

fte: A Coalescent Simulator

Mark Reppell, Michael Boehnke, Sebastian Zöllner

University of Michigan

January 11, 2012

This document is meant to outline how to use the fte program for coalescent simulation. This program allows the user to generate sample haplotypes from populations which have undergone a range of demographic histories including instantaneous growth or contraction events, population subdivision with migration, and continuous growth slower, equal to, or faster than an exponential rate. Samples are generated using the standard coalescent approach where the time to coalescent, recombination, and migration events are independent exponential variables, with mutation events placed on the tree after a sample's ancestry has been simulated [?]. For continuous growth models the ideas of Donnelly and Tavaré [?] have been employed to shift times from a neutral model to those for a population with deterministically varying past size. The program was written in C++ and runs from the command line.

Downloading and Compiling fte

The latest version of the fte program is available from the University of Michigan Center for Statistical Genetics wiki. The source files and instructions on how to compile the program can be found at <http://genome.sph.umich.edu/wiki/Fte>.

Command Overview

The following sections list and describe the parameters recognized by `fte`. Program parameters can be entered in any order on the command line. For each parameter listed the number of dashes shown before the corresponding parameter flag represents the exact number of dashes required at the command line. For example `-N` requires a single dash while `--time` requires two dashes. Some parameters can be entered multiple times on the command line in order to vary their value over time. Parameters that can be entered multiple times are indicated as such in the text, and are an exception to the rule that parameters can be entered in any order on the command line. Parameters that can be entered multiple times need to be entered in chronological order backwards in time from the present.

Basic Commands

- n** The total number of haplotype samples to model. This is a required value.
- x** The number of independent simulation repetitions to perform. This value defaults to one.
- N** The *present day* effective size (diploid) of the population samples are being drawn from. Time for the simulations will be scaled in coalescent units of $2N$ generations, where N is the value entered with this flag. This is a required value and is also referred to as the initial population size in this document.
- time** Boolean, prints the final running time for current series of simulations.
- m** Mutation rate, this value should be entered as the per base mutation rate times the number of bases being simulated per sample haplotype. The program will calculate the value of θ as $4Nm$ where N is the initial population size you specified with the `-N` command. This is a required value.
- r** Recombination rate, the `fte` program handles recombinations as a uniform process across the haplotypes simulated. The recombination rate parameter should be entered as the per base recombination rate times the

number of bases being simulated per haplotype sample. The program will calculate ρ as $4Nr$ where N is the initial population size specified with the $-N$ command. This value defaults to zero.

- L** Boolean, prints time to most recent common ancestor and total tree length for each simulation. Values are given in coalescent units of $2N$ generations.
- h** Boolean, this displays all command line parameters recognized by the program, with a brief description of each.
- seed** Random number generator seed, this value allows for repetition of the same experiment by entering the same parameters along with the same seed value. If nothing is entered a random number based on the current time will be used. The seed value used for a given series of simulations can be seen on the second line of program output.

Population Growth

Both instantaneous and continuous population size changes can be modeled using fte.

Continuous Growth

To model a population which has expanded at a faster than exponential rate, we begin with a differential equation relating the rate of change in population size to current population size [?]

$$\frac{dN}{dt} = -\alpha N^c$$

Here N is population size, t is time, α and c are constants, with $\alpha < 0$ due to the coalescent modeling time backwards from the present. When we use the current population size N_0 as an initial condition, this equation has a general solution

$$N(t) = \begin{cases} \left[\frac{N_0^{c-1}}{1 + N_0^{c-1}(c-1)\alpha t} \right]^{\frac{1}{c-1}}, & \text{for } c \neq 1 \\ N_0 e^{-\alpha t}, & \text{for } c = 1 \end{cases} \quad (1)$$

For $c > 1$ equation (1) models a population expanding at a faster than exponential rate. $N(t)$ can be transformed as $\Lambda(t) = \int_0^t \frac{N_0}{N(s)} ds$ to “shift” coalescent times from the standard model into coalescent times for a population with varying size [?].

Commands

- g Type of continuous growth, recognized values of this parameter are 1 or 2. If set to 1 the program models exponential growth, if set to 2 the program will model faster or slower than exponential growth. If no value is entered population size is modeled as constant backwards in time.
- t Time of growth, the value entered is the amount of time backwards from the present, in coalescent units of $2N$, during which the population being sampled from underwent continuous growth. A coalescent unit is equivalent to $2N$ generations, where for our model N is the current diploid population size entered using the $-N$ flag.
- a Growth constant α , this value controls the rate of growth in continuous growth models.
- c Faster/slower than exponential growth exponent. If the $-g$ parameter is entered as 2 the $-c$ value represents the exponent c in equation 2, controlling how much faster or slower than the exponential growth is occurring. An entry of 1 is invalid, to model exponential growth set the $-g$ flag to 1 and refrain from entering a $-c$ value.

