Genet. Res., Camb. (2007), **89**, *pp.* 231–244. © 2007 Cambridge University Press
doi:10.1017/S0016672307008877 Printed in the United Kingdom

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

SEBASTIÁN E. RAMOS-ONSINS[1]*†, SYLVAIN MOUSSET[2]‡, THOMAS MITCHELL-OLDS[1]§ AND WOLFGANG STEPHAN[2]
[1] *Max Planck Institute of Chemical Ecology, Hans-Knöll-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 θ, 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 θ value (or alternatively, procedures that condition on *S* and take into account the uncertainty of θ) 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 θ (where $\theta = 4N\mu$, N being the effective population size and μ 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 θ (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 θ (Hudson, 1993). However, using this approach, population samples with short genealogical trees tend to exhibit high

* Corresponding author. e-mail: sramosonsins@ub.edu

† Present address: Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain.

‡ Present address: Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, 43 boulevard du novembre 1918, Villeurbanne 69622, France.

§ Present address: Department of Biology, Duke University, Durham, NC 27708, USA.

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

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 θ. A procedure that conditions on S (instead of fixing S) considers for a given θ all possible trees weighted by their probabilities of giving rise to the observed number of segregating sites (Tavaré *et al.*, 1997; Markovtsova *et al.*, 2000, 2001; Jakobsson *et al.*, 2006), and thus takes into account the uncertainty of θ (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) θ value, and (ii) as an approximation of a rigorous procedure that conditions on S and takes into account the uncertainty of θ.

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, 1982*a*, *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 θ value is well supported by the data (Depaulis *et al.*, 2001; Markovtsova *et al.*, 2001; Wall & Hudson, 2001). Thus, the probability of obtaining a θ 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:

$$\left.\begin{array}{ll} N(t)=N_0 & \text{if } t\leqslant t_0, \\ N(t)=N_0+\dfrac{N_1-N_0}{1+e^{-\gamma\left(t-t_0-\left(\frac{t_1-t_0}{2}\right)\right)}} & \text{if } t_0<t<t_1, \\ N(t)=N_1 & \text{if } t\geqslant t_1, \end{array}\right\} \tag{1}$$

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 $t_1-t_0=0{\cdot}1\,N_0$ 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_f). 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 $t_f = 0$, $4Ns = 2 \times 10^4$, $N = 10^6$ and $\varepsilon = 1/4N$.

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

(*b*) The *FS* 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 *Fθ* procedure.

(*c*) Procedure based on fixing the number of segregating sites and the value of *θ* (*FSθ*): 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 *θ* ($FS\theta_{prior}$ procedure): We use the *RAU* procedure (**R**ejection **A**lgorithm using a **U**niform prior). This procedure employs the rejection algorithm #2 described by Tavaré *et al.* (1997) but sampling of *θ* is done from a given prior distribution instead of having a fixed *θ* value.

(iii) *Mutational parameter* θ

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

$$g_u(\theta) = \frac{1}{\theta_{\max} - \theta_{\min}}. \qquad (2)$$

When we do not have information about the value of *θ*, the assumption of a uniform density for any value of *θ* is reasonable. This is a commonly used strategy to estimate *θ* 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 *θ*, 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 *θ*, we used as a minimum bound of *θ* 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 *TD*, Fu & Li's *D* and *F* (Fu & Li, 1993), here named *FD* and *FF*, 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 *R2* (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 *θ* 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* Fθ

We have used the approach described in Wall & Hudson (2001) to determine the level of accuracy of the *FS* procedure compared with the *Fθ* 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' *θ* value for the same evolutionary model is obtained, which is the value that gives an average of $S = 20$ in 10 000 iterations

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

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 θ. 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' θ 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 θ (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 *procedure in relation to* $FS\theta_U$

Here we assume that the θ value is unknown. The $FS\theta_U$ 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 $FS\theta_U$ 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 $FS\theta_U$ 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 θ (given the neutral equilibrium model) is then approximately 0·005 (which has been used in a number of the theoretical population genetics studies). All θ 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_k, then $L_n=\sum_{k=2}^{n} kT_k$ 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 $\overline{L_n}$ of the m independent coalescent trees will converge towards the normal distribution with mean $a_1(n)$ and variance $a_2(n)/m$ (defined in equations (A3) and (A4) of Appendix 1, respectively). This is,

$$\lim_{R\to\infty} f_{\overline{L_n}}(t)=\sqrt{\frac{m}{2\pi a_2(n)}}e^{-\frac{m}{2a_2(n)}(t-a_1(n))^2}, \tag{3}$$

where $f_{\overline{L_n}}(t)$ is the density of the average length $\overline{L_n}$. This last equation can be used in equations (A7)–(A8) (Appendix 1) to address the effects of the simulation procedures $F\theta$, FS, $FS\theta$, and $FS\theta_{prior}$ 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_n (Fig. 1). The mean value of this ratio is given by

$$E\left(\frac{X}{Y_n}|n,S\right)=\frac{1}{\eta(n,S)} \times \int_0^{\infty}\int_0^{\infty}\frac{x}{y}f_X(x)f_{Y_n}(y)\mathbf{P}_{\mathbf{accept}}(x,y)\mathrm{d}x\mathrm{d}y, \tag{4}$$

where f_X and f_{Y_n} denote the prior densities of X and Y_n (see equations (A5) and (A6)) and

$$\mathbf{P}_{\mathbf{accept(x,y)}}=\begin{cases}1 & (FS \text{ procedure})\\ \frac{\theta^S(x+y)^S}{S!}e^{-(x+y)\theta} & (FS\theta \text{ procedure})\\ \frac{(x+y)^S}{S!}\int_0^{\infty}e^{-(x+y)\theta}g(\theta)\mathrm{d}\theta & (FS\theta_{prior} \text{ procedure})\end{cases}$$

and $\eta(n,S)=\int_0^{\infty}\int_0^{\infty}f_X(x)f_{Y_n}(y)\mathbf{P}_{\mathbf{accept}}(x,y)\mathrm{d}x\mathrm{d}y$ is a normalizing constant.

