Estimating Recombination Rates Using Three-Site Likelihoods

Jeffrey D. Wall

Program in Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089

Manuscript received December 11, 2003
Accepted for publication March 22, 2004

ABSTRACT

We introduce a new method for jointly estimating crossing-over and gene conversion rates using sequence polymorphism data. The method calculates probabilities for subsets of the data consisting of three segregating sites and then forms a composite likelihood by multiplying together the probabilities of many subsets. Simulations show that this new method performs better than previously proposed methods for estimating gene conversion rates, but that all methods require large amounts of data to provide reliable estimates. While existing methods can easily estimate an “average” gene conversion rate over many loci, they cannot reliably estimate gene conversion rates for a single region of the genome.

In standard population genetic models, levels of linkage disequilibrium (LD) are primarily determined by the compound parameter \(\rho = 4Na\), where \(N\) is the (diploid) effective population size and \(a\) is the recombination rate per generation per base pair (Hudson 1985; Pritchard and Przeworski 2001). It is well known that, all else being equal, more recombination leads to less LD (e.g., Ohta and Kimura 1969, 1971); in addition, larger population sizes lead to less genetic drift, thus lower levels of LD. From extant sequence polymorphism data alone, neither \(N\) nor \(a\) can be estimated individually. Estimates of \(\rho\) can be thought of as multisite measures of LD (Pritchard and Przeworski 2001; Wall 2001). Accurate estimates of \(\rho\) would be useful for the future design of large-scale association studies (e.g., Kruglyak 1999; Pritchard and Przeworski 2001; Wall and Pritchard 2003), as well as for distinguishing between alternative evolutionary models (Wall 2001; Wall et al. 2002).

The development of methodology for estimating \(\rho\) from patterns of variation has long been an active research area, and there is a range of different methods for estimating \(\rho\) (under a simple model where all recombination events are crossovers and the phase of all double heterozygotes is known). The first methods developed were ad hoc moment estimators (Hudson 1987; Hey and Wakeley 1997; Wakeley 1997). These are quick and easy to use, but do not efficiently utilize the available information. In contrast, full maximum-likelihood techniques are attractive because they make full use of the available haplotype information. Several different implementations have been proposed, which use either importance sampling (Griffiths and Marjoram 1996; Fearnhead and Donnelly 2001) or Markov chain Monte Carlo (Kuhner et al. 2000; Nielsen 2000) to estimate likelihoods. However, the state spaces for these methods are extremely complex and have many dimensions, so it is hard to evaluate when a program has run for long enough to obtain an accurate estimate of the likelihood. In practice, these methods are computationally infeasible for medium- and large-size data sets (e.g., data sets where \(p \geq 10\), which corresponds to ~15 kb in humans; Wall 2000; Fearnhead and Donnelly 2001). Even with the approximately exponential increase in computer processor speeds, full-likelihood methods will be of limited applicability for the foreseeable future.

To date, several compromise approaches have been proposed, which seek to avoid the computational burdens of calculating the exact likelihood of haplotype configurations while keeping the statistical rigor of a likelihood framework. One method involves describing the data with one or more summary statistics and then estimating \(\rho\) using maximum likelihood on the reduced data (Wall 2000). The success of this approach relies on finding summaries of the data that are highly sensitive to the recombination rate. Both the number of distinct haplotypes and the minimum number of inferred recombination events (cf. Hudson and Kaplan 1985) seem to work reasonably well (Wall 2000; Hudson 2001).

Another effective approach exploits the fact that the likelihood of \(\rho\) can be written as the product of conditional distributions of one haplotype given a sample of the other haplotypes (Li and Stephens 2003). If \(x_1, x_2, \ldots, x_r\) are a sample of \(n\) haplotypes, then

\[
\Pr(x_1, x_2, \ldots, x_r | \rho) = \Pr(x_1 | \rho) \Pr(x_2 | x_1, \rho) \ldots \Pr(x_r | x_1, \ldots, x_{r-1}, \rho).
\]

Li and Stephens proceed by obtaining approximations for these conditional distributions, substituting them...
into the right-hand side of the above equation, and estimating the value of \(p\) that maximizes this product. This method is sensitive to the order of the haplotypes, so the authors estimate their approximate likelihood by averaging over several possible orders (Li and Stephens 2003).

A third approach utilizes a composite likelihood (Hudson 1993, 2001; McVean et al. 2002; Ptak et al. 2004). For a given data set, this method calculates the likelihood (as a function of \(p\)) of the haplotype data at each pair of segregating sites and then forms a composite likelihood by multiplying all these pairwise likelihood curves together. The maximum (composite) likelihood \(\hat{\rho}_{\text{cl}}\) is a good point estimate of the recombination rate (Hudson 2001), but, due to dependency among pairs, the standard asymptotic maximum-likelihood assumptions (for the likelihood ratio) do not apply. Hence, the uncertainty in estimates can be assessed only by simulation. In essence, Hudson’s method breaks up the data into small subsets (in this case pairs of sites), calculates the likelihood (as a function of \(p\)) for these subsets, and then multiplies the likelihood functions together to form a composite likelihood. The probabilities for all possible two-site haplotype configurations can be calculated in advance, so the analysis of any particular data set is quite fast once the appropriate file has been generated.

