LFMM version 1.2 - Reference Manual
(Graphical User Interface version)

Eric Frichot[1], Sean Schoville[1], Guillaume Bouchard[2] , Olivier François[1]*

1. Université Joseph Fourier Grenoble, Centre National de la Recherche, TIMC-IMAG UMR 5525,
Grenoble, France
2. Xerox Research Center Europe, Meylan, France

Please, print this reference manual only if it is necessary.

Contents

1 Overview

We proposed an integrated framework based on population genetics, ecological modeling and machine learning techniques for screening genomes for signatures of local adaptation. We implemented fast algorithms using a hierarchical Bayesian mixed model based on a variant of principal component analysis in which residual population structure is introduced via unobserved factors. These algorithms can detect correlations between environmental and genetic variation at the same time as they infer the background levels of population structure. A description of the method is available in our paper:

Eric Frichot, Sean Schoville, Guillaume Bouchard, Olivier François, 2013. *Landscape genomic tests for associations between loci and environmental gradients* Molecular Biology and Evolution, in press.

LFMM has been implemented using the C and C++ programming languages. It contains a command-line engine and a Graphical User Interface (GUI) shell.

The command-line engine is mainly designed for expert users who demand simplicity and flexibility and for users who need to batch-analyze a large amount of data. It accepts data files in lfmm format, including individual environmental information in a separate file. Perl scripts are available to create lfmm format file from ped, eigenstratgeno, or ancestrymap format. The command-line engine produces output in textual format. The textual format stores z-scores and *P*-values for each individual. A set of R and perl scripts is available to compute graphical results (Manhattan plots) without the GUI shell. Textual results can also be inserted in the GUI to analyze them and output graphical results (Manhattan plots). A short manual dedicated to the command-line engine is available.

The LFMM GUI shell is similar with the TESS GUI shell. The LFMM GUI shell can help newbies to familiarize themselves with the software. It is also generally a convenient way to use the LFMM program. It provides facilities for creating and managing projects. A project is a coherent unit which groups the input data, the algorithmic parameter settings, and the output results altogether. Projects are saved in chosen folders automatically. By interacting with the GUI shell, users can check their data, specify the parameter settings, run the MCMC algorithm, and visualize the results without mastering the usage of the command-line engine. The LFMM GUI also provides a couple of additional features. It can perform multiple runs with distinct numbers of latent factors and distinct environmental variables. It can display a summary of all runs and sort them by their values of the Deviance Information Criterion (DIC).

We advise you to read the documentation and in particular, to follow the tutorial.

2 Method Description

Adaptation to local environments often occurs through natural selection acting on large number of alleles, each having a weak phenotypic effect. One way to detect those alleles is by identifying genetic polymorphisms that exhibit high correlation with some environmental gradient or with the variables used as proxies for ecological pressures. Here we proposed an integrated framework based on population genetics, ecological modeling and machine learning techniques for screening genomes for signatures of local adaptation. We implemented fast algorithms using a hierarchical Bayesian mixed model based on a variant of principal component analysis in which residual population structure is introduced via unobserved or latent factors.

Our algorithms can detect correlations between environmental and genetic variation at the same time as they infer the background levels of population structure. We provided evidence that latent factor models efficiently estimated random effects due to population history and isolation-by-distance mechanisms when computing gene-environment correlations, and that they decreased the number of false-positive associations in genome scans for selection. We applied these models to plant and human genetic data and we detected several genes with functions related to multicellular organismal development exhibiting unusual correlations with climatic gradients.

How to choose K, the number of latent factors. The number of latent factors is the number of principal components (or latent factors) that is required to describe the neutral structure of the data. Several values should be tested. A too small value of K leads to liberal tests and may generate False Positive results. A too large value of K leads to conservative tests and may generate False Negative results. In our paper, we used the number of significative principal components in the Tracy-Widom test of `SmartPCA` (http://helix.nih.gov/Applications/eigensoft.html) [3]. This heuristic may be a bit too conservative. We also used the Bayesian clustering programs `STRUCTURE` (http://pritch.bsd.uchicago.edu/software/structure2_1.html) [5] and `TESS` (http://membres-timc.imag.fr/Olivier.Francois/tess.html) [1, 2] to find K the number of components which could better describe our simulated data. We advise you to be really careful in the choice of K and to test several values of K.

3 Installation

R is mandatory to display manhattan plot.
Tips: Linux and Mac versions may be faster than Windows version because they are compiled on your computer. Also, the command-line version is faster than the Graphical User Interface version (especially for big data sets, more than 10 000 SNPs).

3.1 under Linux

To install LFMM GUI version, you just have to execute the install script (install.sh) in LFMM main directory. To do this, you can double-click on it or execute it in a terminal shell.

- In the case you double-click on the install file, a box is displayed with written "Do you want to run "install.sh", or display its contents?". Click on "Run in Terminal". Then, LFMM is installed with the terminal. An executable called `LFMM_GUI` should be displayed in the main directory. If the box is not displayed, it may be due to the fact the the script in not executbable. To make install.sh executable, right-click on the install file. Select "Properties". In panel "Permissions", select "Allow executing file as program".

- In the case you execute it in a terminal shell, go to LFMM main directory and write "./install.sh". If the script is not executable, type "chmod +x install.sh" and then "./install.sh".

