Estimating meiotic gene conversion rates from population genetic data

Gay, J.*, Myers, S.† and McVean, G.*

*Department of Statistics, University of Oxford, Oxford, United Kingdom
†The Broad Institute of MIT and Harvard, Cambridge, MA

July 16, 2007
Running Head: Estimating gene conversion rates

Key Words: Gene conversion, recombination, coalescent, PAC likelihood

Corresponding Author:

Gil McVean

Department of Statistics
University of Oxford
1 South Parks Road
Oxford OX1 3TG
Tel: +44 1865 272860 Fax: +44 1865 272595
mcvean@stats.ox.ac.uk
Abstract

Gene conversion plays an important part in shaping genetic diversity in populations, yet estimating the rate at which it occurs is difficult because of the short lengths of DNA involved. We have developed a new statistical approach to estimating gene conversion rates from genetic variation, by extending an existing model for haplotype data in the presence of crossover events. We show, by simulation, that when the rate of gene conversion events is at least comparable to the rate of crossover events, the method provides a powerful approach to the detection of gene conversion and estimation of its rate. Application of the method to data from the telomeric X chromosome of Drosophila melanogaster, in which crossover activity is suppressed, indicates that gene conversion occurs approximately 400 times more often than crossover events. We also extend the method to estimating variable crossover and gene conversion rates and estimate the rate of gene conversion to be approximately 1.5 times higher than the crossover rate in a region of human chromosome 1 with known recombination hotspots.
INTRODUCTION

An important concept in the description of genetic variation is linkage disequilibrium (LD), the non-random association of alleles at different locations along the genome. Disease association studies rely heavily on knowledge of patterns of LD, both in pinpointing complex disease genes precisely, and in performing genome-wide studies (Pritchard and Przeworski, 2001). Over long ranges, LD is mainly affected by crossover, which has been studied and modelled by many authors and is reviewed by Stumpf and McVean (2003).

A less well known form of recombination is homologous gene conversion, a non-reciprocal process acting on short lengths of DNA, where genetic material from one parental chromosome is incorporated into the alternate chromosome during meiotic exchange (Szostak et al., 1983). Crossover events in fact include a gene conversion tract but this cannot be detected using population-based methods, and we will therefore use the term gene conversion only to refer to gene conversion events which are not accompanied by crossover.

In humans, gene conversion is thought to occur approximately 4-15 times as frequently as crossover (Jeffreys and May, 2004), but is more difficult to detect due to the short lengths of DNA transferred - estimates of the tract length vary between studies and between organisms/regions but tend to lie between 50 and 2000 bp (e.g. Borts and Haber, 1989; Hilliker et al., 1994; Jeffreys and May, 2004). For a full description of the gene conversion process see Stahl (1994) and references therein.

Patterns of linkage disequilibrium in humans can only be satisfactorily explained by models including gene conversion (Frisse et al., 2001). Simulations show that in genomic regions which have been subject to gene conversion, estimates of the crossover rate are inflated when gene conversion is ignored (Smith and Fearnhead, 2005). Przeworski and Wall (2001) showed, using human population genetic data, that gene conversion is likely to be an important factor in explaining a marked difference between estimates of the population recombination rate obtained through comparing genetic and physical maps, and those found through analysis of nucleotide sequence polymorphism data. These factors have made gene conversion the subject of much investigation in recent years.
Although highly localised, the effects of gene conversion may also have a significant impact on association studies, which seek a genotyped marker which is in strong LD with an untyped allele responsible for the phenotype of interest. If gene conversion is ignored, the extent of LD over short distances is likely to be overestimated, while LD at longer distances will be underestimated due to the inflated rate of crossover needed to explain the short-range LD (Frisse et al., 2001). These two influences on LD may affect the choice of the number of markers to genotype for a study (Schork, 2002).

Gene conversion also affects our ability to detect the effects of natural selection on a population (Andolfatto and Nordborg, 1998). Tests for deviation from the null model typically rely on an estimate of the recombination rate in a region, and ignoring the effects of gene conversion will reduce the power of tests for selection and can also increase the false positive rate of such tests.

Finally, learning about gene conversion could help us to gain biological and mechanistic insights into recombination.

It is therefore desirable to be able to estimate the frequency at which gene conversion events occur, at a fine scale, over genomic regions many megabases in length, and to detect variation in gene conversion rates within such a genomic region.

**Approaches to estimation of gene conversion rates:** Rates of both crossover and gene conversion can be estimated directly using sperm-typing experiments such as those of Jeffreys and May (2004) which give highly accurate fine-scale rates, but cannot be performed on a genomic scale or on the X chromosome, or in females.

Pedigree studies (e.g. Kong et al., 2002) can give further information such as sex-specific differences in crossover rates, but because of the infrequency of events, cannot give accurate fine-scale maps.

For genome-wide fine-scale characterisation of recombination rates a practical solution is statistical modelling of genetic data, based on simplified assumptions about the historical
processes which resulted in the population genetic data seen today. Methods of inference can be performed in many different ways:

- Summary statistics (e.g. Wiehe et al., 2000; Padhukasahasram et al., 2006) can sometimes be quick to calculate but only make use of partial information and are not able to detect fine-scale variation.

- Composite likelihoods calculated using pairs or triplets of segregating sites (e.g. Ptak et al., 2004; Frisse et al., 2001) can provide a ‘reasonable’ estimate of the gene conversion rate (i.e. within a factor of two of the truth) given sufficient data (Wall, 2004). Fearnhead et al. (2005) applied one such method (Hudson, 2001) to bacterial datasets and obtained some interesting results, including tract length estimates. However, for densely typed SNP data where there are likely to be high levels of LD, composite likelihood methods may be unsuitable as they ignore the dependency between nearby pairs/triplets of SNPs.

- Full likelihood methods approximate the probability of the data under the assumed population genetic model (exact probabilities are not available due to the unknown history of the sample). Some use techniques such as importance sampling (Fearnhead and Donnelly, 2001) to make the approximation, whilst others use a simplified model under which exact probabilities can be found (e.g. Li and Stephens, 2003; Helenthal, 2006). The main benefit of the full likelihood approach is to make use of as much of the information in the data as possible, and in the case of gene conversion we expect this to be important.

In this article we describe a statistical model of population genetic data which includes both crossover and gene conversion, where a gene conversion tract can include any number of markers. The model can be used to estimate the rates of crossover and/or gene conversion in a given region using maximum likelihood techniques, or could be implemented in a bayesian framework. The model does not require that either rate be constant across the region of interest and could for example be used to obtain an estimate of the gene conversion rate in a
region known to include a crossover hotspot. As well as performing tests on simulated data, we examine single nucleotide polymorphism (SNP) data from a genomic region thought to be free from crossover hotspots, and then consider a region of the human genome which contains several crossover hotspots (JEFFREYS et al., 2005).

