The Detection and Measurement of Recombination From Sequence Data

J. Maynard Smith

School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom

Manuscript received March 22, 1999
Accepted for publication June 7, 1999

ABSTRACT

There are two types of recombination that we may wish to detect: rare recombinants between members of different populations or species and repeated recombination within a population. Methods appropriate in the former context are inappropriate in the latter because they depend on recognizing the existence of runs of nucleotides with similar ancestry. If recombination is sufficiently frequent, no such runs will be present. Several methods, including the homoplasy test and the incompatibility test, are described that are appropriate for detecting repeated recombination and for measuring its importance, relative to mutation, in causing genetic change. The sensitivity of these tests is investigated by simulating populations with varying frequencies of mutation and recombination and calculating the various statistics on samples.

THIS article is concerned with the use of DNA sequence data to detect and measure the role of homologous recombination in natural populations. Its most obvious relevance is to bacteria, which vary from the clonal to the effectively panmictic (Maynard Smith and Smith 1998; Suerbaum *et al.* 1998) and in which recombination is decoupled from reproduction. However, the methods are also relevant to eukaryotes in cases in which the role of recombination is uncertain—for example, between mitochondrial genomes (Eyre-Walker *et al.* 1999; Hagelberg *et al.* 1999) and in organisms whose reproduction is apparently parthenogenetic.

The main point to be made is that there are two very different contexts in which we may wish to detect recombination. Methods appropriate in one context may be ineffective in the other. We may be seeking either (i) unique or rare recombination between members of genetically different populations—for example, between different species or between different parts of the genome; or (ii) repeated recombination between homologous sites in members of a single population.

Unique or rare events lead to linked runs of nucleotides within a sequence whose ancestry is different from other nucleotides in the same sequence. Several methods exist for detecting such runs: they are described only briefly.

With repeated recombination, such runs may not exist or may be hard to detect. If recombination is frequent enough, the sites within a gene will be in linkage equilibrium: that is, there will be no association between neighboring nucleotides. Even with rates of recombination far too low to generate linkage equilibrium, it is shown that methods aimed at detecting runs of nucleotides with similar ancestry are ineffective. Methods of detecting repeated recombination are described. Ultimately, they depend on the fact that, if there has been recombination, the pattern of variation in a set of sequences is incompatible with the hypothesis of clonal descent: in this article, "clonal" is taken to mean reproduction without genetic recombination. In principle, these methods can be used not only to detect recombination but also to measure its importance, relative to mutation, in generating change. The effectiveness of these tests is investigated by applying them to simulated populations of varying size, mutation rate, and recombination rate.

RARE OR UNIQUE RECOMBINATION

Several methods exist for detecting runs of nucleotides with similar ancestry. Stephens (1985) examined sites that are associated with particular phylogenetic partitions of the set of sequences into two groups: thus a site is associated with a partition if it has one allele in one subset and a different allele in the other. He describes statistical tests to determine whether a set of sites associated with a particular partition are clustered. Sawyer (1989) gives a more general method for deciding whether the differences, or similarities, between a pair of sequences occur in runs: the method can also be applied to a set of sequences.

If a visual inspection of the polymorphic sites in a set of sequences suggests that one or more recombinational events have occurred, the maximum chi-square method (Maynard Smith 1992) will locate the most likely positions of the crossovers and test their statistical significance. The method was used to identify the role of horizontal gene transfer in the spread of antibiotic resistance in Streptococcus (Dowson *et al.* 1989) and Neisseria (Spratt *et al.* 1992). Hein (1990) showed how maximum-likelihood methods can be used to locate the position of crossovers. Holmes *et al.* (1999), in an analysis of recombination in Dengue virus, described a maximum-likelihood method that can be used if a putative

recombinant and two parental sequences are suggested by the data.