A binary called `LFMM_GUI` should be created in LFMM main directory. Please, do not move `LFMM_GUI` file from LFMM main directory.

3.2 under Mac

To install LFMM GUI version, you just have to double-click the install script (`LFMM_GUI.install`) in LFMM main directory. A binary called `LFMM_GUI` should be created in LFMM main directory.

3.3 under Windows

Windows version is a compiled version. There is no installation. Just double-click on `LFMM_GUI.exe` in LFMM main directory. In compensation, this version may be slower than Linux or Mac versions. Please, do not move `LFMM_GUI.exe` file from LFMM main directory. Windows version is not multithreaded.

4 Data Format

Input files are composed of two mandatory files (a **genotype file** and an **environmental variable file**) and one optional file (the **snp information file**). The snp file is interesting to analyze zscore results and to display results with manhattan plots. It is not necessary to provide information about individuals. All data formats are described with the same example. These files are available in `example/format_example/`. Each file should end with its format name.

4.1 Genotype Data

- lfmm (example.lfmm)
  The genotype file is 1 line per individual. There is 1 genotype column for each SNP (in the order the SNPs are specified in the snp file). Each element can be 0, 1 or 2. A missing element is identified by the value 9 or −9. Each element of the matrix is separated by a single space. There should be no space after the last value of each line. Lines containing only missing data (−9 or 9) should be removed.

```
1 0 0 1
1 1 -9 2
2 0 1 1
```

- ped (example.ped)
  The genotype file is 1 line per individual. Each line contains 6 columns of information about the individual, plus two genotype columns for each SNP in the order the SNPs are specified in the snp file. Genotype format must be either 0ACGT or 01234, where 0 means missing data. The first 6 columns of the genotype file are: 1st column is family ID, 2nd column is sample ID, 3rd and 4th column are sample IDs of parents, 5th column is gender (male is 1, female is 2), 6th column is case/control status (1 is control, 2 is case), quantitative trait value or population group label. In the two genotype columns for each SNP, missing data is represented by 0.

```
     1      SAMPLE0 0 0 2 2  1 2  3 3  1 1  2 1
     2      SAMPLE1 0 0 1 2  2 1  1 3  0 4  1 1
     3      SAMPLE2 0 0 2 1  2 2  3 3  1 4  1 2
```

- ancestrymap (example.ancestrymap)
  The genotype file contains 1 line per valid genotype. There are 3 columns: 1st column is SNP name,

2nd column is sample ID, 3rd column is number of reference alleles (0 or 1 or 2), Missing genotypes are encoded by the value −9 or 9 in the genotype file.

```
              rs0000              SAMPLE0   1
              rs0000              SAMPLE1   1
              rs0000              SAMPLE2   2
              rs1111              SAMPLE0   0
              rs1111              SAMPLE1   1
              rs1111              SAMPLE2   0
              rs2222              SAMPLE0   0
              rs2222              SAMPLE1   -9
              rs2222              SAMPLE2   1
              rs3333              SAMPLE0   1
              rs3333              SAMPLE1   2
              rs3333              SAMPLE2   1
```

- eigenstratgeno (example.eigenstratgeno)
  The genotype file contains 1 line per SNP. Each line contains 1 character per individual: 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.

```
112
010
091
121
```

Tips: As LFMM does not model allele frequencies, genotype file can be the number of copy of either the reference allele or the derived allele.

4.2 Snp Data

Warning: SNP data information should be in the same order as in the genotype data file.

- pedsnp (example.pedsnp or example.map)
  The snp file contains 1 line per SNP. There are 6 columns (last 2 optional): 1st column is chromosome. Use X for X chromosome, 2nd column is SNP name, 3rd column is genetic position (in Morgans), 4th column is physical position (in bases), Optional 5th and 6th columns are reference and variant alleles.

```
11       rs0000     0.000000            0 A C
11       rs1111     0.001000       100000 A G
11       rs2222     0.002000       200000 A T
```

- snp (example.snp)
  The snp file contains 1 line per SNP. There are 6 columns (last 2 optional): 1st column is SNP name, 2nd column is chromosome. Use X for X chromosome, 2nd column is SNP name, 3rd column is genetic position (in Morgans) (If unknown, ok to set to 0.0), 4th column is physical position (in bases), Optional 5th and 6th columns are reference and variant alleles.

```
               rs0000  11       0.000000            0 A C
               rs1111  11       0.001000       100000 A G
               rs2222  11       0.002000       200000 A T
```

- lfmmsnp (example.lfmmsnp)
  The snp file contains 1 line per SNP. There are 3 columns: 1st column is SNP name, 2nd column is chromosome, 3th column is physical position (in bases).

```
               rs0000  11       0
               rs1111  11       100000
               rs2222  11       200000
```

Tips: SNP data information is not mandatory. But if you have it, we advise you to provide it. It is useful for post-treatment of LFMM analysis.

4.3 Environmental Data

The variable file is a vector composed of n lines and D columns. Each line is the values of the D variables for the corresponding individual. Below, an example of variable file for $n = 3$ individuals and $D = 2$ covariables. Environmental information has to be in the same order of individuals as the one provided in the genotype data file.

```
0.252477 0.0259401
0.216618 0.00908548
-0.47509 0.979297
```

4.4 Output File

Output file for 1 environmental variable is composed of 1 line for each SNP. Each line is composed of 3 columns. 1st column z-score, 2nd column is the $-log_{10}(pvalue)$, and 3rd column is the pvalue for the corresponding loci. Below, an example of output file for $L = 4$ loci.

```
0.0259401 0.00908548 0.979297
0.0616506 0.02191 0.950802
0.0210902 0.0073732 0.983166
0.00991587 0.00346154 0.992061
```

5 Using the GUI Shell

Screenshots presented in this section were taken under Linux. The actual GUI may differ visually a little bit under Mac interface.

