Genet. Res., Camb. (1999), **74**, *pp.* 65–79. *With* 10 *figures.* Printed in the United Kingdom © 1999 Cambridge University Press

Recombination and the power of statistical tests of neutrality

JEFFREY D. WALL*

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

(*Received 20 October 1998 and in revised form 19 January 1999*)

Summary

Two new test statistics were constructed to detect departures from the equilibrium neutral theory that tend to produce genealogies with longer internal branches (e.g. population subdivision or balancing selection). The new statistics are based on a measure of linkage disequilibrium between adjacent pairs of segregating sites. Simulations were run to determine the power of these and previously proposed test statistics to reject an island model of geographic subdivision. Unlike previous power studies, this one uses a coalescent model with recombination. It is found that recombination rates on the order of the mutation rate substantially reduce the power of most test statistics, and that one of the new test statistics is generally more powerful than the others. Two suggestions are made for increasing the power of the statistical tests examined here. First, they can be made more powerful if critical values are obtained from simulations that condition on a lower bound for the population recombination rate. Secondly, for the same total length sequenced, power is increased if independent loci are considered instead of a single contiguous stretch.

1. Introduction

One of the fundamental goals of evolutionary genetics is to determine what forces in the past have influenced the genetic variation observed in the present. For sequence data, some researchers have approached this problem by developing statistical tests to detect departures from a constant size, panmictic, no recombination, neutral Wright–Fisher model (e.g. Hudson *et al*., 1987, 1994; Tajima, 1989; McDonald & Kreitman, 1991; Fu & Li, 1993; Fu, 1996, 1997; McDonald, 1996). This null model is widely used because it makes simple, testable predictions; it is one way of modelling Kimura's (1968, 1983) neutral theory of molecular evolution. When one of these tests rejects the null hypothesis, it is likely that at least one of the assumptions of the equilibrium neutral model has been violated. Possible alternatives include more complex demographic histories (e.g. population structure, changes in population size), linkage to sites under selection of some kind (e.g. balancing, fluctuating, directional or purifying selection), or selection operating directly on the sites in question. Ideally one would like to determine which alternatives are likely if the null model is rejected; however, this is difficult to do, partly because of our ignorance of the patterns of genetic variability expected under many of the alternative models, but also because multiple alternatives can produce patterns that are similar and thus difficult to distinguish. For polymorphism data from a single species, Fu (1996) categorized alternative models as those that tend to produce an excess of 'new' mutations (e.g. linkage to a recent selective sweep, population growth) and those that tend to produce an excess of 'old' mutations (e.g. population subdivision, balancing selection). Alternative models in the same class are expected to produce data sets that differ from equilibrium neutral expectations in similar ways.

The usefulness of these statistical tests depends on how often they reject the null hypothesis when it is actually false. This is often tested by simulating data under some model other than the equilibrium neutral model and documenting what power various statistical tests have to reject the null model (e.g. Braverman *et*

* Telephone: +1 (773) 702 9477. Fax: +1 (773) 702 9740. e-mail: jdwall@midway.uchicago.edu

al., 1995; Simonsen *et al*., 1995; Fu, 1996, 1997). The same approach is followed in this paper. We concentrate on ways of analysing sequence polymorphism data from a single species, and on the effects of two departures from the standard equilibrium neutral model: recombination and geographic subdivision. In particular, we find the power of several tests to reject the null model when data are simulated using a symmetrical island model of geographic subdivision (Wright, 1931) with recombination. Although some researchers note that most test statistics are conservative with respect to recombination (e.g. Tajima, 1989; Fu & Li, 1993; Fu, 1996), no one has documented how strong the effect actually is. Determining the magnitude of this effect is one of the primary aims of this paper. It is expected that models of linkage to a site under balancing selection (Hudson & Kaplan, 1988) and models of deterministically decreasing population size (e.g. Griffiths & Tavaré, 1994) would produce similar results, as might other models of geographic subdivision (e.g. Whitlock & McCauley, 1990). This work is presented in three parts: first, two new test statistics B and Q are developed (see Section 2). Then, their powers are compared with the powers of previous test statistics to detect geographic subdivision in the presence of recombination. The other test statistics considered are Tajima's (1989) D (which hereafter is called T), Fu & Li's (1993) D^* (here called D), and Fu's (1996) W and G_η (here called G). Finally, an example from the literature is analysed.

A different method of demographic inference involves maximum likelihood (see, e.g., Griffiths & Tavaré, 1995; Kuhner *et al*., 1995). If likelihoods can be calculated for alternative models, then a likelihood-ratio test may be used to discriminate between hypotheses. Likelihood methods are appealing because they make full use of the available data, unlike summary statistics. However, they are computationally intensive, and it is unclear how sensitive they are to model assumptions. Algorithms for calculating likelihoods exist for finite-island models without recombination (e.g. Nath & Griffiths, 1996) and for panmictic models with recombination (Griffiths & Marjoram, 1996). Although in theory it should be straightforward to calculate likelihoods for a model with both geographic subdivision and recombination, it is not yet computationally practical to do so for data sets of reasonable size. A program (recom58, provided by R. C. Griffiths) for calculating likelihoods under a panmictic model with recombination (which should be faster than a program that calculates likelihoods for a finite-island model with recombination) takes several weeks of computing time on a 400 MHz Pentium II processor to calculate maximum likelihood estimates for a single data set with sample size $n = 30$, $S = 20$ segregating sites, and recombination parameter $4Nr = 5$ (results not shown). Even then, the result would be hard to interpret; for single-locus data it is unclear how to obtain the critical values of a likelihood-ratio test without extensive simulation. The standard χ^2 approximation for the distribution of $2 \log (L_1/L_0)$ (where L_1 and L_0 are the likelihoods under the alternative and null models) is not necessarily applicable. Data sets considered for this paper's tests are this size or larger, so for them summary statistics may be the only viable alternative.

Another approach to the problem of inferring geographic structure from sequence data can be found in the permutation tests of Hudson *et al*. (1992). When adequate sample sizes are obtained from more than one island, their tests are often much more powerful than the tests surveyed in this paper (Hudson *et al*., 1992; Fu, 1996). This is not surprising since Hudson *et al*.'s tests explicitly use the information of where each sequence was sampled whereas the other tests do not. However, there are situations when their test should not be used. For example, this permutation test cannot be used whenever individuals are sampled from only a single locality, and might be significantly less effective if the population structure does not correspond in a simple way to geographic location (see, e.g., Hilton & Hey, 1996, 1997), since a sample from multiple localities would not necessarily include different putative islands. This lack of correspondence between population structure and geographic location might also apply to species such as *Drosophila melanogaster* that are thought to have originated in a particular area (e.g. Africa for *D. melanogaster*) and recently expanded their range.

Most of the following simulations concentrate on situations when only a single population has been sampled. Some researchers have chosen to analyse single-locality samples under the assumption that local populations are in equilibrium, even if the species as a whole shows evidence of geographic structure. One of the conclusions of this study is that the above assumption is not conservative: if there actually is population structure, then assuming panmixia for a sample from a single locality could lead to false positive test results.

2. New test statistics

Under the standard Wright–Fisher neutral model, genealogies of a sample can be simulated using the coalescent (Kingman, 1982*a*, *b*; Hudson, 1990), and the genealogical relationships between different members in the sample can be represented pictorially by a tree (for a review see Hudson, 1990). Departures from the standard neutral model can be thought of in light of the effect they have on the shape of the genealogical tree. For example, the presence of

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

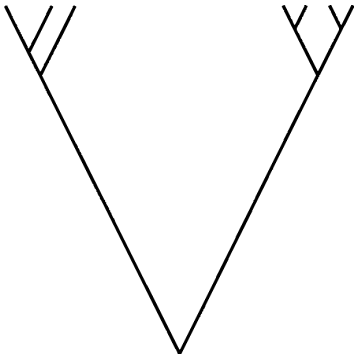

Fig. 1. An example of a coalescent tree that might arise under geographic subdivision.

geographic subdivision might produce a genealogy like the one shown in Fig. 1. This genealogy differs from the neutral one in one branch (the oldest one) being much longer than its standard neutral expectation. (For a detailed explanation see, e.g., Takahata, 1988.) It has been shown that linkage to a site under balancing selection can produce similarly shaped genealogies (Hudson & Kaplan, 1988).

The new test statistics to be introduced use the information in adjacent pairs of segregating sites. Suppose each new mutation happens at a different nucleotide site from all previous mutations (i.e. the infinite-sites assumption). Call a pair of segregating sites *congruent* if the subset of the data consisting of the two sites contains only two different haplotypes. If there has been no recombination between the two segregating sites, they will be congruent if and only if both mutations lie on the same branch of the unrooted tree. Another way of thinking about this is to consider each segregating site as an unordered partition of the sample, where the subsets correspond to those individuals that have the same allele at that particular site. (A partition of the sample consists of two disjoint subsets whose union is the set of individuals in the sample.) Two segregating sites are congruent if and only if their corresponding partitions are identical.

