \tag{13}$$

$$\hat{\theta}_{S_{-\xi_1}} = \frac{S_{-\xi_1}}{(a_n - 1)} \tag{14}$$

$$\hat{\theta}_{\pi_{-\eta_1}} = \pi_{-\eta_1} \times \frac{(n-1)}{(n-3)} \tag{15}$$

$$\hat{\theta}_{\pi_{-\xi_1}} = \pi_{-\xi_1} \times \frac{n}{(n-2)}. \tag{16}$$

Expected impact of sequencing errors on *D* and *F*: Since the overestimation on θ is not the same on all estimators, it has to be true that all neutrality tests based on a difference between $\hat{\theta}_S$, $\hat{\theta}_\pi$, $\hat{\theta}_{\xi_1}$, and $\hat{\theta}_{\eta_1}$ will be biased. Here, we focus on how tests based on *D* (Tajima 1989) and *F* (Fu and Li 1993) are affected by sequencing errors. Please note that all tests that use $\hat{\theta}_{\eta_1}$ or $\hat{\theta}_{\xi_1}$ (*i.e.*, all Fu and Li 1993 ones) behave very similarly to the *F* test. Therefore, for clarity purposes, we choose to present results for *F* only but all our conclusions apply similarly to the other tests.

Introducing sequencing errors, *D* and *F* become

$$D_{\text{err}} = \frac{\pi_{\text{err}} - S_{\text{err}}/a_n}{\sqrt{e_1 S_{\text{err}} + e_2 S_{\text{err}}(S_{\text{err}} - 1)}} \tag{17}$$

$$F_{\text{err}} = \frac{\pi_{\text{err}} - \eta_{1\text{err}}}{\sqrt{\upsilon_F S_{\text{err}} + \nu_F S_{\text{err}}^2}}, \tag{18}$$

where the constants e_1 and e_2 are defined in Tajima (1989) and υ_F and ν_F in Fu and Li (1993).

Even though the whole *D* probability distribution is difficult to derive, one can derive an approximation of the average D_{err} as

$$\begin{aligned} E[D_{\text{err}}] &\approx \frac{E[\pi_{\text{err}}] - E[S_{\text{err}}]/a_n}{(e_1 - e_2)E[S_{\text{err}}] + e_2 E[S_{\text{err}}]} \\ &= \frac{E[\pi] - E[S]/a_n + \mu_{\text{err}}(2 - n/a_n)}{(e_1 - e_2)(E[S] + n\mu_{\text{err}}) + e_2(\text{Var}[S_{\text{err}}] + E[S_{\text{err}}]^2)} \\ &= \frac{\mu_{\text{err}}(2 - n/a_n)}{((e_1 - e_2)E[S] + e_2 E[S^2]) + n\mu_{\text{err}}(e_1 + e_2(2a_n\theta + n\mu_{\text{err}}))} \\ &= \frac{\mu_{\text{err}}(2 - n/a_n)}{\text{Var}[d] + n\mu_{\text{err}}(e_1 + e_2(2a_n\theta + n\mu_{\text{err}}))}. \end{aligned} \tag{19}$$

Similarly, one can show that:

$$E[F_{\text{err}}] \approx \frac{\mu_{\text{err}}(2 - n)}{\text{Var}[f] + n\mu_{\text{err}}(\upsilon_F + \nu_F(1 + 2a_n\theta + n\mu_{\text{err}}))}. \tag{20}$$

As expected, sequencing errors cause negative values in D_{err} and F_{err}. Indeed, the numerator of Equation 19 is always negative (or null when $n = 2$) and even more dramatically so in Equation 20. From the comparison of the numerators of Equations 20 and 19, we can suspect that the impact will be stronger on F_{err} than on D_{err}. This negative bias is increased by both the sequencing error rate, μ_{err}, and the number of sampled sequences, *n*. Interestingly enough, the estimated variances in the denominators are larger with sequencing errors, which fits the intuition, since adding another random variable (the number of sequencing errors) overall will inflate the variance. The left terms are the usual variances whereas the right terms are the expected increase, which grow with *n* and μ_{err}. Increasing the denominator tends to diminish the magnitude of the bias on *D* and *F* (without changing their signs).

Simulated impact of sequencing errors on *D* and *F*: To assess more precisely the impact of sequencing errors on *D* and *F*, we ran simulations of a standard model with extra singletons that were added to the resulting sequences. All coalescent simulations were performed using a standard algorithm that depends only on θ and *n*. A detailed description of such algorithms is given in reviews or books such as Hudson (1990) and Hein *et al.* (2005). Here, all sequencing errors are assumed to be singletons. Therefore to simulate data with sequencing errors, we simulate a regular coalescent tree with "true" mutations and add some "artifactual" singletons whose number ε is computed as a Poisson random variable with parameter $\mu_{\text{err}} \times n$. Errors are distributed uniformly among all sequences.

In this study, we analyze the power of several statistics (*i.e.*, *D*, *F*, *Y**, and *Y*) to detect a departure from a standard model. We use here the largest and most robust confidence interval that is constituted by the most extreme values of the statistics for a given *n* but for $\theta \in [0, \infty[$. We present in appendix a a method to find the confidence-interval limits and discuss other possible methodologies.

Results (Figure 1) show that even with a moderate rate of sequencing errors, the distributions of D_{err} and F_{err} are skewed toward significant negative values. More precisely, when the μ_{err} is in the vicinity of $[\theta/100, \theta/10]$, the D_{err} and the F_{err} distributions tend to be significantly negative. In other words, if there is 1 sequencing error for ~10–100 singletons (or pairwise differences) in the sample ($E[\xi_1] = E[\pi] = \theta$), D_{err} and F_{err} are always "too" negative. As expected, this effect gets stronger when *n* becomes large. Because the test based on *F* directly uses the number of singletons, the bias is stronger for F_{err} than for D_{err}.

This shows that what really matters for the impact of sequencing error is the natural diversity (*i.e.*, θ) of the population and the sample size, *n*. Even with a low rate of sequencing error, results from large data sets (*i.e.*, $n >$ 100) or from species with low diversity can be strongly affected by sequencing errors and therefore should be interpreted with caution. Thus, we derive two new tests that are immune to sequencing errors.