The GUI shell provides a convenient way to use LFMM. It also helps newbies to familiarize themselves with the software. With help of the GUI shell, there is no need for users to understand and remember the command-line options. The GUI shell calls the command-line engine internally and present the analytical results to users visually. The GUI shell can be launched by double-clicking on `LFMM_GUI` in the LFMM GUI home directory.

5.1 The User Interface

As shown in Figure 1, the GUI is composed of a menu, a toolbar, a text information box, a project tree, and a central zone.

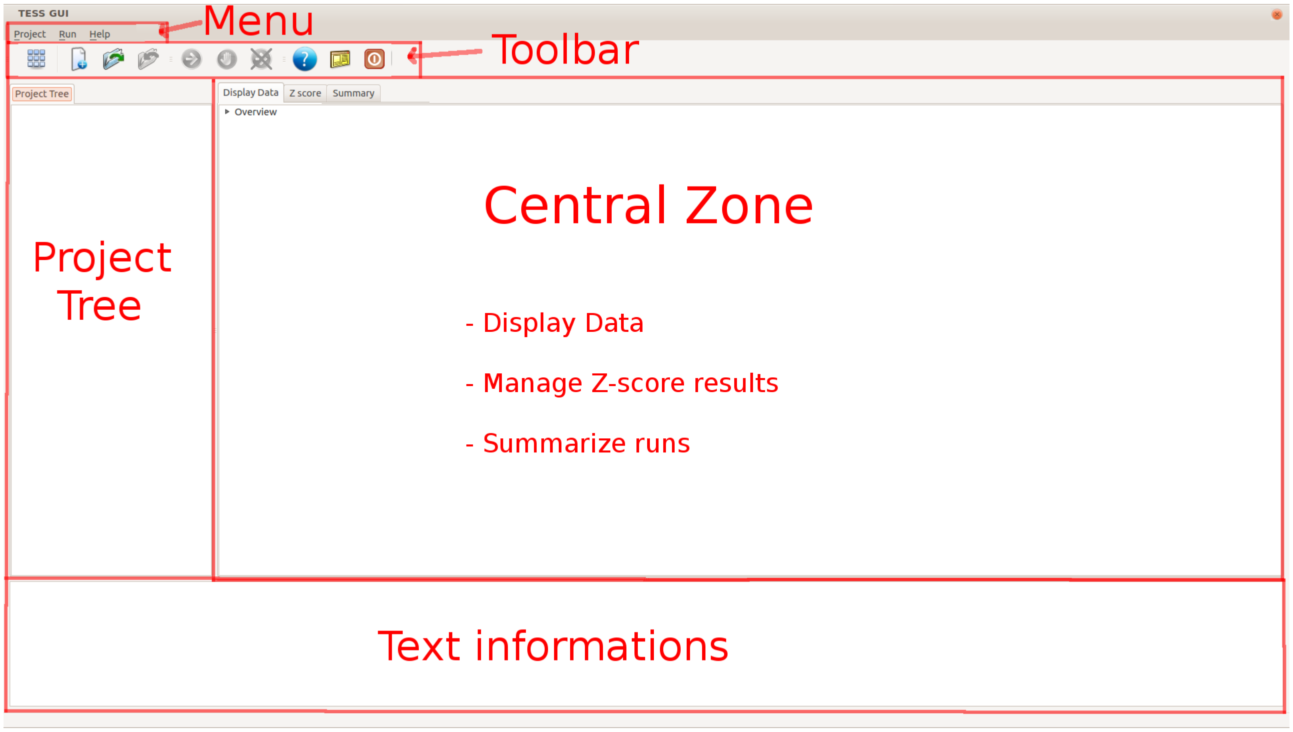


Figure 1: Main Window

The following parts describe each part of LFMM GUI.

5.1.1 The Menu

Figure 2: Menu description

The menu contains a set of possible actions. These actions are organized in 3 groups:

- a Project manager: to create, open, save, and close projects, and to open textual results,
- a Run manager: to set, abort and remove runs,
- a Menu.

5.1.2 The Toolbar

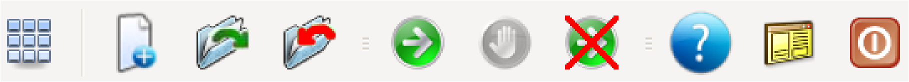


Figure 3: Toolbar description

The toolbar contains the same set of possible actions as the menu. Here is a description of the action associated with each button:

- ▦ opens any data file,
- 🗋 starts a new project,
- 📂 opens an existing project,
- 📂 closes the current project (It is not mandatory to close the current project to open or create a project),
- ⏻ saves the current project and closes the GUI,
- ➡ starts a new run,
- ✋ aborts the current run,
- ✖ removes a run from the current project (To do it, just select in the browse panel, the directory to remove),
- ? displays the reference manual,
- 🗔 displays general information about LFMM GUI.

5.1.3 Textual Information

The textual information displays all information printed by programs called by the GUI.

