SeleDiff Manual

Xin Huang

Friday, March 16, 2018

Contents

`SeleDiff` is implemented with a probabilistic method for estimating and testing selection (coefficient) differences between populations[1].

If you use `SeleDiff`, please cite

```
Huang X, Jin L, He Y. 2018. SeleDiff: A fast and scalable tool for
estimating and testing selection differences between populations. *In submission*.
```

1 The Model

Consider a bi-allelic locus in population i, let $p_i(t)$ and $q_i(t)$ denote the derived and ancestral allele frequencies at time t, respectively. We can define the absolute fitnesses of the derived and ancestral alleles as w_D and w_A, respectively. Thus, the ratio of the derived and ancestral allele frequencies at time $t+1$ is

$$\frac{p_i(t+1)}{q_i(t+1)} = \frac{w_D}{w_A}\frac{p_i(t)}{q_i(t)}.$$

We then define the relative fitness as

$$e^s = \frac{w_D}{w_A}.$$

We define s as selection coefficient here. Based on our previous study[1], the selection (coefficient) difference between populations i and j is

$$\begin{aligned}\Delta s_{ij} &= s_i - s_j \\ &= \frac{1}{t}\left[\ln\frac{p_i(t)}{q_i(t)} - \ln\frac{p_j(t)}{q_j(t)} + \Omega\right],\end{aligned}$$

where Ω follows a normal distribution with mean zero. We call Ω as population demography parameter, which measures the change of allele frequencies caused by factors except selection. Thus, the expectation of Δs_{ij} is

$$E(\Delta s_{ij}) = \frac{1}{t}\ln\frac{p_i(t)q_j(t)}{q_i(t)p_j(t)} = \frac{1}{t}\ln\frac{N_i(t)M_j(t)}{M_i(t)N_j(t)},$$

where $M_i(t)$ is the count of ancestral allele in population i, and $N_i(t)$ is the count of derived allele in population i. Here, $\ln\frac{N_i(t)M_j(t)}{M_i(t)N_j(t)}$ is the log Odds ratio (OR). The variance of Δs_{ij} is

$$\mathrm{var}(\Delta s_{ij}) = \frac{1}{t^2}\left[\mathrm{var}\left(\ln \mathrm{OR}\right) + \mathrm{var}(\Omega)\right].$$

We can estimate $\mathrm{var}(\ln \mathrm{OR})$ as

$$\mathrm{var}(\ln \mathrm{OR}) \approx \frac{1}{M_i(t)} + \frac{1}{N_i(t)} + \frac{1}{M_j(t)} + \frac{1}{N_j(t)},$$

and $\mathrm{var}(\Omega)$ as

$$\mathrm{var}(\Omega) = \mathrm{median}\left\{\frac{E[\ln^2 \mathrm{OR}(l)]}{0.455} - \mathrm{var}[\ln \mathrm{OR}(l)], n \geq l \geq 1\right\},$$

when there are n neutral loci in the sample, and n is large. If any count of allele is less than 5, then we add 0.5 into it for continuity correction.

We proposed a statistic δ for testing selection in a candidate locus:

$$\delta = \frac{E(\Delta s_{ij})^2}{\mathrm{var}(\Delta s_{ij})}.$$

Under the null (neutral) hypothesis, $E(\Delta s_{ij}) = 0$, thus, δ follows a central χ^2 distribution with one degree of freedom. Under the alternative (selective) hypothesis, δ follows a non-central χ^2 distribution with non-centrality parameter $E(\Delta s_{ij})^2$ with one degree of freedom.

2 Usages

2.1 Environment Setting

To use SeleDiff, you should install Java SE Runtime Enviroment 8 first.

After the installation, you can check Java version in the command line (command starts by “>” prompt).

```
> java -version
java version "1.8.0_25"
Java(TM) SE Runtime Environment (build 1.8.0_25-b17)
Java HotSpot(TM) 64-Bit Server VM (build 25.25-b02, mixed mode)
```

2.2 Installation

To install SeleDiff, you first clone the SeleDiff repository from GitHub.

```
> git clone https://github.com/xin-huang/SeleDiff
```

Then you can enter the SeleDiff directory and use gradlew to install SeleDiff.

```
> cd ./SeleDiff
> ./gradlew build
> ./gradlew install
```

