User Manual for *MULLET* v1.0

Xiaoheng Cheng, Michael DeGiorgio

May 11, 2018

1. Introduction

MULLET is a program to perform genome-wide scans of ancient trans-species balancing selection without the inclusion of trans-species polymorphisms by using the spatial distribution of within-species polymorphism and between species substitution in the genome. The current version of the software implements the $T_{1,\text{trans}}$ and $T_{2,\text{trans}}$ test statistics of Cheng and DeGiorgio (2018).

2. Installation

MULLET should run on any UNIX system. It requires that the GNU Scientific Library be installed (http://www.gnu.org/s/gsl/) to running the following compilation steps. To extract and compile the program, enter on the command line:

```
>tar -xzvf mullet_v1.0.tar.gz
>cd mullet_v1.0
>gcc mullet_v1_0.c -o MULLET -lm -lgsl -lgslcblas
```

3. Input file format

3.1. Unrooted tree configuration file

The unrooted tree configuration file summarizes the topology and branch lengths in coalescent time units (generations divided by twice the effective population size), and is tab-delimited. For an unrooted bifurcating tree of K species, there are a total of $2K-3$ branches. That is, for $K=2$, there is a single branch separating the two species, for $K=3$, there are three branches, and for $K=4$, there are five branches. Figure 1 depicts examples for two, three, and four-species trees.

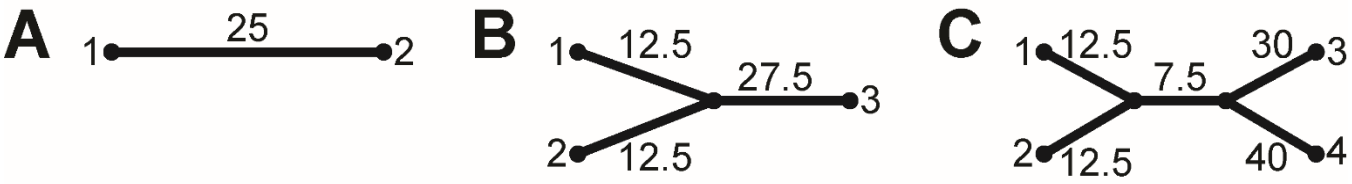

Figure 1: Example unrooted bifurcating trees for (A) $K=2$, (B) $K=3$, and (C) $K=4$ species.

Each of these $2K-3$ branches indicates a location in which a single mutation could be included on a tree that would lead to an informative site within the set of sampled species. That is, in Figure 1A, any mutation along the branch would either lead to a substitution between species 1 and 2, or a polymorphism in species 1 or 2. For Figure 1C, a single mutation on any branch would lead to a substitution between one set of species and another, while a mutation on any of the external branches would lead to a within-species polymorphism.

In general, for K species, the unrooted tree configuration file will have $2K-3$ rows (representing each branch of the tree that a single mutation can arise), and $K+1$ columns, where columns 1 through K represent species 1 through K, respectively, and column $K+1$ represents the length in coalescent time units for the given branch (*i.e.*, row). For the first K columns, we use 0s and 1s to indicate a split partition of the given branch, such that species indicated by 0 are in one partition and species indicated by 1 are in another partition. Moreover, the first set of rows will represent external branches, and the remainder will represent internal branches.

Consider the different K-species bifurcating unrooted trees in Figure 1. An example input unrooted tree configuration file for $K=2$ species using the tree from Figure 1A is

1 0 25

Here, there is a single branch of length 25 coalescent units separating out species 1 and 2, and therefore the file only has a single row. Therefore, there is only one partition that is consistent with this tree, such that species 1 belongs to one part of the partition (indicated here by a 1), and species 2 belongs to the other part of the partition (indicated here by a 0).

An example input unrooted tree configuration file for $K = 3$ species using the tree from Figure 1B is

1 0 0 12.5
0 1 0 12.5
0 0 1 27.5

Here, there are three branches with lengths 12.5, 12.5, and 27.5 coalescent units, separating out species 1, 2, and 3 with the remainder of the other species, respectively. Because there are three branches, the file has three rows. The first row is a partition of species 1 (here indicated by 1) with species 2 and 3 (here indicated by 0), and represents the branch leading to species 1. The second row is a partition of species 2 (here indicated by 1) with species 1 and 3 (here indicated by 0), and represents the branch leading to species 2. The third row is a partition of species 3 (here indicated by 1) with species 1 and 2 (here indicated by 0), and represents the branch leading to species 3.