- it checks that input files were read correctly,
- it informs you step-by-step of the execution of the current run,
- it checks that a manhattan plot is correctly generated.

Please, if anything went wrong, check the textual information box to get more information.
Tips: When you start a run, the corresponding command line for the run is printed in the textual information panel.

5.1.4 The project Tree

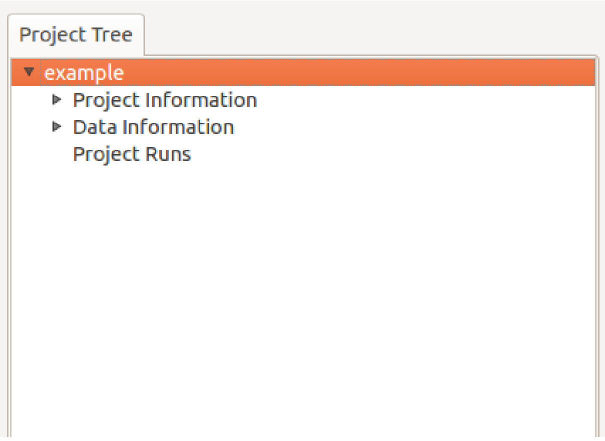


Figure 4: Project Tree global view

The project tree summarizes the current project information. As show in figure 4, the project tree information is grouped in

- the project information: name of the project, date of creation, path to the project, the data file in lfmm format, the data file in original format, the environmental file, and if provided, the SNP name file. the data information, and information for each run.
- the data information: number of individuals, and number of loci.
- information for each run: algorithm, number of latent factors K, total number of sweeps, burn in number of sweeps, and number of processes used, the textual result file, deviance criteria, and DIC criteria.

Tips 1: To display a data file, you can double-click on it in the project tree. It is displayed in the "Display data" tab of the central zone.
Tips 2: If you double-click on a textual result file, it loads this result file and displays it in the Z-score tab of the central zone.

5.1.5 The Central Zone

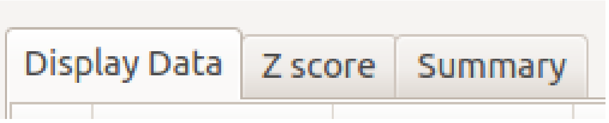
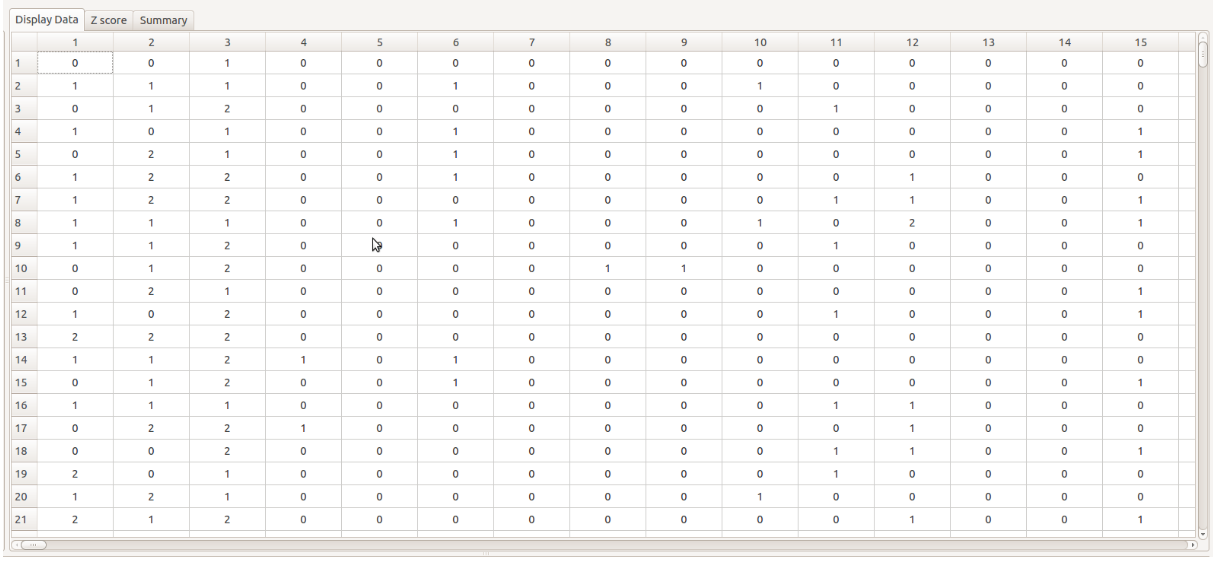


