Population genetic inference using a fixed number of segregating sites: a reassessment

SEBASTIÁN E. RAMOS-ONSINS*, SYLVAIN MOUSSET†, THOMAS MITCHELL-OLDS§ AND WOLFGANG STEPHAN‡
1 Max Planck Institute of Chemical Ecology, Hans-Knoell-Strasse 8, 07745 Jena, Germany
2 Biocenter, Department of Biology II, University of Munich, 82152 Planegg-Martinsried, Germany

(Received 1 August 2006 and in revised form 1 August 2007)

Summary

Coalescent theory is commonly used to perform population genetic inference at the nucleotide level. Here, we examine the procedure that fixes the number of segregating sites (henceforth the FS procedure). In this approach a fixed number of segregating sites ($S$) are placed on a coalescent tree (independently of the total and internode lengths of the tree). Thus, although widely used, the FS procedure does not strictly follow the assumptions of coalescent theory and must be considered an approximation of (i) the standard procedure that uses a fixed population mutation parameter $\theta$, and (ii) procedures that condition on the number of segregating sites. We study the differences in the false positive rate for nine statistics by comparing the FS procedure with the procedures (i) and (ii), using several evolutionary models with single-locus and multilocus data. Our results indicate that for single-locus data the FS procedure is accurate for the equilibrium neutral model, but problems arise under the alternative models studied; furthermore, for multilocus data, the FS procedure becomes inaccurate even for the standard neutral model. Therefore, we recommend a procedure that fixes the $\theta$ value (or alternatively, procedures that condition on $S$ and take into account the uncertainty of $\theta$) for analysing evolutionary models with multilocus data. With single-locus data, the FS procedure should not be employed for models other than the standard neutral model.

1. Introduction

Monte Carlo simulation based on the coalescent is widely used in population genetics. This approach enables researchers to generate data for a given sample size under a panmictic neutral model or other evolutionary models. Using tests of neutrality or other summary statistics, the observed and simulated data can then easily be compared.

In order to simulate a sample under a strict neutral panmictic model, it is necessary to know the population mutation parameter $\theta$ (where $\theta = 4Nu$, $N$ being the effective population size and $u$ the mutation rate). The population mutation parameter, however, is generally unknown and must be estimated (Watterson, 1975; Tajima, 1983; Fu, 1994; Griffiths & Tavaré, 1994; Kuhner et al., 1995). To avoid the uncertainty in using an estimate of $\theta$ (usual estimation methods are mostly inefficient, e.g. mean pairwise difference is inconsistent, Watterson’s estimate converges only asymptotically), several approaches have been proposed (e.g. see Hudson, 1993; Markovtsova et al., 2000; Simonsen et al., 1995). The most popular approximate procedure simulates samples by fixing the number of segregating sites ($S$) instead of using the mutational parameter $\theta$ (Hudson, 1993). However, using this approach, population samples with short genealogical trees tend to exhibit high
mutation rates, and conversely, long genealogical trees have low mutational rates.

This latter procedure should not be confused with approaches that condition on $S$ taking into account the parameter $\theta$. A procedure that conditions on $S$ (instead of fixing $S$) considers for a given $\theta$ all possible trees weighted by their probabilities of giving rise to the observed number of segregating sites (Tavare\' et al., 1997; Markovtsova et al., 2000, 2001; Jakobsson et al., 2006), and thus takes into account the uncertainty of $\theta$ (see Kelly, 1997; Depaulis et al., 2001, equation 2). In the case of conditioning on $S$, the internode times are not independent. Therefore, the procedure proposed by Hudson (1993) might be considered from two different points of view: (i) as an approximation of the standard procedure that is based on a (known) $\theta$ value, and (ii) as an approximation of a rigorous procedure that conditions on $S$ and takes into account the uncertainty of $\theta$.

Although fixing the number of segregating sites does not strictly follow the assumptions of coalescent theory (e.g. independence of the genealogical and mutational phases: see Kingman, 1982a, b; Hudson, 1990; Donnelly & Tavaré, 1995; Nordborg, 2001), this approximate procedure is reasonably accurate for obtaining critical values of statistics for the standard neutral model, provided that the true $\theta$ value is well supported by the data (Depaulis et al., 2001; Markovtsova et al., 2001; Wall & Hudson, 2001). Thus, the probability of obtaining a $\theta$ value that is not supported by the data is expected to be low (Depaulis et al., 2001), and the critical values of tests obtained by fixing the number of segregating sites appear to be quite accurate. However, for alternative evolutionary models the accuracy of this approximation is unknown, although many publications have used a fixed number of segregating sites for testing purposes (e.g. Braverman et al., 1995; Depaulis & Veuille, 1998; Wall, 1999; Fay & Wu, 2000; Przeworski, 2002; Ramos-Onsins & Rozas, 2002; Glinka et al., 2003). It is therefore necessary to study the accuracy of this procedure for a wide range of models.

2. Simulation methods

Simulations have been performed with the program mlcoalsim (Ramos-Onsins & Mitchell-Olds, 2007), which is available at http://www.ub.edu/softevol/mlcoalsim.

(i) Evolutionary models

The following models are considered:

(a) The neutral panmictic model with constant population size.

(b) The symmetric finite-island model (Wright, 1943) with a constant number of populations ($d$), equal population size for each population, and a symmetric and constant migration parameter $M$ among islands (with $M=4Nm$, where $N$ is the effective population size of a deme and $m$ the migration rate).

(c) The logistic growth model (Fu, 1997). In this the population size changes with time as follows:

$$
N(t) = \begin{cases} 
N_0 & \text{if } t \leq t_0, \\
N_0 + \frac{N_1 - N_0}{1 + e^{-\gamma(t-t_1-\frac{t}{t_0})}} & \text{if } t_0 < t < t_1, \\
N_1 & \text{if } t \geq t_1.
\end{cases}
$$

(1)

Here $N(t)$ is the population size at time $t$ (expressed in $N_0$ generations), $N_0$ is the population size at time $t_0$, and $N_1$ is the population size at time $t_1$. We used $\gamma=10/(t_1-t_0)$. Coalescence times were calculated by integrating (1) over $t$. Expansion and reduction of population size was studied using this model. We used 10- and 100-fold differences between $N_1$ and $N_0$. We also used recent expansion processes, where $t_1-t_0=0.1N_0$ generations, and $t_0=0$.

