SELAM

Matt Jones
Russ Corbett-Detig

October 16, 2015

1 Overview

SELAM (Simulation of Epistasis, Local adaptation with Admixture and Mate choice) is a forward time population genetic simulation for studying admixture between ancestral subpopulations. This program tracks local ancestry along chromosomes. SELAM supports complex demography scenarios, including rapid changes in population sizes, migration rates, and arbitrary numbers of subpopulations. This program can also accommodate sophisticated selective regimes, including dominance, interlocus epistasis, local adaptation, and mate choice.

In a **forward simulation**, every individual is accounted for explicitly. Therefore, forward time simulations require much more computational power than coalescent approaches. This compounded with the large computational requisites of following ancestry blocks through each individual creates a practical challenge. For this reason, SELAM has been optimized with high-throughput algorithms and time-saving tactics to enable simulations of relatively large populations in a comparatively short amount of time. This manual will outline how to compile and run SELAM and it includes a number of example simulations.

2 Simulation Methods

SELAM is based on the **Wright-Fisher model** with selection. The Wright-Fisher assumes discrete population sizes with non-overlapping generations. That is, parents are selected with probabilities proportional to their fitness to reproduce. After the offspring are created, the previous generation or parents dies, and the process repeats.

SELAM can accommodate arbitrarily sophisticated demographic scenarios. Users may define any number of subpopulations, within each offspring are produced following a Wright-Fisher model with selection as described above. Migration may occur at any rate (less than 1) between any of the subpopulations. In addition, subpopulations may experience ongoing migration from either of the ancestral (non-recombinant) populations throughout the simulation.

If **mate choice** loci are specified all female parents are selected as described above, and then potential fathers are selected based on their fitness as before. However, males may be rejected

by females based on their genotypes at mate choice loci. That is, a mother's genotype at one locus gives them some probability of rejecting potential mates based on the male's genotype at a given locus. If a male is rejected, another is proposed proportionally to his fitness, and the mate choice process is repeated until a male is found that the female does not reject. See section 4.2.1 which describes how to specify mate choice loci in simulations.

When **selected loci** are specified, SELAM will select parents proportionally to their fitness. Each individual's fitness is computed based on their genotypes at selected loci specified by the user. This can include both single-locus and epistatic selection. These fitness effects can be sex-specific, population specific dominant, or applied universally. Between selected loci, fitness effects are multiplicative unless specified as epistatic. Individuals are selected to reproduce based on multinomial sampling.

In SELAM's **Recombination Model**, when gametes are created, the number of chiasma sites per chromosome is drawn from a Poisson distribution. Chiasma sites are then selected based on a uniform (0, chromosome length) distribution. Chromosome number and lengths can be specified by the command line using -c (see section 5.1.1 below). All positions of mutations, recombination tract start and stop sites are given in genetic distance. An important note is that this simulation accounts for sex chromosomes. In simulations with separate sexes (which is SELAM's default), the last chromosome is an X chromosome, and all others autosomes. Males will have only a single copy of the X chromosome, and there is no recombination at this locus in males. If the population is hermaphroditic, no sex chromosomes are allowed, and all chromosomes will be autosomes.

SELAM achieves a portion of its efficiency by breaking chromosome into blocks of uniform length, known as **Ancestry Blocks**. Ancestry blocks record the genotypes at selected mutations and ancestry tract information for a portion of the chromosome. Thus, an individual's chromosomes are represented by a list of pointers to ancestry blocks. When a gamete is produced, all blocks that do not recombine may simply copy the pointers to the child chromosome, which is generally much more efficient that copying all ancestry tract information. For a details about the ancestry block model, refer to Section 6.2 below.

At any generation during the simulation run, SELAM may output a sample of individual from a user specified subpopulation.

3 Downloading and Compiling SELAM

SELAM is a command line program written in the C++ programming language. The user will also need to have the GNU Scientific Library (GSL) installed on the system. Included in the SELAM src directory is a MAKEFILE - to use it, just run the following command via command-line:

make -f MAKEFILE

As a note, the MAKEFILE will adapt to the default compiler on your computer's operating

system. If google-perftools is not installed, read the MAKEFILE for information how to compile without tcmalloc.