Sneath *et al.* (1975) proposed a method that depends on identifying incompatible pairs of sites. They considered protein sequences, but the method is equally applicable to nucleotide sequences and is discussed here in that context. Two sites are incompatible if they cannot be incorporated into a phylogenetic tree without assuming that one of the sites has changed twice. If each site is present in the data set in only two states, as is typical for nucleotide data, the pair are incompatible if all four genotypes, 00, 01, 10, and 11, are present. An incompatibility matrix can be plotted for a set of sequences, for all phylogenetically informative sites (that is, sites at which both alleles are present in at least two strains). If there are n informative sites, an $n \times n$ matrix is plotted, with black squares for incompatible sites. Jakobsen and Easteal (1996) describe a program to construct incompatibility matrices and to test whether incompatible sites cluster. Jakobsen *et al.* (1997) describe a program that plots partition matrices: in effect, this is a visualization of Stephens' method described above. It is not clear that this offers any advantage over eyeballing a simple listing of polymorphic sites, combined with using the maximum chi-square method to test the significance of possible recombinants.

All the methods described in this section depend on the existence of runs of polymorphic sites with a similar ancestry. They are therefore unsuitable for detecting repeated recombination within a population.

REPEATED RECOMBINATION WITHIN A POPULATION

There are several methods for detecting repeated recombination. First, the logic of these tests is described, and then their sensitivity is investigated by applying them to simulated populations.

The homoplasy test: This test was described by Maynard Smith and Smith (1998). It can be applied to a set of sequences of a single gene or of several unlinked genes from the same strains. The logic is as follows. Construct a maximum-parsimony tree for a set of sequences. If there are v polymorphic sites, and t steps in the tree, then $h = t - v$ is the number of homoplasies, or double events. If, as is usual, there are several equally parsimonious trees, this does not matter because it is only the total number of steps that matters. If descent is clonal and the number of sites infinite, then $h = 0$. If there has been recombination, however, in general $h > 0$. This fact was used by Hudson and Kaplan (1985) to estimate recombination rate.

Difficulties arise when mutation is so common that there can be repeated mutations at the same site. If so, there may be homoplasies even in a clonal population. Maynard Smith and Smith discuss how this difficulty can be overcome. It is usually best to confine attention to synonymous third sites: little information is lost. If there are S synonymous sites, all equally likely to change, it is easy to calculate exph, the expected number of homoplasies with clonal reproduction, given v polymorphic sites. If sites are not equally likely to change, an estimate of exph requires an estimate of S_e, the effective site number, defined as follows. Consider two identical copies of a gene, obeying the same evolutionary rules, and let each undergo a random substitution. Let p_s be the probability that the two substitutions are identical. Then $S_e = 1/p_s$. Clearly, if all sites are equally likely to change, $S_e = S$; otherwise, $S_e < S$.

For synonymous sites, codon bias is the most likely reason why the probability of change should vary between sites. Given a known pattern of codon usage, Maynard Smith and Smith describe a method for calculating S_e: applied to *Escherichia coli*, this gives $S_e = 0.73$ S for genes with a very high codon adaptation index and $S_e = 0.83$ S for genes with a medium high index (data from Bulmer 1988). Alternatively, S_e can be estimated using an outgroup. Further difficulties arise if some sites are hypermutable. Methods for detecting hypermutability, if it exists, are described by Eyre-Walker *et al.* (1999): in general, such methods require an outgroup.

If the observed number of homoplasies, obsh, is significantly greater than exph, the plausible explanation is recombination. The extent of recombination is measured by the "homoplasy ratio,"

$$H = (\text{obsh} - \text{exph})/(\text{shh} - \text{exph}),$$

where shh is the number of homoplasies for a population with the observed variation at each site, but in linkage equilibrium (estimated by randomizing the alleles at each site between sequences and recalculating h). H is a number whose expectation varies from 0 (clonal) to 1.0 (complete linkage equilibrium).

The incompatibility ratio: If more than a very few recombinational events have taken place in the ancestry of a set of sequences, patterns in the incompatibility matrix become hard to interpret. However, for a given degree of polymorphism, recombination increases the proportion of incompatible sites. This suggests the use of the incompatibility ratio, (IR) as a statistic, where IR = (number of pairs of sites that are incompatible) ÷ (number incompatible in a shuffled matrix). The data set is the matrix of phylogenetically informative sites, and, as for the homoplasy ratio, a shuffled matrix is one in which, at each site, the alleles have been randomly shuffled between strains.

