

Detecting Recombination from Gene Trees

J. Maynard Smith and N. H. Smith

School of Biological Sciences, University of Sussex, Brighton, United Kingdom

In this article, a method is proposed for detecting recombination in the sequences of a gene from a set of closely related organisms. The method, the Homoplasy Test, is appropriate when the sequences are rather similar, differing by 1%–5% of nucleotides. It is effective in detecting relatively frequent recombination between a set of rather similar strains, in contrast to previous methods which detect rare or unique transfers between more distant strains. It is based on the fact that, if there is no recombination and if no repeated mutations have occurred (homoplasy), then the number of polymorphic sites, v , is equal to the number of steps, t , in a most-parsimonious tree. If the number of “apparent homoplasies” in the most-parsimonious tree, $h = t - v$, is greater than zero, then either homoplasies have occurred by mutation or there has been recombination. An estimate of the distribution of h expected on the null hypothesis of no recombination depends on S_e , the “effective site number,” defined as follows: if p_s is the probability that two independent substitutions in the gene occur at the same site, then $S_e = 1/p_s$. S_e can be estimated if a suitable outgroup is available. The Homoplasy Test is applied to three bacterial genes and to simulated gene trees with varying amounts of recombination. Methods of estimating the rate, as opposed to the occurrence, of recombination are discussed.

Introduction

We have previously shown that the population structures of microorganisms can range from highly sexual to almost strictly clonal (Maynard Smith et al. 1993). This analysis was based on data generated by multilocus enzyme electrophoresis, a technique that has been highly successful in determining population structures for a variety of organisms. However, multilocus enzyme electrophoresis data sets are difficult and expensive to establish, and recently, nucleotide sequence data of the same gene derived from many, closely related, organisms have become available.

We have developed a test, the Homoplasy Test, for the presence of recombination in the type of sequence data sets that are currently being generated from nucleotide-sequencing population genetic studies. The Homoplasy Test determines if there is a statistically significant excess of homoplasies in the phylogenetic tree derived from the data set, compared to an estimate of the number of homoplasies expected by repeated mutation in the absence of recombination. An excess of homoplasies is considered a hallmark of recombination. The test requires an outgroup sequence and is likely to be most effective for a set of sequences differing by 1%–5% of nucleotides, among which recombination has been frequent. Techniques previously proposed to detect recombination in nucleotide sequences (Stephens 1985; Sawyer 1989; Maynard Smith 1992) are more effective in detecting rare recombination events between sequences differing by 5% or more of nucleotides. The Homoplasy Test is, therefore, complementary to these earlier techniques rather than alternative. We applied the test to three sets of bacterial genes, and the results were broadly in line with expectations from other data. Simulations

show that the test is reasonably sensitive and is unlikely to give a “false positive.”

The Method

We propose a modification of the “cladistic” method suggested by Hudson and Kaplan (1985). Suppose that in a set of homologous gene sequences, there are v polymorphic sites. Hudson and Kaplan assume an “infinite-sites” model, which is equivalent to assuming that no site mutates more than once. Therefore, on the null hypothesis of no recombination, if t is the number of steps in a maximum-parsimony tree (MPT), then $v = t$. If there has been recombination, then, in general, $t > v$.

The method can be applied only if no site mutates twice. For most data sets, this will not be the case. For example, to generate 30 polymorphisms in a set of sequences consisting of 200 sites at which two bases are equally likely will, on average, require 32.5 mutations. In other words, if the number of sites at risk is finite, we expect to find homoplasies in an MPT even if there has been no recombination.

In defining a homoplasy, we make two assumptions: (1) the sequences analyzed arose by binary replication from a single common ancestor, without recombination, and (2) each site can exist in only one of two states; this assumption is discussed further below. The number of homoplasies in a tree can then be defined as follows.

True Homoplasies

The number of occasions on which the same site mutated independently in different links of the phylogenetic tree. Note that (1) if the same site mutates twice in the same link, this double event would not generate a polymorphism and is not counted as a homoplasy, and (2) if the same site mutated in three different links, this is counted as two homoplasies, and so on.

Apparent Homoplasies

If t is the number of mutational changes (events) in an MPT, and v is the number of polymorphisms, then

Key words: Homoplasy Test, effective number of sites, recombination, gene trees.

Address for correspondence and reprints: J. Maynard Smith, School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, United Kingdom.

Mol. Biol. Evol. 15(5):590–599. 1998

© 1998 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

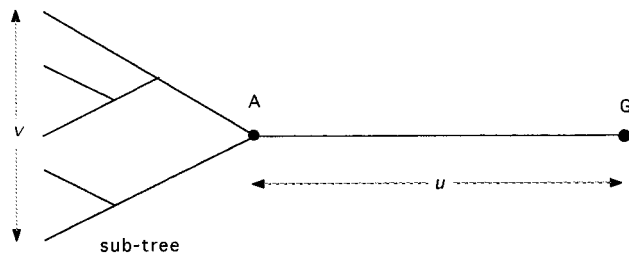


FIG. 1.—Use of an outgroup to estimate the effective number of sites, S_e . A, root of subtree; G, outgroup; u , number of site differences between A and G; v , number of sites polymorphic in subtree.

the number of apparent homoplasies is $h = t - v$. One can calculate h from the data (for example, by using the PAUP program [Swofford 1996]). We ask whether h is greater than the number expected in the absence of recombination. The expected number depends critically on S , the number of sites at risk. If S is large, then h will be small, and vice versa.

A further complication arises because not all sites at risk are equally likely to change. For example, a change at a third site is less likely if there is strong codon bias for that amino acid. We therefore define S_e , the “effective number of sites,” as follows. Suppose that we have two identical genes, obeying the same evolutionary rules concerning change per unit time (e.g., the same per-site mutation rates and codon preferences). A random evolutionary change occurs at one site in each of them. Let p_s be the probability that the same site changes in each of them. Then, $S_e = 1/p_s$; obviously, if there are S sites equally likely to change, then $S_e = S$.

The notation is then as follows:

v = number of sites polymorphic in a set of sequences,

t = number of steps in a maximum parsimony tree,

h = minimum number of apparent homoplasies = $t - v$,

S = number of sites at risk, and

S_e = effective number of sites, as defined above.

