msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots

Garrett Hellenthal & Matthew Stephens

August 8, 2008

"msHOT"

This addition to Hudson's (2002) ms, called msHOT, allows for implementation of multiple crossover hotspots and/or multiple gene conversion hotspots in the simulated genetic region. Crossover hotspots may overlap with gene conversion hotspots, but crossover hotspots may not overlap with each other and gene conversion hotspots may not overlap with each other. After extracting from the .tar file, compile in the following way:

```
gcc -o msHOT msGCHOT.c streecGCHOT.c rand1.c -lm
                        or
gcc -o msHOT msGCHOT.c streecGCHOT.c rand2.c -lm
```

To generate hotspots, use the following switches:

- **-v n $\mathbf{a_1}$ $\mathbf{b_1}$ $\lambda_\mathbf{1}$ ... $\mathbf{a_n}$ $\mathbf{b_n}$ $\lambda_\mathbf{n}$**: The '-v' switch specifies the simulation of n crossover hotspots in the region of interest, the first one having intensity λ_1 (=0,...,∞) between basepairs a_1 and b_1 (input as integers), and the n^{th} one having intensity λ_n between basepairs a_n and b_n (with $1 \leq a_i < b_i \leq$ the total number of basepairs in the region). The *intensity* refers to the multiple of the "background crossover rate" by which the crossover rate is increased (or decreased, if you want a "coldspot") in the hotspot region. Note that unless you toggle the '-V' switch below, the rate of gene conversion is not any different in this region compared to outside it. Hotspot locations must be listed in order (i.e. $b_1 < a_2, b_2 < a_3, ..., b_{n-1} < a_n$).

- **-V m $\mathbf{a_1}$ $\mathbf{b_1}$ $\lambda_\mathbf{1}^{\text{GC}}$ ... $\mathbf{a_m}$ $\mathbf{b_m}$ $\lambda_\mathbf{m}^{\text{GC}}$**: The '-V' switch specifies the simulation of m gene conversion hotspots in the region of interest, the first one having intensity λ_1^{GC} between basepairs a_1 and b_1, and the m^{th} one having intensity λ_m^{GC} between basepairs a_m and b_m (again with $1 \leq a_i < b_i \leq$ the total number of basepairs in the region). *Intensity* has a similar definition

as before, only here referring to the multiple of the "background gene conversion rate." Note that unless you toggle the '-v' switch above, the rate of crossover is not any different in this region compared to outside it. Hotspot locations must be listed in order (i.e. $b_1 < a_2, b_2 < a_3, ..., b_{m-1} < a_m$). Again we note that crossover and gene conversion hotspots may overlap.

WARNING: Another important difference between ms and msHOT is that the assumed structure of gene conversion events is different. In Hudson's ms, a gene conversion event initiates at some basepair x and spreads, on average, t basepairs to the right (what we call t here is denoted as λ in Hudson's ms documentation), following a geometric distribution. That is to say, the expected *tract length* of gene conversion events is t. With msHOT, a gene conversion event initiates at some basepair x and spreads, on average, t^* basepairs to EACH the left and the right. Thus the total expected tract length is $2t^*$. The user specifies t^* in msHOT. Thus you expect the outcome of each gene conversion event to affect $2t^*$ basepairs.

Examples of usage

The basic usage for msHOT can be found in the documentation to ms by Hudson. The only addition is the incorporation of hotspots. As an example, to create one 25000 basepair sequence, with 10 haplotypes, 20 SNPs, and a background (across region) crossover rate $\rho = 10.0$, you would, as with ms, type the following:

```
msHOT 10 1 -r 10.0 25000 -s 20
```

To add two crossover hotspots, one between basepairs 100-200, in which the crossover rate is 10 times the background rate, and the other between basepairs 7000-8000, in which the crossover rate is 20 times the background rate, type the following:

```
msHOT 10 1 -r 10.0 25000 -s 20 -v 2 100 200 10 7000 8000 20
```

In this example, the "background crossover rate," i.e. the rate at which crossover occurs between two adjacent basepairs in the region, is $10.0/25000 = 0.0004$. (To be precise, one should actually use 25001 basepairs to get this value, as recombination only occurs *between* basepairs in msHOT and ms. However, using 25000 provides a very close approximation.) This is the rate at which crossovers occur between adjacent basepairs *outside* of any hotspot in the 25 kb region. In contrast, the crossover rate between adjacent basepairs from the 100th to the 200th basepair will be $(\lambda_1 * \rho)/25000 = (10 * 10.0)/25000 = 0.004$, 10 times that between basepairs outside of any hotspot region. Analagously, the crossover rate between adjacent basepairs from the 7000th to the 8000th basepair will be $(\lambda_2 * \rho)/25000 = (20 * 10.0)/25000 = 0.008$, 20 times that between basepairs outside of any hotspot region.