IR has one advantage and one disadvantage compared to H. The advantage is that it is easier to calculate the proportion of incompatible sites than to find a maximum-parsimony tree, particularly for a large data set. The disadvantage is that its expected value for a clonal population is not known and is certainly not zero so that, in most cases, it cannot be used to test departure

from clonality. However, if the effective site number S_e is very large compared to the number of polymorphic sites, then the expected number of incompatible pairs in the absence of recombination is close to zero, and this difficulty does not arise.

The index of association: A third possible measure is related to the index of association, (I_A), which has been used to analyze multiple-locus enzyme polymorphism (Brown *et al.* 1980; Maynard Smith *et al.* 1993). This is based on V_{obs}, the variance of the genetic distance between pairs of strains, compared to V_{exp}, the corresponding variance in a shuffled matrix. The expected value of the ratio V_{obs}/V_{exp} is 1.0 for complete linkage equilibrium and >1.0 if recombination is absent or infrequent. It is most useful as a measure of departure from linkage equilibrium because its expected value is then known, and an expression for its error variance has recently been published (Haubold *et al.* 1998). It is less useful for detecting departure from clonality because its expected value in clonal populations is unknown unless S_e is very large relative to the number of polymorphic sites. A second difficulty with I_A is that its expected value for a clonal population increases with the number of loci analyzed. Burt *et al.* (1999) suggest a related statistic, which increases monotonically with I_A, but whose expectation is independent of the number of loci analyzed.

The coefficient of linkage disequilibrium: Lewontin (1964) suggested the coefficient $D = (P_{AB} \cdot P_{ab} - P_{Ab} \cdot P_{aB})/(P_{AB} \cdot P_{ab} + P_{Ab} \cdot P_{aB})$, where P_{AB} is the frequency of *AB* haplotypes, and similarly for *Ab*, *aB*, and *ab*, as a measure of departure from linkage equilibrium. Because the choice of the symbols *A* and *B* is arbitrary, it is customary to take the absolute value of *D*, a number whose expectation varies from 0 (linkage equilibrium) to 1 (complete association). Conway *et al.* (1999) have recently used *D* to demonstrate recombination from population data in *Plasmodium falciparum*. They show that values of *D* significantly different from zero are frequent for bases <1 kb apart (demonstrating the power of the test to detect disequilibrium) but absent for sites further apart. The method is appropriate provided that none of the frequencies of the four gametic types are too low: with rare alleles or small samples, values of $D = 1$ will occur by chance, even with frequent recombination. Conway *et al.* analyzed samples varying from 66 to 124 isolates from single geographic regions and included only loci at which the frequency of the common allele did not exceed 0.9. Provided that data of this type are available, the method is an effective one, but it lacks sensitivity applied to more restricted data sets. For the samples of 20 or 30 individuals from the simulated populations analyzed below, it failed to distinguish between linkage equilibrium and clonality (data not shown).

TESTING FOR RECOMBINATION IN SIMULATED POPULATIONS

Values of the three statistics, *H*, IR, and Sawyer's ratio, were calculated for samples drawn from simulated populations. Simulations were carried out as follows:

1. Each population was haploid, of *N* individuals (varying from 50 to 1000), each with 100 sites equally likely to mutate between two alleles, 0 and 1 (thus the simulations are of single-nucleotide polymorphisms for which only two nucleotides are usually found at a site). Each new generation was formed by sampling with replacement from the previous one.
2. In each generation, *m* mutations occurred, each at a random site in a random individual.
3. In each generation, *r* recombinations occurred. A random donor and recipient were chosen, and all sites beyond a random crossover point were exchanged.
4. For each set of parameter values, starting from a population with only 0 alleles, a foundation population was formed by iterating 3*N* generations. Starting from this foundation population, five simulations were made, each of 3*N* further generations.
5. From each final population, two random samples (usually of 20 or 30 individuals) were drawn, and statistics calculated.

Different statistics were calculated on different simulated populations. This was not necessary but arose because the investigation of *H* was completed before the investigation of IR started.

In Figure 1, the statistics are plotted against R/M, where *R* is the probability that a particular site in a gene is altered by recombination and *M* the probability that the site is altered by mutation. The use of this measure of recombination is discussed further below.