IMPORTANT When entering t , c , and a on the command line it is important to remember that during faster than exponential growth, unlike exponential growth, a population shrinking from $2N$ to N over a given time does not have the same growth constant α as a population shrinking from $4N$ to $2N$ over the same time interval, so make sure to enter effective population size in *chromosomes* and *not* individuals.

Instantaneous Population Size Changes

The *fte* program allows population sizes to change instantaneously in the past.

Commands

These commands can be entered multiple times for multiple instantaneous size changes. The sizes entered with `-b` will correspond in order to the times entered with `--t_instant`.

- b** Effective population size after contraction/expansion event, as a proportion of initial population size (N_0).
- t_instant** Time of contraction/expansion event in coalescent units of $2N$. If multiple instantaneous size change events are modeled their times should be entered in chronological order backwards from the present. In the current version of the program if continuous growth is being modeled instantaneous events can only occur further in the past than the conclusion of continuous growth.

Population Subdivision

The `fte` program has the ability to simulate a two subpopulation island model of population subdivision. Mutation and recombination rates are assumed to be the same in the two populations. With the exception of the command `-n`, which remains the *total* number of haplotypes to sample, the values entered for any of the parameter labels in previous sections of this document are taken to be values for population one, with parameters for population 2 entered using the commands in this section.

0.1 Commands

- s** Boolean, indicates current series of simulations involve population subdivision.
- migration** Migration rate between populations, entered as the fraction of chromosomes leaving each population each generation. If no `--migration2` value is entered the value given to `--migration` is taken to be the rate at which chromosomes move in both directions between populations 1 and 2. If a `--migration2` value is entered, `--migration` controls the rate of chromosomes moving from population 1 to population 2. For migration rate changes over time this parameter can be entered multiple times along with `--migration_time` values. The first `--migration`

value will be taken as the starting migration rate and each additional `--migration` value should have a corresponding `--migration_time` entry. This value defaults to zero.

- migration2** Migration rate for individuals migrating from population 2 to population 1, entered as the fraction of chromosomes leaving each population each generation. If no value is entered for this parameter but a value is given to `--migration` that value is taken to be the equal migration rate between populations. For migration rate changes over time this parameter can be entered multiple times along with `--migration_time2` values. The first `--migration2` value will be taken as the starting migration rate and each additional `--migration2` value should have a corresponding `--migration_time2` entry.
- migration_time** Time at which migration rate changes, entered in coalescent units of $2N$ generations. Requires entry of multiple `--migration` values, with a total of one more `--migration` value than `--migration_time` values. If multiple `--migration_time` values are entered they should be in chronological order backwards from the present.
- migration_time2** Time at which migration rate from population 2 to population 1 changes, entered in coalescent units of $2N$ generations. Requires entry of multiple `--migration2` values, with a total of one more `--migration2` value than `--migration_time2` values. If multiple `--migration_time2` are entered they should be in chronological order backwards from the present.
- merge_time** Time until the divided subpopulations merge into a single population, in coalescent units of $2N$ generations. If no value is entered the two subpopulations never merge.
- merge_size** The size of combined population at the time the two subpopulations merge, as a proportion of the initial population 1 size (N_0). If a value for `--merge_time` is entered but this parameter is not set the default sets the effective population size at the time they merge as the sum of the effective sizes of the two subpopulations.
- p2prop** Current effective size of population 2 as a proportion of initial population 1 size (N_0).