Label the branch lengths in the unrooted genealogy $l_1, l_2, \ldots, l_x$. Then, for two segregating sites with the same genealogical history (e.g. two segregating sites with no recombination between them), the probability that they are congruent is

$$P(\text{congruent}) = \frac{\sum_{j=1}^{x} l_j^2}{\left(\sum_{j=1}^{x} l_j\right)^2}. \quad (1)$$

For genealogies with one very long branch (such as Fig. 1), this probability is higher than for most neutral genealogies, because it is quite common for both mutations to lie on the one very long branch. The probability of congruence is much more complicated when there is recombination between the two segregating sites. However, since not all recombination events affect the longest branch, it is expected that compared with the neutral case, alternative models that tend to produce genealogies that look like Fig. 1 will still have a higher probability of congruence of segregating sites even when there is recombination between the two sites. Set S as the number of segregating sites in the sample. Then, define

B' = the number of pairs of adjacent segregating sites that are congruent,

$$B = B'/(S-1).$$

B has been scaled so that its minimum value is 0 and its maximum is 1. It can be thought of as a measure of linkage disequilibrium among the segregating sites. For the reasons outlined above, $E(B)$ (the expectation of B) should be higher under a geographic subdivision model than the standard neutral model with the same level of recombination. Simulation results confirm this for a finite-island model of geographic subdivision (results not shown). Thus, B can be used as a one-tailed test statistic where values that are too high reject the standard neutral model; such values are suggestive of geographic subdivision, or some other force that tends to distort genealogies into having one or more branches that are much longer than the others. Unfortunately, analytical results are difficult to obtain even in the simplest cases. If there is no recombination and $S = 2$, then finding $E(B)$ (i.e. finding the average of (1) over all possible genealogies) would still require knowledge akin to the expectations and variances of all the relative branch lengths in an unrooted tree. Even for a sample size of $n = 3$, this expectation may have to be found by numerical integration. The critical values of B are therefore found by simulation instead.

As the recombination rate increases, $E(B)$ is expected to decrease; it becomes less and less likely that adjacent segregating sites share the same genealogy, and the probability of congruence is less for those that do not than for those that do. In contrast, one quantity that increases with increasing recombination rate is the number of different partitions defined by adjacent pairs of congruent segregating sites. Although the absolute probability of congruence decreases, those pairs that are congruent are more likely to have a genealogy that is not shared by other congruent pairs, and hence are more likely to induce a unique partition. Let A = the set of all distinct partitions induced by congruent pairs of segregating sites. Then, define

$$Q = (B + |A|)/S,$$

where $|A|$ is the size of the set A. Like B, Q is scaled to be between 0 and 1, and is also expected to be larger

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

Table 1. *Data set used as an example* (*see text*)

	Segregating site									
	1	2	3	4	5	6	7	8	9	10
Seq1	a	c	c	t	a	g	a	c	t	a
Seq2	g	.	.	.	.	t	.	.	c	g
Seq3	g	g	t	g	c	t	.	.	c	.
Seq4	g	.	.	.	c	.	.	.	.	.
Seq5	g	g	t	g	c	t	t	g	.	.

Table 2. *Summary of the statistical tests considered*

Test statistic	Source
W	Fu (1996)[a]
B	See Section 2
Q	See Section 2
G	G_{η}, from Fu (1996)
$D(1)$	D^*, from Fu & Li (1993)[b]
$D(2)$	D^*, from Fu & Li (1993)[c]
$T(1)$	D, from Tajima (1989)[d]
$T(2)$	D, from Tajima (1989)[c]

[a] See also Strobeck (1987), Depaulis & Veuille (1998).
[b] One-tailed test of when D^* is significantly positive.
[c] Two-tailed test.
[d] One-tailed test of when D is significantly positive.

under a geographic subdivision model than under the panmictic neutral model. It too can be used as one-tailed test statistic with the critical values determined by simulation. Although B should be conservative in the presence of recombination, it is not as clear what effect recombination has on the distribution of Q, since Q is the sum of one term that is positively correlated with the recombination rate and one that is negatively correlated with the recombination rate. Both Q and B are attempts to use some of the information captured in the phylogeny besides the number of descendants of each mutation. They are both *ad hoc*, but are easy to calculate.

As an example, consider the sample of sequences shown in Table 1. There are 10 segregating sites, so nine pairs of adjacent segregating sites to consider. Three of these pairs are congruent: sites 2 and 3, sites 3 and 4, and sites 7 and 8. Of these, the first two induce the same partition while the last one induces a separate partition. Thus, for this data set, $S = 10$, $B' = 3$, $|A| = 2$, $B = 0{\cdot}333$ and $Q = 0{\cdot}5$.

3. Simulations

Random sequence samples were generated using a modification of a coalescent program with recombination and geographic subdivision kindly provided by R. R. Hudson. This program assumes an infinite-sites model, so all segregating sites are biallelic. The values of different test statistics were then calculated using these simulated samples. A summary of the test statistics considered can be found in Table 2. (The notation differs from that of some authors.) The powers of these eight statistics were then compared under various scenarios. All simulations were run conditional on the number of segregating sites, not $\theta = 4N\mu$. (θ is the population mutation rate, N is the diploid effective population size and μ is the total mutation rate per generation.) The rationale for this is that we can observe the number of segregating sites in a sample but we must estimate θ. Power simulations conditional on θ are problematic since there is no way of constructing an appropriate null distribution without knowing the true value of θ. A more thorough argument can be found in Hudson (1993). When the number of segregating sites is fixed, D is equivalent to counting the total number of singletons, and W is equivalent to counting the number of haplotypes.

Critical values for the test statistics were estimated from 100000 simulations of a panmictic, no recombination model with the sample size and the number of segregating sites fixed, and significance defined at the 5% level. (Critical values for Q often conditioned on low levels of recombination instead of no recombination to be conservative. This is discussed in Section 4.) The power was determined by counting how often the null model was rejected out of 100000 replicates of an alternative model. These latter models conditioned on the same sample size and number of segregating sites, and specified a particular symmetrical island model (fixing the number of islands and the migration rate between them) and a recombination rate. The scale migration and recombination rates are defined as $4Nm$ and $4Nr$ respectively, where m is the proportion of migrants per generation between each pair of islands and r is the recombination rate per generation. Since there are at least six free variables (sample size, recombination rate, number of segregating sites, number of islands, migration rate between islands, and distribution of sampled individuals within islands), it is computationally unfeasible to test power across all the parameter space. There was no attempt to be exhaustive; instead, examples are shown that are thought to be indicative of general patterns. These often involve changing one or two variables while holding the others constant. Many more simulations were run than can be described in this paper; the details and results of these additional simulations, as well as all computer programs used, are available from the author on request.

Conditioning on the number of segregating sites makes most of the variables examined (W, D, Q and B) have few possible values. In order to compare powers more accurately, a randomized test was used

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

(see, e.g., Lehmann, 1986, p. 71). Suppose, for example, 100000 replicates are run, resulting in 2000 trials with $W < 5$, 5000 trials with $W = 5$, and 93000 trials with $W > 5$. What P value should then be assigned to those trials having $W = 5$? The conservative approach would assign $P = 0{\cdot}07$. A drawback of this approach is that a one-tailed test would reject only 2000 trials at the 5% level; furthermore, the P values would not be uniformly distributed on (0, 1) under the null hypothesis. The approach taken here is to choose the P value for a given trial with $W = 5$ uniformly from (0·02, 0·07). This way, exactly 5% of the trials under the null hypothesis will be rejected at the 5% level, making it easier to compare powers. In practice, the conservative approach would be used, leading to a loss of power of W, D, Q and B. This loss of power decreases as the sample size and number of segregating sites increase, and is an inherent problem of simulating conditional on the number of segregating sites (see Section 6). As mentioned before, simulating conditional on θ has its own problems.

4. Results

I first tested the effect of recombination alone on statistics, when panmixia is assumed. The purpose of this was to examine how conservative the test statistics are with respect to recombination. The rejection probabilities of each test statistic as a function of the recombination rate are shown in Fig. 2. Here, the x-axis is $4Nr$ for the entire simulated region and the y-axis is the rejection probability in per cent. The sample size is $n = 30$ and the number of segregating sites is $S = 40$. Note that when $4Nr \geqslant \theta_W$, recombination reduces the rejection probability of most tests by more than half.

$$\theta_W = S \Big/ \sum_{j=1}^{n-1} j^{-1}$$

is the estimate of θ based on the number of segregating sites (Watterson, 1975). For example, when $4Nr \approx \theta_W$ (i.e. $4Nr = 10$ in Fig. 2), the actual rejection probability is $< 4{\cdot}1\%$ for Q, $< 3\%$ for $D(1)$, and is $< 1{\cdot}7\%$ for the other six statistics. The decrease in rejection probability due to recombination is monotonic for all variables except Q, which contains a quantity that positively correlates with the recombination rate.