The procedure is as follows. First estimate S_e . Then estimate $P(h)$, the probability of having $\geq h$ apparent homoplasies on the null hypothesis of no recombination and given v and S_e . If $P(h)$ is sufficiently small, the null hypothesis can be rejected.

As stated above, we assume that each site can be in only one of two states. In the data sets analyzed below, the majority of changes are synonymous transitions (as opposed to transversions) at third sites, which fits the assumption. Third sites at which more than two synonymous bases occur have been classified into two classes, “commonest base” and “all others.” Codons at which there has been an amino acid change, or a change at one of the first two positions of the codon, have been omitted from the analysis. Little information has been lost by these simplifications.

Estimating S_e

S_e can be estimated if we have an outgroup (G, fig. 1), a sequence from a strain whose common ancestor

with the subpopulation is more distant than the common ancestor of the subpopulation. Let A be the root of the subtree. Let u = number of differences between A and G. A simple method for estimating u is given in appendix 1. Our estimate of S_e is

$$\hat{S}_e = 2u.$$

This estimate is based on three assumptions:

1. An equilibrium has been reached between forward and backward mutations in the lineage A–G: that is, “saturation” has been reached.
2. The value of u has been correctly estimated.
3. The “evolutionary rules” (i.e., the likelihood that a particular base will change) are the same within the subtree and in lineage A–G.

Consider these assumptions in turn. If saturation has not been reached in lineage A–G, then $2u$ is an underestimate of S_e . Also, the method of estimating u may lead to an underestimate, even if A–G is saturated, for reasons explained in appendix 1. There is one fact that could lead to $2u$ being an overestimate of S_e . This is that some amino acids are fourfold redundant. However, as explained above, we assume when estimating the number of homoplasies in the tree that each site exists in only two states. It is hard to estimate the exact effect of this assumption, but we think that, when it is taken into account, $S_e = 2u$ remains a reasonable assumption, even for fourfold redundant sites. Hence, provided the evolutionary rules have not changed, the method will underestimate S_e and, hence, overestimate the expected number of homoplasies in the absence of recombination. It is therefore conservative, in that it may fail to detect recombination when it has in fact happened.

An alternative method of estimating S_e , not using an outgroup, by allowing for codon bias is described in appendix 2. Applied to two classes of genes in *Escherichia coli*, genes with very high (VH) and medium high (MH) codon adaptation indices (Bulmer 1988), the method gives the following values:

VH genes: $\hat{S}_e = 0.73S$;

MH genes: $\hat{S}_e = 0.83S$.

We do not recommend the use of this method if the aim is to demonstrate the occurrence of recombination, because it ignores site-specific bias (Maynard Smith and Smith 1996) and hence may overestimate S_e .

However, the method is useful in two contexts. First, if the use of an outgroup leads to an estimate of $2u$ less than, say, $0.6S$, this suggests that lineage A–G is far from saturation, and a more distant outgroup should be sought. Second, if the occurrence of recombination has been established, and the aim is to determine its relative frequency by calculating a “homoplasy ratio” (see below), it may be better to estimate S_e by the method of appendix 2, because the outgroup method has been deliberately planned to underestimate S_e and, hence, be conservative.

If the “evolutionary rules” are different in the lineage A–G, this will lead to $2u > S_e$. An obvious sign that the rules have changed would be that $2u > S$. Then, either a different outgroup should be chosen, or S_e should be estimated as in appendix 2. More generally, it is important to choose an outgroup that is not too distantly related, and to check that G + C ratios and codon bias are similar.

Estimating $P(h)$ Given S_e and v A Simple Method

Draw sites at random, with replacement, from a set of S_e sites until v different sites have been drawn. If w is the number of draws needed, then the number of double hits is $d = w - v$. By repeating the sampling procedure, it is possible to calculate $P(h)$, the proportion of trials in which $d \geq h$.

The validity of using S_e has been checked by simulation. Consider, for example, a gene with 200 sites, of which 160 have probability p of changing, and 40 have probability $4p$ of changing. Then, $S_e = 128$. In 10,000 trials, the mean numbers of double hits when generating 40 polymorphisms were 7.32 for the real gene and 7.78 for a gene with 128 equally likely sites. This and other simulations showed that using S_e slightly overestimates the expected number of homoplasies and, hence, is conservative.

A More Precise Method

In effect, the simple method estimates the number of true homoplasies given v and S_e . The method is conservative, in that it may fail to demonstrate recombination even when it has occurred. This is because the number of apparent homoplasies, h , in an MPT may be lower than the number of true homoplasies estimated by the simple method. The reason is explained in figure 2. If the same site changed in two neighboring links of the tree, this will be interpreted as a single change in the “third link.” This error can be corrected by a method described in appendix 3. In marginal cases, the more precise method can demonstrate recombination when the simple method cannot. To summarize the method:

1. Estimate S_e , the effective site number. Take $S_e = 2u$, where u is calculated as in appendix 1. If $S_e < 0.6S$, then either choose a more distant outgroup or estimate S_e without using an outgroup but allowing for codon bias, as explained in appendix 2.
2. Given S_e and the number of polymorphic sites, v , estimate $P(h)$, either by the simple method or, if statistical significance is marginal, by the method of appendix 3.

QBASIC programs to calculate S_e can be obtained from http://epunix.biols.susx.ac.uk/Home/John_Maynard_Smith/.

Sources of Data

We applied the method to three examples. *Borrelia* was chosen because Dykhuizen et al. (1993) showed that the phylogenetic trees for two chromosomal genes were similar, suggesting that recombination is rare. In contrast, there is known to be extensive recombination