An example input unrooted tree configuration file for $K = 4$ species using the tree from Figure 1C is

1 0 0 0 12.5
0 1 0 0 12.5
0 0 1 0 30
0 0 0 1 40
1 1 0 0 7.5

Here, there are five branches with lengths 12.5, 12.5, 30, 40, and 7.5 coalescent units, where the first four branches are external branches separating out species 1, 2, 3, and 4 with the remainder of the other species, respectively, and the fifth branch is an internal branch separating out species 1 and 2 from species 3 and 4. Because there are five branches, the file has five rows. The first row is a partition of species 1 (here indicated by 1) with species 2, 3, and 4 (here indicated by 0), and represents the external branch leading to species 1. The second row is a partition of species 2 (here indicated by 1) with species 1, 3, and 4 (here indicated by 0), and represents the external branch leading to species 2. The third row is a partition of species 3 (here indicated by 1) with species 1, 2, and 4 (here indicated by 0), and represents the external branch leading to species 3. The fourth row is a partition of species 4 (here indicated by 1) with species 1, 2, and 3 (here indicated by 0), and represents the external branch leading to species 4. The fifth row is a partition of species 1 and 2 (here indicated by 1) with species 3 and 4 (here indicated by 0), and represents the internal branch separating species 1 and 2 from species 3 and 4.

3.2. Frequency file

The allele frequency input file is tab-delimited. Each row represents the within-species polymorphism or between-species substitution status for a position in the genome that is consistent with a given input unrooted tree configuration (section 3.1), and the rows are ordered by increasing position along the chromosome. There should be a separate polymorphism and substitution file for each chromosome when performing a scan for balancing selection. For K-species, $K = 2, 3, \ldots$, there are a total of $2K + 2$ columns in the input file. At each row, column 1 is the position on the chromosome, column 2 is the population-scaled recombination rate ($\rho = 2N_e r$) relative to the prior row, column $2k + 1$, $k = 1, 2, \ldots, K$, is the ancestral allele count (a_k) for species k,

and column $2k+2$, $k = 1, 2, \ldots, K$, is the sample size (n_k) of species k. Note that alleles must be polarized using an outgroup, and the ancestral allele count for species k, $k = 1, 2, \ldots, K$, can take on values $a_k = 1, 2, \ldots, n_k - 1$ for polymorphic sites, and $a_k = 0$ in some species and $a_k = n_k$ for other species at sites that are substitutions. Note that the methods of Cheng and DeGiorgio (2018) do not consider sites in which $a_k = 0$ or $a_k = n_k$ for all species k, $k = 1, 2, \ldots, K$. Furthermore, all substitutions must be consistent with the unrooted tree defined by the configuration file (section 3.1). For the first position in the input file, the population-scaled recombination rate is $\rho = 0$. An example input file with $K = 4$ species consistent with the unrooted tree in Figure 1C is

```
460000      0.0   9     100   100   100   100   100   100   100
460010      0.002 100   100   0     100   100   100   100   100
460210      0.04  100   100   100   100   100   100   30    78
463000      0.558 100   100   94    100   100   100   100   100
473700      0.14  100   100   100   100   0     100   0     100
474000      0.06  0     100   0     100   100   100   100   100
478020      0.804 0     100   100   100   100   100   100   100
480000      0.396 100   0     100   0     100   100   100   100
…           …     …     …     …     …     …     …     …     …
```