Our results on simulated data show that gene conversion rates can be estimated fairly accurately from population genetic data, and the inclusion of gene conversion in our model results in improved estimates of the crossover rate, particularly when gene conversion is present at high levels. In a region near the telomere of the X chromosome of D. melanogaster we find that gene conversion events occur over 400 times as frequently as crossovers, whilst in a region of human chromosome 1, there is only 1.5 times as much gene conversion activity as crossover.

MODEL

Our model is an extension of the coalescent-based model of LI and STEPHENS (2003) (henceforth abbreviated to LS) to include gene conversion as well as crossover. LS modelled the probability of seeing a particular chromosomal segment, given any other homologous segments already seen, and given the rates of mutation $\theta$ and crossover $\rho = 4N_e c$, where $N_e$ is the effective population size and $c$ is the per-generation probability of crossover between adjacent base pairs. We will use the terms haplotype and chromosome interchangeably to refer to a chromosomal segment, and assume the method will be applied to resequenced or densely genotyped SNP data, although it could also be applied to microsatellite data with a suitably adjusted emission probability.

Our approach has the following properties, some of which are novel:

- Gene conversion tract lengths may be arbitrarily long
- SNPs can be arbitrarily densely situated in the region of interest, allowing for multiple-SNP gene conversion tracts
- Crossover and gene conversion rates may vary across the region of interest
• Estimates can be obtained jointly for the crossover rate and gene conversion rate (and in theory, also for the gene conversion tract length, but in settings where the tract length is short relative to the average SNP spacing there is little information in the data to pinpoint this)

• It is model-based and calculates (an approximation to) the likelihood, so can provide estimates of uncertainty

We chose the LS model because it doesn’t rely on summary statistics, but attempts to use all the available information, albeit under an approximation to the likelihood, making it an ideal candidate for extension to the gene conversion model. We expect the trace of gene conversion to be difficult to detect and therefore wish to use the maximum information which can be extracted from the data.

We will first introduce briefly the LS model for crossover alone and then describe the addition of gene conversion to this model. We validate our method using tests on data simulated with a range of parameter values, and evaluate its robustness to deviations from our assumptions about population demographics. Finally we generalise the model to allow for variation in the rate of gene conversion.

**Modelling crossover using a likelihood-based approach:** The objective of maximum likelihood methods is to maximise the function $L(\Theta) = \Pr(H|\Theta)$, i.e. the likelihood of a set of model parameters $\Theta$ given the sampled data (haplotypes) $H = h_1, h_2, \ldots, h_n$.

If we knew the underlying genealogy of the sampled individuals, this could be calculated directly. However, this information, in a population genetic sample of unrelated individuals, is not available. In the presence of recombination, the individuals sampled may be related by a different (correlated) phylogenetic tree at each polymorphic site along the sequence (which, together, form the ancestral recombination graph (ARG) of Griffiths and Marjoram (1997)), and phylogenetic methods are unreliable under these circumstances (Schierup and Hein, 2000). It is therefore useful to develop an approximation to $L$ which is not conditional
on the ARG $G$ relating the sampled individuals, using

$$\Pr(H|\Theta) = \int \Pr(H|G, \Theta)\Pr(G)dG$$

(1)

where $\Pr(G)$ is the probability density function of the ARG relating the haplotypes $H$. One highly robust and flexible way to model $\Pr(G)$ is the coalescent with recombination (Kingman, 1982; Hudson, 1983; Griffiths and Marjoram, 1997). This assumes a panmictic population of constant size, undergoing only neutral evolution. We base our model and the majority of our simulations on the standard coalescent, but we also investigate the accuracy of our method when it is applied to data which deviate from the assumed coalescent model.

LS noted that

$$\Pr(h_1, ..., h_n|\rho) = \Pr(h_1|\rho)\Pr(h_2|h_1, \rho)...\Pr(h_n|h_1, ..., h_{n-1}, \rho)$$

(2)

where $h_i$ denotes the $i$th haplotype in the dataset of $n$ haplotypes, and $\rho = 4N_c$ is the population crossover rate. By approximating each of the terms on the RHS in turn, they arrived at an approximation to the likelihood known as a Product of Approximate Conditionals (PAC) model.

Their approximation $\hat{\pi}_A(k+1)$ to the conditional probability $\Pr(h_{k+1}|h_1, ..., h_k, \rho)$ is a modification of the imperfect mosaic model of Fearnhead and Donnelly (2001). Haplotype $k + 1$ is considered to be made up of segments copied from any or all of the preceding $k$ haplotypes, and at marker $l$ the haplotype being copied from is known as the “nearest neighbour”. The copying process can also be imperfect, giving rise to a difference between the new haplotype and its nearest neighbour; this is considered to be a mutation. When the nearest neighbour changes between marker $i$ and marker $i + 1$ this is considered to be a crossover. The sequence of nearest neighbours taken when traversing the $(k+1)^{th}$ haplotype haplotype from one end to the other can be modelled as a Markov chain where the nearest neighbour at a given marker is dependent only on that at the previous marker and on the
Figure 1: Illustration of the imperfect mosaic model with gene conversion. We construct the new haplotype $h_4$ as a mosaic of pieces copied from existing haplotypes $h_1$-$h_3$. The haplotype copied at a particular point is known as the nearest neighbour at that point, and the nearest neighbour can change when we encounter a gene conversion event such as the one between (a) and (b), or a crossover event (c).

crossover probability. The likelihood given a particular value of the parameter $\rho$ (which may vary across the region) is then calculated by summing over all possible mosaic structures and a maximum likelihood estimate $\hat{\rho}$ can therefore be found.

It is worth noting that this approximation to the likelihood is dependent on the order in which the haplotypes are observed. This unwelcome influence can be greatly reduced by averaging the likelihood over a number of different random orders. We find 20 orders sufficient to ensure that our estimates were consistent between different runs of the program, and more than 20 to be cumbersome in terms of computational time. All results shown in this document are based upon 20 orders chosen uniformly at random except where stated otherwise.

The LS model was previously extended by Hellenthal (2006) to include gene conversion, assuming each gene conversion tract only includes one SNP. In essence, the emission probability for the Markov chain is modified to mimic a gene conversion. This has the benefit of keeping the computational cost the same ($O(N^2)$), but suffers from a difficulty distinguishing gene conversion and genotyping error. Our adaptation of the LS model is much more computationally intensive but can be applied to more densely typed SNP data.
Modelling gene conversion: We now consider our sample to have been affected both by crossover and by gene conversion events throughout its history. This scenario is well modelled by the coalescent model with gene conversion developed by Wiuf and Hein (2000). The imperfect mosaic model can be easily adapted to allow for gene conversion, by allowing a second process to alter our nearest neighbour from $x$ to $x'$, with the proviso that we must eventually return to copying from $x$. See fig. 1 for an illustration of this. The distribution of lengths of gene conversion tracts can be approximated by a geometric distribution (Hilliker et al., 1994), giving a constant probability of ending the tract at any particular position and returning to copy from $x$, irrespective of the length of the tract so far. This lack of memory property allows us to use a Markov chain implementation of the model as above.