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

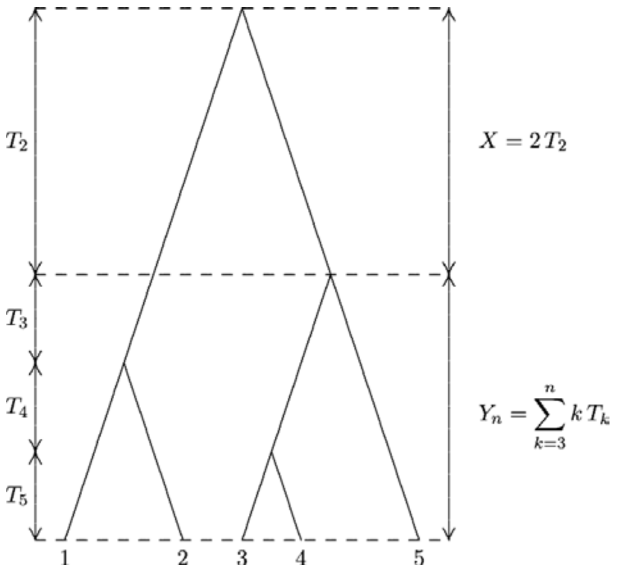

Fig. 1. Shape index of a coalescent tree. The shape index is the ratio of the branch lengths in the upper and lower parts of the tree, X/Y_n (see text).

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 θ; the *FS* procedure) in comparison with the two rigorous procedures: (i) the standard procedure fixing θ ($F\theta$) and (ii) procedures that while conditioning on *S* take into account the uncertainty of θ ($FS\theta_{prior}$, here using a uniform prior, named $FS\theta_U$). 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* Fθ

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 $\theta = 0{\cdot}0057$ for the $F\theta$ procedure; i.e. the estimate of θ for $S = 20$). Fig. 2*A* 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 observed 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. 2*B* 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* θ

We have examined the accuracy of *FS* (the nominal procedure) by comparison with the procedure $FS\theta_U$ (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 $FS\theta_U$ 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*

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

Table 1. *Difference* (*in percentage*) *between the* FS *and* Fθ *procedures for several critical values*[a]

$n=20$		*TD*	*Fs*	*FD*	*FF*	*H*	*B*	*Kw*	*Hw*	*R2*
Neutral										
$4Nr=0$	<2·5%	−0·08	0·23	0·86	0·06	−0·14	−0·01	−0·55	−0·17	0·25
	>97·5%	−0·15	−0·11	0·97	−0·21	0·13	0·15	0·12	0·70	−0·18
$4Nr=10$	<2·5%	0·08	0·18	0·71	0·14	0·06	−0·04	0·22	−0·25	0·05
	>97·5%	0·05	−0·10	1·32	−0·02	0·23	0·22	1·07	−0·33	0·05
Subdivision[b]										
$4Nr=0$	<10%	0·5	6·1	2·9	0·7	−0·4	−0·4	−2·0	−2·6	3·4
	<50%	4·3	13·0	13·3	7·8	−0·1	15·8	2·5	−10·5	11·0
	>90%	−0·8	−2·3	8·0	−1·2	4·9	−7·2	3·2	5·8	−0·4
$4Nr=10$	<10%	0·4	4·3	2·8	0·6	0·0	4·4	2·3	−0·4	1·0
	<50%	1·1	5·4	7·7	1·3	−0·7	7·3	−0·4	−2·9	2·1
	>90%	0·2	−1·4	9·2	−0·4	2·5	2·3	4·8	4·1	0·1
Expansion[c]										
$4Nr=0$	<10%	0·1	0·1	2·4	0·2	−0·4	0·0	−2·0	−0·2	0·4
	<50%	0·7	0·4	6·6	0·8	−0·1	16·1	−2·7	−0·1	0·9
	>90%	−0·5	−0·2	1·6	−0·7	0·3	0·6	1·1	1·4	−0·6
$4Nr=10$	<10%	0·5	0·2	2·8	0·4	−0·2	0·0	2·8	−0·7	0·6
	<50%	1·1	0·6	7·3	1·2	−0·7	16·7	9·6	−2·0	1·1
	>90%	−0·3	−0·3	1·9	−0·5	0·5	2·2	−1·2	1·3	−0·4
Contraction[d]										
$4Nr=0$	<10%	−0·2	0·0	2·5	−0·1	−0·3	−2·1	−0·3	−0·2	0·3
	<50%	1·0	0·8	7·0	1·5	0·1	12·0	−7·7	0·5	1·1
	>90%	−0·4	0·0	2·8	−0·7	0·6	0·5	−0·4	0·4	−0·4
$4Nr=10$	<10%	−0·1	0·1	3·0	−0·1	−0·1	−0·4	−0·5	−0·4	−0·2
	<50%	0·1	−0·1	7·5	0·0	0·1	16·8	4·3	5·6	0·1
	>90%	0·0	0·0	4·2	0·0	0·1	4·0	1·3	0·0	−0·1
Bottleneck[e]										
$4Nr=0$	<10%	−0·2	0·1	2·9	0·4	−0·3	−0·3	−1·5	−0·3	0·7
	<50%	1·1	0·2	7·1	1·6	−0·3	10·7	2·4	−1·0	1·6
	>90%	−0·2	0·1	1·8	−0·4	1·6	0·6	−1·0	0·9	−0·2
$4Nr=10$	<10%	3·1	−2·7	5·4	3·6	0·5	−0·3	4·7	6·6	2·7
	<50%	8·4	−13·3	12·8	7·4	−13·5	17·3	14·4	15·4	6·3
	>90%	2·8	8·3	4·2	1·9	4·8	5·9	−6·1	−5·0	3·0
Hitchhiking[f]										
$4Nr=0$	<10%	10·0	8·5	8·1	7·6	1·4	7·9	5·2	3·1	7·0
	<50%	9·3	12·8	13·3	8·5	1·9	11·4	−5·7	−0·3	8·8
	>90%	0·1	0·5	3·6	0·6	3·9	−1·6	6·4	6·7	−2·0
$4Nr=10$	<10%	10·0	9·0	9·2	8·7	1·3	8·8	5·2	3·4	7·8
	<50%	11·8	12·8	18·4	13·9	−0·2	11·1	−12·3	0·3	1·1
	>90%	0·0	1·3	4·2	1·9	2·4	−0·6	8·9	7·2	−1·2