The same decrease due to recombination can be seen when population structure is simulated instead of panmixia. Fig. 3 shows the power under a two-island model with $4Nm = 0{\cdot}5$, $n = 30$, $S = 40$, and all individuals sampled from the same island. Critical values are determined from 100000 panmictic simulations with $n = 30$ and $S = 40$. Since Q does not decrease monotonically with increasing $4Nr$ in Fig. 2, it would not be conservative to obtain critical values for Q using no recombination simulations. The recombination rates for the null simulations were therefore set at the values for which the rejection probabilities in Fig. 2 were maximal (i.e. $4Nr = 2$ for Q, and $4Nr = 0$ for the other test statistics). That way, if the null model were true, the rejection probabilities would all be $\leqslant 5\%$ regardless of the actual recombination rate. Further geographic subdivision simulations with different sample sizes and number of segregating sites (many of which are described below) suggest that some patterns in Fig. 3 are quite robust.

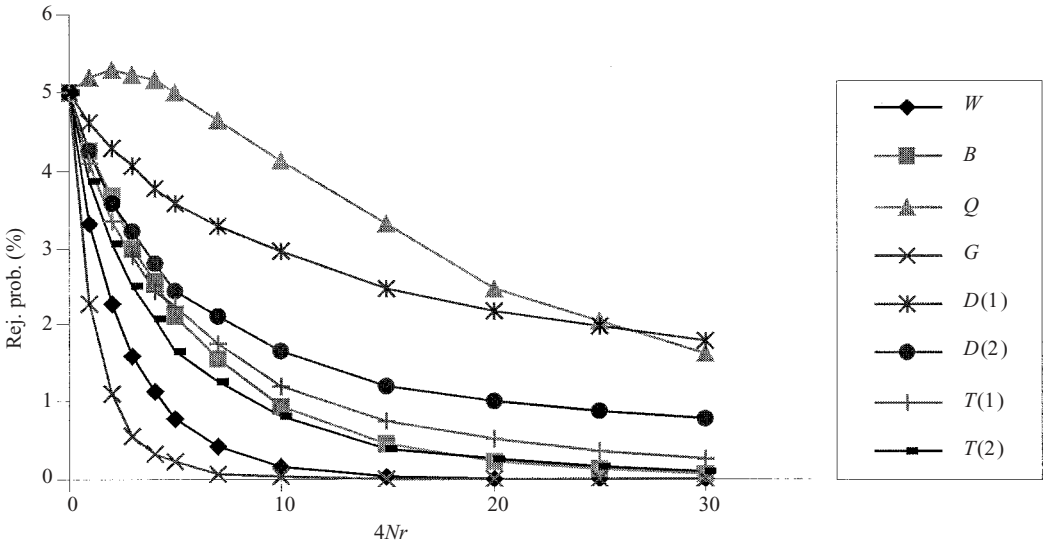

Fig. 2. The decrease in rejection probability due to the presence of recombination. Critical values are obtained from 100000 panmictic, no recombination, infinite-sites stimulations with sample size $n = 30$ and the number of segregating sites $S = 40$. The eight test statistics examined are listed in Table 2.

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

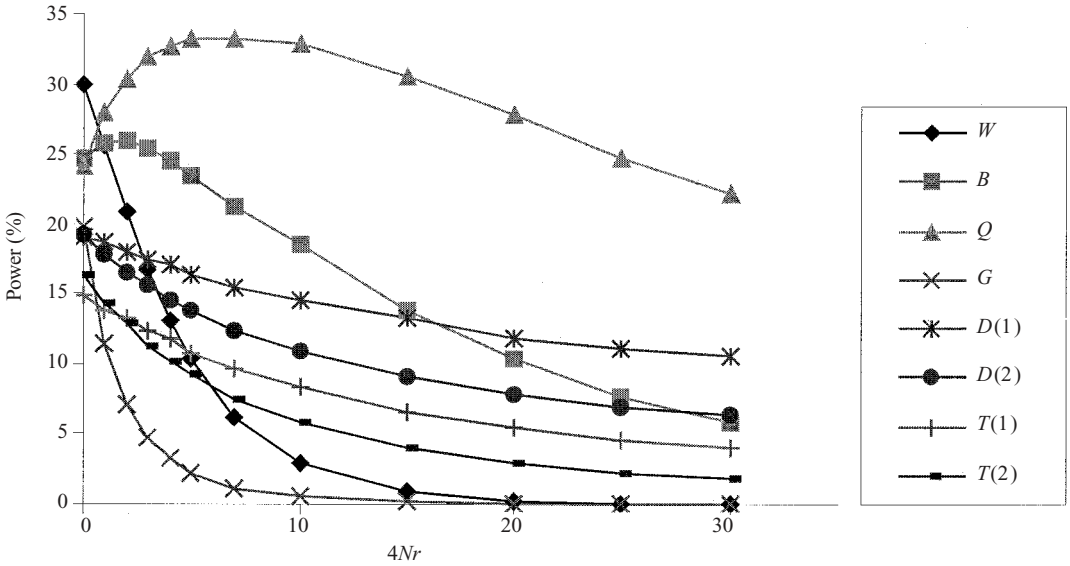

Fig. 3. The effect of recombination on the power to reject the standard neutral model when simulations are run using a symmetrical island model of geographic subdivision. One hundred thousand replicates were run for each value of $4Nr$. A two-island model is used, with $4Nm = 0{\cdot}5$, $n = 30$, $S = 40$, and all individuals sampled from the same island. Critical values are obtained from the panmictic, infinite-sites simulations shown in Fig. 2 using the recombination rate that is the most conservative (see text). The eight test statistics examined are listed in Table 2.

First, it should be noted that all measures have relatively low power for almost all sets of parameter values tested; also, most tests show a monotonic decrease in power with increasing recombination rate. When $4Nr > 0$, the most powerful statistic is Q, except when there are few segregating sites and the recombination rate is very high, in which case $D(1)$ is more powerful. In most simulations, W is the most powerful statistic when $4Nr = 0$; however, W and G are the weakest statistics as soon as $4Nr$ is not extremely small. This implies that Fu's (1997) F_S (which is equivalent to the other tail of W) is strongly non-conservative in areas where recombination is present.

One interesting facet of the data is the large contrast between $4Nr = 0$ and $4Nr > 0$. $D(2)$ and $T(2)$ are more powerful than $D(1)$ and $T(1)$ respectively when $4Nr = 0$, but the situation is reversed as $4Nr$ increases. A possible explanation is as follows: for the parameters in Fig. 3, it is common for there to be more than two migration events in the history of a particular site. When this happens, a positive shift is expected in both D and T. When $4Nr$ is small, the increased variance leads to both tails being large, but as $4Nr$ increases, the expected variance of D and T decreases, leading to substantial weight at only one of the tails. Even more striking is the extreme sensitivity of both W and G to recombination. Though both perform well when $4Nr = 0$ in Fig. 3, they are consistently the worst two measures for medium and high levels of recombination. This result is not surprising. Recombination leads to multiple trees for the segregating sites, and mutations on different genealogies lead to new haplotypes. Low W values are thus quite rare when there is appreciable recombination. Also, a significant G test requires an extreme distortion in the frequency spectrum, such as a majority of mutations occurring in a certain type class. (The type of a mutation, as defined by Fu (1996), is the number of sampled individuals in the smaller of the two allelic classes.) This is very unlikely to occur if the whole sequence does not share the same genealogy.

The change in power is explored under a variety of different scenarios, displayed in Fig. 4–9. Critical values are determined from 100000 panmictic simulations conditional on the same sample size and number of segregating sites. Fig. 4–8 condition on the most conservative recombination rate for each test statistic as was done for Fig. 3; in Fig. 9, critical values are obtained instead from simulations that condition on the actual recombination rate.

