MMC-ABC Manual

I. Introduction

This manual gives instructions for the use of the software package Multiple-Merger Coalescent Approximate Bayesian Computation (MMC-ABC). The software is designed to jointly infer genome-wide values of population size (N) and ψ with site-specific selection coefficients (s). The sample data consist of time-sampled allele frequencies. The analysis is performed with SLiM version 3 (Haller and Messer 2018), and requires that this software be installed prior to use.

The specifics of MMC-ABC are described in detail by Sackman, Harris, and Jensen (*in review*). Briefly, the ABC runs in a two-step process. In Step 1, it simulates populations with the same initial mutational frequencies as in the sample data with N and ψ drawn from their priors. A summary statistic, Jordy and Ryman's (2008) unbiased estimator of N_e, is measured for the sample data and each simulated population. The top 1% of simulations are retained to generate a joint posterior for N and ψ.

In Step 2, for each mutation in the sample data, populations of size N and ψ drawn from the joint posterior are initiated with a mutation of effect size s beginning at the observed initial frequency. A posterior for s is generated for each site using the summary statistics F_{si} and F_{sd}.

For the diploid model, we define relative fitness as $w_{AA} = 1+s$, $w_{Aa} = 1+sh$ and $waa = 1$ where h denotes the dominance ratio (1 = dominant, 0.5 = codominance, 0 = recessive). For the haploid model, we define relative fitness as $w_A = 1+s$ and $wa = 1$.

II. Data Format

Included in this package is a sample data file, test_data.txt. The data are in the format required for correct parsing and analysis by MMC-ABC. An example is shown here:

```
1000 21
0,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,95,100,
250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,
63,45,26,32,14,12,11,4,16,10,7,6,3,2,1,0,0,0,0,0,0,
250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,
51,91,123,120,124,156,154,186,144,123,92,144,133,106,101,123,109,147,165,134,139,
250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,
16,38,38,64,53,43,35,37,38,19,55,33,52,36,42,46,45,31,32,35,47,
250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,
226,219,220,197,192,212,215,216,221,224,183,205,188,204,205,204,213,232,224,205,214,
250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,250,
151,153,139,114,123,152,104,125,126,119,92,112,123,99,98,130,114,140,164,153,145,
```

The first line gives the number of loci (1000) and the number of time points (21). The second line indicates the time of the 21 time points, measured in generations. The software uses the difference in generations between each time point, so giving generations of 1, 11, 21, is effectively the same as 101, 111, 121.

The rest of the file contains two lines for each of the 1000 loci. The first line gives the sample size for each locus at each time point. The second line gives the frequency of the mutation at each time point (as a count out of the total sample size). For example, for the first site in our test data, the mutation began at a frequency of 63 out of 250 at generation 0, and had frequency 45/250 at generation 5. Also, each site need not have a non-zero frequency at the first time point. For example, the last several sites in the provided test data have a count of 0 for the first time point.

Each value in the line of generations and in the lines of sample sizes and counts should be followed by a comma, including the final value in each line. Additionally, there should be an empty line at the end of the text file.

Note: MMC-ABC requires at least three non-zero, un-fixed time points at each site (i.e., there must be a non-zero and un-fixed value for each site by at least the third-to-last time point, or else the software will not function properly). This is because a reliable estimate of *s* relies upon having sufficient data for comparing the observed data with simulated data.

III. Basic usage of MMC-ABC

MMC-ABC is run from the command line using Python 2.7 and is not compatible with Python 3. It additionally requires that SLiM 3 be installed (Haller and Messer 2018), as well as the Python packages scipy and numpy.

Your data file should be moved to the directory containing the MMC-ABC files.