[a] 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, $4Nm=0{\cdot}5$ and $4Nr=10$. The samples are all obtained from a single population. In the $FS\theta_U$ procedure, the bounds of θ were reduced d times for a better comparison with the other models. In the $F\theta$ procedure, $\theta=0{\cdot}00062$.

[c] Logistic expansion model with a 10-fold growth of population size. The expansion process started $0{\cdot}4N_0$ generations ago and finished at present. $4N_0r=10$. In the $F\theta$ procedure, $\theta=0{\cdot}03384$.

[d] Logistic contraction model: In contrast to the logistic expansion model, the population is reduced 10 fold. In the $F\theta$ procedure, $\theta=0{\cdot}00059$.

[e] Bottleneck model: Population size is maintained constant during $0{\cdot}5N_0$ generations since present. It follows a logistic, 100-fold reduction of population size for $0{\cdot}01N_0$ generations, maintained for $0{\cdot}01N_0$ generations, then an instantaneous recovery to the present population size. $4N_0r=10$. In the $F\theta$ procedure, $\theta=0{\cdot}00860$.

[f] Hitchhiking model: The selective phase is completed at present time ($t_f=0$). $N=10^6$, $4Ns=10^4$, the population recombination parameter between the selected and observed locus is $4Nc=400$, $c/s=0{\cdot}02$, and intragenic recombination is indicated in the table. In the $F\theta$ procedure, $\theta=0{\cdot}00127$.

