SELAM

Russ Corbett-Detig and Matt Jones

January 25, 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, 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 to enable simulations of 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 non-overlapping generations. That is, parents are selected with probabilities proportional to their fitness to reproduce. After the offspring are created, the previous generation 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 between any of the subpopulations. In addition, subpopulations may experience ongoing migration from the ancestral (non-recombinant) populations throughout the simulation.

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.

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 selected 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.

In SELAM's recombination model, 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. Coordinates are therefore reported in morgans. 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 morgans. 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 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, 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 compile SELAM, run the following command via command-line:

make

If google-perftools is not installed, read the Makefile for instructions on how to compile without tcmalloc. Note that in general, linking tcmalloc during compilation will reduce runtime.

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. However, the user may specify as many ancestral populations as they desire (using the designations, a0,a1,a2..aN). These names must be used to specify the ancestral populations. Additionally, the user must define subpopulation sizes by using a line with the same to and from populations; the demography must also start at generation 0. Below is a simple example of a demography file for one population.

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. In this example, 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 individuals in the subpopulations in 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, hence male and female migration rates may be designated separately. Below, we provide an example of a more sophisticated demography file.

Pop0	Pop1	Sex	0	1	10
0	0	A	1000	1000	1000
0	a0	A	0.5	0	0
0	a1	A	0.5	0	0.1
0	1	M	0	0.1	0.1
1	a0	A	0.75	0	0
1	a1	A	0.25	0	0
1	1	A	1000	2500	2500

Note that in this demography file, at generation 1, 10% of the males (specified by M in column 3) in population 0 are migrants from population 1. The population sizes are specified via a migration from the same subpopulation - the first line translates to a population size of 1000 males and 1000 females for population 0. In generation 1, subpopulation 1 changes in size from 1000 males and 1000 females to 2500 males and 2500 females. Also, observe that in generation 10, new demographic changes take effect – 10% of males and 10% of females in population 0 are migrants from ancestral populations 0. All other parameters are unchanged, but must be specified in each new demography column.

To summarize:

1. There must be at least 2 ancestral populations in the demography file. Ancestral populations should be numbered sequentially starting from 0.

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 that is affected (A, M, or F).

4.2 The Output File

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

0	1	0	10	Output1.txt
1	0	10	0	Output1.txt
5	1	10	0	Output1.txt
50	1	10	0	Output2.txt

The first column denotes the generation, the second the subpopulation; the third and fourth are the amount of females and males to be output respectively and the file specified in the final column 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, as well as each genotype at selected sites. SELAM will cease running in the final output generation. The table below gives an example of SELAM's output.

5	0	0	1	0	0	0	0	0.683
5	0	0	1	0	0	1	0.683	1
5	0	0	1	0	1	1	0	1

Here, a single female individual during the fifth generation is represented. This list explains each number in the first row by column:

1. "5" denotes the generation output.
2. "0" indicates that the individual was selected from subpopulation "0."
3. "0" indicates 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). The index is one-based.
5. "0" indicates 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 anything from 0-N where N is the number of ancestral populations included in the simulation.
8. The next two numbers respectively (0 and 0.683) are the start and stop positions of this tract.

When selected mutation are specified (see sections 4.3 and 4.4), the output will also contain information about which selected sites are present within that individual. This information is output to the same file as the ancestry tract lengths and represented by a single line for each chromosome. This line begins with '##', and the following six columns are identical to the six columns describing the individual, population, generation and chromosome shown above. In the remaining columns is a complete list of the mutations present on that chromosome.

4.3 The Selection File

A selection file can also be provided to SELAM. The user may specify epistatic, population-specific selection, 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. Below we give an example of a selection file.

D	A	0	0	0.5	0.9	1	1	0.9	1	1	1	0.8	1	1
S	M	0	0.2	1	1	0.9								
P	F	1	0.4	0.9	0.8	0.8	1	1						
M	F	0	0	0.4	0.5	1	1	0.5	1	1	0.4	1	1	0.3

The example above gives information to specify selected sites distributed over two chromosomes (0, and 1) and provides examples of how to specify epistatic, single locus selection, populations specific selection and mate choice. The following describes the details pertinent to each line in detail.