Figure 5: Tabs of central zone

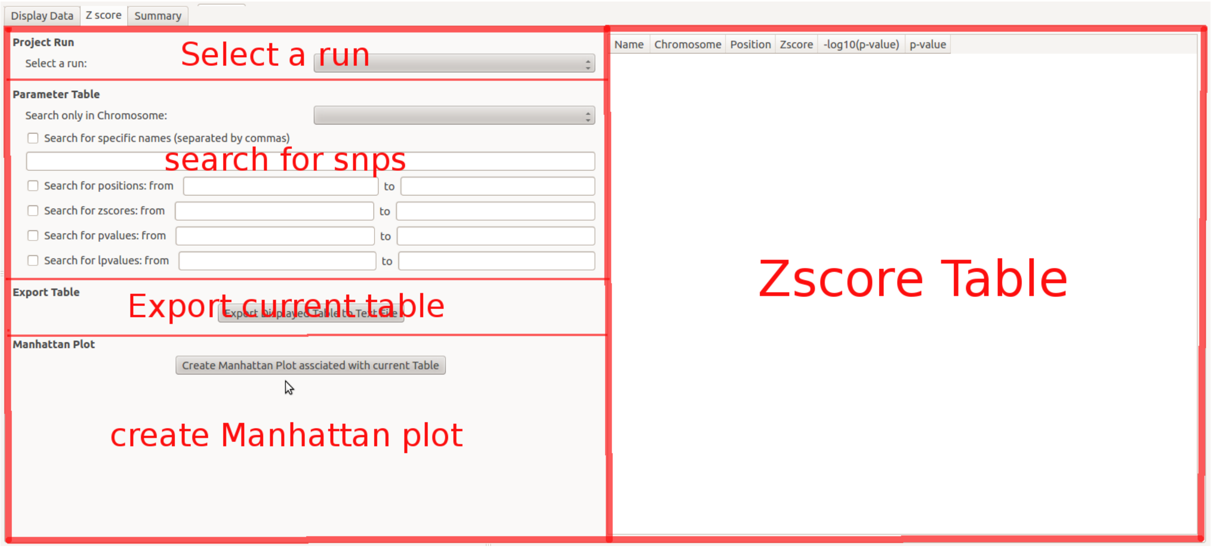
The central zone is the main part of the GUI. As shown in figure 5, it is divided in 3 different panels (Display Data, Z-score, and summary). Each panel is described below.

Display Data This panel displays the data files. Each time you click on a data file in the Project Tree, the data is displayed here.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	0	0	1	0	0	0	1	0	0	0	0	0
3	0	1	2	0	0	0	0	0	0	0	1	0	0	0	0
4	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1
5	0	2	1	0	0	1	0	0	0	0	0	0	0	0	1
6	1	2	2	0	0	1	0	0	0	0	0	1	0	0	0
7	1	2	2	0	0	0	0	0	0	0	1	1	0	0	1
8	1	1	1	0	0	1	0	0	0	1	0	2	0	0	1
9	1	1	2	0	0	0	0	0	0	0	1	0	0	0	0
10	0	1	2	0	0	0	0	1	1	0	0	0	0	0	0
11	0	2	1	0	0	0	0	0	0	0	0	0	0	0	1
12	1	0	2	0	0	0	0	0	0	0	1	0	0	0	1
13	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	2	1	0	1	0	0	0	0	0	0	0	0	0
15	0	1	2	0	0	1	0	0	0	0	0	0	0	0	1
16	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
17	0	2	2	1	0	0	0	0	0	0	0	1	0	0	0
18	0	0	2	0	0	0	0	0	0	0	1	1	0	0	1
19	2	0	1	0	0	0	0	0	0	0	1	0	0	0	0
20	1	2	1	0	0	0	0	0	0	1	0	0	0	0	0
21	2	1	2	0	0	0	0	0	0	0	0	1	0	0	1

Figure 6: Display data panel



Project Run

Select a run:

Parameter Table

Search only in Chromosome:

□ Search for specific names (separated by commas)

