Consequences of Recombination on Traditional Phylogenetic Analysis

Mikkel H. Schierup and Jotun Hein

Department of Ecology and Genetics, University of Aarhus, DK-8000 Aarhus C., Denmark

Manuscript received February 25, 2000
Accepted for publication June 8, 2000

ABSTRACT

We investigate the shape of a phylogenetic tree reconstructed from sequences evolving under the coalescent with recombination. The motivation is that evolutionary inferences are often made from phylogenetic trees reconstructed from population data even though recombination may well occur (mtDNA or viral sequences) or does occur (nuclear sequences). We investigate the size and direction of biases when a single tree is reconstructed ignoring recombination. Standard software (PHYLIP) was used to construct the best phylogenetic tree from sequences simulated under the coalescent with recombination. With recombination present, the length of terminal branches and the total branch length are larger, and the time to the most recent common ancestor smaller, than for a tree reconstructed from sequences evolving with no recombination. The effects are pronounced even for small levels of recombination that may not be immediately detectable in a data set. The phylogenies when recombination is present superficially resemble phylogenies for sequences from an exponentially growing population. However, exponential growth has a different effect on statistics such as Tajima's *D*. Furthermore, ignoring recombination leads to a large overestimation of the substitution rate heterogeneity and the loss of the molecular clock. These results are discussed in relation to viral and mtDNA data sets.

WITH automatic, PCR-based sequencing, the amount of population DNA sequence data is rapidly increasing and a number of microevolutionary hypotheses can be tested. Sequences sampled from populations differ from sequences sampled from different species in that population genetics models can be used to analyze their relationships and they can potentially recombine.

The coalescent (Kingman 1982) describes the genealogy of a sample of nonrecombining sequences in a panmictic population with random mating and no selection. It has been extended to include recombination (Hudson 1983), gene conversion (Wiuf and Hein 2000), population growth (Slatkin and Hudson 1991), selfing (Nordborg and Donnelly 1997), and population subdivision (Notohara 1990). Various parameters can then be estimated from the sampled sequences under the model chosen (Griffiths and Tavare 1994; Kuhner *et al.* 1995, 1998; Griffiths and Marjoram 1996; Beerli and Felsenstein 1999; Griffiths 1999; Stephens and Donnelly 2000).

If intragenic recombination occurs, different parts of the sequence have different phylogenetic histories. This is an advantage because different parts of the sequence represent different, although correlated, realizations of the evolutionary process. Each realization is associated with a large variance. Recombining sequences should therefore provide an estimate of an evolutionary parameter of interest with a smaller variance than an estimate from a set of nonrecombining sequences. However, the occurrence of recombination also complicates analysis. First, recombination implies that the sequences under study are not related by a single phylogenetic tree, but rather by a set of correlated trees over the sequence (Hudson 1983). This can be viewed as unfortunate because a phylogenetic tree is a visually appealing way of representing the data. Second, the power to detect recombination in a data set is limited (Hudson and Kaplan 1985; Wiuf and Hein 1999), estimated recombination rates have very large variances (Hey and Wakeley 1997; Wall 1999), and many deviations from an infinite-sites model of mutation can mimic the effect of recombination by causing more instances of parallel evolution (Eyre-Walker *et al.* 1999). Thus, even fairly high rates of recombination cannot be detected statistically. Third, the phylogeny contains information that is not captured by simpler methods independent of recombination, such as the pairwise distances between sequences (Griffiths and Tavare 1994). Because of these complications in dealing with recombination, phylogenetic trees are often reported even when sequences can potentially recombine. This is particularly true for viral species and bacterial species, where many conclusions are based on phylogenetic patterns, *e.g.*, estimates of the scaled mutation rate, $\theta = 4Nu$ (Kelsey *et al.* 1999) and dating of lineage splitting and origin of diversity assuming a molecular clock (Zhu *et al.* 1998; Holmes *et al.* 1999a). Furthermore, recent studies report evidence for recombination (or rather gene conversion)

Corresponding author: Mikkel Heide Schierup, Department of Ecology and Genetics, University of Aarhus, Bldg. 540, Ny Munkegade, DK-8000 Aarhus C., Denmark. E-mail: mikkel.schierup@biology.au.dk

in mtDNA of humans (Awadalla *et al.* 1999; Eyre-Walker *et al.* 1999). If these reports prove correct, then the nonpseudoautosomal part of the heterogametic sex chromosomes appears to be the only case of nonrecombining DNA. Accordingly, methods based on phylogeny that ignore recombination may have a very limited use in molecular population genetics unless we can show that the amount of recombination typically estimated in population data sets has only a negligible effect on evolutionary inferences.

Our goal was therefore to quantify how ignoring recombination affects inferences based on phylogenetic trees. We do this by simulating sequences under the coalescent with recombination and subsequently by reconstructing a single phylogenetic tree from these sequences. We study quantitatively how recombination affects various parameters that can be estimated from the inferred phylogenetic trees. We then use statistics that summarize the shape of these trees and compare the values to the values expected without recombination, *i.e.*, under the standard coalescent. Furthermore, we investigate how robust our results are to common deviations from the simple Jukes-Cantor substitution model, such as rate heterogeneity and transition-transversion bias.

We were motivated by the fact that many published analyses of data suggest that the sequences do not conform to the neutral coalescent. In a neutral coalescent, most coalescence events in the history of the sample happen very fast, but in many data sets the terminal branches (connecting to the "twigs") appear very long. This empirical pattern is often interpreted as evidence for population expansion (Slatkin and Hudson 1991; Holmes *et al.* 1999a) and has been investigated mainly through the distribution of pairwise differences, also termed the mismatch distribution (Slatkin and Hudson 1991). However, recombination has a similar effect because shuffling parts of the sequences should make distances between the sequences more similar to each other, thus causing the inferred tree to approach a star phylogeny. We show here that this expected pattern is evident in simulated data sets in which trees from data sets with recombination have long terminal branches and are less clock-like than the trees from sequences without recombination (see also Schierup and Hein 2000). The main question is how much recombination is needed before these effects are of a detectable magnitude. The apparent similarity between the effect of recombination and exponential growth led us to investigate the effect of exponential growth on the shape of the phylogenetic tree and search for statistics that can distinguish the effects of recombination and exponential growth. Finally, we discuss implications of our results for timing of events and estimation of evolutionary parameters and some experimental data sets from mitochondrial DNA and viruses in the light of our findings.

SIMULATION OF DATA SETS

Coalescent simulations: We simulate samples of k sequences under the coalescent with recombination, based on Hudson's (1983) algorithm. The population consists of N diploid individuals. We use the continuous time approximation and scale time in $2N$ generations, and recombination rate as $\rho = 4Nr$. A given sequence thus has a recombination length of $\rho/2$. The coalescent is constructed by waiting for recombination or coalescence until all ancestral material in the k sequences has found a common ancestor. With k sequences, the waiting time for coalescence is exponentially distributed with parameter $k(k-1)/2$. The waiting time until a sequence is created by recombination is exponentially distributed with parameter $\rho/2$ (Hudson 1983). For the k extant sequences, the exponential rate of recombination is thus $R = k\rho/2$. For ancestral sequences, R also includes nonancestral material if it is "trapped" by segments of ancestral material. This is because recombinations in trapped nonancestral material will split two blocks of ancestral material and therefore will have an effect on the coalescence process (see also Wiuf and Hein 1997). Since coalescence and recombination events are independent, the time to one of the events happening is exponentially distributed with parameter $R + k(k-1)/2$.

Simulation of a single outcome of this process is performed by starting with k sequences of recombination length $\rho/2$ and determining by drawing a random number from the exponential distribution when the first event happens. According to what happened, the parameter of the exponential distribution for the next event is then updated. If a coalescence event happened, the number of sequences with ancestral material is reduced by one; if it was a recombination event, it is increased by one. A recombination point is chosen uniformly over the ancestral material and the nonancestral material trapped by two blocks of ancestral material. Coalescences may increase the intensity of recombination if nonancestral material is trapped by two blocks of ancestral material. The process is continued until each point on the extant sequences has found a most recent common ancestor. With recombination, different parts of the sequence are likely to have different coalescent trees (and different times to most recent common ancestor). For each continuous segment of ancestral material, all information about times of coalescence events is kept in memory until mutations are added.