The runnable SeleDiff is in ./build/install/SeleDiff/bin/. You can add this directory into your system environment variable PATH by

```
> export PATH="/path/to/SeleDiff/build/install/SeleDiff/bin/":$PATH
```

You can get help information by typing

```
> SeleDiff
```

and you will get the following:

```
Usage: SeleDiff [command] [command options]
  Commands:
    compute-var      Sub-command for estimating variances of population
            demography parameters
      Usage: compute-var [options]
        Options:
        * --geno
            The EIGENSTRAT GENO file stores allele counts: 0, zero copy of the
            reference allele; 1, one copy of the reference allele and one copy
            of the alternative allele; 2, two copies of the reference allele;
            9, missing values.
        * --ind
            The EIGENSTRAT IND file stores information of individuals and
            populations.
        * --output
            The output file.
        * --snp
            The EIGENSTRAT SNP file stores information of variants.

    compute-diff      Sub-command for estimating selection differences of loci
      Usage: compute-diff [options]
        Options:
        * --geno
            The EIGENSTRAT GENO file stores allele counts: 0, zero copy of the
            reference allele; 1, one copy of the reference allele and one copy
            of the alternative allele; 2, two copies of the reference allele;
            9, missing values.
        * --ind
            The EIGENSTRAT IND file stores information of individuals and
            populations.
        * --output
            The output file.
        * --snp
            The EIGENSTRAT SNP file stores information of variants.
        * --time
            The file stores divergence times between populations. A divergence
            time file is space delimited without header, where the first
            column is the population ID of the first population, the second
            column is the population ID of the second population, the third
            column is the divergence time of this population pair. This file
            is needed when estimating selection differences.
        * --var
            The file stores variances of population demography parameters,
            which is space delimited without header the first column is the
            first population ID the second column is the second population ID
            the third column is the variance of drift of this population pair.
            This file is needed when estimating selection differences.
```

* indicates required options

There are two sub-commands in `SeleDiff`. The first sub-command `compute-var` is used for estimating variances of population demography parameter Ω[1], which are required for the second sub-command `compute-diff`.

2.3 Input Files

2.3.1 EIGENSTRAT

`SeleDiff` accepts EIGENSTRAT format of genetic data as inputs. These files can be compressed by `gzip`.

For EIGENSTRAT format, there are 3 files: **SNP** file, **IND** file and **GENO** file. Consider an example dataset containing 3 unrelated individuals (Ind1, Ind2 & Ind3) from 3 populations (Pop1, Pop2 & Pop3) that were typed on 3 SNPs (SNP1, SNP2 & SNP3):

```
        SNP1 SNP2 SNP3
Ind1    T/T  A/T  T/T
Ind2    C/G  C/G  C/G
Ind3    C/C  A/A  ?/?
```

The **SNP** file describes the information of each SNP. The SNP file corresponding to the example dataset is:

```
SNP1 1 0.1 100 A T
SNP2 1 0.2 101 C G
SNP3 1 0.3 103 C A
```

Each row corresponds to a SNP. The 6 columns are:

1. SNP ID
2. Chromosome number
3. SNP genetic position
4. SNP physical position
5. Reference allele
6. Alternative allele

The **IND** file describes the information of each individual. The IND file corresponding to the example dataset is:

```
Ind1 M pop1
Ind2 F pop2
Ind3 U pop3
```

Each row corresponds to an individual. The 3 columns are:

1. Individual ID
2. Sex: M for male, F for female and U for unknown
3. Population ID

The **GENO** file contains genetic data. The GENO file corresponding to the example dataset is:

```
010
111
209
```

Each row corresponds to a SNP, and each column corresponds to an individual. The characters, 0, 1, 2, 9, correspond to an individual's genotype:

- 0 means zero copies of reference allele.
- 1 means one copy of reference allele.
- 2 means two copies of reference allele.
- 9 means missing data.

2.3.2 Var File

The Var file is the output file from the first sub-command `compute-var`, which stores variances of pairwise population demography parameters. When using sub-command `compute-diff` to estimate selection differences, `SeleDiff` uses `--var` option to accept a a *SPACE* delimited file without header that specifies variances of population demography parameter between two populations.

```
YRI CEU 1.547660
YRI CHS 1.639591
CEU CHS 0.989241
```

The first two columns are the population IDs, and the third column is the variances of population demography parameter of the two populations.