Each row displays polymorphism and substitution data for positions on a chromosome (positions 460000, 460010, 460210, 463000, 473700, 474000, 478020, and 480000) as well as the population-scaled recombination rates ($\rho = 2N_e r$) for each position and the preceding position in the file (0.0, 0.002, 0.04, 0.558, 0.14, 0.06, 0.804, and 0.396). Each row indicates the number of ancestral alleles observed (and the total number of alleles observed) at the given chromosomal position in each of species 1, 2, 3, and 4. At position 460000, 9 ancestral alleles were observed out of 100 total alleles (50 diploid individuals) for species 1, and 100 ancestral alleles were observed out of 100 total alleles in each of species 2, 3, and 4, leading to an observed within-species polymorphism in species 1. At position 460010, no ancestral alleles were observed out of 100 total alleles in species 2, and 100 ancestral alleles were observed out of 100 total alleles in each of species 1, 3, and 4, leading to an observed substitution on the external branch leading to species 2. At position 460210, 30 ancestral alleles were observed out of 78 total alleles (39 diploid individuals) in species 4, and 100 ancestral alleles were observed out of 100 total alleles in each of species 1, 2, and 3, leading to an observed within-species polymorphism in species 4. At position 463000, 94 ancestral alleles were observed out of 100 total alleles in species 2, and 100 ancestral alleles were observed out of 100 total alleles in each of species 1, 3, and 4, leading to an observed within-species polymorphism in species 2. At position 473700, no ancestral alleles were observed out of 100 total alleles in each of species 3 and 4, and 100 ancestral alleles were observed out of 100 total alleles in each of species 1 and 2, leading to an observed substitution on the internal branch separating species 1 and 2 from species 3 and 4. At position 474000, no ancestral alleles were observed out of 100 total alleles in each of species 1 and 2, and 100 ancestral alleles were observed out of 100 total alleles in each of species 3 and 4, leading to an observed substitution on the internal branch separating species 1 and 2 from species 3 and 4. At position 478020, no ancestral alleles were observed out of 100 total alleles in species 1, and 100 ancestral alleles were observed out of 100 total alleles in each of species 2, 3, and 4, leading to an observed substitution on the external branch leading to species 1. At position 480000, no ancestral alleles were observed out of 100 total alleles in each of species 1 and 2, and 100 ancestral alleles were observed out of 100 total alleles in each of species 3 and 4, leading to an observed substitution on the internal branch separating species 1 and 2 from species 3 and 4.

4. Helper files (required before using `-T1trans` and `-T2trans` options)

It is necessary to generate these files prior to using the `-T1trans` and `-T2trans` options to scan for trans-species balancing selection. For all helper files, it is required that the user first combined their frequency files

into a single frequency file with the same format as the example in section 3.2, and we will refer to this file as `CombinedSNPFile`.

4.1. Generating configuration file (`-config`)

To compute the proportion of within-species polymorphisms and between-species substitutions consistent with a given bifurcating unrooted tree, use the `-config` option. The command is:

```
./MULLET -config K TreeFile CombinedSNPFile ConfigFile
```

where `K` is the number of species K, $K = 2,3, \ldots$, `TreeFile` is the unrooted tree configuration file (section 3.1), `CombinedSNPFile` is a frequency file (section 3.2) combined across all chromosomes in the analysis (to get a genome-wide estimate), and `ConfigFile` is the name of a file where the results will be printed.

4.2. Generating empirical frequency spectra (`-spect`)

To compute the marginal empirical frequency spectra for the set of K species, use the `-spect` option. The command is:

```
./MULLET -spect K CombinedSNPFile SpectFile1 SpectFile2 … SpectFileK
```

where `K` is the number of species K, $K = 2,3, \ldots$, `CombinedSNPFile` is a frequency file (section 3.2) combined across all chromosomes in the analysis (to get a genome-wide estimate), `SpectFile1` is the name of a file where the results will be printed for species 1, `SpectFile2` is the name of a file where the results will be printed for species 2, and `SpectFileK` is the name of a file where the results will be printed for species K.

5. Scanning for ancient trans-species balancing selection

5.1 Scan using $T_{1,\text{trans}}$ test statistic (`-T1trans`)

To perform a scan for ancient balancing selection with the $T_{1,\text{trans}}$ statistic of Cheng and DeGiorgio (2018), use the `-T1trans` option. The command to perform this scan is

```
./MULLET -T1trans W K TreeFile ConfigFile SNPFile OutFile
```

where `W` is a user-defined window size (W polymorphisms or substitutions directly upstream of a test site and W polymorphisms or substitutions directly downstream of a test site) to compute the test statistic, `K` is the number of species K, $K = 2,3, \ldots$, `TreeFile` is the unrooted tree configuration file (section 3.1), `ConfigFile` is an input file containing the proportion of within-species polymorphisms and between-species substitutions consistent with the given bifurcating unrooted tree calculated using the `-config` option in section 4.1, `SNPFile` is the frequency file (section 3.2), and `OutFile` is the name of a file where the results will be printed. Here, the `SNPFile` would be for a specific chromosome (or region of the genome) rather than combined across all chromosomes.