Figure 1A shows Sawyer's ratio, a measure based on Sawyer (1989). This test depends on the sum of squares of the lengths of runs in the data; Figure 1A shows the ratio of this sum calculated for the real data and for a randomized matrix with the same allele frequencies at each site. A value of 1.0 indicates that there is no tendency for differences to occur in runs. As expected for frequent recombination, a test based on the occurrence of runs is unable to distinguish between clonality and complete linkage equilibrium, although for low values of R/M the ratio is usually >1.0.

Figure 1B plots the homoplasy ratio, *H*. The value rises continuously with R/M. Although the range of values for a given *r* and *m* is rather large, no overlap occurred between values for $R/M = 0$, 20, and 80. At least one can use *H* to distinguish between no recombination, some recombination, and linkage equilibrium. $H = 0.5$, a value not atypical for bacteria, implies that $R/M \sim 20$. The value is very approximate, but it does

confirm the conclusion of Guttman and Dykhuizen (1994) that, in *E. coli*, recombination is more important than mutation in generating genetic change in bacteria in the short term.

Finally, Figure 1C plots IR. Like *H*, this statistic rises continuously with *R*/*M*, but, as expected, its value in clonal populations is not zero. Table 1 shows how, in a clonal population, the value varies with different levels of genetic variability. Without an estimate of its expected value in clonal populations, the test cannot be used to distinguish between clonality and low levels of repeated recombination unless an "infinite sites" assumption is justified, in which case any incompatible pairs are evidence for recombination.