(i) *Different migration rates*

Fig. 4 shows how power is affected by changes in the migration rate. Fig. 4*a* has a low migration rate ($4Nm = 0{\cdot}1$), while Fig. 4*b* displays a high migration rate ($4Nm = 2{\cdot}0$). All other variables are the same as in Fig. 3. All test statistics are more powerful when the migration rate is low, and Q is the most powerful except when $4Nr < 2$ or when both $4Nr$ and $4Nm$ are large. Increasing the migration rate causes the power of both Q and B to decrease more quickly as $4Nr$ increases, while both T and D seem much less sensitive to recombination (regardless of the migration rate). As before, both W and G are almost powerless for

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

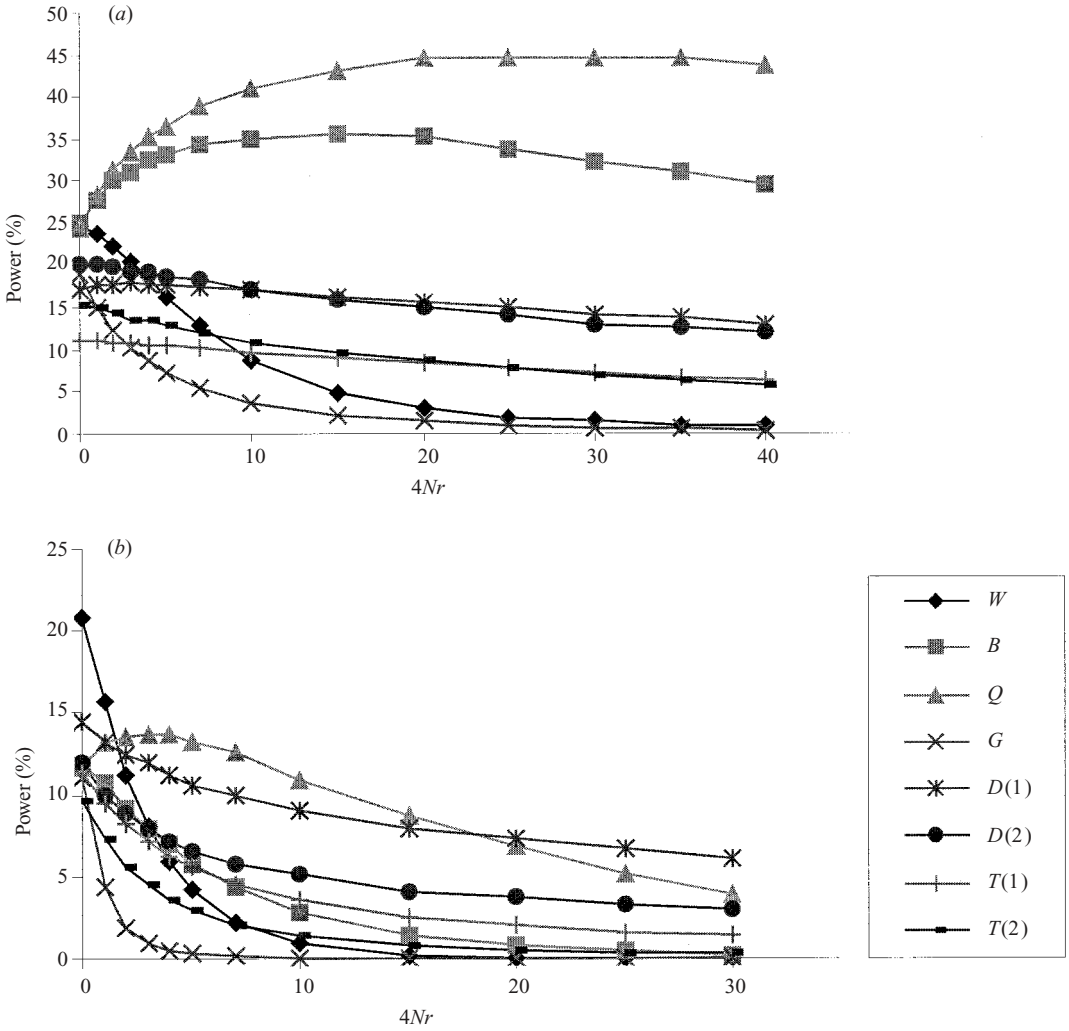

Fig. 4. The recombination rate versus power for different levels of migration. One hundred thousand replicates were run for each value of $4Nr$. A two-island model is used, with $n = 30$ and $S = 40$. All individuals are sampled from the same island. The migration parameters are: (*a*) $4Nm = 0{\cdot}1$ (low migration); (*b*) $4Nm = 2{\cdot}0$ (high migration). Critical values are the same as in Fig. 3. The eight test statistics examined are listed in Table 2.

medium or high levels of recombination. When all individuals are sampled from the same island (as they are in Fig. 4), there is a rather narrow range of migration parameter values that lead to a reasonable chance of detecting the structure using any of the test statistics. When migration is too high (e.g. $4Nm > 5$), migrants are so common that the population behaves similar to a panmictic one. When the migration rate is low, many samples have no migrants in their history (which is also close to the panmictic case). As a result, the power of most test statistics starts decreasing when the migration rate is too low (e.g. $4Nm < 0{\cdot}1$) (results not shown).

(ii) *Different sample configurations*

Fig. 5*a* shows the effect of having five islands instead of two. The sample size, number of segregating sites and migration rate are the same as in Fig. 3, and all individuals are sampled from the same island. Increasing the number of islands from two to five makes all the statistical tests substantially more powerful; for $4Nr > 1$, Q is the most powerful test, and it is more than twice as powerful as it is in Fig. 3. Population structure is often easier to detect because those trials with at least one migration event often have very deep (thus distorted relative to neutral expectations) genealogies; it takes longer for two individuals in different islands to coalesce because most migration events move one individual to a third island instead of to the same island as the other individual. Also, in Fig. 5*a*, all test statistics (except W and G) retain most of their power even when the recombination rate is quite high.

Fig. 5*b* shows how sampling equally from two islands (instead of sampling all individuals from the

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

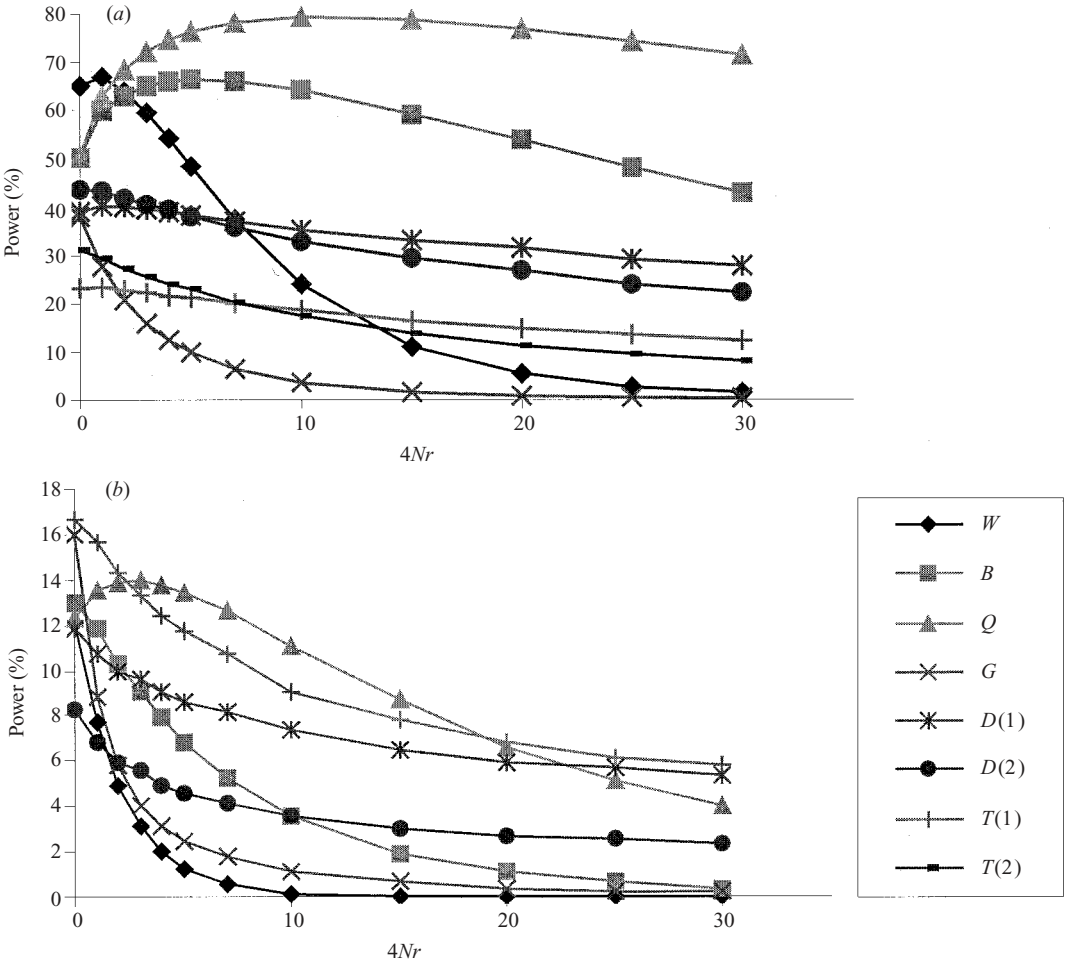

Fig. 5. The recombination rate versus power for two different sample configurations. One hundred thousand replicates were run for each value of $4Nr$. For both graphs, $4Nm = 0{\cdot}5$, $n = 30$ and $S = 40$. In (*a*) there are five islands, and all individuals are sampled from the same island. For (*b*), there are two islands, and 15 individuals are sampled from each island. Critical values are the same as in Fig. 3. The eight test statistics examined are listed in Table 2.

same island) affects the power of the different statistics. The total sample size, number of segregating sites, number of islands and migration rate are as in Fig. 3. All statistics except for $T(1)$ show a decrease in power when multiple islands are sampled. Equal sampling leads to proportionally longer external branches, since fewer pairs of individuals are in the same island (a necessary prerequisite for coalescence). This leads to more expected singletons and a smaller expected probability of congruence. $T(1)$ fares well only because both islands were sampled equally, leading to a rise in intermediate frequency polymorphisms. If the islands are sampled unequally, there is no increase in power (results not shown).

