LASSI-Plus: haplotype frequency spectrum analysis for selective sweeps

V1.0.0

Zachary A. Szpiech, PhD

Introduction

LASSI-Plus implements several haplotype frequency spectrum statistics for detecting and classifying hard and soft selective sweeps. These statistics include the LASSI composite likelihood ratio statistic (Harris and DeGiorgio), H12 (Garud, et al.), H2/H1 (Garud, et al.), G123 (Harris, et al.), and G2/G1 (Harris, et al.). LASSI-Plus accepts VCF and gzipped VCF files, and outputs all files as gzipped flat text.

Source code, precompiled binaries, and example data are available at https://github.com/szpiech/lassip.

Statistics

All of the statistics implemented in LASSI-Plus concern the ranked haplotype frequency spectrum. Define the ranked haplotype frequency spectrum as $\boldsymbol{h} = (h_1, h_2, \cdots, h_N)$ where h_i represents the frequency of the ith most common haplotype in a region, N represents the total number of unique haplotypes, and $h_1 \geq h_2 \geq \cdots \geq h_N$. H12 (Garud, et al.) is then defined as $H_{12} = (h_1 + h_2)^2 + \sum_{i=3}^{N} {h_i}^2$, and H2/H1 is defined as $H_{2/1} = h_2/h_1$. (Harris, et al.) construct G123 and G2/G1 as generalizations of H12 and H2/H1 for unphased genotype data or "multilocus genotypes". A multilocus genotype can be conceptualized as a sequence of homo- and heterozygotes in a region, without regard to phase. The ranked multilocus genotype spectrum is defined as $\boldsymbol{g} = (g_1, g_2, \cdots, g_N)$, where g_i represents the frequency of the ith most common multilocus genotype in a region, N represents the total number of unique multilocus genotypes, and $g_1 \geq g_2 \geq \cdots \geq g_N$. Then, $G_{123} = (g_1 + g_2 + g_3)^2 + \sum_{i=4}^{N} {g_i}^2$, and $G_{2/1} = g_2/g_1$.

(Harris and DeGiorgio) develop LASSI () a composite likelihood ratio statistic based on the ranked truncated haplotype frequency spectrum to compare the genome-wide haplotype spectrum to an alternate haplotype spectrum where mass has been redistributed to m higher frequency classes, attempting to capture the distortion of the haplotype frequency spectrum under a true selective sweep. Full details are available in (Harris and DeGiorgio), but the LASSI method provides both a test statistic T and an inferred parameter $\widehat{m}$, which represents the number of sweeping haplotypes with $\widehat{m} = 1$ indicating a hard sweep and $\widehat{m} > 1$ indicating a soft sweep.

Command Line Options

`--hapstats` <bool>
: Set this flag to calculate haplotype statistics Nuniqe (the number of uniq haplotypes at the locus), and H12 and H2/H1 or G123 and G2/G1 depending on whether –unphased is set.

`--help` <bool>
: Prints help dialog.

`--k` <int>
: Truncation parameter for LASSI statistic. Uses the top K haplotypes for inference.
Default: 10

`--lassi` <bool>
Set this flag to use the LASSI method. Will create a spectra.gz file with truncated haplotype frequency spectra for all windows and, if `--hapstats` is set, other statistics too. Use these files directly or pass all contigs back with `--spectra` to compute m and the CLR test statistic.

`--lassi-choice` <int>
Set this flag to change the way LASSI distributes mass across sweeping haplotype classes. Takes an integer 1-5 which represent options A-E, respectively from (Harris and DeGiorgio).
Default: 4

`--out` <string>
The basename for all output files.

`--pop` <string>
A file containing <ind ID> <pop ID> for all the individuals in the VCF you want to use for computations. Will compute each population separately.

`--spectra` <string1> ... <stringN>
A list of spectra.gz files (typically all contigs that were separately run previously) for finalization of the LASSI inference algorithm. This step computes the genome-wide ranked haplotype frequency spectrum for use as the null hypothesis, and then computes the maximization steps and likelihood ratio for all windows.

