A Comparison of Estimators of the Population Recombination Rate

Jeffrey D. Wall

Department of Ecology and Evolution, University of Chicago

Three new estimators of the population recombination rate $C = 4Nr$ are introduced. These estimators summarize the data using the number of distinct haplotypes and the estimated minimum number of recombination events, then calculate the value of C that maximizes the likelihood of obtaining the summarized data. They are compared with a number of previously proposed estimators of the recombination rate. One of the newly proposed estimators is generally better than the others for the parameter values considered here, while the three programs that calculate maximum-likelihood estimates give conflicting results.

Introduction

Two important scaled parameters in classical panmictic theory without selection are the population mutation parameter $\theta = 4N\mu$ (where N is the effective population size and μ is the per-locus mutation rate per generation) and the population recombination parameter $C = 4Nr$ (where r is the per-locus recombination rate per generation). New mutations are the ultimate source of all variation; under the infinite-sites model (where every new mutation is assumed to occur at a previously unmutated site), the expected number of segregating sites in a sample is directly proportional to θ (Watterson 1975). The recombination rate has a more subtle effect on the expected amount of linkage disequilibrium between segregating sites and the expected correlation between genealogies of nearby sites (e.g., Griffiths 1981; Hudson 1983; Kaplan and Hudson 1985; Pluzhnikov and Donnelly 1996).

Two main approaches have been employed for estimating C. One method observes the frequency of sequence exchange between distant markers (e.g., Ashburner 1989; True, Mercer, and Laurie 1996; Bouffard et al. 1997; Nagaraja et al. 1997), while the other method estimates C from the patterns of sequence variation expected in random population samples (e.g., Hudson and Kaplan 1985; Hudson 1987; Griffiths and Marjoram 1996; Hey and Wakeley 1997; Wakeley 1997; Kuhner, Yamato, and Felsenstein 1999). If there is local variation in recombination rates (e.g., Fullerton et al. 1994), then the former method might be quite inaccurate for a particular region of interest. In this paper, I focus on ways of estimating C from sequence variation in random samples.

Researchers have used different aspects of the expected patterns of sequence variation to estimate C. For example, Hudson (1987) used the observed variance of the number of pairwise differences to construct an estimator of C (here called C_{hud}). This variance is a measure of the amount of linkage disequilibrium in a sample; as the recombination rate increases, the expected amount of linkage disequilibrium between segregating sites decreases, as does the expected variance of the number of pairwise differences. Wakeley (1997) made two small technical improvements in C_{hud} to obtain a new estimator, which I call C_{wak}. Although these two estimators are easy to calculate, they ignore most of the available information by essentially describing the data with a single summary statistic. Another concern is that their method of moments approach is probably not optimal; maximum-likelihood methods might be more appropriate. Perhaps the most elegant approach is to condition on the complete data set to obtain a maximum-likelihood estimate of C (Griffiths and Marjoram 1996; Kuhner, Yamato, and Felsenstein 1999; Nielsen, personal communication). The full likelihood approach has the advantage of using all of the information that is available, but it is extremely computationally intensive. For example, it is not yet computationally feasible to estimate C for many human polymorphism data sets (e.g., Harding et al. 1997; Zietkiewicz et al. 1997; Nickerson et al. 1998) using maximum-likelihood estimation programs provided by R. C. Griffiths and M. Kuhner. One possible compromise is the approach of Hey and Wakeley (1997). Their estimator γ is close to the average of maximum-likelihood estimates of many small subsets of the data. Another possible compromise (and the one adopted here) is to describe sequence data using summary statistics, but then to use maximum-likelihood methods on these summaries of the data to estimate C (see below).

Recently, some researchers have proposed statistical tests based on the observed number of distinct haplotypes in a sample (Strobeck 1987; Fu 1996, 1997; Depaulis and Veuille 1998). The motivation is that some forces tend to decrease the expected number of distinct haplotypes (e.g., population subdivision, balancing selection, or decreasing population size) (Fu 1996; Wall 1999), while other forces act in the opposite direction (e.g., linkage to a selective sweep or increasing population size) (Fu 1997). However, the latter forces are confounded with the effect of intragenic recombination: the expected number of distinct haplotypes increases rapidly as the recombination rate increases. As a result, the above statistical tests are inappropriate when the recombination rate is unknown.

Instead, I use the number of distinct haplotypes (denoted hereinafter as H) to estimate C. The general approach is to find what value of C maximizes the probability of obtaining the observed value of H. This method of using maximum likelihood on summary statistics

Key words: recombination, coalescent simulations, evolutionary inference, statistical tests of neutrality.