(iii) *Power versus the number of segregating sites*

The effect of the number of segregating sites on the power of the different test statistics is shown in Fig. 6. Under the assumption that the mutation rate and the recombination rate do not vary between nucleotide sites, Fig. 6 can also be interpreted as showing power versus increasing length sequenced (since then the number of segregating sites will be proportional to the length in base pairs). The three graphs are for no ($4Nr = 0$), medium ($4Nr = 0{\cdot}25 * S$) and high ($4Nr = 0{\cdot}75 * S$) levels of recombination. The latter two correspond to $4Nr \approx \theta_W$ and $4Nr \approx 3\theta_W$. Again, all other variables have the same value as in Fig. 3. The most powerful measures are W (Fig. 6*a*) and Q (Fig. 6*b*, *c*). An ideal test statistic would become more powerful when more information (i.e. more segregating sites) is available. In Fig. 6*a*, all eight test statistics have this property. However, this is under the assumption of no recombination, which is unreasonable for most nuclear gene sequences. For medium and high levels of recombination (Fig. 6*b*, *c*), only Q becomes more powerful as the number of segregating sites increases. The others start losing power after an intermediate peak. This observation is somewhat surprising (see Section 6).

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

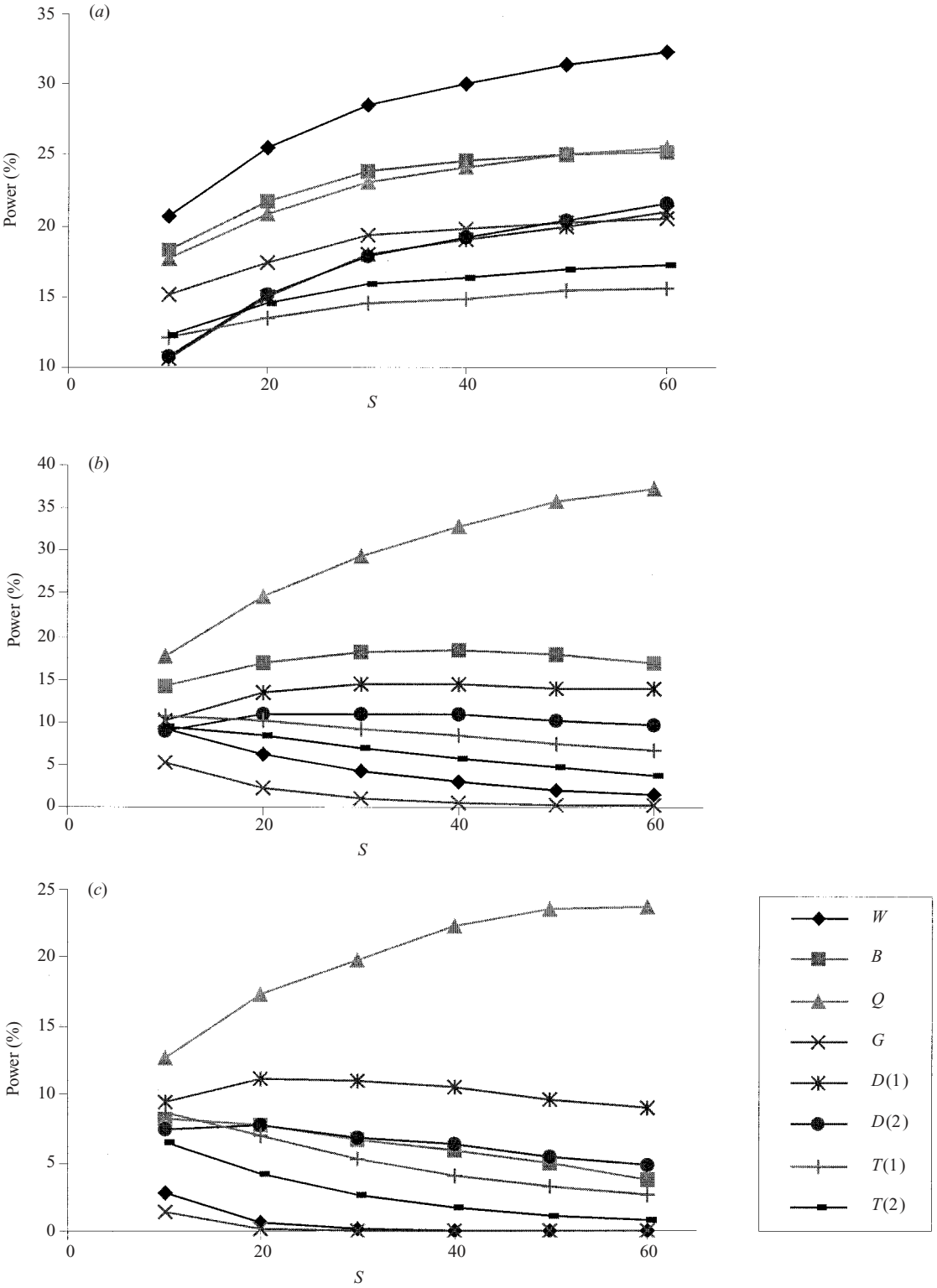

Fig. 6. The number of segregating sites (S) versus power. Each data point is based on 100000 replicates. A two-island model is used, with $4Nm = 0{\cdot}5$, $n = 30$, and all individuals sampled from the same island. The recombination rates are: (*a*) no recombination; (*b*) $4Nr = 0{\cdot}25 * S$ (i.e. $4Nr \approx \theta_W$); (*c*) $4Nr = 0{\cdot}75 * S$ (i.e. $4Nr \approx 3\theta_W$). Critical values are obtained from 100000 panmictic, infinite-sites simulations that condition on the same sample size and number of segregating sites, and use the recombination rate that is the most conservative (see text). The eight test statistics examined are listed in Table 2.