The methods described above assume that all recombination events are crossovers. However, this simple model of recombination is not biologically realistic. Current meiotic recombination models allow for different forms of recombination (e.g., Szostak et al. 1983). Both involve some copying of DNA from one chromosome to another, but they differ in whether the copying is accompanied by the reciprocal exchange of flanking markers (i.e., crossing over) or not. We call these two forms of recombination “crossing over” and “gene conversion,” respectively. There is ample experimental evidence for gene conversion in higher eukaryotes (e.g., Hilliker et al. 1994; Zangenberg et al. 1995; Guillot and De Massy 2002; Jeffreys and May 2004), and some evidence that the empirical patterns of LD also reflect the effects of gene conversion (Langley et al. 2000; Ardlie et al. 2001; Frisse et al. 2001; Przeworski and Wall 2001; Andolfatto and Wall 2003). Unfortunately, direct (laboratory-based) estimation of gene conversion rates is technically challenging and extremely laborious, even for a single locus.

Recently, theoretical models have been developed that incorporate both crossing over and gene conversion (Andolfatto and Nordborg 1998; Wiuf and Hein 2000). Under certain assumptions, it is possible to generalize the composite-likelihood approach of Hudson (2001) to jointly estimate crossing over and gene conversion rates from polymorphism data (Frisse et al. 2001). Given specific crossing-over and gene conversion rates, as well as the distribution of gene conversion tract lengths, one can calculate the function that relates physical to genetic distance. From this, it is straightforward to calculate likelihoods for pairs of sites and to construct a composite likelihood.

In practice, it is very difficult to accurately estimate gene conversion rates from a single genetic region. The effects of crossing over vs. gene conversion on patterns of LD are quite subtle and are often overwhelmed by the inherent stochasticity of the evolutionary process. Generally, accurate estimation of crossing-over and gene conversion rates requires polymorphism data from multiple evolutionarily independent regions. Frisse et al. (2001) assume that recombination rates do not vary across loci and simply multiply the composite likelihoods at each locus to form a composite likelihood for a collection of loci. An alternative approach was proposed by Ptak et al. (2004); they assume that the ratio of gene conversion events to crossing-over events is constant across loci, but that the crossing-over rate varies across loci. Ptak and colleagues argue that crossing-over rates generally vary across loci and that local rates are not necessarily accurately estimated by a comparison of physical and genetic maps.

In this article, we present a new method for estimating recombination rates. We take a composite-likelihood approach similar to Hudson (2001), but we consider triplets of sites instead of pairs of sites. This should allow us to discriminate more efficiently between the effects of crossing over and of gene conversion. Due to memory constraints, we can calculate only three-site likelihoods for small sample sizes. For larger sample sizes, we consider random subsets of the sample and form a composite likelihood by multiplying together the (composite) likelihoods of many subsets (see Methods). Our approach can be used to estimate crossing-over rates only (comparable to Wall 2000, Hudson 2001, and Li and Stephens 2003) or to jointly estimate crossing-over rates and gene conversion rates (comparable to Frisse et al. 2001 and Ptak et al. 2004). We run simulations to compare the relative accuracies of our and previous methods.

**METHODS**

The model: Our approach involves estimating the likelihoods of three-site sample configurations under specified models of recombination. Here “site” refers to a segregating site. We consider a simple neutral model, with no population structure and no changes in population size, and assume we have haplotype polymorphism data from a sample of \(n\) chromosomes. As with Hudson (2001), we assume the limiting case when \(\theta = 4N_{\mu}\), where \(\mu\) is the mutation rate per site per generation) is near \(\theta\) and condition on there being exactly two alleles at each site. Our model of recombination assumes that gene conversion and crossing over are mechanistically independent and that gene conversion tract lengths follow a geometric distribution (Hilliker et al. 1994; Frisse et al. 2001). The model has three parameters: a scaled crossing-over rate \(\rho_{\text{co}}\) (= \(4N_{\text{co}}\),
where \( r_0 \) is the crossing-over rate per base pair per generation), a scaled gene conversion rate \( \rho_0 \) \( (= 4N_{GC} \), where \( N_{GC} \) is the gene conversion initiation rate per base pair per generation), and the average gene conversion tract length \( t \) (in base pairs). As with previous studies (e.g., Frisse et al. 2001; Ptak et al. 2004), we replace \( \rho_0 \) with \( f = \rho_0 / \rho_0 \) so that \( f \) is the ratio of gene conversion initiation events (per base pair) to crossing-over events (per base pair). Simulation under this model is a straightforward generalization of the standard coalescent (Wiuf and Hein 2000; Przeworski and Wall 2001). We assume no variation in recombination rates across the sequence, although this can be relaxed.

There are two different ways of calculating two-site sample probabilities. One method utilizes a recursion described by Goldberg (1984) and Ethier and Griffiths (1990), where exact probabilities can be obtained by solving a large (but sparse) system of linear equations. The other method involves generating a large number of random genealogies using standard coalescent machinery (e.g., Hudson 1983) and estimating the probabilities of different sample configurations separately for each replicate. The latter method proves to be more practical for larger sample sizes (Hudson 2001) and we adopt the same Monte Carlo approach here.