`--threads` <int>
The number of threads to spawn during computations.
Default: 1

`--unphased` <bool>
Set this flag to indicate data are unphased. If set output files will have the `mlg` tag applied, otherwise they will have the `hap` tag applied.

`-vcf` <string>:
A VCF file containing haplotype data of a single contig you wish to analyze. Multiple contigs should be provided separately. Variants should be coded 0/1 or missing.

`--winsize` <int>
The window size in SNPs within which to calculate statistics.
Default: 0

`--winstep` <int>
The sliding window step size.
Default: 0

Output Files

If only `--hapstats` is given, files are named
`<basename>.lassip.[hap|mlg].stats.gz` with format:

```
<chr> <start> <end> <nSNPs> <nHaps> <uniqHaps> <h12|g123> <h2h1|g2g1>
```

The final 4 columns are repeated for each population in the analysis, with the population code prepended on the appropriate header label.

If only `--lassi` is given with `--vcf`, files are named `<basename>.lassip.[hap|mlg].spectra.gz` with format:

```
<header line for use with lassip when reading with --spectra>
<chr> <start> <end> <nSNPs> <nHaps> <uniqHaps> <hfs1> ... <hfsK>
```

The final K+2 columns are repeated for each population in the analysis, with the population code prepended on the appropriate header label. hfsN gives the frequency of the Nth most common haplotype in the given window.

If both `--lassi` and `--hapstats` are given with `--vcf`, files are named `<basename>.lassip.[hap|mlg].spectra.gz` with format:

```
<header line for use with lassip when reading with --spectra>
<chr> <start> <end> <nSNPs> <nHaps> <uniqHaps> <h12|g123> <h2h1|g2g1> <hfs1> ... <hfsK>
```

The final K+4 columns are repeated for each population in the analysis, with the population code prepended on the appropriate header label. hfsN gives the frequency of the Nth most common haplotype in the given window.

Passing `*.spectra.gz` files to `lassip` with `--spectra <file1> ... <fileN>` will compute the LASSI CLR and output a single file concatenating the results from all contigs, named `<basename>.lassip.[hap|mlg].out.gz` with one of two formats:

```
<chr> <start> <end> <nSNPs> <nHaps> <uniqHaps> <h12|g123> <h2h1|g2g1> <m> <T>
```

or

```
<chr> <start> <end> <nSNPs> <nHaps> <uniqHaps> <m> <T>
```

Depending on whether `--hapstats` was set when the `*.spectra.gz` files were generated. The final 4/6 columns are repeated for each population in the analysis, with the population code prepended on the appropriate header label. `m` gives the inferred number of sweeping haplotypes, and `T` gives the CLR test statistic.

An Example

The phased whole genome sequence of chromosome 22 of 108 Yoruban individuas from the 1000 Genomes Project is available in the repository to illustrate how to use `lassip`. To test the program and ensure it is functioning execute:

```
lassip --vcf YRI.chr22.vcf.gz --hapstats --lassi --winsize 117 --winstep 12 --out YRI.chr22 --pop
YRI.ids.pop.txt
lassip --spectra YRI.chr22.lassip.hap.spectra.gz --lassi --out YRI.chr22
```

These commands will produce two files: `YRI.chr22.lassip.hap.spectra.gz` and `YRI.chr22.lassip.hap.out.gz`.

A full pipeline for whole autosome analysis of the YRI population, might take the form of the following bash script.

```
for c in $(seq 1 22)
do

  lassip --vcf YRI.chr${c}.vcf.gz --hapstats --lassi --winsize 117 --winstep 12 –out \
  YRI.chr${c} --pop YRI.ids.pop.txt

done

lassip --spectra YRI.chr*.lassip.hap.spectra.gz --lassi --out YRI.chraut
```

References

Garud, N.R.*, et al.* Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet* 2015;11(2):e1005004.

Harris, A.M. and DeGiorgio, M. A likelihood approach for uncovering selective sweep signatures from haplotype data. *Mol Biol Evol* 2020.

Harris, A.M., Garud, N.R. and DeGiorgio, M. Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity. *Genetics* 2018;210(4):1429-1452.