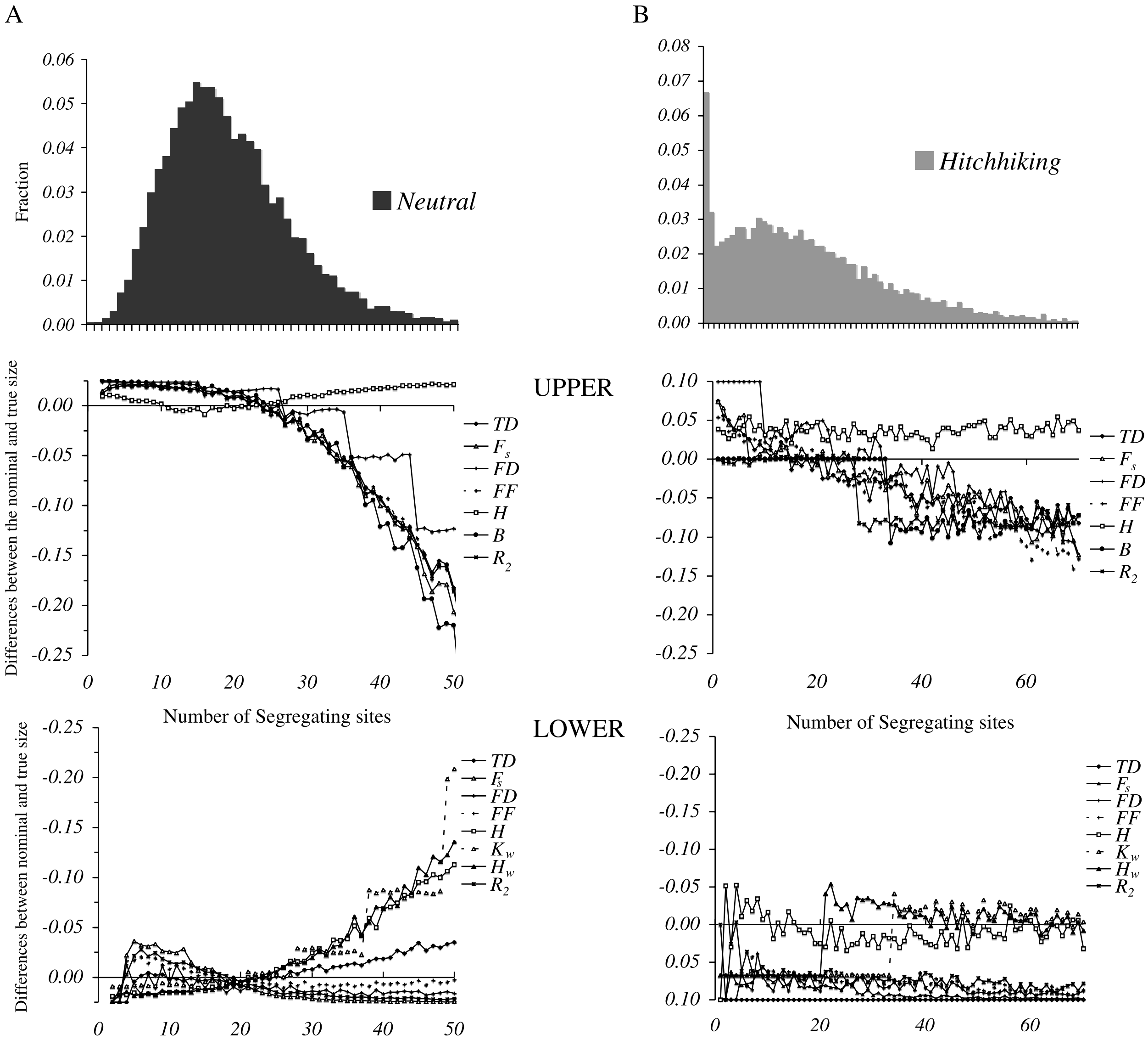

Fig. 2. Differences between the sizes of *FS* and *Fθ* procedures for each *S* separately. (*A*) Neutral model: In the upper panel the distribution of *S* values obtained with the *Fθ* procedure using $\theta = 0{\cdot}0057$ is shown. In the middle and lower panels the size differences for statistics 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 *Fθ* procedure using $\theta = 0{\cdot}0127$ 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.

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

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

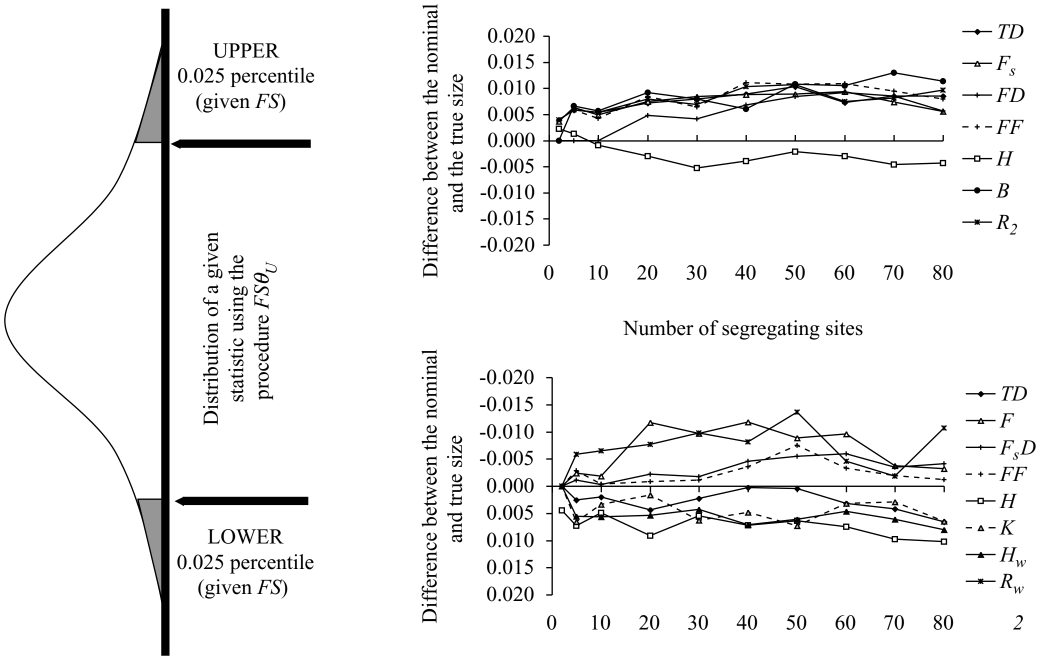

Fig. 3. Differences between the sizes of the *FS* and $FS\theta_U$ procedures. Critical values are obtained for each of the nine neutrality tests for the 2·5% upper and lower tails. Abbreviations are indicated in Section 2.

$FS\theta_U$ become very important when multilocus studies are performed. Some statistics, like K_w, H_w and Fay & Wu's *H*, are extremely conservative for the lower tail, as are most statistics (except for Fay & Wu's *H*) for the upper tail (see Fig. 4*A*). On the other hand, statistics such as Tajima's *D*, Fu's F_S, Fu & Li's *D* and *F*, and Ramos-Onsins & Rozas' *R2* 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 θ value ($F\theta$) and with a procedure that conditions on *S* taking into account the uncertainty of θ ($FS\theta_U$). 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* Fθ *or* $\mathrm{FS\theta_U}$ *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 $FS\theta$ procedures can explain the inaccuracy of *FS*, given that the critical values of *FS* are compared with the $FS\theta$ 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 θ (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 θ, when *S* is lower than the average for that θ, 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 θ. Thus, these results show how the simulation procedure 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 $FS\theta_U$

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

Table 2. *Difference* (*in percentage*) *between the* FS *and the* FSθ_U *procedures for several critical values*[a]