Address for correspondence and reprints: Jeffrey D. Wall, Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637. E-mail: jdwall@midway.uchicago.edu.

Mol. Biol. Evol. 17(1):156–163. 2000

has been used before (e.g., Fu and Li 1997; Takahata and Satta 1997; Weiss and von Haeseler 1998), and it is expected to be useful if data sets have large sample sizes and relatively few segregating sites. These criteria are satisfied by most recent studies of human genetic variation (e.g., Harding et al. 1997; Zietkiewicz et al. 1997; Nickerson et al. 1998), but not by most studies of *Drosophila* sequence variation. Although exact results are available for the joint distribution of H and S (the observed number of segregating sites) when there is no recombination (Griffiths 1982), it is difficult, if not impossible, to obtain analytic results in the presence of intragenic recombination. Therefore, simulations are used to estimate likelihoods (see below).

Other summary statistics can also be used for maximum-likelihood estimation. Hudson and Kaplan (1985) proposed using the minimum number of recombination events (R_{M}) to estimate C. Recombination was inferred to have occurred between a pair of segregating sites when all four gametic types were present. This implicitly assumes an infinite-sites model of mutations (i.e., that each new mutation occurs at a previously unmutated site). Similar to the above, I construct an estimator that maximizes the joint probability of obtaining the observed values of both R_{M} and H.

The simulations that are used to estimate likelihoods can be run in two different ways. They can be run with θ as an unknown parameter or they can be run conditional on S (Hudson 1990, 1993). Hudson (1993) proposed using the latter method, since S is known, whereas θ must be estimated. The drawback is that conditioning on S results in a null model that is slightly different from the standard coalescent one. Instead of a constant mutation rate per generation, there is a constant number of mutations regardless of the length of the tree. I run both types of simulations and thus directly test the effect of conditioning on S. The new estimators of C are compared with many previously proposed ones (Hudson 1987; Griffiths and Marjoram 1996; Hey and Wakeley 1997; Wakeley 1997; Kuhner, Yamato, and Felsenstein 1999; Nielsen, personal communication).

Materials and Methods

Coalescent theory (Kingman 1982*a*, 1982*b*) provides an efficient framework for simulating neutral population histories. I assume that there is a large, constant-sized, panmictic population, no selection, and an infinite-sites model. The mutation rate and recombination rate are taken to be constant per base pair (cf. Hudson 1983, 1990). Simulations either condition on S or treat θ as a parameter (see below), and are run using modifications of programs kindly provided by R. R. Hudson.

One or more summary statistics are used to characterize the data. The ones considered are S, R_{M}, and H. Most simulations are run conditional on S (cf. Hudson 1993). The estimators C_H and C_{HRM} are defined as the values of C that maximize the likelihoods $L(C | H)$ and $L(C | H, R_{\mathrm{M}})$, respectively. For each possible value of S, a range of C values from 0.0 to 300.0 is considered, and 10^5 trials are run for each value of C. $\Pr(H | C_0)$ and $\Pr(H, R_{\mathrm{M}} | C_0)$ are estimated as the proportion of trials with $C = C_0$ that show a particular configuration of $\{H\}$ or $\{H, R_{\mathrm{M}}\}$, respectively.

Standard coalescent simulations (cf. Hudson 1990) are also run. The estimator C_{SH} is defined as the value of C that maximizes the joint likelihood $\mathrm{L}(\theta, C | S, H)$. A range of parameter values is considered, with θ varying from 0.5 to 20.0 and C varying from 0.0 to 260.0. For each combination of θ and C, 10^5 trials are run, and $\Pr(S, H | \theta_0, C_0)$ is estimated as the proportion of trials with $\theta = \theta_0$ and $C = C_0$ showing a particular configuration of S and H. C_{SH} is very similar in motivation to C_H. However, it is much quicker to calculate C_H, since simulations are needed for a single value of S instead of for many possible values of θ.

