KLinterSel Manual v 0.1

Antonio Carvajal-Rodríguez
Facultad de Biología, Campus Lagoas-Marcosende
Departamento de Bioquímica Genética e Inmunología
Universidad de Vigo, Vigo 36310, Spain
Email: acraaj@uvigo.es
Web: http://acraaj.webs.uvigo.es

Table of Contents

Versions

Version KLinterSel 0.1 (26 July 2025):

- Multiprocessor options to manage large amounts of memory.

Version KLinterSel 0.0 (July 2025):

- This is the first version.

Introduction

KLinterSel is a Python project for calculating intersections between selective sites detected by different methods. The main script performs operations on genomic data from selective scans and supports statistical tests and plotting.

Pre-built Binaries Requiring No Installation

Pre-built KLinterSel binaries are available at https://github.com/noosdev0/KLinterSel/releases/tag/v1.0. Binaries are provided for Windows, Linux, and macOS (arm64) and should work on most versions of these operating systems.To run them:

Linux: ./KLinterSel_U totsnps.map sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path ./ --perm 1000 --dist 10 --paint

macOS: ./KLinterSel_OS totsnps.map sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path ./ --perm 1000 --dist 10 --paint

Windows: Double click just for running the program under default options. The program will ask for the names of the files. Alternatively, you can go to the command prompt (cmd.exe) and type

```
C:\KLinterSel\KLinterSel.exe totsnps.map sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path ./ --perm 1000 --dist 10 --paint
```

or you can also access the Run command by pressing the Windows logo key ⊞+r

then drag and drop the .exe file from your folder and add the desired arguments, e.g.

```
C:\KLinterSel\KLinterSel_Win.exe totsnps.map sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path ./ --perm 1000 --dist 10 --paint
```

Installation

Clone the KLinterSel repository or download the files. To use the KLinterSel script (KLinterSel.py), you need to have Python installed (version 3.7 or higher). To install the necessary dependencies, navigate to the folder containing the KLinterSel.py script where the requirements.txt file should also be located, and run the following command in the terminal:

```
pip install -r requirements.txt
```

This command will install all required libraries as specified in the requirements.txt file, including numpy, pandas, matplotlib, seaborn, and scipy, along with any other dependencies listed.

Input (Data Format)

The script requires two types of input files:

1.- Original Positions Data File: The first file contains the positions of all analyzed SNPs. A CSV, TSV or text file with the following structure:

```
CHR POS

1 12345

1 12367

.

.
```

The first column identifies the chromosome, and the second column lists the position of the SNP within the chromosome.

In addition, a map file without a header, where the first column corresponds to chromosome numbers and the fourth to physical positions is also accepted:

```
1 rs10181821 5703 5703

.

2 rs11901199 8856 8856

2 rs4637157 19443 19443

.
```

2.- Sites Results Files: Files containing the candidate positions from each method. These files should have the same structure as the original data file but may contain specific formats (e.g., norm, hapflk). Files with a norm extension correspond to the output files from the norm-selscan program. In this case, a single chromosome is assumed, and physical positions with a value of one in the last column are selected. Finally, files with the hpflk extension are also accepted, where the chromosome number column is the second and the positions are in the third column.

Command Line Arguments

Here is a breakdown of the various command line arguments that can be used with the script:

- **Basic Command**

  ```
  python3 KLinterSel.py totsnps.csv sigsmethod1.csv sigsmethod2.csv sigsmethod3.csv
  ```

  This command processes the data without additional options it is equivalent to

  ```
  python3 KLinterSel.py totsnps.csv sigsmethod1.csv sigsmethod2.csv sigsmethod3.csv --path . --perm 10000 --dist 10000 --SL 0.05.
  ```

- **Path Specification**

  --path ./home/b/results/data Specify the directory where the input files are located.

- **Distance option**

  --dist 10000 Specify the distance threshold (default 10,000).

- **Kmax option**

  --Kmax undefined If defined and the number of candidate SNPs in any file is greater than Kmax, it filters SNP positions to group those that are at distance <=D, leaving only the first and last of each group.

- **Number of permutations**

  --perm 10000 Specify the number of permutations to compute the expected distribution of distances (default 10,000).

- **Significance level**

  --SL 0.05 Set the significance level for the permutation test (default 0.05).

- **Plotting**

  --paint Enable the generation of distribution plots. In this case, intersection computations are skipped. The program draws two histograms, the one on the left with the observed distances and the one on the right with the expected distribution.

- **Stats**

  --stats Generates statistics for each chromosome in each file, including minimum, maximum, mean, standard deviation, and median values. No other calculations are performed in this case.

- **No test option**

  --notest Computes intersecctions without performing statistical test.

- **Random control**

  --rand From the SNP file and the input files with the candidate lists, it generates random candidate lists and performs the permutation test, generating the observed distribution (with random controls) and the expected distribution.

Output

The KLinterSel.py script generates several output files that contain results and analyses based on the input genomic data and the methods compared.

Test Results File

- This file contains the results of the statistical tests performed by the script. It includes data related to the significance of intersections between different methods. The contents may include statistical measures such as p-values, scores, and other relevant statistics used for assessing the significance of the intersections found.

Intersections Result File

- This file contains information on the intersections between significant sites detected by the different methods. It typically includes details such as the chromosomal positions where intersections occur, the number of intersections per chromosome, and any other relevant intersection metrics. This file is essential for understanding the overlap and agreement between different detection methods.

Plots (if enabled)

- If the plotting option is enabled, the script generates graphical representations of the observed and expected distance distributions.