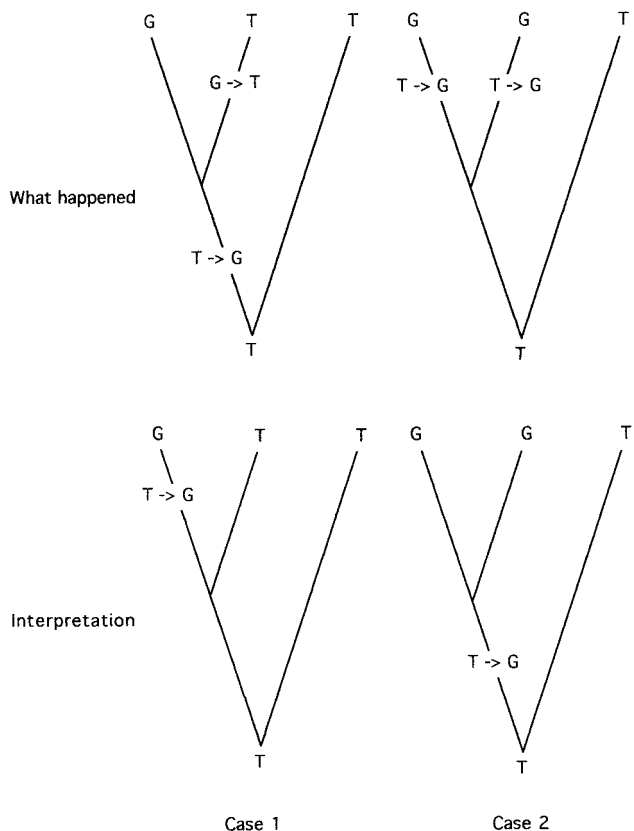


FIG. 2.—Underestimation of homoplasies. If the same site mutates twice in neighboring links, it will be interpreted as a single mutation in the “third link.”

in *Neisseria*, at least between closely related species (Spratt et al. 1995). A gene from *E. coli* was included, because there is some disagreement about the extent of recombination in this group of organisms (Guttman 1997).

The *Escherichia coli mdh* data set consisted of 12 unique partial sequences of *mdh* derived from strains of *E. coli* (subtree strains) and 1 sequence derived from *Salmonella typhimurium* (outgroup sequence). The sequences were 288 codons long and could be aligned without indels. GenBank accession numbers: *S. typhimurium*—M95049 (Lu and Abdelal 1993); *E. coli*—U04742, U04744, U04745, U04746, U04747, U04750, U04752, U04753, U04756, U04757, U04759, U04770 (Boyd et al. 1994).

The *Borrelia* flagellin data set consisted of seven unique sequences of flagellin genes derived from *Borrelia burgdorferi* and *Borrelia garinii* (subtree strains) and one sequence derived from *Borrelia hermsii* (outgroup sequence). The alignment of the outgroup sequence with the subtree sequences was 337 codons long and required the insertion of three codons in the outgroup sequence and the removal of one codon from the subtree sequences. Because our analysis is critically dependent on the correct alignment of the sequences, the region containing these two indels (21 codons) was deleted from the entire data set. The segments analyzed consisted of 316 codons and represented codons 1–202

Table 1
Analysis of Three Bacterial Genes

	<i>Bor- rel- ia Flag- ellins</i>	<i>Esch- eri- chia mdh</i>	<i>Neis- seria recA</i>
Strains in subtree	7	12	15
Total codons	316	288	237
S^a	274	266	210
Variable third sites			
In subtree			
Unique	4	18	18
Informative	11	14	19
Total in subtree (v)	15	32	37
Fixed in subtree but different in outgroup	79	83	53
Total	94	115	90
Steps to outgroup (u)	87	93	67
Homoplasies in subtree MPT (h)	1	9	18
Effective sites (\hat{S}_e)	174	186	134
\hat{S}_e/S	0.63	0.70	0.64
$P(\geq h \text{ homoplasies}) P(h)$	0.48	0.003	0.001
Homoplasies in randomized matrix			
\hat{h}_r (mean)	3.7	13.8	24.8
\hat{h}_r (range)	3–5	12–17	22–26
Homoplasies expected if clonal (\hat{h}_c)	0.68	3.08	6.09
Homoplasy ratio: $(h - \hat{h}_c)/(\hat{h}_r - \hat{h}_c)$	0.11	0.55	0.64

^a Sites with variable amino acids, changes at the first site of a codon, and sites coding for tryptophan and methionine residues excluded.

and 223–336 of the sequence of the *B. burgdorferi* flagellin. GenBank accession numbers: *B. burgdorferi*—L42881 (Livey et al. 1995), X69598 (Jauris-Heipke et al. 1993), X69610, X69609, X69608, X69614 (Dykhuisen et al. 1993); *B. hermsii*—M86838; *B. garinii*—X75203 (Noppa et al. 1995).

The *Neisseria recA* sequences consisted of 15 unique partial sequences of *recA* derived from strains of *Neisseria* closely related to *Neisseria meningitidis* (subtree sequences) and one sequence derived from a strain of *Neisseria mucosa* (outgroup sequence). The sequences were 237 codons long, representing codons 18–254 of the partial sequence of the *recA* gene of *Neisseria gonorrhoeae* strain FA19 and could be aligned without indels. GenBank accession numbers: *N. gonorrhoeae*—X64842; *N. meningitidis*—X64849, X64848, X64844, X64850, X64846, X64843; *Neisseria lactamica*—U57905, Y11818, Y11819; *Neisseria polysaccharea*—U57904, Y11814, Y11815, Y11816, Y11817; *N. mucosa*—U57908 (Zhou and Spratt 1992; unpublished data).

Applying the Method to Three Bacterial Genes

Table 1 shows the results of applying the method to three bacterial genes. In the analysis, only synonymous changes at third sites were included. The number of homoplasies observed by retaining the observed numbers of the two bases at each site but assigning them randomly to strains is given by \hat{h}_r . That is, they are the numbers expected if there is complete linkage equilibrium. The expected number of homoplasies in S_e sites if mutations are randomly applied to sequences to generate the number of polymorphic sites seen in each data

set is given by \hat{h}_c . The values of $P(h)$ and \hat{h}_c were calculated without the correction shown in appendix 3. The final line in table 1 calculates a “homoplasy ratio,” $(h - \hat{h}_c)/(\hat{h}_r - \hat{h}_c)$, where \hat{h}_r is the mean value of \hat{h}_r in 10 trials. This number is expected to be approximately 0 if there is no recombination and 1.0 if there is random assortment. It is, therefore, a measure of the importance of recombination, relative to mutation, in determining the pattern of variation in a population.

The results were as expected. There was no sign of recombination in *Borrelia*. In *Neisseria*, the evidence for recombination was overwhelming, although the number of homoplasies was not as high as it would be with complete linkage equilibrium between sites. The evidence for recombination in the *mdh* gene of *E. coli* was also strong, as suggested by Boyd et al. (1994), who detected a significant clustering of polymorphic sites in one of the *mdh* alleles by the application of Stephens’ test. However, it is interesting to note that the Homoplasy Test also detects evidence for recombination in the *mdh* alleles of *E. coli* if the recombinant sequence detected by Boyd et al. (1994) is eliminated from the data set (data not shown). Thus, Stephens’ test detected recombinant fragments originating from outside the data set, while the Homoplasy Test detects recombination among the sequences under analysis.

