A coalescent-based method for detecting and estimating recombination from gene sequences

Gil McVean¹, Philip Awadalla² and Paul Fearnhead³

¹Department of Statistics, University of Oxford, U.K.
²Section of Evolution and Ecology, University of California, Davis, U.S.A.

December 11, 2001
Running title: Detecting recombination

Keywords: Recombination
  Recombination
  Coalescent
  Homoplasy
  Mitochondria
  HIV

Corresponding author: Gilean McVean
  Department of Statistics
  1 South Parks Road
  Oxford OX1 3TG
  UK
  Tel: +44 1865 281881
  Fax: +44 1865 272595
  email: mcvean@stats.ox.ac.uk
Abstract

Determining the amount of recombination in the genealogical history of a sample of genes is important to both evolutionary biology and medical population genetics. However, recurrent mutation can produce patterns of genetic diversity similar to those generated by recombination, and can bias estimates of the population recombination rate. Hudson (2001) has suggested an approximate likelihood method based on coalescent theory to estimate the population recombination rate, $4N_e r$, under an infinite-sites model of sequence evolution. Here we extend the method to the estimation of the recombination rate in genomes, such as those of many viruses and bacteria, where the rate of recurrent mutation is high. In addition, we develop a powerful permutation-based method for detecting recombination that is both more powerful than other permutation-based methods, and is robust to mis-specification of the model of sequence evolution. We apply the method to sequence data from viruses, bacteria, and human mitochondrial DNA. The extremely high levels of recombination detected in both HIV1 and HIV2 sequences demonstrates that recombination cannot be ignored in the analysis of viral population genetic data.
1 Introduction

Recombination breaks down the correlation in genealogical history between different regions of a genome and shuffles genetic diversity among chromosomes. In evolutionary biology, the importance of recombination is the generation of novel gene combinations that enables the spread of multiple beneficial mutations (Muller, 1932; Fisher, 1932), and prevents the accumulation of deleterious ones (Muller, 1964). In medical genetics, associations between disease phenotypes and genetic markers that build up through genetic drift and are broken down by recombination are central to the mapping of disease-associated mutations (Pritchard and Przeworski, 2001).

The occurrence of recombination also has practical implications for evolutionary inference. For population geneticists, recombination reduces the effects of evolutionary stochasticity, averaging out genealogical histories over a genome. In contrast, traditional methods of phylogenetic inference typically assume the absence of recombination. If the assumption is incorrect, inferences about the evolutionary history of gene sequences may be misleading (SCHIERUP and HEIN, 2000). Recombination is therefore a critical issue for analyses of within species variation.

A variety of non-parametric methods have been developed to detect recombination from gene sequences, without estimating the rate at which it occurs. Some use phylogenetic methods to ask whether different regions of a gene have dif-
different histories (Grassly and Holmes, 1997; McGuire et al., 2000), and are targeted at identifying rare recombinant genotypes. Other methods are aimed at inferring the presence of recurrent recombination, such as occurs among the genes of most eukaryote species. Among these methods, some consider summary statistics that are sensitive to recombination, such as the relationship between physical distance and measures, or indicators of linkage disequilibrium (Lewontin, 1964; Maynard Smith, 1999). Other methods consider properties of phylogenetic trees inferred under the assumption of no recombination (Maynard Smith and Smith, 1998; Worobey, 2001). The methods vary in their ability to detect recombination under different conditions, and their sensitivity to an accurate characterisation of the underlying model of sequence evolution (Maynard Smith, 1999; Meunier and Eyre-Walker, 2001).

The inability of such methods to estimate the rate at which recombination occurs is a serious limitation. Characterising the rate of recombination is important for analysing the power of association studies, assessing the reliability of phylogenetic methods and predicting the rate at which advantageous mutations, such as those conferring drug resistance, can spread between genetic backgrounds. Some non-parametric methods for detecting recombination, such as the homoplasy test (Maynard Smith and Smith, 1998) and derivatives (Worobey, 2001), provide a characterisation of how far the data are from the extremes of free recombination, and complete clonality. But there is no straightforward relationship between such a property and the parameters of any underlying evolutionary model.
As a result, comparison between genes or species is problematic, and there is little or no way of statistically testing whether data sets have different levels of recombination. Model-based estimation of the rate of recombination does rely on an underlying model that is almost certainly a simplification of reality. However, the benefits gained are the ease of comparison between different data sets, the ability to make predictions about the question of interest and the potential to test whether the model of evolution is an adequate characterisation of the underlying processes. In addition, parametric models can be used to test for the presence of recombination by comparing the likelihood of the data under models with and without recombination (BROWN et al., 2001).

What evolutionary model is appropriate for describing the effects of recombination on gene sequences? Coalescent theory provides a statistical description of the genealogical history of sequences sampled from large, Fisher-Wright populations with non-overlapping generations, constant population size, and no selection or migration (KINGMAN, 1982; HUDSON, 1991). Within this framework, the effects of recombination on sample history are a function not of the absolute recombination rate, but of the product of the per gene per generation rate of crossing-over (genetic map length), $r$, and the effective population size, $N_e$ (GRIFFITHS and MARJORAM, 1996b). Without prior information about one of these parameters, it is only possible to estimate the product of these parameters, often written as $\rho = 4N_e r$ (equivalently, one can estimate the ratio of the recombination rate and the mutation rate, $r/\mu$ and the population mutation rate $\theta = 4N_e \mu$).
lescent can readily be extended to include time-varying population size, migration and some forms of selection (Hudson and Kaplan, 1994; Braverman et al., 1995). Under these more complex situations, the effects of recombination on gene samples also depend on other parameters. In general, however, the product of the current effective population size of the population and the absolute recombination rate is the key determinant of the impact of recombination on patterns of genetic diversity.