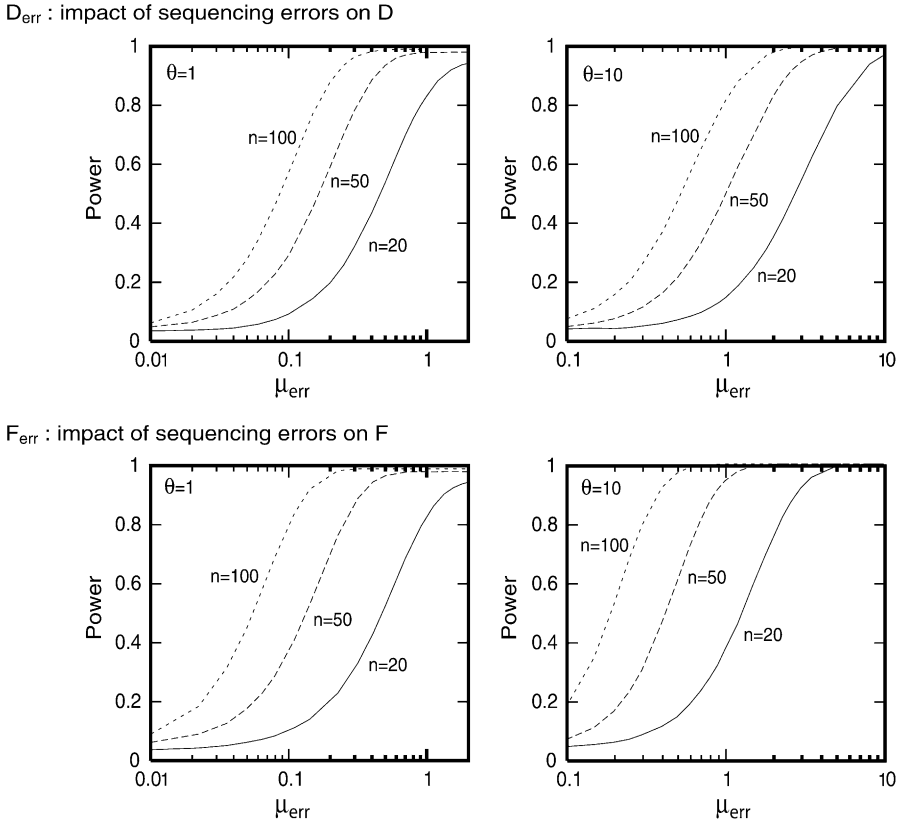

FIGURE 1.—Strong impact of sequencing errors on D and F, noted as D_{err} and F_{err}, when $\mu_{err} > 0$. All 10^5 simulations were performed with $n = 20$, 50, 100, with $\theta = 1$ (left) or $\theta = 10$ (right) and a variable μ_{err}. This rate of sequencing errors is defined for one sequence and for the whole locus, so that the number of errors is given by a Poisson law with mean $n\mu_{err}$. Sequencing errors artificially steer the statistics to negatives values. We report the power of the tests to reject the standard model. This shows that even when the sequencing error rate is moderate (0.01–0.1 of θ), the effect can be strong (especially when n is large).

Derivation of *Y and *Y*:** As in the standard D from Tajima, we define here two new statistics Y and Y^* to be a difference between θ-estimators. Following FU and LI's (1993) notation, Y requires the use of an outgroup whereas Y^* does not. As we have shown, θ can be estimated either from the number of nonsingleton segregating sites ($S_{-\eta_1}$ or $S_{-\xi_1}$) or from the average pairwise differences, excluding singletons ($\pi_{-\eta_1}$ or $\pi_{-\xi_1}$). Therefore, we define Y and Y^* as

$$Y = \frac{\pi_{-\xi_1} - fS_{-\xi_1}}{\sqrt{\mathrm{Var}[\pi_{-\xi_1} - fS_{-\xi_1}]}}, \quad \text{where} \quad f = \frac{(n-2)}{(n(a_n - 1))} \tag{21}$$

and

$$Y^* = \frac{\pi_{-\eta_1} - f^* S_{-\eta_1}}{\sqrt{\mathrm{Var}[\pi_{-\eta_1} - f^* S_{-\eta_1}]}}, \quad \text{where} \quad f^* = \frac{(n-3)}{(a_n(n-1) - n)}. \tag{21}$$

The details of the derivations for both variances in the denominators are presented in APPENDIX B. They can be expressed as $\mathrm{Var}[\pi_{-\xi_1} - fS_{-\xi_1}] = \alpha_n\theta - \beta_n\theta^2$ and $\mathrm{Var}[\pi_{-\eta_1} - f^*S_{-\eta_1}] = \alpha_n^*\theta - \beta_n^*\theta^2$, where α_n, β_n, α_n^*, and β_n^* are constants that depend on n only. As for regular D, we show that θ and θ^2 are unknown but can be estimated from $S_{-\eta_1}$ or from $S_{-\xi_1}$. Please refer to APPENDIX B for more details.

Importantly, these new statistics are totally immune to sequencing errors provided that these errors are singletons.

Violation of the standard model: We choose to explore the relative power of D, F, Y^*, and Y to detect violation of the standard model in three different scenarios: bottleneck, hitchhiking along with a selective sweep, and isolation. In all three scenarios, we used dedicated coalescent simulations for both $n = 20$ and $n = 50$ and used $\theta = 10$ with either no sequencing errors ($\mu_{err} = 0$) or a low rate of sequencing error ($\mu_{err} = 0.1$).

It is important to keep in mind that it is always possible to increase θ for the locus by increasing sequence length. This will result in an improvement of the power of the tests. Actually, more mutations (larger θ) in the tree make it easier to recover its shape from polymorphisms and, therefore, to unravel its nonneutral distortions. This effect of θ is especially important when we consider cases where the tree is reduced overall.

As a consequence, we do not discuss the absolute power of the tests based on Y and Y^* but rather their relative power compared to the one based on D or F. Since we disregard some of the information carried by the sequences (*i.e.*, the singletons), we expect tests based on Y^* and Y to have less power when compared to the tests based on D and F. Importantly, adding sequencing errors will not affect the tests based on Y and Y^*, but will change both D and F into D_{err} and F_{err}. It is

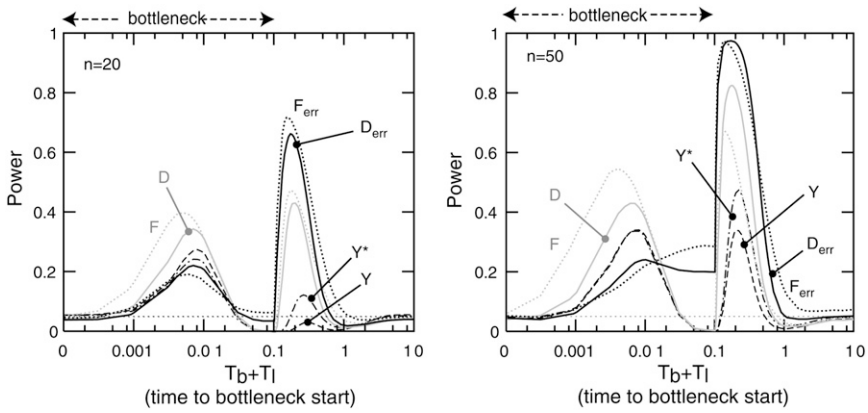

FIGURE 2.—Power of all tests to detect population expansion or decline. All 10^5 simulations were performed with $\theta = 10$ and $n = 20$ (left) or $n = 50$ (right). The sequencing error rate was set either to $\mu_{err} = 0$ (D and F) or to $\mu_{err} = 0.1$ (D_{err} and F_{err}). We report the power of all tests as a function of the total time since the bottleneck started. Bottlenecks are characterized by two times: T_l, the length of the bottleneck, and T_b, the time after the bottleneck. Here, the population size reduction is 1/100th and lasts for at most $T_l = 0.1$. A sample can be taken during the bottleneck ($T_l + T_b \leq 0.1$) or after it has ended ($T_l + T_b > 0.1$). The graphs illustrate that depending on the how the frequency spectrum is skewed, the new tests performed either poorly (*i.e.*, excess of low frequency: star-like trees) or honorably (*i.e.*, excess of medium frequency: trees with stretched internal branches). They also illustrate that sequencing errors mask an excess of medium frequency and artificially enhance an excess of low frequency.

important to note that, with the parameters we used ($\theta = 10$ and $\mu_{err} = 0.1$), the power of the test based on D is not affected by sequencing errors only and the test based on F is only weakly affected (Figure 1). We report the power of the tests with and without sequencing errors in all scenarios.