Exponential growth was simulated following Slatkin and Hudson (1991) closely. The population size at generation t in the past is $N(t) = N_0 e^{-bt}$, and the scaled growth rate is defined as $\beta = Nb$. Sequences are assumed nonrecombining. The times between coalescence events can be simulated according to the following recursion: $t_i = \ln(1 - \beta \exp(-\tau_i)(2/i(i-1))\ln(U))$,

where $\tau_i = \Sigma_{k=i+1}^{n} t_k$ and U is a uniformly distributed random variable on the interval (0, 1). In this case, time is measured in units of $1/b$ (Slatkin and Hudson 1991).

Creation of nucleotide data sets: From the simulated genealogy, a data set of nucleotide sequences of length L can easily be generated. The sequences are divided into L equally sized fragments. A substitution process is then performed in the left endpoint (hereafter termed the nucleotide position) of each of these fragments. Each nucleotide position from left to right is considered separately, assuming that nucleotides mutate independently. First, a nucleotide is assigned to the most recent common ancestor (MRCA) with probabilities according to the equilibrium frequencies of nucleotides under the substitution model. The evolution of the nucleotide is then followed down the genealogical tree at this position. For a given branch, the number of mutations is Poisson distributed with a parameter m that may depend on the nucleotide state at the beginning of the branch. Two mutation models were used, Jukes-Cantor and Kimura's two-parameter model (see Li 1997). For the Jukes-Cantor model, the probability that the nucleotide changes along a branch of length t is $P(\text{change}) = \frac{3}{4} - \frac{3}{4}e^{-4/3mt}$. The parameter m is related to the population mutation rate, θ, as $m = 0.5\theta$. The models both assume equal equilibrium nucleotide frequencies, so the nucleotide at the MRCA was assigned randomly.

Mutation rate heterogeneity in different sites was modeled using a gamma distribution with both parameters being equal to α; that is, the mean equals one. The rate of a given site was then determined by multiplying a random number drawn from this distribution with the mean rate m.

In many cases an outgroup was simulated to enable subsequent analysis programs to root the inferred phylogenetic tree. The outgroup sequence was similarly simulated from the nucleotide at the MRCA, with a predetermined branch length and the same substitution model.

ANALYSIS OF SEQUENCE DATA

Construction of genealogy: The simulated data sets were analyzed using published programs for the construction of phylogenetic trees. We used both distance-based methods and maximum-likelihood methods. We constructed the distance matrix among sequences using the DNAdist program of the PHYLIP (Felsenstein 1995) package, assuming in all cases the Jukes-Cantor model. This distance matrix was then used as input to either the Fitch or Kitsch program from the PHYLIP package for construction of a phylogenetic tree. Both programs implement the Fitch-Margoliash least-squares methods, but differ in the assumption of a molecular clock; the clock is assumed in Kitsch but not in Fitch. The simulated outgroup was used to root the tree in Fitch. For a maximum-likelihood estimation of the tree, we used fastDNAml (Olsen *et al.* 1994), also set to assume the Jukes-Cantor model. FastDNAml is a speed up of DNAml of the PHYLIP package (Felsenstein 1981) and works reasonably fast with 20 sequences.

Measures on the tree: From these trees, several statistics were recorded from the branch lengths, which are given in units of m:

D: The time to the most recent common ancestor
P: The average time to the most recent common ancestor of two genes
T: The total length of the genealogy
S: The sum of the length of the terminal branches
B: The average length of basal branches emanating from the root

Under the neutral coalescent without recombination, the expected values of these are

$$E(D) = 2(1 - 1/n)$$
$$E(P) = 1$$
$$E(T) = 2a_n$$
$$E(S) = 2$$
$$E(B) = 2b_n$$

(Kingman 1982), where $a_n = \Sigma_{i=1}^{n-1}(1/i)$ and $b_n = (1/n) + \Sigma_{i=1}^{n-1}(1/i^2)$ depend on the number of sampled sequences, n, only. Uyenoyama (1997) defined the four ratios

$$R_{\text{PT}} = \frac{2Pa_n}{T},\ R_{\text{ST}} = \frac{Sa_n}{T},\ R_{\text{SD}} = \frac{S(1 - 1/n)}{D},\ R_{\text{BD}} = \frac{B(1 - 1/n)}{Db_n}$$

and showed by simulation that they are almost independent of the mutation rate. The scaling assures that if the ratios are viewed as ratios of expectations they each have an expected mean of one under the neutral coalescent. However, the distribution of the ratios over replicates may have a different mean dependent on the joint distribution of the numerator and denominator. The ratios were calculated from the branch lengths of the inferred trees. The outgroup was used to root the tree and the height of the tree, D, was calculated as the average height from the root to the tips.

We also calculated the time between subsequent coalescence events. Under the neutral coalescent, the waiting time F_i while there are i sequences in the sample is exponentially distributed with mean $E(F_i) = 2/(i(i - 1))$, in units of $2N$. These waiting times are independent of each other. Thus, we can define $G_i = F_i i(i - 1)/2$ with an expected $E(G_i) = 1$, for all i. Plotting G_i as a function of i can thus visualize systematic deviations from neutral expectations. Values of F_i were calculated from trees reconstructed using Kitsch instead of Fitch, because they can be unambiguously defined only for a phylogenetic tree with a molecular clock. Tajima's D was calculated from the simulated data sets as $D = (\pi - S_G/a_n)/\sqrt{\text{Var}(\pi - S_G/a_n)}$ (Tajima 1989), where

the average pairwise distances π were estimated using DNAdist and the number of segregating sites, S_G, was counted.

The computer program for simulating sequences and for calculating the various statistics can be accessed through http://www.bioinf.au.dk/~mheide.

RESULTS

All results are based on simulations of 1000-bp sequences. The recombination rate ρ is for the whole sequence and can thus be converted to the per base pair recombination rate by dividing by 1000. Thus, the results can be directly compared to experimental data sets even if their sequences have different length, because it is the number of recombination events over the whole sequence that is important.

Figure 1 shows two typical trees of simulated data sets of 20 sequences for the case of no recombination and for ρ = 8. The mutation rate, $m = 0.05$, and trees were reconstructed by the distance-based method. For 20 sequences, ρ = 8 is equivalent to an expected $8 \sum_{i=1}^{19} (1/i) = 28.8$ recombination events in total in the history of the sequences back to the most recent common ances-