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

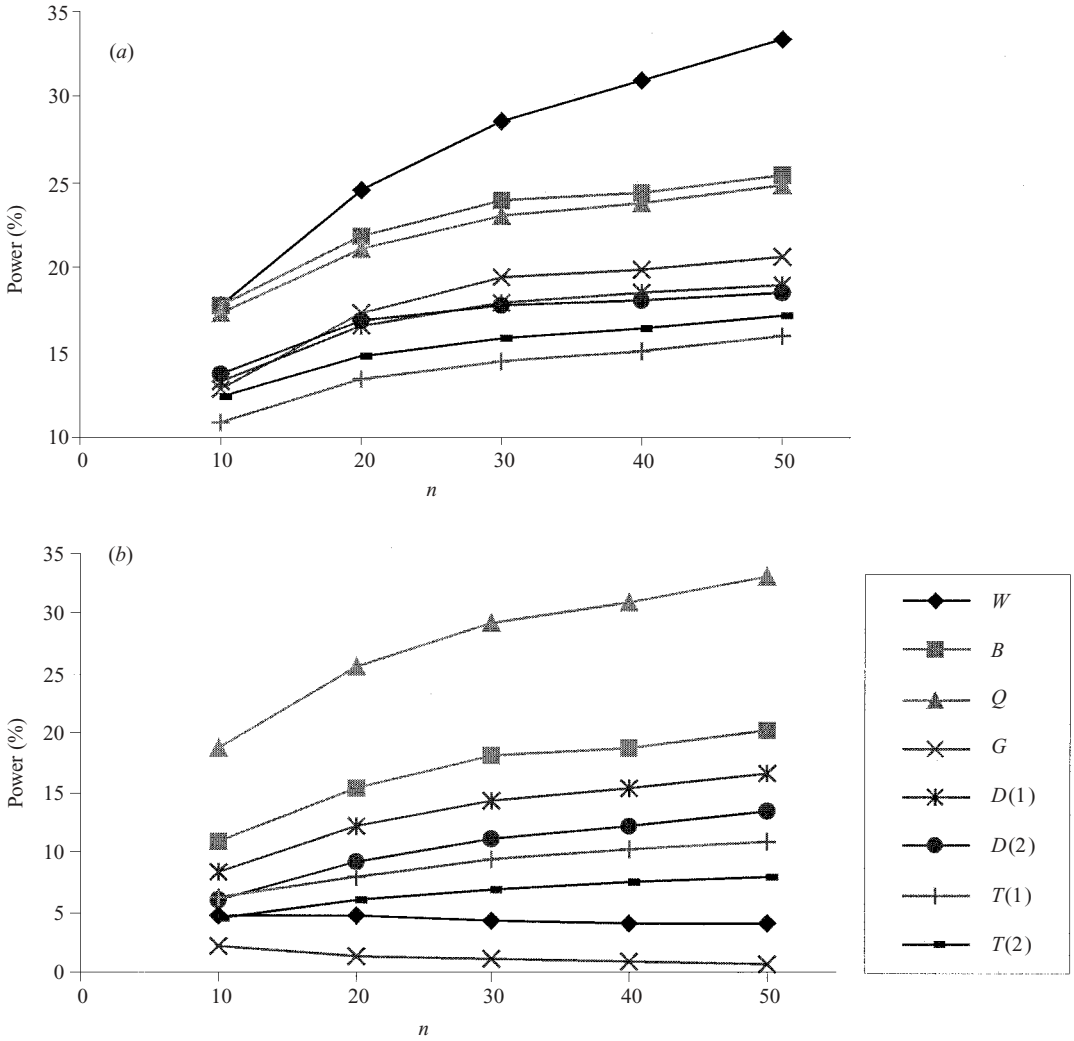

Fig. 7. The sample size (n) versus power for different values of $4Nr$. Each data point is based on 100000 replicates. A two-island model of geographic subdivision is used, with $4Nm = 0{\cdot}5$ and all individuals sampled from the same island. The number of segregating sites was chosen to keep θ_W as close to constant as possible over a fixed length of simulated sequence. The sets of parameters are: $n = 10$ and $S = 21$; $n = 20$ and $S = 27$; $n = 30$ and $S = 30$; $n = 40$ and $S = 32$; $n = 50$ and $S = 34$. The total recombination rates are: (*a*) $4Nr = 0$; (*b*) $4Nr = 7{\cdot}5$ (i.e. $4Nr \approx \theta_W$). Critical values are obtained from 100000 panmictic, infinite-sites simulations that condition on the same sample size and number of segregating sites, and use the recombination rate that is the most conservative (see text). The eight test statistics examined are listed in Table 2.

(iv) *Power versus sample size*

Fig. 7 shows what effect a change in the sample size has on the powers of the different test statistics. The model of subdivision is the same as in Fig. 3, and the number of segregating sites is chosen to keep θ_W (for the whole gene) as close to constant as possible. This involved simulating with the following pairs: $n = 10$ and $S = 21$; $n = 20$ and $S = 27$; $n = 30$ and $S = 30$; $n = 40$ and $S = 32$; $n = 50$ and $S = 34$. Fig. 7*a* and *b* show the results for no recombination ($4Nr = 0$) and medium levels of recombination ($4Nr = 7{\cdot}5$; equivalently, $4Nr \approx \theta_W$). The most powerful statistics are W (no recombination) and Q (medium recombination). Higher recombination rates yield graphs similar to Fig. 7*b* (results not shown). As above, good test statistics should increase in power as more individuals are sampled. The only statistics that do not are W and G in Fig. 7*b*. Once again, W and G are very sensitive to recombination; though W is the most powerful in the no recombination case, it is one of the worst once the recombination rate is on the same order as the mutation rate. The superiority of Q in Fig. 7*b* arises because it is less affected by recombination than the other statistics.

(v) *Trade-off between sample size and length sequenced*

Under the constraint that a fixed total length could be sequenced, simulations were run to see whether it were better to sequence large stretches of few individuals or

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

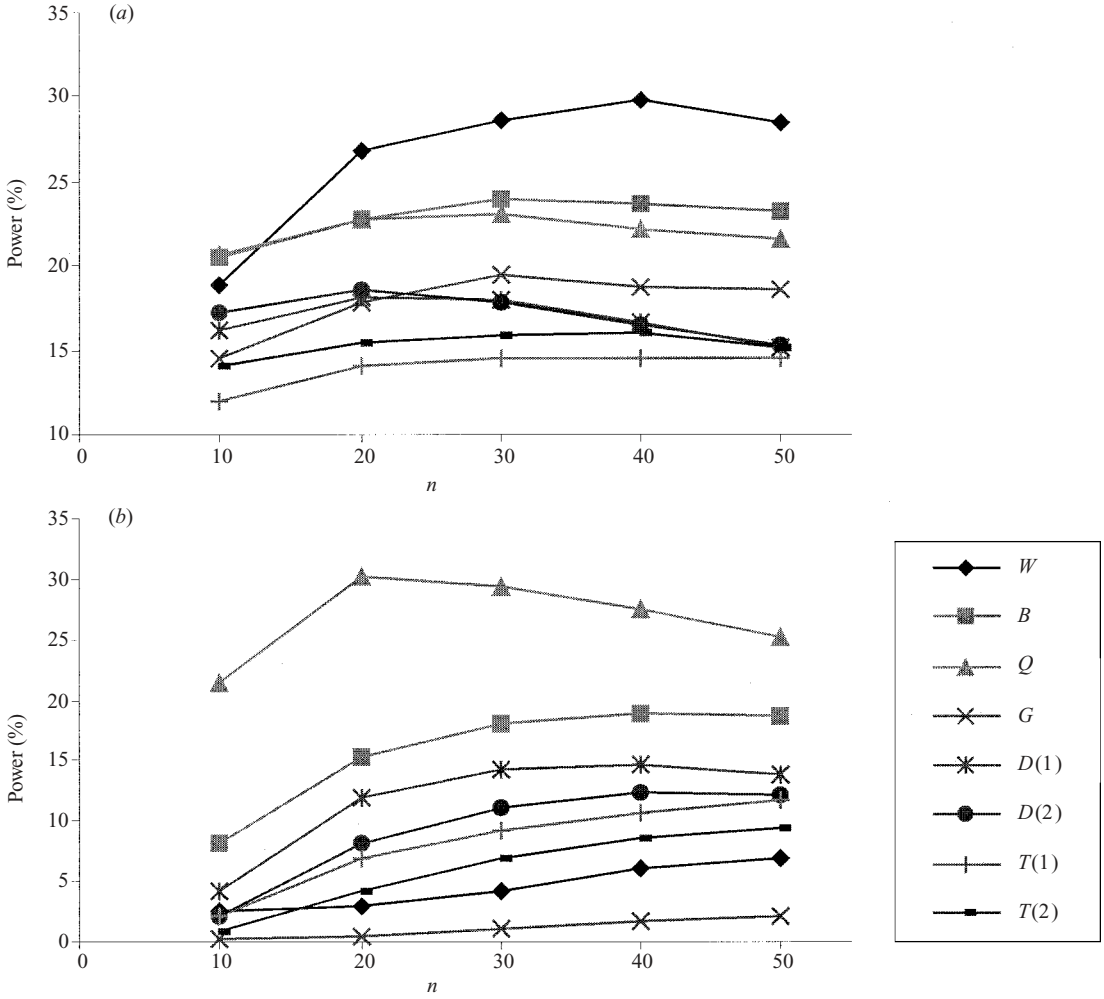

Fig. 8. The sample size (n) versus power for different values of $4Nr$. Each data point is based on 100000 replicates. A two-island model of geographic subdivision is used, with $4Nm = 0{\cdot}5$ and all individuals sampled from the same island. Simulations are meant to mimic a situation where a fixed total length is sequenced and θ_W per site is (as close to) constant (as possible); it is used to study the trade-off between sample size and length sequenced. The specific parameter values are: $n = 10$ and $S = 64$; $n = 20$ and $S = 40$; $n = 30$ and $S = 30$; $n = 40$ and $S = 24$; $n = 50$ and $S = 20$. The recombination rates are: (*a*) $4Nr = 0$; (*b*) $4Nr = 225/n$ (i.e. $4Nr \approx \theta_W$). Critical values are obtained from 100000 panmictic, infinite-sites simulations that condition on the same sample size and number of segregating sites, and use the recombination rate that is the most conservative (see text). The eight test statistics examined are listed in Table 2.

a smaller length from more individuals. Since both time and money are limited, most researchers face this trade-off. For a panmictic population and a similar question, a detailed discussion of optimal sequencing strategies can be found in Pluzhnikov & Donnelly (1996). Fig. 8*a* and *b* show power versus sample size for no recombination ($4Nr = 0$) and moderate recombination ($4Nr = 225/n$). As before, the medium rate corresponds to $4Nr \approx \theta_W$. This implicitly assumes that the recombination rate and the proportion of sites that are segregating are constant per base pair. The model of subdivision was taken to be the same as in Fig. 3, and the number of segregating sites was chosen to keep θ_W (per base pair) as close to constant as possible. The following parameter pairs were used: $n = 10$ and $S = 64$; $n = 20$ and $S = 40$; $n = 30$ and $S = 30$; $n = 40$ and $S = 24$; $n = 50$ and $S = 20$. As in Figs. 6 and 7, the most powerful test statistics are W (in Fig. 8*a*) and Q (in Fig. 8*b*). For most simulations, statistics show maximum power under intermediate values of sample size and length sequenced, but this intermediate value depends both on the particular statistic and on the recombination rate. In general, as the recombination rate increases, the optimal strategy is to sequence a smaller length from more individuals.