Change in population size: We considered a population that experiences a severe bottleneck. The simulations were performed using a change in timescale (GRIFFITHS and TAVARÉ 1994). The population has a regular size N, but shrinks at size $N/100$ during the bottleneck that goes for a time T_l (at most $T_l = 0.1$). Some time T_b can have elapsed after the bottleneck ended. Looking at this process in reverse, any coalescent time that falls between T_b and $T_b + T_l/f$ needs to be shortened adequately (SIMONSEN *et al.* 1995). Here we consider the cases where the sample is taken during the bottleneck ($T_b = 0$; $T_l \leq 0.1$) or after the bottleneck ($T_b > 0$; $T_l = 0.1$).

Results from simulations (Figure 2) show that deviations from the standard model are observed whenever the sample is taken either close to the beginning of the bottleneck ($0 < T_l < 0.01$ and $T_b = 0$) or just when it finishes ($T_l = 0.1$ and $0 < T_b < 1$). On one hand, if the sample is taken after the bottleneck (*i.e.*, population expansion), the tree will have internal branches that are "too short," and therefore all statistics will exhibit negative values. On the other hand, if the sample is taken during the bottleneck (*i.e.*, population decline), the resulting coalescent tree will be shorter and with internal branches that are "too long"; this will lead to an excess of medium-frequency polymorphisms and therefore to positive statistics. In both cases, the signal lasts for N (or $N/100$) generations. Whereas the new test performs poorly when detecting an excess of low-frequency polymorphisms, it performs well when detecting an excess of medium-frequency polymorphisms. Adding sequencing errors reduces the power to detect an excess of medium-frequency polymorphisms but enhances the ability to detect an excess of low-frequency polymorphisms. This highlights that the bias induced by sequencing errors tends to lower the positive deviations of the statistics and enhance the negative ones. It should also be noted that the loss of power to detect positive deviations is stronger for F than for D. This is in good agreement with a stronger impact of sequencing errors on F as shown above.

Hitchhiking along with a selective sweep: Another interesting situation is when the locus of interest is neutral but linked to a nearby locus that experienced a selective sweep. This effect has long been referred as the hitchhiking effect (MAYNARD SMITH and HAIGH 1974; STEPHAN *et al.* 1992; KIM and STEPHAN 2002). We used the simplified version of the model proposed by BRAVERMAN *et al.* (1995), described in FAY and WU (2000). From the end of the sweep to its beginning, the frequency of the selected allele decreases deterministically from $1 - 10^{-4}$ to 10^{-4} with a selection coefficient equal to $\alpha = 2Ns = 1000$. We started the simulation at $T_s = 0.01$ after the sweep end (*i.e.*, a very recent sweep). In Figure 3 we report results for a variable range of recombination rates ($R = 4Nc$) expressed as the ratio between recombination and selection (c/s).

When the c/s ratio is very small, the resulting star tree has almost only singletons. In the case of $n = 20$, it has only a few singletons (*i.e.*, no power for any tests) whereas it has several when $n = 50$. In this case, the tests based on Y and Y^* have no power to detect the deviation. Importantly, adding sequencing errors greatly increases the excess of singletons and therefore artificially enhances the power of the test based on D_{err} and F_{err} at a low c/s ratio. When the ratio is in the vicinity of 0.01, an important fraction of the trees shows only 1 lineage that has escaped the sweep through recombination. This lineage will have ancestral singletons (ξ_{n-1}). In this case, the test based on Y performs well whereas the test based on Y^* performs poorly. When the ratio becomes high, most of the lineages escape the sweep through recombination and the process is assimilated

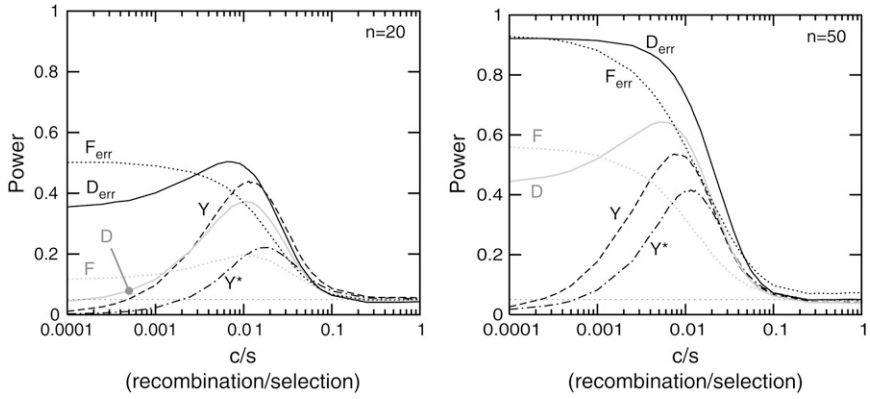

FIGURE 3.—Impact of a selective sweep near the neutral locus under study. All 10^5 simulations were performed with $\theta = 10$, $\alpha = 2Ns = 1000$, $T_s = 0.001$ (the sample was taken right after the sweep) and $n = 20$ (left) or $n = 50$ (right). The sequencing error rate was set either to $\mu_{err} = 0$ (D and F) or to $\mu_{err} = 0.1$ (D_{err} and F_{err}). Power is given as a function of c/s, the ratio between recombination and selection coefficients (although s is fixed). The graphs illustrate that when the recombination rate is very small, the new statistics have little or no power to detect a deviation, but, when the ratio c/s is in the order of 1/100, the new test with outgroup (based on Y) performs well. Finally, when it is very large, there is no more deviation from the standard model.

to a standard process; there is no more deviation to be detected besides a residual bias due to sequencing errors on F.

Isolated populations: An extreme case of population structure is isolated subpopulations, *i.e.*, a population complete split. In a simple model (SIMONSEN *et al.* 1995), an isolation event happened at some time T_i in the past, after which both populations (size $N/2$) did not mix anymore. Before T_i, the ancestral population (size N) is panmictic and simply follows a standard model. We analyze two types of sampling: (1) both populations have been sampled with equal size and (2) one population is largely underrepresented in the sample. As we will see, the tests behave differently in each situation.

Results for the case of equal sample size (Figure 4a) show that all tests have almost the same power to detect a violation from the standard expectations. With $T_i \geq 1$, all distributions are skewed on high values in a very similar fashion. Here, there is a long internal branch that splits the sample in two subtrees with an equal number of lineages. Since singletons do not have any particular role, all tests perform identically when there are no sequencing errors. Adding sequencing errors lowers the power of tests based on D and F since it reduces their positive deviations.

On the contrary, results for samples of unequal sizes (Figure 4b) show that the new tests based on Y and Y^* perform better than the original test based on D and F. Despite the difference of power, all distributions are skewed toward negative values. Having unequal sample sizes creates a stretch of the branch that splits the tree into two subtrees, one with few individuals and the other one with all other individuals. This stretch induces an excess of both high- and low-frequency polymorphisms.

DISCUSSION

Sequencing errors are very often encountered in data sets used for population genetics inferences. As a result, geneticists can be misguided by the effect of sequencing errors that artificially increase the number of singletons. This overabundance of singletons leads to overestimations of θ, which is typically used to compute the effective population size. We show that, even with a moderate rate of errors, the overestimation can be significant. As a consequence, we propose new estimators of θ that are insensitive to sequencing errors provided that these errors are singletons. We therefore highly recommend estimating θ with these new estimators in suspicious data sets.

Another very important consequence of the presence of sequencing errors is their effects on neutrality tests. Indeed, singletons are usually the class of polymorphisms that has the greatest impact on neutrality tests. It is mainly the excess or the deficiency of singletons that steers the statistics outside of their confidence intervals. We have shown (Figure 1) that an artifactual excess of singletons will easily alter the rejection of the standard model. More precisely, it will enhance an excess of singletons caused by some evolutionary scenarios (*e.g.*, population expansion or selective sweep) or mask any deficiency of singletons caused by other scenarios (*e.g.*, population decline or population isolation with equal sampling). As a result, it becomes hard to distinguish the effect of sequencing errors from the ones due to biological departures of the standard model. Tests that use the θ-estimators based on singletons (*i.e.*, the FU and LI 1993 ones) are the most affected tests.