n = *20*		*TD*	F_s	*FD*	*FF*	*H*	*B*	K_w	H_w	*R2*
Subdivision										
4Nr = *0*	< 10 %	0·7	2·8	0·6	0·3	1·2	2·4	2·6	3·4	2·6
	< 50 %	1·2	−10·7	−1·7	−2·2	0·4	−5·5	14·1	11·6	−1·2
	> 90 %	2·1	3·4	0·0	−1·9	2·4	3·1	3·3	2·6	2·2
4Nr = *10*	< 10 %	−0·5	−2·5	0·1	0·0	0·5	2·1	0·5	1·2	−0·5
	< 50 %	0·1	−2·2	−0·6	−0·4	2·2	7·6	2·4	2·0	−0·4
	> 90 %	−1·2	−0·2	0·0	0·3	−1·3	−1·1	−1·0	−1·1	−1·0
Expansion										
4Nr = *0*	< 10 %	0·6	0·3	0·1	0·4	−0·2	0·0	0·8	0·3	0·6
	< 50 %	0·0	0·0	−0·8	−0·8	0·4	−0·2	0·3	0·5	−1·2
	> 90 %	0·6	1·2	1·0	0·8	−0·4	1·5	0·2	−0·3	0·9
4Nr = *10*	< 10 %	0·0	0·1	0·1	0·0	0·1	0·0	−0·4	−0·1	−0·2
	< 50 %	−1·2	−0·8	−0·9	−0·6	1·5	0·0	0·5	0·2	−1·3
	> 90 %	0·9	−0·2	0·7	0·7	−0·4	0·6	0·1	0·0	1·0
Contraction										
4Nr = *0*	< 10 %	−0·5	−4·5	−1·6	−1·8	3·0	−2·5	2·3	3·5	−2·7
	< 50 %	−5·2	−9·6	−6·8	−7·3	7·3	−9·0	10·3	8·6	−6·8
	> 90 %	4·1	5·0	3·6	4·2	−3·0	4·5	−3·7	−3·6	4·4
4Nr = *10*	< 10 %	0·0	1·7	−0·3	−0·2	−0·5	0·0	−1·1	−1·1	0·0
	< 50 %	0·3	3·2	−0·4	0·1	−0·6	1·8	−2·9	−2·7	0·6
	> 90 %	−0·1	−1·8	−0·1	−0·4	0·8	−0·6	1·3	1·2	−0·2
Bottleneck										
4Nr = *0*	< 10 %	−0·5	−3·8	−1·2	−2·4	3·8	0·0	3·4	4·0	−2·9
	< 50 %	−9·0	−12·6	−9·6	−10·1	10·2	−12·5	12·3	10·7	−11·4
	> 90 %	4·4	4·7	3·6	4·6	−5·1	4·5	−3·1	−2·7	4·5
4Nr = *10*	< 10 %	−2·3	−3·6	−2·6	−3·7	1·9	0·0	1·8	1·8	−3·8
	< 50 %	−7·4	−6·6	−6·4	−6·9	5·8	−6·9	4·5	4·3	−7·6
	> 90 %	3·0	2·8	1·5	2·7	−4·4	2·2	−1·1	−1·0	3·0
Hitchhiking										
4Nr = *0*	< 10 %	0·0	8·0	5·7	5·8	−1·6	5·8	0·0	0·0	5·7
	< 50 %	1·5	3·3	2·1	1·5	−4·5	3·2	−6·0	−5·8	2·4
	> 90 %	0·8	−0·2	−0·1	0·3	6·3	0·0	7·4	7·0	0·1
4Nr = *10*	< 10 %	0·0	7·5	6·3	5·6	−2·5	6·5	0·0	0·0	6·6
	< 50 %	−1·6	5·6	1·6	0·7	−8·3	9·1	−9·4	−9·0	7·2
	< 90 %	0·5	−1·2	−1·5	−0·6	6·4	0·0	7·7	7·3	−0·2

[a] As in Table 1.

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

Figure 6 shows the L_n distribution for four of the alternative models analysed with the *FS* and $FS\theta_U$ procedures. There are strong differences in the distribution of L_n for the *FS* procedure in comparison with $FS\theta_U$ 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 $FS\theta_U$. 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

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

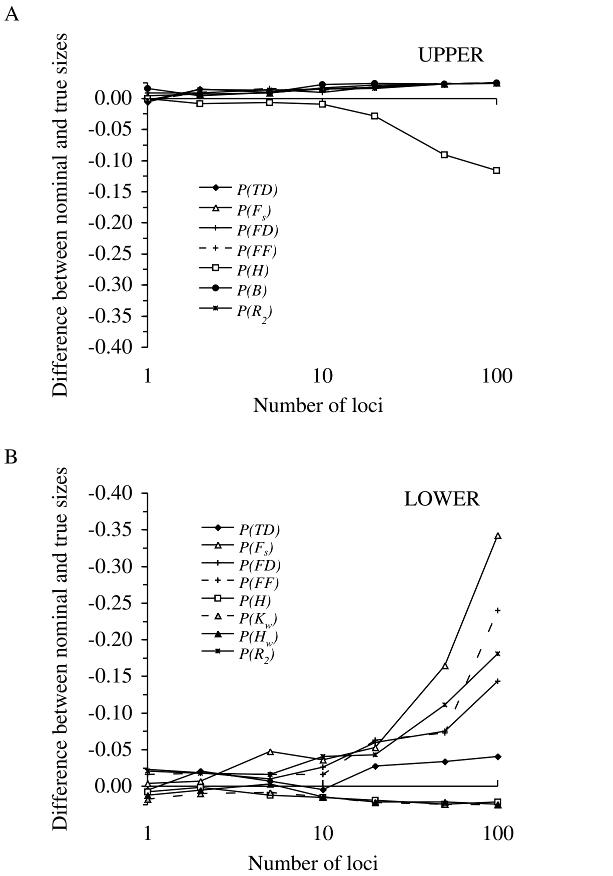

Fig. 4. Effect of the number of loci on the probability of rejecting the neutral panmictic model for nine neutrality tests. (*A*) Differences between the sizes of *FS* and $FS\theta_U$ procedures with no recombination in the upper 2·5% tail, given different numbers of loci. (*B*) Differences between the sizes of *FS* and $FS\theta_U$ procedures with no recombination in the lower 2·5% tail, given different numbers of loci. 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. *S* was fixed at 20 for each locus. Plots obtained using *S* values from a distribution compatible with $\theta = 0{\cdot}0057$ gave equivalent results (not shown) although for a small number of loci we observed a large variance.

