msHOT: Simulating Crossover and Gene Conversion Hotspots with the ms Simulator

Garrett Hellenthal & Matthew Stephens

Department of Statistics
University of Washington, Box 354322, Seattle WA 98195-4322, USA
hellenth@stats.ox.ac.uk

Variable Rates in Recombination

Rates of **recombination** vary considerably on fine-scales, i.e. kilobases, across genetic regions (Crawford et al. (2004),Jeffreys/Kauppi/Neumann (2001),McVean et al. (2004)).

In particular, **crossover** events appear to cluster into narrow, 1-2kb regions of intense activity, known as **hotspots**. These hotspots may also be highly active for allelic **gene conversion**, i.e. nonreciprocal exchange of genetic material between homologous regions without crossing-over (Jeffreys/May (2004),Jeffreys/Neumann (2005)) (see Fig.1).

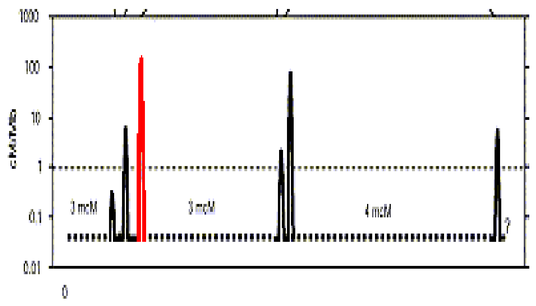


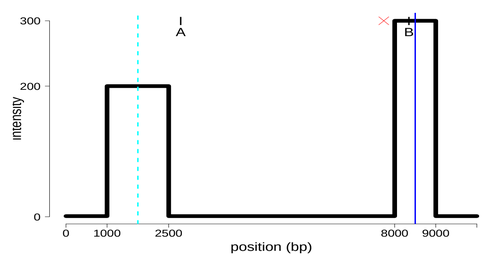
Figure 1: *A 200-kb region of the Major Histocompatibility Complex (MHC) in humans (chromosome 6), explored via sperm analysis by Jeffreys/Kauppi/Neumann (2001). The y-axis depicts the rate of crossover activity in cM/Mb. Crossover activity here appears to primarily occur in six regions of very localized, intense activity, known as hotspots. The DNA3 crossover hotspot, highlighted in red, was explored via sperm analysis by Jeffreys/May (2004) jointly for crossover and gene conversion activity. The authors found DNA3 to have relatively large rates of each process. (Image taken from Jeffreys/Kauppi/Neumann (2001).)*

Furthermore, crossover hotspots appear to be a common feature of the human genome, occuring every ≈30-50kb genomewide (Fearnhead/Smith (2005),Myers et al. (2005)).

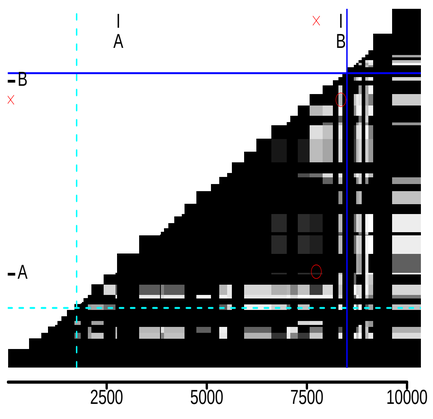
Hotspots and LD

The effect of such hotspots on patterns of genetic variation in humans, e.g. **Single Nucleotide Polymorphism (SNP)** data, is to break down the associations, or **Linkage Disequilibrium (LD)**, amongst locations even in close proximity to one another.

This can have vast consequences on the methodologies that use such associations, in particular **association studies**, which rely on finding SNP markers strongly associated with a disease-causing location (see Fig. 2).



(a)



(b)

Figure 2: *The effects of hotspots on patterns of LD.*
(a) msHOT *was used to simulate a 10kb region with two hotspots, one of length 1.5kb with recombination (crossover and gene conversion) activity 200 times that of the background rate (dashed line) and one of length 1kb with recombination activity 300 times that of the background rate (solid line). Consider a disease-causing SNP at × in this region, and SNP markers at positions* **A** *and* **B** *as shown.*
(b) D' *association statistic values across 50 haplotypes for all pairs of SNPs in the simulated region, where a high value of D' is indicative of a stronger association (low D' to high D' = light to dark). The dashed and solid lines mark the centers of each hotspot depicted in (a). Though marker* **B** *is in considerably closer physical proximity to the disease-causing SNP (×), it appears marker* **A** *has a higher value of D' with the disease-causing SNP (D' between the disease-causing location and each marker are highlighted with a red circle ◯). This is because marker* **B** *resides in a hotspot, which diminishes any associations between marker* **B** *and its surrounding locations. In contrast, marker* **A** *and the disease-causing location both fall between the two hotspots and thus appear to be more correlated to each other.*

This simple simulated example illustrates how leaving variable recombination rates unaccounted for can sometimes lead to dubious inference, in this case perhaps concluding that the disease-causing location is physically closer to one SNP when it is in reality closer to a different one.

msHOT: **Simulating Hotspots**

As recombination hotspots significantly influence patterns of LD and occur frequently across the genomes of humans and other organisms, it seems sensible to include them when simulating "realistic" genetic data to test new methods.