We show here that the tests based on D and F are skewed even when the error rate is low. Namely, if the rate of artifactual singletons per sequence represents $\sim$0.01–0.1 of θ (the number of singletons or the average pairwise difference), D and F are often significantly negative. Furthermore, a low rate of sequencing error (0.01 of θ) can alter strongly the power of the tests when the scenario is not a standard neutral one. Using the average pairwise difference, we can calculate the error rate above which a data set becomes suspicious. In the human population, we have $\theta \approx 0.001$/site

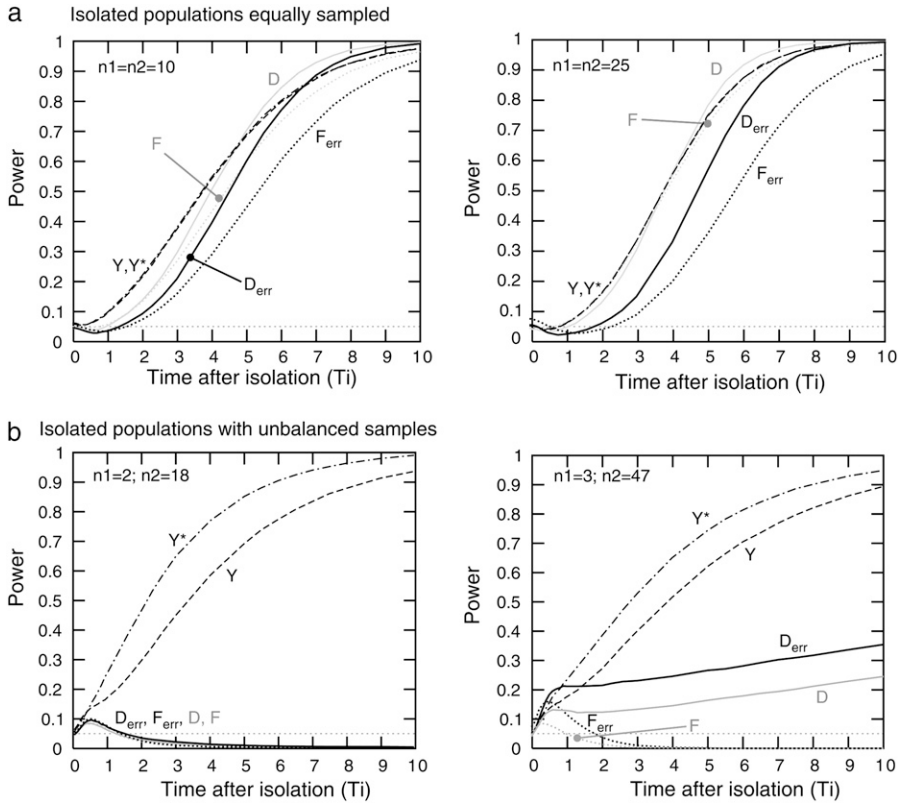

FIGURE 4.—Power of all tests to detect deviation due to isolation (an extreme case of population structure). All 10^5 simulations were performed with $\theta = 10$, $N_1 = N_2 = N_{anc}/2$, and $n = 20$ (left) or $n = 50$ (right). The sequencing error rate was set either to $\mu_{err} = 0$ (D and F) or to $\mu_{err} = 0.1$ (D_{err} and F_{err}). (a) Deviation when the sampling is equilibrated between the populations ($n_1 = n_2 = 10$ or $n_1 = n_2 = 25$), as a function of the time to the isolation event (T_i). All tests exhibit very similar power to detect deviation from a standard model. (b) The sampling scheme is very unbalanced ($n_1 = 2$, $n_2 = 18$ or $n_1 = 3$, $n_2 = 47$). Interestingly, the test based on Y or Y^* exhibits a stronger power than the one based on D.

(SACHIDANANDAM *et al.* 2001); this translates into a rate of $\sim 10^{-5}$–10^{-4} errors/bp. In a population of HIV-1 infecting a patient, we observe $\theta \approx 0.01$/site (ACHAZ *et al.* 2004); this leads to an error rate of 10^{-4}–10^{-3} errors/bp. When these rates are compared to the ones observed either for pyrosequencing techniques [up to 10^{-3} errors/bp (WANG *et al.* 2007)] or for cloned PCR products [7×10^{-4} errors/bp (EYRE-WALKER *et al.* 1998; TIFFIN and GAUT 2001)], we have to acknowledge that sequencing errors are a major issue in an uncured data set. Consequently, only data sets where sequencing errors have been carefully removed should be analyzed by standard neutrality tests. The step to remove sequencing errors usually requires independent cloning and/or sequencing, which is highly time and resource consuming. In this context, we think our new tests can be very useful to many population geneticists, since they offer a safe and quick way to test for neutrality despite the presence of sequencing errors.

If there are no sequencing errors, we expect tests based on Y and Y^* to show less power than regular tests. However, more than a radical loss of power, ignoring part of the suspicious data leads to a shift in the sensitivity of the tests. We show that in any situation where the height of the tree is greatly reduced and the tree shape converges to a star tree (*i.e.*, population expansion and selective sweep with no recombination), the new tests show less power to detect departure from the standard model. This fits perfectly with intuition since the signal lies on derived singletons, which here are completely ignored. On the contrary, we show that for scenarios where the internal branches of the tree are stretched (*i.e.*, population decline and isolated populations with an equilibrated sample) all the tests behave very similarly. In this case, most of the segregating sites will be at intermediate frequency and all tests are more or less equally able to detect departures from a standard model. Finally, the new tests can outperform the original test based on D in two situations: (i) when there is an excess of high-frequency polymorphisms (hitchhiking along with a sweep—especially Y—and isolated populations with an unbalanced sample) and (ii) when there is an excess of low-frequency, though nonsingleton, polymorphisms (isolated populations with an unbalanced sample).

We have assumed throughout this study that the sequencing errors were singletons. If these errors are genuinely distributed uniformly along long enough sequences, it is very likely that our assumption will hold. It is, however, known that some errors are more commonly encountered than others. For example, microsatellites of mononucleotides are often increased or decreased by 1 unit during the cloning/sequencing step. These small indels are, however, not a problem since they can be easily corrected manually. On the other hand, if the sequencing errors rate depends on the nucleotide context, *i.e.*, some sites being more mutated than others, this can create nonsingleton

sequencing errors that will also affect tests based on Y and Y^*. In any case, if there are some nonsingleton sequencing errors, our tests will be much less sensitive to those few rare events than the original D is to sequencing errors in general. We note that the equivalent model with finite sites (Johnson and Slatkin 2008) is more appropriate when the density of polymorphic sites is high and/or when the sequences are small. This model, however, requires a prior knowledge of the error rate to handle them properly.

We ignored the possibility of recombination events within the locus under study. It has been reported, however, that recombination tends to decrease the variance of both θ-estimators $\hat{\theta}_S$ and $\hat{\theta}_\pi$ (Hudson 1983). As a consequence it decreases the variance of D and therefore, if the same confidence interval is kept, there is an important loss of power (Wall 1999). We expect that the same will happen with tests based on Y and Y^*. Here, we decided to keep the most conservative confidence interval for the test (using $\theta \in [0, \infty[$ and $R = 0$). However, one can define a confidence interval for a different range of parameters that can be calculated from the data. Using this last strategy, we should be able to recover most of the power loss.