(d) A bottleneck model using a constant population size initially for a period of time $t_d$ (expressed in $4N$ generations), then a sharp reduction for a time $t_b$ followed by a size increase. Reduction and expansion of population size also follow a logistic model. The conditions used here are bottlenecks of 10- or 100-fold differences in effective population size.

(e) Hitchhiking model: We followed essentially the algorithm described in Braverman et al. (1995) to generate hitchhiking genealogies, and allowed recombination within the locus of interest during the selective and neutral phases (see also Fay & Wu, 2000; Kim & Stephan, 2002; Przeworski, 2002). As in Fay & Wu (2000), we used as parameters the selection coefficient ($s$), the recombination rate between the selected locus and the studied locus ($c$), the intragenic recombination rate ($r$), and the time at which an advantageous mutation is fixed ($t_j$). For the selective phase, we calculated the time to the most recent common ancestor (instead of checking at small time increments) for both the selected and unselected ‘subpopulations’ using the reasoning of Nordborg (2001, equation 7). The selective phase starts at time $t_f$ with a frequency of the selected allele of $1-1/2N$, and ends when the frequency of $x(t)<1/N$. The value $x(t)$ was calculated deterministically using equation 1 in Kim & Stephan (2002) (see also equation 3a in Stephan et al., 1992). The computer code was tested by comparing the results with those of table 2 in Stephan.
et al. (1992) and also by comparison with the ssw program (Kim & Stephan, 2002). The parameter values used in this work were $t_f=0$, $4N\theta = 2 \times 10^4$, $N=10^4$ and $\epsilon = 1/4N$.

(ii) Monte Carlo methods

In all procedures, we used sample data ($n$ lines) obtained from a diploid species. The procedures are as follows:

(a) The $F\theta$ procedure (Fixed $\theta$ procedure): This is the original standard coalescent procedure. We used a fixed value of the population mutation parameter $\theta$. We placed a number of mutations on the tree according to the Poisson distribution with the mean value $\theta$ times the total length of the tree.

(b) The $F\bar{S}$ procedure (Fixed $S$ procedure): We placed a fixed number of segregating sites uniformly on each tree generated under the models mentioned above (Hudson, 1993). The procedure for generating trees is identical to the $F\theta$ procedure.

(c) Procedure based on fixing the number of segregating sites and the value of $\theta$ ($F\bar{S} \theta$): This procedure employs the rejection algorithm #2 justified and described by Tavaré et al. (1997).

(d) Procedure based on fixing the number of segregating sites but taking into account the uncertainty of the value of $\theta$ ($F\bar{S} \theta_{prior}$ procedure): We use the RAU procedure (Rejection Algorithm using a Uniform prior). This procedure employs the rejection algorithm #2 described by Tavaré et al. (1997) but sampling of $\theta$ is done from a given prior distribution instead of having a fixed $\theta$ value.

(iii) Mutational parameter $\theta$

We assume a uniform distribution over some arbitrarily chosen interval $[\theta_{\text{min}} ; \theta_{\text{max}}]$ (Depaulis et al., 2001). Thus the prior density of $\theta$ is

$$g_\theta(\theta) = \frac{1}{\theta_{\text{max}} - \theta_{\text{min}}}.$$  \hspace{1cm} (2)

When we do not have information about the value of $\theta$, the assumption of a uniform density for any value of $\theta$ is reasonable. This is a commonly used strategy to estimate $\theta$ given observed data (e.g. Watterson, 1975; Wright & Charlesworth, 2004; Wright et al., 2005; Haddrill et al., 2005). If the researcher has available information about the distribution of $\theta$, then this information should be used instead of assuming a uniform distribution (e.g. Pritchard et al., 1999; Przeworski, 2003), but we do not consider other prior distributions in this paper.

In order to avoid biologically unrealistic values of $\theta$, we used as a minimum bound of $\theta$ per nucleotide a value of $0.0005$, and as a maximum bound a value of $0.05$. Numerous publications show that these numbers are realistic. Different bounds might modify some of our results, but do not change the main conclusions.

(iv) Statistical methods

From the simulated trees we calculated Tajima’s $D$ (Tajima, 1989), here named $T_D$, Fu & Li’s $D$ and $F$ (Fu & Li, 1993), here named $F_D$ and $F_F$, respectively, and Fay & Wu’s $H$ (Fay & Wu, 2000), abbreviated as $H$, the statistic $B$ (Wall, 1999) (considered sensitive to structured populations), $F_S$ (Fu, 1997) and $R_2$ (Ramos-Onsins & Rozas, 2002), which are sensitive to population size expansion, as well as the number of haplotypes (here divided by the sample size) $K_w$ (Strobeck, 1987; Fu, 1996; Depaulis et al., 2001; Wall, 1999) and the haplotype diversity $H_w$ (Depaulis & Veuille, 1998). In total, nine summary statistics were computed. The calculation of these statistics was examined with the software package DnaSP 3.51 (Rozas & Rozas, 1999; Rozas et al., 2003).

For multilocus analyses, we treated each locus independently because we assumed that the studied loci are unlinked. When we used the RAU procedure, we chose independent $\theta$ values for each locus in order to perform simulations. For a given statistic, we recorded the $P$ value independently for each locus and combined them as in Voight et al. (2005) but calculated each tail separately.

To avoid excessively liberal critical values (95% interval of the null sampling distribution) for discrete distributions, we have used the following procedure: for the 2.5% interval of the upper tail, we took the first value that is larger than the observed value at 2.5% of the distribution (see Ramos-Onsins & Rozas, 2002). The same logic was applied to the 2.5% interval of the lower tail and also for comparing discrete distributions. Because of this issue, the realized type I error rates are slightly lower than the expected 5% and have to be assessed for every statistic and method.

(v) Determining the accuracy of the $FS$ procedure in relation to $F\theta$

We have used the approach described in Wall & Hudson (2001) to determine the level of accuracy of the $FS$ procedure compared with the $F\theta$ procedure. In this approach, a number of simulations using a large range of values of $S$ (e.g. from $S=1$ to $S=120$) is obtained by the $FS$ procedure, and its critical values and the false positive rate (called the ‘nominal’ size) for each statistic are calculated for a given evolutionary model. Then, the ‘true’ $\theta$ value for the same evolutionary model is obtained, which is the value that gives an average of $S=20$ in 10 000 iterations.
under the $F\Theta$ procedure. Next, a simulation with the $F\Theta$ procedure (i.e. the null hypothesis) is performed using the ‘true’ fixed value of $\theta$. For each iteration, the value of $S$ is calculated, and acceptance or rejection of the null hypothesis is determined for every statistic by comparing its value with the critical values of the corresponding $FS$ distribution. The rejection rate (i.e. the false positive rate given the $FS$ critical values, here called the ‘true’ size for each statistic) is stored. The choice of the ‘true’ $\theta$ value, although somewhat arbitrary, is suggested by the design of the study, which fixes the number of segregating sites instead of fixing the value of $\theta$ (which is unknown to the researcher). Using the following method, we try to be conservative in the sense of minimizing the differences between the two different procedures (see Wall & Hudson, 2001; Depaulis et al., 2003, 2005).