4 Input Files

4.1 The Demography File

This table must specify the initial composition of each subpopulation in terms of the ancestral populations. By default, there should be a minimum of two ancestral populations that constitute each subpopulation - **a0** and **a1**. These names must be used to specify the ancestral populations. Additionally, the user must define subpopulation sizes by specifying "migration rates" from to the same population; the demography must also start at generation 0. Below is a simple example of a demography file for one population.

Table 1: Simple Demography File

pop0	pop1	sex	0	1
0	0	A	1000	1000
0	a0	A	0.5	0.1
0	a1	A	0.5	0.1

The example above produces a population initially made up of half of both ancestral populations. By convention, the first line should specify the population size, and the subsequent lines should specify the proportion of ancestral individuals migrating into each new population. It is important to recognize that generation 0 only serves to outline the proportion of the ancestral individuals in each subpopulation at the start of the simulation. From generation 1 onwards, 10% of the individuals in the subpopulation will be migrants from each ancestral population while 80% will be offspring from the previous generation.

A couple of other important notes include the migration semantics - **pop0** specifies the receiving subpopulation whereas **pop1** specifies from which population individuals are migrating. The capital letters A, M, and F denote the sex that the proportions apply to. Table 2 provides a more sophisticated demography file.

Table 2: Demography File

pop0	pop1	sex	0	1	10
0	0	A	1000	1000	1000
0	a0	A	0.5	0.1	0.1
0	a1	A	0.5	0.01	0.01
0	1	M	0	0.25	0.4
0	1	F	0	0.3	0.3
1	1	A	1000	1000	1000
1	a0	A	0.25	.01	.01
1	a1	A	0.75	.2	.2
1	0	A	0	0.5	0.5

Note that in Table 2, by generation 1, 25% in population 0 are migrants from population 1, while 30% of in population 0 are migrants from population 1. In column 3, **A** signifies that the proportion is universal and thus applies to both males and females. The population sizes are specified via a migration from the same subpopulation - the first line translates to a population size of 1000 for population 0. Note that there is no migration between the subpopulations 0 and 1 during this initial generation. Also, observe that in generation 10, new demographic changes take effect - instead of 25%, now 40% of males in population 0 are migrants from population 1. All other parameters are unchanged, but must be specified in each new demography column.

To summarize:

1. There must be 2 ancestral populations in the demography file.
2. The first generation must be 0 and specifies each subpopulation's ancestral composition.
3. Each subpopulation used must have a population size, specified via a self-migration. See lines 2 and 7 in Table 2.
4. Each line must specify the sex (A, M, or F).

4.2 The Output File

The user must provide a file specifying how many individuals of each sex from a certain subpopulation should be outputted in a given generation. Provided in Table 3 is an example.

Table 3: Output File

0	1	10	10	test1.txt
1	0	10	10	test1.txt
5	1	15	15	test1.txt
50	1	10	10	test2.txt

The first column denotes the generation, the second the subpopulation; the third and fourth are the amount of females and males to be studied respectively and the .txt file specified is the output file name. Note that multiple output lines can output to the same file; SELAM will append to the file specified. For the number of individuals specified, SELAM will print each chromosome's admixture tracts in a tab delimited file. The table below gives an example of SELAM's output; post processing scripts can be used to extrapolate useful information from this output.

Table 4: Example Output

5	0	0	1	0	0	0	0	0.683841
5	0	0	1	0	0	1	0.683841	1
5	0	0	1	0	1	1	0	1

In Table 4, a single female individual during the fifth generation is represented. The first line can illustrate the foundation of the format; the following list explains each number by column:

1. "5" denotes the generation output.
2. "0" means that the individual was selected from subpopulation "0."
3. "0" means that the individual is female; a "1" would indicate that the individual is male.
4. "1" is the individual index (this is the first individual printed).
5. "0" means the tract is on the first chromosome.
6. "0" indicates that the chromosome was inherited from the mother ("1" would mean that the chromosome was inherited from the father).
7. "0" gives the ancestry type, where the tract may be either a "0" or a "1."
8. The next two numbers respectively (0 and 0.683841) represent the start and stop positions of the tract.

4.2.1 The Selection File

Should the user decide to specify selected sites in a population, a selection file can be provided to SELAM. The user may specify epistatic or single locus selection, as well as determining sex-specific effects and mate-choice. The basic structure of both of these specifications require a chromosome that the site is located on, the position (a number between 0 and the length of the chromosome), and selection coefficients. Table 5 gives an example of a selection file.