We make the following additional assumptions:

1. Crossover events occur independently of gene conversion events

2. Gene conversion events cannot overlap or be nested

3. The gene conversion rate may change instantaneously at each typed marker but cannot change within the interval between adjacent SNPs.

The first of these assumptions allows us to separate the gene conversion and crossover processes in our model, which simplifies the calculation of the transition probabilities in our Markov chain. It is also biologically reasonable in that we would not expect that the fact that a crossover had once occurred in a particular region to influence the probability that a gene conversion occurs in any given meiosis in that region, except in that a higher rate of crossover might point to a potentially higher rate of gene conversion. Our assumption does not disallow dependence between rates of gene conversion and crossover, only between events.

In our second assumption, we specify the conditions on entering and exiting a gene conversion event. We may only begin a new gene conversion tract when we are not already in a tract, and we may only end a tract by returning to copy from the haplotype we were copying from before the tract began. Allowing tracts to be nested and/or to overlap would violate
the Markovian property of our model or necessitate the addition of one or more further dimensions to the model. There is no clear biological interpretation of this assumption. It is certainly reasonable to state that any gene conversion event taking place during a particular meiosis cannot overlap with or be nested within another tract occurring in the same meiosis. The trace left in population genetic data by many independent gene conversion events over a long period of time, perhaps occurring in hotspots and likely to overlap with previous events, is less obvious. When tract lengths are short compared to SNP spacing, we do not expect this assumption to have any effect (two or more SNPs must be in a gene conversion tract for overlapping or nesting to be detectable). When tract lengths are long, we might expect to miss some overlapping gene conversion tracts, or see them as crossovers, thus giving a slight underestimate of the gene conversion rate and overestimate of the crossover rate.

Our final assumption is also one of convenience. We have no information about any variation in the rates of gene conversion and crossover in the gap between any pair of adjacent typed SNPs, and we therefore assume the rate is constant. In this article we are mainly considering regions where the rates of crossover and gene conversion are considered to be uniform, but the method also allows for rate variation. When rate variation exists, in this model the rate is only permitted to change instantaneously at a typed marker, and the rate in an interval will correspond to the average rate over the gap between the SNPs.

Details of our implementation of this model are in the appendix. In the next section we describe the results of applying this model to simulated data with the aim of jointly estimating $\gamma = 4N_e g$ and $\rho = 4N_e c$ where $g$ denotes the gene conversion rate per meiosis per unit distance.
Figure 2: Comparison of maximum likelihood estimates of $\rho$ on data simulated with different values of $f$ (all data simulated with $\rho = \theta = 1$ per kb) using our model (a) as compared to the LS model without bias correction (b). Our model gives good estimates of the crossover rate regardless of the amount of gene conversion present. It shows little bias, and when gene conversion is present (at least in simulated data), estimates of crossover rates can be inflated when gene conversion is not taken into account.
RESULTS

Simulation study: To test the performance of our method we undertook a simulation study. Datasets were simulated using the program ms (Hudson, 2002). Each dataset contains 50 haplotypes of length 20kb; simulated with mutation rate $\theta = 0.5$, 1, or 2.5; crossover rate $\rho = 0$, 0.5, 1 or 2.5; and gene conversion rate $\gamma = 0$, 1 or 10 (per kb). We will focus on the datasets with $\theta = 1$ as this corresponds to the human population-scaled mutation rate of approximately 0.7 - 1.0 per kb (Ptak et al., 2004). The mean gene conversion tract length, $\frac{1}{\lambda}$, was fixed at 500 bp (c.f Frisse et al., 2001) for the simulations and during estimation of parameters. The number of SNPs in each dataset varied with the mutation rate and, when $\theta = 1$, averaged 89 SNPs per simulated dataset.

Our estimates $\hat{\rho}$ and $\hat{\gamma}$ are shown in figs. 2(a) and 3, and summarised in table 1. In each case we fixed the gene conversion tract length parameter at the value used to simulate data.

Estimation of $\rho$: The distributions of our estimates $\hat{\rho}$ for datasets simulated with no gene conversion, equal rates of gene conversion and crossover ($f = 1$), and more gene conversion than crossover ($f = 10$) are shown in fig. 2(a). The presence of gene conversion does not seem to have a detrimental effect on our ability to estimate $\rho$, and estimates of $\rho$ are within a factor of 2 of the truth more than 90% of the time for each value of $f$. All the simulations shown used $\rho=1$ per kb, simulations with $\rho = 0.5$ and $\rho = 2.5$ per kb had similar results with slightly reduced accuracy. For comparison, estimates for $\rho$ obtained using LS (without bias correction) on the same datasets are shown in fig. 2(b). In the case where $f = 10$, these estimates are highly inflated, not surprisingly since this method is not intended for datasets where gene conversion is present. However, the fact that for all 1000 datasets with $f = 10$, this method gives an estimate for $\rho$ which is more than twice the true value, serves as a reminder of the effect that undetected gene conversion can have on estimates of the crossover rate.
Figure 3: Maximum likelihood estimates of $\gamma$ on data simulated with $\rho = \theta = 1$ per kb using our model. Estimates obtained for data with high levels of gene conversion activity are very encouraging (999 out of 1000 within a factor of two of the truth), but we tend to overestimate the level of gene conversion present when it is low or non-existent. This is inevitable due to the true value being at the boundary of the range of possible values.

*Estimation of $\gamma$:* The distributions of our estimates $\hat{\gamma}$ for the same datasets as above are shown in fig. 3. In the case where $f = 10$, our estimated $\gamma$ was within a factor of two of the value used to simulate data, for 999 out of the 1000 simulated datasets. Results are summarised in table 1.

*Estimation of $f$:* We used our estimates of $\rho$ and $\gamma$ for each dataset to obtain an estimate $\hat{f} = \frac{\hat{\gamma}}{\hat{\rho}}$. These estimates were also generally close to the truth but suffered slightly from being a ratio of two other estimates with uncertainty in both. For datasets simulated with $\theta = 1$ our median estimates of $f$ were 0.55, 1.45, and 9.54 for datasets simulated with true $f$ 0, 1, and 10 respectively.