Simulated Trees

To estimate the likely effectiveness of the Homoplasy Test, we generated trees by simulation, and then analyzed them. The trees were generated as follows:

1. Sixteen sequences were produced from a single ancestor by four successive doublings with no deaths.
2. Thirty-six mutational changes were simulated at sites chosen randomly, with replacement, from 200 equivalent sites, at each of which two bases were possible. The mutations occurred in randomly chosen links in the tree.
3. Varying numbers, R , of recombinational events were simulated, each involving the transfer of 100 sites from a randomly selected donor to a randomly selected recipient, with the requirement that the donor and recipient were not identical.
4. $P(h)$ was calculated assuming 200 effective sites, using the simple method and 1,000 trials.

Some results are shown in figure 3 and table 2. The following conclusions can be drawn:

1. The method is unlikely to give a “false positive”—that is, to indicate that recombination has happened when reproduction was in fact clonal (table 2).
2. If the number of recombinational events affecting the sequence is at least one half the number of mutations per sequence ($R = 14, 28$, or 56 ; table 2), the method is likely to detect recombination. The number of polymorphic sites decreased with R , because nonreciprocal recombination is a homogenizing force. The effect is substantial only because of the small size of the simulated populations; it would be negligible in real populations.

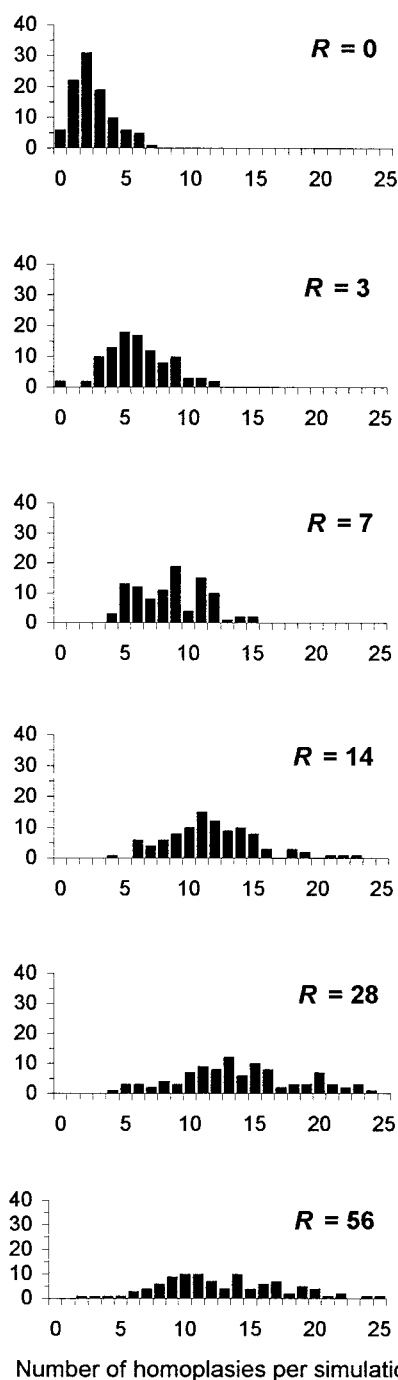


FIG. 3.—Simulations to show the likely effectiveness of the Homoplasy Test. Each simulation generated 16 sequences of 200 equivalent sites produced by four replications from a single ancestor and incorporating 36 mutational changes. Varying numbers, R , of non-reciprocal recombination events were simulated, each involving the transfer of 100 sites. The simulations were repeated 100 times for each value of R . The number of sequences with the corresponding number of homoplasies is plotted.

3. With lower numbers of recombinations, it will sometimes be possible to detect recombination.

We also compared the number of homoplasies in simulated trees with the number expected with free recombination and, hence, no linkage disequilibrium.

Table 2
Application of the Homoplasy Test to Simulated Phylogenies^a

NUMBER OF RECOMBINATIONS (R)	MEAN NUMBER OF:			EVIDENCE FOR RECOMBINATION ^b
	Polymorphic Sites (v)	Apparent Homoplasies (h)	Homoplasies Expected if Clonal (\hat{h}_c)	
0	34.0	2.48	3.00	1
3	30.6	6.03	2.76	20
7	30.4	8.60	2.52	61
14	28.3	11.80	2.12	90
28	24.5	13.78	1.57	97
56	20.4	12.75	0.94	99

^a Ten simulations for each value of R involving 200 equivalent sites and 36 mutational changes. One hundred sites transferred per recombination.

^b Number of cases in 100 simulations in which there was evidence ($P(h) < 0.01$) of recombination. $P(h)$ is the probability of getting at least the observed number of homoplasies on the hypothesis of no recombination using the "simple" method.

Figure 4 shows the results of some simulations in which, after a tree had been generated, and the number of apparent homoplasies had been estimated, the genotypes of the strains were randomized, retaining the observed numbers of bases at each site but assigning them randomly to sequences. Except for 7 of 100 cases when $R = 56$, the number of homoplasies in the randomized tree was always greater than the number in the simulated one.

Discussion

This paper suggests a method of detecting whether recombination has occurred, or, more precisely, whether it has played an important role in determining the pattern of variation in a population. The method is primarily aimed at detecting recombination in prokaryotes, or other populations (e.g., mitochondria) in which a recombinational event consists of the replacement of a relatively small block of DNA (10^2 – 10^5 nucleotides) in a recipient by a homologous block from a donor, but it could also be applied to reciprocal recombination in sexual diploids.

Several methods of detecting recombination are possible:

1. Compare two sequences, and look for a "mosaic structure" of blocks of high similarity interspersed with blocks of high sequence divergence (Maynard Smith 1992). Methods of this type have been applied to *E. coli* (e.g., Milkman and Crawford 1983; Du Bose, Dykhuizen, and Hartl 1988; Stoltzfus and Milkman 1988; Guttman and Dykhuizen 1994). The effects are more dramatic and hence easier to detect in *Streptococcus* (Dowson et al. 1989) and *Neisseria* (Spratt et al. 1995), because transfers have occurred between strains differing by 20% of nucleotides. In *Neisseria*, it has been possible to identify both the donor and the recipient species, as well as the "hybrid." Given sufficient nucleotide difference, this is an effective method, but it is hard to apply when differences are 1% or less.