Within the framework of the coalescent, several methods have been proposed as estimators of the population recombination rate. Hudson (1987) derived a moment estimator based on the variance in pairwise differences. Hey and Wakeley (1997) developed a method based on combining analytically derived likelihoods for all pairs of sites and sets of four sequences. Wall (2000) proposed to find the value of $4N_e r$ that maximises the likelihood of observing the number of haplotypes and inferred minimum number of recombination events (Hudson and Kaplan, 1985). More recently, full likelihood estimators of the population recombination rate, based on the coalescent, have been developed. These use computationally intensive Monte Carlo methods; Griffiths and Marjoram (1996a) described a method based on importance sampling while Kuhner et al. (2000) developed a Metropolis-Hastings rejection Monte Carlo Markov Chain (MCMC) method. Recently, Fearnhead and Donnelly (2001) have improved the importance sampling method considerably. Even so, full-likelihood methods are computationally intensive, and practically impossible for many data sets.
Recently, Hudson (2001) has suggested an *ad hoc* method for estimating the population recombination rate based on combining the coalescent likelihoods of all pairwise comparisons of segregating sites. Estimation of $4N_e r$ is rapid, and the method performs well in terms of bias and variance in comparison to Hudson’s earlier moment estimator (Hudson, 1987) and other *ad hoc* approaches (Hudson, 2001). The method does not use all available information in the sequence data, and introduces non-independence in the combination of multiple comparisons, but is flexible and can potentially be expanded to incorporate deviations from the standard coalescent. Hudson’s (2001) estimator of $4N_e r$ has been termed the composite likelihood estimate, or CLE.

In this paper we consider a problem of critical importance to the analysis of recombination; the detection and estimation of recombination in genomes, such as those of many viruses and bacteria, where the rate of substitution is sufficiently high that some sites have experienced multiple mutations in the history of the sample. The issue is important because recurrent mutation can generate patterns of genetic variability that resemble the effects of recombination (Figure 1); in particular, the presence of all four haplotypes for a pair of segregating sites. Under the infinite-sites model, any such incompatibilities would be interpreted as evidence for recombination, and hence will bias estimates of the recombination rate upwards. Similarly, the likelihood ratio test for the presence of recombination will be sensitive to mis-specification of the mutation model, particularly the presence of unacknowledged rate heterogeneity.
To address these problems we have extended Hudson’s composite likelihood method (Hudson, 2001) to allow for finite-sites mutation models. In addition, we propose a permutation-based test (the likelihood permutation test) to detect the presence of recombination. We find that Hudson’s approach performs well even when most sites analysed have experienced multiple mutations, and that the likelihood permutation test is more powerful than previous permutation-based methods for detecting recombination. We also consider the effect of mis-specification of the model of sequence evolution on both the test for recombination, and estimation of $4N_e r$. We show that the likelihood permutation test is robust to mis-specification, unlike the homoplasy test (Maynard Smith and Smith, 1998) or the informative sites test (Worobey, 2001), and that estimation of $4N_e r$ is also robust to minor mis-specification of the model of sequence evolution. We apply the likelihood permutation test and estimation procedure to several empirical data sets from viruses, bacteria and human mitochondria.

2 Methods

2.1 Composite likelihood estimation of $4N_e r$

First, we wish to outline our implementation of the approach of Hudson (2001) for estimating the population recombination rate under the standard Fisher-Wright population model. The central difference between the method of Hudson (2001)
Figure 1: Recurrent mutation (A) and recombination (B) can generate similar patterns of genetic variability. The upper half of the picture shows the genealogies and occurrence of mutations, while the lower half depicts the resulting sampled gene sequences.

and that presented here is that we allow for models of sequence evolution in which multiple mutations may occur at a site during the history of the sample. Although it is possible to use an arbitrary model of sequence evolution, we make the simplifying assumption that all sites in a sequence conform to a two-allele model with reversible, symmetric mutation, such that the rate of mutation per site per generation is $\mu$, and is constant across sites. Consequently, we restrict analysis to sites at which there are no more than two alleles segregating. The extension of the method to more complex models of sequence evolution is left to future research, however it is worth noting that the method appears to perform well, even when the true model of sequence evolution is considerably more complex than that assumed (see below).
The estimation procedure has four stages. The initial step is to estimate the population mutation rate per site, $\theta = 4N_c\mu$, from an approximate finite-sites version of the Watterson estimate

$$\hat{\theta}_W = \left( \sum_{k=1}^{n-1} \frac{1}{k} \right)^{-1} \ln \left( \frac{L}{L - S} \right)$$  \hspace{1cm} (1)$$

where $S$ is the number of segregating sites, $L$ is the total length of sequence analysed, and $n$ is the number of sampled gene sequences. The second stage is to consider every pair of segregating sites in the data (excluding sites with more than two alleles) and classify them into equivalent sets. For example, under the assumed mutation model, if one pair had the ordered data $\{AA, AT, TA, TA, AA\}$ and another $\{GG, CC, CG, GG, CG\}$, these are equivalent to the unordered sequence $\{00, 00, 10, 10, 01\}$ where 0 represents the rare allele at each site. The number of types (hence the execution time of the program) depends on the number of sequences, the level of diversity, and the complexity of the assumed mutation model.