(vi) *Conditioning on the actual recombination rate*

Fig. 9 shows power versus recombination rate when the critical values are obtained from simulations that condition on the actual recombination rate. All other parameter values are the same as in Fig. 3. This

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

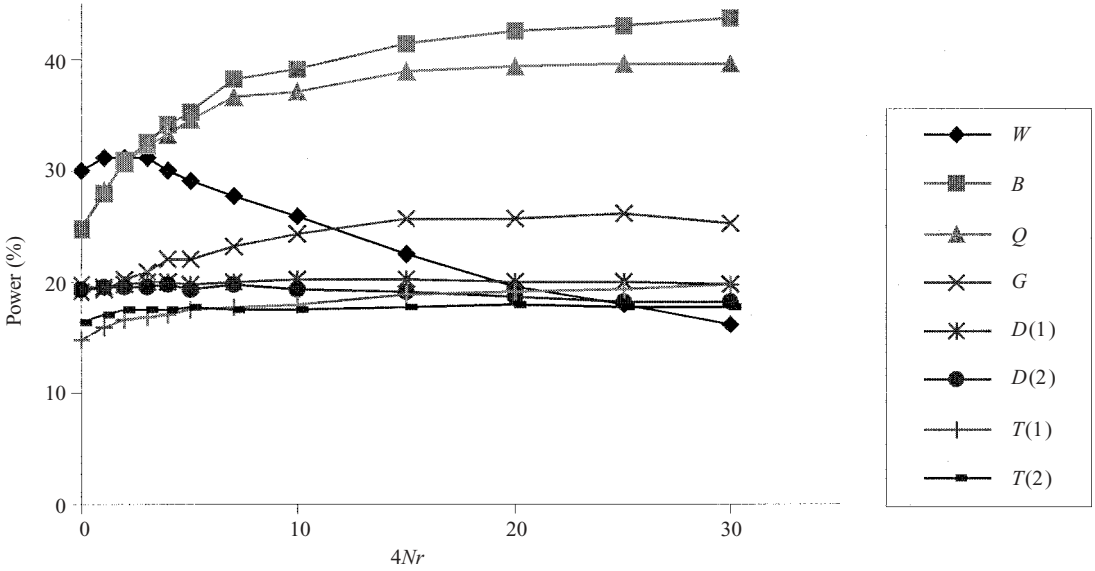

Fig. 9. The recombination rate versus power with critical values obtained from simulations that condition on the actual recombination rate. One hundred thousand replicates were run for each value of $4Nr$, and all other parameter values are the same as in Fig. 3. The eight test statistics examined are listed in Table 2.

situation is unrealistic because the intragenic recombination rate cannot yet be measured directly but must be estimated from the patterns of variation. Current estimators of $4Nr$ from sequence data are all biased and have large variances (e.g. Hudson, 1987; Hey & Wakeley, 1997; Wakeley, 1997). Conditioning on the actual recombination rate noticeably increases the power of all the test statistics relative to Fig. 3. The most powerful test statistic is W (for $4Nr < 3$) or B (for $4Nr \geqslant 3$). Q is less sensitive to recombination than B, so it performs better when critical values are determined using simulations with a conservative recombination rate (e.g. as in Fig. 3).

Table 3. P *values for the different test statistics and the data set of Innan* et al. (1996)

	P value	
Test statistic	$4Nr = 0$	$4Nr = 8{\cdot}9$
W	0·579	0·444
B	0·062	0·019
Q	0·024	0·007
G	0·759	0·304
$D(1)$	0·148	0·087
$D(2)$	0·296	0·175
$T(1)$	0·297	0·196
$T(2)$	0·594	0·392

5. An example

The results of Fig. 9 show that conditioning on a positive recombination rate for the null distribution increases the power of all test statistics. We demonstrate this by analysing a data set taken from a recently published study of *Adh* in *Arabidopsis thaliana* (Innan *et al.*, 1996). Although *A. thaliana* is mostly self-fertilizing, a coalescent model is still reasonable (with a change in time-scaling) since each sequence was sampled from a different individual (Nordborg & Donnelly, 1997). The purpose of this example is pedagogical, not explanatory. Thus, the facts that the sample locations might not be random and that selection might be operating on *Adh* will be ignored.

Table 2 from Innan *et al.* (1996) was culled to include only biallelic single-nucleotide polymorphisms. There are 17 individuals in the sample, 75 segregating sites and 13 distinct haplotypes. The average number of nucleotide differences between two sampled individuals is 19·06, $B' = 31$, $|A| = 9$, $B = 0{\cdot}419$, $Q = 0{\cdot}533$ and $\eta = (38, 7, 9, 5, 0, 1, 5, 10)$. The last is Fu's (1996) notation for the frequency spectrum of a data set with no outgroup. One hundred thousand no recombination, coalescent simulations were run for a panmictic population conditional on $n = 17$ and $S = 75$. The results are shown in the second column of Table 3. The only statistic that is significant (at the 5% level) is Q, with $P = 0{\cdot}024$.

The data from Innan *et al.* (1996) show some evidence of recombination at *Adh* in *A. thaliana*. The estimated minimum number of recombination events in the sample (R_M, from Hudson & Kaplan, 1985) is 7. We construct a lower bound $C_{\min}$ for $4Nr$ as follows: we take the largest value of $4Nr$ such that simulations with this recombination rate are unlikely to produce data sets with $R_M \geqslant 7$ (i.e. they do so less than 2·5% of the time). This is roughly the same method as in Hudson & Kaplan (1985) and Hudson (1987). For our example, the estimated lower bound is

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

$C_{\min} \approx 8{\cdot}9$. Simulations were then run with the most conservative value of $4Nr$ that was greater than or equal to $C_{\min}$ (in this example, $4Nr = C_{\min}$ for all test statistics). Although this method has two sources of Type I error, further simulations show that it is still conservative (results not shown). The P values are displayed in the third column of Table 3. All the P values are lower, and two statistics are significant; Q has $P = 0{\cdot}007$ and B has $P = 0{\cdot}019$. These significant P values support the claim of Innan *et al*. (1996) that the observed haplotypic structure is suggestive of population subdivision or selection.

6. Discussion

One observation that is evident from the simulations is that all test statistics have poor power to reject the null hypothesis, even when large sample sizes and many segregating sites are considered. In reality, the power is even lower for actual data sets because a randomized test would not be used. However, the order, (e.g. that Q is usually the most powerful while W and G are often the worst) would be essentially the same. As discussed in Section 1, part of the problem is that the information contained in which individuals were sampled from which localities is not used by any of the test statistics. In fact, all the test statistics except for $T(1)$ perform better when all individuals are sampled from the same island, instead of comparable sampling from all islands (see Figs. 3 and 5*b*). However, there is still some new information that these statistical tests provide. Though the permutation tests of Hudson *et al*. (1992) are constructed to determine whether samples from different localities are different from each other, the results presented here suggest that samples from a single locality already often do not conform to equilibrium neutral expectations. Researchers who overlook possible subdivision in their samples underestimate the variance in possible outcomes that can arise due to non-selective factors.

The conclusions that can be drawn from a discrepancy between results and neutral expectations are far from obvious; the test statistics described were all constructed to test the consistency of a given data set with the standard equilibrium neutral model. Those data sets that are consistent with the neutral model provide at best only indirect evidence that the region in question is actually evolving neutrally; such data sets may also be consistent with selective alternatives. Conversely, a data set that is inconsistent with the null model might also be inconsistent with many alternative hypotheses (see, e.g., Wayne & Simonsen, 1998). Because of this difficulty, studies of genetic variation that use 'statistical tests of neutrality' without explicit *a priori* alternative hypotheses are hard to interpret. A significant test result without any additional information does not help to distinguish between possible alternatives. Even if the data are unlikely to have arisen under the null model, they may be even more unlikely to have arisen under most or all alternative models. If one had two easily simulated hypotheses and access to powerful computers, then a likelihood approach (that compares the likelihoods of the data under each model) might be appropriate. However, this might not be computationally practical for large data sets or for models (such as island models of geographic subdivision) that require extensive parameterization.

Another problem with *post hoc* analysis using statistical tests of neutrality is that most researchers do not correct for multiple tests. In practice, researchers apply a number of different statistical tests, and consider their data set 'non-neutral' if at least one test is significant. The probability that at least one test rejects neutrality at the 5% level is clearly much higher than 5%. When the P values in Table 3 are recalculated (by simulation) to correct for multiple tests, only Q (when $4Nr = C_{\min}$) remains significant.

