

Peter Calabrese
petercal@usc.edu
January 12, 2007

This C++ code has been developed on a computer running the UNIX operating system, and these instructions apply to such a system. Download the files "heterohot.h" and "heterohot.cpp" and put them in the same directory. At the command line type,

```
g++ -o heterohot heterohot.cpp
```

this creates an executable file called "heterohot". Here then is an example command,

```
.\ heterohot 1234 100 10000 .30 30 .001 .0005 .001 1 100 200 50150
```

And here are an explanation of these parameters, in order,

1. 1234 is a random number seed: different numbers give different seeds.
2. 100 is the number of haploids in the sample.
3. 10000 is the diploid population size.
4. .30 is the frequency of the hotspot in the current population (the frequency in the past is stochastic conditioned on it originating on a single chromosome at some point in the past).
5. 30 is the number of haploids with the hotspot in the sample.
6. .001 is the mutation rate per haploid per generation.
7. .0005 is the double strand break probability per haploid per generation: the location of such a possible break is uniform in the haploid's length (for a haploid not harboring the hotspot motif the recombination probability is two times this probability, since a break in either of the two haploid copies results in a recombination event).
8. .001 is the additional probability of a break per haploid per generation if the haploid harbors the hotspot motif: the location of this break is in the center of the haploid.
9. 1 is the probability a double strand break is resolved as a crossover (it is resolved as a conversion with probability one minus this parameter).
10. The simulated fragment has fixed length 100,000 bases (for other lengths just scale the parameters appropriately). 100, 200 for each double strand break event select a uniform random number between these two parameters: this number of bases is then lost (and subsequently copied from another haploid) from each side of the break.

11. 50150 this is the location of the hotspot motif (the location of the break caused by this motif is fixed at position 50000).

The output is the number of segregating sites, positions of these sites (total fragment is 1 – 100,000), and the haplotypes of the sample.

Please reference the PNAS paper, "A Population Genetics Model with Recombination Hotspots that are Heterogeneous Across the Population" by Peter Calabrese.