□ Search for positions: from to

□ Search for zscores: from to

□ Search for pvalues: from to

□ Search for lpvalues: from to

Export Table

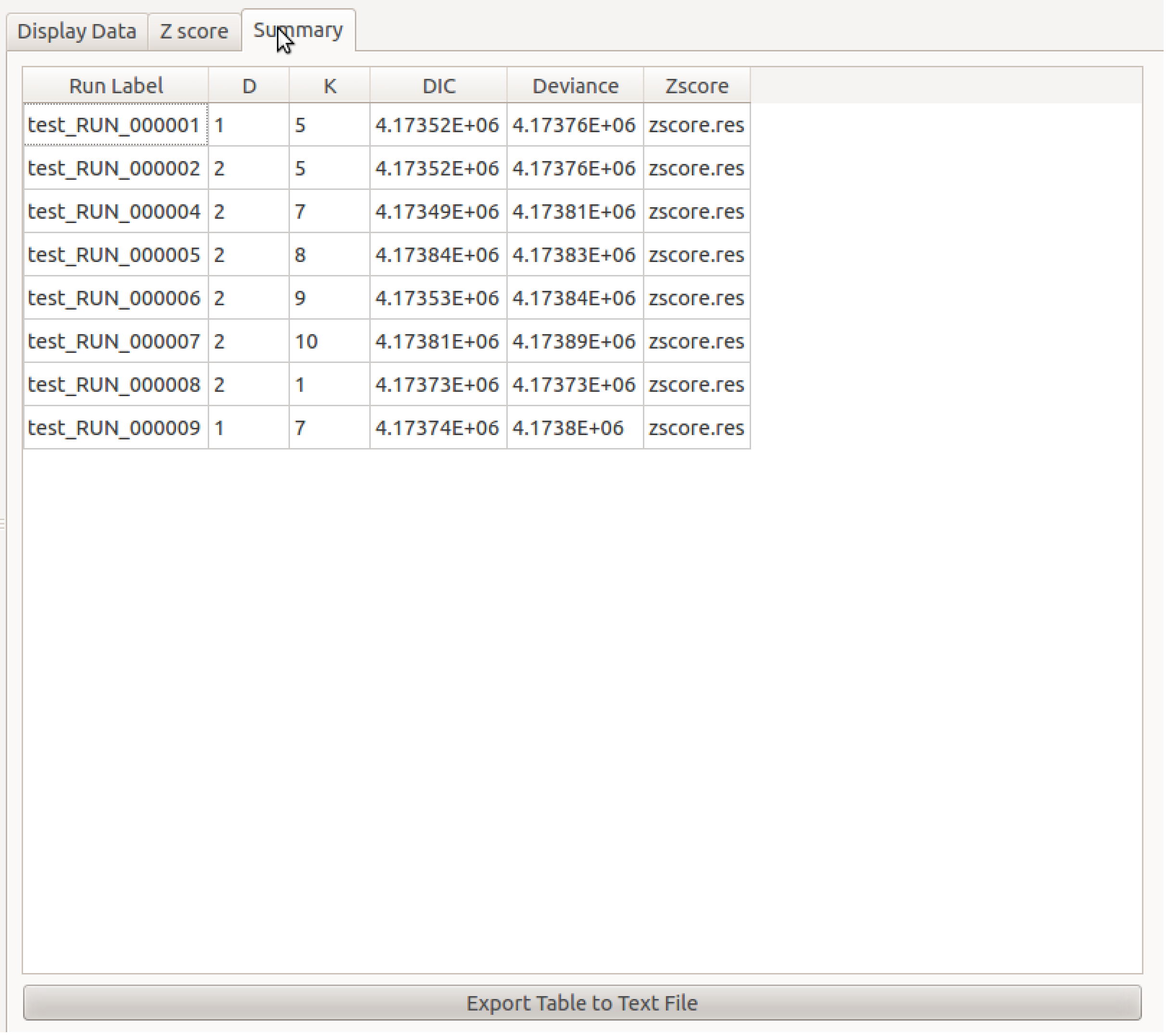
Manhattan Plot

Name	Chromosome	Position	Zscore	-log10(p-value)	p-value

Figure 7: Z-score panel

Z-score Results can be analysed in the panel. As shown in figure 6, you can display your results in a table.
Tips: If you did not provide a snp data file, SNPs are ordered in the same order as they are given in the genotype data file. The i^{th} SNP is called SNP_*i* at position *i*. In this case, all SNPs are in chromosome 0.
With this tool, you can

- select the run of the project that you want to analyze. If you double-click on a specific zscore file in the Project tree, it automatically shows the associated zscore table.
- know the number of SNPs currently displayed in the table.
- search for a specific snp or a group of SNPs in a single chromosome by selecting its names or a range of positions, zscores, pvalues, or $-log_{10}$(pvalue). You can select or unselect a criteria by using the checkbox at the beginning of each criteria. For example, if want to select all SNPs in with a zscore upper than 3. You write 3 in the first bound box of zscore line and check the corresponding checkbox.
- order them by name, Chromosome Position, Zscore, pvalue, and $-log_{10}(pvalue)$.
- export the current displayed zscore table in a text file.
- display the manhattan plot associated with the current zscore table. The manhattan plot is displayed with a R script in pdf. If the pdf is not displayed, you can probably find it in your directory of you project with the name "manhattan_plot.pdf". If the pdf is displayed, we advise you to use your own pdf viewer to register it in a proper name and space. The upper button for manhattan plot displays a manhattan plot with only SNPs currently in the zscore table. For example, if you selected a specific chromosome and a specific range of positions, it displays only these SNPs. The lower button for manhattan plot displays all the SNPs and black and grey and it underlines in green those which are currently in the zscore table. For example, it can underline a manhattan of SNPs (ie a group of SNPs in the same region with significant correlation).



Run Label	D	K	DIC	Deviance	Zscore
test_RUN_000001	1	5	4.17352E+06	4.17376E+06	zscore.res
test_RUN_000002	2	5	4.17352E+06	4.17376E+06	zscore.res
test_RUN_000004	2	7	4.17349E+06	4.17381E+06	zscore.res
test_RUN_000005	2	8	4.17384E+06	4.17383E+06	zscore.res
test_RUN_000006	2	9	4.17353E+06	4.17384E+06	zscore.res
test_RUN_000007	2	10	4.17381E+06	4.17389E+06	zscore.res
test_RUN_000008	2	1	4.17373E+06	4.17373E+06	zscore.res
test_RUN_000009	1	7	4.17374E+06	4.1738E+06	zscore.res

Figure 8: Summary panel

Summary This panel provides a global view of the runs. As shown in figure 8, it displays a summary table of the runs. In the table, you can find information about each run, such as the label, the variable used, the number of latent factors, the DIC criteria, the deviance criteria, and the name of the Zscore file. You can also export this table in text format. We advise you to be really careful in the interpretation of DIC criteria. For example, we do not advise you to use DIC criteria to select the number of latent factors K.

5.2 New Project

When using the GUI shell, users work with projects. A project is a coherent unit which groups the input data, the algorithmic parameter settings, and the output results altogether. To create a new project, access the menu "Project=New Project..." or the corresponding button on the tool bar (see Figure 3). The GUI shell shows the "New Project" dialog box (see Figure 9).