The third stage is to estimate the likelihood of each equivalent set under the estimated value of $\theta$, the symmetric, reversible mutation model, and a range of recombination rates (typically $0 \leq 4N_c r \leq 100$), using the importance sampling method of Fearnhead et al. (FEARNHEAD and DONELLY, 2001). We have also used a simple Monte Carlo scheme for estimating the likelihood, similar to that implemented in (HUDSON, 2001), to check the accuracy of likelihoods estimated by the importance sampling method (results not shown).
In the final stage, an estimate of the population recombination rate for the entire sequence \(4N_e r\) is obtained by combining the likelihoods from all pairwise comparisons. The composite likelihood is given by

\[
\ell_C(4N_e r) = \sum_{i,j} \ell(X_{ij}|4N_e r_{ij})
\]  

(2)

where \(\ell(X_{ij}|4N_e r_{ij})\) is the log likelihood of the data for segregating sites \(i\) and \(j\) given

\[
r_{ij} = \frac{r d_{ij}}{L - 1}
\]

(3)

where \(d_{ij}\) is the physical distance (in nucleotides) separating sites \(i\) and \(j\) and \(L\) is the total length of the sequence (i.e. we assume a constant rate of recombination over the gene). The estimate of \(4N_e r\) is taken as the value that has the highest composite log likelihood.

For genomes, such as viruses and bacteria, in which a gene-conversion model for recombination is more appropriate than a crossing-over model, the relationship between physical distance and recombination rate is modelled as

\[
r_{ij} = 2c\bar{T}(1 - e^{-d_{ij}/\bar{T}})
\]

(4)

where \(c\) is the per base rate of initiation of gene conversion and \(\bar{T}\) is the average gene conversion tract length (assuming an exponential distribution) (FRISSE et al., 2001). This type of model can also be applied to circular genomes, such as that of the mitochondria, where \(d_{ij}\) is the minimum distance between two points on the circle (WIUF, 2001). While it is possible to coestimate both the rate of gene
conversion and the average tract length, in practice we fix the average tract length and estimate the compound parameter

$$\gamma = 8N_e c \tilde{l}$$ (5)

which can be thought of as the population rate of recombination between two distantly linked loci caused by gene conversion.

For simple data sets, and low values of $4N_e r$, it is possible to compare the composite likelihood surface with the full likelihood surface estimated by the method of Fearnhead and Donnelly (2001). Figure 2 shows a comparison of the two surfaces for a single case, and the joint distribution of the maximum likelihood (MLE) and composite likelihood (CLE) point estimates of $4N_e r$ for 100 simulated data sets with $n = 50$ and $\theta = 4N_e r = 3$. For the single example (Figure 2A), the composite likelihood curve has a very similar point estimate to the ML estimate, but is more highly curved because of the nonindependence introduced by multiple comparisons. Statistics for the two estimators (full likelihood/composite likelihood) are: median 2.4/3.8, variance 9.1/15.6, proportion within a factor of two from the true value 0.50/0.52. The correlation between the composite and maximum likelihood estimates is 0.78 (Figure 2B).

Hudson (2001) has characterised the composite likelihood estimator for the case where data conform to the infinite-sites model. In terms of bias and variance, the CL estimator is one of the better ad hoc methods for estimating the population recombination rate, although the estimator has considerable variance. However,
Figure 2: A) The composite (CLR) and full (LR) relative likelihood surface for a single simulated data set. B) The joint distribution of the maximum likelihood estimate (MLE) of $4N_e r$ and the composite likelihood estimate (CLE). Likelihoods were calculated with $\theta = 0.01$ per site.
this is also true of the MLE (Figure 2), and, to a large extent, is a reflection of inherent stochasticity in the genealogical process. However, while full likelihood provides an estimate of the relative likelihood of different values, there is no easily interpretable meaning of the composite likelihood curve. Confidence intervals for the estimate of $4N_e r$ can only be obtained by extensive simulation (HUDSON, 2001).

2.2 The likelihood permutation test

We propose a simple test for the presence of recombination. Under a model of no recombination, and assuming a uniform mutation rate, sites are exchangeable (this is also true if there is free recombination). That is, the likelihood of observing the data is independent of the order in which sites occur. If there is some recombination, sites are no longer exchangeable, because closely linked sites have correlated genealogies. Consequently, the likelihood of observing the data is dependent on the order of sites. The likelihood permutation test for recombination is based on this property; we find the maximum composite likelihood for a data set (estimating $4N_e r$ in the process), then permute segregating sites by location and for each permutation find the maximum composite likelihood (and the corresponding value of $4N_e r$). The proportion of permuted data sets with a composite likelihood equal to or greater than that of the original data is calculated. If this proportion is lower than a chosen significance level, we conclude that there is evidence for
recombination.

There are several methods for detecting recombination based on the permuta-
tion of segregating sites. Permutation tests for recombination aimed at detecting a
decay of a summary statistic of linkage disequilibrium ($r^2$ or $|D'|$) with distance
have been used to suggest the presence of recombination in human mitochon-
dria (AWADALLA et al., 2000), Plasmodium falciparum (CONWAY et al., 1999),
and regions of low recombination in the D. melanogaster genome (MIYASHITA
and LANGLEY, 1988). Another permutation test (referred to as $G4$) has been
suggested by Meunier and Eyre-Walker (2001), which compares the sum of dis-
tances between all pairs of sites that have all four possible haplotypes, with the
distribution in permuted data sets. We have compared the power of the likelihood
permutation test with these other permutation-based tests.