Table 5: Selection File

D	A	0	0	0.5431	0.5221	1	1	0.9	1	1	1	0.8	1	1
S	M	0	0.2305	1	1	0.9								
P	F	0	0.4829	1	0.9	0.8	0.8	1	1					
M	F	0	0	0.4899	0.5555	1	1	0.5	1	1	0.4	1	1	0.3

The example above gives information to specify selection over two chromosomes and provides examples of how to specify epistatic and single locus selection. The following discussion breaks down the file line by line:

1. The D prefix denotes epistatic selection. As for the demography section, the A denotes that the selection applies for both males and females. The following two columns designate the chromosome and the next two columns specify the position of the interaction sites - for example, the first site is on chromosome 0 at position 0.5431. The next 9 values are selection coefficients for all possible genotypes - AA/BB, AA/Bb, etc. Since each individual is diploid, it has the possibility of being homozygous or heterozygous at any of these selected positions. Selection coefficients must be supplied in the following order: AA/BB, AA/Bb, AA/bb, Aa/BB, Aa/Bb/ Aa/bb, aa/BB, aa/Bb, aa/bb. In this particular example, AA/BB would have a fitness of 1, AA/bb would have a fitness of 0.9, and aa/BB would have fitness 0.8.

2. The S prefix denotes single locus selection; M means that the selection only applies to males. The following two numbers designate the chromosome and position that is selected. Like in the epistatic example above, the individual can be either homozygous or heterozygous for this particular allele - here the fitnesses are provided as AA, Aa, and aa. In this particular example, the genotype aa has fitness of 0.9 in males.

3. The P prefix denotes population specific selection (i.e. as in local adaptation); F means that the selection only applies to females. The following two numbers define the chromosome and position, respectively, of the selected locus. As in the single selection case, an individual can be homozygous or heterozygous; the only difference is that the selection is different for each subpopulation and thus we must specify 3 values for each subpopulation. Each triplet is ordered by subpopulation - the first triplet corresponds to the "0" subpopulation, the second triplet to the "1" subpopulation, etc.

4. The M prefix is used to specify mate choice: for the purposes of this simulation, mate selection will only take place by females. The first pair of chromosome-position pairs (i.e. 0 & 0.4899) refers to the preferred site on females, whereas the second pair (i.e. 0 & 0.5555) refers to the site being chosen in males. Much like in epistatic selection, the next 9 values are coefficients dependent on whether or not the female and potential male mate is homozygous or heterozygous for the particular allele. For example, if the female is AA for the choice loci, and the male selected is Bb, the probability of the two mating is 1; if the female is AA and the male is bb, the probability of mating is 0.5. Likewise, if the female is Aa, and the male is bb, the probability of two mating is 0.4. Much like the epistatic selection line, each value is a map value of two genotypic pairings (AA/BB, AA/Bb, etc. in the same order as epistatic selection).

Mate choice is an important extension of SELAM - it allows the user to designate selected alleles that either inhibit or facilitate mating between parents. For example, if a female-male pair have a coefficient of 0.80 as a consequence of a mate choice parameter, they will only mate 80% of the time - 20% of the time, the female will reject the male and select another potential mate.

5 Using SELAM

SELAM is run via the command line. The most basic usage is to supply only a demography file to the program:

./SELAM -d ⟨ demography file name ⟩

5.1 Parameters

Running the program consists of specifying a list of parameters that direct the simulation; the list of parameters enables the program to be dynamic enough to accommodate the user's specific simulation settings. The demography file lays the foundation of the simulation, giving the program the amount of subpopulations, the size of each subpopulation, and the migration rates from each subpopulation in between each generation.

A more exhaustive list of parameters is listed in the table that follows in Table 6:

Table 6: SELAM Parameters

Parameter	Command Line Prefix	Significance
Seed	- -seed	Random number generator seed
Generations	-g	Number of generations simulated
Output File	-o	User defined output file
Selection File	-s	User defined selection file
Demography File	-d	User defined demography file
Chromosome Lengths	-c	User defined chromosome Lengths
Male Recombination Scalar	-m	Control the map length of male chromosomes
Hermaphroditic Simulation	-h	All individuals are hermaphrodites

In order to set a parameter, just use the command line prefix and then insert the value after a space. Any ordering of the parameters can be used via the command line.