To conclude, sequencing errors can easily misguide interpretations of the data. In particular, they can make regular statistics (D, F among others) significantly negative, a signal that is usually interpreted as a star-like tree. In this case, if there is no departure from a standard model despite sequencing errors, tests based on Y and Y^* could be helpful to avoid spurious interpretation of the data. However, because the new tests do not have much power to detect star-like trees, an absence of a significant result should not systematically mean that the standard model is correct. We can already envision that new tests inspired from the Fu and Li (1993) ones using θ-estimators based on polymorphisms at frequencies $2/n$ and $n - 2/n$ should perform better than Y and Y^* for detecting star-like trees.

The source code was designed as a C++ library and is available upon request. I thank J. Wakeley, who motivated this study, for his scientific generosity and his priceless advice. I also thank F. Depaulis, M. Tenaillon, P. Nicolas, and D. Higuet for giving me constructive comments on the manuscript; N. Bierne and O. Tenaillon for suggesting the name of the new statistics; and T. Treangen for improving the English of this manuscript. Finally, I thank the three anonymous reviewers for their constructive suggestions that improved the manuscript.

LITERATURE CITED

Achaz, G., S. Palmer, M. Kearney, F. Maldarelli, J. W. Mellors *et al.*, 2004 A robust measure of HIV-1 population turnover within chronically infected individuals. Mol. Biol. Evol. **21:** 1902–1912.

Berger, R., and D. Boos, 1994 P values maximized over a confidence set for the nuisance parameter. J. Am. Stat. Assoc. **89:** 1012–1016.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

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.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95:** 4441–4446.

Fay, J., and C. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

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

Fu, Y., and W. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

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

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

Green, R. E., J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan *et al.*, 2006 Analysis of one million base pairs of neanderthal DNA. Nature **444:** 330–336.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344:** 403–410.

Hein, J., M. H. Schierup and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution.* Oxford University Press, London/New York/Oxford.

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

Hudson, R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanism of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer, Sunderland, MA.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Innan, H., B. Padhukasahasram and M. Nordborg, 2003 The pattern of polymorphism on human chromosome 21. Genome Res. **13:** 1158–1168.

Johnson, P. L., and M. Slatkin, 2008 Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. **25:** 199–206.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Knudsen, B., and M. M. Miyamoto, 2007 Incorporating experimental design and error into coalescent/mutation models of population history. Genetics **176:** 2335–2342.

Markovtsova, L., P. Marjoram and S. Tavare, 2001 On a test of Depaulis and Veuille. Mol. Biol. Evol. **18:** 1132–1133.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Palmer, S., A. P. Wiegand, F. Maldarelli, H. Bazmi, J. M. Mican *et al.*, 2003 New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. J. Clin. Microbiol. **41:** 4531–4536.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

Simonsen, K., G. Churchill and C. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Souciet, J., M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara *et al.*, 2000 Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett. **487:** 3–12.

Stephan, W., T. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

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.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

Teshima, K. M., G. Coop and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. **16:** 702–712.

Tiffin, P., and B. S. Gaut, 2001 Molecular evolution of the wound-induced serine protease inhibitor wip1 in zea and related genera. Mol. Biol. Evol. **18:** 2092–2101.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch *et al.*, 2004 Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:** 66–74.

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.

Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi and R. W. Shafer, 2007 Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. **17:** 1195–1201.

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

Communicating editor: M. Nordborg

APPENDIX A: POWER OF THE TESTS

To test whether an observed D value significantly differs from the neutral expectations, one needs to compare the observation to limit values beyond which one can reject the regular coalescent model (for a given α-risk). Even though we do not mention it each time (for readability purposes), all the following equally applies to F, Y, and Y^*. We are interested in finding the confidence interval on D ($[D_{\text{low}}, D_{\text{up}}]$) that contains a fraction $1 - \alpha$ of all values under a standard model. Outside of this interval, we shall reject neutrality. There are two sources of variance in the frequency spectrum of the polymorphic sites. The first one comes from the variance in the shape and in the branch length of the genealogy; the second one derives from the number of mutations and their locations in the tree. Importantly, the mutation locations are dependent only on the underlying tree, which itself, if expressed in units of N generations, depends only on n (the sample size). On the other hand, the number of mutations depends on θ, the population mutation rate, and on the total tree length (expressed in N generations).

When a geneticist samples sequences from the wild and plans to use a neutrality test, he/she knows n but ignores θ; he/she can, however, easily measure some values from the sequences (*e.g.*, S, the number of polymorphic sites). There are therefore several possible strategies to compute the confidence interval $[D_{\text{low}}, D_{\text{up}}]$. The most conservative one that defines the largest confidence interval assumes that only n is known and that all θ-values are possible (*i.e.*, $\theta \in [0, +\infty[$). A second strategy assumes that some part of the data is known (typically S). Using S, we can either give a single estimate of θ (*i.e.*, $\hat{\theta}_S$) or give a confidence interval on θ. This latter option was retained by Simonsen *et al.* (1995) to compute a confidence interval $[\theta_{\text{low}}, \theta_{\text{up}}]$, using Tavaré (1984), with a small β-risk (with $\beta < \alpha$) and, using the method proposed by Berger and Boos (1994), to compute the confidence interval of D for all θ-values in $[\theta_{\text{low}}, \theta_{\text{up}}]$. Although it sets the confidence interval only for "likely" values of θ, this method is extremely time consuming. Finally, a last strategy is to assume that S is known and that the D confidence interval can be computed only for a given S (Hudson 1993; Depaulis and Veuille 1998). This last strategy gets rid of all the variance that comes from the mutation rate so it will give the narrowest confidence interval and therefore increase the power of the test. It implicitly assumes that the generated genealogies have to be weighted by their probability knowing S (Markovtsova *et al.* 2001), but the shortcut of using S instead of θ is robust under almost all neutral genealogies (Depaulis *et al.* 2001; Wall and Hudson 2001). Here, we chose to use the most conservative confidence interval (*i.e.*, for $\theta \in [0, +\infty[$). Even though one cannot scan the whole θ-space, we show that there is a relatively quick method to compute this conservative confidence interval.

To get a sense of how the confidence interval of D ($[D_{\text{low}}, D_{\text{up}}]$) varies as a function of θ, we explore, for a given n, a large range of θ-values (*i.e.*, 0–100, by steps of 0.1); they include most of the values typically encountered. For each set of n and θ, we ran 10^5 regular coalescent simulations, built the empirical distribution of D, and reported the 0.95 confidence interval. Results for $n = 10$, 50, 100, and 300 (Figure A1a) show that D_{up} exhibits a peak for low values of θ (usually $\theta < 5$). Similarly, D_{low} reaches its lowest value also for small θ-values. This suggests that the most extreme limits of D are observed for low θ-values, regardless of n. Importantly this property holds for F, Y^*, and Y.

From there, we set up a strategy to find D_{low} and D_{up}, for a given n but for $\theta \in [0, +\infty[$. For D_{up}, we want to find the most extreme value (which is reached for a small θ). Therefore, we start from $\theta = 0$ and run 10^5 standard coalescent simulations with increasing θ-values (by steps of 0.1) until the newly calculated D_{up} is less than the most extreme value of D_{up} encountered so far (by at least 0.001). The same can be applied to find the most extreme D_{low}. Both D_{up} and D_{low} are considered individually. We therefore explore until we capture the summit of the peak or the bottom of the well. It usually corresponds only to a reasonable number of steps (extreme values are observed for low θ-values). The limits computed in this way depend only on n and no longer on the unknown parameter θ. A graphical representation of the confidence interval for D, Y^*, and Y is given in Figure A1b for n ranging from 5 to 500.