### 2.3 Models of sequence evolution

We wish to characterise both the composite likelihood estimator and likelihood
permutation test under a range of models of sequence evolution that reflect genomes
experiencing high mutation rates at some, or all sites. We have chosen four cari-
cature models to represent the diversity of possible situations:

- **Infinite sites:** all sites have the same, low mutation rate ($\theta = 0.01$) and
  conform to the two-allele symmetric, reversible mutation model used in the
  likelihood estimation stage. This represents the best-case scenario (effec-
tively infinite-sites), as might be assumed for nuclear loci in humans (ex-
cluding hypermutable CpG dinucleotides).

- **Hypermutable:** most sites (99.5%) effectively conform to the infinite-sites
  model ($\theta = 0.005$), but a fraction (0.5%) have a 100-fold higher mutation
  rate. All sites conform to the symmetric, reversible mutation model. This is
  chosen to reflect extreme rate variation, as occurs when hypermutable CpG
  dinucleotides are included in an analysis, or in the mitochondrial genome
  of mammals.

- **Complex:** characterised by strong base composition variation and mutation
  rate variation. Specifically, an HKY mutation model Hasegawa et al.
  (1985), with base frequencies $\pi_T = 0.4, \pi_C = 0.1, \pi_A = 0.4, \pi_G = 0.1$, a
  transition-transversion ratio of 2 and an exponential distribution of mutation
  rates with a base-averaged mutation rate of $\tilde{\theta} = 0.1$, where

  \[ \tilde{\theta} = 4N_e \sum_i \pi_i \sum_{j \neq i} \tilde{\mu}_{ij} \]  

  and $\tilde{\mu}_{ij}$ is the average per generation mutation rate from base $i$ to base
  $j$ (from the exponential distribution). This model is chosen to reflect the
  complexity of sequence evolution in prokaryote genomes with strong base
  composition bias.

- **Finite sites:** all sites have the same, high mutation rate ($\theta = 0.5$) and con-
  form to the two-allele symmetric, reversible mutation model. In this case,
each segregating site experiences, on average, 2.6 mutations in the history of the sample. This model represents the extreme levels of polymorphism as occur at synonymous sites in retroviruses such as HIV.

Data are simulated under the null $4N_c r = 0$ and $4N_c r = 10$, for $n = 50$ and the length of sequence chosen such that the average number of segregating sites is in the range 40-50. Ideally, for each simulated data set the likelihoods should be calculated for the value of $\theta$ estimated from the data. However, for the large number of replicates required to provide an accurate characterisation of the estimator’s properties, calculating the likelihoods for each data set is practically unfeasible. Instead, we have estimated likelihoods under three different values of $\theta$: 0.01, 0.1 and 0.5, and present the results for each along with mean and standard deviation of the values of $\theta$ estimated from the simulated data. One advantage of this approach is that it allows us to characterise the severity of model mis-specification on the detection and estimation of recombination.

2.4 Empirical data

We have applied both the likelihood permutation test and estimation of the population recombination rate to a series of empirical data sets from viruses, bacteria and human mtDNA. Previous analyses (Worobey, 2001; Worobey et al., 1999; Suerbaum et al., 1998) of these data sets has revealed a range of levels of recombination, from effectively clonal in HCV (Worobey, 2001) to freely
recombining in *H. pylori* (Suerbaum et al., 1998). While none of these data sets represent random samples from Fisher-Wright populations, as is supposed by the coalescent methods of analysis, the results are likely to be indicative of the situation in more appropriate samples.


**Bacterial genomes.** *Helicobacter pylori*: 33 sequences of the flaA gene (worldwide) (Suerbaum et al., 1998).

**Mitochondrial genomes.** 45 partial genome sequences from the analysis of Awadalla et al. (worldwide) (1999), 53 complete genome sequences from the analysis of Ingman et al. (2000).
3 Results

3.1 Estimating $4N_e r$ with recurrent mutation

To date, estimators of the population recombination rate have typically been characterised under the infinite-sites assumption; that each segregating site is the result of a single mutation. In many biologically realistic situations this assumption can not be justified, even though the infinite-sites model is superficially plausible. For example, if 20 mutations occur in a genealogy of 500 linked sites (the expected number for $n = 50$ and $\theta = 0.009$), the probability that at least one site experiences recurrent mutation is over 30% and will be higher if there is recombination, or any variation between sites in the mutation rate. In organisms with high mutation rates, such as many viruses and bacteria, a large proportion of sites may have experienced multiple mutations.

Because recurrent mutation can create patterns of genetic variability that resemble the effects of recombination (Figure 1), it is important to develop methods for estimating the recombination rate that can account for finite-sites models of sequence evolution. We have extended Hudson’s (2001) composite likelihood method for estimating the population recombination rate, $4N_e r$, within a coalescent framework, to incorporate models in which sites may experience multiple mutations in the history of the sample. Our approach is to use the simplest possible model of finite-sites evolution (two-allele system with symmetric reversible mutation, and a constant mutation rate across sites) and to investigate how the
method performs under a variety of caricature models of sequence evolution chosen to reflect biological diversity.

