MStree-User Guide-V1.1

1) Quick Reference

The C program MStree is used to estimate ancestral population sizes and species divergence times for three species during speciation with gene flow (Fig 1). For the three species, species **1** and species **2** are closely related species, and species **3** is an out-group species. There may be gene flow between species **1** and species **2**, but no gene flow occurs between the ancestral species of two closely related species and the species **3**. The input file of MStree is gene tree, which is in Newick format, and the gene tree should be estimated from the observed sequence alignments where there must be three sequences, with one sequence from each species, at each locus. Furthermore, the observed sequences should be available from multiple neutral loci. Therefore, we need a program, such as PHYLIP, to convert to gene tree when applying MStree to experimental data. It is noteworthy that the three species names should be converted to **1**, **2** and **3** when using MStree.

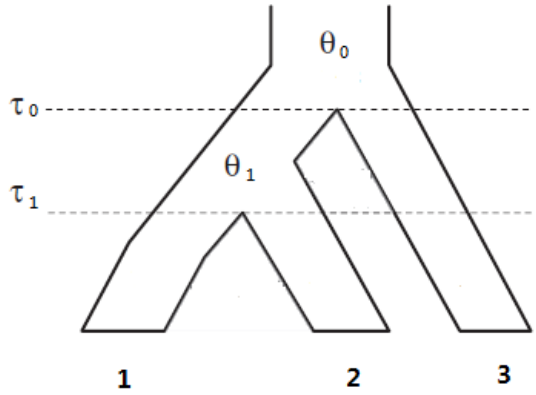

Fig.1. Species tree ((1,2),3) for three species. The species divergence times are denoted as τ_0 and τ_1, respectively. The ancestral species population sizes are denoted as θ_0 and θ_1.

(1) Compiling the program

You should use gcc or any ANSI C-compatible compiler for UNIX or Windows. The source codes are from my MStree package. The compile command is as follows.

Usage: gcc -o mstree mstree.c tools.c -lm

(2) Running the program

The running command is as follows.

Usage: mstree mstree.ctl

(3) The format of control file mstree.ctl

The format of control file mstree.ctl is as follows.

```
seed = -1                    * for generating random number, no change
outfile = example_out        *  the name of output file
seqfile  = example_tree      *  the name of input file
getSE = 1000 1000 10 1000    * four numbers are integer and are used to adjust output
```

Note. For the term "getSE" in the control file, the value of the first number divide by the second number **is usually 1**. **If the branch length in gene tree is not the product of the number of generations and mutation rate**, the first number divide by the second number should be the value which can convert the branch length to the product of the number of generations and mutation rate by multiplying it. For example, the branch length in gene tree obtained by ms is in units of $4N$ (N is effective population size) generations, so the value of the first number divide by the second number should be $4N\mu$ (μ is the mutation rate), that is the value of parameter t in ms. The value of the third number divide by the fourth number is the threshold value ε in MStree, we suggest the threshold value should be between 0.01 and 0.03.

(4) The format of input file

The input file of MStree is gene tree, which is in Newick format. Gene trees for all loci are in one file, one locus after another. For example, gene trees for eight loci are as follows.

(3:1.787,(1:0.671,2:0.671):1.116);

(3:2.236,(1:1.288,2:1.288):0.948);

(3:1.390,(1:0.665,2:0.665):0.725);

(3:1.732,(1:0.853,2:0.853):0.879);

(1:1.249,(2:1.218,3:1.218):0.031);

(3:1.383,(1:0.875,2:0.875):0.508);

(1:1.429,(2:1.233,3:1.233):0.196);

(2:3.617,(1:1.544,3:1.544):2.073);

Note. The format ((#:##,#:##):##,#:##) is not available in mstree. For example, for the above eighth gene tree, the format ((1:1.544,3:1.544):2.073, 2:3.617) is not available in mstree.