5.2. Scan using $T_{2,\text{trans}}$ test statistic (`-T2trans`)

To perform a scan for ancient balancing selection with the $T_{2,\text{trans}}$ statistic of Cheng and DeGiorgio (2018), use the `-T2trans` option. The command to perform this scan is

```
./MULLET -T2trans W K TreeFile ConfigFile SpectFile1 … SpectFileK SNPFile
     PATH OutFile
```

where W is a user-defined window size (W polymorphisms or substitutions directly upstream of a test site and W polymorphisms or substitutions directly downstream of a test site), K is the number of species K, $K = 2,3, \ldots$, TreeFile is the unrooted tree configuration file (section 3.1), ConfigFile is an input file containing the proportion of within-species polymorphisms and between-species substitutions consistent with the given bifurcating unrooted tree calculated using the –config option in section 4.1, SpectFile1 through SpectFileK are input files containing the marginal ancestral frequency spectra for species 1 through K, respectively, calculated using the –spect option in section 4.2, SNPFile is the frequency file (section 3.2), PATH is the path to the directory containing the simulated frequency spectra under a model of balancing selection (included in simulated_spectra.tar.gz), and OutFile is the name of a file where the results will be printed. Here, the SNPFile would be for a specific chromosome (or region of the genome) rather than combined across all chromosomes.

To extract the simulated spectra, type the command

```
tar –xzvf simulated_spectra.tar.gz
```

The directory structure for the decompressed folder has the form

```
simulated_spectra/nX/
```

for $X = 2, 3, \ldots, 200$, where nX is the directory containing simulated spectra for a sample of size X. When using the –T2trans option, the PATH variable should be set to

```
PATH = /[paths]/simulated_spectra/
```

For example, suppose the simulated_spectra.tar.gz file was extracted to directory /home/user/project/. Then the appropriate command would look like

```
./MULLET –T2trans W K TreeFile ConfigFile SpectFile1 … SpectFileK SNPFile
     /home/user/project/simulated_spectra/ OutFile
```

where in this example we set the PATH variable to

```
PATH = /home/user/project/simulated_spectra/
```

The program will now know to look for simulated frequency spectra in /home/user/project/simulated_spectra/.

Note: Over time we will update the simulated_spectra.tar.gz file with larger sample sizes. Currently, the maximum is 200 alleles (100 diploid individuals).

6. Output file format

The output file is tab-delimited. Each row represents the calculation of either the $T_{1,\text{trans}}$ or $T_{2,\text{trans}}$ test statistic at an informative site (within-species polymorphism and between-species substitution). The first column is the position on the chromosome of the test site, and the second column is the value of the $T_{1,\text{trans}}$ or $T_{2,\text{trans}}$ test statistic at that site. For convenience, in subdirectory /example_input/Kspecies/ for input with K species, we provide example output files for $T_{1,\text{trans}}$ and $T_{2,\text{trans}}$, respectively labeled T1trans_2Sp.out and T2trans_2Sp.out in directory /example_input/2species/ for $K = 2$ species, T1trans_3Sp.out and T2trans_3Sp.out in directory /example_input/3species/ for $K = 3$ species, and T1trans_4Sp.out and T2trans_4Sp.out in directory /example_input/4species/ for $K = 4$ species.

7. Examples

The `example_input` directory provides example input files for $K = 2$ (`2species` subdirectory), $K = 3$ (`3species` subdirectory), and $K = 4$ (`4species` subdirectory) species. For the following commands, we assume that executable *MULLET* is located in the same directory as the example files, and that the set of simulated spectra is in a subdirectory called `example_input/simulated_spectra/`. This set of simulated spectra can be downloaded from the same website that *MULLET* was downloaded.

For an example of K species (directory `example_input/Kspecies/`), there are five sets of files: a file describing the bifurcating unrooted tree (section 3.1) named `Clades_forKsp.txt`, an input frequency file (section 3.2) named `forKSp-HCGOg_15MYA_s.01_h100_63input.txt` of a simulated replicate in which a selected allele introduced 15 million years ago (assuming a generation time of 20 years) in the center of the sequence evolved under heterozygote advantage with per-generation selection coefficient $s = 0.01$ and dominance parameter $h = 100$, a configuration file (section 4.1) named `forKSp-HCGOg_Neut_transConfig.txt`, and the K marginal frequency spectra (section 4.2) named `forKSp-HCGOg_Neut_spect1.txt` through `forKSp-HCGOg_Neut_spectK.txt`.