Coalescent trials were run with fixed values of θ and C, and the distributions of estimator values for these trials are compared with each other. The other estimators of C considered were C_{hud} (Hudson 1987), C_{wak} (Wakeley 1997), C_{GM} (Griffiths and Marjoram 1996; the program recom58 uses Monte Carlo recursion methods to estimate C and was kindly provided by R. C. Griffiths), C_{KYF} (Kuhner, Yamato, and Felsenstein 1999; the program RECOMBINE, version 1.0, incorporates a Metropolis-Hastings [Metropolis et al. 1953; Hastings 1970] Markov chain Monte Carlo [MCMC] algorithm with a finite-sites model and estimates the likelihood surface using importance sampling), C_{N1} (R. Nielsen, personal communication; a preliminary version of the program Baysim was kindly provided by R. Nielsen. Baysim also uses Metropolis-Hastings MCMC, but with an infinite- sites model and a Bayesian approach to parameter estimation assuming a uniform prior. C_{N1} is the maximum-likelihood estimate.), C_{N2} (R. Nielsen, personal communication; C_{N2} is the posterior mean for C), and γ (Hey and Wakeley 1997; the program SITES is available at http://heylab.rutgers.edu). C_H, C_{SH}, C_{hud}, C_{wak}, and γ were compared using a sample size of $n = 50$ and $C = \theta = 1.0, 2.0, \ldots, 12.0$, $C = 2\theta = 2.0, 4.0, \ldots, 24.0$, $C = 4\theta = 4.0, 8.0, \ldots, 48.0$, and $\theta = 5$ with $C = 0.0, 5.0, 10.0, \ldots, 45.0$. Ten thousand trials were run for each estimator and each pair of parameter values. Due to computational constraints, the other estimators (C_{HRM}, C_{GM}, C_{KYF}, C_{N1}, and C_{N2}) were compared only with $n = 50$ and $C = \theta = 3.0$. C_{HRM} was calculated for 1,000 different trials, while 100 trials were run for the other four estimators. For each trial, C_{GM} was estimated from 500,000 replicates with a single set of generating parameters ($C = \theta = 3.0$). Trials for C_{KYF} had 10 short chains (40,000 iterations each), one long chain (600,000 iterations), and starting parameters $C = \theta = \theta_{\mathrm{W}}$ (Watterson's [1975] estimate of θ), while those for C_{N1} and C_{N2} had a burn-in time of 10^6 iterations and a single chain of 5×10^6 iterations. Although more replicates would have been desirable, the amount of time required to run these simulations was already substantial (i.e., 5–6 h per trial of C_{GM} or C_{KYF} on a 400-MHz Pentium II processor). Note that the simulations conducted for this study were meant to be exploratory, not exhaustive.

Table 1
Comparison of Estimators with *n* 50 and *C* = θ = 3.0

Estimator	Trials	Mean	Mean Squared Error	Frequency Undefined	g[a]	Confidence Internal[b] 95%
C_H	10^4	3.86	36.6	0.0008	0.4219	0.0–15.9
C_{SH}	10^4	4.20	34.6	0.0003	0.3506	0.0–16.0
C_{HRM}	10^4	3.20	14.6	0	0.381	0.0–14.1
C_{hud}	10^4	55.5[c]	2.19×10^5[c]	0.0519	0.1992	0.0–∞
C_{wak}	10^4	35.8[c]	2.36×10^5[c]	0.0267	0.2092	0.0–∞
C_{GM}	10^2	0.645	5.80	0	0.10	0.0–1.5
C_{KYF}	10^2	2.50	4.53	0	0.54	0.0–6.9
C_{N1}	10^2	2.43	68.1	0.13	0.05	0.0–∞
C_{N2}	10^2	3.87	42.1	0.13	0.44	0.0–∞
γ	10^2	1.57	13.7	0.0475	0.2066	0.0–∞

NOTE.—See text for definitions of estimators.

[a] The proportion of trials whose estimate is contained in [1.5, 6.0].

[b] The 2.5th and 97.5th percentiles. Estimators that were undefined were taken to have returned +∞.

[c] Arbitrarily large (but finite) estimates were given a value of 20,000.

The mean and the mean squared error (MSE) are calculated for the estimators' distributions. These two summaries are inadequate, both because the distributions are not approximately normal (see Results) and because estimators do not always return defined and finite values. C_{hud}, C_{wak}, and γ are undefined when there is not enough information (i.e., not enough segregating sites) or when the patterns of variation are consistent with free recombination. Similarly, C_H, C_{SH}, and C_{HRM} are undefined whenever $S < 2$ or $H \approx 2^S$. In addition, sometimes C_H, C_{SH}, and C_{HRM} could not be calculated (and were considered undefined) because simulations with high enough values of C were not run. This latter case was always rare. It did not occur in any of the trials shown in table 1 and does not affect the summary statistic described below. I present histograms of the distributions when $C = \theta = 3.0$, and for the other simulations I summarize the results using

$$g = \Pr\left(0.5 \leq \frac{\hat{C}}{C} \leq 2.0\right),$$

where g is the probability that a particular estimator will return a good estimate (defined arbitrarily as the probability that the estimate is within a factor of 2 of the actual value). g might be a more appropriate measure than the bias or MSE for comparing highly skewed distributions.