(5) The format of output file

The format of output file is as follows.

mstree (Version 1.1, July 2018)

The number of tree ((1,2),3)is 7496

The number of tree ((1,3),2)is 1217

The number of tree ((2,3),1)is 1287

theta0 theta1 tau0 tau1 are 0.54292 0.51804 0.59731 0.43006

Note. θ=4 $N\mu$ and τ= $t\mu$, where N is effective population size and t is the number of generations. The values of theta0, theta1 and tau0 are the 100 times of θ_0 , θ_1 and τ_0. The value of tau1 is τ_1 divide by τ_0. For the above example, θ_0=0.54292/100, θ_1=0.51804/100, τ_0=0.59731/100 and τ_1=0.43006/100.

2) Example

We simulated gene tree data at multiple loci under isolation-with-initial-migration model, secondary contact model and isolation-with-migration model using the program ms. The two sets of parameter values were used, roughly based on estimates from hominoids and mangroves. They are as follows: θ_0=θ_1=0.005, τ_0=0.006 and τ_1=0.004 (hominoids); θ_0=θ_1=0.01, τ_0=0.02, and τ_1=0.01 (mangroves). For the three kinds of models, gene flow is symmetrical and the migration rate (the expected number of migrants per generation) is 1. The number of loci is 10,000 and the number of replicates is 1000.

(1) Hominoids

The simulation commands are as follows.

Isolation-with-initial-migration model:

./ms 3 10000 -t 0.005 -T -I 3 1 1 1 -m 1 2 0 -m 2 1 0 -em 0.533 1 2 4 -em 0.533 2 1 4 -em 0.8 1 2 0 -em 0.8 2 1 0 -ej 0.8 2 1 -ej 1.2 3 1 | tail -n +4 | grep -v // > tree

Secondary contact model:

./ms 3 10000 -t 0.005 -T -I 3 1 1 1 -m 1 2 4 -m 2 1 4 -em 0.267 1 2 0 -em 0.267 2 1 0 -ej 0.8 2 1 -ej 1.2 3 1 | tail -n +4 | grep -v // > tree

Isolation-with-migration model:

./ms 3 10000 -t 0.005 -T -I 3 1 1 1 -m 1 2 4 -m 2 1 4 -em 0.8 1 2 0 -em 0.8 2 1 0 -ej 0.8 2 1 -ej 1.2 3 1 | tail -n +4 | grep -v // > tree

The control file mstree_H.ctl is as follows.

```
seed = -1                  * for generating random number, no change
outfile = out_H            *  the name of output file
seqfile  = tree            *  the name of input file
getSE = 5 1000 10 1000     * four numbers are integer and are used to adjust output
```

(2) Mangroves

The simulation commands are as follows.

Isolation-with-initial-migration model:

./ms 3 10000 -t 0.01 -T -I 3 1 1 1 -m 1 2 0 -m 2 1 0 -em 0.667 1 2 4 -em 0.667 2 1 4 -em 1 1 2 0 -em 1 2 1 0 -ej 1 2 1 -ej 2 3 1 | tail -n +4 | grep -v // > tree

Secondary contact model:

./ms 3 10000 -t 0.01 -T -I 3 1 1 1 -m 1 2 4 -m 2 1 4 -em 0.333 1 2 0 -em 0.333 2 1 0 -ej 1 2 1 -ej 2 3 1 | tail -n +4 | grep -v // > tree

Isolation-with-migration model:

./ms 3 10000 -t 0.01 -T -I 3 1 1 1 -m 1 2 4 -m 2 1 4 -em 1 1 2 0 -em 1 2 1 0 -ej 1 2 1 -ej 2 3 1 | tail -n +4 | grep -v // > tree

The control file mstree_M.ctl is as follows.

```
seed = -1                    * for generating random number, no change
outfile = out_M              *  the name of output file
seqfile  = tree              *  the name of input file
getSE = 10 1000 30 1000      * four numbers are integer and are used to adjust output
```