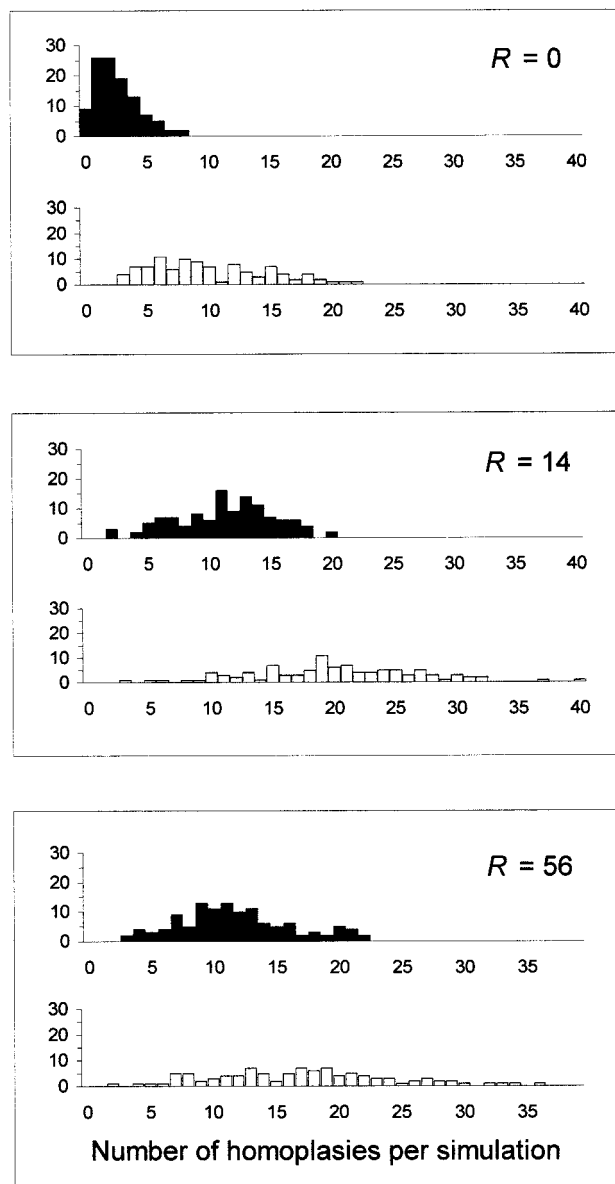


FIG. 4.—Comparison of the observed number of homoplasies with the number expected with linkage equilibrium. Populations were simulated as in figure 3. For each simulation, the number of homoplasies was calculated, as was the number of homoplasies in a population obtained by randomizing the matrix of genotypes while retaining the numbers of each allele at each site. Filled boxes: observed homoplasies. Open boxes: homoplasies in randomized matrix.

Sawyer's (1989) test for recombination calculates an index of the mosaic structure within a set of sequences and is used to identify data sets that contain significant runs of similar bases. Runs of similar bases in sequences that are otherwise dissimilar are presumed to be generated by recombination. We suggest that Sawyer's test be interpreted with care. The test is conservative in that it will fail to find evidence for recombination for data sets in which the polymorphic sites are in linkage equilibrium. Furthermore, a data set containing two very similar sequences and a divergent sequence may score more highly in Sawyer's test than if a seg-

ment of sequence had been transferred nonreciprocally from the divergent sequence to one of the other sequences.

2. Compare phylogenies derived from different blocks of linked loci. This method has been applied in three contexts:

- Stephens (1985) proposed the following method of analyzing a set of sequences of the same gene. Identify sites that support different phylogenetic partitions of the data. If sites supporting the same partition are linked, this suggests that there has been horizontal transfer of blocks of linked sites. This approach has been used to detect recombination between distantly related strains of enteric bacteria (Nelson and Selander 1992; Boyd et al. 1994).
- Dykhuizen and Green (1991) suggested that phylogenetic trees be constructed using several genes from the same group of strains. If the same tree is obtained for different genes, recombination between genes must be rare, and hence it is reasonable to place the strains in different species, whereas if different trees are obtained, recombination has been common, and the strains belong to a single species. The method has been used by Dykhuizen et al. (1993) to demonstrate the absence of recombination in *Borrelia*.
- Recently, Grassly and Holmes (1997) described a method to detect regions of genes that do not fit in with the maximum-likelihood evolutionary model of the entire data set. Subsequent reanalysis of the suspect region may suggest that recombination is the cause of the anomaly. The authors used this approach to confirm that regions of the *argF* gene of *Neisseria* had been subject to recombination.

3. Given a set of sequences, compare the number of apparent homoplasies (that is, two changes at a single site) in a most-parsimonious tree with the number expected in the absence of recombination. This method was first suggested by Hudson and Kaplan (1985). They analyzed an "infinite-alleles" model, which is equivalent to assuming no homoplasies. Hence, any apparent homoplasies in the tree refute the null hypothesis of no recombination. Hudson and Kaplan applied the method to Kreitman's (1983) data on *Adh* in *Drosophila*, in which there were only 43 polymorphic sites in 2,700 bases (900 third sites). Hence, provided that changes at different third sites were approximately equally likely, the infinite-sites model is a reasonable approximation. But in many data sets, including those used as illustrative examples in this paper, the assumption is not justified. We therefore suggest a method of estimating the effective site number, S_e . If there were S_e sites at which change was equally likely, then the expected number of homoplasies (in the absence of recombination) would be the same as the number expected for the real data, with different probabilities of change at different sites. Given S_e , one can estimate the probability, $P(h)$, of getting at least as high a number of homoplasies as is actually observed, on the null hypothesis of no recombination.

If $P(h)$ is sufficiently small, the null hypothesis can be rejected.

These three methods detect departures from clonality. A complementary method is available to detect departures from random assortment. The method is based on the value of S_e^2 , the variance in the number of site differences between pairs of sequences in a sample. It was first proposed by Sved (1968) and was applied by Brown, Feldman, and Nevo (1980) to plant data, and by Whittam, Ochman, and Selander (1983) and Maynard Smith et al. (1993) to bacteria.