**Robustness to deviation from assumed demography:** In this section we use additional simulated data to evaluate the robustness of our model to deviations from the assumed neutral model. We consider our three major demographic assumptions: Constant population size, panmictia (random mating), and neutral evolution. In each case, we simulated 100
<table>
<thead>
<tr>
<th>Simulated parameters</th>
<th>Deviation from model</th>
<th>median $\hat{\gamma}$</th>
<th>median $\hat{\rho}$</th>
<th>median $\hat{f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma = 0$, $\rho = 1$ ($f = 0$)</td>
<td>None</td>
<td>0.5332 (11.5)</td>
<td>0.9706 (96.9)</td>
<td>0.55 (11.5)</td>
</tr>
<tr>
<td>Growth</td>
<td>0.5863 (16)</td>
<td>0.9045 (89)</td>
<td>0.586 (16)</td>
<td></td>
</tr>
<tr>
<td>Bottleneck</td>
<td>0.4777 (18)</td>
<td>0.8678 (95)</td>
<td>0.516 (18)</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>0.5472 (5)</td>
<td>0.9415 (97)</td>
<td>0.634 (5)</td>
<td></td>
</tr>
<tr>
<td>Selection</td>
<td>0.6031 (8)</td>
<td>0.9464 (98)</td>
<td>0.667 (8)</td>
<td></td>
</tr>
<tr>
<td>$\gamma = 1$, $\rho = 1$ ($f = 1$)</td>
<td>None</td>
<td>1.348 (77.3)</td>
<td>0.9323 (95.8)</td>
<td>1.45 (62.9)</td>
</tr>
<tr>
<td>Growth</td>
<td>1.344 (71)</td>
<td>0.8907 (94)</td>
<td>1.36 (55)</td>
<td></td>
</tr>
<tr>
<td>Bottleneck</td>
<td>1.286 (75)</td>
<td>0.8868 (97)</td>
<td>1.43 (64)</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>1.401 (84)</td>
<td>0.9077 (98)</td>
<td>1.48 (67)</td>
<td></td>
</tr>
<tr>
<td>$\gamma = 10$, $\rho = 1$ ($f = 10$)</td>
<td>None</td>
<td>9.428 (99.9)</td>
<td>0.9708 (94.0)</td>
<td>9.54 (90.1)</td>
</tr>
<tr>
<td>Growth</td>
<td>7.805 (97)</td>
<td>0.8738 (92)</td>
<td>9.08 (87)</td>
<td></td>
</tr>
<tr>
<td>Bottleneck</td>
<td>8.315 (97)</td>
<td>0.8997 (94)</td>
<td>8.91 (82)</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td>8.455 (99)</td>
<td>0.8927 (93)</td>
<td>9.50 (89)</td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td>7.49 (97)</td>
<td>1.21 (93)</td>
<td>6.24 (68)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Summary of results of testing done on simulated data. We first simulated datasets according to standard model assumptions (constant sized, panmictic population with neutral evolution), with $\theta = 1$, and a variety of values of $f$. For each set of parameters $\hat{\gamma}$, $\hat{\rho}$, and $\hat{f}$, the median estimate for 1000 independent simulations is given, with the proportion of datasets for which the estimate lies within the range $[\text{truth}/2, \text{truth}*2]$ in brackets. In the case $f = 0$, we see that in 11.5% of cases we found $\hat{\gamma} = 0$. However, for 82.8% of these datasets, $\hat{\gamma} < 1$ per kb. We also show the corresponding results for 100 datasets simulated with a range of deviations from the above assumptions (see text). These results are broadly similar to those with no deviation, except in the case of rephased data where the crossover rate is overestimated.
datasets with 50 haplotypes, 20kb in length, with mutation and crossover rates of 1 per kb, and various gene conversion rates.

Variation in population size: To test our model in the presence of population size variation, we simulated 100 datasets with a bottleneck ~0.15 $N_e$ generations ago which reduced diversity to 85% of that expected without the bottleneck, and a further 100 datasets with scaled exponential population growth parameter 1 (c.f. McVean et al., 2004). Results under these demographic variations (summarised in table 1) do not deviate far from those obtained on data simulated under the standard model, although we see a slight increase in our underestimation of the gene conversion rate when $f$ is high.

Population structure: To test our model in a non-random mating scenario, we simulated 100 datasets, where 25 of the 50 chromosomes were sampled from each of two subpopulations corresponding to a level of population differentiation of $F_{ST} \approx 0.2$ (c.f. Pritchard and Przeworski, 2001). Properties of the resulting estimates for $\hat{\gamma}$ are shown in table 1. Although in each simulation, the variance (not shown) of $\hat{\gamma}$ was higher than that for a single-population dataset, this did not have a big effect on the median estimate or the proportion of results within a factor of 2 of the truth. The estimates for $\hat{\rho}$ were similarly affected.

Selection: As a final test we simulated 100 datasets where a positive selective sweep had just finished. This data was generated using the program SelSim (Spencer and Coop, 2004). The strength of selection was chosen to be $\sigma = 2N_e s = 50$, where $s$ is the selective coefficient between homozygotes (c.f. Smith and Fearnhead, 2005), and was applied to a single site in the centre of the 20kb region. Again we saw no major difference in our results, implying that this method is robust to low to moderate levels of selection (see table 1).

Genotype data: In order to use our method on genotype data it is necessary to first phase the data Currently available programs to phase genotype data do not take gene conversion into account, so we investigated the effect of performing this pre-processing of the data. For
100 of the above datasets simulated with \( f = 10 \), we randomly paired the 50 haplotypes into 25 individuals, and then used the program PHASE (Stephens et al., 2001; Stephens and Scheet, 2005) to rephase the data. We then obtained estimates for the gene conversion and crossover rates on these datasets, which are summarised in table 1. We found that our method overestimates the crossover rate under these circumstances, as well as underestimating the gene conversion rate, which leads to an underestimate of \( f \).

**Comparison with other methods** We now compare our results with two other methods: Hudson’s pairwise composite likelihood method (Hudson, 2001), and a method based on the summary statistics method of Padhukasahasram et al. (2006). For the former, we used the program maxhap, freely available from the author’s website. For the latter, we adapted a program (also available on the author’s website). Our implementation only differs from that described in Padhukasahasram et al. (2006) in that we do not fix the positions of the segregating sites in our simulations. Results are shown in table 2. For these calculations, maxHap took about 1.7 seconds per dataset, summStat between 5 and 24 hours (depending on \( f \)), and GenCo just under 1 hour on a standard desktop computer.