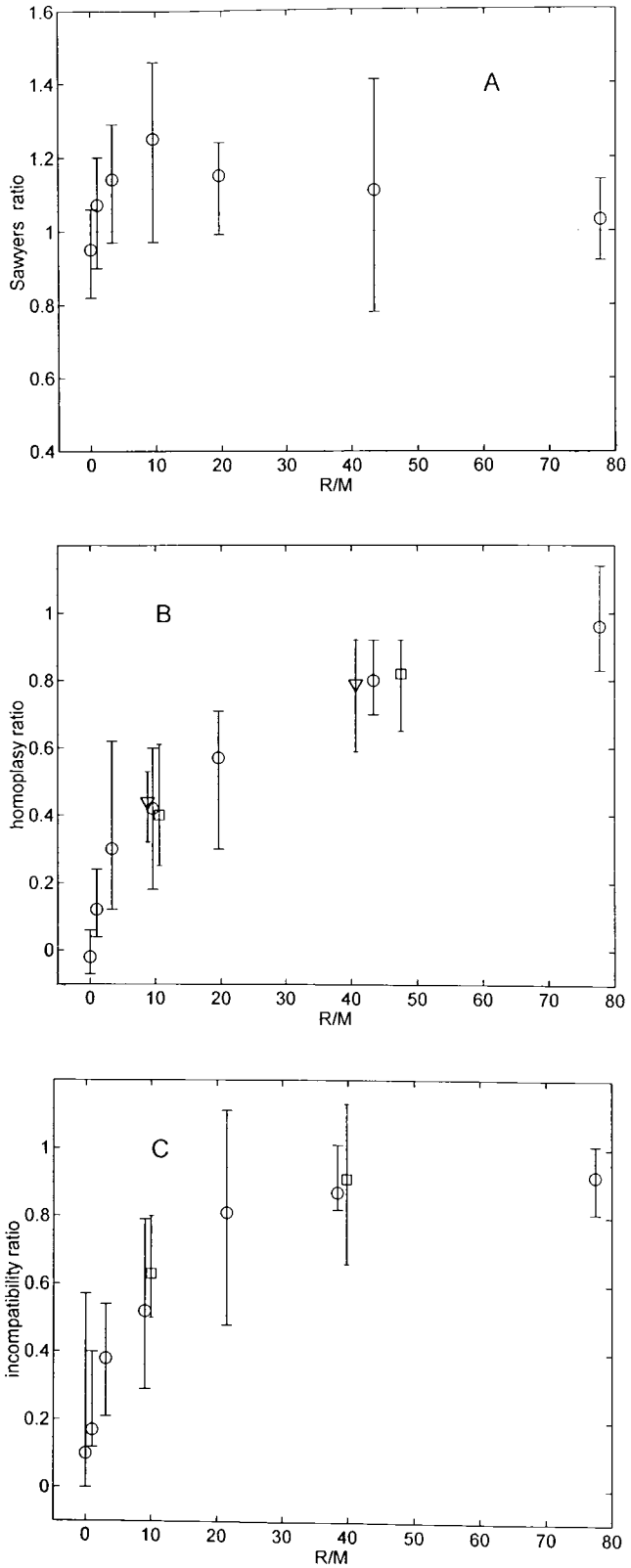

TABLE 1

Effect of mutation rate on the incompatibility ratio in clonal populations

Mutation rate (*m*)	Polymorphic sites (*v*)	Incompatibility ratio (IR)	
		Mean	Range
2	17.4	0.127	0–0.353
4	31.0	0.128	0–0.422
8	47.3	0.158	0.065–0.270
16	68.1	0.357	0.197–0.469

Population size is 100; number of sites is 100; sample size is 30.

Note that recombination was reciprocal as for chromosomal genes of eukaryotes. Recombination in prokaryotes differs in two respects: it is nonreciprocal and involves the insertion of relatively short pieces of DNA. It was not practicable to simulate nonreciprocal recombination because it would have a large effect in reducing genetic variability, given the small population sizes: this effect would be negligible in the large populations characteristic of bacteria. However, the results for *H* and IR should hold for bacterial populations. Populations were simulated in which short regions of 50 sites were reciprocally exchanged. *H* and IR were then plotted against *R*/*M*, with results (not shown) very similar to those in Figure 1. There is, however, one context in which the prokaryotic type of recombination is likely to have results different from the eukaryotic type. This is for the rate of decline in linkage disequilibrium with distance, which will depend on the size of the pieces transferred: this problem is worth further investigation.

The effect of population size on the variance of these estimates is of some interest. Table 2 shows the effect on IR of varying *N* in clonal populations and in populations

Figure 1.—Statistics calculated for simulated populations. Each point is the mean value of 10 samples, 2 from each of five populations; vertical lines indicate the range of values. Population size, *N*: □, 50; ○, 100; ▽, 200. Each population was run for 6*N* generations before sampling. Sites per gene, 100; sample size, 20; *R*, number of changes in a gene caused by recombination; *M*, number of changes caused by mutation. *R* and *M* are summed over all individuals and generations. (A) Sawyer's ratio; (B) homoplasy ratio; (C) incompatibility ratio.

TABLE 2

Effect of population size on estimates of the incompatibility ratio

Population size (N)	Recombination rate (r)	Polymorphic sites (v)	Incompatibility ratio (IR)			
			Mean	Range	Variance × 100	
					V_{within} [a]	$V_{between}$ [b]
50	0	22.2	0.135	0–0.426	0.534	5.39
150	0	27.4	0.108	0.031–0.264	0.316	0.89
500	0	26.7	0.158	0.088–0.286	0.313	0.63
1000	0	28.6	0.074	0.013–0.296	0.502	1.04
50	5	26.2	0.527	0.405–0.667	0.234	1.94
150	5	27.7	0.438	0.200–0.624	0.850	2.08
500	5	28.0	0.441	0.298–0.553	0.531	0.77
1000	5	26.7	0.575	0.433–0.757	0.172	1.79

Sample size is 30; number of sites is 100; mutations per generation is 4.
[a] The differences between two samples from the same population.
[b] The differences between five simulated populations.

with some recombination; similar results (not shown) were obtained for the effect on H of varying N. The variance between populations does not decrease with N. For clonal populations, this is not unexpected. Statistics such as H and IR depend on the form (topology and branch lengths) of the phylogenetic tree. This does not become uniform as the population gets larger. If one imagines looking at the (true) phylogenetic tree of a large population in successive generations, its form would not remain constant. Consider, for example, the coalescence time of all members of a clonal population. This will usually increase by one in each generation but occasionally decrease by some large number. That is, the tree changes discontinuously. It is therefore not surprising that statistics that depend on the tree (as do both H and IR) vary between large populations with the same parameters. It was, at least to me, more surprising that the variance of IR does not decline with N in populations with recombination.