A full list of command line options for MMC-ABC is given further below. Here, we will demonstrate a worked example using the provided test_data.txt file.

```
python MMC-ABC.py --part1sims 100000 --part2sims 10000 --numthreads 4 --n_min 250 --n_max 2000 test_data.txt
```

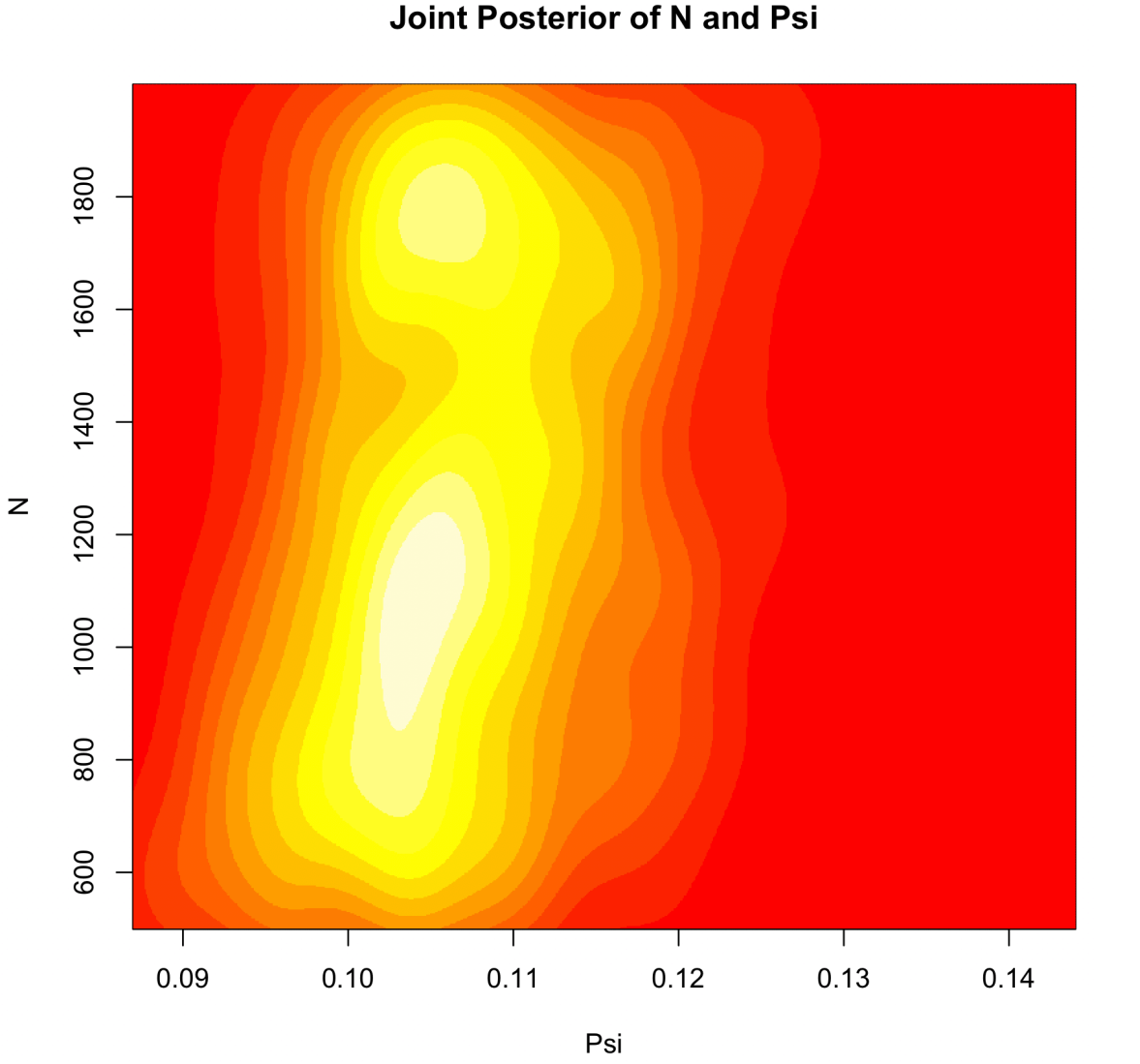
This command runs MMC-ABC for the file 'test_data.txt', with 100,000 simulation runs for Part 1 and 10,000 for Part 2, run over 4 threads. We use a uniform prior for *N* of $\sim U[250, 2000]$ (though for this test data, the population is known to be a diploid population of *N*=1000). We use the default priors of $\sim U[0, 0.3]$ and $\sim U[-0.2, 0{,}6]$ for ψ and *s*, respectively.