Figure 3 shows the distribution of point estimates for $4N_c r$ for data simulated under the four caricature models ($n = 50$ and $4N_c r = 10$) and likelihoods estimated under three different values of $\theta$; 0.01, 0.1 and 0.5. In Table 1 we also present the median and proportion of estimates that are with a factor of two from the true value, along with the mean and standard deviation of estimates of $\theta$ obtained from Equation (1).

As expected, when there is a considerable discrepancy between the true value of $\theta$ and that used to estimate likelihoods, estimates of $4N_c r$ are strongly biased. When the true value of $\theta$ is lower than the value used to estimate likelihoods, estimates of $4N_c r$ are downwardly biased. In contrast, when the true value of $\theta$ is greater than the value used to estimate likelihoods, estimates of $4N_c r$ are biased upwards. However, it is encouraging to find that when likelihoods are estimated under the correct value of $\theta$, the estimator performs almost as well when the mutation rate is very high, as it does when the mutation rate is low (bottom right versus top left panels).

The middle two rows of Figure 3 and Table 1 show the effects of applying the simplistic mutation model to data simulated under models representing some degree of biological complexity. For both the hypermutable and complex models there is strong rate variation across sites, yet the estimator properties are hardly worse than under the best-case scenario, and the estimates of $\theta$ are well within the
Figure 3: The distribution of CLEs of the population recombination rate under different models of sequence evolution. Each chart represents the results from 1000 data sets simulated with $4N_c r = 10$. The model of sequence evolution used to simulate data is on the left and the value of $\theta$ used to calculate likelihoods under two-allele symmetric reversible model is at the top of the column.

range that leads to sensible estimates of $4N_c r$. In short, the composite likelihood estimator of the population recombination rate is robust to minor mis-specification of the underlying mutation model. This conclusion is of great importance as it provides a justification of the use of the CL estimator on real data sets.
<table>
<thead>
<tr>
<th>Mutation model</th>
<th>$\theta$ = 0.01</th>
<th>$\theta$ = 0.1</th>
<th>$\theta$ = 0.5</th>
<th>$\bar{\theta}_w(\pm sd)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>infinite sites: $\theta$ = 0.01</td>
<td>9.4 (0.77)</td>
<td>6.6 (0.68)</td>
<td>0.0 (0.08)</td>
<td>0.010 (0.002)</td>
</tr>
<tr>
<td>hypermutable: $\tilde{\theta}$ = 0.008</td>
<td>12.2 (0.83)</td>
<td>8.6 (0.82)</td>
<td>1.2 (0.13)</td>
<td>0.006 (0.001)</td>
</tr>
<tr>
<td>complex: $\tilde{\theta}$ = 0.1</td>
<td>11.8 (0.80)</td>
<td>8.4 (0.79)</td>
<td>1.2 (0.15)</td>
<td>0.073 (0.015)</td>
</tr>
<tr>
<td>finite sites: $\theta$ = 0.5</td>
<td>32.6 (0.13)</td>
<td>24.0 (0.34)</td>
<td>9.8 (0.71)</td>
<td>0.337 (0.078)</td>
</tr>
</tbody>
</table>

Table 1: Median of estimates of $4N_e r$ (proportion of estimates within a factor of two of $4N_e r = 10$) for the data presented graphically in Figure 3. The last column is the mean and standard deviation of estimates of $\theta$ obtained from applying Equation 1 to the simulated data.

### 3.2 Detecting recombination

The results presented above may give us some confidence that the value of $4N_e r$ estimated by the composite likelihood method is meaningful, even in genomes where the rate of recurrent mutation is high. However, one important question that is difficult to address within the CLE framework is whether one can reject the hypothesis that $4N_e r = 0$. Direct experimental evidence for recombination may be difficult to obtain for many genomes (particularly if genetic exchange is very rare), thus it is important to have indirect, population genetic based methods for detecting recombination. And it is equally important that such methods should not create false-positives through mis-specification of the model of sequence evolution.
We have proposed the likelihood permutation test as a means of testing for the presence of recombination. Table 2 shows the results of the power analysis carried out on the same four caricatures of sequence evolution, and again estimating likelihoods under the three values of $\theta$. We also compare the power of the likelihood permutation test with other permutation-based tests for recombination that consider summaries of the data sensitive to the presence of recombination.

The key result is that the likelihood permutation test is consistently the most powerful permutation-based method for detecting recombination from population genetic data. In the case of infinite-sites data, recombination is detected in almost 96% of cases, compared to approximately 80% for the other tests. Even when the model used to estimate likelihoods is very different from the true model, the power of the test is considerable. For example, with data generated by the finite-sites model with $\theta = 0.5$, recombination is detected in 83% of cases when the correct value of $\theta$ is used to calculate likelihoods, compared to 82% of cases when $\theta = 0.01$ is used to estimate likelihoods. In contrast, those methods that rely heavily on the distribution of pairs at which all four gametes are present ($|D'|$ and $G4$) have greatly reduced power under such high levels of mutation (51% and 39% respectively). The one situation where the likelihood permutation test has reduced power is when the true value of $\theta$ is much lower than that used to estimate likelihoods, however, such a situation is unlikely to occur for empirical data. It is also worth noting that the power to detect recombination using the correlation between $r^2$ and physical distance is consistently greater than with either $|D'|$ and
4 Discussion and application

The composite likelihood method and likelihood permutation test together present a powerful approach for assessing the influence of recombination on patterns of genetic variability. Even when the mutational and substitutional processes affecting gene sequence evolution are complex, and unlikely to be fully characterised by any simple model, the use of simple models provides a remarkably robust way of detecting recombination and estimating the population recombination rate. To investigate how the new approach performs on real data, we have applied the methods to samples of gene sequences from the viruses HIV1, HIV2, Hepatitis C, Dengue-1, and measles, the bacterium Helicobacter pylori and human mitochondrial DNA sequence. We also discuss possible limitations of the approach, in particular mis-specification of the population model used to estimate the likelihoods.