The method proposed here is likely to be useful when sequences are available from a set of closely related organisms, differing by approximately 1%–5% of nucleotides. The estimate of S_e depends on the use of an outgroup and on the assumption that the evolutionary rules (likelihood of change per unit time at different sites) were the same in the lineage leading to the outgroup as within the tree. This means that an outgroup should not be too distantly related and should have the same G + C ratio and codon bias. It also makes the method unsuitable for DNA coding for RNA with a secondary structure, because, in such cases, change at one nucleotide will alter the likelihood of change at other sites.

One alternative to the present method is that proposed by Stephens (1985). The present method is easier to use, and we think it will prove to be more sensitive. This may seem surprising, since our method ignores proximity of sites within a gene. However, for the data sets analyzed here, simulated and real, we performed the following test of the effect of proximity. After finding a maximum-parsimony tree, we counted the number of cases in which neighboring polymorphic sites occurred in the same link of the tree (as expected from recombination) and compared this number with those of trees in which the same “events” (changes at sites) were applied randomly to a tree with the same topology. The observed number of neighbors within a link was not significantly greater in the real trees than in the random trees. This suggests, although it does not prove, that information about proximity of sites is not particularly informative.

Our method does not measure the rate of recombination, but only its occurrence. However, it is possible, as shown in table 1, to compare the observed number of apparent homoplasies with the numbers expected for the null hypothesis of no recombination, and for free recombination.

Is it possible to measure a recombination rate? It is not easy. A preliminary difficulty is to decide what it is we want to measure. In classical diploid genetics, the recombination rate is defined as the probability that two genes, or sites, on a chromosome inherited from a single parent will be transmitted to different offspring. It can be measured by counting offspring from doubly heterozygous parents. In bacteria, a similar measure is the cotransduction frequency of two genes. However, measurements of cotransduction frequency will not illuminate the impact of recombination on the population structure or evolution of bacteria. Instead, we can ask

how the frequency of horizontal transfer depends on genetic distance between donor and recipient: for example, Roberts and Cohan (1993) showed that, in *Bacillus*, the frequency of transformation declines, in a log linear fashion, with sequence divergence. We can also ask (e.g., McKane and Milkman 1995) what is the size of the DNA fragments transferred in a single event and compare this to the scale of the mosaic structure found in natural populations. Neither type of experiment tells us the absolute rate of recombination in nature. The difficulty is that the relevant rate of recombination (relevant in the sense of determining patterns of variation in nature) depends not only on the way in which the frequency of transfer depends on genetic distance and on the size of fragments transferred, but also on the frequency with which different strains, perhaps adapted to different niches, actually meet in nature. How often does *N. gonorrhoeae* meet *N. meningitidis*?

In models of bacterial evolution (e.g., Cohan 1994; Maynard Smith 1994), the relevant parameter is the probability, per unit time and per individual, that a particular DNA region will be replaced by homologous DNA from a donor compared with the probability that the same region will change by mutation. Unfortunately, what we need is not a single value, but some idea of how this probability of replacement varies with genetic distance, because the probability of genetic change depends both on the probability of transfer and on the probability that the introduced block differs from the block it replaces. For example, if transfer were frequent but occurred only between identical sequences, it would have no genetic effects. Perhaps, as suggested by Guttman and Dykhuizen (1994), the number that would be most useful to estimate for a bacterial population is the probability that a nucleotide at a site will alter as a result of recombination compared with the probability that the site will alter by mutation.

Acknowledgments

N.H.S. was supported by a Wellcome Trust grant to Professor B. G. Spratt. We thank Dr. Adam Eyre-Walker for stimulating discussion.

APPENDIX 1

Estimating S_e Using an Outgroup

The number of differences between the common ancestor of the subtree (A) and the outgroup (G) is u (fig. 1). An estimate of S_e is $\hat{S}_e = 2u$.

For any data set, the values in table 3 can be calculated using a program such as MEGA (Kumar, Tamura, and Nei 1993). An “informative” site is one for which the rarer allele is present in at least two strains in the relevant data set; thus, g_1 is the number of sites that are informative in the full data set, including the outgroup, and g_2 is the number of informative sites in just the ingroup. Hence, a site that is merely variable in the ingroup may become informative when the outgroup is added. A “variable” site is any site, informative or not, that varies in the relevant data set. The number of

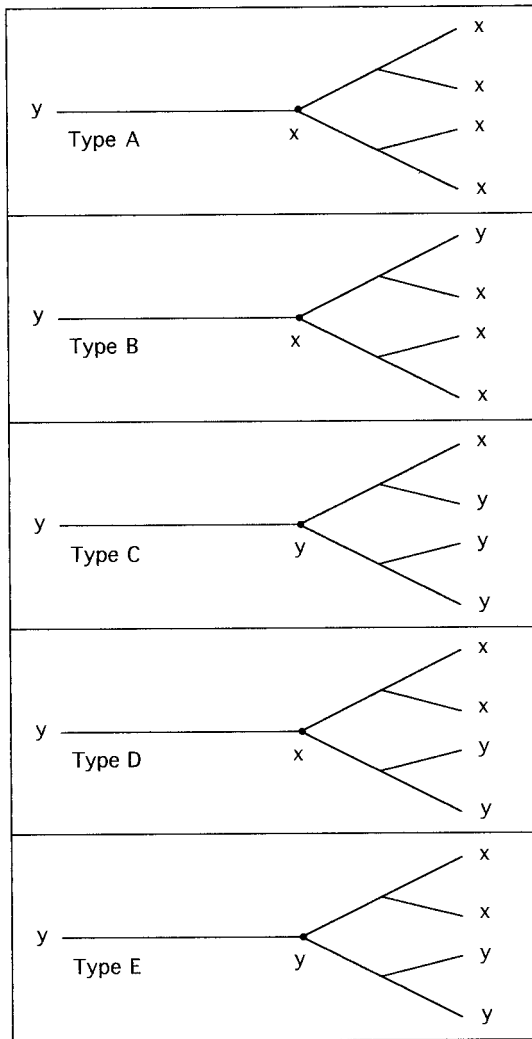


FIG. 5.—The five types of variable sites considered when estimating S_e using an outgroup. The numbers of sites of types A, B, C, D, and E are a , b , c , d , and e . For sites of types B and C there is a unique ingroup strain with an allele different from all other ingroup strains.

steps in an MPT is for informative sites only: that is the number delivered by the MEGA program.