Results

Figure 1 shows histograms of the distributions of all 10 estimators when $n = 50$ and $C = \theta = 3.0$. Summary measures of these distributions are presented in table 1. As can be seen, there is a wide variation in the shapes of the distributions and the qualities of the estimators. The 95% confidence intervals (i.e., the range between the 2.5th and 97.5th percentiles) suggest that an actual value of $C = 3.0$ is consistent with a wide range of estimated values. In fact, $C = 3.0$ is consistent with any possible observed value of C_{hud}, C_{wak}, C_{N1}, C_{N2}, or γ. The best estimator is C_{KYF}; it has the second smallest bias, the smallest MSE, and the greatest proportion of estimates within a factor of two of the actual value. C_{N2} is also generally a good estimator, although a few extreme values greatly increase its MSE. On the other end of the spectrum are the other two full likelihood estimators, C_{GM} and C_{N1}; they almost always underestimate the actual value.

This variation in the qualities of maximum-likelihood estimators is surprising. In principle, C_{KYF} and C_{N1} are computing the same quantity, so they should converge to the same value as the number of iterations increases. C_{GM} assumes an outgroup (unlike the other two), so it should converge to a different number. Table 2 shows the actual values of C_{KYF}, C_{GM}, and C_{N1} for the first 20 trials. Although the values of C_{GM} and C_{N1} are often quite similar to each other, neither seem to be strongly correlated with C_{KYF}. To determine whether this discrepancy was due to C_{KYF}'s initial parameter values, all 100 trials were rerun with starting values of $5C = \theta = \theta_{\mathrm{W}}$. Having a lower starting value of C has very little effect on the final estimate of C (the mean was 2.31, compared with 2.50 in table 1). There is also some concern that the choice of generating parameters for C_{GM} might have some influence on the final estimate. The first 20 trials were rerun (with 10^6 replicates) using the estimated values of θ and C as new generating parameters. This also had a minor effect on the estimate of C_{GM} (results not shown). Thus, sensitivity to initial parameter values explains very little if any of the difference in the distributions of C_{KYF} versus C_{GM} or C_{N1}.

The most disparate trial in table 2 (trial 11) was chosen for further simulations. To ascertain the effect of a limited number of iterations on the results, all estimators were rerun with 10 times the number of iterations. The new estimates from these runs (shown in table 3) are quite similar to the old ones. In addition, trial 11 was rerun 20 times (using the same parameters as in table 1) with different random number seeds. The mean and variance of the estimated values for trial 11 are shown in table 3. The variance gives some idea of how far the estimators are from converging to a single estimate (presumably the MLE). Table 3 shows that these variances are not that large, which reinforces the ob-