We also mention that we chose here to consider the "conservative" confidence interval for all statistics to reduce the computation time. However, our strategy to find this conservative confidence interval could be easily changed to find

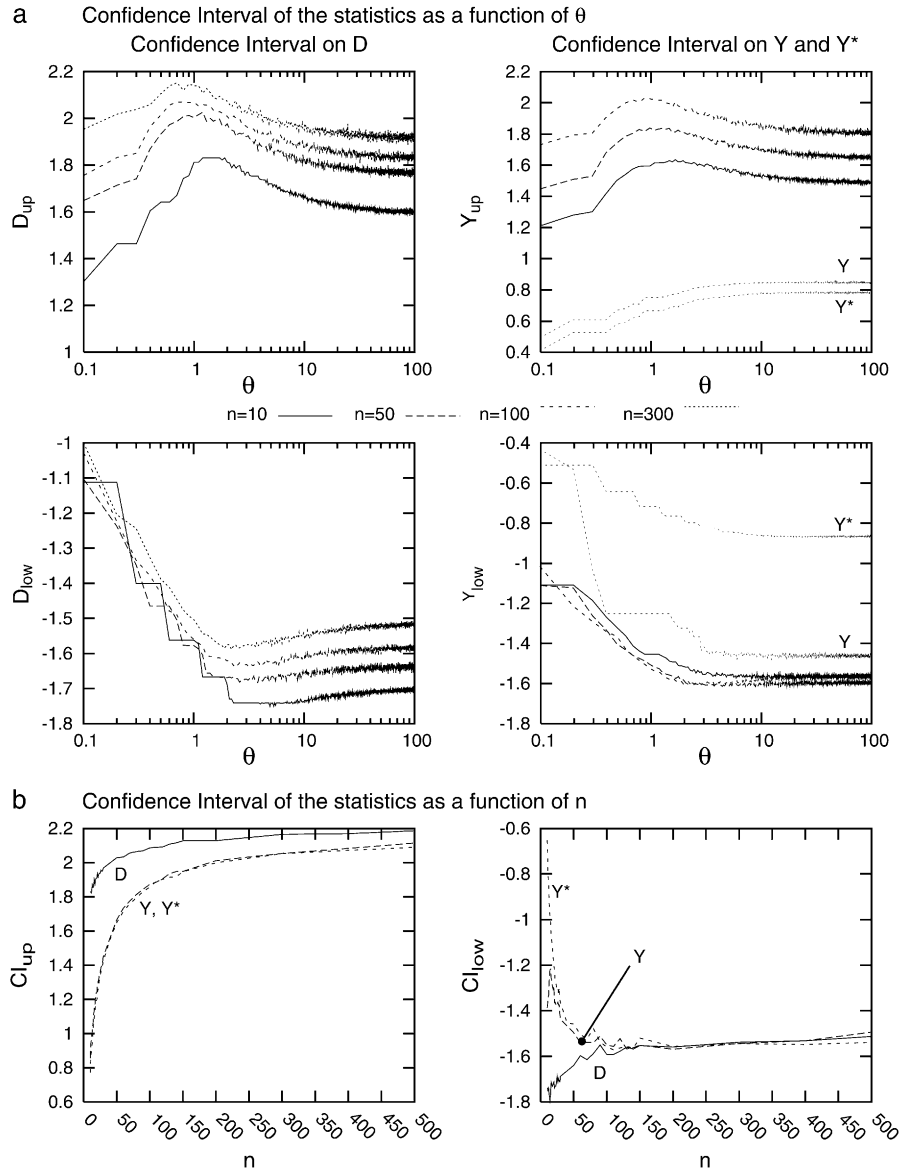

FIGURE A1.—Limits of the 95% confidence interval of of *D*, *Y*, and *Y** as a function of θ and *n*. (a) The upper and lower limits of *D* (left), *Y* (right), and *Y** (right). For $n = 10$, 50, 100, 300, we report the limits for θ varying between 0 and 100 with steps of 0.1. Values for *Y* and *Y** for $n > 10$ are so similar that we cannot distinguish one from the other. This shows that, typically, the limits peak for small θ-values. (b) We report the most extreme limits for $n \leq 500$. The conservative confidence-interval limits we computed, using our strategy, for $n \leq 500$. Note that the upper limit of *Y* and *Y** is different only for small values of $n \leq 15$.

a narrower confidence interval that considers only likely values of θ (SIMONSEN *et al.* 1995). This interesting idea can be exploited through a minor modification of our algorithm that will greatly reduce the original computation time. A modified version of Simonsen *et al.*'s strategy could be defined as follows:

1. Find D_{up} and D_{low} and their associated $\theta_{D_{\text{up}}}$ and $\theta_{D_{\text{low}}}$ as we do currently (this is affordable since only a few runs are needed from $\theta = 0$ to the peak).
2. Compute the confidence interval for θ ($[\theta_{\text{low}}, \theta_{\text{up}}]$) using *S* (TAVARÉ 1984) (this can be performed very quickly using a numerical approach).
3. Compare $\theta_{D_{\text{low}}}$ to $[\theta_{\text{low}}, \theta_{\text{up}}]$ to set the final value of the lower boundary on *D* ($D_{\text{low}}^{\text{final}}$). If $\theta_{D_{\text{low}}}$ is inside the interval $[\theta_{\text{low}}, \theta_{\text{up}}]$, set $D_{\text{low}}^{\text{final}}$ as D_{low}; otherwise perform one last run of simulations using θ_{low} (if $\theta_{D_{\text{low}}} < \theta_{\text{low}}$) or θ_{up} (if $\theta_{D_{\text{low}}} > \theta_{\text{up}}$). In the two last cases, set $D_{\text{low}}^{\text{final}}$ as the low boundary from this last run.
4. Perform almost the same procedure as in step 3 to find $D_{\text{up}}^{\text{final}}$.

Obviously the same four steps can be done for *F*, *Y**, and *Y*. This would narrow the confidence interval and therefore improve the power of the tests especially when θ is large. It, however, implies that the confidence interval has to be computed from *S* for each data set individually; this seems feasible for a given data set but inadequate for extensive simulations.

APPENDIX B: VARIANCES OF THE NEW ESTIMATORS AND STATISTICS

Elementary variances and covariances: Here we give all elementary variances of π, S, ξ_1, and η_1 as well as their covariances. These are necessary to derive the variances of $S_{-\xi_1}$, $\pi_{-\xi_1}$, $S_{-\eta_1}$, $\pi_{-\eta_1}$, Y, and Y^*.