Results for all example scans are highlighted in Figure 2.

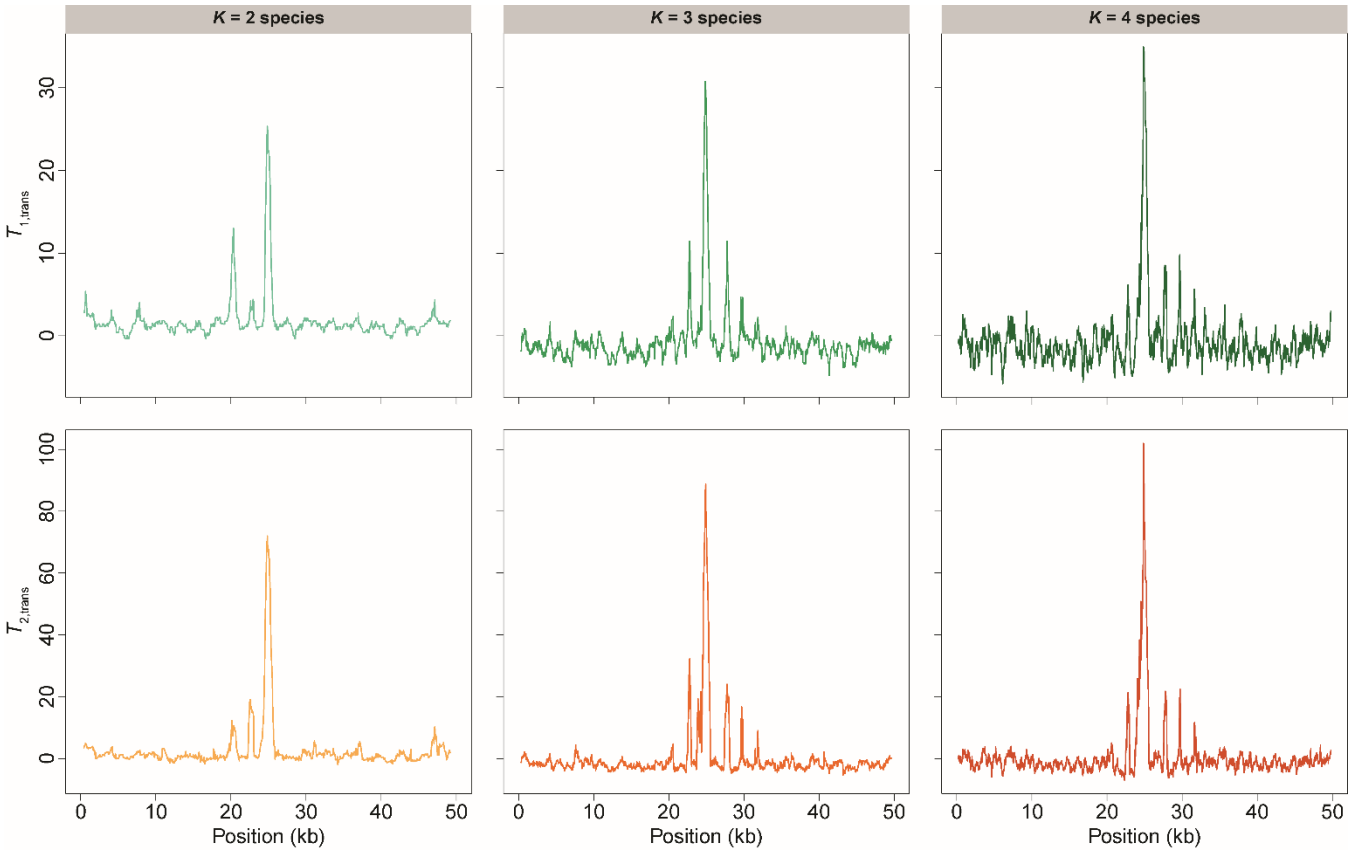

Figure 2: *MULLET* applied to data simulated under long-term under balancing selection, in which a selected allele arose 15 million years ago (assuming a generation time of 20 years) in the center of a sequence, and evolved under heterozygote advantage with per-generation selection coefficient $s = 0.01$ and dominance parameter $h = 100$. Results are displayed for both $T_{1,\text{trans}}$ and $T_{2,\text{trans}}$ for $K = 2$, $K = 3$, and $K = 4$ species.