4.1 Empirical data

The empirical data sets were chosen to reflect a diversity of levels of recombination, as had been estimated from previous studies (Maynard Smith et al., 1993; Worobey et al., 1999; Worobey, 2001; Suerbaum et al., 1998). For the HIV data sets, we analysed 3rd position sites in the coding region separately.
## Power analysis of permutation tests for detecting recombination

| Mutation model   | $4N_e r$ | $Lk_{\theta=0.01}$ | $Lk_{\theta=0.1}$ | $Lk_{\theta=0.5}$ | $r^2$ | $|D'|$ | G4 |
|------------------|---------|---------------------|-------------------|-------------------|------|-------|----|
| infinite sites   | 0       | 0.053               | 0.057             | 0.058             | 0.046| 0.018 | 0.019 |
| hypermutable     | 0       | 0.025               | 0.045             | 0.055             | 0.019| 0.030 | 0.015 |
| complex          | 0       | 0.048               | 0.060             | 0.061             | 0.038| 0.027 | 0.031 |
| finite sites     | 0       | 0.041               | 0.052             | 0.053             | 0.049| 0.051 | 0.046 |
| infinite sites   | 10      | 0.958               | 0.931             | 0.447             | 0.783| 0.797 | 0.796 |
| hypermutable     | 10      | 0.969               | 0.957             | 0.560             | 0.856| 0.740 | 0.717 |
| complex          | 10      | 0.969               | 0.958             | 0.890             | 0.838| 0.790 | 0.767 |
| finite sites     | 10      | 0.824               | 0.849             | 0.834             | 0.712| 0.514 | 0.393 |

Table 2: 1000 data sets were simulated for each set of mutation models and combination of parameter values.
from the first two positions, in order to investigate whether different results were obtained from using data with different levels of diversity. In addition, we analysed two human mtDNA data sets that have been used to provide evidence for (Awadalla et al., 1999) and against (Ingman et al., 2000) recombination. In all cases, a gene-conversion model for recombination is more appropriate than a crossing-over model, and we have fixed the average tract length of gene conversion to 100 bp for the viral and bacterial data sets, and 500 bp for the mtDNA data sets. These numbers are arbitrary, although in the microbial and viral data sets, the composite likelihood increases for small tract lengths (data not shown). In one of the few cases in eukaryotes where gene conversion tract lengths have been estimated, the best fit to the data was a geometric distribution with mean tract length of 352 bp (Hilliker et al., 1994).

Table 3 presents the results of these analyses, and the estimate of the population recombination rate, $\gamma$, under a gene conversion type model; see Equation (5). In addition, we carried out the same analyses, but filtering out SNPs for which the minor allele was at a frequency less than 0.1; the results are presented in Table 4. For the HCV and Dengue virus data sets the results from the filtered analysis are identical to those in Table 2 as the sample sizes are less than 10. We have also ommitted the results for the test of Meunier and Eyre-Walker (2001) as it behaves in an almost identical fashion to $|D'|$.

From Table 3 and, more noticeably, from Table 4, we find evidence for recombination in almost all data sets, and levels of recombination that range from
### Detecting recombination in empirical data

| Genome     | Gene\(^1\) | \(L\) | \(n\) | \(\hat{\theta}_W\) | \(D^2\) | \(P_{Lk}\) | \(P_{\gamma}\) | \(P|D|\) | \(\gamma^3\) | Ref                  |
|------------|------------|-------|-------|-----------------|--------|----------|----------|---------|------------|----------------------|
| HCV        | CC         | 8922  | 6     | 0.325           | 0.496  | 0.142    | 0.011    | 0.957   | 0.84       | WOROBHEY (2001)       |
| measles    | \(H\)     | 1830  | 50    | 0.089           | -2.061 | 0.170    | 0.005    | 0.573   | 3.0        | WOELK et al. (2001)   |
| \(H. pylori\) | flaA | 471   | 33    | 0.045           | -0.531 | 0.000    | 0.146    | 0.547   | 41         | SUERBAUM et al. (1998) |
| HIV2       | \(env_{12}\) | 1364  | 21    | 0.102           | -1.136 | 0.000    | 0.036    | 0.002   | 43         | KUIKEN et al. (2000)  |
| HIV2       | \(env_3\) | 682   | 21    | 0.302           | -0.457 | 0.250    | 0.042    | 0.320   | 26         | KUIKEN et al. (2000)  |
| Dengue     | \(C,prM/M,E\) | 2322  | 7     | 0.053           | 0.316  | 0.000    | 0.000    | 0.000   | 60         | WOROBHEY et al. (1999) |
| HIV1B      | \(env_{12}\) | 1316  | 93    | 0.144           | -2.185 | 0.512    | 0.418    | 0.194   | > 100      | KUIKEN et al. (2000)  |
| HIV1B      | \(env_3\) | 658   | 93    | 0.333           | -1.878 | 0.479    | 0.393    | 0.130   | > 100      | KUIKEN et al. (2000)  |
| Human mtDNA | \(CG\) | 8458  | 45    | 0.0056          | -2.473 | 0.000    | 0.000    | 0.014   | 9.0        | AWADALLA et al. (1999) |
| Human mtDNA | \(CG\) | 16581 | 52    | 0.0071          | -2.233 | 0.102    | 0.241    | 0.502   | 2.6        | INGMAN et al. (2000)  |

\(^1\) CC: Complete coding sequence; \(CG\): Complete genome; subscripts indicate positions in coding sequences.