(vi) **Determining the accuracy of the $FS_{U}$ procedure in relation to $FS_{0U}$**

Here we assume that the $\theta$ value is unknown. The $FS_{0U}$ procedure is considered the ‘true’ procedure because it takes into account the uncertainty assumed by the researcher. In this approach, a simulation for a given evolutionary model is performed for the $FS$ procedure (for a fixed $S = 20$) and its critical values and the false positive rate (called the ‘nominal size’) of each statistic are stored. Then, a simulation using the $FS_{0U}$ procedure (for $S = 20$) is performed for the same evolutionary model (the null hypothesis). The ‘true’ false positive rate is calculated in the following way: for each iteration calculated with the $FS_{0U}$ procedure, rejection of the null hypothesis is determined for every statistic by comparing its value with the critical values of the $FS$ distribution. Finally, the rejection rate for each statistic is stored.

(vii) **Parameter values**

We used $n = 20$ for a locus of 1000 nucleotides. For procedures that fix $S$, we preferentially used $S = 20$. The reason for choosing $n = 20$ and $S = 20$ is largely historical, as the value of $\theta$ (given the neutral equilibrium model) is then approximately 0.005 (which has been used in a number of the theoretical population genetics studies). All $\theta$ values under any evolutionary model studied here were in the range we considered biologically realistic. For $n = 20$ and $S = 50$ we found similar results.

3. Analytical approaches

(i) **Effects of recombination**

Recombination is difficult to take into account in analytical models. However, one can consider the extreme case of free recombination (e.g. a sequence consisting of $m$ freely recombining fragments of equal size such as $m$ nucleotides). This is the limiting case when the recombination parameter $R = 4Nr$ becomes very large. Let the parameter $T$ be the vector of the coalescent times $T_h$, then $L_n = \sum_{h=1}^{m} T_h$ is the total length of the coalescent tree (by summing the length of all branches) measured in units of $4N$ generations. Assuming $m$ is large, the central limit theorem shows that the prior distribution of the average length $\bar{T}_n$ of the $m$ independent coalescent trees will converge towards the normal distribution with mean $\eta(n)$ and variance $\sigma(n)/m$ (defined in equations (A3) and (A4) of Appendix 1, respectively). This is,

$$\lim_{R \to \infty} f_{\bar{T}_n}(t) = \sqrt{\frac{m}{2\pi\sigma(n)}} \frac{m}{2\sigma(n)} (t - \eta(n))^2 ,$$

where $f_{\bar{T}_n}(t)$ is the density of the average length $\bar{T}_n$. This last equation can be used in equations (A7)–(A8) (Appendix 1) to address the effects of the simulation procedures $F\Theta$, $FS$, $F\Theta_0$, and $FS_{0prior}$ on the posterior density of the length of the coalescent tree in the limiting case of freely recombining sequences.

(ii) **Shape of the coalescent trees**

In the present work, we focus on procedures where simulations use a fixed number of segregating sites $S$. In these procedures, unlike in the $F\Theta$ procedure, the total length of a tree does not play a role as long as the shape of the tree remains the same (trees are only scaled). In order to assess the impact of the rejection algorithm on the shape of the trees, we study the ratio of the branch lengths in the upper and lower parts of the tree $X/Y$ (Fig. 1). The mean value of this ratio is given by

$$E\left(\frac{X}{Y_n} \mid n, S \right) = \frac{1}{\eta(n, S)} \times \int_0^\infty \int_0^\infty \frac{-f_X(x)f_Y(y)}{y} \mathcal{P}_{accept}(x,y) dxdy,$$

where $f_X$ and $f_Y$ denote the prior densities of $X$ and $Y_n$ (see equations (A5) and (A6)) and

$$\mathcal{P}_{accept}(x,y) = \begin{cases} \frac{1}{S!} \theta^S e^{-(x+y)\theta} & (FS \text{ procedure}) \\ \frac{(x+y)^S}{S!} \int_{\theta}^{\infty} e^{-(x+y)\theta} f(\theta) d\theta & (F\Theta_0 \text{ procedure}) \end{cases}$$

and $\eta(n, S) = \int_0^\infty \int_0^\infty f_X(x)f_Y(y) \mathcal{P}_{accept}(x,y) dxdy$ is a normalizing constant.
4. Simulation results

We determined the accuracy of the approximate procedure proposed by Hudson (1993) (which fixes the number of segregating sites while ignoring the value of the population mutation value \( \theta \); the FS procedure) in comparison with the two rigorous procedures: (i) the standard procedure fixing \( \theta \) (FS\theta) and (ii) procedures that while conditioning on \( S \) take into account the uncertainty of \( \theta \) (FS\theta\prior), here using a uniform prior, named FS\theta\Uc. We consider several different evolutionary models including the neutral panmictic model, population subdivision, population size bottleneck, expansion and genetic hitchhiking. As a measure of accuracy, we calculated the difference in the false positive rate between procedures.

(i) Comparison of the FS procedure with the standard procedure F0

We examined the accuracy of the FS procedure for different alternative models. To do this we also re-examined the type I error for the standard neutral model (neutral panmictic population with zero recombination) for the nine statistics studied here, since the type I error for single-locus data was also studied in Depaulis et al. (2001), Markovtsova et al. (2001) and Wall & Hudson (2001).