Upon running this command, the program will output to the console, "Beginning Part 1 of MMC-ABC: Estimation of N and Psi." It will generate temporary text files, which are outputs from SLiM and are subsequently used by ABC. After Part 1 is finished and posteriors for *N* and ψ have been generated, it will output the means of those posteriors to the console, before printing: "Beginning part two of MMC-ABC: Estimation of site-specific selection coefficients using the joint posterior of N and Psi." The program will output the number of each site as it progresses through Part 2.

Depending on the number of sites and the number cores being used, Part 1 will require anywhere from a few minutes to a few days to complete. In our experience, the program generates very accurate estimates of *N* and ψ even with 1,000 replicate simulations. Part 2 requires a similar amount of time, depending on the number of replicates used for the ABC.

The joint posterior of N and ψ is output to a file named '[data filename]_N_psi_joint_posterior.csv'. For our test data, the code above generates a joint posterior for N and ψ with means of 1241.04 and 0.108, respectively. We can plot the joint posterior in R using the following code:

```
> require(MASS)
Loading required package: MASS
> x=read.csv('test_data_N_psi_joint_posterior.csv')
> attach(x)
> z<-kde2d(Psi,N,n=1000)
> image(z,xlab='Psi',ylab='N',main="Joint Posterior of N and Psi")
```



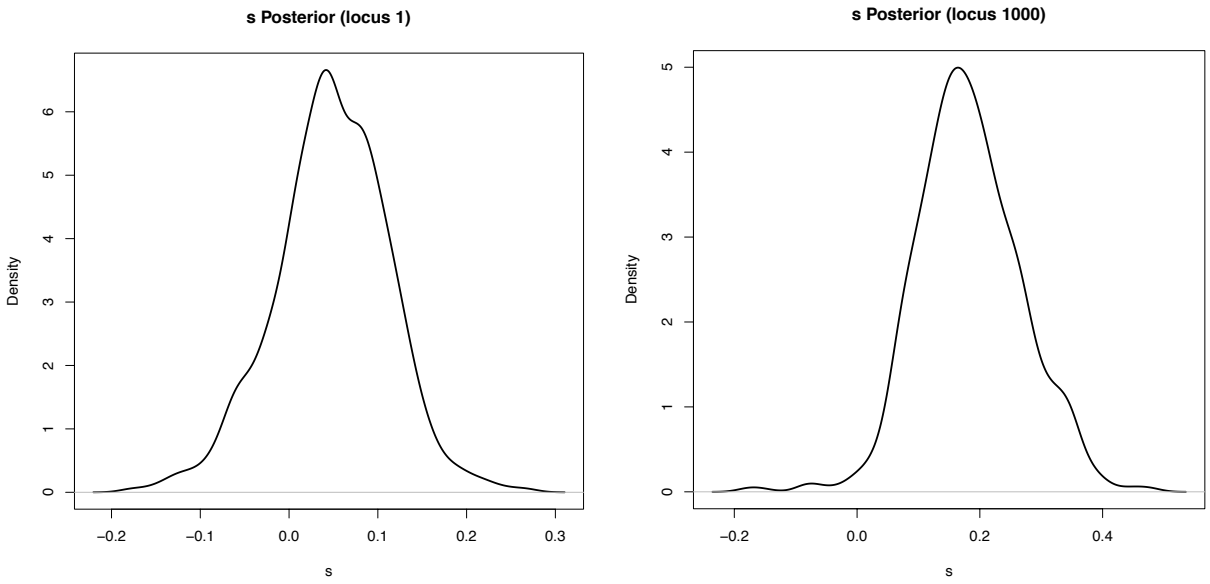
Due to the non-linear scaling of N and N_e under sweepstakes reproduction, the posterior covers a broad range of N and a narrow range of ψ, with the posterior for N centered at the true value of N = 1,000. The posterior of N covers a much more narrow range in cases where ψ is low.