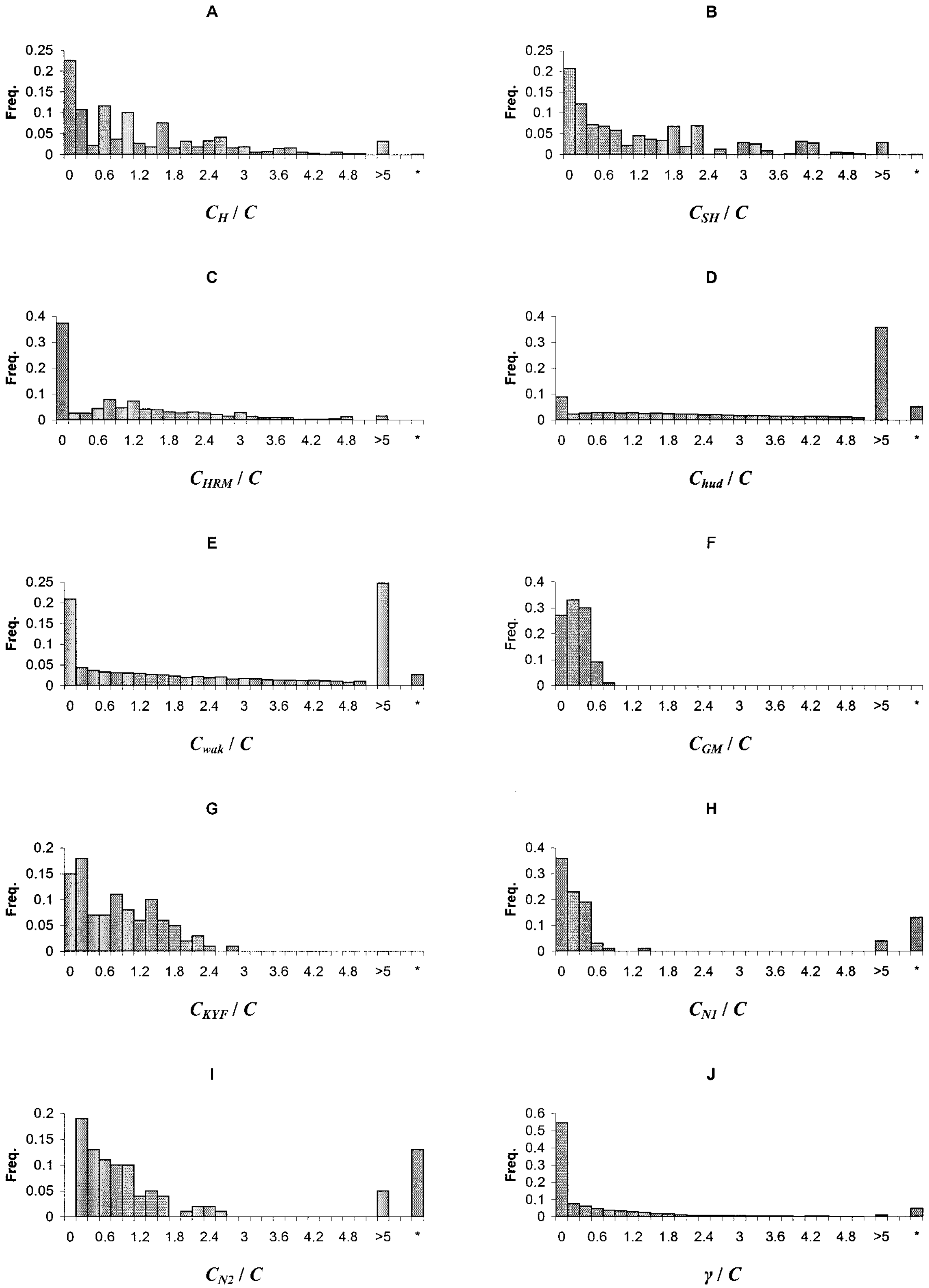

FIG. 1.—Histograms of the distributions of all 10 estimators when $n = 50$ and $C = \theta = 3.0$. *A–J* show, in order, C_H, C_{SH}, C_{HR_M}, C_{hud}, C_{wak}, C_{GM}, C_{KYF}, C_{N1}, C_{N2}, and γ, each normalized by the actual value of *C*. The number of independent trials run for each estimator is shown in table 1. The "*" column displays the frequency of trials that are undefined.

Table 2
Comparison of Maximum-Likelihood Estimators for the First 20 Trials of Table 1

Trial No.	C_{GM}	C_{KYF}	C_{N1}
1	0.0	2.06	0.02
2	1.2	3.16	0.74
3	1.2	6.22	1.38
4	0.3	2.12	—[a]
5	2.1	2.76	0.94
6	0.0	0.63	0.02
7	0.0	0.22	0.02
8	0.6	2.68	0.58
9	0.9	4.58	0.90
10	0.0	0.80	0.02
11	0.6	6.59	0.42
12	0.0	1.42	0.02
13	1.2	0.00	0.86
14	0.9	5.30	1.14
15	0.6	1.19	0.78
16	0.9	0.54	1.98
17	0.9	3.27	0.38
18	0.6	2.08	0.14
19	0.3	3.33	0.74
20	0.9	3.36	1.42

[a] Baysim did not return a value for this trial.

Table 3
Comparison of Maximum-Likelihood Estimators for Trial 11 (See Text)

	C_{GM}	C_{KYF}	C_{N1}
Original estimate[a]	0.6	6.59	0.42
Revised estimate[b].	1.2	5.42	0.82
Mean[c]	0.95	4.75	0.75
Variance[c]	0.048	2.16	0.064
Range[c]	0.6—1.5	1.16–6.70	0.06—1.18

[a] Values for trial 11 of table 2.

[b] Simulations were rerun with 10 times the number of iterations for each program.

[c] Based on 20 replicates with different random number seeds, each run with the same data set and the same parameter values as in tables 1 and 2.

served discrepancies between the different estimators. All 20 of the trials for C_{GM} and C_{N1} returned values that were less than all but one of the 20 different C_{KYF} values.