Suppose we have three (segregating) sites, \( s_1 \), \( s_2 \), and \( s_3 \) (in that order along a chromosome), and assume \( N \), \( n \), \( \rho_{GC} \), \( f \), and \( t \) are fixed, where \( \rho_{GC} \) refers to the scaled crossing-over rate per base pair and the distances between \( s_1 \), \( s_2 \), and \( s_3 \) are fixed. We wish to calculate the probability of a particular three-site configuration \( \mathbf{n} \), which we write as \( \Pr(\mathbf{n} | \theta, \rho_{GC}, f, \ell) \). We run coalescent simulations with gene conversion and construct three-locus genealogies \( \mathbf{g} \) by considering the trees at the positions corresponding to the locations of \( s_1 \), \( s_2 \), and \( s_3 \). At site \( s_1 \), the genealogy is a tip-labeled tree with \( 2n - 2 \) branches. Arbitrarily label these 1, 2, 3, \ldots, \( 2n - 2 \) and denote the length (in units of \( 4N \) generations) of the \( i \)th branch as \( \theta_0 \). (Note that for notational convenience our scaling differs by a factor of 2 from the standard scaling.) The branch lengths at \( s_2 \) and \( s_3 \) are defined analogously and are labeled \( \{b_i\} \) and \( \{c_i\} \). We assume that the number of mutations that occur on a branch of length \( x \) is Poisson, with mean \( \theta_0 x \), and set \( \sum x_i^2 = \tau_1 \), \( \sum x_i^2 = \tau_2 \), and \( \sum x_i^2 = \tau_3 \). Define the indicator variable \( I(\mathbf{g}, \mathbf{n}, i, j, k) \) to be 1 if mutations on the \( i \)th, \( j \)th, and \( k \)th branches of the \( s_1 \), \( s_2 \), and \( s_3 \) trees produce the haplotype configuration \( \mathbf{n} \), and 0 otherwise. It is easy to determine the value of \( I(\mathbf{g}, \mathbf{n}, i, j, k) \) from the genealogy \( \mathbf{g} \).

Now, suppose \( I(\mathbf{g}, \mathbf{n}, i, j, k) = 1 \). Then, the configuration \( \mathbf{n} \) would arise if a mutation happened on the \( i \)th, \( j \)th, and \( k \)th branches of the \( s_1 \), \( s_2 \), and \( s_3 \) trees, but no mutations occurred on any of the other branches. The probability of this is

\[
I(\mathbf{g}, \mathbf{n}, i, j, k) (1 - e^{-\theta}) (1 - e^{-\theta}) e^{-\theta(\tau_2 - \theta)}.
\]

Thus, to obtain \( \Pr(\mathbf{n} | \theta, \rho_{GC}, f, \ell) \), we sum over all possible trios of branches (from the three trees) and take the expectation over random genealogies (assuming the standard equilibrium neutral model):

\[
\Pr(\mathbf{n} | \theta, \rho_{GC}, f, \ell) = E \left( \sum_{i,j,k} I(\mathbf{g}, \mathbf{n}, i, j, k) (1 - e^{-\theta}) (1 - e^{-\theta}) \times e^{-\theta(\tau_2 - \theta)} e^{-\theta(\tau_2 - \theta)} e^{-\theta(\tau_2 - \theta)} \right)
\]

\[
= E \left( \sum_{i,j,k} I(\mathbf{g}, \mathbf{n}, i, j, k) a_b c_j \right).
\]

Here \( i, j, \) and \( k \) index over the branches of the \( s_1 \), \( s_2 \), and \( s_3 \) trees, respectively, and the approximation is derived by expanding the exponentials and ignoring higher-order \( \theta \) terms. As \( \theta \to 0 \), we consider the scaled likelihood function

\[
h_a(n, \rho_{GC}, f, \ell) = \lim_{\theta \to 0} \Pr(\mathbf{n} | \theta, \rho_{GC}, f, \ell) / \theta^5
\]

\[
= E \left( \sum_{i,j,k} I(\mathbf{g}, \mathbf{n}, i, j, k) a_b c_j \right).
\]

If we run \( y \) different replicates, \( h_a \) can be estimated by

\[
h_a(n, \rho_{GC}, f, \ell) = \frac{1}{y} \sum_{y=1} I(\mathbf{g}_y, \mathbf{n}, i, j, k) a_b c_j(h),
\]

where \( \mathbf{g}_y \) is the three-site genealogy for the \( y \)th replicate, \( a_b(h) \) is the length of the \( b \)th branch of the tree for \( s_1 \) for the \( h \)th replicate, and so on. One benefit of this simulation scheme is that the scaled likelihood of all possible configurations \( \mathbf{n} \) can be estimated simultaneously: for fixed \( \mathbf{g}, i, j, \) and \( k, I(\mathbf{g}, \mathbf{n}, i, j, k) \) equals 1 for exactly one configuration and must equal 0 for all others.

Our main interest is in the probabilities of sample configurations conditional on there being two alleles at each site. In the limit as \( \theta \to 0 \), the conditional probability of the configuration \( \mathbf{n} \) is

\[
\lim_{\theta \to 0} \Pr(3\text{-site configuration}|2 \text{ alleles at each site}) = \frac{\Pr(3\text{-site configuration and 2 alleles at each site})}{\Pr(2 \text{ alleles at each site})}
\]

\[
= \frac{h_a(n, \rho_{GC}, f, \ell)}{\sum_{\mathbf{m}} h_a(m, \rho_{GC}, f, \ell)},
\]

where the summation in the denominator is over all possible sample configurations \( \mathbf{m} \). Analogous to Equation 1, the denominator in (2) can be estimated from simulation:

\[
\sum_{\mathbf{h}} h_a(m, \rho_{GC}, f, \ell) = \frac{1}{y} \sum_{y=1} I(\mathbf{g}_y, \mathbf{m}, i, j, k) a_b(h) b_j(h) c_j(h)
\]

\[
= \frac{1}{y} \sum_{y=1} I(\mathbf{g}_y, \mathbf{m}, i, j, k) a_b(h) b_j(h) c_j(h)
\]

\[
= \frac{1}{y} \sum_{y=1} I(\mathbf{g}_y, \mathbf{m}, i, j, k) a_b(h) b_j(h) c_j(h)
\]