Table 1 shows the difference between the nominal size (the false positive rate of a test given the FS procedure) and the ‘true’ size (the false positive rate of a test given the F\theta procedure) for a 5% critical region. Our analysis confirmed the small difference in the type I error rate observed by Wall & Hudson (2001) and Depaulis et al. (2001) under a neutral model for a single locus, although we used some other statistics and slightly different conditions (\( n = 20 \) and \( \theta = 0.0057 \) for the F0 procedure; i.e. the estimate of \( \theta \) for \( S = 20 \)). Fig. 2A shows the differences in size for each \( S \) value among the procedures for the nine studied statistics under the neutral model. Negative values indicate that the critical values of the FS procedure are more liberal than expected (increased type I error). The large fluctuations observed in Fig. 2 are also a consequence of the discrete distribution of values obtained when the number of segregating sites is fixed (see Ramos-Onsins & Rozas, 2002). Important differences are observed in the proportion of acceptance/rejection when the number of mutations is far from the average (\( S = 20 \)), as indicated in Wall & Hudson (2001) (see also Depaulis et al., 2003, 2005). Nonetheless, the more extreme values contribute very little such that when summed the differences cancel each other out, thus leading to a good accuracy of the FS procedure for the standard neutral model (Table 1). Therefore, we consider the FS procedure as sufficiently accurate under the neutral panmictic model for single-locus data.

Next we consider the accuracy for alternative models. Table 1 shows the difference between nominal (FS procedure) and true (F\theta procedure) distributions of statistical tests for certain critical values. The reason to use different critical values for alternative models (10%, 50% and 90% instead 2.5% and 97.5%) is because we are interested in whether the probability distribution for a given statistic can be accurately represented using the FS procedure. The analysis is performed for five different alternative models (subdivision, expansion, contraction, bottleneck and hitchhiking) and for some arbitrary parameters. The strong differences observed between the two procedures indicate that the FS procedure is inaccurate when alternative models are used. Fig. 2B shows considerable differences between nominal and true sizes for several of the studied statistics for each \( S \) separately, leading to a large difference in total (Table 1).

(ii) Comparison of the FS procedure with a procedure that conditions on \( S \) and takes into account the uncertainty of \( \theta \)

We have examined the accuracy of FS (the nominal procedure) by comparison with the procedure FS\theta\Uc (the ‘true’ procedure, see Section 2). With regard to the type I error for the standard neutral model, we have studied the 95% interval of the nine neutrality test statistics for the FS procedure and the procedure FS\theta\Uc by simulation for \( n = 20 \) and a range of \( S \) values from \( S = 2 \) to 80. Fig. 3 shows the difference between the nominal and ‘true’ sizes. Liberal tests are Fay & Wu’s \( H \) with regard to the upper tail, and Fu’s \( F_{S} \), Ramos-Onsins & Rozas’ \( R_{2} \), and Fu & Li’s \( D \) and \( F \).
Table 1. Difference (in percentage) between the FS and F0 procedures for several critical values\(^a\)

<table>
<thead>
<tr>
<th>n = 20</th>
<th>TD</th>
<th>Fs</th>
<th>FD</th>
<th>FF</th>
<th>H</th>
<th>B</th>
<th>Kw</th>
<th>Hw</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4Nr = 0</td>
<td>&lt;2.5%</td>
<td>-0.08</td>
<td>0.23</td>
<td>0.86</td>
<td>0.06</td>
<td>-0.14</td>
<td>-0.01</td>
<td>-0.55</td>
</tr>
<tr>
<td></td>
<td>&gt;97.5%</td>
<td>-0.15</td>
<td>-0.11</td>
<td>0.97</td>
<td>-0.21</td>
<td>0.13</td>
<td>0.15</td>
<td>0.12</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>4Nr = 10</td>
<td>&lt;2.5%</td>
<td>0.08</td>
<td>0.18</td>
<td>0.71</td>
<td>0.14</td>
<td>0.06</td>
<td>-0.04</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>&gt;97.5%</td>
<td>0.05</td>
<td>-0.10</td>
<td>1.32</td>
<td>-0.02</td>
<td>0.23</td>
<td>0.22</td>
<td>1.07</td>
<td>-0.33</td>
</tr>
</tbody>
</table>

\(h\)Neutral

\(4Nr = 0\)

\(h\)Subdivision\(^b\)

\(4Nr = 0\)

\(h\)Expansion\(^c\)

\(4Nr = 0\)

\(h\)Contraction\(^d\)

\(4Nr = 0\)

\(h\)Bottleneck\(^e\)

\(4Nr = 0\)

\(h\)Hitchhiking\(^f\)

\(4Nr = 0\)

\(h\)Abbreviations are explained in Simulation methods. The critical values studied are 2.5% and 97.5% for the neutral model and 10%, 50%, and 90% for alternative models (see Simulation methods). The maximum precision error detected is around ±1%.

\(b\) Island model with \(d = 10\) subpopulations, \(4Nm = 0.5\) and 4Nr = 10. The samples are all obtained from a single population. In the FSO\(_U\) procedure, the bounds of \(\theta\) were reduced \(d\) times for a better comparison with the other models. In the F0 procedure, \(\theta = 0.00062\).

\(c\) Logistic expansion model with a 10-fold growth of population size. The expansion process started 0.4\(N_e\) generations ago and finished at present. 4\(N_e\)\(g^f\) = 10. In the F0 procedure, \(\theta = 0.03384\).

\(d\) Logistic contraction model: In contrast to the logistic expansion model, the population is reduced 10 fold. In the F0 procedure, \(\theta = 0.00059\).

\(e\) Bottleneck model: Population size is maintained constant during 0.5\(N_e\) generations since present. It follows a logistic, 100-fold reduction of population size for 0.01\(N_e\) generations, maintained for 0.01\(N_e\) generations, then an instantaneous recovery to the present population size. 4\(N_e\)\(g^f\) = 10. In the F0 procedure, \(\theta = 0.00860\).

\(f\) Hitchhiking model: The selective phase is completed at present time (\(t = 0\)). \(N = 10^4\), \(4Ns = 10^4\), the population recombination parameter between the selected and observed locus is 4\(Nc = 400\), \(c/s = 0.02\), and intragenic recombination is indicated in the table. In the F0 procedure, \(\theta = 0.00127\).
for the lower tail. In the case of Fu & Li's $F$, the test is, in fact, less liberal than shown, because the critical value for the $FS$ procedure is sometimes much lower than 2.5% (given the discrete distribution of values; not shown). We observed that all statistics at intermediate $S$ values show a difference lower than 1.5%. Therefore, in this comparison the $FS$ procedure is sufficiently accurate for statistical inferences under the neutral panmictic model for single-locus data.