Example commands for using *MULLET* with the $T_{1,\text{trans}}$ and $T_{2,\text{trans}}$ statistics to identify ancient trans-species balancing selection using the tree in Figure 1A for $K = 2$ species with a window of size $W = 10$ informative sites upstream and downstream of a test site.

```
./MULLET -T1trans 10 2 example_input/2species/Clades_for2sp.txt
example_input/2species/for2Sp-HCGOg_Neut_transConfig.txt
example_input/2species/for2Sp-HCGOg_15MYA_s.01_h100_63input.txt
T1trans_2Sp.out

./MULLET -T2trans 10 2 example_input/2species/Clades_for2sp.txt
example_input/2species/for2Sp-HCGOg_Neut_transConfig.txt
example_input/2species/for2Sp-HCGOg_Neut_spect1.txt
example_input/2species/for2Sp-HCGOg_Neut_spect2.txt
example_input/2species/for2Sp-HCGOg_15MYA_s.01_h100_63input.txt
example_input/simulated_spectra/ T2trans_2Sp.out
```

Examples of `T1trans_2Sp.out` and `T2trans_2Sp.out` are already located in the `example_input/2species/` directory with the *MULLET* download. Example commands for using *MULLET* with the $T_{1,\text{trans}}$ and $T_{2,\text{trans}}$ statistics to identify ancient trans-species balancing selection using the tree in Figure 1B for $K = 3$ species with a window of size $W = 10$ informative sites upstream and downstream of a test site.

```
./MULLET -T1trans 10 3 example_input/3species/Clades_for3sp.txt
example_input/3species/for3Sp-HCGOg_Neut_transConfig.txt
example_input/3species/for3Sp-HCGOg_15MYA_s.01_h100_63input.txt
T1trans_3Sp.out

./MULLET -T2trans 10 3 example_input/3species/Clades_for3sp.txt
example_input/3species/for3Sp-HCGOg_Neut_transConfig.txt
example_input/3species/for3Sp-HCGOg_Neut_spect1.txt
example_input/3species/for3Sp-HCGOg_Neut_spect2.txt
example_input/3species/for3Sp-HCGOg_Neut_spect3.txt
example_input/3species/for3Sp-HCGOg_15MYA_s.01_h100_63input.txt
example_input/simulated_spectra/ T2trans_3Sp.out
```

Examples of `T1trans_3Sp.out` and `T2trans_3Sp.out` are already located in the `example_input/3species/` directory with the *MULLET* download. Example commands for using *MULLET* with the $T_{1,\text{trans}}$ and $T_{2,\text{trans}}$ statistics to identify ancient trans-species balancing selection using the tree in Figure 1C for $K = 4$ species with a window of size $W = 10$ informative sites upstream and downstream of a test site.

```
./MULLET -T1trans 10 4 example_input/4species/Clades_for4sp.txt
example_input/4species/for4Sp-HCGOg_Neut_transConfig.txt
example_input/4species/for4Sp-HCGOg_15MYA_s.01_h100_63input.txt
T1trans_4Sp.out

./MULLET -T2trans 10 4 example_input/4species/Clades_for4sp.txt
example_input/4species/for4Sp-HCGOg_Neut_transConfig.txt
example_input/4species/for4Sp-HCGOg_Neut_spect1.txt
example_input/4species/for4Sp-HCGOg_Neut_spect2.txt
example_input/4species/for4Sp-HCGOg_Neut_spect3.txt
example_input/4species/for4Sp-HCGOg_Neut_spect4.txt
```

example_input/4species/for4Sp-HCGOg_15MYA_s.01_h100_63input.txt
example_input/simulated_spectra/ T2trans_4Sp.out

Examples of T1trans_4Sp.out and T2trans_4Sp.out are already located in the example_input/4species/ directory with the *MULLET* download.

8. References

X Cheng, M DeGiorgio (2018) Detection of shared balancing selection in the absence of trans-species polymorphism. *bioRxiv* doi:10.1101/320390.