Here \( \tau_i(h) \) is the sum of the branch lengths at site \( s_i \) for the \( h \)th simulation, etc. Note that the conditional
Estimating recombination rates for multiple loci: For joint estimates of crossing-over and values. We take as a point estimate the parameter combination that produces the maximum value of clik(D; pco, f, t). Call this \( \hat{p}_{\text{joint}} \).

If \( n > 10 \) we adopt a subsampling approach. We choose a large number of subsets of 10 sequences. Denote \( V = \{v_1, v_2, \ldots, v_\ell\} \) as the collection of subsets. For each subset we consider all triplets of segregating sites. Then, we calculate a composite likelihood by multiplying together the likelihoods of all of the triplets in each of the subsets,

\[
\text{clik}(D; p_{\text{co}}, f, t) = \prod_{\mathbf{n} \in \mathcal{S}} \text{Pr}(\mathbf{n}; p_{\text{co}}, f, t),
\]

where \( \mathbf{n} \) is the configuration formed by sites \( s_1, s_2, \) and \( s_3 \). We take as a point estimate the parameter combination that produces the maximum value of \( \text{clik}(D; p_{\text{co}}, f, t) \).

For each triplet, we estimate its likelihood over a grid of recombination values \((p_{\text{co}}, f, t)\). To estimate these likelihoods, we use linear interpolation over the calculated likelihoods (described above) whose parameter values \((e.g., p_{\text{co}}, d_{\text{po}}, d_{\text{po}} f, t)\) are closest to the actual ones. For our simulation study (described below), we fix \( t \) at a plausible value (either 125 or 500 bp; see Hilliker et al. 1994; Frisse et al. 2001; Jeffreys and May 2004) and consider only a grid of \( p_{\text{co}} \) and \( f \) values. We decided to calculate likelihoods over a grid of values, rather than use various iterative procedures (cf. Press et al. 1992), because we did not know \textit{a priori} whether the composite-likelihood surfaces would have a single local maximum.

Estimating recombination rates for multiple loci: Suppose we have multiple unlinked loci and we wish to jointly estimate recombination parameters from these data. If the relevant parameters are assumed to be constant across loci, then we can simply multiply the composite-likelihood functions at each individual locus to form a composite likelihood for the combined data. If \( D = \{D_1, D_2, \ldots, D_N \} \) is a collection of loci, we set

\[
\text{clik}(D; p_{\text{co}}, f, t) = \prod_{D \in D} \text{clik}(D; p_{\text{co}}, f, t).
\]

Here we utilize the fact that unlinked loci are evolutionarily independent. We consider the parameters that maximize \( \text{clik}(D; p_{\text{co}}, f, t) \) as a point estimate. Call these \( \hat{p}_{\text{co}} \), \( \hat{f} \), and \( \hat{t} \), and denote the maximum composite likelihood as \( \text{MCL}(D; p_{\text{co}}, f, t) \). We call this method the “joint” method and note that it is analogous to the approach of Frisse et al. (2001).

We also implement an alternative approach that is analogous to Prak et al. (2004). We form a different composite likelihood assuming that \( f \) and \( t \) are fixed across loci (\textit{e.g.}, at \( f_k \) and \( t_l \)), but allowing \( p_{\text{co}} \) to vary.
plik(\textit{D}; \hat{f}_c, t_0) = \prod_{i \in D} \text{clik}(D_i; \rho_{\text{co}}, f_0, t_0). \tag{4}

Thus, for each locus, we take the maximum (composite) likelihood, constrained so that \( f = \hat{f}_c \) and \( t = t_0 \). For our simulation study, we first fix \( t_0 \) at either 125 or 500 bp, then maximize \text{plik} as a function of \( f \). We call the value of \( f \) that maximizes (4) \( \hat{f}_c \). Following Ptak et al. (2004), we call this method of estimating \( f \) the “profile” method, due to its similarity to profile likelihood.

Extensions: We note in passing that it should be straightforward to extend our methods to handle diploid (i.e., unphased) data, data without ancestral-derived information, ascertainment bias, and/or missing data. For example, given unphased data from three sites, we would first enumerate all possible (haploid) haplotype configurations that are consistent with the diploid data, and then calculate the probability that each haplotype configuration would produce the observed diploid configuration (assuming random assortment of gametes). The likelihood of the diploid configuration \( \mathbf{M} \) is then

\[ \sum_{\mathbf{n}} \text{Pr}(\mathbf{n}; \rho_{\text{co}}, f, t) \text{Pr}(\mathbf{M}|\mathbf{n}). \]

Here the summation is over all haplotype configurations \( \mathbf{n} \) that are consistent with the diploid data.