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 0.5. 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. If one of the loci is on the X chromosome, males

will be given a selection coefficient equivalent to the diploid homozygous genotype-i.e. males that are AA/B will be given fitness of AA/BB individuals.

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).

4.4 The Frequency File

An important feature in SELAM is the ability to specify more than two ancestral populations and the frequency of selected mutations in each ancestral population. With no frequency file given to SELAM, the simulation will default to selected sites in first ancestral population having a frequency of zero, selected sites in the second ancestral population having a frequency of one, and all other ancestral populations having a frequency of zero for all selected sites. However, in the case that the user chooses to specify the frequency of the selected sites, a simple file can be provided to do so. It is important to note that if a frequency file is used, all selected sites must be specified – even if the specifications do not differ from the default behavior. An example frequency file is given below.

0	0.3	0.2	0.8	0.5
1	0.5	0	1	0.9

In the above table, there are two selected mutations. The following list explains each item:

1. The chromosome that the selected site is on.

2. The position of the selected site.

3. The next values correspond to the frequency of the selected site in each ancestral population. In the first line, this mutation is frequency 0.2 in ancestral population 0, 0.8 in ancestral population 1, and 0.5 in ancestral population 2. The first value corresponds to the first ancestral population, the second to the second population, etc. A frequency must be specified for each population.

5 Using SELAM

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

./SELAM -d ⟨ demography file name ⟩ -o ⟨ output 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 below.

Parameter	Command Line	Significance
Seed	--seed	Random number generator seed
Frequency File	-f	File to specify frequency of selected sites
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	Scalar to change the map length of male chromosomes relative to female map length.
Hermaphroditic Simulation	-h	All individuals are hermaphrodites

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 chromosomes, each with length of 1 morgans. As noted before, if the population is dioecious, the second chromosome will be an X. Otherwise, if -h is used, this will be two autosomes each of length 1 and the population is monoecious. If sex chromosomes are not desired in a dioecious population, the command –c 2 1 0 would create a simulation with a single autosome. If this parameter is not specified, the default chromosome will be an X in a dioecious population.

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

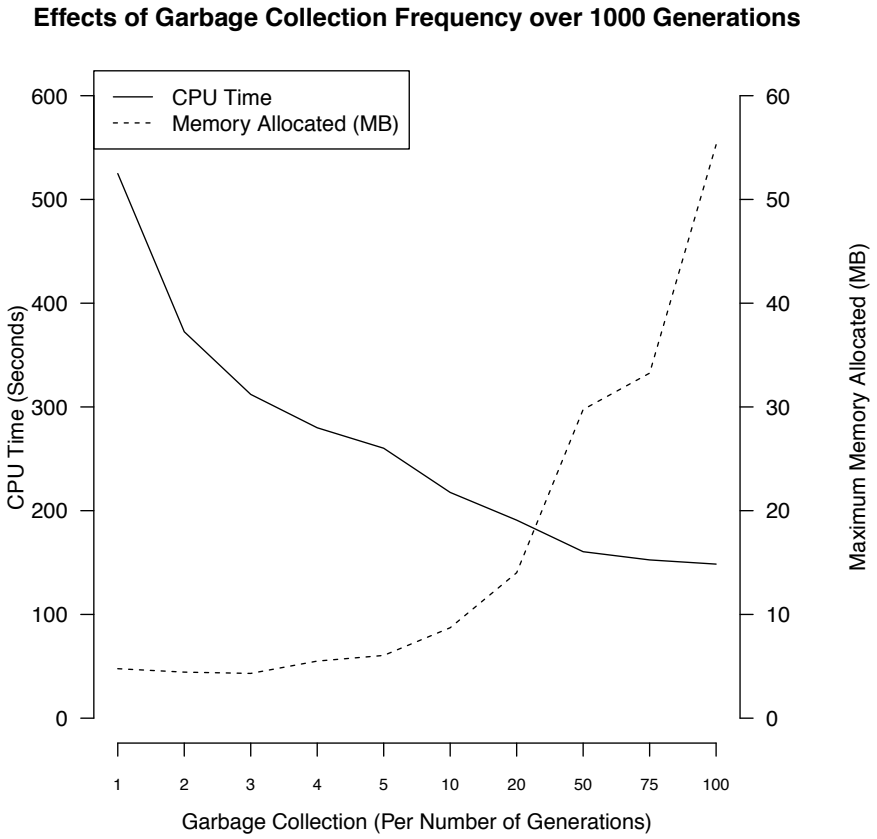
If the user specifies −h on the command-line, the simulation will be populations of hermaphrodites. All selection and demographic parameters must be specified by setting the sex column to ”F.” During hermaphroditic simulations, there will be no sex chromosomes.