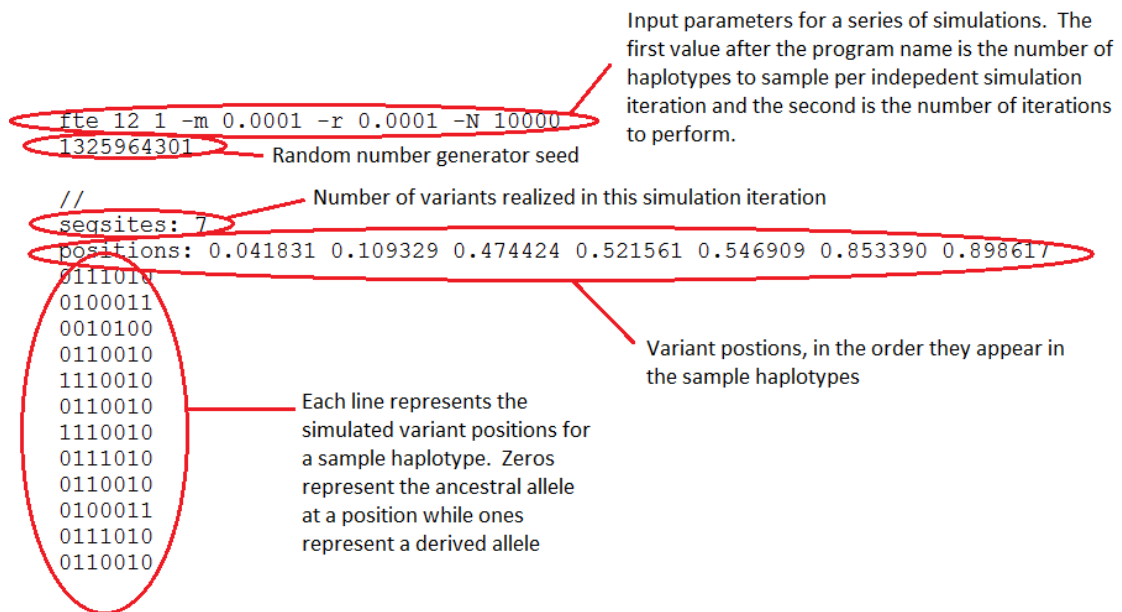
- p2n** Number of simulated samples originating in population 2. fte subtracts --p2n from the -n value to calculate the number of samples originating in population 1.
- g2** Type of continuous growth in population 2. A value of 1 indicates population 2 grew at an exponential rate, a value of 2 indicates population 2 grew at a faster or slower than exponential rate. If nothing is entered population 2 is modeled with as having a constant size backwards in time.
- t2** Time of growth in population 2, the amount of time backwards from the present, in coalescent units of $2N$, during which population 2 underwent continuous growth.
- a2** Growth constant α for population 2, this value controls the rate of growth in continuous growth models for population 2.
- c2** Faster/slower than exponential growth exponent. If the --g2 parameter is entered as 2 the --c2 value represents the exponent c in equation 2, controlling how much faster or slower than the exponential growth is occurring. An entry of 1 is invalid, to model exponential growth set the --g2 flag to 1 and refrain from entering a --c2 value.
- b2** Population 2 size after instantaneous expansion/contraction event, as proportion of population 1 initial size (N_0). This parameter can be entered multiple times for multiple instantaneous size changes in population 2.
- t_instant2** Time of contraction/expansion event in population 2 in coalescent units of $2N$. If multiple instantaneous size change events are modeled in population 2 their times should be entered in chronological order backwards from the present. To model an instantaneous event after two subdivided populations have merged use the -b and --t_instant values.

Output

Figure 1 demonstrates the format of fte output. Output can be captured from standard out. The segregating site positions, which are randomly and

independently chosen, have values between 0 and 1, the arbitrary range covered by the simulated haplotypes. These values are important only in how they relate to each other, for example recombination events are more likely to occur between variants at positions 0.05 and 0.95 than between variants which occur at positions 0.05 and 0.06.

Figure 1: fte simulator output



Example 1

Simulating two independent samples of 5000 haplotypes, each haplotype approximately 50kb in length, from a current day diploid population with a size of $N = 5,000,000$ that has grown at a faster than exponential rate from an ancestral population of size $N = 9,000$ over the last 500 generations

- 500 generations is $\frac{500}{10,000,000} = 5 \times 10^{-5}$ in $2N$ coalescent units.

- If we use a faster than exponential exponent c of 1.25 we can solve for the growth constant:

$$18000 = \left(\frac{10000000^{(1.25-1)}}{1+\alpha(10000000^{(1.25-1)})(5 \times 10^{-5})} \right)^{\frac{1}{1.25-1}} \Rightarrow \alpha = 5484.1$$

- Assume a per base mutation and recombination rate of 10^{-8} , then the m and r values become $50000 \times 10^{-8} = 0.0005$

This gives us the command line input:

```
./fte -N 5000000 -n 5000 -x 2 -m .0005 -r .0005 -g 2 -c 1.25 -t 5e-5 -a 5484.1
```

Example 2

Simulating 1,000 repetitions of 2,000 sample haplotypes from two divided subpopulations with 1,000 haplotypes coming from each subpopulation, and each haplotype approximately 25kb in length. Assume the underlying populations have effective sizes $N = 5,000,000$ and $N = 1,000,000$ currently but were both 10,000 individuals in size 200 generations in the past. Assume there is no migration between the populations currently but before 100 generations ago the chance of a chromosome moving between populations was 0.0001 per generation.

- Assume a per base mutation and recombination rate of 10^{-8} , then the m and r values become $25000 \times 10^{-8} = 0.00025$
- In coalescent units of $2N$, 100 generations = $\frac{100}{10000000} = 10^{-5}$ and 200 generations = 2×10^{-5}
- The second population is $\frac{1}{5}$ the size of the first to begin and both populations have size $\frac{10000}{5000000} = 0.002$ times the initial population 1 size 200 generations in the past

This gives us the command line input:

```
./fte -N 5000000 -n 2000 -x 1000 -m .00025 -r.00025 -s --p2prop .2 --p2n 1000 -migration 0 --migration .0001 --migration_time 1e-5 -b .002 --t_instant 2e-5 --b2 .002 --t_instant2 2e-5
```