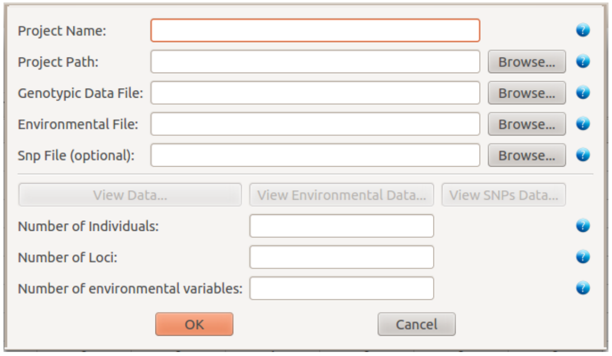


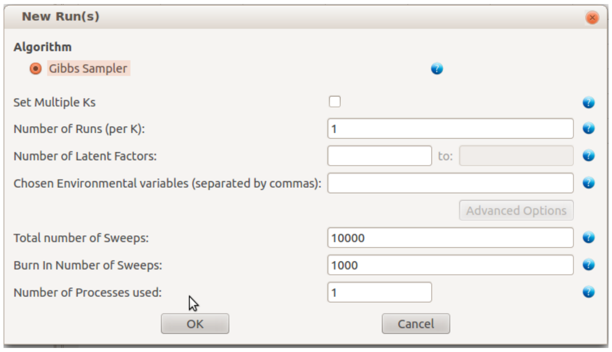
Figure 9: new project

The user should name his/her project and choose the project path and data (tip: using the "Browse..." is a convenient way to input this information). Note that LFMM requests its user to put his/her data in the same directory of the project file itself. We also recommend users to store different projects using separate directories for clear organization of information, although LFMM does not require this.

The user also needs to input the number of individuals, number of loci and number of environmental variables. He/she can also enter a snp data file. This file is optional. If the user is not clear of this information, he/she can always click the "View Data..." button to check the data first (see Figure 9). Users should input information and the format of the data with care. Although the software can catch most common errors, wrong inputs may result in strange analytical results.
Tips: If you did not provide a snp data file, SNPs are ordered in the same order as they are given in the genotype data file. The i^{th} SNP is called SNP_i at position i. In this case, all SNPs are in chromosome 0.

When inputs are finished, click on the "OK" button to confirm the creation of the new project. When the project is created, the GUI shell checks that all requirements are followed. Then, it converts the data file in lfmm format, automatically load it, and show its data to the user. If the dataset is a bit large (more than 5000 SNPs), the conversion in lfmm format can take a few minutes.

5.3 New Run



Algorithm

● Gibbs Sampler

Set Multiple Ks □

Number of Runs (per K): 1

Number of Latent Factors: to:

Chosen Environmental variables (separated by commas):

Total number of Sweeps: 10000

Burn In Number of Sweeps: 1000

Number of Processes used: 1

Figure 10: New run box

The user can then start a run by accessing the menu "Project=Run..." or the corresponding button on the tool bar (see Figure 3). The GUI shell shows the "New Run(s)" dialog box asking the user to key in the required run information (see Figure 9). This box allows users to choose

- the algorithm
- the number of latent factors that you want to use. You can launch several runs at the same time with the same value of K or different values of K.
- a list of environmental variable number. A run is launched sequentially for each environmental variable number listed (Ex: 1,2,5. A run is launched for the first, the second and the fifth environmental variable).
- the total and burn in number of sweeps.
- the number of processes. Choosing several processes multithreads the algorithm and by consequence speed up the runs. The result is the same with one or several processes. To ensure the algorithm to be speeded up, we advise you to choose the number of processes at most equal to the number of processes of your computer.

6 An example/Tutorial

All examples are available in the directory example. The goal of this tutorial is to introduce the potential of this software.

6.1 Data Set

The data set that we analyze in this tutorial is an Asian human data set of SNPs data. This data is a worldwide sample of genomic DNA (10757 SNPs) from 388 individuals, taken from the Harvard Human Genome Diversity Project - Centre Etude Polymorphism Humain (Harvard HGDP-CEPH)2 . In those data, each marker has been ascertained in samples of Mongolian ancestry (referenced population HGDP01224) [4].

We selected all samples from Asia. Using Tracy-Widom tests implemented in SmartPCA [3], we found that the number of principal components with P-values smaller than 0.01 was around $KTW = 10$. Using the Bayesian clustering programs STRUCTURE [5] and TESS [1, 2], we found that $K = 7$ components could better describe our simulated data. We extracted climatic data population samples using the WorldClim data set at 30 arcsecond (1km2) resolution (Hijmans, Cameron, Parra, Jones, and Jarvis (2005)). We summarized the climatic variables by using the first axis of a principal component analysis for temperature variables and for precipitation variables. The data set is in directory examples/human_example/. The genotype information are in panel11_Asia.lfmm. the SNPs information is in panel11.pedsnp. The environmental file is cov_panel11_Asia.env. There are 2 variables, one proxy for temperature and one for precipitation.