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

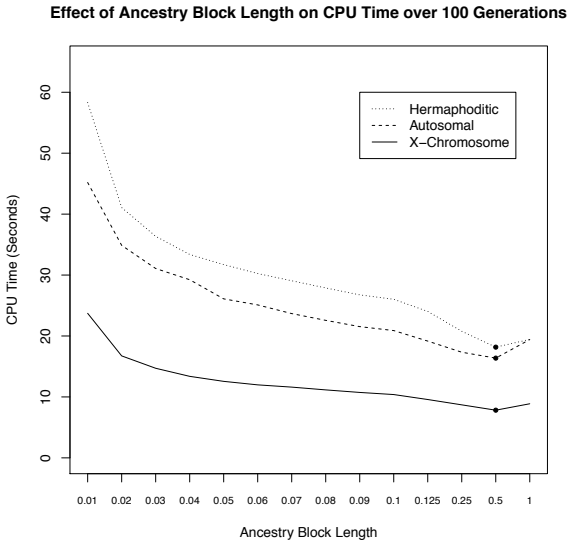
hand, in terms of memory allocation, more frequent garbage collection reduces the amount of memory used. The option is measured in generations per garbage cleanup. For garbage collection to occur every generation, simply use the command --garbage 1. Likewise if the user wants to turn off garbage collection, simply use the command --garbage off. By default, the program undergoes garbage collection every 20 generations. See below for an example of tradeoffs associated with this parameter.