In contrast, for alternative models there are strong differences between the nominal ($FS$ procedure) and true ($FS\theta_U$ procedure) distributions of statistical tests for 10%, 50% or 90% critical values (Table 2). The important differences observed between the two procedures indicate that the $FS$ procedure is also inaccurate when it is compared with the $FS\theta_U$ procedure and when alternative models are used.

We performed a multilocus analysis for the nine test statistics using the combined $P$ values for different numbers of loci (see Section 2) for the standard neutral model. The difference between the nominal and true size is examined in Fig. 4. The results indicate that for a larger number of loci the $FS$ procedure becomes inaccurate (through the accumulation of small departures in the single-locus tests). Thus, the small differences observed for one locus between $FS$ and

---

**Fig. 2.** Differences between the sizes of $FS$ and $F\theta$ procedures for each $S$ separately. (A) Neutral model: In the upper panel the distribution of $S$ values obtained with the $F\theta$ procedure using $\theta = 0.0057$ is shown. In the middle and lower panels the size differences for statistics for the upper (97.5%) and the lower (2.5%) tails, respectively, are presented. Note that large $S$ values, although causing large differences, contribute very little to the total. (B) Hitchhiking model: In the upper panel the distribution of $S$ values obtained with the $F\theta$ procedure using $\theta = 0.0127$ is shown. In the middle and lower panels the size differences for statistics for the upper (90%) and the lower (10%) tails, respectively, are presented. Abbreviations are explained in Section 2. Critical values are not calculated for $B$ (lower tail), and for $K_w$ and $H_w$ (upper tail), because these statistics are not conservative with recombination.
neutrality tests for the 2.5% upper and lower tails. Abbreviations are indicated in Section 2.

Statistics such as Tajima’s $D$, Fu’s $F$, Fu & Li’s $D$ and $F$, and Ramos-Onsins & Rozas’ $R^2$ are too liberal for the lower tail of the distribution, and Fay & Wu’s $H$ is also too liberal for the upper tail of the distribution.

5. Discussion

We have compared the $FS$ procedure with the standard procedure using a fixed $\theta$ value ($F\theta$) and with a procedure that conditions on $S$ taking into account the uncertainty of $\theta$ ($FS\theta_U$). Our results show that the $FS$ procedure is inaccurate in both comparisons when alternative models are used.

(i) Causes of the discrepancy between the $FS$ and the $F\theta$ or $FS\theta_U$ procedures

In the comparison of $FS$ with $F\theta$, differences in the length ($L_n$) and topology ($X/Y_n$) of the trees for the $FS$ and $F\theta$ procedures can explain the inaccuracy of $FS$, given that the critical values of $FS$ are compared with the $F\theta$ distribution for each $S$. The distribution of the total length of the coalescent trees of a non-recombining sequence is shown in Fig. 5. The shapes of the $FS$ distribution and the distribution obtained with the $FS\theta$ procedure (taking $S=20$) are clearly different. The distribution of $L_n$ obtained with the $FS\theta$ procedure has lower variance than the $FS$ procedure, which can be explained by the lack of uncertainty in the value of $\theta$ (see also Table 3). The mean and standard deviation of the shape index $X/Y_n$ for the different simulation procedures are shown in Table 4. Higher values are obtained for trees with long internal branches whereas smaller ratios denote shorter internal branches. We observed that for $S=20$ the trees obtained with the $FS\theta$ procedure have smaller ratios $X/Y_n$ than the trees obtained with the $FS$ procedure. These observations can be explained as follows. For a given $\theta$, when $S$ is lower than the average for that $\theta$, this will often have been caused by a tree that is shorter than average. Although the internode times are a priori independent (because the total tree length is dominated by the last few internode times, given a short tree), it will often be that the final internode time is particularly short. Thus $X/Y_n$ is smaller than the unconditional expectation. The situation is reversed for $S$ greater than the average for a given $\theta$. Thus, these results show how the simulation procedures may affect the shape of the sampled coalescent trees and may lead to different outcomes of neutrality tests.

To understand the causes of the inaccuracy of the $FS$ procedure in comparison with the procedure $FS\theta_U$, we examined differences in the trees by considering the total length of the tree ($L_n$) analytically and by simulations. The distribution of the total length of the coalescent tree of a non-recombining sequence is shown in Fig. 5. Some inaccuracies of the $FS$ procedure in comparison with the procedure $FS\theta_U$
Population genetic inference

Table 2. Difference (in percentage) between the FS and the FS₀_U procedures for several critical values

