KLinterSel Manual v 0.2

Antonio Carvajal-Rodríguez
Centro de Investigación Mariña (CIM)
Departamento de Bioquímica, Genética e Inmunología
Facultad de Biología
Universidad de Vigo, Vigo 36310, Spain
Email: acraaj@uvigo.es
Web: http://acraaj.webs.uvigo.es

Table of Contents

Versions

Version KLinterSel 0.2 (November 2025):

- Major speed-up of the KL computation thanks to a more efficient RelEntr implementation.

- New command-line options:

    --permissive to disable the strict redundancy filter between candidate files.

    --uniform perform resampling assuming a uniform distribution of SNPs.

- Sorted intersection output: elements within each intersection are now returned in sorted order for consistent and cleaner reporting.

- Bug fix: corrected an issue in the resampling test where the comparison of medians introduced a slight conservative bias.

Version KLinterSel 0.1 (August 2025):

- Improvements for figure customization. Multiprocessor options.

Version KLinterSel 0.0 (July 2025):

- This is the first version.

Introduction

KLinterSel is a Python project that performs a Monte Carlo–based test to evaluate whether the distance patterns among candidate sites detected by different methods are closer than expected by chance. It also calculates intersections between these selective candidates and provides complementary plotting and statistical summaries.

Pre-built Binaries Requiring No Installation

Pre-built KLinterSel binaries are available at https://github.com/noosdev0/KLinterSel/releases/tag/v0.2. Binaries are provided for Windows, Linux, and macOS (arm64) and should work on most versions of these operating systems. To run

them (assuming that the files totsnps.txt, sigsmethod1.norm, sigsmethod2.tsv and sigsmethod3.norm exist and are in the same KLinterSel folder as the executable) type in the command-line interface:

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

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

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.txt sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path ./ --perm 1000 --dist 10

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.txt sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path ./ --perm 1000 --dist 10

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, scipy and psutil, along with any other dependencies listed.

To run the .py script type in the command-line interface (assuming that the files totsnps.txt, sigsmethod1.norm, sigsmethod2.tsv and sigsmethod3.norm exist and are in the same folder as the script):

```
python3 KLinterSel.py totsnps.txt sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm
```

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 (.txt) 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.

2.- Candidate Site Results Files: Files containing the candidate positions for each method. These files can have the same format as the original data file, but in addition files with a norm extension are also accepted. These files 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.

Command Line Arguments

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

- **Basic Command**

  python3 KLinterSel.py totsnps.txt sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm

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

  python3 KLinterSel.py totsnps.txt sigsmethod1.norm sigsmethod2.tsv sigsmethod3.norm --path . --perm 10000 --dist 10000 --SL 0.05.

- **Distance option**

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

- **Intersections**

  --chr-id The program calculates intersections for all chromosomes by default. If a valid chromosome number is specified with --chr-id only intersections for that chromosome will be calculated.

- **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 $\leq D$, leaving only the first and last of each group.

- **No test option**

  --notest Computes intersecctions without performing statistical test.

- **Number of permutations**

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

- **Path Specification**

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

- **Permissive with redundant methods**

  --permissive Disables the new default rule that removes candidate files showing high redundancy (i.e., when all or most candidate sites are identical across methods).

- **Plotting**

  --paint Enable the generation of plots. In this case, intersection computations are skipped. The program draws two histograms: the one on the left shows the observed distances, and the one on the right shows the expected distance profile if the matches between methods were generated at random given the genomic locations of SNPs in the input data. If there are multiple chromosomes, the program displays the graph for each chromosome and, after closing it, continues with the next chromosome. If you only want a graph for a specific chromosome, for example chromosome 7, you can specify --paint 7.

  The scale of the X and Y axes can be controlled with the --max-xvalue and --max-yvalue arguments, respectively. The first is the Mb scale (--max-xvalue 20) and the second is a frequency between 0 and 1 (--max-yvalue 0.25).

- **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 (that are random controls) and the expected profiles.

- **Significance level**

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

- **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.

- **Uniform**

  --uniform Perform resampling assuming a uniform distribution of SNPs.

Output

The KLinterSel.py script generates two output files.

Test Results File (testTKL__)

- This file contains, for each chromosome, the results of the statistical test performed by the script. It includes a Kullback-Leibler-like discrepancy measure between the observed and expected ordered distance profiles, its p-value, and the observed (oQ2) and expected (eQ2) medians.

Intersections Result File (INTERSEC_D1E4_)

- This file contains, for each chromosome, the intersections between the different methods. The distance at which the intersections are defined is indicated in the file name itself; for example, _D1E4_ indicates a distance of 10 kb. Intersections are provided for all combinations of methods. If, for example, there are three methods (K=3), the pairwise intersections and the intersection with all three methods are provided. The last column lists the sites involved in the intersection type involving all methods.

Plots (if enabled)

- If the plotting option is enabled, the script generates graphical representations of the observed and expected distance profiles that can be saved to the user's computer.