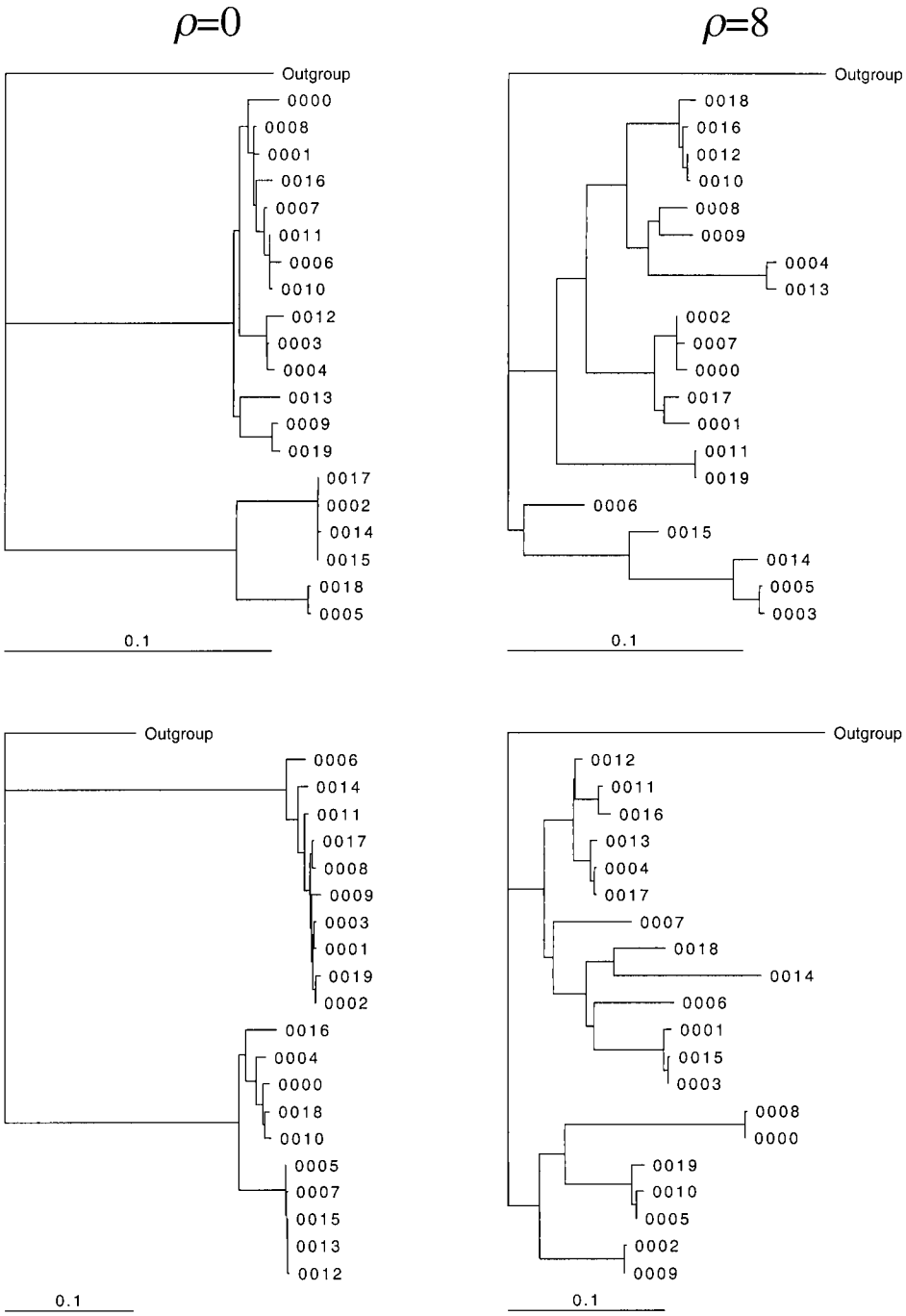


FIGURE 1.—Phylogenetic trees reconstructed for randomly selected simulated data sets. The two trees to the left are reconstructed from sequences simulated with no recombination and the two trees to the right from sequences simulated with ρ = 8. Twenty sequences and one outgroup were simulated with $m = 0.05$. Phylogenetic trees were reconstructed using DNAdist and Fitch of PHYLIP (FELSENSTEIN 1995).

tor. It is evident that recombination affects the inferred trees in at least two ways in that the terminal branches leading to the tips appear longer and the tree appears less clock-like (see also SCHIERUP and HEIN 2000). It is these two basic observations that are the basis of the following quantitative investigations.

We set out by studying the effects of ignoring recombination for sequences simulated under the simplest possible substitution model, the Jukes-Cantor model. We then investigate, one at a time, how common deviations from the Jukes-Cantor model affect these results. The most focus is on the commonly used distance-based methods of inferring genealogies, but we also compare them with maximum-likelihood-based methods because these are expected to be used more in the future.

Measures derived from the genealogy by distance-based methods: Figure 2 shows the values of *D*, *P*, *T*, *S*, and *B* inferred from the reconstructed tree as functions of the recombination rate assumed in the simulations. Samples of 20 sequences were simulated and results (± SD) are shown based on 3000 replicates. An outgroup was simulated with an average distance of $0.5m$ from the root. The Jukes-Cantor model with $m = 0.05$ was used. Trees were reconstructed using the distance-based method. The values of the five statistics for $\rho = 0$ closely match the values expected for sequences evolving under the neutral coalescent with $m = 0.05$. As recombination increases, each of the five quantities gets increasingly biased. The time to the most recent common ancestor and the length of the basal branches decrease, whereas the total length of the tree and the length of the terminal branches increase. The estimated average pairwise distances decrease slightly. Since recombination should not affect this quantity, the decrease is caused by the reconstruction method. The increasing length of the tree is caused by the incompatibilities in the data set caused by recombination (EYRE-WALKER *et al.* 1999). The distance-based method postulates more mutation events in the tree than have actually happened in order to accommodate these incompatibilities. It is a property of the distance-based method that the height of the tree and the pairwise distances are reduced with increasing recombination. This is most likely because the number of extra mutations needed to make the data compatible with a single tree can be limited by reducing the height of the tree, which in turn reduces the pairwise distances.

Figure 3a shows Uyenoyama's four ratios for the same data set (± SD) and Figure 3b shows the ratios for the smaller mutation rate $m = 0.01$. As expected, the ratios are almost independent of the mutation rate. Their values for $\rho = 0$ are close to one. The deviation from one is caused by the fact that, in general, the expectation of a ratio is different from the ratio of the expectation of its components, but in this case, the difference is slight. The mutation rates of Figure 3 were chosen to represent a typical nuclear data set ($m = 0.01$ corresponds to a nucleotide diversity $\pi = 2\%$) and $m = 0.05$ is more likely for viral data sets. Figure 3 also shows that the standard deviation of the ratios is very large. However, noting that the variances in Figure 3, a and b, are very similar, it can be concluded that this variation is caused mainly by the large variance in the genealogical process and not in the mutation process. Thus, in a data set of an average pairwise difference of 20 sites (corresponding to $m = 0.01$), the ratios are expected

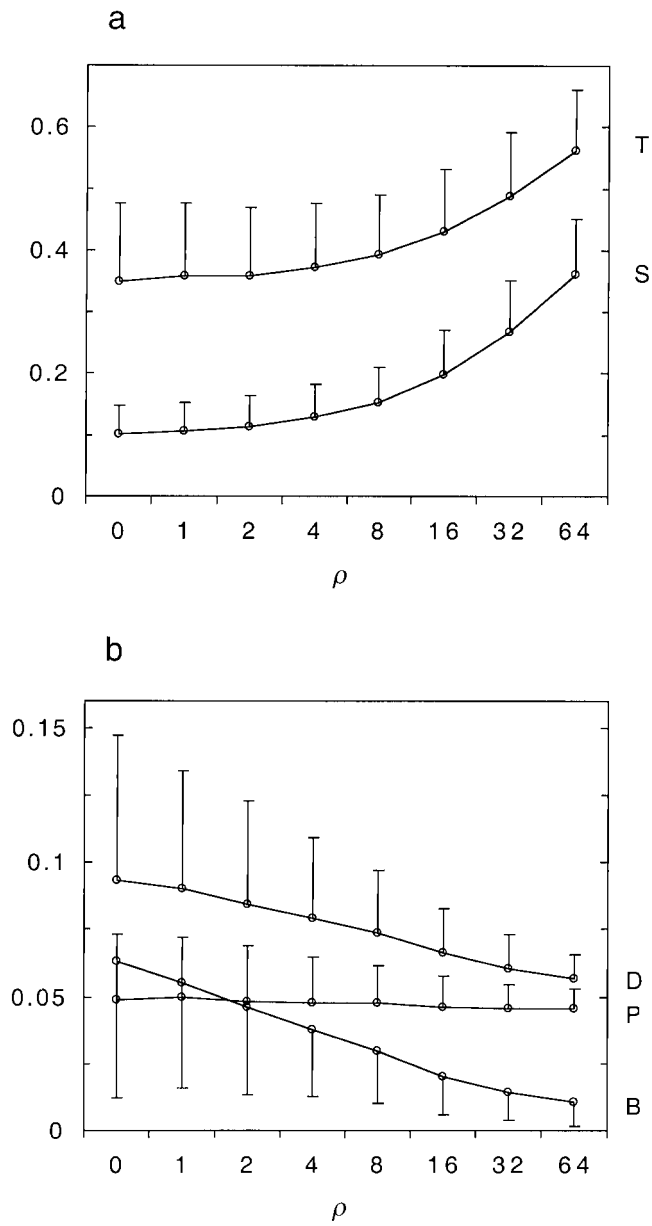
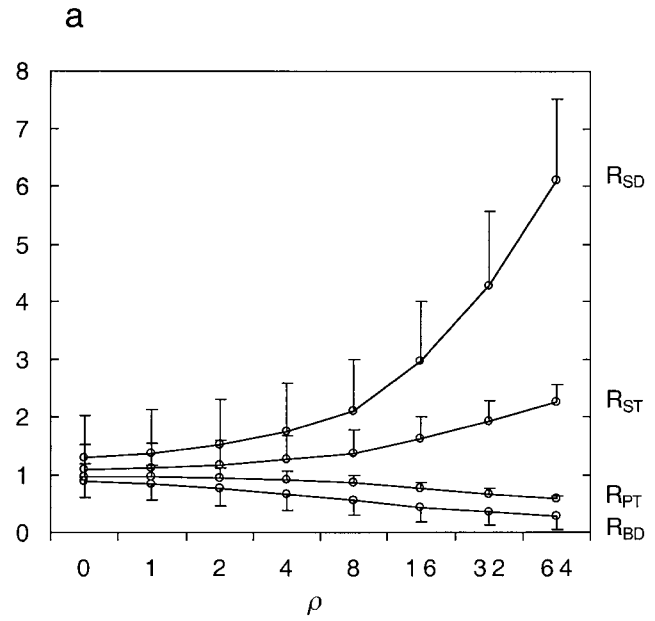