<table>
<thead>
<tr>
<th>n = 20</th>
<th>TD</th>
<th>Fₛ</th>
<th>FD</th>
<th>FF</th>
<th>H</th>
<th>B</th>
<th>Kₑ</th>
<th>Hₑ</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subdivision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4Nₑ = 0</td>
<td>&lt; 10%</td>
<td>0.7</td>
<td>2.8</td>
<td>0.6</td>
<td>0.3</td>
<td>1.2</td>
<td>2.4</td>
<td>2.6</td>
<td>3.4</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>1.2</td>
<td>-10.7</td>
<td>-1.7</td>
<td>-2.2</td>
<td>0.4</td>
<td>-5.5</td>
<td>14.1</td>
<td>11.6</td>
<td>-1.2</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>2.1</td>
<td>3.4</td>
<td>0.0</td>
<td>-1.9</td>
<td>2.4</td>
<td>3.1</td>
<td>3.3</td>
<td>2.6</td>
<td>2.2</td>
</tr>
<tr>
<td>4Nₑ = 10</td>
<td>&lt; 10%</td>
<td>-0.5</td>
<td>-2.5</td>
<td>0.1</td>
<td>0.0</td>
<td>0.5</td>
<td>2.1</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>0.1</td>
<td>-2.2</td>
<td>-0.6</td>
<td>-0.4</td>
<td>2.2</td>
<td>7.6</td>
<td>2.4</td>
<td>2.0</td>
<td>-0.4</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>-1.2</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.3</td>
<td>-1.3</td>
<td>-1.1</td>
<td>-1.0</td>
<td>-1.1</td>
<td>-1.0</td>
</tr>
<tr>
<td>Expansion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4Nₑ = 0</td>
<td>&lt; 10%</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
<td>-0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>0.3</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.8</td>
<td>-0.8</td>
<td>0.4</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>0.6</td>
<td>1.2</td>
<td>-1.0</td>
<td>0.8</td>
<td>-0.4</td>
<td>1.5</td>
<td>0.2</td>
<td>-0.3</td>
<td>0.9</td>
</tr>
<tr>
<td>4Nₑ = 10</td>
<td>&lt; 10%</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.4</td>
<td>-0.1</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>-1.2</td>
<td>-0.8</td>
<td>-0.9</td>
<td>-0.6</td>
<td>1.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.2</td>
<td>-1.3</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>0.9</td>
<td>-0.2</td>
<td>0.7</td>
<td>0.7</td>
<td>-0.4</td>
<td>0.6</td>
<td>0.1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Contraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4Nₑ = 0</td>
<td>&lt; 10%</td>
<td>-0.5</td>
<td>-4.5</td>
<td>-1.6</td>
<td>-1.8</td>
<td>3.0</td>
<td>-2.5</td>
<td>2.3</td>
<td>3.5</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>-5.2</td>
<td>-9.6</td>
<td>-6.8</td>
<td>-7.3</td>
<td>7.3</td>
<td>-9.0</td>
<td>10.3</td>
<td>8.6</td>
<td>-6.8</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>4.1</td>
<td>5.0</td>
<td>3.6</td>
<td>4.2</td>
<td>-3.0</td>
<td>4.5</td>
<td>-3.7</td>
<td>-3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>4Nₑ = 10</td>
<td>&lt; 10%</td>
<td>0.0</td>
<td>1.7</td>
<td>-0.3</td>
<td>-0.2</td>
<td>-0.5</td>
<td>0.0</td>
<td>-1.1</td>
<td>-1.1</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>0.3</td>
<td>3.2</td>
<td>-0.4</td>
<td>0.1</td>
<td>0.6</td>
<td>1.8</td>
<td>-2.9</td>
<td>-2.7</td>
<td>0.6</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>-0.1</td>
<td>-1.8</td>
<td>-0.1</td>
<td>-0.4</td>
<td>0.8</td>
<td>-0.6</td>
<td>1.3</td>
<td>1.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>Bottleneck</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4Nₑ = 0</td>
<td>&lt; 10%</td>
<td>-0.5</td>
<td>-3.8</td>
<td>-1.2</td>
<td>-2.4</td>
<td>3.8</td>
<td>0.0</td>
<td>3.4</td>
<td>4.0</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>-9.0</td>
<td>-12.6</td>
<td>-9.6</td>
<td>-10.1</td>
<td>10.2</td>
<td>-12.5</td>
<td>12.3</td>
<td>10.7</td>
<td>-11.4</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>4.4</td>
<td>4.7</td>
<td>3.6</td>
<td>4.6</td>
<td>-5.1</td>
<td>4.5</td>
<td>-3.1</td>
<td>-2.7</td>
<td>4.5</td>
</tr>
<tr>
<td>4Nₑ = 10</td>
<td>&lt; 10%</td>
<td>-2.3</td>
<td>-3.6</td>
<td>-2.6</td>
<td>-3.7</td>
<td>1.9</td>
<td>0.0</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>-7.4</td>
<td>-6.6</td>
<td>-6.4</td>
<td>-6.9</td>
<td>5.8</td>
<td>-6.9</td>
<td>4.5</td>
<td>4.3</td>
<td>-7.6</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>3.0</td>
<td>2.8</td>
<td>1.5</td>
<td>2.7</td>
<td>-4.4</td>
<td>2.2</td>
<td>-1.1</td>
<td>-1.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Hitchhiking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4Nₑ = 0</td>
<td>&lt; 10%</td>
<td>0.0</td>
<td>8.0</td>
<td>5.7</td>
<td>5.8</td>
<td>-1.6</td>
<td>5.8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>1.5</td>
<td>3.3</td>
<td>2.1</td>
<td>1.5</td>
<td>-4.5</td>
<td>3.2</td>
<td>-6.0</td>
<td>-5.8</td>
<td>2.4</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>0.8</td>
<td>-0.2</td>
<td>-0.1</td>
<td>0.3</td>
<td>6.3</td>
<td>0.0</td>
<td>7.4</td>
<td>7.0</td>
<td>0.1</td>
</tr>
<tr>
<td>4Nₑ = 10</td>
<td>&lt; 10%</td>
<td>0.0</td>
<td>7.5</td>
<td>6.3</td>
<td>5.6</td>
<td>-2.5</td>
<td>6.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>&lt; 50%</td>
<td>-1.6</td>
<td>5.6</td>
<td>1.6</td>
<td>0.7</td>
<td>-8.3</td>
<td>9.1</td>
<td>-9.4</td>
<td>-9.0</td>
<td>7.2</td>
</tr>
<tr>
<td>&gt; 90%</td>
<td>0.5</td>
<td>-1.2</td>
<td>-1.5</td>
<td>-0.6</td>
<td>6.4</td>
<td>0.0</td>
<td>7.7</td>
<td>7.3</td>
<td>-0.2</td>
</tr>
</tbody>
</table>

As in Table 1.

are observed under a panmictic neutral model with no recombination. Differences in the distributions of Lₑ are small but apparent, with longer and slightly broader distributions for the procedure FS₀_U than for FS. It is noteworthy that this procedure samples shorter trees with a lower variance of Lₑ than the FS procedure (Table 3). When we analysed the shape of the coalescent trees (Table 4), the coalescent trees obtained by the FS₀_U procedure are skewed towards shorter trees with smaller internal branches than trees sampled with the FS procedure.

Figure 6 shows the Lₑ distribution for four of the alternative models analysed with the FS and FS₀_U procedures. There are strong differences in the distribution of Lₑ for the FS procedure in comparison with FS₀_U for most of the alternative models. In particular, in the case of subdivided populations, large differences are observed for the parameter values studied. Bottleneck and hitchhiking models show also apparent differences in the trees that are used in FS in relation to FS₀_U. On the other hand, models of population size expansion exhibit smaller differences among procedures. In summary, differences in the length (and in the topology) of the trees explain the inaccuracy of the FS procedure and these differences are large in most of the alternative models studied.

(ii) The effect of recombination

Analysing the effect of recombination is also important, as the true value of the recombination parameter is generally unknown. Recombination has an important effect on the distribution of segregating sites because it breaks up the correlation among
contiguous positions. Differences in \( L_n \) are observed when recombination is added to the model (Fig. 7). Average values of the parameter \( L_n \) are constant under the \( FS \) procedure, as expected. Indeed, the \( FS \) procedure uses all output trees, like the standard coalescent procedure uses a fixed \( \theta \) value. The expected average value of \( L_n \) for the \( FS \) procedure value is given by equation (A3) in Appendix 1. Therefore, for \( n = 20 \), \( \mathbb{E}(L_n) = 3.548 \). On the other hand, the procedure \( FS\theta_U \) has lower average \( L_n \) values for zero recombination (\( L_n = 3.16 \) for \( n = 20 \)), and this value increases for larger recombination values. The procedure \( FS\theta \) leads to a similar pattern as that shown for \( FS\theta_U \), but with fewer differences relative to \( FS \).