Figure 5 shows five kinds of sites, A–E, with numbers a – e . For sites of kinds B and C, there is a unique ingroup strain that differs from all other ingroup strains. Then, $u = a + b + d$. Clearly, $a = v_1 - v_2$. To find $b + d$ we proceed as follows: For the full data set, $g_1 =$ (number of steps involving sites D and E) + (number of steps involving sites B). If we choose an MPT which includes a subtree of minimal length g_2 , then $g_1 = g_2 + d + 2b$ or $g_1 - g_2 = d + 2b$.

Now,

$$h_1 = g_1 - c_1 = g_1 - b - d - e,$$

and

$$h_2 = g_2 - c_2 = g_2 - d - e.$$

Hence, $h_1 - h_2 = g_1 - g_2 - b = b + d$. It follows that

Table 3
Estimating S_e Using an Outgroup

	Full Data Set (including outgroup)	Subtree
Variable sites	v_1	v_2
Informative sites	c_1	c_2
Steps in an MPT ^a	g_1	g_2
Homoplasies	$h_1 = g_1 - c_1$	$h_2 = g_2 - c_2$

^a For informative sites only.

$$a + b + d = v_1 - v_2 + h_1 - h_2.$$

Therefore, to estimate $S_e = 2u$, we calculate the values in table 3 and take

$$u = v_1 - v_2 + h_1 - h_2.$$

This method may underestimate u . If we insist that the tree for the full data set is an MPT, this will result in a bias in favor of sites of type E, whereas at saturation, sites of types D and E are equally likely. Hence, we will underestimate $u = a + b + d$. In other words, the maximum-parsimony assumption forces us to assume that sites are of type D rather than type E and so to underestimate u .

APPENDIX 2

Estimating S_e Allowing for Codon Bias

Suppose there are S sites in a gene, at each of which two bases are possible, with frequencies p of the common allele and q of the rare allele. At equilibrium, the probability that a random substitution will affect a common allele is 0.5, and the probability is similar for a rare allele. If two random changes occur, the probability that both affect a common allele, or that both affect rare alleles, is 0.25. Hence, the probability P that two random substitutions affect the same site is $0.25(1/Sp + 1/Sq) = 0.25/Spq$. Hence, $\hat{S}_e = 4Spq$.

Thus, for no bias, $p = q = 0.5$, and $\hat{S}_e = S$, as expected. If, for example, $p = 0.8$ and $q = 0.2$, then $\hat{S}_e = 0.64S$.

In practice, not all sites have the same bias. Suppose that, for a given amino acid, the bias is $p : q$, where $p > q$. Then (Kimura 1957), the probability that a mutation from a common to a rare base is established is $P_s = 2rs/(1 - r)$, and of a mutation from a rare to a common base is $P_r = 2s/(1 - r)$, where s is the selective advantage of the favored codon, and $r = e^{-2Ns}$. Then, at equilibrium, $p = 1/(1 + r)$ and $q = r/(1 + r)$.

Let m be the mutation rate, and let P_c be the probability that a site changes per unit time. If there are n codons for a particular amino acid, then, summing over all codons,

$$\begin{aligned} \sum_n P_c^2 &= \frac{n}{1+r} \left(\frac{2msr}{1-r} \right)^2 + \frac{nr}{1+r} \left(\frac{2ms}{1-r} \right)^2 \\ &= \frac{nr(2ms)^2}{(1-r)^2}, \end{aligned}$$

$$\text{and } \sum_n P_c = \frac{4mnsr}{1-r^2}.$$

Summing over the 20 amino acids,

$$\sum P_c^2 = \sum_i \frac{n_i(2ms_i)^2 r_i}{(1 - r_i)^2} = \frac{m^2}{N^2} \sum_i \frac{n_i(2Ns_i)^2 r_i}{(1 - r_i)^2},$$

and $\sum P_c = \frac{m}{N} \sum_i \frac{2n_i(2Ns_i)r_i}{1 - r_i^2}.$

$$\text{Then, } \hat{S}_e = \frac{\left(\sum P_c\right)^2}{\sum P_c^2}.$$

Knowing for each amino acid the values of n and p , we can calculate $r = 1/(1 - p)$ and $2Ns = \ln(1/r)$, and hence \hat{S}_e . The calculation has been carried out for the observed amino acid frequencies and codon usages for genes with VH and MH codon adaptation indices (Bulmer 1988), with the following results:

for VH genes: $\hat{S}_e = 0.736S$;

for MH genes: $\hat{S}_e = 0.828S$.

There are two sources of inaccuracy in this method of estimation. First, it ignores site-specific bias (Maynard Smith and Smith 1996). This will lead to an overestimate of S_e and hence will be nonconservative. Second, it assumes the same mutation rate for all changes. Since for fourfold and sixfold redundant sites, we classify bases as “commonest” and “all others,” this is inaccurate: it is not clear what type of error this introduces.

APPENDIX 3

A More Precise Method of Estimating the Expected Number of Apparent Homoplasies

As explained in the main text, the number of apparent homoplasies in an MPT may be less than the true number, because if the same site changes in two neighboring links (i.e., links originating in the same node) in the tree, this will be interpreted as a single nonhomoplastic change (fig. 2). The approximate magnitude of this effect can be calculated as follows. Let there be v polymorphic sites, of which c are informative: then there will be $v - c$ unique changes in the terminal links of the phylogenetic tree. If there are f different strains, there will be f terminal links and $f - 3$ other links, which we will call “interior.” Each terminal link has two neighbors, and each interior link has four neighbors.

We want to estimate the probability, P_t , that if the same site changes twice in different links, the two links are neighbors. Note that if the same site changes twice in the same link, it will not appear as a polymorphism and can be ignored. Then, $P_t = P(\text{first change is in a terminal link}) \times P(\text{second change is in a neighboring link} \mid \text{first change is in a terminal link}) + P(\text{first change is in an interior link}) \times P(\text{second change is in a neighboring link} \mid \text{first change is in an interior link})$. If we assume changes are equally likely to occur in any of the $2f - 3$ links, then:

$$P_t = \frac{v - c}{v} \times \frac{2}{2f - 4} + \frac{c}{v} \times \frac{4}{2f - 4} = \frac{v + c}{v(f - 2)}.$$

If the true number of homoplasies was h_v , then the expected number of homoplasies in the most parsimonious

tree is $h_v(1 - P_t)$. If h_v is estimated from S_e and v , the correction $(1 - P_t)$ can be applied; it is only approximate, because it is based on the assumption that changes are equally likely to occur in any link.