From Watterson (1975), we know that

$$\begin{aligned}\mathrm{Var}[S] &= a_n\theta + b_n\theta^2\\ a_n &= \sum_{i=1}^{n-1}\frac{1}{i}\\ b_n &= \sum_{i=1}^{n-1}\frac{1}{i^2}.\end{aligned} \tag{B1}$$

Tajima (1983) showed that

$$\mathrm{Var}[\pi] = \frac{n+1}{3(n-1)}\,\theta + \frac{2(n^2+n+3)}{9n(n-1)}\,\theta^2. \tag{B2}$$

From Tajima (1989), we know that

$$\mathrm{Cov}[\pi,\, S] = \theta + \left(\frac{1}{2}+\frac{1}{n}\right)\theta^2. \tag{B3}$$

Fu and Li (1993) showed that

$$\mathrm{Var}[\xi_1] = \theta + \frac{2na_n - 4(n-1)}{(n-1)(n-2)}\,\theta^2 \tag{B4}$$

$$\mathrm{Cov}[\pi,\xi_1] = \frac{2(n+1)}{(n-1)^2}\left(a_n + \frac{1}{n} - \frac{2n}{n+1}\right)\theta + \frac{1}{n-1}\,\theta^2 \tag{B5}$$

$$\mathrm{Cov}[S,\xi_1] = \theta + \frac{a_n}{n-1}\,\theta^2 \tag{B6}$$

$$\mathrm{Var}[\xi_{n-1}] = \frac{\theta}{n-1} + \frac{n-2}{(n-1)^2}\,\theta^2 \tag{B7}$$

$$\mathrm{Cov}[\pi,\xi_{n-1}] = \frac{2}{n(n-1)}\,\theta + \frac{4b_n - 7 + 8/n}{(n-1)}\,\theta^2 \tag{B8}$$

$$\mathrm{Cov}[S,\xi_{n-1}] = \frac{1}{(n-1)}\,(\theta+\theta^2) \tag{B9}$$

$$\mathrm{Cov}[\xi_1,\xi_{n-1}] = \frac{1}{n-1}\left(\frac{3}{2} - \frac{2(a_n+1/n)-3}{n-2} - \frac{1}{n}\right)\theta^2. \tag{B10}$$

Using the variances of ξ_1 and ξ_{n-1} and the covariance between both, one can directly derive variances and covariances that include η_1. Here, we have expressed them in an alternative form (usually more compact) to that in Fu and Li (1993).

$$\mathrm{Var}[\eta_1] = \frac{n}{n-1}\,\theta + \left(a_n\frac{2}{(n-1)} - \frac{1}{(n-1)^2}\right)\theta^2 \tag{B11}$$

$$\mathrm{Cov}[S,\eta_1] = \frac{n}{n-1}\theta + \left(\frac{a_n+1}{n-1}\right)\theta^2 \tag{B12}$$

$$\mathrm{Cov}[\pi,\eta_1] = \left(a_n\frac{2(n+1)}{(n-1)^2} - \frac{4}{n-1}\right)\theta + \left(b_n\frac{4}{n-1} - \frac{6n-8}{n(n-1)}\right)\theta^2. \tag{B13}$$

Derivation of *Var*[*Y]:** If we define $f^* = (n-3)/(a_n(n-1)-n)$, it can then be written that

$$\begin{aligned}\mathrm{Var}[Y^*] &= Var[\pi_{-\eta_1} - f^* S_{-\eta_1}] \\ &= \mathrm{Var}[\pi_{-\eta_1}] + f^{*2}\mathrm{Var}[S_{-\eta_1}] - 2f^*\mathrm{Cov}[\pi_{-\eta_1}, S_{-\eta_1}].\end{aligned} \tag{B14}$$

Following the notations of Fu (1995), we define ξ_1 as the number of mutations in external branches and ξ_{n-1} as the number of mutations in the highest branch (whenever it exists); thus both variances as well as the covariance can be rewritten as follows:

$$\begin{aligned}\mathrm{Var}[\pi_{-\eta_1}] &= \mathrm{Var}\left[\pi - \frac{2}{n}(\xi_{n-1}+\xi_1)\right] \\ &= \mathrm{Var}[\pi] + \frac{4}{n^2}\mathrm{Var}[\xi_{n-1}] + \frac{4}{n^2}\mathrm{Var}[\xi_1] + \frac{8}{n^2}\mathrm{Cov}[\xi_{n-1},\xi_1] \\ &\quad - \frac{4}{n}\mathrm{Cov}[\pi,\xi_{n-1}] - \frac{4}{n}\mathrm{Cov}[\pi,\xi_1]\end{aligned} \tag{B15}$$

$$\begin{aligned}\mathrm{Var}[S_{-\eta_1}] &= \mathrm{Var}[S - (\xi_{n-1}+\xi_1)] \\ &= \mathrm{Var}[S] + \mathrm{Var}[\xi_{n-1}] + \mathrm{Var}[\xi_1] + 2\mathrm{Cov}[\xi_{n-1},\xi_1] \\ &\quad - 2\,\mathrm{Cov}[S,\xi_{n-1}] - 2\,\mathrm{Cov}[S,\xi_1]\end{aligned} \tag{B16}$$

$$\begin{aligned}\mathrm{Cov}[\pi_{-\eta_1}, S_{-\eta_1}] &= \mathrm{Cov}\left[\pi - \frac{2}{n}(\xi_{n-1}+\xi_1), S_{-\eta_1}\right] \\ &= \mathrm{Cov}[\pi,S] - \mathrm{Cov}[\pi,\xi_{n-1}] - \mathrm{Cov}[\pi,\xi_1] - \frac{2}{n}\mathrm{Cov}[\xi_{n-1},S] + \frac{2}{n}\mathrm{Var}[\xi_{n-1}] \\ &\quad + \frac{2}{n}\mathrm{Cov}[\xi_{n-1},\xi_1] - \frac{2}{n}\mathrm{Cov}[\xi_1,S] + \frac{2}{n}\mathrm{Cov}[\xi_1,\xi_{n-1}] + \frac{2}{n}\mathrm{Var}[\xi_1].\end{aligned} \tag{B17}$$

Therefore, by replacing the variances (Equations B15 and B16) and the covariance (Equation B17) in the variance of Y^* (Equation B14), one can show that

$$\begin{aligned}\mathrm{Var}[Y^*] &= \mathrm{Var}[\pi] + f^{*2}\mathrm{Var}[S] - 2f^*\mathrm{Cov}[\pi,S] \\ &\quad + \left(f^* - \frac{2}{n}\right)^2[\mathrm{Var}[\xi_1] + \mathrm{Var}[\xi_{n-1}] + 2\,\mathrm{Cov}[\xi_1,\xi_{n-1}]] \\ &\quad + 2\left(f^* - \frac{2}{n}\right)[\mathrm{Cov}[\pi,\xi_1] + \mathrm{Cov}[\pi,\xi_{n-1}]] \\ &\quad + 2f^*\left(\frac{2}{n} - f^*\right)[\mathrm{Cov}[S,\xi_1] + \mathrm{Cov}[S,\xi_{n-1}]].\end{aligned} \tag{B18}$$

Each element of Equation B18 is known or can be easily derived. Therefore, replacing all elementary variances and covariances (Equations B1–B10) in the variance expressed in Equation B18, the variance of Y^* can be expressed as

$$\mathrm{Var}[Y^*] = \alpha_n^*\theta + \beta_n^*\theta^2, \tag{B19}$$

where both coefficients are equal to

$$\alpha_n^* = f^{*2}\left(a_n - \frac{n}{(n-1)}\right) + f^*\left(a_n\frac{4(n+1)}{(n-1)^2} - 2\frac{n+3}{(n-1)}\right) - a_n\frac{8(n+1)}{n(n-1)^2} + \frac{n^2+n+60}{3n(n-1)} \tag{B20}$$

$$\beta_n^* = f^{*2}\left(b_n - \frac{2n-1}{(n-1)^2}\right) + f^*\left(b_n\frac{8}{n-1} - a_n\frac{4}{n(n-1)} - \frac{n^3+12n^2-35n+18}{n(n-1)^2}\right) - b_n\frac{16}{n(n-1)} + a_n\frac{8}{n^2(n-1)} + \frac{2(n^4+110n^2-255n+126)}{9n^2(n-1)^2}. \tag{B21}$$