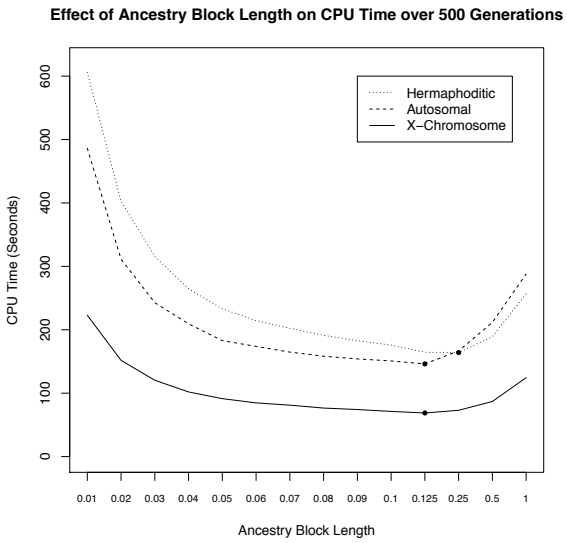


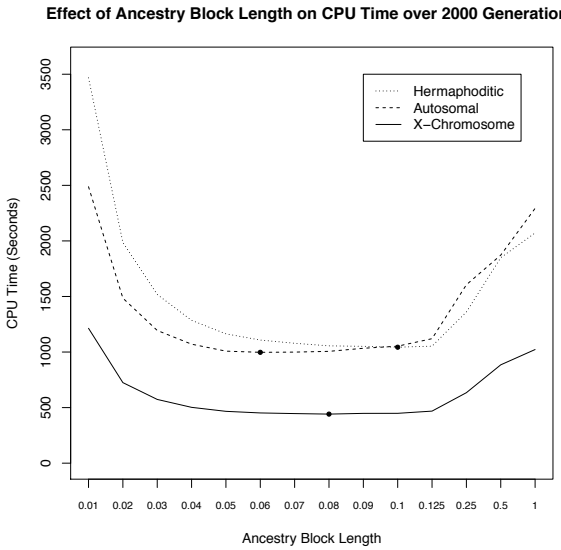
6.2 Ancestry Block Length

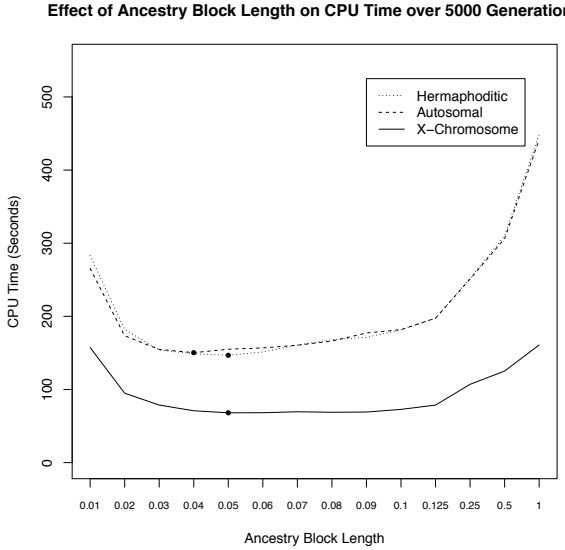
Ancestry Blocks can be thought of as caches that store data pertaining to the recombining chromosomes. The block size of the cache impacts the SELAM's performance. In SELAM, each chromosome is broken up into a number of these ancestry blocks as specified by the ancestry block length (--abl). Given simulation with varying lengths, there is a clear optimal ancestry

block length–very generally, as simulations grow longer, the optimal ancestry block length decreases. Optimizing this parameter can substantially reduce runtimes. The default setting in SELAM is 0.1.





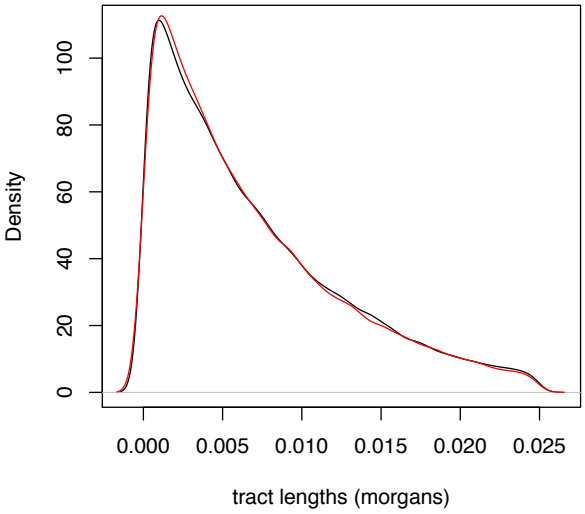


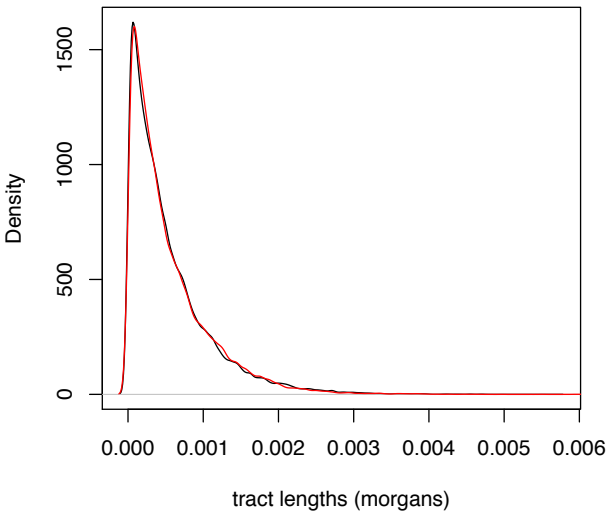


7 Validation

7.1 Comparison to Neutral Coalescent Simulations

To validate that SELAM produces the expected haplotype length distribution in neutral simulations, we first simulated a variety of demographic histories consistent with admixture at varying times in the past using a version of Hudson's MS[1], which has recently been modified to record the local ancestry along a chromosome [2]. We compared across a range of admixture and a variety of different initial ancestry proportions. Shown are two examples for 200 generations of admixture with an initially equal composition of two ancestries (top) and 2000 generations of admixture with a population that is initially 95% ancestry 0 and 5% ancestry 1 (bottom).



