SeleDiff Manual

Xin Huang

Saturday, December 12, 2015

Contents

1 Introduction

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

1.1 Basic Model

Let Φ denote the difference of selection between population A and B, i.e., $\Phi = s_A - s_B$, then

$$\hat{\Phi} = E(s_A - s_B) = \frac{\log(\text{Odds})}{t}$$

where t is the divergence time between A and B, and $\text{Odds} = \frac{C_{A,m}C_{B,w}}{C_{A,w}C_{B,m}}$ (For more details, please see [1]), $C_{A,m}$ denotes the count of derived allele in population A, $C_{A,w}$ denotes the count of ancestral allele in population A, $C_{B,m}$ denotes the count of derived allele in population B, $C_{B,w}$ denotes the count of ancestral allele in population B.

The variance of Φ could be calculated as

$$\text{Var}(\Phi) = \frac{\text{Var}[\log(\text{Odds})]}{t^2} + \text{Var}(\Omega)$$

where Ω is the general effect of genetic drift between population A and B.

Therefore, when a sample has n neutral loci and n is large, the variance of Ω can be estimated as

$$\hat{\text{Var}}(\Omega) = \text{median}\left\{\frac{\hat{\Phi}^2(i)}{0.455} - \frac{\text{Var}[\log(\text{Odds}_i)]}{t^2}\right\}, n \geq i \geq 1$$

where the variance of log-odds ratio could be effectively approximated as

$$\text{Var}[\log(\text{Odds})] = \frac{1}{C_{A,m}} + \frac{1}{C_{A,w}} + \frac{1}{C_{B,m}} + \frac{1}{C_{B,w}}$$

.

When $C < 5$, we do corrections as $C' = C + 0.5$.

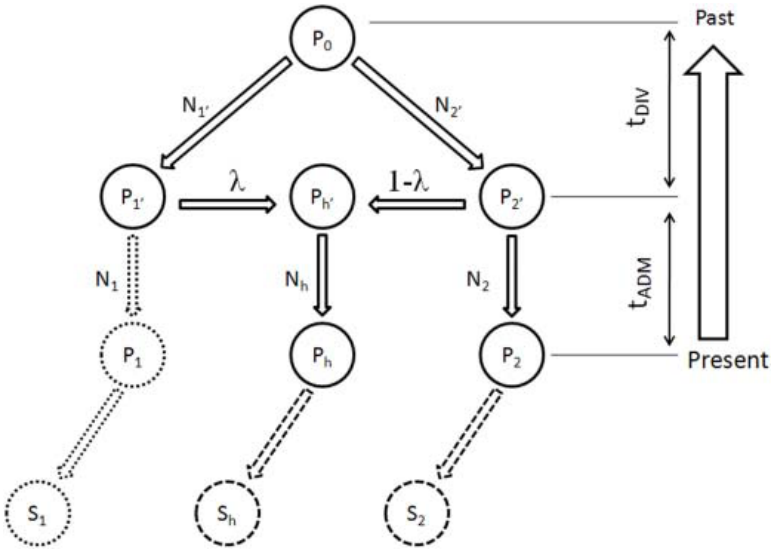
The statistic for natural selection of a candidate locus is

$$\delta = \frac{\hat{\Phi}^2}{\text{Var}(\hat{\Phi})}$$

Under the null hypothesis that differences in natural selection are absent, the statitics δ follows a central chi-square distribution with a degree of freedom = 1. Under the alternative hypothesis with a selection difference the statistic δ has a noncentral chi-square distribution with non-centrality parameter $\hat{\Phi}^2$ and a degree of freedom = 1.