2.3.3 Divergence Time File

When using sub-command `compute-diff` to estimate selection differences, `SeleDiff` uses `--time` option to accept a *SPACE* delimited file without header that specifies divergence times between two populations.

```
YRI CEU 5000
YRI CHS 5000
CEU CHS 3000
```

The first two columns are the population IDs, and the third column is the divergence times of the two populations.

2.4 Output File

The output file from `SeleDiff` is *TAB* delimited. The first row is a header that describes the meaning of each column.

Column	Column Name	Description
1	SNP ID	The name of a SNP
2	Ref	The reference allele
3	Alt	The alternative allele
4	Population1	The first population ID
5	Population2	The second population ID
6	Selection difference	The selection difference between the first and second populations
7	Std	The standard deviation of the selection difference
8	Lower bound of 95% CI	Lower bound of 95% confidence interval of the selection coefficient difference
9	Upper bound of 95% CI	Upper bound of 95% confidence interval of the selection coefficient difference
10	Delta	The δ statistic for selection difference
11	p-value	The p-value of the δ statistic

3 An Example

Here is an example to show how `SeleDiff` estimates and tests selection differences between populations. 4 populations (YRI, CEU, CHB, CHD) from HapMap3 (release3) were extracted. CHB and CHD were merged into one population called CHS. PLINK 1.7 were used to remove correlated individuals and SNPs with minor allele frequencies less than 0.05 and strong linkage disequilibrium. These genome-wide data are stored in `./examples/example.geno` and used for estimating variances of population demography parameters.

Two alternative alleles (rs1800407 and rs12913832) associated with blue eyes were identified in genes *HERC2* and *OCA2*.[2] These candidate data are stored in `./examples/example.candidates.geno` and used for estimating selection differences of these SNPs between populations.

The counts of alleles in our example data were summarized in below.

SNP ID	Population	Reference Allele Count	Alternative Allele Count
rs1800407	YRI	290	0
rs1800407	CEU	207	17
rs1800407	CHS	486	4

SNP ID	Population	Reference Allele Count	Alternative Allele Count
rs12913832	YRI	294	0
rs12913832	CEU	47	177
rs12913832	CHS	491	1

We assume the divergence time of YRI-CEU and YRI-CHS are both 5000 generations, while the divergence time of CEU-CHS is 3000 generations. This information is stored in `./examples/example.time`.

First, we estimate variances of population demography parameters using sub-command `compute-var`.

```
> SeleDiff compute-var --geno ./examples/example.geno \
                       --ind ./examples/example.ind \
                       --snp ./examples/example.snp \
                       --output ./examples/example.var
```

To estimate selection differences, we use the sub-command `compute-diff`.

```
> SeleDiff compute-diff --geno ./examples/example.candidates.geno \
                        --ind ./examples/example.candidates.ind \
                        --snp ./examples/example.candidates.snp \
                        --var ./examples/example.var \
                        --time ./examples/example.time \
                        --output ./examples/example.candidates.results
```

The result is stored in `./examples/example.candidates.results`. The main result is in below.

SNP ID	Population1	Population2	Selection difference	Std	delta	p-value
rs1800407	YRI	CEU	-0.000773	0.000380	4.129	0.042154
rs1800407	YRI	CHS	-0.000336	0.000393	0.731	0.392559
rs1800407	CEU	CHS	0.000728	0.000377	3.730	0.053443
rs12913832	YRI	CEU	-0.001541	0.000378	16.583	0.000047
rs12913832	YRI	CHS	-0.000117	0.000415	0.080	0.777297
rs12913832	CEU	CHS	0.002372	0.000433	30.062	0.000000

From the result, we can see the selection coefficient of rs12913832 in CEU is significantly larger than that in YRI or CHS, which indicates rs12913832 is under directional selection in CEU. While the selection coefficient of rs1800407 in CEU is marginal significantly larger than that in YRI or CHS.

4 Dependencies

- Java 1.8
- Apache Commons Math 3.6
- JCommander 1.72
- t-digest 3.1

5 References

1. He Y, Wang M, Huang X, Li R, Xu H, Xu S, Jin L. 2015. A probabilistic method for testing and estimating selection differences between populations. *Genome Res*, **25**: 1903-1909.

2. Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, Martin NG, Montgomery GW. A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color. *Am J Hum Genet*, **82**: 424-431.