Simulations: To test the accuracy of our method for estimating crossing-over rates (only), we run coalescent simulations (cf. Hudson 1983) with known recombination rates and then tabulate the distributions of estimated recombination rates. These simulations have \( n = 50, \theta = 0.001/\text{bp}, \rho = 0.001/\text{bp} \) or 0.004/\text{bp}, and 2, 4, 6, . . . , 16 kb of simulated sequence. A total of 1000 replicates were run for each parameter combination. For each replicate, we consider 100 subsets of size 20 and all triplets of three segregating sites from each of these subsets. These simulations are directly comparable to those in Figures 7 and 8 of Hudson (2001) and Figure 5 of Li and Stephens (2003).

We take a similar approach for evaluating how accurately our method can estimate \( f \). We simulate 50,000 loci of length 5 kb, each with \( n = 50 \) and \( \theta = \rho_{\text{co}} = 0.001/\text{bp} \). One set of simulations had \( f = 1 \) and a mean conversion tract length of \( t = 500 \) bp while the other had \( f = 4 \) and \( t = 125 \) bp. (Gene conversion tract lengths followed a geometric distribution.) For each locus, we calculate composite likelihoods (using 150 subsets of 10 sequences) over a grid of \( \rho_{\text{co}} \) \{0, 0.0002, 0.0004, . . . , 0.004\} and \( f \{0, 0.35, 0.5, 0.7, 1, 1.4, 2, 2.8, 4\} \) when \( f = 1 \) and \( \{0, 1, 1.4, 2, 2.8, 4, 5.6, 8, 11.2, 16\} \) when \( f = 4 \) values. We consider each locus individually, as well as groups of 2, 5, 10, 20, 50, 100, 200, and 500 loci. (Due to computational constraints, the same 50,000 loci are used for each analysis.) For each collection of loci, we calculate \( \hat{f}_c, \hat{f}_{1\text{r}}, \) and \( \hat{f}_{2\text{r}} \), using the two methods described above, in (3) and (4). For comparison, we also estimated \( f \) using the methods of Frisse et al. (2001) and Ptak et al. (2004). We call these pairwise composite-likelihood estimators \( \hat{f}_1 \) and \( \hat{f}_2 \), respectively. To compare the accuracies of the different estimators of \( f \) we log transform all values and calculate the root mean square error [i.e., the square root of the average squared difference between \( \log(\text{estimated } f) \) and \( \log(\text{actual } f) \)]. Estimates of \( f \) that are 0 are changed to 0.25 times the true value (i.e., 0.25 or 1 depending on the simulations) to ensure that \log \( f \) is well defined.

Following Ptak et al. (2004), we also examine the effects of recombination rate variation (among loci) on estimates of \( f \). We generate 50,000 loci of length 5 kb with \( n = 50, \theta = 0.001/\text{bp}, \rho_{\text{co}} = \text{gamma}(10, 10^{-4}) \) and either \( f = 1 \) with \( t = 500 \) bp or \( f = 4 \) with \( t = 125 \) bp. Tract lengths are geometrically distributed and the loci are analyzed in the same way as described above. The mean value of \( \rho_{\text{co}} \) in these simulations is 0.001/\text{bp} (as before), and the gamma distribution parameters were chosen to approximately match the simulations described in Ptak et al. (2004).

We are also interested in determining how often one could reject a model of crossing over only (i.e., \( f = 0 \)). For each collection of loci with \( \theta = \rho_{\text{co}} = 0.001/\text{bp} \) and \( f = 1 \) or 4, we evaluate

\[
\frac{\text{MCL}(\text{D}; \rho_{\text{co}}, f = 0)}{\text{MCL}(\text{D}; \rho_{\text{co}}, f = 0, 0)}. \tag{5}
\]

Equation 5 is analogous to the standard likelihood-ratio statistic; we reject the hypothesis that \( f = 0 \) when (5) is too large. To determine the appropriate critical values, we run null simulations of 5-kb loci with \( n = 50, \theta = 0.001/\text{bp}, \rho_{\text{co}} = 0.00168/\text{bp} \) (or 0.0018/\text{bp}), and \( f = 0 \). The value of \( \rho_{\text{co}} \) was chosen to produce the same mean \( \rho_{\text{co}} \), estimate as the \( \rho_{\text{co}} = 0.001/\text{bp} \), \( f = 1 \) (or \( f = 4 \)) simulations when we set \( f = 0 \). We generate 50,000 null loci and evaluate loci either singly or in groups of 2, 5, 10, 20, 50, 100, 200, or 500 loci as above. For each one, we take the 95th percentile of the composite-likelihood-ratio distribution as the critical value when analyzing the simulations with gene conversion.

The simulations described in this section took several months of computing time on a pair of 2.4 GHz Xeon processors. All programs used in the analyses were written in C and are available upon request.

RESULTS

Crossing over alone: First we examined how accurately our method can estimate recombination rates under a simplified model with no gene conversion. We ran simulations with known values of \( \rho \) and then tabulated the distributions of estimated values. Figure 1 shows the 10th and 90th percentiles of the distributions of \( \hat{\rho}_{\text{co}} \), with analogous results from Hudson’s (2001) \( \hat{\rho}_{\text{co}} \) shown for comparison. The distributions of the two estimators are quite similar, although the percentiles of the \( \hat{\rho}_{\text{co}} \) distributions tend to be slightly lower than the
Figure 1.—Estimates of the 10th and 90th percentiles of the distributions of \( \hat{\rho}_{CL} \) (cf. Hudson 2001) and \( \hat{\rho}_{W04} \). (A) \( \theta = \rho \); (B) \( \theta = \rho/4 \). Each point is based on 1000 replicates. See methods for further details.