The means and standard deviations of \( L_n \) in the limiting case of free recombination between fragments are given in Table 3. Recombination tends to bring the posterior distributions closer to the prior distribution of \( L_n \). Whereas the difference between the \( FS \) and \( FS\theta_U \) simulation procedures remains clear, the outcomes of the \( FS \) and \( FS\theta_U \) simulation procedures can hardly be distinguished (independent of the assumed prior for \( \theta \)), confirming the observation of Fig. 7.

The consequences of having differences in the average \( L_n \) given no recombination might be important, because it indicates that the zero-recombination neutral panmictic model may have an average of a given statistic that is different from that of a model with recombination, and therefore may not be conservative. Zero recombination in a neutral panmictic model leads to the largest deviation in average \( L_n \) in relation to the \( FS \) procedure. We have observed that for single loci this difference can be tolerated, but becomes too large for multilocus analyses. That is, for multilocus analyses, the \( FS \) procedure should not be used unless recombination is quite high for each locus.

The prior distribution of the mutational parameter \( \theta \) is shown to have an important effect on the posterior distribution of coalescent trees (Fig. 5), although, as previously pointed out by Wall & Hudson (2001), the recombination parameter also has a strong effect on this distribution. The improvement obtained from using sophisticated techniques such as the \( FS\theta_U \) procedure may, however, be negligible in comparison with the errors caused by the uncertainty in the recombination parameter \( R \) (Wall & Hudson, 2001). A procedure considering also the uncertainty in the recombination parameter would give more appropriate distributions, but is beyond the scope of this paper.

Appendix 1

(i) The branch lengths in a coalescent tree

In the standard neutral coalescent process without recombination, the waiting time \( T_n \) (in units of \( 4N \) generations) until two lineages among \( n \) have a common ancestor is exponentially distributed with parameter \( n(n-1) \) (Kingman, 1982b). The probability density of \( T_n \) is thus

\[
    f_{T_n}(t) = n(n-1)e^{-n(n-1)t}.
\]  

(A1)

The density of the sum \( L_n \) of the lengths of the branches of a coalescent tree with \( n \) tips is given by Tavaré (1984, unlabelled equation at the top of p. 153) as

\[
    f_L(t) = (n-1)e^{-t(1-e^{-t})^{n-2}}.
\]  

(A2)
no recombination. Results based on the $FS$ arbitrarily set to $20/a$

As shown in Appendix 2, the density of $Y_n$ is

$$f_{Y_n}(t) = (n-1)(n-2)e^{-2t}(1-e^{-t})^{n-3}.$$  \hfill (A6)

The mean and variance of $L_n$ are

$$E(L_n) = a_1(n) = \sum_{i=1}^{n-1} \frac{1}{i},$$ \hfill (A3)

$$V(L_n) = a_2(n) = \sum_{i=1}^{n-1} \frac{1}{i^2}. $$ \hfill (A4)

In order to characterize the shape of a coalescent tree, we introduce a shape index $X$ where $X$ is the length of the two branches from the second-last coalescent to the last (in the upper part of the coalescent tree) and $Y_n$ is the length of the rest (the branches in the lower part) of the coalescent tree (see Fig. 1). The density of $X$ follows from equation (A1):

$$f_X(t) = e^{-t}. $$ \hfill (A5)

As shown in Appendix 2, the density of $Y_n$ is

$$f_{Y_n}(t) = (n-1)(n-2)e^{-2t}(1-e^{-t})^{n-3}. $$ \hfill (A6)

Table 3. Mean and standard deviation (in parentheses) of the posterior distribution of $L_n$

<table>
<thead>
<tr>
<th>Simulation procedure</th>
<th>$R = 0$</th>
<th>$R = \infty^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$FS$</td>
<td>3.548 (1.262)</td>
<td>3.548 (0.179)</td>
</tr>
<tr>
<td>$FS \theta^b$</td>
<td>$S = 10$</td>
<td>2.358 (0.497)</td>
</tr>
<tr>
<td></td>
<td>$S = 20$</td>
<td>3.493 (0.657)</td>
</tr>
<tr>
<td></td>
<td>$S = 30$</td>
<td>4.800 (0.814)</td>
</tr>
<tr>
<td>$FS \theta_U$</td>
<td>$g_{Uniform}$</td>
<td>3.158 (1.110)</td>
</tr>
</tbody>
</table>

$^a$ Assuming $m = 50$ recombining fragments (equation 3).

$^b$ Assuming $\theta = 20/a_1(20)$.

(ii) Branch lengths in the procedures $F\theta$ and $FS$

In the standard procedure $F\theta$, but also in the $FS$ simulation procedures, all simulated coalescent trees are used and thus the posterior density of tree length is the same as the prior density $f_{L_n}$ (equation A2).

(iii) Branch lengths in the $FS\theta$ and $FS\theta_{prior}$ procedures

In the $FS\theta$ simulation procedure, trees are sampled from the prior distribution according to the probability of observing exactly $S$ mutations given a known mutational parameter $\theta$. The number of mutations in the coalescent tree follows a Poisson distribution with parameter $tL_n$. Thus the probability of observing $S$ mutations in a sample of $n$ lines given $\theta$ is (Tavare’, 1984, unlabelled equation on p. 153)

$$\mathbf{P}(S=k|n, \theta) = \int_{0}^{\infty} \frac{e^{-\theta t}}{k!} f_{L_n}(t) dt,$$

and the posterior density of $L_n$ given $S$ and $\theta$ follows from the definition of the posterior density:

$$f_{L_n|S, \theta}(t) = \frac{(\theta t)^k e^{-\theta t}}{k!} \mathbf{P}(S=k|n, \theta) f_{L_n}(t).$$ \hfill (A7)