FIGURE 2.—The value of tree statistics as a function of recombination rate. Twenty sequences of 1000 bp and outgroup were simulated under the JC model ($m = 0.05$) and trees were reconstructed using DNAdist and Fitch of PHYLIP (FELSENSTEIN 1995). (a) The total length of the terminal branches (*S*) and the total branch length (*T*). (b) The average length of the basal branches (*B*), the average pairwise distance (*P*), and the average height of the tree (*D*). Standard deviations are shown to one side only to avoid overlap. Means are based on 3000 replicates.



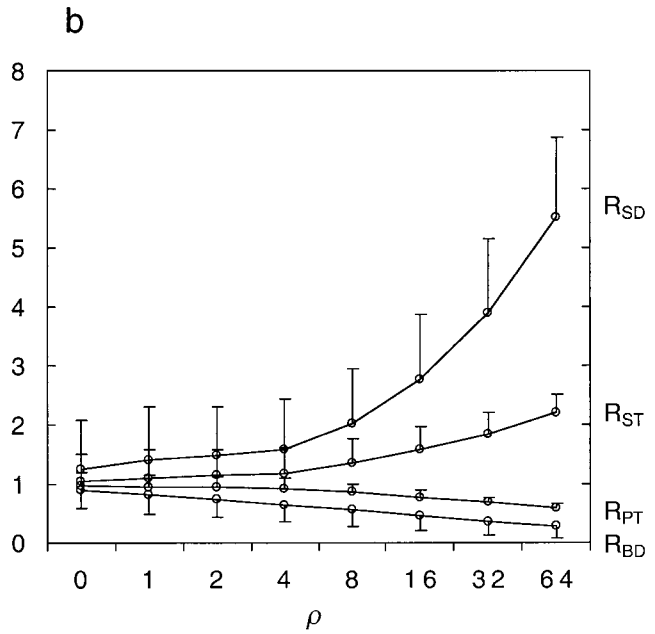


FIGURE 3.—The value of tree-based ratios as a function of recombination rate (with one-sided standard deviations). (a) Simulation parameters as in Figure 2. (b) As in a, except sequences were simulated with a reduced mutation rate of $m = 0.01$.

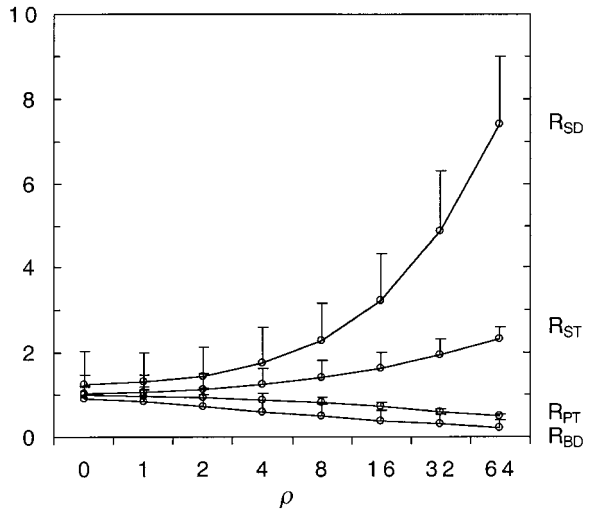


FIGURE 4.—Effect of the number of sequences sampled. The value of tree-based ratios as a function of recombination rate. The simulation parameters used were as in Figure 2 except that 30 sequences were simulated. Results are based on 5000 replicates.

to have a variance in the range shown in Figure 3. The patterns in the ratios are as expected from Figure 2. When recombination exceeds $\rho = 8$, the effect on the ratios is noticeable and can be expected to be significant for many data sets. For example, in a typical *Drosophila melanogaster* nuclear gene, $\rho = 8$ corresponds to just 100 bp (BEGUN and AQUADRO 1995). Overall, R_{SD} appears to be most affected by recombination, agreeing well with the fact that among the four ratios R_{SD} has been found to deviate more significantly in the few studies where it has been used (UYENOYAMA 1997; SCHIERUP *et al.* 1998; MAY *et al.* 1999).

Figure 4 shows results for 30 sequences sampled. The ratios are slightly more affected by a given recombination rate because more recombination events are expected in the history of 30 sequences than with 20 sequences. Likewise, the standard deviations of the ratios are slightly smaller for 30 sequences. However, we conclude that the main source of variance is intrinsic to the coalescent process and that the power of the four ratios is relatively insensitive to the number of sequences and the mutation rate within the range typically observed in experimental data sets.

Comparison with maximum-likelihood methods: Trees were also reconstructed using a maximum-likelihood method as implemented in FastDNAml. Results are shown in Figure 5. For zero recombination, results cannot be distinguished from those of distance-based tree reconstruction (Figure 3), but with increasing recombination there are differences. The deduced time to the most recent common ancestor D is not reduced with this method, in contrast to the distance-based method (Figure 2). Consequently, the total length of the genealogy is relatively larger than under the distance-based methods, and the pairwise differences P are increasing with recombination in this case. Thus, if there is recombination in a data set, the bias by ignoring the recombination is dependent on the method used for tree reconstruction. However, the four ratios have the same qualitative pattern for distance-based and maximum-likelihood methods, with the bias largest for distance-based methods (Figure 5c).

The effect of substitution model and rate variation:

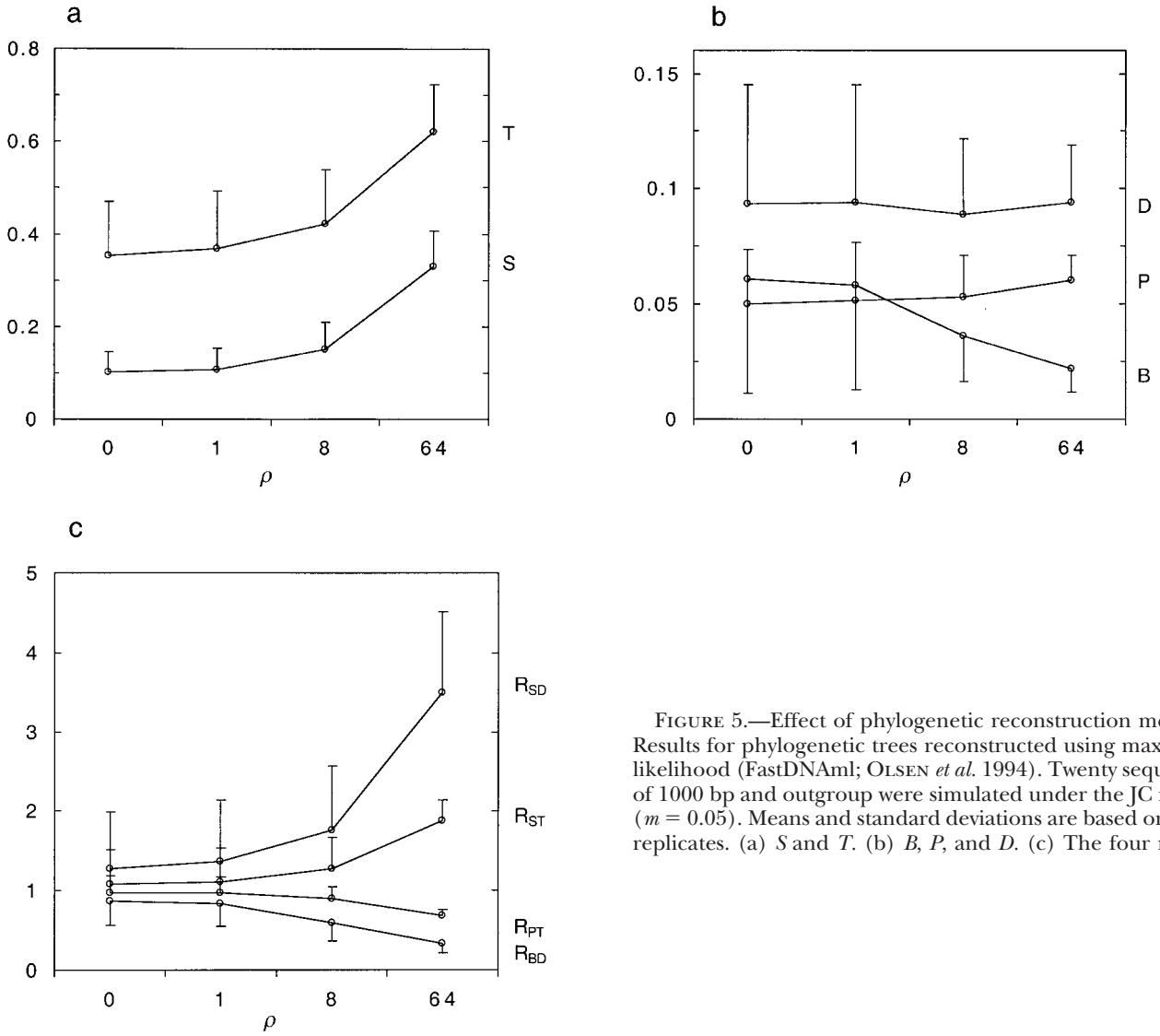


FIGURE 5.—Effect of phylogenetic reconstruction method. Results for phylogenetic trees reconstructed using maximum likelihood (FastDNAml; OLSEN *et al.* 1994). Twenty sequences of 1000 bp and outgroup were simulated under the JC model ($m = 0.05$). Means and standard deviations are based on 1000 replicates. (a) *S* and *T*. (b) *B*, *P*, and *D*. (c) The four ratios.

The Jukes-Cantor model of substitution is extremely simple but inaccurate for most, if not all, real data sets. It is therefore of interest to know how deviations from the Jukes-Cantor model affect the distributions of the ratios, in particular if the deviations are not taken into account during analysis. We focus on two of the most common deviations, namely transition/transversion bias (Kimura's two-parameter model) and heterogeneity in the rate of sequence evolution between sites. We simulated data sets with these deviations, but analyzed the data sets assuming the simple Jukes-Cantor model. We note that this approach maximizes the likelihood that the deviations can mimic the effect of recombination. Rate heterogeneity can be expected to have a similar effect as recombination, because it creates parallel evolution at the fastest evolving sites and this leads to incompatibilities in the data set. Figure 6 shows the value of the four ratios (for $\rho = 0$) as functions of the rate shape parameter α. Even an extremely high rate heterogeneity of $\alpha = 0.125$ has a minor effect on the mean of the ratios compared to the effect caused by recombination. Transition/transversion bias had an even smaller effect; for an extreme bias of 20, the ratios deviated by <2% from their values without transition/transversion bias (results not shown). We conclude that differences in substitution models have small effects compared to the effect of recombination when $\rho > 8.0$.

Comparison with exponential growth: Ignoring recombination leads to long terminal branches and a more star-shaped genealogy and thus superficially resembles the effect of exponential growth. To investigate this quantitatively, we simulated data sets under exponential growth with growth parameter β (see SLATKIN and HUDSON 1991). Figure 7 shows the four ratios as a function of growth. The ratios are affected by exponential growth in much the same way as by recombination.

To attempt to distinguish exponential growth from recombination, we employed several more detailed analyses. The first was to look at the scaled internode distances G_i as a function of the number of ancestral

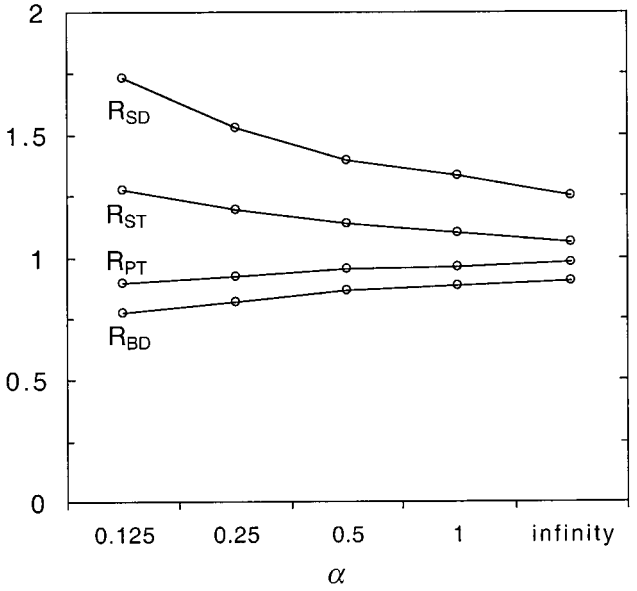


FIGURE 6.—The effect of rate heterogeneity on the four ratios. Twenty sequences of 1000 bp and outgroup were simulated under the JC model ($m = 0.05$) and trees were reconstructed using DNAdist and Fitch of PHYLIP (FELSENSTEIN 1995). $\alpha = \infty$ corresponds to no heterogeneity, and decreasing α implies increasing heterogeneity ($\alpha = 1$ is the exponential distribution). Means are based on 5000 replicates.

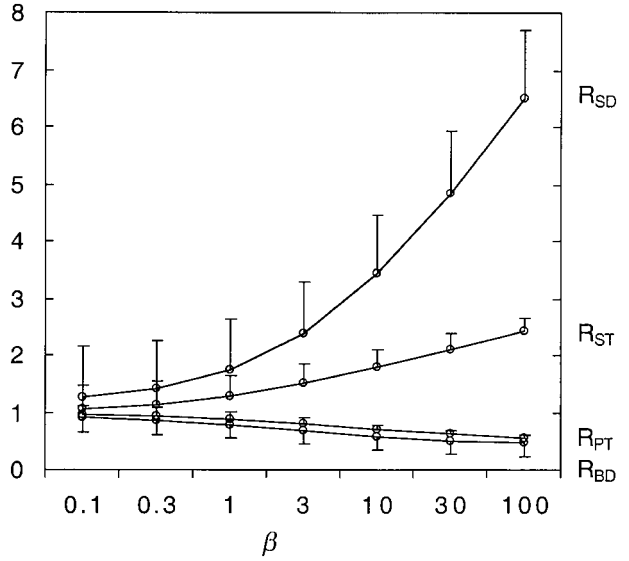


FIGURE 7.—The effect of exponential growth on the four ratios. Twenty sequences, 1000 bp, with an outgroup were simulated with no recombination but different rates of exponential growth, β. Time is rescaled in units $1/b$ and the mutation rate m was chosen so that the average pairwise divergence was 0.1. Means with standard deviations are based on 5000 replicates.

sequences *i*. Figure 8b shows results for various recombination rates (with constant population size) and Figure 8a shows results for several different rates of exponential growth (with no recombination). Figure 8b shows that for $\rho = 0$, the line is close to horizontal at $y = 2m = \theta$, as expected under the neutral coalescent. With increasing recombination, coalescences closest to the root of the tree are much smaller than expected, whereas the most recent coalescence times are much larger than expected. Exponential growth has much the same effects (Figure 8a) except that recent coalescence times are expected to be larger than with recombination. It does not appear likely that this difference is large enough to distinguish the two alternatives, but it does show that the likelihood of two identical sequences in a sample is much higher for the case of recombination. This is also reflected in mismatch distributions constructed under the two hypotheses (results not shown). With exponential growth, the mismatch distribution is closer to a Poisson distribution than with recombination, but again, the difference is very small and would not be statistically detectable for most data sets (results not shown).