corresponding ones for \( \hat{\rho}_{CL} \). The root mean square error for \( \hat{\rho}_{W04} \) is generally slightly smaller than that for \( \hat{\rho}_{CL} \), but both are larger than the root mean square error of the distribution of \( \hat{\rho}_{ML-R} \) (Li and Stephens 2003, Figure 5). A priori we expected that an estimate based on the likelihoods of triplets of sites would be more accurate than an estimate based on the likelihoods of pairs of sites, but that the difference would be mitigated by having to consider subsets of 20 sequences when calculating \( \hat{\rho}_{W04} \) but using all sequences when calculating \( \hat{\rho}_{CL} \). If we compare the two estimators on simulated data with \( n = 20 \), then the difference in accuracy between \( \hat{\rho}_{W04} \) and \( \hat{\rho}_{CL} \) is more substantial (results not shown). We note that it should be possible to modify the methods of McVean et al. (2002) to consider triplets of sites and that such a method may outperform \( \hat{\rho}_{W04} \).

**Crossing over and gene conversion:** Next, we ran simulations of multiple loci, each with the same crossing-over and gene conversion rates, and estimated \( \rho \) using both the joint (\( \hat{f}_{T1} \)) and profile (\( \hat{f}_{P1} \)) methods using triplets of sites. For comparison, we also calculated \( \hat{f}_{T2} \) and \( \hat{f}_{P2} \), the pairwise versions of the joint and profile methods (cf. Frisse et al. 2001; Ptak et al. 2004). Roughly speaking, the joint method assumes that \( \rho \) (per base pair) is constant across loci, while the profile method does not. (For both \( \hat{f}_{T1} \) and \( \hat{f}_{P1} \) we jointly estimate \( f \) and \( \hat{\rho} \) for each collection of loci but present results only for \( \hat{\rho} \).) Figures 2 and 3 show the distribution of estimated \( f \) values for the four different methods: Figure 2 uses 1, 5, 20, and 100 independent loci, respectively, with \( f = 1 \) and \( t = 500 \) bp while Figure 3 uses 1, 5, 20, and 100 independent loci, respectively, with \( f = 4 \) and \( t = 125 \) bp. It is quite striking how poorly all four methods perform when there are few loci. With 1 or 5 loci, none of the methods perform much better than random guessing (among the 9 or 10 possible values of \( f \)), while even with 20 loci there is an \( \sim 50\% \) probability of being off by at least a factor of two. The only scenario where the actual value of \( f \) (\( f = 1 \) in Figure 2 and \( f = 4 \) in Figure 3) was estimated at appreciable frequency (e.g., more than half of all replicates) was with \( \hat{f}_{T1} \) for sets of 100 loci. We note in passing that the distributions for \( \hat{f}_{T1} \) and \( \hat{f}_{P1} \) are substantially more ragged than the corresponding distributions for the pairwise estimators. Presumably this reflects in part the small sample sizes of the subsets and the roughness of the grid over which likelihoods were calculated.

To compare the accuracy of the four methods, we tabulated the root mean square (RMS) error of the logs of the estimated values. These are shown in Figure 4. In general, for the same number of loci, \( \hat{f}_{T1} \) has lower RMS error than the other methods. The true value of \( \rho \) is constant across loci in the simulations, so we might expect the joint methods (which explicitly assume this rate constancy) to perform better than the profile methods. While this is clearly true for the triplet composite-
likelihood methods, the two pairwise composite likelihoods perform comparably. As did Ptak et al. (2004), we find that the profile methods are biased downward; in contrast, \( \hat{f}_1 \) is biased upward slightly, while \( \hat{f}_1 \) is nearly unbiased.

The \( f = 1 \) and \( f = 4 \) simulations are meant to be approximately comparable since the probability (per generation) that a given nucleotide site is covered by a gene conversion tract is the same. Figures 2–4 show that, at least for the parameter values used in this study, estimation (on a log scale) is more accurate when the actual gene conversion rate is higher. Limited laboratory data suggest that \( f \approx 4 \) in both humans (Zangenberg et al. 1995; Jeffreys and May 2004) and Drosophila melanogaster (Hilliker and Chovnick 1981).

Thus far we have concentrated on the problem of estimating \( f \). Ideally, we would want to jointly estimate \( \rho \) and \( f \). In general this is difficult to do, since high values of \( \rho \) and low values of \( f \) are hard to distinguish from lower values of \( \rho \) and higher values of \( f \) (Ptak et al. 2004; results not shown). We tabulate for both joint methods the probability that both \( \rho \) and \( f \) can be estimated “reasonably” well: the probability that both are closer than a factor of 2 away from their true values (i.e., \( 0.5 < \hat{f} < 2 \) or \( 2 < \hat{f} < 8 \) and \( 0.0005 < \rho < 0.002 \)).

Figure 5 plots this as a function of the number of loci. As with the earlier results, the three-site joint method is substantially better than the pairwise joint method, and estimation is more accurate for the \( f = 4 \) simulations compared with the \( f = 1 \) simulations. Both methods perform poorly overall, and \( \sim 10–20 \) loci are needed to have a 50% chance of estimating both \( \rho \) and \( f \) reasonably well. In contrast, under the simpler crossing-over-only model of recombination, only a single locus is necessary for estimating \( \rho \) to the same level of accuracy.