5.1.1 Chromosomes

An important note when specifying chromosomes is that the user must first give the number of chromosomes, and then the length of each. For example, the command -c 2 1 1, translates to 2 pairs chromosomes, each with a length of 1. As noted before, if the population is dioecious, the second pair will be an X. Otherwise, if -h is used, this will be two autosomes each of length 1 and the population is monoecious.

5.1.2 Male Recombination Scalar

This parameter allows the user to vary the male chromosome map length, effectively multiplying the female recombination rate by a scalar to obtain the male map length. This will not redistribute the regional frequencies of recombination events, but it will modify the total rate of recombination in males. For example, $-m\ 0.5$ will produce a male map half the length of the female map.

5.1.3 Hermaphroditic Populations

If the user specifies $-h$ on the command-line, the simulation will be populations of hermaphrodites. All selection and demographic parameters must be specified by setting the sex column to "F." During hermaphroditic simulations, there will be no sex chromosomes.

6 Advanced Options

6.1 Garbage Collection

In an effort to manage memory efficiently, the simulation has a built in garbage collector that will periodically clean up the extinct, unused admixture tracts. There is a heavy tradeoff, however, as frequent garbage collection (every 1-5 generations) dramatically increases the runtime. On the other hand, in terms of memory allocation, more frequent garbage collection reduces the amount of memory used. The option is measured in generations per garbage cleanup - for garbage collection to occur every generation, simply use the command - -garbage 1; likewise if the user wants to turn off garbage collection, simply use the command - -garbage_off. By default, the program undergoes garbage collection every 20 generations. For a graphical representation of this tradeoff, refer to Figure 1.

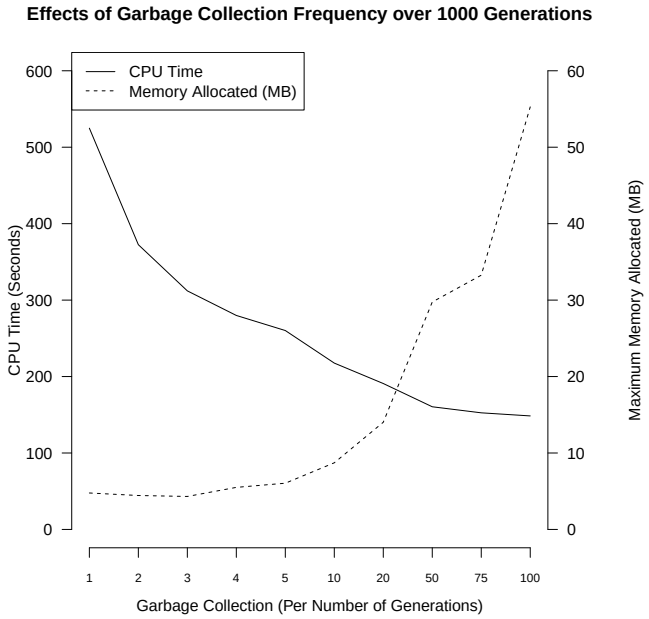

Figure 1: Tradeoffs of garbage collection frequency. Higher frequencies result in less memory allocation and longer run times, whereas the opposite is true for less frequent garbage collection. The optimum for most simulations is near garbage collection once per 20 generations, but will depend on the specific system that is running SELAM.

6.2 Ancestry Block Length

Ancestry Blocks can be thought of as caches that store data pertaining to the hybridization of chromosomes. Much like in the hardware of a computer, the block size of any cache impacts the computer's performance. In SELAM, each chromosome is broken up into a number of these ancestry blocks as specified by the ancestry block length (- -abl). Given simulation with varying lengths, there is a clear optimal ancestry block length - as a general trend, as simulations grow longer, the optimal ancestry block length decreases. The default setting in SELAM is 0.1. Refer to Figure 2 for a graphical representation of this relationship.