What do the statistics *H* and *IR* measure? In Figure 1, H and IR are plotted against R/M, where R is the probability, per generation, that a nucleotide will change because of recombination and M is the probability that it will change by mutation. M is simply the per base mutation rate, but R is more complicated because it depends not only on the recombination rate but also on the genetic distance between recombinants. In a bacterial context, with one-way insertion of DNA fragments, it is easy to interpret R. Imagine an insertion of length d into a region of length s. If $d < s/2$, the new gene will, at least approximately, occupy the same position in a phylogenetic tree as it did before insertion, and the d inserted nucleotides may contribute homoplasies, depending on the number of changed sites introduced: that is, the number of homoplasies depends on Σdh, where h is the heterozygosity. Note that if $d > s/2$, the new gene is more similar to the gene in the donor, and the number of homoplasies depends on $\Sigma(s - d)h$. In the limit, when a whole gene is transferred, $d = s$, and no homoplasies are caused.

A similar approach can be adopted for reciprocal recombination with a single crossover point. The recombination event separated the gene into a shorter piece, length d, and a longer piece, length $s - d$. Again, H will depend on Σdh, but, because the event is reciprocal, each crossover contributes $2dh$ to R. In plotting Figure 1, R was taken as the number of site changes caused by recombination in the shorter gene region, summed over all chromosomes, and divided by Ns, where N is the population size and s the number of sites per gene.

The expected value of R/M can be calculated as follows. Let the per generation mutation rate $= u$; the probability that a site in two random individuals is occupied by a different allele $= h$; and the per gene recombination rate $= sc$. Then $M = u$ and $R = P$ (a gene undergoes a recombination event) $\times P$ (a particular site is included in the inserted region) $\times h = sch/4$. For a haploid population with only two alternative alleles per site, $h = 2Nu/(1 + 4Nu)$, and the expected value of R/M is $Nsc/2(1 + 4Nu)$.

Table 3 shows expected and observed values of H and R/M for simulated populations with varying rates of mutation and reciprocal recombination. With one minor exception, H increases monotonically with R/M. Simulations (not shown) of populations with reciprocal exchange of short regions confirmed that H is a function of R/M. Thus the answer to the question at the head of this section is that H is a measure of R/M.

CONCLUSIONS

Several methods (Stephens 1985; Sawyer 1989) exist for detecting unique or rare recombinations between genetically different populations. They depend on rec-

TABLE 3

Homoplasy ratio in simulated populations with varying values of *r* and *m*

			R/*M*		
r	*m*	*r*/*m*	Observed	Expected	*H*
1	5	0.2	0.89	0.82	0.071
2	10	0.2	1.21	1.43	0.123
2	2	1	1.66	1.85	0.104
5	25	0.2	2.49	2.50	0.332
5	5	1	4.24	4.17	0.370
10	10	1	6.97	7.14	0.395
10	2	5	8.0	9.26	0.460
25	5	5	19.6	20.8	0.654
50	10	5	33.3	35.7	0.774

Population size is 100; number of sites is 100; sample size is 20.

ognizing the presence of runs of linked nucleotides with distinct ancestries. Incompatibility and partition matrices can be used to give a visual impression that such runs are present before statistical testing (Jakobsen *et al.* 1997). However, it is not clear that such matrices offer any advantage over the more obvious procedure of inspecting a printout of all informative sites in a set of sequences. The statistical significance of any runs sugested by such an inspection can be tested by the maximum Chi-square method (Maynard Smith 1992) or in other ways.

Such procedures are already familiar. The main point of this article is to point out that they are ineffective in detecting repeated recombination between the members of a population because repeated recombination breaks up the runs of linked nucleotides on which they depend. In the limit, in a population in linkage equilibrium, there is no association between neighboring nucleotides and hence no runs.

Several methods of detecting repeated recombination are described here, and their effectiveness is compared on simulated populations. The homoplasy test (Maynard Smith and Smith 1998) compares the observed number of homoplasies in a maximum-parsimony tree of the sequences with the number expected in the absence of recombination. It has the advantage that the number of homoplasies expected in an equally variable clonal population can be estimated so that the evidence for recombination can be tested. It has the drawback, however, that it depends on finding a maximum-parsimony tree for the data, which is time consuming and inaccurate for large data sets. This difficulty can be met by analyzing a subset of, say, 30 sequences.

An alternative is the incompatibility ratio, which compares the number of pairs of sites that are phylogenetically incompatible with the number expected in a panmictic population. The statistic is easy to compute but has the drawback that the expected value in a clonal population is unknown unless an infinite sites model is appropriate for the data being analyzed.