1.2 Correction for Admixed Populations

When



Suppose

$$\mathrm{E}(f) = \frac{m_1/(m_1+m_2) - (1-\lambda)n_1/(n_1+n_2)}{\lambda}$$

$$\mathrm{Var}(f) =$$

$$\begin{aligned}
\mathrm{Var}[\log\left(\frac{f}{1-f}\right)] &\approx \mathrm{Var}\left(\frac{f}{1-f}\right)\left[1/\mathrm{E}\left(\frac{f}{1-f}\right)\right]^2 \\
&\approx \frac{\mathrm{E}^2(f)}{\mathrm{E}^2(1-f)}\left[\frac{\mathrm{Var}(f)}{\mathrm{E}^2(1-f)} - 2\frac{\mathrm{Cov}(f,1-f)}{\mathrm{E}(f)\mathrm{E}(1-f)} + \frac{\mathrm{Var}(1-f)}{\mathrm{E}^2(1-f)}\right]\left[1/\mathrm{E}\left(\frac{f}{1-f}\right)\right]^2 \\
&= \left(\frac{1}{\mathrm{E}(f)} + \frac{1}{\mathrm{E}(1-f)}\right)^2 \mathrm{Var}(f) \\
&= \frac{\mathrm{Var}(f)}{\mathrm{E}^2(f)\mathrm{E}^2(1-f)}
\end{aligned}$$

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)
```

Once you have installed Java SE Runtime Environment 8, you can run `SeleDiff.jar` without any parameter in the command line to look at help information.

```
> java -jar SeleDiff.jar
Usage: SeleDiff [options]
  Options:
    --admixed-population
       A file specifies admixed population.
    --all-gen
       A Oxford GEN file contains all the sample SNPs' information and genotype
       data.
    --all-gen-threshold
       A threshold specifes the confidence of genotype in all the sample data,
       if Oxford GEN/SAMPLE format is used.
       Default: 0.9
    --all-geno
       A EIGENSTRAT GENO file contains all the sample genotype data.
    --all-haps
       A HAPS file contains all the sample SNPs' information and genotype data.
    --all-ind
       A EIGENSTRAT IND file contains all the sample individuals' information.
    --all-sample
       A Oxford SAMPLE file contains all the sample individuals' information.
    --all-snp
       A EIGENSTRAT SNP file contains all the sample SNPs' information.
  * --ancestral-allele
       A file specifies ancestral alleles.
    --candidate-gen
       A Oxford GEN file contains the candidate SNPs' information and genotype
       data.
    --candidate-gen-threshold
       A threshold specifies the confidence of genotypes in the candidate data,
       if Oxford GEN/SAMPLE format is used.
       Default: 0.9
    --candidate-geno
       A EIGENSTRAT GENO file contains the candidate genotype data.
    --candidate-haps
       A HAPS file contains the candidate SNPs' information and genotype data.
    --candidate-ind
       A EIGENSTRAT IND file contains the candidate individuals' information.
```

```
  --candidate-sample
     A Oxford SAMPLE file contains the candidate individuals' information.
  --candidate-snp
     A EIGENSTRAT SNP file contains the candidate SNPs' information.
* --divergence-time
     A file specifies divergence time.
  --haplotype
     A file specifies haplotypes.
  --help
     Show SeleDiff's usage.
     Default: false
  --log
     Redirect log into a file.
* --output
     The output file.
```

* indicates required options.

2.2 Options

Option	Parameter	Description
`--geno`	EIGENSTRAT .geno format	Specify .geno file, required
`--ind`	EIGENSTRAT .ind format	Specify .ind file, required
`--snp`	EIGENSTRAT .snp format	Specify .snp file, required
`--output`	CSV format	Specify output file, required
`--mode`	o, a, s, so, sa, sao	Specify analysis mode, required
`--omega`	CSV format	Specify a file contains variances of pairwise population drift, which can be obtained by performing `--mode o` analysis first
`--admixed-population`	CSV format	Specify a file contains admixed populations, required when performing `--mode a` analysis
`--admixed-proportion`	CSV format	Specify a file contains admixed proportions of each admixed population, which can be obtained by performing `--mode a` analysis first
`--divergence-time`	CSV format	Specify a file contains divergence time of each pair of populations, required when performing `--mode {s, so, sa, sao}` analysis

2.3 Input Files

`SeleDiff` accepts 3 kinds of file formats as inputs. They are EIGENSTRAT format, Oxford GEN/SAMPLE format and HAPS/SAMPLE format.

In order to describe each format, 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  ?/?
```

2.3.1 EIGENSTRAT

For EIGENSTRAT format, there are 3 files: **SNP** file, **IND** file and **GENO** file.