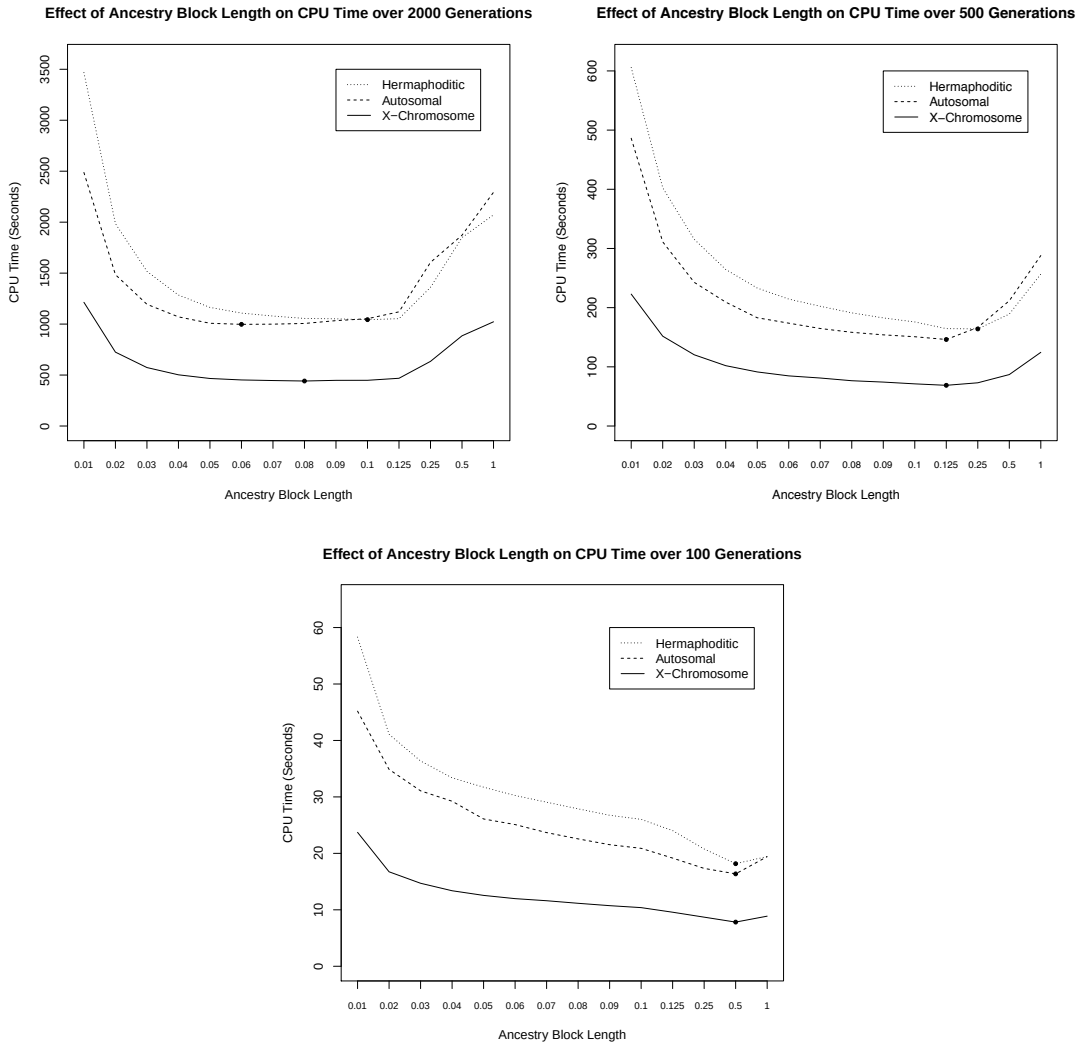

Figure 2: These three figures represent the effect that ancestry blocks have on the CPU time of a simulation; the figures are ordered in decreasing runtimes and include three different types of simulations - autosomal, hermaphroditic, and regular (includes x-chromosome). The graphs serve to illustrate that the optimal ancestry-block length varies as a function of type and length of simulation, which is represented by a point on each curve.

Top Left: Optima over 2000 generations: X-Chromosome - 0.08, Autosomal - 0.06, Hermaphroditic - 0.1.

Top Right: Optima over 1000 generations: X-Chromosome - 0.125, Autosomal - 0.125, Hermaphroditic - 0.25

Bottom: Optima over 100 generation: X-Chromosome - 0.5, Autosomal - 0.5, Hermaphroditic - 0.5