The three new estimators compare favorably (especially using g) with all of the others except C_{N2} and C_{KYF}. C_{HRM} seems to be the best of the three, while the properties of C_H and C_{SH} seem quite similar. Thus, it does not seem to make much difference whether one conditions on S or not. The raggedness of the distributions of C_H and C_{SH} is due to the inherent discreteness of the estimators; there are not that many combinations of S and H that are possible. γ has moderate bias, but its low MSE seems to be an artifact of its downward bias and the consequent artificial truncation at 0. Notice that the estimator with the second highest bias and the highest MSE (C_{wak}) actually has a higher value of g than does γ. This illustrates how looking only at the mean and the MSE can be misleading; these two summary statistics are highly sensitive to the tail of the distribution.

g was calculated for C_H, C_{SH}, C_{hud}, C_{wak}, and γ under a wider range of parameter values. Figure 2*A* shows simulations with $C = \theta$, figure 2*B* assumes $C = 2\theta$, and figure 2*C* assumes $C = 4\theta$. As expected, all estimators improve as θ increases. As above, C_H and C_{SH} are roughly of the same quality, and both usually have higher g values than the other three estimators. γ is poor for small data sets but improves more rapidly with increasing θ than do the others in figure 2*A* and *B*. However, γ is relatively less accurate in figure 2*C* for higher recombination rates. When the actual recombination rate is so high, γ's downward bias has a large influence in lowering the value of g. This trend is highlighted in figure 3, which plots g as a function of C when $\theta = 5$. While other estimators get better as C increases (presumably because less linkage disequilib-

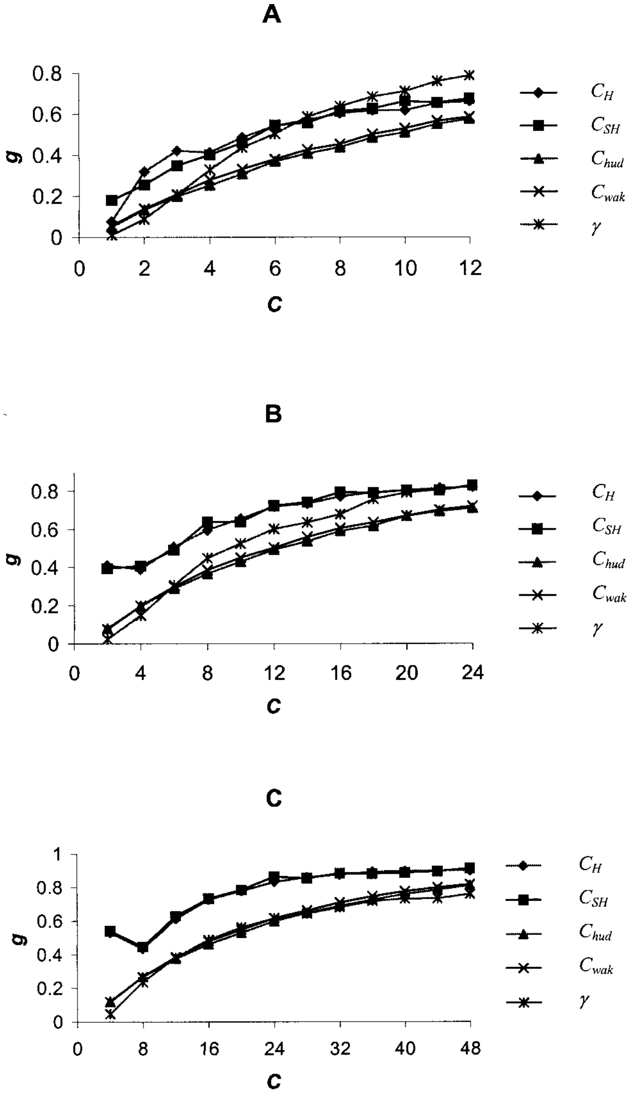

FIG. 2.—The population recombination rate, C, versus g, the proportion of estimated values within a factor of 2 of the true value of C. *A* assumes $C = \theta$, *B* assumes $C = 2\theta$, and *C* assumes $C = 4\theta$. Ten thousand replicates were run for each estimator and each parameter combination.

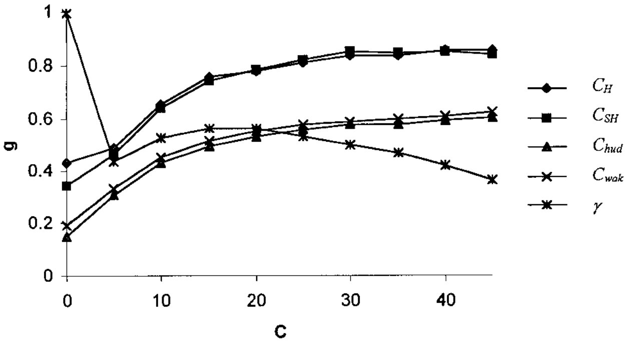

FIG. 3.—The population recombination rate, C, versus g, the proportion of estimated values within a factor of 2 of the true value of C, with $\theta = 5$. Ten thousand replicates were run for each estimator and each parameter combination.