Incorporating gene conversion hospots is done similarly. For a background gene conversion rate that is 2 times that of the crossover rate $\rho = 10.0$, and with a total expected tract length of 200 basepairs, type the following:

msHOT 10 1 -r 10.0 25000 -c 2.0 100 -s 20

This differs from ms in that the mean tract length would be specified as 200 (e.g. "-c 2.0 200") in ms rather than as 100 (e.g. "-c 2.0 100") in msHOT to get an expected total tract length of 200 basepairs.

To add a gene conversion hotspot between basepairs 100 and 200 that is 10 times the background gene conversion rate, type the following:

msHOT 10 1 -r 10.0 25000 -c 2.0 100 -s 20 -V 1 100 200 10

In this example, the "background gene conversion rate," i.e. the rate at which gene conversion occurs between two adjacent basepairs in the region, is (2.0*10.0)/25000 = 0.0008. This is the rate at which gene conversions occur between adjacent basepairs *outside* of any hotspot in the 25kb region. In contrast, the gene conversion rate between adjacent basepairs from the 100th to the 200th basepair will be $\lambda_1^{GC} * c * \rho/25000 = 0.008$, 10 times that between basepairs outside of any hotspot region.

If you want a hotspot between 100 and 200 in this 25 kb region, such that it is a hotspot for both crossover *and* gene conversion, each with equal intensities of 10 (so that "**c**," equal to the relative rate of gene conversion to crossover, remains constant across this 25 kb region), type the following:

msHOT 10 1 -r 10.0 25000 -c 2.0 100 -s 20 -v 1 100 200 10 -V 1 100 200 10

Note that the crossover and gene conversion intensities need not be the same and that you do not need to have the same number of crossover and gene conversion hotspots in a region (nor at the same locations) as in the example above.

Note: msHOT, as in ms, outputs segregating site locations as fractions from 0 (start of genetic region) to 1 (end of genetic region). Hotspot locations are input as integer values, in relation to the specified sequence length (user-input with the "-r" switch). Thus to determine which segregating sites (if any) are in "hot" sequences, you must multiply the fractions by the user-input sequence length. Of course, you need not input the sequence length in terms of basepairs, as we have done here. For example, you may wish a sequence length of 25000 (e.g. "-r 10.0 25000") to refer to 250000 basepairs, implying each unit 1,...,25000 corresponds to 10 basepairs. In this case, be sure to scale your hotspot locations (a_i,b_i) accordingly.

Questions? Bugs? Contact Garrett Hellenthal at hellenth@stats.ox.ac.uk. Eric Anderson contributed to a lot of the hotspot code (though he should in no way

be blamed if something is wrong with it!). The core code of ms was written by Richard Hudson.

Bug Fixes

- **August 8, 2008** – fixed a bug that unequally distributed hotspot probabilities in a manner that reduced the amount of recombination that occured in the right end of the region when hotspots were present. The bug was such that one can quantify exactly how many basepairs in the right side of the region would be affected. Specifically, for a region with L total basepairs and H hotspots, with hotspot h=1,..,H, beginning at basepair a_h, ending at basepair b_h, and having intensity λ_h (where $a_1 < ... < a_H$), the error prohibited recombination from occurring after each basepair $x \in (1, L-1)$ for which:

$$\sum_{z=1}^{x} \big(1 + (\lambda_h - 1) I_{z \in [a_h, b_h]}\big) > \big(L + \sum_{h=1}^{H} (\lambda_h - 1)(b_h - a_h)\big).$$

  This could have a noticeable effect if one generated a relatively strong hotspot in a fairly small region using the old code. The bug could furthermore have shifted the locations of any hotspot located to the right of the leftmost hotspot; this is most likely to have a discernable effect when hotspots were simulated to be close together and the left-most hotspot(s) was intense relative to hotspots to its right. In practice we have noticed no strong effects (e.g. on patterns of linkage disequilibrium) of this bug, for example in simulations with two 1-1.5kb recombination (crossover and gene conversion) hotspots, separated by only 5.5kb in a 10kb region, with background crossover rate 0.4/kb and intensities 200 and 300 times the background rate for the left and right hotspots, respectively. However, intensity values more extreme than this, perhaps in a smaller region, could potentially have significantly influenced the resulting linkage disequilibrium structure in the region. Please email with any specific questions.

Citations

When publishing research that has used msHOT, the appropriate citations are:

1. Hellenthal, G. and M. Stephens, 2006 msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, to appear.

2. Hudson, R.R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:**337-338.