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 θ value. The expected average value of L_n for the *FS* procedure value is given by equation (A3) in Appendix 1. Therefore, for $n=20$, $E(L_n)=3{\cdot}548$. On the other hand, the procedure $FS\theta_U$ has lower average L_n values for zero recombination (L_n is 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 θ), 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 θ 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 4*N* generations) until two lineages among *n* have a common ancestor is exponentially distributed with parameter $n\ (n-1)$ (Kingman, 1982*b*). The probability density of T_n is thus

$$f_{T_n}(t) = n(n-1)e^{-n(n-1)t}. \qquad \text{(A1)}$$

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

$$f_{L_n}(t) = (n-1)e^{-t}(1-e^{-t})^{n-2}. \qquad \text{(A2)}$$

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

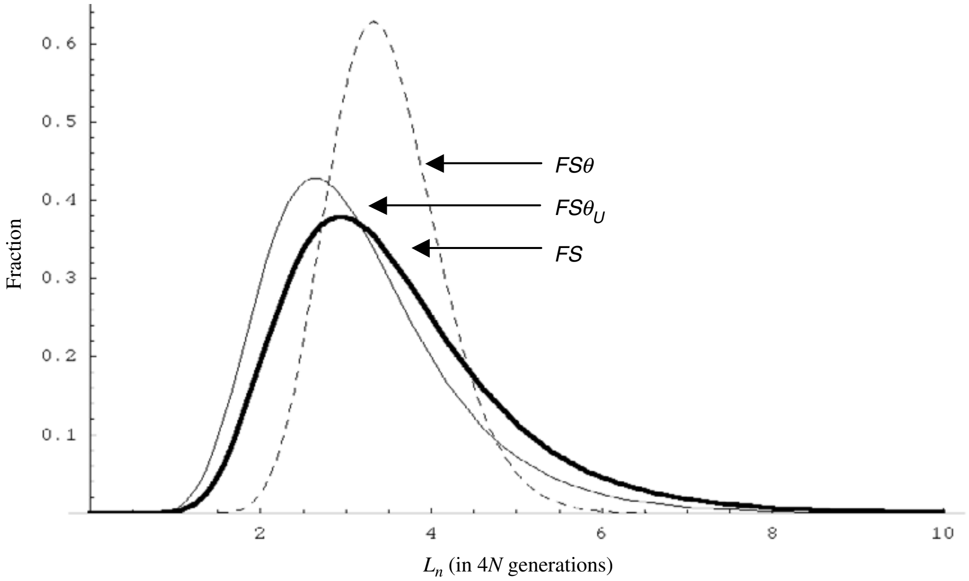

Fig. 5. Posterior densities of the total length L_n of a coalescent tree with $n=20$ are shown for the different procedures and no recombination. Results based on the *FS*, *FSθ* and $FS\theta_U$ procedures are displayed. For *FSθ* the θ parameter was arbitrarily set to $20/a_2(20)$.

Table 3. *Mean and standard deviation* (*in parentheses*) *of the posterior distribution of* L_{20}

Simulation procedure		$R=0$	$R=\infty^a$
FS		3·548 (1·262)	3·548 (0·179)
FS θ^b	$S=10$	2·358 (0·497)	3·460 (0·176)
	$S=20$	3·493 (0·657)	3·548 (0·174)
	$S=30$	4·800 (0·814)	3·632 (0·172)
FS θ_U	$g_{Uniform}$	3·158 (1·110)	3·539 (0·179)

[a] Assuming $m=50$ recombining fragments (equation 3).
[b] Assuming $\theta=20/a_1(20)$.

The mean and variance of L_n are

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

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

In order to characterize the shape of a coalescent tree, we introduce a shape index X/Y_n 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):

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

As shown in Appendix 2, the density of Y_n is

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

(ii) *Branch lengths in the procedures* Fθ *and* FS

In the standard procedure *Fθ*, 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* FSθ *and* $\mathrm{FS\theta_{prior}}$ *procedures*

In the *FSθ* simulation procedure, trees are sampled from the prior distribution according to the probability of observing exactly *S* mutations given a known mutational parameter θ. The number of mutations in the coalescent tree follows a Poisson distribution with parameter θL_n. Thus the probability of observing *S* mutations in a sample of *n* lines given θ is (Tavaré, 1984, unlabelled equation on p. 153)

$$\mathbf{P}(S=k|n,\theta)=\int_0^\infty \frac{(\theta t)^k e^{-\theta t}}{k!} f_{L_n}(t)\mathrm{d}t,$$

and the posterior density of L_n given *S* and θ follows from the definition of the posterior density:

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

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

$$\mathbf{P}(S=k|n,g)=\int_0^\infty \int_{\theta_{\min}}^{\theta_{\max}} \frac{(\theta t)^k e^{-\theta t}}{k!} f_{L_n}(t)g(\theta)\mathrm{d}\theta\mathrm{d}t,$$

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press

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

Simulation procedure		X^a	$Y_n{}^a$	X/Y_n
FS		1·000 (1·000)	2·548 (0·770)	0·429 (0·468)
$FS\theta^b$	$S=10$	0·388 (0·345)	1·970 (0·447)	0·219 (0·225)
	$S=20$	0·879 (0·655)	2·613 (0·622)	0·392 (0·368)
	$S=30$	1·725 (1·020)	3·075 (0·839)	0·675 (0·553)
$FS\theta_U$	$g_{Uniform}$	0·782 (0·815)	2·376 (0·719)	0·355 (0·395)

[a] See Figure 1. Assuming no recombination and $n=20$.
[b] Assuming $\theta=20/\mathrm{a}_1(20)$.

