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

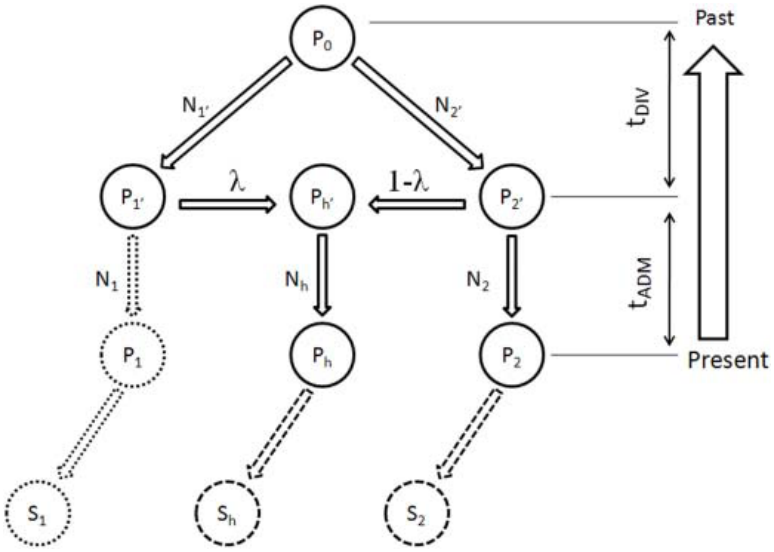
$$\delta = \frac{\hat{\Phi}^2}{\text{Var}(\hat{\Phi})}$$

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

In the basic model, we assume no gene flow in population A and B. When at least one population is admixed, we have to estimate the admixture proportion and use its parental populations instead.

We assume an admixture model as shown in the figure in below. An ancestral population P_0 is split into two parental populations, $P_{1'}$ and $P_{2'}$ (with effective sizes $N_{1'}$ and $N_{2'}$, respectively), which evolve independently for t_{DIV} generations before they contribute genes of proportion λ and $1-\lambda$ to form the hybrid population, $P_{h'}$ with effective sizes N_1, N_2 and N_h, respectively, and evolve independently for t_{ADM} generations before a sample $(S_j, j = 2, h)$ is taken from each of them.[2]



1.2.1 Estimate Admixed Proportion

We estimate admixed proportion for neutral locus i in the admixed population P_h with n loci as

$$\lambda_i = \frac{f_{ih} - f_{i2}}{f_{i1} - f_{i2}}$$

where $f_{ij}(j = 1, 2, h)$ is the allele frequency of locus i in population j.

The overall admixed proportion from population P_1 is estimated as

$$\hat{\lambda} = \text{mean}(\lambda_i)$$

1.2.2 Estimate Allele Frequencies in Missing Parental Populations

When population P_1 is missing, we estimated allele frequency of each SNP in admixed population using an maximum likelihood method. Given a genetic contribution λ and current observation C_{j2} and C_{jh} on locus j for population P_2 and P_h, allele frequency $f_{j1'}$ in the missing ancestral population $P_{1'}$ could be estimated when the recent effective population sizes (N_h and N_2) are large and the admixture event is relatively young.[3]

Given the current observation $C_j(C_{j2}, C_{jh})$ and sample size $S(S_2, S_h)$, we have the probability of observation

$$P(C_{j2}, C_{jh}|f_{j2}, f_{jh}, S_2, S_h) \approx \binom{S_h}{C_{jh}} f_{jh'}(1 - f_{jh'}) \binom{S_2}{C_{j2}} f_{j2'}(1 - f_{j2'})$$

Maximize the probability $P(C_{j2}, C_{jh}|f_{j2}, f_{jh}, S_2, S_h)$, we have estimation for the allele frequency in missing parental population f_{j1} by

$$\hat{f_{j1}} = \max\left(0, \min\left(\frac{C_{jh}/S_h - (1-\lambda)C_{j2}/S_2}{\lambda}, 1\right)\right)$$