LITERATURE CITED

- BOYD, E. F., K. NELSON, F. S. WANG, T. S. WHITTAM, and R. K. SELANDER. 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. *Proc. Natl. Acad. Sci. USA* **91**:1280–1284.
- BROWN, A. H. D., M. W. FELDMAN, and E. NEVO. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* **96**:523–536.
- BULMER, M. 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *J. Evol. Biol.* **1**:15–26.
- COHAN, F. M. 1994. Genetic exchange and evolutionary divergence in prokaryotes. *TREE* **9**:175–180.
- DOWSON, C. G., A. HUTCHISON, J. A. BRANNIGAN, R. C. GEORGE, D. HANSMAN, J. LIÑARES, A. TOMASZ, J. M. SMITH, and B. G. SPRATT. 1989. Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Proc. Natl. Acad. Sci. USA* **86**:8842–8846.
- DU BOSE, R. F., D. E. DYKHUIZEN, and D. L. HARTL. 1988. Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **85**:7036–7040.
- DYKHUIZEN, D. S., and L. GREEN. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J. Bacteriol.* **173**:7257–7268.
- DYKHUIZEN, D. E., D. S. POLIN, J. J. DUNN, B. WILSKA, V. PREAC-MUSRIC, R. J. DATTWYLER, and B. J. LUFT. 1993. *Borrelia burgdorferi* is clonal: implications for taxonomy and vaccine development. *Proc. Natl. Acad. Sci. USA* **90**:10163–10167.
- GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**:239–247.
- GUTTMAN, D. S. 1997. Recombination and clonality in natural populations of *Escherichia coli*. *TREE* **12**:16–21.
- GUTTMAN, D. S., and D. E. DYKHUIZEN. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**:1380–1383.
- HUDSON, R. R., and N. L. KAPLAN. 1985. Statistical properties in the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**:147–164.
- JAUROS-HEIPKE, S., R. FUCHS, M. MOTZ, V. PREAC-MURSIC, E. SCHWAB, E. SOUTSCHEK, G. WILL, and B. WILSKA. 1993. Genetic heterogeneity of the genes coding for the outer surface protein C (OspC) and the flagellin of *Borrelia burgdorferi*. *Med. Microbiol. Immunol. (Berl.)* **182**:37–50.
- KIMURA, M. 1957. Some problems of stochastic processes in genetics. *Ann. Math. Stat.* **28**:882–901.
- KREITMAN, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**:412–417.
- KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. Version 1.01. The Pennsylvania State University, University Park.
- LIVEY, I., C. P. GIBBS, R. SCHUSTER, and F. DORNER. 1995. Evidence for lateral transfer and recombination in OspC variation in Lyme disease *Borrelia*. *Mol. Microbiol.* **18**:257–269.

- LU, C.-D., and A. H. T. ABDELAL. 1993. Complete nucleotide sequence of the *Salmonella typhimurium* gene encoding malate dehydrogenase. *Gene* **123**:143–144.
- MCKANE, M., and R. MILKMAN. 1995. Transduction, restriction and recombination patterns in *Escherichia coli*. *Genetics* **139**:35–43.
- MAYNARD SMITH, J. 1992. Analysing the mosaic structure of genes. *J. Mol. Evol.* **34**:126–129.
- . 1994. Estimating the minimum rate of genetic transformation in bacteria. *J. Evol. Biol.* **7**:525–534.
- MAYNARD SMITH, J., and N. H. SMITH. 1996. Site-specific codon bias in bacteria. *Genetics* **142**:1037–1043.
- MAYNARD SMITH, J., N. H. SMITH, M. O'ROURKE, and B. G. SPRATT. 1993. How clonal are bacteria? *Proc. Natl. Acad. Sci. USA* **90**:4384–4388.
- MILKMAN, R., and I. P. CRAWFORD. 1983. Clustered third-base substitutions among wild strains of *Escherichia coli*. *Science* **221**:378–380.
- NELSON, K., and R. K. SELANDER. 1992. Evolutionary genetics of the proline permease gene (*putP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. *J. Bacteriol.* **174**:6886–6895.
- NOPPA, L., N. BURMAN, A. SADZIENE, A. G. BARBOUR, and S. BERGSTROM. 1995. Expression of the flagellin gene in *Borrelia* is controlled by an alternative sigma factor. *Microbiology* **141**:85–93.
- ROBERTS, M. S., and F. M. COHAN. 1993. The effect of DNA sequence divergence on sexual isolation in *Bacillus*. *Genetics* **134**:401–408.
- SAWYER, S. A. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**:526–538.
- SPRATT, B. G., N. H. SMITH, J. ZHOU, M. O'ROURKE, and E. FEIL. 1995. The Population genetics of the pathogenic *Neisseria*. Pp. 143–160 in S. BAUMBERG, J. P. W. YOUNG, E. M. H. WELLINGTON, and J. R. SAUNDERS, eds. The population genetics of bacteria. SGM Symposium No. 52. Cambridge University Press, Cambridge, England.
- STEPHENS, J. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* **2**:539–556.
- STOLTZFUS, A., and R. MILKMAN. 1988. Molecular evolution of the *Escherichia coli* chromosome II. Clonal segments. *Genetics* **120**:359–366.
- SVED, J. A. 1968. The stability of linked systems of loci with a small population size. *Genetics* **59**:543–563.
- SWOFFORD, D. L. 1996. PAUP: phylogenetic analysis using parsimony (and other methods). Version 4.0. Sinauer, Sunderland, Mass.
- WHITTAM, T. S., H. OCHMAN, and R. K. SELANDER. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **80**:1751–1755.
- ZHOU, J., and B. G. SPRATT. 1992. Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Mol. Microbiol.* **6**:2135–2146.
- STANLEY A. SAWYER, reviewing editor

Accepted January 22, 1998