The homoplasy ratio, *H*, is a number whose expectation varies from 0 (clonality) to 1 (complete linkage equilibrium). It can therefore be used as a measure of the rate of recombination. But just what rate is being measured? In simulated populations with a range of values of recombination and mutation rates, *H* is a function of *R*/*M*, where *M* is the probability that, in a short time interval, a nucleotide will alter as a result of mutation, and *R* is the probability that it will be altered by a recombination event. In a bacterial population, *R* is easy to interpret because recombination events usually consist of the insertion of a short region of DNA. The value of *R* will depend on the frequency of such events, the length of the inserted regions, and the genetic distance between donor and recipient. In eukaryotes, interpretation is less obvious. Recombination gives rise to two new sequences, each consisting of a region from each parent. Which parent, then, is to be regarded as the "donor" of novel nucleotides? The answer is that the shorter of the two regions is to be treated as "donated" DNA. Although an intuitive justification for this procedure can be given, the real justification is that, if *R* is calculated in this way, *H* proves to be a monotonic function of *R*/*M*.

LITERATURE CITED

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? Genetics **96:** 15–26.

Burt, A., V. Koufopanou and J. W. Taylor, 1999 Population genetics of human-pathogenic fungi, in *The Molecular Epidemiology of Infectious Diseases*, edited by R. C. A. Thompson. Chapman & Hall (in press).

Conway, D. J., C. Roper, A. M. J.Oduola, D. E. Arnot, P. G. Kremsner *et al.*, 1999 High recombination rate in *Plasmodium falciparum.* Proc. Natl. Acad. Sci. USA (in press).

Dowson, C. G., A. E. Jephcott, K. R. Gough and B. G. Spratt, 1989 Penicillin-binding protein 2 genes of non-lactamase-producing, penicillin-resistant strains of *Neisseria gonorrhoeae.* Mol. Microbiol. **3:** 35–41.

Eyre-Walker, A., N. H. Smith and J. Maynard Smith, 1999 How clonal are human mitochondria? Proc. R. Soc. Lond. Ser. B **266:** 477–483.

Guttman, D. S., and D. E. Dykhuizen, 1994 Clonal divergence in *Escherichia coli* is a result of recombination, not mutation. Science **266:** 1380–1383.

Hagelberg, E., N. Goldman, P. Lio, S. Whelan, W. Schiefenhovel *et al.*, 1999 Evidence for mitochondrial DNA recombination in a human population of island Melanasia. Proc R. Soc. Lond. Ser. B **266:** 485–492.

Haubold, B., M. Travisano, P. B. Rainey and R. R. Hudson, 1998 Detecting linkage disequilibrium in bacterial populations. Genetics **150:** 1341–1348.

Hein, J., 1990 Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98:** 185–200.

Holmes, E. C., M. Worobey and A. Rambaut, 1999 Phylogenetic evidence for recombination in Dengue virus. Mol. Biol. Evol. **16:** 741–749.

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.

Jakobsen, I. B., and S. Easteal, 1996 A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12:** 291–295.

Jakobsen, I. B., S. R. Wilson and S. Easteal, 1997 The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol. Biol. Evol. **14:** 474–484.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General consideration of heterotic models. Genetics **49:** 49–67.

Maynard Smith, J., 1992 Analyzing the mosaic structure of genes. J. Mol. Evol. **34:** 126–129.

Maynard Smith, J., and N. H. Smith, 1998 Detecting recombination from gene trees. Mol. Biol. Evol. **15:** 590–599.

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.

Sawyer, S. A., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

Sneath, P. H. A., M. J. Sackin and R. P. Ambler, 1975 Detecting evolutionary incompatibilities from protein sequences. Syst. Zool. **24:** 331–332.

Spratt, B. G., L. D. Bowler, Q-Y. Zhang, J. Zhou and J. Maynard Smith, 1992 Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. J. Mol. Evol. **34:** 115–125.

Stephens, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or genic conversion. Mol. Biol. Evol. **2:** 539–556.

Suerbaum, S., J. Maynard Smith, K. Bapumia, G. Morelli, N. H. Smith *et al.*, 1998 Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA **95:** 12619–12624.

Communicating editor: P. L. Foster