As another test, we estimated Tajima's *D* for sequences simulated under exponential growth or recombination. The mean of Tajima's *D* is expected to be independent of recombination, whereas exponential growth should lead to a negative Tajima's *D* because of an excess of singletons. Figure 9 shows Tajima's *D* as a function of recombination (Figure 9a) or of exponential growth rate (Figure 9b). The pattern of the means is as predicted. Noticeable, however, is the decrease of the standard deviation as either growth rate or recombination rate increases. This may increase the power of distinguishing between the two hypotheses. As an example, assume that a data set of 20 sequences shows an R_{SD} value of 4. This value would correspond to either $\rho = 32$ or $\beta = 10$ (compare Figures 3a and 7). For $\rho = 32$, $D \in [-0.5, 0.5]$, and for $\beta = 10$, $D \in [-1.6, -0.9]$ (see Figure 9), so in this case the value of Tajima's *D* may give a good indication of what is causing the deviation from a coalescent tree.

Recombination and rate heterogeneity: With recombination, different segments of the sequence have different phylogenetic trees with different times to the most recent common ancestor and consequently different amounts of sequence variation. Furthermore, when recombination is ignored, parallel mutations need to be postulated to fit the data to a single tree. These two effects are likely to cause apparent mutation rate heterogeneity over the sequence. This was investigated quantitatively by simulating sequences under the Jukes-Cantor model without rate heterogeneity (equivalent to $\alpha = \infty$), but with varying amounts of recombination, and subsequently estimating rate heterogeneity while ignoring recombination. DNAdist and Fitch were used to infer a topology of the phylogenetic tree of sequences and this topology together with the sequences was piped into the program Baseml of PAML (YANG 1999). Baseml was set to find the maximum-likelihood estimate of the shape parameter α of the gamma distribution using a discrete approximation with eight classes. Results are

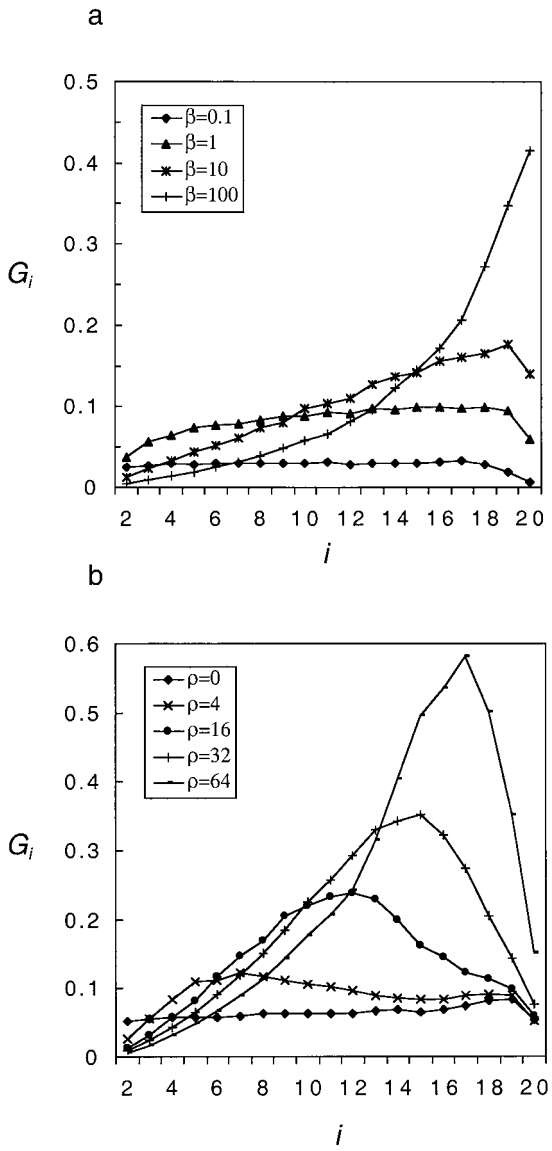


FIGURE 8.—Internode distances with recombination and exponential growth. Shown are the standardized time intervals G_i between coalescent events in the inferred trees, with i referring to the coalescence event when there are i lineages left. Twenty sequences and one outgroup were simulated with $m = 0.05$. Phylogenetic trees were reconstructed using DNAdist and Kitsch (assuming a molecular clock) of PHYLIP (FELSENSTEIN 1995). Means are based on 2000 replicates. (a) Internode distances for sequences simulated under four different exponential growth rates. (b) Internode distances for sequences simulated with five different recombination rates.

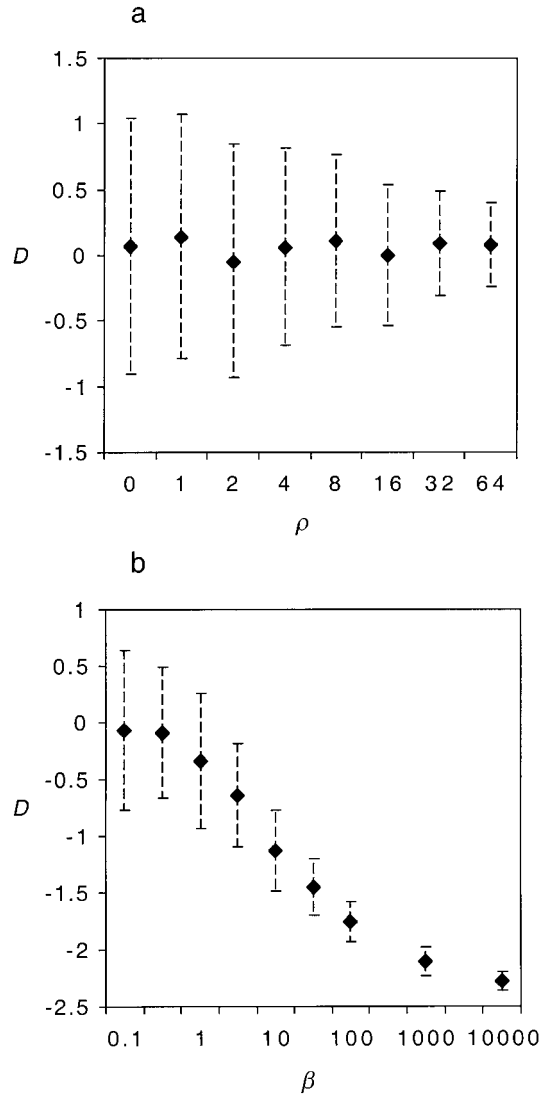


FIGURE 9.—Tajima's D as a function of recombination rate (a) and exponential growth rate (b). Sequences with 10,000 bp were simulated ($m = 0.005$). Means and standard deviations are based on 500 replicates.

shown in Table 1. Because values of infinity are sometimes returned, Table 1 shows the median and the 95% confidence interval for α. When $\rho = 0$, estimates of α are very large, as expected. However, with $\rho > 0$, Baseml infers significant rate heterogeneity. In particular, when $\rho > 16$, $\alpha < 0.5$, which is a very large rate heterogeneity considering that most analyses of interspecific phylogenies, where recombination does not occur, have $\alpha > 0.5$.