\(^2\) Tajima \(D\) values calculated for segregating sites with only two alleles segregating.

\(^3\) \(\gamma = 8N_c\ell\) from Equation (5).

Table 3: Tests for recombination in empirical data sets. Estimates of \(\theta\) and \(\gamma\) are given per base. \(P\) values less than 0.05 are in bold.
## Detecting recombination with mutations at intermediate frequencies

| Genome          | Gene   | S  | $P_{Lk}$ | $P_{r^2}$ | $P_{|D'|}$ | $\hat{\gamma}$ | Ref                  |
|-----------------|--------|----|----------|-----------|------------|----------------|----------------------|
| measles         | $H$    | 59 | 0.067    | 0.048     | 0.002      | 3.0            | Woelk et al. (2001)  |
| *H. pylori*     | flaA   | 30 | 0.000    | 0.000     | 0.000      | 44             | Suerbaum et al. (1998) |
| HIV2            | env$_{12}$ | 97 | 0.000    | 0.059     | 0.242      | $>100$         | Kuiken et al. (2000) |
| HIV2            | env$_3$ | 183 | 0.018    | 0.016     | 0.037      | 36             | Kuiken et al. (2000) |
| HIV1B           | env$_{12}$ | 36 | 0.020    | 0.083     | 0.435      | $>100$         | Kuiken et al. (2000) |
| HIV1B           | env$_3$ | 36 | 0.018    | 0.713     | 0.773      | $>100$         | Kuiken et al. (2000) |
| *H. sapiens* mtDNA | CG    | 12 | 0.197    | 0.006     | 0.442      | 15             | Awadalla et al. (1999) |
| *H. sapiens* mtDNA | CG    | 49 | 0.720    | 0.802     | 0.769      | 1.0            | Ingman et al. (2000) |

Table 4: Tests for recombination in empirical data sets only using SNPs with a minor allele frequency of at least 0.1. Sample details are as for Table 3. *P* values less than 0.05 are in bold.
$\hat{\gamma} = 0.84$ in HCV to $\hat{\gamma} > 100$ in HIV1 ($\gamma = 100$ was chosen as a cut-off as it is the limit for which likelihoods were estimated). In HCV, only the correlation of $r^2$ with distance shows a significant negative relationship, but with six sequences, there is little power in the likelihood permutation test. For the measles data set, only $r^2$ is significant when all data is used, but all tests are either significant, or marginally significant, for the filtered data. The other data sets show evidence for much higher levels of recombination. The estimate of $\gamma$ is over 40 for *H. pylori* and 60 for Dengue. The ratio $\hat{\gamma}/\hat{\theta}_W$ gives an indication of the relative likelihood of a nucleotide experiencing a recombination event relative to mutation. Within the data sets for which there is strong support for recombination, the ratio varies from about 35 in measles to about 1000 in Dengue and *H. pylori* and potentially much higher in HIV1.

The effect of filtering out rare variants is worth noting. Rare variants are largely uninformative about recombination (though not entirely; McVean, 2001), hence their inclusion may obscure the signal of recombination, particularly if there is an excess of rare mutations in the data. Removal of rare variants from the data has little effect on estimates of the population recombination rate both in the empirical data (compare estimates of $\gamma$ from Tables 3 and 4) and in the simulated data. For example, under the finite-sites model, the median of estimates of $\gamma$ was 9.8 when all sites were used (and analysed under the correct mutation model) and 10.2 when the analysis was restricted to sites for which the minor allele frequency was at least 0.1 In the simulated data, no increase in the power of the likelihood
permutation test was found when the analysis was restricted to intermediate frequency variants. However, the simulated data sets have no excess of rare variants, unlike the empirical data.

4.2 Very high levels of recombination in HIV

The results concerning recombination in HIV1 subtype B and HIV2 subtype A sequences are particularly notable. Although recombination between different subtypes is occasionally observed (Kuiken et al., 2000), recombination within subtypes has largely been ignored in phylogenetic analysis of genetic diversity (Nielsen and Yang, 1998; Rambaut et al., 2001). The results presented here support such a conclusion. Using the likelihood permutation test, we find evidence for recombination in both HIV2 and HIV1, though only when SNPs are filtered for the case of HIV1. For HIV1 the estimate of $\gamma$ is beyond the range for which likelihoods were estimated.

Levels of genetic diversity are extremely high in HIV1 and HIV2 (estimates of $\theta$ per site at first/second codon positions of 0.144 and 0.102 respectively). Because recurrent mutation can cause patterns of genetic diversity similar to that caused by recombination, one might be cautious of concluding that recombination is present. However, the estimation of a low level of recombination in HCV which has an even higher level of diversity ($\hat{\theta}_W = 0.325$), and in measles, which has a comparable level of sequence diversity ($\hat{\theta}_W = 0.089$), indicates that high
levels of sequence diversity does not necessarily lead to high estimates of the pop-
ulation recombination rate.