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 Oxford GEN/SAMPLE

For Oxford GEN/SAMPLE format, there are 2 files: **GEN** file and **SAMPLE** file.

The **GEN** file describes the information of each SNPs and its genotypes. The GEN file corresponding to the example dataset is:

```
1 SNP1 100 A T 0 0 1 0 1 0 0 0 1
1 SNP2 101 C G 0 1 0 0 1 0 0 1 0
1 SNP3 103 C A 1 0 0 0 0 1 0 0 0
```

Each row corresponds to a SNP. The first 5 columns are:

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

From the 6th column, every 3 columns correspond to 3 three probabilites of an individual's genotype (AA, AB, BB). If the largest of the probabilities is over the threshold specified by `--all-gen-threshold` or `--candidate-gen-threshold`, then the genotype corresponding to the largest probability is used for `SeleDiff`.

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

```
ID_1 ID_2 missing sex phenotype
0 0 0 D P
Pop1 Ind1 0 M -9
Pop2 Ind2 0 F -9
Pop3 Ind3 0.33 U -9
```

The 1st and 2nd rows are headers. From the 3rd row, each row corresponds to an individual. The 5 columns are:

1. Population ID
2. Individual ID
3. Missing data rate
4. Sex
5. Phenotype

2.3.3 HAPS/SAMPLE

For HAPS/SAMPLE format, there are 2 files: **HAPS** file and **SAMPLE** file.

The **HAPS** file describes the information of each SNPs and its genotypes. The HAPS file corresponding to the example dataset is:

```
1 SNP1 100 A T 1 1 0 1 1 1
1 SNP2 101 C G 0 1 0 1 0 1
1 SNP3 103 C A 0 0 1 1 9 9
```

Each row corresponds to a SNP. The first 5 columns are the same as those in Oxford GEN file.

From the 6th column, every two columns correspond to an individual's genotype. The characters, 0, 1, 9, correspond to alleles of a SNP:

- 0 means reference allele
- 1 means alternative allele
- 9 means missing data

The **SAMPLE** file is the same as the Oxford SAMPLE file.

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/Haplotype ID	The name of a SNP/haplotype
2	Ancestral Allele	The ancestral allele of a SNP/haplotype
3	Derived Allele	The derived allele of a SNP/haplotype
4	Population1	The first population's ID
5	Ancestral Allele Count	The count of the ancestral allele in the first population
6	Derived Allele Count	The count of the derived allele in the first population
7	Population2	The second population's ID
8	Ancestral Allele Count	The count of the ancestral allele in the second population
9	Derived Allele Count	The count of the derived allele in the second population
10	Selection Difference (Population1 - Population2)	The selection difference between the first and second populations
11	Std(Selection Difference)	The standard deviation of the selection difference
12	Divergence Time	The divergence time between the first and second populations
13	log(Odds Ratio)	The logarithm of Odds Ratio
14	Var(log(Odds Ratio))	The variance of the logarithm of Odds Ratio
15	Population Variance	The drift strength $\hat{\text{Var}}(\Omega)$ between the first and second populations without dividing the square of divergence time
16	Delta	The δ statistic for selection difference
17	p-value	The p-value of the δ statistic

Note:

3 Examples

4 References

[1] Yungang He, Minxian Wang, Xin Huang, Ran Li, Hongyang Xu, Shuhua Xu and Li Jin. A Probabilistic Method for Testing and Estimating Selection Differences Between Populations. *Genome Research*, 25:1903-1909, 2015.

[2] Yungang He, Wei R. Wang, Shuhua Xu, Li Jin and Pan-Asia SNP Consortium. Paleolithic Contingent in Modern Japanese: Estimation and Inference using Genome-wide Data. *Scientific Reports*, 2:355, 2012.