**Drosophila melanogaster:** A particularly interesting organism in the study of LD is *Drosophila melanogaster*, because of its unusual patterns of recombination. We applied our method to SNP data from two genes near the telomere of the X chromosome of African *D. melanogaster* (Langley et al., 2000). This dataset consists of 87 SNPs within the su(s) and su(w) genes, which are involved in the regulation of gene expression (Fridell and Searles, 1994). The genes are approximately 4 and 2.5 kb long respectively, and are separated by a region of 400 kb in which no SNPs were typed. Like chromosome 4 in *Drosophila* (Hochman, 1976), the region near the X chromosome telomere is subject to a severely reduced level of crossover per physical length (Aguade and Langley, 1994) compared to the genome-wide average rate of 1.5 cM/Mb (Nachman, 2002), perhaps due to regulation of double-strand-break repair mechanisms (MCKIM et al., 2002).
Table 2: Comparison with other methods. For 1000 datasets simulated with $\rho = 1$ and $\gamma \in (1, 10)$, we compare our results (GenCo) with those from maxHap (Hudson, 2001), and for 100 of the same datasets we also show results obtained using a third method (summStat) based on that of Padhukasahasram et al. (2006) (see text). Maxhap was run over a grid of 11 $f$ values ranging from 0 to 2.5 or 25 (inclusive) for the simulated datasets with $f = 1$ and $f = 10$ respectively. SummStat was run on a coarse grid of $\rho \in (8, 10, 20, 40, 45)$, and $\gamma \in (8, 10, 20, 40, 45)$ for the first test and $\gamma \in (80, 100, 200, 400, 450)$ for the second test. For each parameter $\gamma$, $\rho$ and $f$ we present the proportion of datasets for which the estimated value was within a factor of two of the value used to simulate data.

In addition to obtaining maximum likelihood estimates for $\rho$ and $\gamma$ for this dataset we also constructed a likelihood surface over a grid of values of $\rho$ and $\gamma$, shown in figure 4. We fixed the mean gene conversion tract length at 352 base pairs (Hilliker et al., 1994) and obtained $\hat{\rho} = 0.067$ and $\hat{\gamma} = 26.9$ per kilobase ($f = 432$). Such a strong signal is unlikely to be explained by repeat mutation or genotyping error. These estimates support the conclusion of Langley et al. (2000) that whilst crossover is suppressed in the region, gene conversion is not. This could indicate that gene conversion and crossover are completely separate processes, or, if both are initiated by the same process, that in this region of $D. melanogaster$ there is a strong tendency for recombination events to be resolved as gene conversion rather than crossover. Whether this is the cause of, or a consequence of the suppression of crossover is as yet unknown.

**Variation in the rate of gene conversion:** To date there is little data regarding fine-scale variation in gene conversion rates. The clearest examples of such variation are the
Figure 4: Likelihood surface for D. melanogaster dataset. The maximum likelihood point on the surface is $\gamma = 26.8$, $\rho = 0.062$ per kb. The surface is fairly flat around this region but drops off steeply when $\rho$ gets close to zero or $\gamma$ drops below about 15.

gene conversion hotspots experimentally identified in the centre of two crossover hotspots by Jeffreys and May (2004). Padhukasahasram et al. (2006) estimated non-uniform gene conversion rates in simulated data, and found that their method underestimates the gene conversion rate under these circumstances. Their method produces a single estimate of the total amount of gene conversion in a given region, and does not attempt to pinpoint hotspots or to measure their intensity.

Our model allows for a different rate of gene conversion between each pair of adjacent SNPs, so it was possible to implement an expectation maximisation algorithm to determine $\hat{\gamma}_i$ for each interval $i$. It should be noted that in order to reflect biological reality and maintain symmetry we would prefer to model the rate at which gene conversion initiation sites are encountered (i.e. somewhere around the middle of the tract), but due to the way the model is implemented we are in fact modelling the rate at which the left hand side of gene conversion tracts are encountered. It would be preferable to model a gene conversion tract extending in both directions from an initiation point (Helltenthal and Stephens, 2007), but this would greatly increase the complexity and hence the computation time of our method.
Instead, we map our estimates to the gene conversion rate by assuming the initiation is in the exact centre of the gene conversion tract, according to the following equation:

$$\gamma'(x) = \int_0^{\infty} 2\lambda \hat{\gamma}(x - y)e^{-2\lambda y}dy$$ (3)

where $\hat{\gamma}(x)$ is our MLE of the gene conversion rate at distance $x$ from the beginning of the observed region, and a constant rate $\gamma_0$ is assumed for all $x < 0$.

When the distances between markers are long compared to the length of a gene conversion tract, or when rates change only gradually between intervals, the difference between modelling the centre and modelling the end of a gene conversion tract will be negligible. However, if very narrow hotspots of gene conversion are found, it may be necessary to convert the rate of encountering the left hand side into the rate of gene conversion tract initiation in order to provide a useful gene conversion rate estimation.

To examine the power and reliability of our method when recombination rates vary, we used the program msHOT (Helligenthal and Stephens, 2007) to simulate 100 datasets containing a hotspot for both gene conversion and crossover. Each dataset consisted of 50 haplotypes, 20kb in length, with $\theta = 1$ per kb, mean tract length 500 bp, and $\gamma = 0.5$ and $\rho = 0.05$ per kb ($f = 10$), except in a ‘hotspot’ 2 kb wide in the centre of the region, where $\gamma = 50$ and $\rho = 5$ per kb ($f = 10$).

Assuming that $f$ was constant across the region, we obtained maximum likelihood estimates for $\gamma$ and $f$ for each simulated dataset. The estimates for $\gamma$ and their median (sampled every 100 bp) are shown in fig 5.

Individual estimates of $\gamma$ show high levels of variance, but on average, the position, width and heat of the estimated hotspot are close to the values used to simulate data, and there is little bias in our estimates of $\gamma$. However, estimates of $\rho$ are downwardly biased, resulting in an overestimate of $f$ (median 25.1). More work is needed to develop a method which can produce a less biased estimate of $f$ in this variable rate scenario, even under the restriction that $f$ is constant. To obtain a more reliable gene conversion rate estimate on a single
Figure 5: Variable rate simulations: We estimated gene conversion and crossover rates for 100 datasets, simulated with a hotspot in which the gene conversion and crossover rates are 100 times the background rate (simulated gene conversion rate shown in red). To estimate the gene conversion rate for each of these datasets, we assume the ratio $f$ of gene conversion to crossover is fixed throughout the region, and run our program 20 times with one independent random ordering each time. The results from each run are transformed using equation 3 and we use the median as our best estimate for that dataset. Here we show (in black) the median of these 100 estimates, and the 5th and 95th percentiles (grey).