Analysis of experimental data sets: To investigate the practical implications of our analysis in more detail, we chose to analyze four data sets that have been used to reconstruct phylogenetic trees and where it is unclear whether recombination plays an important role or occurs at all. Results of the various analyses are summarized in Table 2. Included are two data sets from viruses [human immunodeficiency virus (HIV) from North America, 1986–1990 (KORBER *et al.* 1998) and foot and mouth disease from Southern Africa (A. D. S. BASTOS, personal communication)] and two mitochondrial data sets [African humans (VIGILANT *et al.* 1991) and Grant's gazelle from a single population (ARCTANDER *et al.* 1996)]. For comparison, we also analyzed a data set from *D. melanogaster*, the nuclear gene vermillion, which is located in a region of high recombination in the Drosophila genome and shows no evidence of selection

TABLE 1

Recombination and rate heterogeneity

	$\rho = 0$	$\rho = 1$	$\rho = 4$	$\rho = 16$	$\rho = 64$
Median	∞	5.5	1.4	0.47	0.28
95% confidence	[3.2, ∞]	[0.98, ∞]	[0.62, ∞]	[0.24, 0.86]	[0.21, 0.35]

The rate heterogeneity parameter α was estimated using PAML 2.0g (Yang 1999) for sets of 20 sequences simulated under the neutral coalescent with recombination. The Jukes-Cantor model with $\alpha = \infty$ was used for simulation. A total of 100 replicates were run.

(Begun and Aquadro 1995). Distance-based methods were used for reconstruction of the phylogenetic tree (*i.e.*, DNAdist and Fitch, using an outgroup), assuming the HKY85 substitution model. The recombination rate was also estimated according to Hey and Wakeley's (1997) method, calculated using SITES (Hey and Wakeley 1997). This estimator was shown to perform relatively well on simulated data sets (Wall 1999). However, the estimates of ρ should be interpreted cautiously because they may be inflated by multiple substitutions.

The results for vermillion show, as expected, high values of R_{SD} and R_{ST}, compatible with the estimated high value of $\rho = 259$. Tajima's D suggests no exponential growth. Results for the other data sets are remarkably similar to vermillion. The four ratios all show large deviations from the neutral coalescent, and the estimated values of ρ are very high. For human mtDNA, Tajima's D suggests some evidence for population growth. However, $D = -1.27$ is most compatible with a growth rate of only $\beta = 10$ (Figure 9b), but R_{SD} is then expected to be on the order of 4 (Figure 9a), whereas the observed value is 10.8. For Grant's gazelle, Tajima's D does not depart from zero, but the ratios are still different from expectations under the neutral coalescent. Thus, there appears to be a deviation from expectation in both mitochondrial data sets, which cannot be explained by exponential growth alone, but is compatible with recombination. However, the deviation may also be compatible with some substitution process or demographic process not considered here. If we assume that recombination plays a large role in the human mtDNA data set and accept the ρ value estimated (Table 2), then this will have consequences for previous estimates of the age of diversity. Judging from Figure 2, the time to the MRCA estimated when $\rho = 50$ is $\sim$40% lower than the real mean coalescence time over the sequences. If recombination occurs in human mtDNA (Awadalla *et al.* 1999), then the "mitochondrial Eves" must be older than previously estimated.

The results for the two viral data sets also show strong evidence for recombination with very large estimated ρ values and strongly biased ratios. We assume here that selection affects only a small proportion of the segregating nucleotides. For HIV, Tajima's D also provides evidence for exponential growth, which agrees with the rapid spread of this virus compared to the endemic foot and mouth virus. These results are perhaps not surprising since recombination is being reported in many viruses (Robertson *et al.* 1995; Holmes *et al.* 1999b; Santti *et al.* 1999), but the rates appear here to be so high that phylogenetic analysis may be of very limited value. Dating of events from phylogenetic trees of viruses is therefore likely to be associated with much larger variances than is generally appreciated (Zhu *et al.* 1998).

All five data sets show large among-site heterogeneity in mutation rate when estimated from the phylogenetic analysis. Table 1 shows that part of this heterogeneity might be an artifact of ignoring recombination and may not be due to real differences in the rate of substitution over the sequences. This effect of recombination has not been acknowledged much in the past and some of the claims of high rate heterogeneity in viruses and mtDNA (*e.g.*, Yang and Kumar 1996) may also well be artifacts from ignoring recombination.

DISCUSSION

The results of this study show that ignoring recombination can have large effects on the shape of the inferred phylogenetic tree. Recombination makes sequences more equidistant than expected under the neutral coalescent, *i.e.*, their mean pairwise distance is constant but the variance of their pairwise distance decreases with increasing recombination (Hudson 1983). Thus, it is not surprising that a tree reconstructed ignoring recombination appears more star-like than expected under the neutral coalescent with recombination. We chose to quantify the effect through five tree statistics and four ratios as defined by Uyenoyama (1997). The ratios have the advantage that they are independent of the mutation rate and thus truly measure tree shape. The ratios were found to have power to distinguish between presence and absence of recombination when $\rho > 8$. This conclusion is little affected by different substitution models and rate heterogeneity. The power of the ratios depends on the number of sequences sampled and the number of segregating sites. However, when the average pairwise difference is >20 mutations and the number of sampled sequences is

TABLE 2

Summary of analysis of five population data sets

Locus	mtDNA, control region	mtDNA, D-loop	Envelope gene	SAT1, capsid protein	Vermillion nuclear gene
Number of sequences	71	21	35	31	20
Alignment length	764	369	2874	431	2160
Organism	Human	Grant's gazelle	HIV1, subtype B	Foot and mouth disease virus	*D. melanogaster*
Population	African	Nairobi	American, years 1986–1990	Southern Africa	Kenya and Zimbabwe
Outgroup	Chimpanzee	Grant's gazelle, Tsavo population	HIV1, subtype D, African sample, 1986	Sat1 sequence from Uganda	*D. simulans*
Average number of nucleotide differences	15	15	184	87	27.2
R_{SD}	10.8	3	19	8.9	8.3
R_{ST}	2.62	2.25	3.62	2.67	2.88
R_{BD}	0.05	0.79	0.89	0.38	0.41
R_{PT}	0.37	0.57	0.29	0.49	0.47
Rate heterogeneity α	0.13	0.15	0.27	0.35	0.04
Tajima's D	−1.27	−0.59	−1.92	−0.33	−0.48
ρ	43.3	34.8	623	127.3	259
Reference	Vigilant *et al.* (1991)	Arctander *et al.* (1996)	Korber *et al.* (1998)	A. D. S. Bastos (personal communication)	Begun and Aquadro (1995)

Data sets were aligned with ClustalX. The aligned sequences were then run through DNAdist and Fitch of PHYLIP (Felsenstein 1995) using the designated outgroup to root the tree. Substitution model used was HKY85. Tajima's D and ρ based on Hey and Wakeley's (1997) method was calculated using SITES (Hey and Wakeley 1997). Rate heterogeneity was calculated using PAML 2.0g (Yang 1999), using the tree topology suggested from Fitch.

>20, the evolutionary variance dominates and including more segregating sites or sequences would yield little added power. The method of phylogenetic inference, though, does have an effect. We focused the most effort on distance-based methods because they are still more widely used than methods based on maximum likelihood, in particular for large data sets. With distance-based methods, ignoring recombination leads to an underestimate of the inferred time to the most recent common ancestor, whereas with maximum-likelihood-based methods, the total number of inferred mutations is more strongly elevated. However, the qualitative effect on the ratios is similar for the two inference methods.

Almost all DNA in all organisms appears to have the capacity to recombine. Recombination is being reported from bacterial species and viruses, and recent analysis of mammalian mtDNA data suggests that the mitochondrion may be able to recombine, too (Awadalla *et al.* 1999; Eyre-Walker *et al.* 1999). Thus, only Y chromosomes and perhaps cpDNA satisfy the assumption of phylogenetic analysis of intraspecific sequences. This is unfortunate since the phylogeny contains information not contained in unordered summary statistics (Felsenstein 1992) and can be used to date mutations and lineage divergence under the assumption of a molecular clock (Leitner and Albert 1999). For example, if timing of events is to be estimated, then the biases caused by recombination will make it look as if some of the lineages diverged a longer time ago than they actually did but that the polymorphism as a whole is younger. This is important for estimation of the level of *trans*-specific polymorphism from genealogies and thus for the estimation of long-term effective population size of the species in question (Clark 1997, and references therein). It also has important consequences for inferences from sequences of systems under balancing selection, such as self-incompatibility and MHC (Ayala 1995; M. H. Schierup, A. Mikkelsen and J. Hein, unpublished results).