Suppose we want to compare population P_1 with another population P_3 that is without admixture. Then

$$\text{Var}[\log(\text{Odds})] = \text{Var}\left[\log\left(\frac{f_{j1}}{1-f_{j1}}\right)\right] + \text{Var}\left[\log\left(\frac{f_{j3}}{1-f_{j3}}\right)\right]$$

Use Δ-method[4],

$$\begin{aligned}
\text{Var}\left[\log\left(\frac{f_{j1}}{1-f_{j1}}\right)\right] &\approx \text{Var}\left(\frac{f_{j1}}{1-f_{j1}}\right)\left[1/\text{E}\left(\frac{f_{j1}}{1-f_{j1}}\right)\right]^2 \\
&\approx \frac{\text{E}^2(f_{j1})}{\text{E}^2(1-f_{j1})}\left[\frac{\text{Var}(f_{j1})}{\text{E}^2(1-f_{j1})} - 2\frac{\text{Cov}(f_{j1}, 1-f_{j1})}{\text{E}(f_{j1})\text{E}(1-f_{j1})} + \frac{\text{Var}(1-f_{j1})}{\text{E}^2(1-f_{j1})}\right]\left[1/\text{E}\left(\frac{f_{j1}}{1-f_{j1}}\right)\right]^2 \\
&= \left(\frac{1}{\text{E}(f_{j1})} + \frac{1}{\text{E}(1-f_{j1})}\right)^2 \text{Var}(f_{j1}) \\
&= \frac{\text{Var}(f_{j1})}{\text{E}^2(f_{j1})\text{E}^2(1-f_{j1})}
\end{aligned}$$

where

$$\text{E}(f_{j1}) = \hat{f_{j1}}$$

$$\text{Var}(f_{j1}) = \frac{1}{\lambda^2}\left[\frac{C_{jh}(S_h - C_{jh})}{S_h^3} + (1-\lambda)^2\frac{C_{j2}(S_2 - C_{j2})}{S_2^3}\right]$$

While

$$\text{Var}\left[\log\left(\frac{f_{j3}}{1-f_{j3}}\right)\right] = \frac{1}{C_{j3}} + \frac{1}{S_3 - C_{j3}}$$

where C_{j3} is the observation in population P_3 and S_3 is the sample size of P_3.

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 Input Files

`SeleDiff` accepts 3 kinds of file formats of genetic data 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  ?/?
```

2.2.1 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.

2.2.2 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

2.2.3 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.

2.2.4 Ancestral Allele File

`SeleDiff` uses `--ancestral-allele` option (required) to accept a *TAB* delimited file that specifies the ancestral allele of each SNP in the data. The content of the file looks like:

```
rsID    Ancestral Allele
rs001   A
rs002   G
...
```

The first line is a header which will be skipped by `SeleDiff`.

2.2.5 Divergence Time File

`SeleDiff` uses `--divergence-time` option (required) to accept a *TAB* delimited file that specifies divergence time between two populations.

```
Population1 Population2 Time(generations)
EastAfrica  WesAfrica   2000
EastAfrica  EastSouthasia(EastAsia) 3600
...
```

The first line is a header which will be skipped by `SeleDiff`. Here, EastSouthasia is a admixed population. We estimate and use its missing parental population in East Asia instead.

2.2.6 Admixed Population File

`SeleDiff` uses `--admixed-population` option to accept a TAB delimited file that specifies admixed populations in the data. The content of the file looks like:

```
Population1 Population2 Desc
EastAsia    Europe  EastSouthasia
EastAsia    Europe  CentralAsia
...
```

The first line is a header which will be skipped by `SeleDiff`. The first two columns (e.g., EastAsia and Europe) are the parental populations we can sample in the present time. The third column (e.g. CentralAsia) is the admixed population. You can find admixture proportions in the log generated by `SeleDiff`.

2.2.7 Haplotype File

`SeleDiff` uses `--haplotype` option to accept a *TAB* delimited file that specifies haplotypes in the data. The content of the file looks like:

```
rs001   rs002
rs001   rs003   rs004
```

Each row represents a haplotype.

2.3 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: For a admixed population, the allele counts of its missing parental populations are estimated by their estimated allele frequecies multiply by 1000 (See Introduction section for estimating allele frequencies in missing parental populations).

3 Examples

Here is an example to show how `SeleDiff.jar` tests and estimates selection differences between populations. 5 populations (YRI, CEU, CHB, CHD, ASW) from HapMap3 (release3) were extracted. CHB and CHD were merged into one population called CHS. Correlated individuals and SNPs which major allele frequencies are less than 0.05 were removed by PLINK 1.07(`--geno 0.01 --maf 0.05`). SNPs in strong linkage disequilibrium were removed, applying a window of 50 SNPs advanced by 5 SNPs and r^2 threshold of 0.01 (`--indep-pairwise 50 5 0.01`) in PLINK. All the genetic data are stored in EIGENSTRAT format.

3.1 Estimate Selection Differences in SNPs

The SNP rs12913832 in gene *HERC2* is associated with blue/non-blue eyes. Its ancestral allele is A and its derived allele is G. The SNP rs1800407 in gene *OCA2* is also associated with blue/non-blue eyes. Its ancestral allele is C and its derived allele is T. The ancestral allele information is stored in `examples/ancestral_alleles.tsv`.

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

SNP ID	Population	Ancestral Allele Count	Derived Allele Count
rs12913832	YRI	294	0
rs12913832	CEU	47	177
rs12913832	CHS	491	1
rs1800407	YRI	290	0
rs1800407	CEU	207	17
rs1800407	CHS	486	4

We assume the divergence time of YRI-CEU and YRI-CHS are both 3600 generations, while the divergence time of CEU-CHS is 2000 generations. This information is stored in `examples/divergence_times.tsv`.

To estimate selection differences, in the command line, we type

```
java -jar SeleDiff.jar --all-geno example.geno --all-ind example.ind --all-snp example.snp --candida
```

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

SNP ID	Population1	Population2	Selection Difference (Population1 - Population2)	Std(Selection Difference)	de
rs12913832	YRI	CEU	-0.00214	3.96E-4	16
rs12913832	YRI	CHS	-1.63E-4	4.54E-4	0.
rs12913832	CEU	CHS	0.003558	4.17E-4	30
rs1800407	YRI	CEU	-0.001073	3.99E-4	4.
rs1800407	YRI	CHS	-4.67E-4	4.15E-4	0.
rs1800407	CEU	CHS	0.001091	2.68E-4	3.

From the result, we can see the selection coefficient of rs12913832 in CEU is significantly higher than that in

YRI or CHS, which indicates rs12913832 is under positive selection in CEU. While the selection coefficient of rs1800407 in CEU is marginal significantly higher than that in YRI or CHS.

When estimating selection differences in admixed populations, we have to correct for its admixed proportions from parental populations. In our example, ASW is an admixed population. We assume its parental populations are YRI and CEU. This information is stored in `examples/admixed_populations.tsv`.

To estimate selection differences, in the command line, we type

```
java -jar SeleDiff.jar --all-geno example.geno --all-ind example.ind --all-snp example.snp --candida
```

From the log information generated by `SeleDiff`, we can see the admixed proportion of ASW from YRI is approximately equal to 0.8. The result is stored in `examples/example.admixed.result.tsv`.

3.2 Estimate Selection Differences in Haplotypes

In the command line, we type

```
java -jar SeleDiff.jar --all-geno example.geno --all-ind example.ind --all-snp example.snp --candida
```

4 Dependencies

- Java 1.8
- Apache Commons Math 3.5
- JCommander 1.48

5 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 Jinliang Wang. Maximum-Likelihood Estimation of Admixture Proportions From Genetic Data. *Genetics*, 164:747-765, 2003.

3 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.

4 Alex Papanicolaou. Taylor Approximation and the Delta Method. http://web.stanford.edu/class/cme308/OldWebsite/notes/TaylorAppDeltaMethod.pdf, 2009.