**Rate variation simulations**: The simulations in the previous section assumed that recombination rates did not vary across loci. While this might be a reasonable assumption for some species (e.g., \( D. \) simulans, cf. Charlesworth 1996; Wall et al. 2002), it is demonstrably false for other species, most notably humans (Yu et al. 2001; Kong et al. 2002). In humans, estimates of recombination rates from comparisons of physical and genetic maps have limited accuracy due to the small number of meioses used in the genetic maps (Weber 2002; Ptak et al. 2004). To account for local recombination rate variation, we simulated loci with the same mean recombination rate, but with rates for each locus chosen
Figure 3.—Estimates of $f$ using the three-site composite-likelihood joint ($\hat{f}_{T1}$) and profile ($\hat{f}_{T2}$) methods and the two-site composite-likelihood joint ($\hat{f}_{P1}$) and profile ($\hat{f}_{P2}$) methods. (A–D) Results for 1, 5, 20, and 100 independent loci, respectively. The recombination parameters are the same for all loci ($p_o = 0.001/bp$, $f = 4$, and $t = 125$ bp). See Methods for further details.

from a gamma distribution. We then performed the same analyses as before, assuming either that rates were constant across loci (joint method) or that they varied (profile method). Figure 6 shows the RMS error for the four different methods. A comparison with Figure 4 shows that all methods are less accurate when the actual data have recombination rate variation. This is not surprising, since the variation adds an extra degree of uncertainty to the calculations. Further simulations confirm that as rate variation becomes more extensive the estimators become even less accurate (results not shown). As with the previous simulations, all methods perform poorly when there are few loci, $\hat{f}_{T1}$ is more accurate than the other methods, and the $f = 4$ simulations have lower RMS error than the $f = 1$ simulations. Also, as before, the only scenario where the true value of $f$ was estimated at appreciable frequency (e.g., >50% of replicates) was using $\hat{f}_{T1}$ and sets of 100 loci.

In contrast to the fixed recombination rate simulations, we might expect the profile methods to be more accurate than the joint methods in Figure 6, since the former correctly assume that there is variation in crossing-over rates across loci, while the latter erroneously assume that $p_o$ is constant across loci. However, this does not turn out to be the case. $\hat{f}_{P1}$ and $\hat{f}_{P2}$ show approximately equal bias and RMS error (on log-transformed data), with $\hat{f}_{T2}$ biased downward and $\hat{f}_{T1}$ biased upward. For the three-site likelihoods, $\hat{f}_{T1}$ has a smaller bias and RMS error than $\hat{f}_{T2}$.

We also examined how well the two joint methods perform at jointly estimating $p_o$ and $f$. The probability that both parameters can be estimated reasonably well (see previous section for details) is slightly lower for loci with (crossing-over) rate variation than in Figure 5 for loci with no rate variation (results not shown).

Rejecting $f = 0$: The gene conversion results presented above examine the question of how well we can estimate $p_o$ and $f$. Here we address a simpler question and look at how often we can reject a model of no gene conversion (i.e., $f = 0$). To do this, we use the equivalent of a likelihood-ratio statistic, with critical values determined from (null) simulations without gene conversion (see Methods for details). We tabulate the proportion of simulations (with gene conversion) where $f = 0$ can be rejected at the 5% level. Figure 7 shows these proportions as a function of the number of independent loci considered, for the joint and profile methods, using either pairs or triplets of sites. Figure 7A shows the results when $f = 1$ and $t = 500$ bp while Figure 7B shows the results when $f = 4$ and $t = 125$ bp. The number of
loci needed to have appreciable power to reject the no gene conversion model is substantial; in Figure 7A >10 loci are necessary to have 50% power (5 or more loci in Figure 7B), while ~50 loci are required to achieve >80% power (~10 loci in Figure 7B). Since estimators of \( f \) are more precise for the \( f = 4 \) simulations (see Figures 4 and 5), it is not surprising that the probability of rejecting \( f = 0 \) is higher in Figure 7B than in Figure 7A. However, it is somewhat surprising that for \( \geq 20 \) loci the pairwise methods have greater power to reject the null model than the three-site methods, even though \( \hat{f}_{T1} \) is the most accurate and most precise of the four estimators (see Figures 2–6).

One would expect that the more loci (with gene conversion) one has, the easier it should be to reject a model of no gene conversion. Figure 5 shows that this is generally true, although there is an unexpected dip in power for the joint method (using pairs of sites) when there are \( \geq 100 \) loci. We stringently checked our code and reran our simulations to verify that this odd result was not an error. The dip in power is due to a sharp increase in the critical values for sets of \( >50 \) loci. Slightly <1 locus out of every 1000 simulated with no gene conversion is extremely unusual, with a composite likelihood (based on pairs of sites) that strongly supports \( f > 0 \). Equivalently, the distribution of (5) for each locus taken individually is highly skewed, with a long tail to the right (results not shown). These unusual loci in the tail dominate when calculating (5); since \( >5\% \) of groups of 100 loci contain one of these unusual loci, the estimated critical value is quite large, and the power is low. Similarly, since \( <5\% \) of groups of 50 loci contain one
**DISCUSSION**

Three-site likelihoods have the potential to be much more informative about recombination rates (and in particular gene conversion rates) than two-site likelihoods. In our implementation, this is counteracted by the decreased efficiency of having to consider subsets of 10 sequences, rather than the full sample size. Nonetheless, overall, Figures 2–6 show that using three-site likelihoods (rather than existing two-site likelihood methodology) leads to a substantial improvement in accuracy when estimating gene conversion rates. Since the relevant likelihood files need to be calculated only once, this method provides an efficient and widely applicable method for estimating gene conversion rates. (All software and likelihood files are available online at http://www.cmb.usc.edu/~jeffwall).