Estimates of site-specific selection coefficients are output to a file named "s_outputs.txt" with the mean of the posterior for each site output on a separate line. The posteriors are output into a

file named '[data filename]_s_posteriors.csv'. We can plot the output for particular sites in R via the following code:

```
> x=read.csv('test_data_s_posteriors.csv')
> plot(density(t(Site_1)),lwd=2,main="s Posterior (locus 1)",xlab="s")
> plot(density(t(Site_1000)),lwd=2,main="s Posterior (locus 1000)",xlab="s")
```

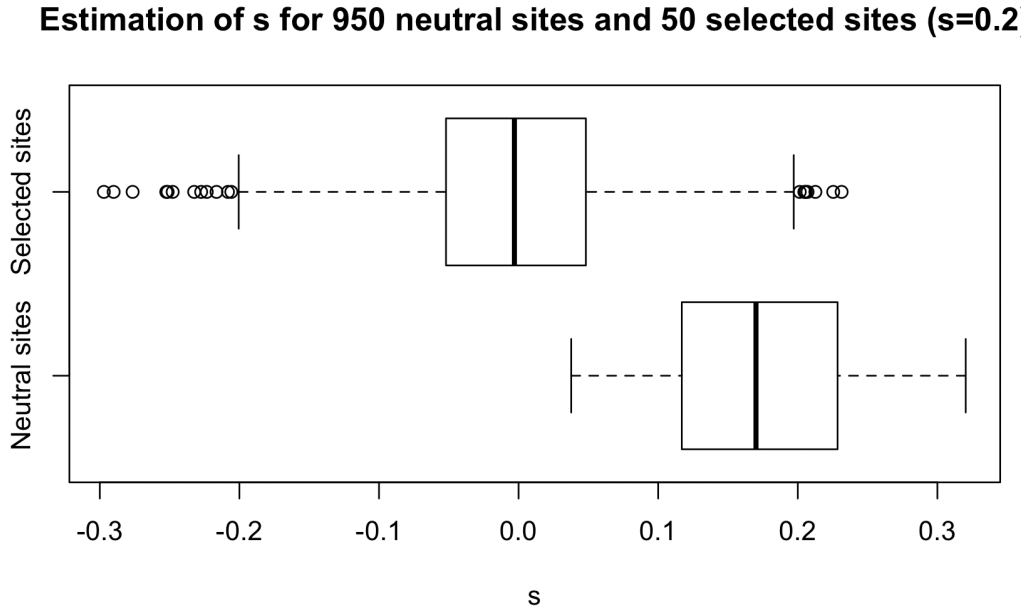
The first 950 sites in the test data are neutral, and the last 50 sites are under selection with $s = 0.2$ (1,000 sites total). Plotting the posteriors for the first and last site in the data yields the following:



We can also take a look at estimates of s for all sites in this way:

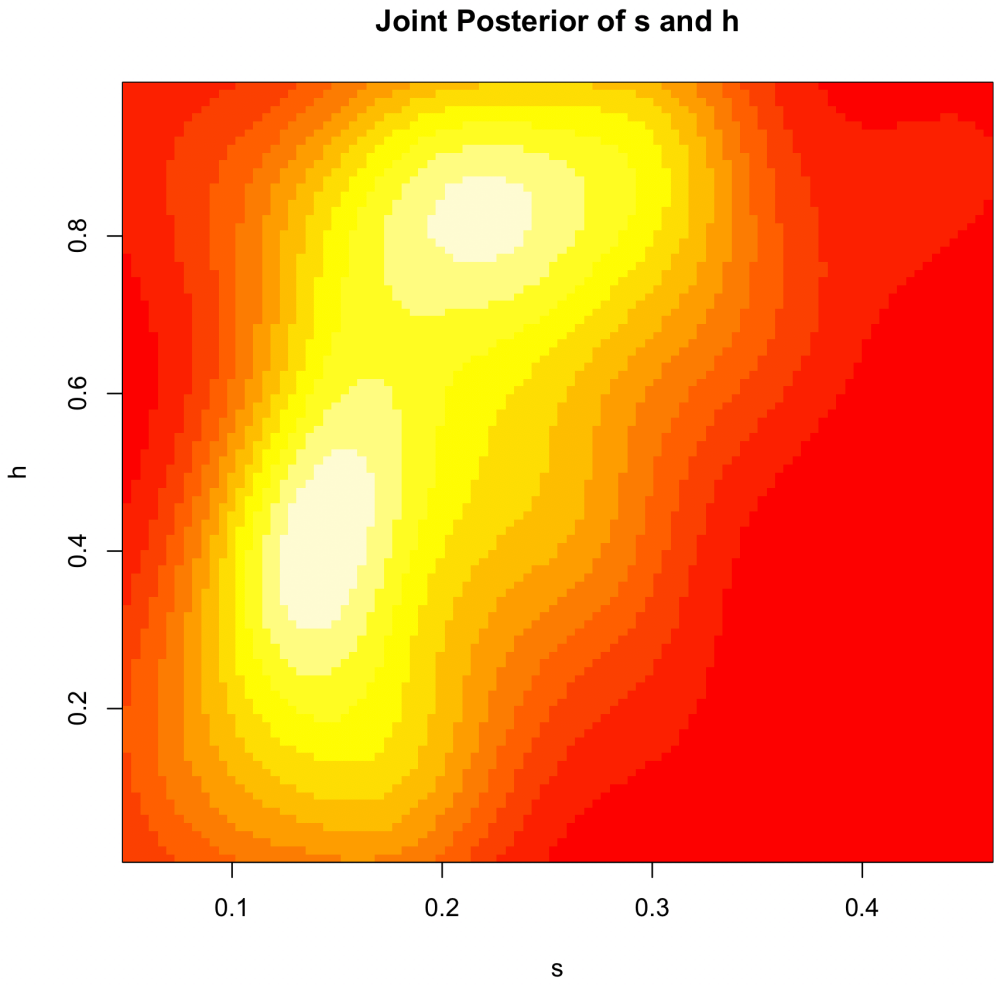
```
> x=read.csv('test_data_s_posteriors.csv')
> boxplot(colMeans(x[951:1000]),colMeans(x[1:950]),horizontal=T,xlab='s',names=c('Selected
sites','Neutral sites'),main='Estimation of s for 950 neutral sites and 50 selected sites (s=0.2)')
```