rium provides more useful information), γ's g value starts decreasing when $C > 20$. In both figures 2 and 3, C_{hud} and C_{wak} perform almost identically, with the latter having marginally higher g values. In summary, all of the test statistics are not very good unless both C and θ are reasonably high. For example, in figure 2*A*, when $C = \theta = 5.0$ (and the expected number of segregating sites is roughly 22), all estimators are within a factor of 2 of the true value less than half the time. In contrast, a simple estimator of θ (based on the observed value of S; cf. Watterson 1975) for these simulations is within a factor of 2 of the true value of θ almost all of the time.

Discussion

One of the fundamental goals of molecular population genetics is to infer demographic and selective history from the patterns observed in current sequence data. Evolutionary inference requires an accurate estimate of the recombination rate, since the relative likelihoods of the data under various models are a function of C. In particular, the standard assumption of no recombination ($C = 0$) can lead to misleading conclusions. For example, consider the *Runt* data set from *Drosophila simulans* (Labate, Biermann, and Eanes 1999). This data set has a sample size of $n = 11$ and $S = 20$. Tajima's (1989) D statistic for this data set is -1.47. If we assume panmixia and no recombination, then $P = \Pr(D \leq -1.47) = 0.065$ (10^5 coalescent simulations conditional on $S = 20$ were run to obtain this probability and the others described below). If, instead, we assume a symmetric island model (with two islands, $4Nm = 0.5$, $C = 0$, and roughly equal sampling from each island), then $P = \Pr(D \leq -1.47) = 0.036$. Hence, with a no-recombination model, the value of Tajima's D statistic is negative (but not significantly so) and is slightly less likely under a simple island model than under panmixia. However, this assumption of no recombination is known to be incorrect. Our best guess for the scaled recombination rate of the *Runt* gene is $\hat{C} = 84.6$; this assumes a conservative value of $N = 10^6$ for the population size of *D. simulans* and estimates the recombination rate from the observed rate of exchange between flanking markers (True, Mercer, and Laurie 1996). If null simulations are rerun with $C = 84.6$, then the resultant probabilities are $P = \Pr(D \leq -1.47) = 0.00087$ under panmixia and $P = \Pr(D \leq -1.47) = 0.00001$ under the island model. Not only are both probabilities drastically lower (suggesting that the data are not very consistent with either model), but the ratio of the two probabilities has changed as well. Once recombination has been included, the particular island model considered is much less likely than the panmictic model. In practice, the island model would probably be discarded, both because of the low absolute P value and the low relative P value. However, when $C = 0$, the island model would remain a plausible option, since it is not much less likely than the panmictic model. Which alternatives are plausible depends on what recombination rate is assumed.

Previous researchers have noted that C_{hud} and C_{wak} are effective only for very large data sets (e.g., $C = \theta \approx 100$) (Hudson 1987; Wakeley 1997). To my knowledge, there are no sequence polymorphism data sets with sample sizes greater than three and $\theta > 50$. Thus, there is a premium on estimators that are still effective even when θ is relatively small. The results obtained here suggest that while some estimators are reasonable for intermediate-sized data sets, all of them (with the possible exceptions of C_{N2} and C_{KYF}) behave poorly for small data sets (e.g., $\theta < 5$).

Figure 2 implies that for most of the parameter values considered, all ad hoc estimators use no more information than a simple count of the number of distinct haplotypes that are present (i.e., other estimators are generally not as good as C_H). Although more clever ways of summarizing the data will lead to better (yet still easy to calculate) estimators of the population recombination rate, it will always be easier to estimate θ than to estimate C. This is not only because it is conceptually more difficult to understand the effects of recombination on the patterns of observed sequence variation, but also because given data sets can be consistent with a wide range of possible recombination rates. For example, nearly half of the trials shown in figure 1 and table 1 are consistent with no recombination at all.

One open question is whether estimators that summarize the data (e.g., C_H, C_{SH}, C_{HRM}, $\mathrm{C}_{\mathrm{hud}}$, and C_{wak}) or estimators that only consider subsets of the data (e.g., γ) are throwing away valuable information. As computational power increases, methods that condition on all of the data will become feasible for a wider range of data sets. In addition, technical advancements will increase the efficiency of existing likelihood algorithms. A program written by P. Fearnhead which incorporates an importance sampling scheme appears to be a substantial improvement over the Griffiths and Marjoram (1996) method (results not shown; P. Fearnhead, personal communication). Although in certain circumstances full maximum-likelihood estimators are expected to be asymptotically optimal, there is no guarantee that they are the best estimators for actual data sets (for which there is dependency and the small sizes are far from the asymptotic limit). It is possible that the MLE is highly biased (see, e.g., C_{GM} and C_{N1} in tables 1 and 2), so other likelihood-based methods like the Bayesian approach of C_{N2} might be best. In any case, table 2 suggests that there are probably still some methodological problems with more than one of the programs used.