Since an accurate correction for multiple tests requires extensive computer simulations, it might be best if only a single statistical test (chosen before the data is collected) is used for analysing a given data set. The choice of test should depend on what alternative hypotheses the investigator thinks are the most reasonable. A statistical test that is good at detecting geographic subdivision, for example, might not be particularly effective at detecting other types of departures from the null model such as recent bottlenecks or linkage to a recent selective sweep. Knowing which test statistic to choose requires more work to be done investigating the predictions of common alternative models to the standard neutral theory.

To highlight the degree of overlap between test statistics, 100000 genealogies were simulated using the same parameter values as in Fig. 2; these simulated data sets were then analysed using W, Q, $D(1)$ and $T(1)$. These four are generally superior (for detecting geographic subdivision) to G, B, $D(2)$ and $T(2)$ respectively. It was found that 12·8% of the replicates had at least one significant test (7·8% with one, 3·2% with two, 1·4% with three, and 0·5% with all four). To have a 5% chance of obtaining at least one significant test statistic, one could run each test with a nominal rejection probability of 1·75%. (This rejection probability was determined by simulation.) However, this composite test statistic is no more powerful than any of the other statistics (results not shown).

There are many other ways one could construct a composite test statistic from W, Q, $D(1)$ and $T(1)$. If the composite is defined to be significant when one or more of the component test statistics are significant,

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

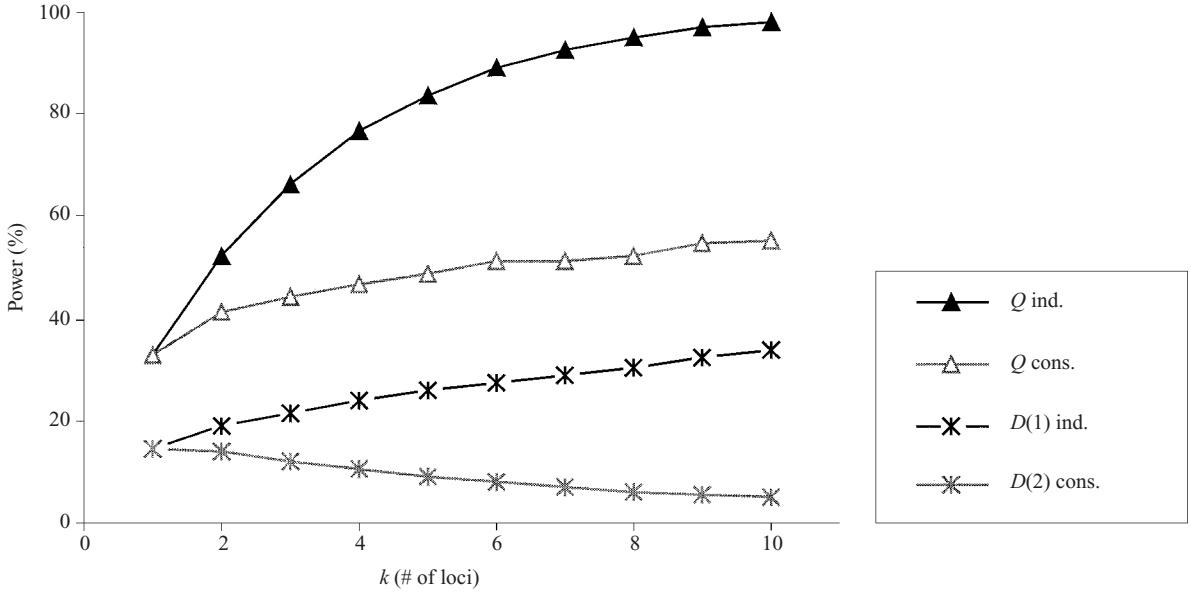

Fig. 10. The power of $D(1)$ and Q as a function of the number of loci considered. Each locus has parameters as in Fig. 3, with $4Nr = 10$, and critical values determined from panmictic simulations of k loci using a conservative recombination rate. Here ind. stands for independent, and cons. stands for consecutive (see text).

then the nominal rejection probabilities could differ among the four component test statistics. For example, since Q is more powerful than the others in Fig. 3, it might make sense to 'weight' Q more heavily than the others by giving it a higher rejection probability while lowering that for the others. A composite test could also be devised that requires more than one test statistic to be significant at a certain level. The power of a number of different composite statistics was compared with the eight test statistics described in Section 1, and none of them performed better than Q (for low levels of recombination) and $D(1)$ (for high levels of recombination) (results not shown).

This last observation is rather disappointing, because it suggests that many test statistics use similar aspects of the data. Perhaps even more disturbing is the observation that increasing the amount of data (i.e. increasing the sample size or the length sequenced) does not lead to a large increase in power. In fact, when recombination rates are high, increasing the length sequenced actually decreases the power of most test statistics (see Fig. 6*c*). As the length sequenced increases, so does $4Nr$; tests that use critical values from no recombination simulations become increasingly conservative (and thus less powerful). However, when the parameter combinations in Fig. 6*c* were rerun with critical values determined from simulations that conditioned on the actual recombination rate (cf. Fig. 9), the shape of the power curves was more like Fig. 6*a* (i.e. power increased with increasing numbers of segregating sites for all test statistics except W) (results not shown). Thus, it is not the presence of recombination itself that decreases power, but the difference between the actual recombination rate and the rate used in simulating the null distribution.

Demographic departures from the standard null model are expected to affect the patterns of observed variability over the whole genome. One way of increasing the power to detect geographic subdivision of all the test statistics examined is to sequence independent loci instead of one large contiguous stretch. The advantage then is that we *know* when there is free recombination between the loci, and hence can condition on it. Simulations were run that modelled k independent loci ($1 \leqslant k \leqslant 10$), each with $n = 30$, $S = 40$, $4Nr = 10$, and the same island model as in Fig. 3. Critical values were determined from simulations with k independent panmictic loci, each with conservative recombination rate. The power of Q is shown in Fig. 10 as a function of k. Also shown in Fig. 10 is the power of Q as a function of k when the k loci are consecutive, not independent (i.e. $S = 40k$, $4Nr = 10k$). As can be seen, the increase in power that arises from being able to condition on free recombination between loci is substantial. For comparison, Fig. 10 includes the power of $D(1)$ (the best of the old test statistics in Fig. 3) as a function of k when the k loci are independent or consecutive.

In the future, the amount of sequence data available will not be the limiting factor, and information will be available on the patterns of variation at many unlinked 'neutral' areas (or at least areas with no functional significance or obvious signs of selection). This information will be used to construct likely demographic scenarios; these scenarios will then be used as null models in tests for selection in specific areas. Instead of just accepting or rejecting the equilibrium

https://doi.org/10.1017/S0016672399003870 Published online by Cambridge University Press

neutral model, we will be able to infer which selective and demographic forces have shaped the observed patterns of sequence variability.

I thank R. R. Hudson for helpful discussions as well as B. Charlesworth, M. Kreitman, T. F. C. Mackay, M. S. McPeek, M. Przeworski and two anonymous referees for comments on an earlier version of this manuscript. This work was supported by the University of Chicago Division of Biological Sciences Unendowed Fund.

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.

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.

Griffiths, R. C. & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.

Griffiths, R. C. & Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London*, *Series B* **344**, 403–410.

Griffiths, R. C. & Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical Biosciences* **127**, 77–98.

Hey, J. & Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846.

Hilton, H. & Hey, J. (1996). DNA sequence variation at the *period* locus reveals the history of species and speciation events in the *Drosophila virilis* group. *Genetics* **144**, 1015–1025.

Hilton, H. & Hey, J. (1997). A multilocus view of speciation in the *Drosophila virilis* species group reveals complex histories and taxonomic conflicts. *Genetical Research* **70**, 185–194.

Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**, 245–250.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7 (ed. D. Futuyma & J. Antonovics), pp. 1–44. New York: Oxford University Press.

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

Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R. R. & Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Hudson, R. R., Boos, D. D. & Kaplan, N. L. (1992). A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution* **9**, 138–151.

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. (1994). Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**, 1329–1340.

Innan, H., Tajima, F., Terauchi, R. & Miyashita, N. T. (1996). Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics* **143**, 1761–1770.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

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

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

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

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. New York: Wiley.

McDonald, J. H. (1996). Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Molecular Biology and Evolution* **13**, 253–260.

McDonald, J. H. & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654.

Nath, H. B. & Griffiths, R. C. (1996). Estimation in an island model using simulation. *Theoretical Population Biology* **50**, 227–253.

Nordborg, M. & Donnelly, P. (1997). The coalescent process with selfing. *Genetics* **146**, 1185–1195.

Pluzhnikov, A. & Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–1262.

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

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

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research* **52**, 213–222.

Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genetical Research* **69**, 45–48.

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

Wayne, M. L. & Simonsen, K. L. (1998). Statistical tests of neutrality in the age of weak selection. *Trends in Ecology and Evolution* **13**, 236–240.

Whitlock, M. C. & McCauley, D. E. (1990). Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution* **44**, 1717–1724.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.