In the $FS\theta$ procedures that consider the uncertainty of the value of $\theta$ ($FS\theta_{prior}$), a prior distribution of $\theta$ with density $g$ is assumed. Thus the probability of observing $S$ mutations in a sample of $n$ lines is

$$\mathbf{P}(S=k|n, g) = \int_{0}^{\infty} \int_{0}^{\theta_{max}} \frac{(\theta t)^k e^{-\theta t}}{k!} f_{L_n}(t) g(\theta) d\theta dt,$$
and the posterior density of \( L_n \) given \( S \) and the prior density \( g \) is

\[
f_{L_n|S,g}(t) = \frac{f_{L_n}(t)}{P(S=k|n,g)} \times \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \frac{\theta^k e^{-\theta t}}{k!} g(\theta) d\theta.
\]

(A8)

If a uniform density \( g_u \) is assumed for \( \theta, g_n \) is constant (equation 2) and the posterior density in equation (A8) becomes

\[
f_{L_n|S,g}(t) = \frac{f_{L_n}(t)}{\int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \frac{\theta^k e^{-\theta t}}{k!} g(\theta) d\theta} \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \frac{\theta^k e^{-\theta t}}{k!} f_{L_n}(t) d\theta d\theta.
\]

Using \( \theta_{\text{min}} = 0 \), we see that for \( n \geq 3 \) a limit of the posterior distribution of \( L_n \) exists when \( \theta_{\text{max}} \) is very large (this condition is necessary for the integral in the

Table 4. Mean and standard deviation (in parentheses) of the tree shape

<table>
<thead>
<tr>
<th>Simulation procedure</th>
<th>( X^\alpha )</th>
<th>( Y_n^\alpha )</th>
<th>( X/Y_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td>1.000 (1.000)</td>
<td>2.548 (0.770)</td>
<td>0.429 (0.468)</td>
</tr>
<tr>
<td>FS( \theta^b )</td>
<td>( S=10 )</td>
<td>0.388 (0.345)</td>
<td>1.970 (0.447)</td>
</tr>
<tr>
<td></td>
<td>( S=20 )</td>
<td>0.879 (0.655)</td>
<td>2.613 (0.622)</td>
</tr>
<tr>
<td></td>
<td>( S=30 )</td>
<td>1.725 (1.020)</td>
<td>3.075 (0.839)</td>
</tr>
<tr>
<td>FS( \theta_U )</td>
<td>( g_{\text{Uniform}} )</td>
<td>0.782 (0.815)</td>
<td>2.376 (0.719)</td>
</tr>
</tbody>
</table>

\( a \) See Figure 1. Assuming no recombination and \( n = 20 \).

\( b \) Assuming \( \theta = 20/n_0(20) \).
denominator of equation A9 to converge):

$$\lim_{n \to \infty} f_{L_n|S_d}(t) = \frac{t^{-1}f_{X}(t)}{\int_0^x t^{-1}f_{X}(t)\,dt}. \quad (A9)$$

It is noteworthy that this distribution is independent of the observed number of mutations $S$ and may be denoted as $f_{L_n|S}$. 

**Appendix 2. Proof of equation (A6)**

(i) **Notation and preliminary results**

In this section the density function of a random variable $X$ will be denoted as $f_X$. We introduce a new random variable $P_k = kT_k$ (see equation A1). It is straightforward to show that $P_k$ is exponentially distributed with parameter $k-1$, thus

$$f_{P_k}(t) = (k-1)e^{-(k-1)t}, \quad (A10)$$

and $Y_n$ is defined by $Y_n = \sum_{k=2}^{n} P_k$ (see Fig. 1). $Y_n$ is the sum of $n-2$ independent exponentially distributed random variables with parameters 2, ..., $(n-1)$. As we show below, the density function of $Y_n$ can be obtained using simple order statistics.

The following properties of the exponential distribution will be used:

- Minimum of independent exponentially distributed random variables: Consider $k$ independent random variables $X_k$, exponentially distributed with parameters $\lambda_i$ $(i = 1, \ldots, k)$. Then $\min(X_k)$ is exponentially distributed with parameter $\lambda = \sum_{i=1}^{k} \lambda_i$.

- Memoryless property of the exponential distribution: Consider a random variable $X_\lambda$ exponentially distributed with parameter $\lambda$:

$$\forall x \geq 0, \quad P(X_\lambda - x \leq t | X_\lambda \geq x) = 1 - e^{-\lambda t} = P(X_\lambda \leq t).$$

(ii) **Density function of $Y_n$**

We consider $n-1$ independent random variables $E_i$ $(i = 1, \ldots, n-1)$, exponentially distributed with parameter 1, and denote the smallest one as $E_{(1)}$, $E_{(1)}$ is exponentially distributed with parameter $n-1$, and the $n-2$ remaining random variables are independent and such that $E_1 - E_{(1)}$ is exponentially distributed with parameter 1. Then the difference between the second and the first smallest random variables $E_{(2)} - E_{(1)}$ is exponentially distributed with parameter $n-2$ and the $n-3$ remaining random variables are independent and such that $E_1 - E_{(2)}$ is exponentially distributed with parameter 1. If we define $E_0 = 0$, it is straightforward to show that the difference between two successive sorted random variables $(E_k - E_{(k-1)})$ is exponentially distributed with parameter $n-k$.

$E_{(n-2)}$ can be rewritten as $E_{(n-2)} = \sum_{k=2}^{n-2} (E_k - E_{(k-1)})$. This is the sum of $n-2$ independent exponentially distributed random variables with parameters 2, ..., $(n-1)$, thus $E_{(n-2)}$ and $Y_n$ have the same distribution. Because $E_{(n-2)}$ is the $(n-2)^{th}$ random variable among the $n-1$ random variables $E_i$ sorted in increasing order, the density function of $E_{(n-2)}$ and $Y_n$ is (Pitman, 1992, p. 326):

$$f_{Y_n}(t) = (n-1)(n-2)e^{-(n-2)(1-e^{-t})^{n-3}}. \quad (A11)$$

We thank J. M. Comeron, J. Rozas, K. Schmid, J. Zavala and A. Navarro-Quezada for helpful comments on the manuscript. S. R. O. thanks S. Gebauer-Jung and S. Wright for help with the calculation of Hudson’s recursive equation, J. Fay for supplying information to check the selective model and Y. Kim for his assistance in the analysis of the selective model. Special thanks go to two anonymous reviewers for their input into this manuscript. This research was supported by the Max Planck Society, a Marie Curie fellowship from the EU (MCFI 2002 01461) to S.M., and grant Ste 325/5 from the DFG to W.S.

**References**