The implications of such a high level of recombination in HIV1 are consid-
erable. Not only does it question the validity of conclusions about the age and
timings of events in the history of the virus that have been made assuming an ab-
sence of recombination (Rambaut et al., 2001; Nielsen and Yang, 1998), but it has practical implications for predicting how fast mutations (such as drug resis-
tance) may spread across different genetic backgrounds. Analysis of genetic data
from appropriate samples taken at different population scales will be essential for
inferring the extent and consequences of recombination.

4.3 Recombination in human mtDNA?

Another issue of considerable importance is whether there is evidence for recom-
bination in human mtDNA. The data set of Awadalla et al. (1999) clearly shows
evidence for recombination when all data are used, irrespective of the test em-
ployed (for \( r^2 \) and the likelihood permutation test this is also true for over 90% of
random subsets of 35 of the 45 sequences). In direct contrast, the data of Ingman
et al. (2000) show no evidence for recombination, irrespective of the test used.
When the frequency filter is applied, only one statistic, \( r^2 \), still shows evidence for
recombination in the first data set (and this is sensitive to the removal of a single
segregating site). These results are in direct contrast to those from the viral and
bacterial sequences, where the frequency filter increases the power of almost all tests. Taken together, the results suggest a lack of evidence for recombination in human mtDNA.

Why should low frequency variants create the impression of recombination? Hey (2000) has suggested that sequencing protocols might lead to the propagation of correlated errors. Such an effect may be enhanced by the combination of sequences from multiple laboratories (because recurrent errors will be strongly correlated), and for this reason, the data collected and sequenced by Ingman et al. is preferable. Given that sequencing errors tend to be at low frequency, this may explain why three of the four tests are only significant if all the data is analysed, but it does not explain (beyond chance) why $r^2$ still shows a significant relationship with distance when only high frequency variants are used. McVean (2001) has suggested that bouts of local adaptive evolution might lead to correlated mutations and a relationship between physical distance and linkage disequilibrium as measured by $r^2$. How adaptive evolution influences patterns of linkage disequilibrium and the measurement and detection of recombination is an important problem.

4.4 Mis-specification of the population model

While the properties of the composite likelihood estimator of the population recombination rate have been examined across a variety of models of sequence evo-
lution, no mention has been made so far as to how robust the methods described here may be to deviations from the population model. Coalescent estimation of likelihoods assumes that a random sample has been taken from a population of constant size, with random mating, no migration to or from different populations, and no natural selection. In reality, none of these assumptions are tenable, although several deviations from the standard neutral model (such as fluctuating population size) can be approximated as having an effect on the effective population size, $N_e$.

Population growth, strong geographical structuring and non-random representation of gene sequences in the databases are potentially important concerns for the use of coalescent methods. Sampling of sequences specifically for population genetic analysis will overcome the problems of non-random database representation, however inadequacies in the demographic model are more problematic. Population growth tends to decrease linkage disequilibrium while population structure tends to increase linkage disequilibrium (e.g. Pritchard and Przeworski, 2001). Consequently, one might expect estimates of the population recombination rate (and the ability to detect recombination) to be sensitive to the demographic history of the population.

While no exhaustive attempt will be made here to characterise the behaviour of the CL estimator under mis-specified population models, it is possible to ask whether the data sets analysed show evidence for deviation from neutral model in terms of the allele frequency spectrum. This can most simply be assessed
through the use of Tajima’s $D$ statistic, which compares estimates of the population mutation rate derived from the number of segregating sites, and the average pairwise differences. A negative value of the statistic indicates an excess of rare variants and the possibility of population growth, a positive value suggests population structure may be important.

Table 3 includes the value of Tajima’s $D$ statistic for the data sets analysed, and indicates the significance level estimated assuming no recombination. While the statistic is negative for all data sets, it is only significantly so for measles, HIV1 and the two mtDNA data sets. However, the variance of the statistic is reduced by recombination (so reducing the confidence limits under the null model). Other data sets (particularly the HIV2 data) may therefore also reflect significant deviations from the standard neutral model. However, those data sets that show evidence for a departure from the standard neutral model also reflect the full diversity of estimated recombination rates. In short, while departure from the assumed demographic model may have some influence on the estimate of the population recombination rate, it is unlikely to be confused with the signal of recombination.

**Acknowledgements** We would like to thank Michael Worobey for the generous supply of empirical data sets and important insights. In addition, we thank Dick
Hudson, Molly Przeworski and two reviewers for discussion and comments on the manuscript. GM is funded by the Royal Society and PA is funded by the Wellcome trust.

**Software** The programs *pairwise* and *permute* used to estimate the population recombination rate and test for recombination are available within the *LDhat* package, which can be downloaded from http://www.stats.ox.ac.uk/~mcvean.

**References**


Maynard Smith, J., N. H. Smith, M. O’Rourke, and B. G. Spratt. 1993

McGuire, G., F. Wright, and M. J. Prentice. 2000 A Bayesian model for
detecting past recombination in DNA multiple alignments. J. Comput. Biol. 7:
159–170.

McVean, G. A. T. 2001 What do patterns of genetic variability reveal about
mitochondrial recombination? Heredity in press.

Meunier, J. and A. Eyre-Walker. 2001 The correlation between linkage
disequilibrium and distance. Implications for recombination in Hominid mito-

Miyashita, N. and C. H. Langley. 1988 Molecular and phenotypic variation


Res 1: 2–9.

Nielsen, R. and Z. Yang. 1998 Likelihood models for detecting positively se-
lected amino acid sites and applications to the HIV-1 envelope gene. Genetics