Ad hoc estimators will still be necessary for analyzing larger data sets. Of these, compromise approaches that use likelihood methods on summaries of the data (such as C_H, C_{SH}, C_{HRM}, and γ) look promising. An estimator of C based on the likelihood of observed pairwise sample configurations appears to be superior to C_H (Hudson 1993; R. Hudson, personal communication). Of the estimators considered here, C_H or C_{HRM} is preferred as long as the sample size is not too small; otherwise, γ is probably the best, although one should keep in mind that γ's downward bias is large when the recombination rate is high.

Acknowledgments

I thank R. Hudson, M. S. McPeek, and M. Przeworski for helpful discussions and R. Hudson, M. Przeworski, S. Sawyer, and two anonymous reviewers for comments on an earlier version of this manuscript. P. Fearnhead, B. Griffiths, J. Hey, R. Hudson, M. Kuhner, and R. Nielsen generously provided computer programs.

LITERATURE CITED

Ashburner, M. 1989. Drosophila: a laboratory handbook. Cold Spring Harbor Laboratory Press, New York.

Bouffard, G. G., J. R. Idol, V. V. Braden et al. (14 coauthors). 1997. A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. Genome Res. **7**:673–692.

Depaulis, F., and M. Veuille. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15**:1788–1790.

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

———. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**:915–925.

Fu, Y.-X., and W.-H. Li. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. **14**:195–199.

Fullerton, S. M., R. M. Harding, A. J. Boyce, and J. B. Clegg. 1994. Molecular and population analysis of allelic sequence diversity at the human betaglobin locus. Proc. Natl. Acad. Sci. USA **91**:1805–1809.

Griffiths, R. C. 1981. Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19**:169–186.

———. 1982. The number of alleles and segregating sites in a sample from the infinite-alleles model. Adv. Appl. Probab. **14**:225–239.

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

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox, J. A. Schneider, D. S. Moulin, and J. B. Clegg. 1997. Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60**:772–789.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**:97–109.

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

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

———. 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50**: 245–250.

———. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyma and J. Antonovics, eds. Oxford surveys in evolutionary biology. Vol. 7. Oxford University Press, New York.

———. 1993. The how and why of generating gene genealogies. Pp. 23–36 *in* N. Takahata and A. G. Clark, eds. Mechanisms of molecular evolution. Sinauer, Sunderland, Mass.

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.

Kaplan, N. L., and R. R. Hudson. 1985. The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. Theor. Popul. Biol. **28**:382–396.

Kingman, J. F. C. 1982*a.* On the genealogy of large populations. J. Appl. Probab. **19A**:27–43.

———. 1982*b*. The coalescent. Stochastic Proc. Appl. **13**:235–248.

Kuhner, M. K., J. Yamato, and J. Felsenstein. 1999. RECOMBINE. Version 1.0. Available at http://evolution.genetics.washington.edu/lamarc.html.

Labate, J. A., C. H. Biermann, and W. F. Eanes. 1999. Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans.* Mol. Biol. Evol. **16**:724–731.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. **21**: 1087–1091.

Nagaraja, R., S. Macmillan, J. Kere et al. (25 co-authors). 1997. X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. Genome Res. **7**:210–222.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. Stengard, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C. F. Sing. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat. Genet. **19**:233–240.

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

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

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

Takahata, N., and Y. Satta. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc. Natl. Acad. Sci. USA **94**:4811–4815.

True, J. R., J. M. Mercer, and C. C. Laurie. 1996. Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics **142**:507–523.

Wakeley, J. 1997. Using the variance of pairwise differences to estimate the recombination rate. Genet. Res. **69**:45–48.

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

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256–276.

WEISS, G., and A. VON HAESELER. 1998. Inference of population history using a likelihood approach. Genetics **149**: 1539–1546.

ZIETKIEWICZ, E., V. YOTOVA, M. JARNIK et al. (11 co-authors). 1997. Nuclear DNA diversity in worldwide distributed human populations. Gene **205**:161–171.

STANLEY SAWYER, reviewing editor

Accepted October 6, 1999