LASSI-Plus: haplotype frequency spectrum analysis for selective sweeps

V1.1.1

Zachary A. Szpiech, PhD and Michael DeGiorgio, 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, 2020), the saltiLASSI extension (DeGiorgio and Szpiech, 2021), H12 (Garud, et al., 2015), H2/H1 (Garud, et al., 2015), G123 (Harris, et al., 2018), and G2/G1 (Harris, et al., 2018). 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., 2015) is then defined as $H_{12} = (h_1 + h_2)^2 + \sum_{i=3}^{N} h_i^2$, and H2/H1 is defined as $H_{2/1} = \sum_2^N h_i^2 / \sum_1^N h_i^2$. Harris, et al. (2018) 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 $G_{2/1} = \sum_2^N g_i^2 / \sum_1^N g_i^2$.

Harris and DeGiorgio (2020) develop LASSI (Likelihood-based Approach for Selective Sweep Inference) 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, 2020), but the LASSI method provides both a test statistic T and an inferred parameter $\hat{m}$, which represents the number of sweeping haplotypes with $\hat{m} = 1$ indicating a hard sweep and $\hat{m} > 1$ indicating a soft sweep.

DeGiorgio and Szpiech (2021) develop an extension to LASSI called saltiLASSI (Spatially Aware Likelihood Test for Improving LASSI) which incorporates a spatial component to the CLR statistic developed by (Harris and DeGiorgio). saltiLASSI reports the test statistic Λ as well as $\hat{m}$. Furthermore, the method infers $\hat{A}$, which is related to the "width" of the sweep. See DeGiorgio and Szpiech (2021) for details.

Command Line Options

`--avg-spec` <bool>
> Set this flag along with `--spectra` to compute and output the average K-truncated HFS from a set of .spectra files. This average spec file can then be passed with `--null-spec` when computing LASSI or saltiLASSI stats.

`--calc-spec` <bool>
> Set this flag to compute K-truncated HFS in windows along the genome. 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 for your own purposes or pass all to lassip with `--spectra` to compute LASSI/saltiLASSI test statistics.

`--dist-type` <stringl>
> Set this flag to `bp`, `cm`, or `nw` to use different distance measures when using saltiLASSI. Stands for using basepairs (physical distance), centiMorgans (genetic distance, must provide a genetic map file with `--map`), or number of windows, respectively. Use at the same time as `--spectra` and `--salti`.
> Default: bp.

`--filter-level` <int>
> Determines how filtering of loci is handled. Filter level 0 means none. Filter level 1 means that any loci that are monomorphic in all samples marked for analysis (without regard to populations) are filtered. E.g. a locus that is monomorphic in all populations analyzed will be filtered, but a locus with a singleton will be retained in all populations. Filter level 2 will filter monomorphic loci at the population level. This almost certainly creates uneven loci numbers across populations, unequal numbers of analysis windows, and window coordinates that do not align across populations. As a results filter level 2 will output a separate file for each population.
> Default: 2

`--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 with `--spectra` to compute LASSI test statistics.

`--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 (2020).
> Default: 4

`--map` <string>
> A genetic map file for saltiLASSI when using centimorgans distance. Should be formatted <chr#> <locusID> <genetic pos> <physical pos>. Sites in data and not present in file will be linearly interpolated.

`--max-extend-bp <double>`
>   Maximum distance in basepairs from core window to consider for saltiLASSI.
>   Default: 100000

`--max-extend-cm <double>`
>   Maximum distance in centiMorgans from core window to consider for saltiLASSI.
>   Default: 0.05

`--max-extend-nw <double>`
>   Maximum distance in number of windows from core window to consider for saltiLASSI.
>   Default: 5

`--null-spec <bool>`
>   A file containing a null K-truncated haplotype spectrum for use computing LASSI or
>   saltiLASSI. Could be created with `--avg-spec`.

`--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.

`--salti <bool>`
>   Set this flag with `--spectra` to compute saltiLASSI test statistics.

`--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 `--calc-spec` 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 `--calc-spec` 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 if `--lassi` is set and the saltiLASSI CLR if `--salti` is set. Output is a single file concatenating the results from all contigs, named `<basename>.lassip.[hap|mlg].out.gz`. For LASSI computations, one of two formats is output:

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

or

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

For saltiLASSI computations, one of two formats is output:

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

or

```
<chr> <start> <end> <nSNPs> <nHaps> <uniqHaps> <m> <A> <L>
```

Depending on whether `--hapstats` was set when the `*.spectra.gz` files were generated. The final columns are repeated for each population in the analysis, with the

4

population code prepended on the appropriate header label. `m` gives the inferred number of sweeping haplotypes, `A` gives the sweep "width" and `T/L` gives the CLR test statistic.

If --filter-level 2 is given (default), loci are filtered on a population basis and all monomorphic snps in a given population are filtered. As a consequence, total number of windows and the window coordinates will not necessarily match between pops, and so each population has results output in a separate file.

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

DeGiorgio, M. and Szpiech, Z.A. A spatially aware likelihood test to detect sweeps from haplotype distributions. *bioRxiv* 2021:2021.2005.2012.443825.
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.