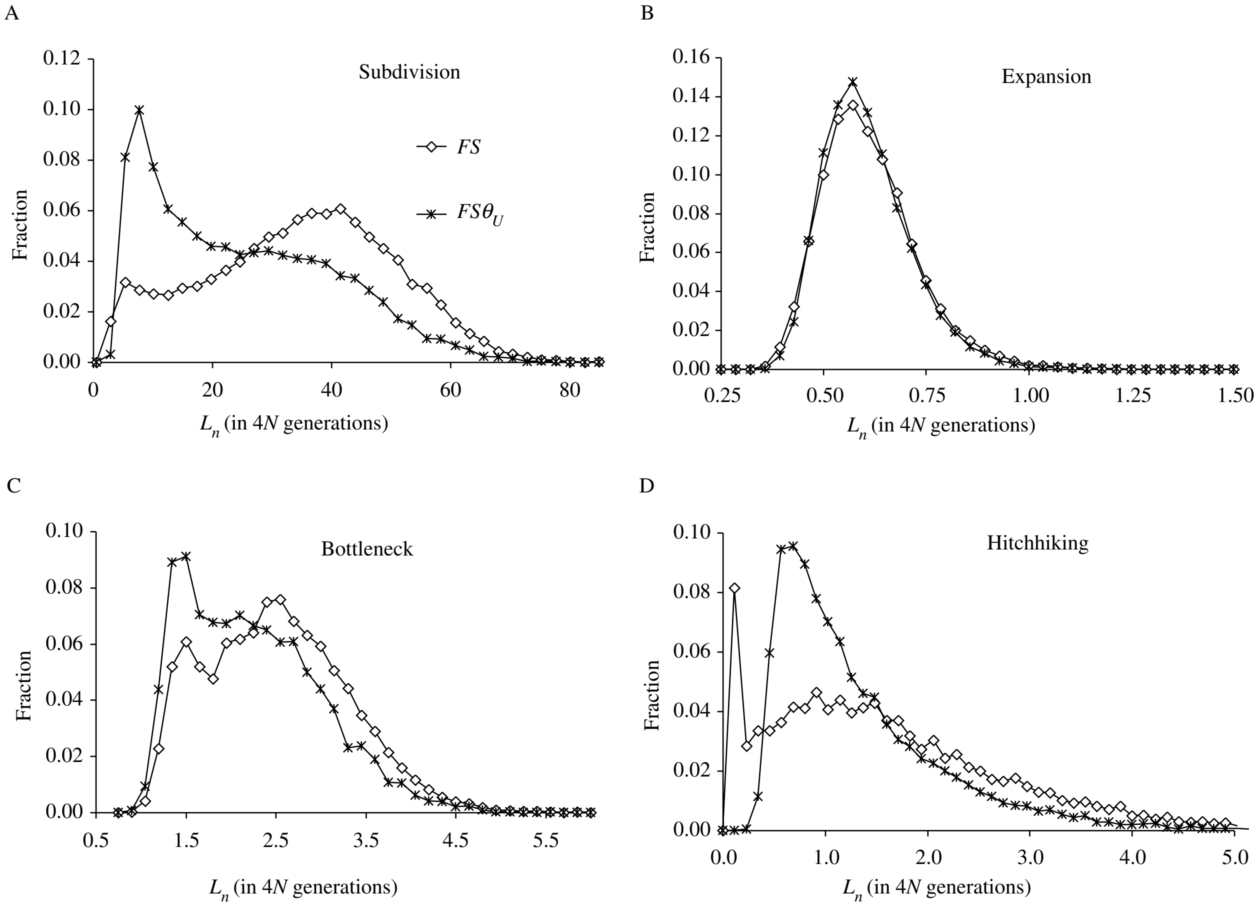

Fig. 6. Probability distribution of L_n for four alternative models using $n=20$ and $S=20$. The *FS* and $FS\theta_U$ procedures are displayed in each case. Parameter values are the same as in Table 1 but intragenic recombination was set to $4Nr=10$ (except for panel *D* where it was zero). (*A*) Island model. (*B*) Logistic expansion model. (*C*) Bottleneck model. (*D*) Hitchhiking model.

and the posterior density of L_n given S and the prior density g is

$$f_{L_n|S,g}(t)=\frac{f_{L_n}(t)}{\mathbf{P}(S=k|n,g)}\times\int_{\theta_{\min}}^{\theta_{\max}}\frac{(\theta t)^k e^{-\theta t}}{k!}g(\theta)\mathrm{d}\theta. \tag{A8}$$

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

$$f_{L_n|S,g}(t)=\frac{f_{L_n}(t)\int_{\theta_{\min}}^{\theta_{\max}}\frac{(\theta t)^k e^{-\theta t}}{k!}\mathrm{d}\theta}{\int_0^{\infty}\int_{\theta_{\min}}^{\theta_{\max}}\frac{(\theta t)^k e^{-\theta t}}{k!}f_{L_n}(t)\mathrm{d}\theta\mathrm{d}t}.$$

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

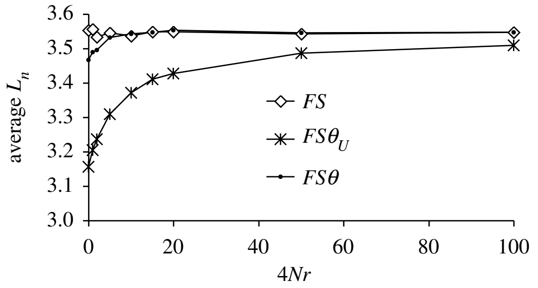

Fig. 7. Effect of recombination on average L_n for the *FS*, $FS\theta$ ($\theta = 0{\cdot}0057$) and $FS\theta_U$ procedures in a neutral panmictic population using $n = 20$ and $S = 20$.