From a more practical point of view, it is of interest to ask whether the amounts of recombination shown to have a large effect are likely in typical data sets. We find that $\rho = 4Nr = 8$ has a large effect and want to translate this number into the number of base pairs in different organisms. Here r is the recombination rate for the whole gene, *i.e.*, $r = Lr'$, where L is the number of base pairs and r' the recombination rate per base pair per generation in morgans. In humans, *D. melanogaster*, and *Arabidopsis thaliana*, r' can be calculated as an average over the whole genome to be $\sim 10^{-8}$, 2×10^{-8}, and 4×10^{-8}, respectively. Estimates of N are much less accurate, but current consensus appears to favor $N \approx 10^6$ for *D. melanogaster* and $N \approx 10^5$ for humans. Thus, $\rho = 8$ equals 100 bp in *D. melanogaster* and 2000 bp in humans. These are average numbers, which vary extensively with the recombination rate over the genome. However, the numbers illustrate that recombination rates with large consequences for phylogenetic inferences will be common in typical nuclear data sets, and that recombination has a large effect even when it is difficult to prove statistically (as within a 100 bp segment of a Drosophila gene). This is an important observation in relation to noneukaryotes such as bacteria, viruses, and organelles, where recombination rates are much more difficult to estimate. Our simple analysis of four data sets showed that recombination may be sufficiently high to invalidate the use of phylogenetic trees in many population studies.

In its effect on the phylogeny, estimated from a sample of allelic sequences, recombination mimics exponential growth and the ratios alone cannot distinguish between these two alternatives. Thus, claims about exponential growth, *e.g.*, through mismatch distributions (Slatkin and Hudson 1991), are also compatible with recombination. This may be important for interpretations of mitochondrial data sets. However, it should be possible to distinguish the two forces. Recombination causes many apparent homoplasies, high apparent rate heterogeneity, loss of a molecular clock (Schierup and Hein 2000), and an expected decay of linkage disequilibrium with distance (Miyashita and Langley 1988; Awadalla *et al.* 1999). None of these are expected under exponential growth; this can be detected mainly by negative Tajima's D values.

CONCLUSIONS

Ignoring recombination in tree-based analysis of sequence data from populations may lead to the following important artifacts:

1. Underestimation of the time to most recent common ancestor
2. Underestimation of the amount of recent divergence (long terminal branches)
3. Overestimation of the number of mutations
4. Apparent signs of exponential growth
5. Apparent substitution rate heterogeneity among sites
6. Apparent parallel substitutions
7. Loss of a molecular clock
8. More apparent ancient polymorphism (*trans*-specific evolution)

Methods that include recombination in phylogenetic estimation of evolutionary parameters are needed before full use can be made of population sequence data. Such work has recently started (*e.g.*, Griffiths and Marjoram 1996; Stephens and Donnelly 2000), but needs further development before it can be applied to the standard experimental data set of today.

We thank Thomas Christensen and Anders Mikkelsen for excellent computer programming, and the Department of Computer Science for computing resources. Deborah Charlesworth, Gilean McVean, Carsten Wiuf, Xavier Vekemans, Philip Awadalla, Roald Forsberg, and two anonymous reviewers all made very valuable comments to a previous version of the manuscript. Amanda Bastos, Onderstepoort

Veterinary Institute, kindly provided the unpublished SAT1 data set. The study was supported by grant no. 9701412 from the Danish Natural Sciences Research Council and by the Basic Research in Computer Science (BRICS) Centre of the Danish National Research Foundation.

LITERATURE CITED

Arctander, P., P. W. Kat, R. A. Aman and H. R. Siegismund, 1996 Extreme genetic differences among populations of Gazella granti, Grant's gazelle, in Kenya. Heredity **76:** 465–475.

Awadalla, P., A. Eyre-Walker and J. M. Smith, 1999 Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286:** 2524–2525.

Ayala, F. J., 1995 The myth of Eve: molecular biology and human origins. Science **270:** 1930–1936.

Beerli, P., and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Begun, D. J., and C. F. Aquadro, 1995 Molecular variation at the vermilion locus in geographically diverse populations of *Drosophila melanogaster* and *Drosophila simulans*. Genetics **140:** 1019–1032.

Clark, A. G., 1997 Neutral behavior of shared polymorphism. Proc. Natl. Acad. Sci. USA **94:** 7730–7734.

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

Felsenstein, J., 1981 Evolutionary trees from DNA-sequences—a maximum-likelihood approach. J. Mol. Evol. **17:** 368–376.

Felsenstein, J., 1992 Estimating effective population-size from samples of sequences—inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

Felsenstein, J., 1995 *PHYLIP (Phylogeny Inference Package) Version 3.572.* Distributed over the World Wide Web, Seattle.

Griffiths, R. C., 1999 The time to the ancestor along sequences with recombination. Theor. Popul. Biol. **55:** 137–144.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479–502.

Griffiths, R. C., and S. Tavare, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

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

Holmes, E. C., O. G. Pybus and P. H. Harvey, 1999a. The molecular population dynamics of HIV-1, pp. 177–207 in *The Evolution of HIV*, edited by K. A. Crandall. The Johns Hopkins University Press, Baltimore.

Holmes, E. C., M. Worobey and A. Rambaut, 1999b Phylogenetic evidence for recombination in dengue virus. Mol. Biol. Evol. **16:** 405–409.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

Kelsey, C. R., K. A. Crandall and A. F. Voevodin, 1999 Different models, different trees: The geographic origin of PTLV-I. Mol. Phylogenet. Evol. **13:** 336–347.

Kingman, J. F. C., 1982 The coalescent. Stochastic Process. Appl. **13:** 235–248.

Korber, B., C. L. Kuiken, B. Foley, B. Hahn, F. McCutchan *et al.*, 1998 *Human Retroviruses and AIDS: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences.* Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

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

Kuhner, M. K., J. Yamato and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429–434.

Leitner, T., and J. Albert, 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. USA **96:** 10752–10757.

Li, W.-H., 1997 *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

May, G., F. Shaw, H. Badrane and X. Vekemans, 1999 The signature of balancing selection: fungal mating compatibility gene evolution. Proc. Natl. Acad. Sci. USA **96:** 9172–9177.

Miyashita, N., and C. H. Langley, 1988 Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured population. J. Math. Biol. **29:** 59–75.

Olsen, G. J., H. Matsuda, R. Hagstrom and R. Overbeek, 1994 FastDnaML: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10:** 41–48.

Robertson, D. L., P. M. Sharp, F. E. McCutchan and B. H. Hahn, 1995 Recombination in HIV-1. Nature **374:** 124–126.

Santti, J., T. Hyypia, L. Kinnunen and M. Salminen, 1999 Evidence of recombination among enteroviruses. J. Virol. **73:** 8741–8749.

Schierup, M. H., and J. Hein, 2000 Recombination and the molecular clock. Mol. Biol. Evol. **17**(9).

Schierup, M. H., X. Vekemans and F. B. Christiansen, 1998 Allelic genealogies in sporophytic self-incompatibility systems in plants. Genetics **150:** 1187–1198.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial-DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics. R. Stat. Soc. Ser. B (in press).

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

Uyenoyama, M. K., 1997 Genealogical structure among alleles regulating self-incompatibility in Angiosperms. Genetics **147:** 1389–1400.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson, 1991 African populations and the evolution of human mitochondrial-DNA. Science **253:** 1503–1507.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

Wiuf, C., and J. Hein, 1997 On the number of ancestors to a DNA sequence. Genetics **147:** 1459–1468.

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. **55:** 248–259.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Yang, Z., 1999 *Phylogenetic Analysis by Maximum Likelihood (PAML), Version 2.0g.* University College, London.

Yang, Z., and S. Kumar, 1996 Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13:** 650–659.

Zhu, T. F., B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp *et al.*, 1998 An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature **391:** 594–597.

Communicating editor: W. Stephan