Despite some evidence that $f$ may vary between regions in humans (HELLENTHAL and STEPHENS, 2006; PADHUKASAHASRAM et al., 2006), we do not consider this scenario - mainly because population genetic data is unlikely to contain sufficient information to obtain an accurate fine-scale map of gene conversion rates independently of crossover rates, but also because the existence of gene conversion hotspots within crossover hotspots implies some correlation between the two rates, and finally, the processing time needed to maximise such a likelihood would be immense.
**Human Chromosome 1** Many crossover hotspots have been identified in the human genome, but of particular interest is the MS32 hotspot on chromosome 1, the existence of which is supported by strong experimental evidence, but has not left a significant imprint on LD (Jeffreys et al., 2005). We applied our method to a 206 kb region, including this hotspot and several others. The SNP data (Jeffreys et al., 2005) consists of 214 SNPs on 80 genotypes; we used PHASE v2.1.1 (Stephens et al., 2001; Stephens and Scheet, 2005) to infer the haplotypes and missing data, and averaged our results over 20 independent random subsamples of 50 haplotypes, each taken in 10 random orders. For comparison, the crossover rate for this region, estimated using LDhat (McVean et al., 2002), is shown below. LDhat was run for $10^8$ iterations with block penalty 5, results were sampled every $10^4$ iterations and the first 100 samples discarded. To reflect the idea that crossover and gene conversion hotspots tend to coincide (Jeffreys and May, 2004), we allowed gene conversion rates to vary independently in each interval between SNPs whilst keeping $f$ constant everywhere in the region. Our median estimated gene conversion rate is shown in fig 6. Our median estimate of $f$ for this region was 1.5. This estimate is strongly influenced by the gene conversion tract length parameter. In this study we assumed the mean tract length was 100 bp, but note that a longer mean tract length would lead to a lower $\hat{\gamma}$ and a correspondingly lower $\hat{f}$.

For comparison, we also analysed the TAP2 region of the human MHC, and found $f$ to be higher in this region, approximately 9. This lies centrally within the range of 4-15 suggested by Jeffreys and May (2004) although as above, is dependent on the gene conversion tract length. The difference between this estimate, and the one for the MS32 region, could reflect variation in $f$ between different regions of the human genome, also reported by Hellenthal and Stephens (2006).

**DISCUSSION**

We have developed a powerful and robust method for estimating gene conversion rates from population genetic data. Our accuracy is at its best when analysing data which has been
Figure 6: Median maximum likelihood estimate for the gene conversion rate (blue) in 206 kb region around the MS32 gene. For comparison, the crossover rate for the same region is shown in green. The gene conversion rate estimate assumes that $f$ is constant throughout the region, and the mean tract length is 100 bp. Red triangles at the top and bottom show the centres of hotspots identified experimentally. Assuming no gene conversion, the hotspot found experimentally at MS32 (Jeffreys et al., 2005) shows little signal in population genetic data. However, under the gene conversion model this hotspot can clearly be seen.

affected by fairly high levels of gene conversion, and where the mean tract length is at least comparable to the mean SNP spacing. Our model also provides a reliable estimate of the rate of crossover in a region, regardless of the gene conversion rate. Our results are not seriously damaged by the most common deviations from standard model assumptions: non-random mating, changing population size, and non-neutral evolution.

Our model allows multiple SNPs to be included in a gene conversion tract. SNP density varies widely between datasets, but also within datasets: for example, in the MS32 dataset analysed above there are 214 SNPs in a 206 kb region, giving an average interval of 967 bp between adjacent markers. However, 45 intervals (21%) are less than 100 bp long and 133 (62%) are less than 500 bp long. Our simulations show that when the mean tract length is 100 bp, 9% of gene conversion tracts which initiate within this region will encompass the positions of two or more markers. With 500 bp tracts, this rises to 38%.

When applied to data with little or no history of gene conversion, our model tends to overestimate the gene conversion rate. This is mainly due to the true value lying on or near the boundary of the parameter space. Simulation results (figure 2, first column) demonstrate
that including gene conversion in the model results in improved estimation of the recombination rate, suggesting that modelling errors are preferentially interpreted as gene conversion events. This implies that the use of our model could result in improvements to the estimation of the underlying crossover rate, even in the case where gene conversion is not occurring.

Estimates of uncertainty cannot be obtained directly from this method, due to the approximate likelihood used. To obtain confidence intervals it is necessary to perform a simulation study tailored to the specifics of a given dataset (such as the number of haplotypes and rate of mutation).

We analysed a region of the X chromosome in *D. melanogaster* and found that, under the assumption that the rates of crossover and gene conversion are constant across the region, gene conversion events occur over 400 times as often as crossovers. Application of this model to additional regions of the *Drosophila* genome could enhance our understanding of the unusual patterns of recombination in this species.

We also analysed a region of human chromosome 1, around the MS32 gene, a region containing several known crossover hotspots. Our analysis, allowing the rates of crossover and gene conversion to vary across the region whilst keeping their ratio \( f \) constant, shows that the MS32 hotspot, previously difficult to detect using population genetic methods, appears highly active under a gene conversion model. This could indicate that this hotspot is more active in gene conversion than in crossover, but it seems to contradict conclusions that the hotspot has only recently emerged (Jeffreys et al., 2005).

The maximum likelihood estimate \( \hat{f} = 1.5 \) for this region should be treated with caution for two reasons: simulations show that this is a biased estimator (see above), but also \( \hat{f} \) is highly dependent on the accuracy of our tract length estimate (here 100 bp). As with the method of Ptak et al. (2004), experiments with our method have shown that there is a strong correlation between the estimated values of \( \lambda \) and \( \gamma \) (data not shown).

For this reason, we do not attempt to estimate the gene conversion tract length from the data. By misspecifying the mean tract length parameter for datasets simulated with known
tract length $t_0$, we find that using a mean tract length estimate $t^\star$ which is double that simulated ($t^\star = 2t_0$) results in a slightly lowered gene conversion rate estimate, while using an estimate $t^\star = \frac{t_0}{2}$ causes us to overestimate the gene conversion rate by approximately a factor of 2 (data not shown). These results will vary depending on the SNP spacing and the actual gene conversion rate. Fortunately, estimates of the gene conversion tract length are available for several organisms (e.g. Nishant et al., 2004; Palmer et al., 2003; Jeffreys and May, 2004), although little is yet known about whether there is heterogeneity in this length between different genomic regions.