Despite the improvement over previous methods, our three-site composite-likelihood approach does not perform very well on an absolute scale. For the parameter values considered here, accurate estimates of crossing-over and gene conversion rates require data from ≥10 independent loci, whereas the same level of accuracy in estimating crossing-over rates alone (assuming no gene conversion) requires just a single locus. While the poor performance in Figures 2–6 reflects in part the limitations of our methods, the main difference (compared with the results in Figure 1) lies in the added

---

**Figure 6.** Root mean square (RMS) error of log-transformed estimates of $f$ as a function of the number of loci. (A) Results for simulations with $f = 1$ and $t = 500$ bp; (B) results for simulations with $f = 4$ and $t = 125$ bp. For each locus, $\rho_{co}$ is chosen from a gamma distribution with mean 0.001/bp. See METHODS for further details.
difficulty in estimating multiple recombination rates instead of just a single one. We fear that future methods will yield only incremental improvements because the information content of sequence polymorphism data is so poor for the question at hand.

Since accurate estimates of gene conversion (and crossing-over) rates require polymorphism data from multiple, evolutionarily independent regions, we must make explicit assumptions regarding the extent to which recombination rates vary across different regions. The most restrictive assumption, namely that rates are constant across loci (i.e., the joint method), may not be generally appropriate for analyzing real data (but see below). If the goal is to estimate $f$, then variation in crossing-over rates can be accounted for mostly by using the profile method. However, both the joint and profile methods assume that $f$ is constant across loci. Essentially nothing is known about whether $f$ varies across the genomes of higher eukaryotes, but several authors have suggested that $f$ may be higher in regions of reduced crossing over in *D. melanogaster* (Langley et al. 2000; Andolfatto and Wall 2003). In principle, we could test this hypothesis by jointly estimating $f$ and $p_w$ for individual regions and determining whether estimates of $f$ vary in a systematic way. We caution though that the results of such an analysis would be difficult to interpret, since the rate estimates for any particular region are likely to be inaccurate. Nonetheless, it may be possible to determine whether estimates of $f$ are higher at some set of loci compared to others, even if the actual values cannot be estimated accurately.

Another approach for dealing with rate heterogeneity across loci is to apply the joint or profile methods to the data anyway, viewing the estimates obtained as “averages” across many loci. Some systematic biases may be associated with such an approach, such as the upward bias to $f_1$ when the actual data have rate variation. In our simulations, the joint methods show equal or less bias than the profile methods regardless of whether or not there is variation in recombination rates in the underlying data (Figures 2–4; results not shown). We also ran simulations with more extreme rate variation (gamma(4, 0.00025)-distributed rates for $p_w$) but still found that $f_{12}$ performed as well as or better than $f_{11}$ when the underlying data had variation in recombination rates (Ptak et al. 2004). The reason for this discrepancy between the two studies is unclear, but the performance of different estimators (of $f$) may be sensitive to the particular parameter values used in the simulations (in particular, to the precise extent of recombination rate variation).

It should also be noted that nothing is yet known about the effect of variation in $f$ (across loci) on the various methods for estimating $f$ discussed here. Further simulation studies will address this question.

**Figure 7.**—Probability of rejecting $f = 0$ when data are simulated with gene conversion, as a function of the number of loci considered. Results are shown for the three-site and two-site joint and profile methods. (A) Results when $f = 1$ and $t = 500$ bp; (B) results when $f = 4$ and $t = 125$ bp. See Methods for further details.
Finally, we emphasize that our simulation study assumed we have perfect data quality. Real data have sequencing or genotyping errors, and these errors can artificially inflate estimated gene conversion rates (Prat et al. 2004). Recurrent mutations (i.e., multiple mutations at the same nucleotide site) can also have the same effect. Gene conversion can be thought of as a process where small stretches of sequence (i.e., conversion tracts) look different because they were inherited from a different ancestral chromosome. Similarly, sequencing errors and recurrent mutations can be thought of as processes where small stretches of sequence (i.e., single SNPs) look different because of experimental error or mutation. The extent to which these processes “look” different (on the basis of patterns of variation and LD) depends in part on the density of single-nucleotide polymorphisms (SNPs) and the mean gene conversion tract length. In species where the typical gene conversion tract might contain no more than one SNP, we expect that it will be extremely difficult to distinguish between the effects of sequencing errors, recurrent mutation, and gene conversion on the basis of patterns of LD alone. However, in most cases sequencing error rates can be estimated directly, given sufficient time and motivation. Under simple models of the error process, recombination rates could then be estimated assuming a fixed rate of sequencing errors. In addition, in species with higher SNP densities and/or longer gene conversion tract lengths, it may be possible to distinguish between the effects of gene conversion and sequencing errors (or even jointly estimate both rates) using a generalization of the methods described in this article.

We thank R. Hudson and M. Przeworski for helpful discussions and M. Przeworski and M. Stephens for comments on an earlier version of this manuscript. This work was supported in part by NIH grant no. 1P50HG027900-01 to M. Waterman.

LITERATURE CITED
Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein and F. W. Stahl,


Communicating editor: D. Charlesworth