SELAM Stats

Matt Jones
Russ Corbett-Detig

November 3, 2015

1 Overview

SELAM_STATS can be used to summarize SELAM output in a few different ways. This program can provide windowed summaries of allele frequencies, windowed linkage disequilibrium statistics, and genotypes in windows or user-specified positions.

2 Compilation

SELAM_STATS can be compiled using the same makefile as the SELAM program itself. No additional software packages should be necessary. To use the makefile, run the following command via command-line:

$$make \quad -f \quad MAKEFILE$$

3 Using SELAM_STATS

Various statistical values can be computed by specifying options via command line. Table 1 summarizes these options.

Table 1: SELAM_STATS Options

Statistic	Command Line Prefix	Description
Input File	-i	Specify SELAM output file
Windowed Genotypes	-g	Output ancestry types at window intervals across chromosome
Position Genotypes	-G	Output genotype at specific sites
Allele Frequency	-a	Output allele frequency and tract length information in windows
Fisher's Exact Test	-f or - -fet	Compute Fisher's Exact Test over a subset of genotypes
Ancestry Linkage	- -ld	Compute linkage disequilibrium between pairwise combinations of sites
Paternal Linkage Disequilibrium	- -lp	Compute LD over paternally inherited chromosomes at windowed intervals
Maternal Linkage Disequilibrium	- -lm	Compute LD over maternally inherited chromosomes at windowed intervals
Parental Linkage Disequilibrium	- - lpm	Compute LD between parents of a gamete

3.1 Windowed Genotypes

SELAM_STATS can provide genotypes either in evenly spaced windows across the genome or at user specified sites. In the most basic usage,

$$./SELAM\_STATS \quad -i \quad \langle \text{SELAM output file} \rangle \quad -g \quad \langle \text{window size} \rangle$$

SELAM_STATS will output ancestry types across all chromosomes at evenly spaced intervals. For example, this command:

$$./SELAM\_STATS \quad -i \quad \text{selam\_output\_example.txt} \quad -g \quad 0.1$$

Would produce this output (which is abridged for the simplicity of this manual):

Table 2: Example Output

##positions:		0	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
30	0	0	1	0	0	1	0 1	1	0	1	1	1	1	1
30	0	0	1	0	1	0	0 0	0	0	1	1	1	1	1
30	0	0	2	0	0	0	0 0	0	0	0	1	1	1	0
30	0	0	2	0	1	0	1 1	1	1	1	1	0	0	0

Here, the first line denotes the positions of sites at which genotypes are output. It begins with '##positions,' the chromosome, and the positions of genotyped sites on the chromosome. Note that if the SELAM output file contains more than one chromosome, SELAM_STATS will output the positions along each chromosome on two separate lines. The subsequent lines contain individual information, and genotypes. The first column is the generation in which the individual was printed. The second column is the subpopulation from which the individual was sampled. The third column is the sex of the individual (where 0 is female and 1 is male). The fourth column is the index of the individual. The fifth is the chromosome. The sixth column indicates whether the chromosome was inherited from the mother (0) or the father (1), and the next columns correspond to the ancestry type at the positions listed on the first line of the output.

3.2 Genotypes at Specific Positions

Alternatively, SELAM_STATS can output ancestry types at specified sties. To do this, the user must provide SELAM_STATS with a file that specifies the positions of sites for which genotypes will be output. The format of the genotype positions file is a tab delimited file with two values per line, where the first is the chromosome and the second is the position, and each line corresponds to one site for which genotypes will be output. To run, use this command:

$$./SELAM\_STATS \quad -i \quad \langle \text{SELAM output file} \rangle \quad -G \quad \langle \text{genotype position file} \rangle$$

The output will then be identical to that of the windowed genotypes output with the exception that the '##positions' row will be determined by the user, rather than by evenly spaced intervals.

3.3 Ancestry Frequencies and Mean Tract Lengths

SELAM_STATS will also output allele frequencies and tract length information in windows across the genome. Note that SELAM_STATS will compute these population summary statistics across all samples in the input file, regardless of whether they are sampled from the same subpopulation or generation. To compute allele frequency statistics, use the following command:

```
./SELAM_STATS   -i   ⟨SELAM output file⟩   -a   ⟨window size⟩
```

For example, the command:

```
./SELAM_STATS   -i   selam_output_example.txt   -a   0.2
```

would produce the following output:

```
0   0     0.445   0.312526   0.265296   0.104239    0.114652
0   0.2   0.545   0.380198   0.467662   0.0849813   0.0995096
0   0.4   0.51    0.428401   0.493338   0.0924005   0.0977238
0   0.6   0.45    0.466634   0.506144   0.0916934   0.080751
0   0.8   0.495   0.40518    0.449264   0.0876032   0.102302
0   1     0.5     0.312573   0.332555   0.108619    0.121431
```

This output has the following format: The first column is the chromosome, followed by the position, followed by the frequency of ancestry 0 at that site. The fourth and fifth columns are the mean tract length for ancestry 0 and ancestry 1. The sixth and seventh columns are the variance in tract lengths for each ancestry type.

3.4 Ancestry Linkage Statistics

SELAM_STATS can produce a variety of windowed linkage statistics. The basic usage is as follows

```
./SELAM_STATS   -i   ⟨SELAM output file⟩   --ld   ⟨window size⟩
```

The program will then compute linkage disequilibrium statistics between all pairwise combinations of sites genome-wide. For example, the following command

```
./SELAM_STATS   -i   selam_output_example.txt   --ld   0.2
```

would produce this output:

0	0	0	0.2	0.445	0.545	200	0.062475	0.25245
0	0	0	0.4	0.445	0.51	200	0.05305	0.213538
0	0	0	0.6	0.445	0.45	200	0.02975	0.12033
0	0	0	0.8	0.445	0.495	200	0.034725	0.139755
0	0	0	1	0.445	0.5	200	0.0025	0.010061
0	0.2	0	0.4	0.545	0.51	200	0.07705	0.309518
0	0.2	0	0.6	0.545	0.45	200	0.04475	0.180635
0	0.2	0	0.8	0.545	0.495	200	0.000225008	0.000903743
0	0.2	0	1	0.545	0.5	200	0.0175	0.0702853
0	0.4	0	0.6	0.51	0.45	200	0.0855	0.343792
0	0.4	0	0.8	0.51	0.495	200	0.04755	0.190248
0	0.4	0	1	0.51	0.5	200	0.03	0.120024
0	0.6	0	0.8	0.45	0.495	200	0.08725	0.350776
0	0.6	0	1	0.45	0.5	200	0.06	0.241209
0	0.8	0	1	0.495	0.5	200	0.0875	0.350018

The format is also tab-delimited in the following order: chromosome 1, position 1, chromosome 2, position 2, frequency of ancestry 0 at position 1, frequency of ancestry 0 at position 2, sample size, d, and r^2. If the option $-f$ is specified on the command line, SELAM_STATS will output an additional column, which is the p-value obtained by Fisher's Exact Test for the association between ancestry at the two sites. Additionally, the user may wish to compute linkage disequilibrium solely between paternally inherited or maternally inherited chromosomes. In this case, the command line is identical, except the user must specify - -lp for paternal and - -lm for maternal. Finally, the option - -lpm will compute the linkage statistics between paternally and maternally inherited chromosomes.