Our implementation of this model is of order approx $N^3$ and is linear in both the number of markers and the number of orders used. It takes approximately 30 mins of processing time on a standard desktop machine to jointly calculate the constant MLEs for $\rho$ and $\gamma$ when $N = 50$, $L = 100$, $\lambda$ is fixed, and 10 random orderings are used. However, when $\gamma$ and $\rho$ are allowed to vary, with their ratio $f$ constant, using the same dataset it would take around 9 hours to obtain $\hat{\gamma}$ and $\hat{f}$ (and therefore $\hat{\rho}$). This time could be reduced by improvements to the implementation and by running on a faster computer. For very large datasets the method is impracticable with present computers, but results can still be obtained by taking several subsamples of the data and taking the median result. Subsamples should be as large as possible as smaller subsamples tend to produce underestimates of the gene conversion rate and have higher variance (see Supplementary Table 1), but when averaging over several subsamples, one order is sufficient. For reasonable sized datasets we believe the accuracy obtained by this model’s usage of all the information contained in the data makes its slow speed a worthwhile penalty.

In this article we do not consider the effects of genotyping error on our results. Genotyping error can have a similar effect on the patterns of genetic diversity to that of gene conversion (Ptak et al., 2004). In densely typed SNP data, allowing for multi-SNP gene conversion tracts should reduce the impact of genotyping error.

We anticipate that this model could be adapted to detect the signature of non-allelic gene conversion in population genetic data, which would be particularly useful when considering
the evolution of multi-gene families such as the histones (see NEI and ROONEY, 2005).

As well as finding fine-scale variations in $\gamma$, it would be straightforward to adapt this model to compare estimates of $f$ for different regions or genes within a given organism, allowing production of a broad-scale map of $f$. This could be used to discover whether large-scale genomic features such as proximity to centromeres affect this ratio.

Whether $f$ is constant at a fine scale remains an open question.

ACKNOWLEDGMENTS

The authors would like to thank Daniel Falush, Chris Spencer, Matthew Stephens and one anonymous reviewer for helpful comments on earlier versions of this work. Many of the simulations and data analyses were performed on the multi-node computing cluster that was funded by a grant from the Wolfson Foundation to Peter Donnelly. JG is funded by the Engineering and Physical Sciences Research Council through the University of Oxford Life Sciences Interface Doctoral Training Centre.

LITERATURE CITED


LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. M. BRAVERMAN, 2000 Linkage disequilibria and the site frequency spectra in the su(s) and su(wa) regions of the drosophila melanogaster x chromosome. Genetics 156: 1837–52.


APPENDIX

Hidden Markov model implementation: Here we detail the implementation of the model described above, including the state transition probabilities and some algorithmic shortcuts used to reduce the computation time. The ‘hidden’ data in our model is the ‘true’ underlying mosaic structure of the current haplotype. Under the true coalescent genealogy, a mosaic consisting of haplotypes we have already seen may not exist, but under our approximation it always does. We do not try to infer this but sum over all mosaic structures using the HMM formulation. Each of the terms $\Pr(h_{k+1}|h_1, ..., h_k, \rho, \gamma)$ is approximated using its own HMM with $k(k+1)$ states and its own emission and transition probabilities which depend on $k$.

Our HMM has $k(k+1)$ distinct states ($X = x, G = g$) where $1 < x \leq k$ denotes our nearest neighbour, unless $g \neq 0$, in which case we are in a gene conversion, and the current nearest neighbour is $g$. Starting at the left-most marker, we calculate the likelihood of the data for each possible state $(X_j, G_j)$ at each marker $j$, based on the $k(k+1)$ distinct state probabilities at the previous marker, and the transition ($t$) and emission ($e$) probabilities (see below).
**Transition Probabilities:** We model the process of crossover and that of gene conversion as separate processes, happening independently of each other. The variable $X_j$ can only be modified by crossover, while $G_j$ is only affected by gene conversion.

\[ \Pr(X_{j+1}, G_{j+1} | X_j, G_j) = \Pr(X_{j+1} | X_j) \Pr(G_{j+1} | G_j) \]  

(A1)

**Initial state probabilities:** Since we have no data outside our region, $\Pr(X_1 = x) = \frac{1}{k}$ for all $x$. The probability that we start our Markov chain inside a gene conversion tract depends on how the rate of starting a gene conversion tract compares to the rate of ending one.

\[ \Pr(G_1 = g) = \begin{cases} \frac{\lambda k}{(k \lambda + \gamma)} & (g = 0) \\ \frac{\gamma k}{(k \lambda + \gamma)} & (g \neq 0) \end{cases} \]  

(A2)

**$X$ transition probabilities:** The first term on the RHS of equation A1 is given by Li and Stephens as:

\[ \Pr(X_{j+1} = x | X_j = x') = \begin{cases} \frac{1}{k} \left(1 - e^{-\frac{\rho_j d_j}{k}}\right) & (x \neq x') \\ e^{-\frac{\rho_j d_j}{k}} + \frac{1}{k} \left(1 - e^{-\frac{\rho_j d_j}{k}}\right) & (x = x') \end{cases} \]  

(A3)

when we are considering the $(k + 1)$th haplotype and the distance between marker $j - 1$ and marker $j$ is $d_j$. Informally, in the case where $x \neq x'$, we must have had at least one crossover between the two sites, and the probability that the new nearest neighbour was $x'$ is $1/k$. When $x = x'$, we may have had no crossover event, or we may have crossed-over one or more times but in the last event chose the same nearest neighbour.

**$G$ transition probabilities:** To calculate $\Pr(G_{j+1} | G_j)$ we must consider not only the probability of beginning a gene conversion event within the current interval but also that of ending one. The rate of terminating a gene conversion tract is fixed at $\lambda$, regardless of the length the tract has so far covered. This geometric model allows us to consider the ending of a gene conversion event as a process in its own right, which goes on all the time, independently of
our current state, and resets the state of the system to the base (non-gene-conversion) state. This is reasonable because although in biological terms there is no event corresponding to the end of a gene conversion which can occur outside of a gene conversion tract, any such event occurring when we are already in the base state has no effect and thus has no effect on our model.

For each type of transition between gene conversion states, we will describe the sequence of events which could cause that transition to occur, and give the probability of undergoing this transition. In each case, any events occurring before the last reset event within the interval will have no effect on the state at the right hand side of the interval. We therefore integrate back from the right hand side of the interval, over possible positions of the last reset event, so that we do not have to explicitly consider how many gene conversion events may have taken place prior to the final reset event.

