Directions for simulated sequence generation for MASSPRF analysis

Requires ms/msHOT, Indelible, MASSPRF, and MATLAB.

Please forward any issues to Daniel Lee at [dslee.sm@gmail.com](mailto:dslee.sm@gmail.com).

1. **Determine ms parameters and generate tree.**

General form of input:

./ms.exe nsam nreps -T -t θ -I npop n1 n2 -ej t i j -en t i x –r ρ nsites > outfilename

nsam=101 copies of each locus per tree (to generate 1 divergent and 100 polymorphic sequences)

nreps=20 replicates

-T command instructs ms to output trees (Newick format)

-t indicates the next input is mutation parameter θ≡4N0μ

Where N0≡diploid population size =104 μ≡neutral mutation rate for the entire locus ≈10-8 per generation per site(Lynch 2009) \* a simulated gene length of 900.

In this simulation, θ=4(104)(10-8)(900)=**.36.**

-I indicates “island (isolated) model” of subpopulations.

npop=2 populations (early speciation event)

n1=1 set of loci in population 1 (divergent sequence)

n2=100 set of loci in population 2 (polymorphic sequences)

-ej indicates separation event; move all lineages in subpopulation i to subpopulation j at time t (which is in units of 4N0 generations; we assume 1 generation is about 25 years for a human population).

We set speciation event to be at 6mya = **7.5**\*25\*4\* 104, wherein all populations from population 2 (the one comprising 100 species) merges with population 1 6mya, i.e., splits 6mya. (Recall that all ms events are retroactive).

Thus t = 7.5 ; i = 2; x=1.

-en indicates a demographic size change event (i.e., bottleneck or expansion); such that before time t (again expressed in units of 4N generations) population i was x fold times that of N0 (**see page 11 of the ms documentation for more detail; it is self-contradictory as to whether the population before t is 4xN0 or xN0)**.

Thus, we model a demographic event .1\*4\*104\*25=100kya occurring in population 2 (the one comprised of 100 loci).

In expansion, we claim that the population before the event was .2 times what it is currently, i.e., expanded from 2000 to 10000.

In bottleneck, we claim that the population before the event was 40000 (i.e., 4 times greater).

This section is omitted otherwise.

-r indicates the inclusion of recombination/crossing over events.

ρ=4N0r where r is the probability of cross over per generation between the ends of the loci being simulated; used ρ=0.6 based on Wang & Rannala [2009], who also had Ne=104 .(<http://www.pnas.org/content/106/15/6215.full>).

Nsites=number of loci in gene between which recombination can occur, i.e., all of them, so this is equal to gene length, which is 900 here.

Thus the following parameters were to simulate bottleneck, expansion, and constant population with and without recombination, assuming N0=104 and gene length =900 bp:

Expansion

./ms.exe 101 20 -T -t 0.36 -I 2 1 100 -ej 7.5 2 1 -en .1 2 .2 >Expan\_NR.out

Bottleneck

./ms.exe 101 20 -T -t 0.36 -I 2 1 100 -ej 7.5 2 1 -en .1 2 4 >Bottle\_NR.out

Constant (no demographic event)

./ms.exe 101 20 -T -t 0.36 -I 2 1 100 1 -ej 7.5 2 1 >Const\_NR.out

Expansion +recombination

./ms.exe 101 20 -T -t 0.36 -I 2 1 100 -ej 7.5 2 1 -en .1 2 1 -r .06 900 >Expan\_RE.out

Bottleneck +recombination

./ms.exe 101 20 -T -t 0.36 -I 2 1 100 -ej 7.5 2 1 -en .1 2 .2 -r .06 900 >Bottle\_RE.out

Constant (No demographic event) +recombination

./ms.exe 101 20 -T -t 0.36 -I 2 1 100 1 -ej 7.5 2 1 -r .06 900 >Const\_RE.out

1. **Using script ‘mstreeread.m’ load in tree file and output control file.** 
   1. If using parameters including recombination, use the SECOND part of mstreeread
   2. If not including recombination in your ms tree generation, use the FIRST part.
2. **Manually input model parameters into control text file and run Indelible manually in /bin/.**

Model used:

Neutral: M1

2.5 .743 0.5 1 (kappa p0 w0 w1, based on citation  μr*/*μs default = .345 [(Nei and Gojobori, 1986)])

Since p0 and p1 are proportions of ALL codon substitutions that are respectively unconstrained and constrained (i.e., synonymous or replacement) we simply calculate that since μr*/*μs =.345, p0 = 1/(.345+1)=.743.

Positive selection: M2 model

2.5 .212 .2565 0.5 1 2 (p0 p1; w0 w1 w2; p2=.531=1-p0-p1)

1. **Read Indelible outputs.** 
   1. If including recombination in the simulation, run those files (manually inputting their heading, in this case “Expan\_RE”) through the ‘Recombconcat.m’ to output ‘.fas’ files. Proceed to step 4b.
   2. Run the .fas outputs through the script ‘indelibletofas.m’, which will simply separate the larger fas files into pol and div files.
   3. Use the second portion of ‘indelibletofas’ to generate a .pbs file for use on the cluster.
2. **Run MASSPRF as normal using the pol and div files from step 4.**