Which produces the following plot:



By default, diploid loci are assumed to have heterozygosity $h = 0.5$. However, MMC-ABC can accommodate an alternative value of *h,* specified as --h [value of *h*]. MMC-ABC can also jointly infer *h* and *s* for each site. To do this, add --h_yes 1 --h_min [low end of prior] --h_max [upper end of prior] to the command. All haploids have heterozygosity $h = 1$.

The joint posteriors of *h* and *s* can be plotted for each site in a manner similar to the joint posterior of *N* and ψ. Below, $s = 0.2$ and $h = 0.5$:



IV. List of available arguments:

All arguments should start with '--' and should be followed by the specified value. For example, to specify the lower end of the posterior for *N*, add '--n_min 100' to the command line. The name of the data file (test_data.txt in our example) should be at the end of the command.

Sample command: `python MMC-ABC.py --ploidy 1 --numthreads 12 test_data.txt`

--ploidy	1 = haploid, 2 = diploid (default = 2)
--n_min	Lower bound of prior for *N* (default = 500)
--n_max	Upper bound of prior for *N* (default = 10,000)
--psi_min	Lower bound of prior for ψ (default = 0.0)
--psi_max	Upper bound of prior for ψ (default = 0.3)
--s_min	Lower bound of prior for *s* (default = -0.2)
--s_max	Upper bound of prior for *s* (default = 0.6)
--h_yes 1	Jointly infer *h* with *s* (must include '1' after --h_yes)
--h_min	Lower bound of prior for *h* (only for diploids, default = 0)
--h_max	Upper bound of prior for *h* (only for diploids, default = 1)
--part1sims	Number of replicate simulations for Part 1 (default = 10000)
--part2sims	Number of replicate simulations for Part 2 (default = 10000)
--n_apriori	Specifies value for *N* (instead of uniform prior)
--psi_apriori	Specifies value for ψ (instead of uniform prior)
--h	Specifies value for *h* (only for diploids, default = 0.5)
--recomb	Specifies recombination rate for Part 1 (default = 1e-8)
--numthreads	Specifies number of computing threads

V. References

Haller, B. C. and P. W. Messer. 2018. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. bioRxiv doi: 10.1101/418657

Jorde, P. E. and N. Ryman, 2007 Unbiased estimator for genetic drift and effective population size. Genetics 177: 927–935.

Sackman, A. M., R. B. Harris, and J. D. Jensen. 2018. Inferring demography and selection in organisms characterized by skewed offspring distributions. *In review.*