There are 5 distinct types of transition between gene conversion states:

1. We are currently not in a gene conversion, and we were not in one at the last marker. We break this down into two scenarios: There was no reset event in the interval (which happens with probability \( \exp(-\lambda d_j) \)) and also no gene conversion event (probability \( \exp(-\gamma d_j/k) \)); or there was a reset event, and no gene conversion event has taken place since then. This last term can be written as the integral (over all possible places at which the last reset event might have happened) of the probability that no further reset event occurred multiplied by the probability that no gene conversion occurred.

\[
\Pr(G_{j+1} = 0 | G_j = 0) = e^{-\lambda d_j} e^{-\frac{\gamma d_j}{k}} + \int_0^{d_j} \lambda e^{-\lambda x} e^{-\frac{\gamma x}{k}} dx
\]

\[
= e^{-\frac{\lambda k + \gamma}{k}} \left[ 1 - \frac{\lambda k}{\lambda k + \gamma} \right] + \frac{\lambda k}{\lambda k + \gamma} \tag{A4}
\]

2. We have moved from a non-gene conversion state to a gene conversion state:

\[
\Pr(G_{j+1} = g | G_j = 0) = \frac{e^{-\lambda d_j/k} [1 - e^{-\frac{\gamma d_j}{k}}]}{k(\lambda k + \gamma)} \left[ 1 - e^{-\frac{\lambda k + \gamma}{k} d} \right]
\]

\[
= \frac{\gamma}{k(\lambda k + \gamma)} \left[ 1 - e^{-\frac{\lambda k + \gamma}{k} d} \right] \tag{A5}
\]
(either there was no reset event in the interval but there was a gene conversion event which made \(g\) our new nearest neighbour, or there was a reset event and there was a gene conversion event after it)

3. Previously we were in a gene conversion event but now we are not:

\[
Pr(G_{j+1} = 0 | G_j = g) = \int_0^{d_j} \lambda e^{-\lambda x} e^{-\frac{x^2}{2}} \, dx
= \frac{\lambda k}{\lambda k + \gamma} \left[ 1 - e^{-\frac{\lambda k + \gamma}{k} d} \right]
\]  

(A6)

(no gene conversion event has taken place since the last reset event)

4. We were previously in a gene conversion state where we were copying from haplotype \(g\), and we are currently in a similar state:

\[
Pr(G_{j+1} = g | G_j = g) = e^{-\lambda d_j} + \int_0^{d_j} \lambda e^{-\lambda x} \left[ 1 - e^{-\frac{x^2}{2}} \right] \frac{1}{k} \, dx
= \frac{k-1}{k} e^{-\lambda d} + \frac{\lambda}{\lambda k + \gamma} e^{-\frac{\lambda k + \gamma}{k} d} + \frac{\gamma}{k(\lambda k + \gamma)}
\]  

(A7)

(either no reset event has occurred, or there has been a gene conversion event choosing the same value of \(g\) since the last reset event)

5. We were previously in a gene conversion state copying from haplotype \(g\) and we have moved into a gene conversion state where we are copying from haplotype \(g'\) where \(g \neq g'\):

\[
Pr(G_{j+1} = g' | G_j = g) = \int_0^{d_j} \lambda e^{-\lambda x} \left[ 1 - e^{-\frac{x^2}{2}} \right] \frac{1}{k} \, dx
= \frac{\lambda}{\lambda k + \gamma} e^{-\frac{\lambda k + \gamma}{k} d} + \frac{\gamma}{k(\lambda k + \gamma)} - \frac{1}{k} e^{-\lambda d}
\]  

(A8)

(as above but without the option for no event occurring)

These gene conversion state transition probabilities, together with the crossover state transition probabilities above, make up the state transition probabilities for our Markov chain.

We will write \(t_G(g'|g, j) = Pr(G_{j+1} = g'|G_j = g)\) and \(t_X(x'|x, j) = Pr(X_{j+1} = g'|X_j = x)\).
Emission Probabilities: When it is not known, we (as LS) use Watterson’s estimator (WATTERTON, 1975) to approximate the per-site rate of mutation, $\theta/L$:

$$\frac{\theta}{L} = \left(\sum_{m=1}^{n-1} \frac{1}{m}\right)^{-1}\quad (A9)$$

Conditional on the hidden state $(X_j, G_j)$ at marker $j$, we could calculate the emission probability based on whether a mutation had occurred or not, compared to the chromosome $c$ from which we are copying ($c = X_j$ if $G_j = 0$, otherwise $c = G_j$). This is simply

$$e_k(j | X_j, G_j) = Pr(h_{k+1,j} | X_j, G_j) = \begin{cases} \frac{\theta}{2(kL+\theta)} & (h_{k+1,j} \neq h_{c,j}) \\ \frac{2kL+\theta}{2(kL+\theta)} & (h_{k+1,j} = h_{c,j}) \end{cases}\quad (A10)$$

Likelihood: Let $p_{x,g}(j)$ be the relative probability of being in the state $(x, g)$ at the marker $j$, given the data up to that marker. Then

$$p_{x,g}(1) = \Pr(X_1 = x, G_1 = g)\Pr(h_{k+1,1} | X_1 = x, G_1 = g) = \begin{cases} \frac{1}{k^2} \left(\frac{g}{\gamma+k} \right) e_k(1 | x, g) & (g \neq 0) \\ \frac{1}{k} \left(\frac{k\lambda}{\gamma+k\lambda} \right) e_k(1 | x, g) & (g = 0) \end{cases}\quad (A11)$$

and

$$p_{x,g}(j) = \sum_{x',g'} p_{x',g'}(j - 1) t_X(x | x', j - 1) t_G(g' | j - 1) e_k(j | x, g)\quad (A12)$$

and the approximate likelihood of the data, for the $(k + 1)$th chromosome, for this chromosomal order, is given by

$$\pi_C(h_{k+1} | h_1, h_2, ..., h_k, \rho, \gamma, \lambda) = \sum_{x,g} p_{x,g}(L)\quad (A13)$$

In order to overcome the issue of order-dependency, we repeat this calculation $n \geq 10$ times and take the average likelihood.
**Optimisation:** The basic algorithm described above is of approximate order $N^4$, but we were able to reduce this to $N^3$ by calculating sums of several subgroups of the $p_{x,g}(j)$, and using these sums to facilitate the calculation of $p_{x',g'}(j + 1)$. First note that $t_X(x|x',j)$ can only take two possible values, depending on whether $x = x'$. Similarly, $t_G$ can take five values as described above. So, for example, $t_X(x|x',j)$ is the same for all values of $x'$ except $x' = x$, and the sum $\sum_{x'=1}^{k} p_{x',g'}(j)$ gives the total probability of all the states where $G_j = g'$. In order to calculate $p_{x,g'}(j + 1)$, subtracting $p_{x,g'}(j)$ from this sum leaves a group of states at marker $j$ for which the transition probability is the same.

These sums, calculated once for each marker, can be used many times in different combinations.

**Estimating parameters:** To find maximum likelihood estimates we use a direct search algorithm (Hooke and Jeeves, 1961) to find the $\hat{\rho}$ and $\hat{\gamma}$ which maximise the average likelihood. When variable rate estimates are required, we use Expectation Maximisation to find these rates independently in each interval.

The C++ code and windows/linux executables for our implementation of this model are available from http://www.stats.ox.ac.uk/~gay/