We have incorporated hotspots into the widely-used, haplotype simulator program ms developed by Hudson (2002). We call this updated simulator msHOT.

Features:

Both ms and msHOT simulate genetic variation data (SNPs) for chromosomes (haplotypes) randomly sampled from a population.

In generating these datasets, the user is allowed to control a variety of features:

- ms, msHOT:
  - demography (expansions, bottlenecks, etc.)
  - patterns of migration (admixture)
  - rates of mutation
  - rates of recombination (crossover, gene conversion)
- msHOT:
  - variable crossover and gene conversion rates across a region (hotspots)

Algorithm:

The basic algorithm of msHOT utilizes theoretical arguments described in Hudson (1983). In brief, the program generates ancestral recombination graphs for a sample of chromsomes by stochastically determining "events" to occur on the ancestral material of the chromosomes going back in time, until all the material has coalesced into a common ancestor. Potential "events" include the coalesence of two ancestral segments or a recombination event (crossover or gene conversion) occuring in a single ancestral segment. Incorporating hotspots involves changing the rates at which these recombination events occur. The basic features of ms remain intact.

Usage:

Incorporating H hotspots (these can be crossover and/or gene conversion hotspots) requires the user to specify a left endpoint (a_h), right endpoint (b_h), and intensity (λ_h) for each, $h = 1, ..., H$ (see Fig.3). All other input features are the same as in ms.

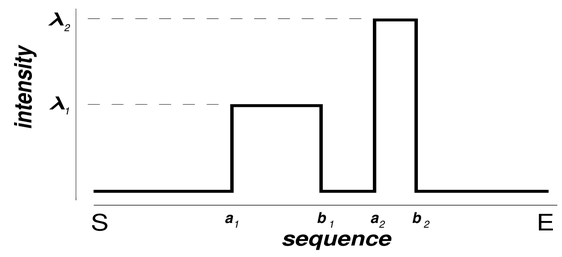


Figure 3: *Illustration of varying crossover and/or gene conversion intensities in a genetic region [S,E]. All variables $\vec{a}, \vec{b}, \vec{\lambda}$ are user-input. Outside any hotspot, the probability of a recombination event (either crossover or gene conversion) occuring between two adjacent basepairs in a single transmission from parent to offspring is a user-specified value x. Inside hotspot h, this probability is $\lambda_h x$. For example, for the simulated data used for Fig.2, which also considered two simulated hotspots, (λ_1,a_1,b_1) = (200,1000,2500) and (λ_2,a_2,b_2) = (300,8000,9000), for a 10kb sequence with $x \approx 10^{-8}$. (The same rates and hotspot parameters were used for crossover and gene conversion in this example, though this need not be the case.)*

Output:

The output from msHOT, as in ms, includes:

1. the SNP data (coded as **0**s and **1**s) for a user-specified number of haplotypes
2. the SNP locations in the genetic region (scaled to be between 0 and 1, with 0 the left endpoint of the region and 1 the right endpoint)

example of first 4 haplotypes (out of 50) for 50 SNP region simulated for Fig.2:

```
segsites: 50
positions: 0.0302 0.0834 0.0903 0.1171 0.1252 0.1283 0.1467 0.1551 0.1832 0.1881
 0.1892 0.1948 0.2116 0.2134 0.2715 0.2724 0.2793 0.3824 0.3851 0.3853 0.4021 0.
4058 0.4325 0.4426 0.4473 0.5008 0.5207 0.5414 0.5607 0.5645 0.6223 0.6254 0.697
4 0.7029 0.7146 0.7407 0.7735 0.8064 0.8232 0.8349 0.8433 0.8613 0.8729 0.8738 0
.8857 0.8896 0.9015 0.9048 0.9270 0.9987
00000101000000000100101100001000100110000010101100
00000000000000000100101100000000010001000011100000
00000000100001000100101100001000100000000010000000
00000000100000100100101100000000100000000010100000
```

Speed:

- msHOT can be used to generate a large number of haplotypes for moderately-sized regions
- Example: simulating 1000 haplotypes, 1000 SNPs, over 1 Mb region with average crossover rate = 1cM/Mb (with N_e=10000), no gene conversion, and 20 crossover hotspots of width 1kb and λ=100, using the standard demographic model, takes just a few minutes with msHOT on standard desktops
- Computation time is typically affected most by the total recombination rate, particularly the rate of gene conversion (is currently not suitable for entire human genome simulation)

Availability:

The source code for msHOT, along with accompanying instructions, is available **for free** by email from hellenth@stats.ox.ac.uk.

Selected References

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

Hudson, R.R.: Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18 2002, Nr. 2, 337–8

Jeffreys, A.J./Kauppi, L./Neumann, R.: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29 2001, Nr. 2, 217–22

Myers, S. et al.: A fine-scale map of recombination rates and hotspots across the human genome. Science 310 2005, Nr. 5746, 321–4