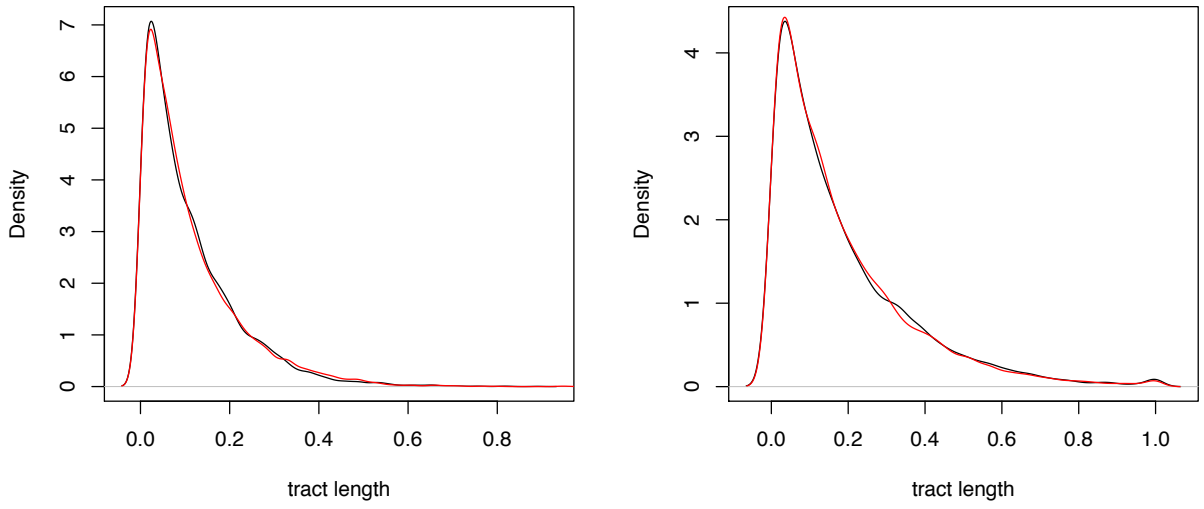


In all simulations considered (additional simulations not shown), the tract length distributions generated using the coalescent and with SELAM are indistinguishable given that a sufficiently large number of tracts were sampled ($p = 0.43$ (top), and $p = 0.78$ (bottom), Kolmogorov-Smirnov test).

7.1 Comparison to Short Term Neutral Forward Simulations

It is well known that the coalescent fails to produce the expected tract length distribution over relatively short time scales. This is due to the relatively small number of individuals within a fixed pedigree[3]. For a detailed discussion of this problem in the context of ancestry tract lengths, see Liang and Nielsen[4]. Hence, to validate SELAM on relatively short time scales, we compare SELAM to the ancestry tract length distribution produced using a simple, and independent, forward time simulation PERL script (provided in the src/ directory within the SELAM repository).

Here, we present results obtained from a population size of 2N=20,000 with 10 generations of admixture at initial ancestry frequencies of 0.1 (right) and 0.5 (left). The tract length distribution obtained using SELAM is in red in both, and the tract length distributions are not significantly different (K-S test, P = 0.635 and P = 0.176). We found that this script and SELAM produce indistinguishable distributions of tract lengths across a variety of admixture times (g = 1—50), and all initial ancestry frequencies considered (0.01-0.5).



Literature Cited:

1. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
2. Sankararaman, S. *et al*. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507,** 354–357 (2014).
3. Wakeley, J., King, L., Low, B. S. & Ramachandran, S. Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent. *Genetics* **190,** 1433–1445 (2012).
4. Liang, M. & Nielsen, R. The Lengths of Admixture Tracts. *Genetics* **197,** 953–967 (2014).