Variances and covariance of $S_{-\eta_1}$ and $\pi_{-\eta_1}$: We can show, from solving Equations B15–B17, that

$$\mathrm{Var}[S_{-\eta_1}] = \theta\left(a_n - \frac{n}{(n-1)}\right) + \theta^2\left(b_n - \frac{2n-1}{(n-1)^2}\right) \tag{B22}$$

$$\mathrm{Var}[\pi_{-\eta_1}] = \theta\times\left(-a_n\frac{8(n+1)}{n(n-1)^2} + \frac{n^2+n+60}{3n(n-1)}\right) + \theta^2\times\left(-b_n\frac{16}{n(n-1)} + a_n\frac{8}{n^2(n-1)} + \frac{2(n^4+110n^2-255n+126)}{9n^2(n-1)^2}\right) \tag{B23}$$

$$\mathrm{Cov}[S_{-\eta_1},\pi_{-\eta_1}] = \theta\times\left(-a_n\frac{2(n+1)}{(n-1)^2} + \frac{n+3}{(n-1)}\right) + \theta^2\times\left(-b_n\frac{4}{n-1} + a_n\frac{2}{n(n-1)} + \frac{n^3+12n^2-35n+18}{2n(n-1)^2}\right). \tag{B24}$$

These values can be used as an alternative way of computing $\mathrm{Var}[Y^*]$ (using Equation B14).

Estimation of θ and θ^2 from $S_{-\eta_1}$: As for the regular D, θ is unknown, but it can be estimated using $E[S_{-\eta_1}]$, since

$$E[S_{-\eta_1}] = \left(a_n - \frac{n}{n-1}\right)\theta = \gamma_n^*\theta.$$

Hence, an adequate unbiased estimator of θ is

$$\hat{\theta}_{S_{-\eta_1}} = \frac{S_{-\eta_1}}{\gamma_n^*}. \tag{B25}$$

Similarly, θ^2 can be estimated using $E[S_{-\eta_1}]$ along with $\mathrm{Var}[S_{-\eta_1}]$. Actually, Equation B22 can be rewritten in a simplified form using adequate γ_n^* and ε_n^* for practical purposes as

$$\mathrm{Var}[S_{-\eta_1}] = \theta\gamma_n^* + \theta^2\delta_n^*.$$

Using these results, we can now express $E[(S_{-\eta_1})^2]$ as

$$\begin{aligned} E[(S_{-\eta_1})^2] &= \mathrm{Var}[S_{-\eta_1}] + E[S_{-\eta_1}]^2 \\ &= \theta\gamma_n^* + \theta^2(\delta_n^* + \gamma_n^{*2}) \\ &= E[S_{-\eta_1}] + \theta^2(\delta_n^* + \gamma_n^{*2}). \end{aligned}$$

Hence, the unbiased estimator of θ^2 is

$$\hat{\theta}^2_{S_{-\eta_1}} = \frac{S_{-\eta_1}(S_{-\eta_1}-1)}{\delta_n^* + \gamma_n^{*2}}. \tag{B26}$$

Derivation of $Var[Y]$: The use of an outgroup allows us to keep the ancestral state at frequency $(n-1)/n$. The derivations are extremely similar to the one above except that ξ_{n-1} is ignored (only ξ_1 is removed from S and π). We then have to define another constant $f = (n-2)/n(a_n - 1)$ that will be used in the Y definition:

$$\begin{aligned}\mathrm{Var}[Y] &= \mathrm{Var}[\pi_{-\xi_1} - fS_{-\xi_1}] \\ &= \mathrm{Var}\left[\pi - \frac{2}{n}\xi_1 - f(S - \xi_1)\right] \\ &= \mathrm{Var}[\pi] + f^2\mathrm{Var}[S] - 2f\,\mathrm{Cov}[\pi, S] + \left(f - \frac{2}{n}\right)^2 \mathrm{Var}[\xi_1] \\ &\quad + 2\left(f - \frac{2}{n}\right)\mathrm{Cov}[\pi, \xi_1] + 2f\left(\frac{2}{n} - f\right)\mathrm{Cov}[S, \xi_1]. \end{aligned} \tag{B27}$$

Replacing all elementary variances and covariances (Equations B1–B6) in Equation B27, the variance of Y can be expressed as

$$\mathrm{Var}[Y] = \alpha_n\theta + \beta_n\theta^2, \tag{B28}$$

where both coefficients are equal to

$$\alpha_n = f^2(a_n - 1) + f\left(a_n\frac{4(n+1)}{(n-1)^2} - \frac{2(n+1)(n+2)}{n(n-1)}\right) - a_n\frac{8(n+1)}{n(n-1)^2} + \frac{n^3 + n^2 + 60n + 12}{3n^2(n-1)} \tag{B29}$$

$$\begin{aligned}\beta_n = f^2&\left(b_n + a_n\frac{4}{(n-1)(n-2)} - \frac{4}{(n-2)}\right) + f\left(-a_n\frac{4(n+2)}{n(n-1)(n-2)} - \frac{n^3 - 3n^2 - 16n + 20}{n(n-1)(n-2)}\right) \\ &+ a_n\frac{8}{n(n-1)(n-2)} + \frac{2(n^4 - n^3 - 17n^2 - 42n + 72)}{9n^2(n-1)(n-2)}. \end{aligned} \tag{B30}$$

Variances and covariance of $S_{-\xi_1}$ and $\pi_{-\xi_1}$: Using individual variances and covariances (Equations B2–B6), we can show that

$$\mathrm{Var}[S_{-\xi_1}] = \theta \times (a_n - 1) + \theta^2 \times \left(b_n + a_n\frac{4}{(n-1)(n-2)} - \frac{4}{(n-2)}\right) \tag{B31}$$

$$\begin{aligned}\mathrm{Var}[\pi_{-\xi_1}] = \theta &\times \left(-a_n\frac{8(n+1)}{n(n-1)^2} + \frac{n^3 + n^2 + 60n + 12}{3n^2(n-1)}\right) \\ &+ \theta^2 \times \left(a_n\frac{8}{n(n-1)(n-2)} + \frac{2(n^4 - n^3 - 17n^2 - 42n + 72)}{9n^2(n-1)(n-2)}\right) \end{aligned} \tag{B32}$$

$$\begin{aligned}\mathrm{Cov}[S_{-\xi_1}, \pi_{-\xi_1}] = \theta &\times \left(-a_n\frac{2(n+1)}{(n-1)^2} + \frac{(n+1)(n+2)}{n(n-1)}\right) \\ &+ \theta^2 \times \left(a_n\frac{2(n+2)}{n(n-1)(n-2)} + \frac{n^3 - 3n^2 - 16n + 20}{2n(n-1)(n-2)}\right). \end{aligned} \tag{B33}$$

Estimation of θ and θ^2 from $S_{-\xi_1}$: We have

$$E[S_{-\xi_1}] = (a_n - 1)\theta = \gamma_n\theta.$$

Hence, the adequate θ unbiased estimator is

$$\hat{\theta}_{S_{-\xi_1}} = \frac{S_{-\eta_1}}{\gamma_n}. \tag{B34}$$

Using γ_n and δ_n as adequate coefficients to write Equation B31 in a simplified form, we can write that

$$\begin{aligned} E[(S_{-\xi_1})^2] &= \mathrm{Var}[S_{-\xi_1}] + E[S_{-\xi_1}]^2 \\ &= E[S_{-\xi_1}] + \theta^2(\delta_n + \gamma_n^2). \end{aligned}$$

Therefore, the estimation of θ^2 can be done by

$$\hat{\theta}^2_{S_{-\xi_1}} = \frac{S_{-\xi_1}(S_{-\xi_1} - 1)}{\delta_n + \gamma_n^2}. \tag{B35}$$