denominator of equation A9 to converge):

$$\lim_{\theta_{\max}\to\infty} f_{L_n|S,g_u}(t) = \frac{t^{-1} f_{L_n}(t)}{\int_0^\infty t^{-1} f_{L_n}(t)\,\mathrm{d}t}. \qquad \text{(A9)}$$

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

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

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

and Y_n is defined by $Y_n = \sum_{k=3}^{n} P_k$ (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_{λ_i} exponentially distributed with parameters λ_i ($i = 1, \ldots, k$). Then $\min(X_{\lambda_i})$ is exponentially distributed with parameter $\lambda = \sum_{i=1}^{k} \lambda_i$.
- Memoryless property of the exponential distribution: Consider a random variable X_λ exponentially distributed with parameter λ:

$$\forall x \geqslant 0,\ \mathbf{P}(X_\lambda - x \leqslant t | X_\lambda \geqslant x) = 1 - e^{-\lambda t} = \mathbf{P}(X_\lambda \leqslant t).$$

(ii) *Density function of* Y$_\text{n}$

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 $E_i - E_{(1)}$ 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 $E_i - E_{(2)}$ 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 $E_{(n-2)} = \sum_{k=1}^{n-2} (E_{(k)} - E_{(k-1)})$. 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 $(n-2)^{\text{th}}$ 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):

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

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

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

Depaulis, F. & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution* **15**, 1788–1790.

Depaulis, F., Mousset, S. & Veuille, M. (2001). Haplotype tests using coalescent simulations conditional on the number of segregating sites. *Molecular Biology and Evolution* **18**, 1136–1138.

Depaulis, F., Mousset, S. & Veuille, M. (2003). Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of Molecular Evolution* **57**(Suppl. 1), S190–200.

Depaulis, F., Mousset, S. & Veuille, M. (2005). Detecting selective sweeps with haplotype tests. In *Selective Sweep* (ed. D. Nurminsky), pp. 34–54. Georgetown, TX: Landes Biosciences.

Donnelly, P. & Tavaré, S. (1995). Coalescent and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.

Fay, J. C. & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

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

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

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

Fu, Y.-X. & Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. (2003). Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. *Genetics* **165**, 1269–1278.

Griffiths, R. & Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* **15**, 790–799.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7 (ed. D. Futuyma & J. Antonovics), pp. 1–45. Oxford: Oxford University Press.

Hudson, R. R. (1993). The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution* (ed. N. Takahata & A. Clark), pp. 23–36. Sunderland, MA: Sinauer Associates.

Jakobsson, M., Hagenblad, J., Tavaré, S., Säll, T., Halldén, C., Lind-Halldén, C. & Nordborg, M. (2006). A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Molecular Biology and Evolution* **23**, 1217–1231.

Kelly, J. (1997). A test of neutrality based on interlocus associations. *Genetics* **146**, 1197–1206.

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

Kingman, J. F. C. (1982*a*). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Kingman, J. F. C. (1982*b*). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.

Kuhner, M. K., Yamato, J. & Felsentein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **140**, 1421–1430.

Markovtsova, L., Marjoram, P. & Tavaré, S. (2000). The effects of rate variation on ancestral inference in the coalescent. *Genetics* **156**, 1427–1436.

Markovtsova, L., Marjoram, P. & Tavaré, S. (2001). On a test of Depaulis and Veuille. *Molecular Biology and Evolution* **18**, 1132–1133.

Mousset, S., Derome, N. & Veuille, M. (2004). A test of neutrality and constant population size based on the mismatch distribution. *Molecular Biology and Evolution* **21**, 724–731.

Nordborg, M. (2001). Coalescent theory. In *Handbook of Statistical Genetics* (ed. D. Balding, M. Bishop & C. Cannings), pp. 179–212. Chichester: Wiley.

Pitman, J. (1992). *Probability*. Berlin: Springer.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, W. (1999). Population growth of human y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189.

Przeworski, M. (2003). Estimating the time since the fixation of a beneficial allele. *Genetics* **164**, 1667–1676.

Ramos-Onsins, S. & Mitchell-Olds, T. (2007). *mlcoalsim*: multilocus coalescent simulations. *Evolutionary Bioinformatics* **2**, 41–44.

Ramos-Onsins, S. E. & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092–2100.

Rozas, J. & Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175.

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497.

Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Sokal, R. R. & Rohlf, F. J. (2001). *Biometry*. New York: W. H. Freeman.

Stephan, W., Wiehe, T. & Lenz, M. W. (1992). The effect of strongly selected substitutions on neutral polymorphism analytical results based on diffusion theory. *Theoretical Population Biology* **41**, 237–254.

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

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. *Theoretical Population Biology* **26**, 119–164.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., & Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the USA* **102**, 18508–18513.

Wall, J. (1999). Recombination and the power of statistical tests of neutrality. *Genetical Research* **74**, 65–79.

Wall, J. & Hudson, R. R. (2001). Coalescent simulations and statistical tests of neutrality. *Molecular Biology and Evolution* **18**, 1134–1135.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.

Wright, S. I. & Charlesworth, B. (2004). The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**, 1071–1076.

Wright, S. I., Vroh Bi, I., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D. & Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314.

https://doi.org/10.1017/S0016672307008877 Published online by Cambridge University Press