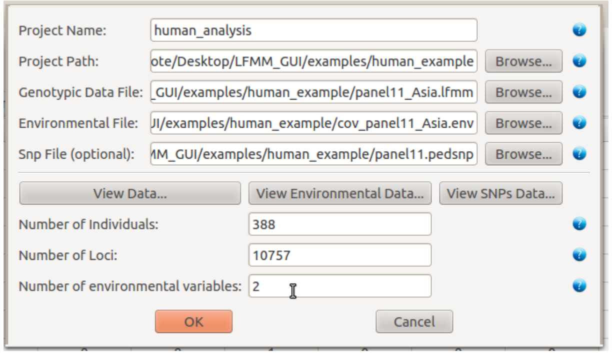
6.2 Launch LFMM

Let us start to analyze this data. The first step is to install LFMM. You can find all explanations in the section Installation of the documentation. Once it is installed, you can run LFMM GUI by clicking on the created executable. You can recognize the menu and the toolbar at the top, the project tree on the left, the text information box at the bottom, and the central zone.

6.3 New project

To analyze our dataset, we create a new project. You can click on the new project button in the toolbar:

A new box is displayed. It asks you some information to start your project. We fill it in as show in the following figure.



Project Name: human_analysis

Project Path: ote/Desktop/LFMM_GUI/examples/human_example

Genotypic Data File: _GUI/examples/human_example/panel11_Asia.lfmm

Environmental File: JI/examples/human_example/cov_panel11_Asia.env

Snp File (optional): MM_GUI/examples/human_example/panel11.pedsnp

Number of Individuals: 388

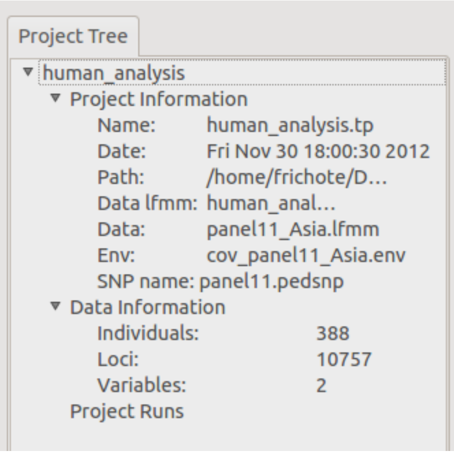
Number of Loci: 10757

Number of environmental variables: 2

The name of the data project is human_analysis. The project directory is human_example. Each time that you need to fill a directory or an existing file, you should fill the entire path to it. That is why we advise you to use the "Browse" button. This directory should contain the data set to analyze. The genotype data set is panel11_Asia.lfmm. The environmental file is cov_panel11_Asia.env. The snp file is panel11.pedsnp. You can use the "view" buttons to have a preview of these files. For the genotype file, it can be a bit slow. The number of individuals is 388. The number of loci is 10757. The number of

environmental variables is 2.
Tips: If you have forgotten what a field should contain, you can click on the interogation point buttons.

We are ready to create this new project. You can click on "OK". The software checks the consistency of your information and create the new project. If anything goes wrong, the detected inconsistency is displayed in a warning message. The genotype data information is displayed in the central zone and the project tree is filled with information you provided. you can click on the elements of the project tree to see them. In the project tree, you can click on any file to obtain its preview.

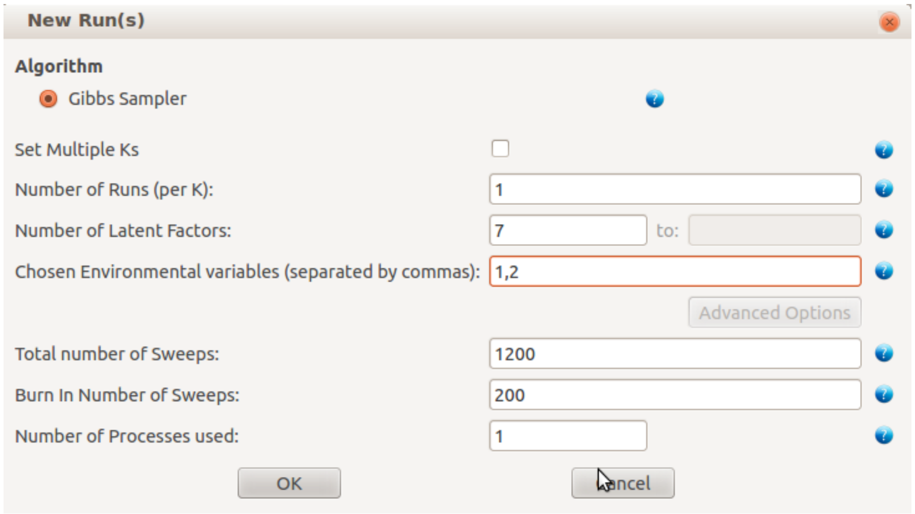


```
Project Tree
▾ human_analysis
    ▾ Project Information
        Name:       human_analysis.tp
        Date:       Fri Nov 30 18:00:30 2012
        Path:       /home/frichote/D...
        Data lfmm:  human_anal...
        Data:       panel11_Asia.lfmm
        Env:        cov_panel11_Asia.env
        SNP name: panel11.pedsnp
    ▾ Data Information
        Individuals:            388
        Loci:                   10757
        Variables:              2
      Project Runs
```

6.4 New Runs

Now that we have created a project, we can analyze our data and launch LFMM. To start it, you can click on the new run button in the toolbar:

A new box pops up. It asks you some information to start your project. We fill it as show in the following figure.



New Run(s)

Algorithm

● Gibbs Sampler

Set Multiple Ks □

Number of Runs (per K): 1

Number of Latent Factors: 7 to:

Chosen Environmental variables (separated by commas): 1,2

Total number of Sweeps: 1200

Burn In Number of Sweeps: 200

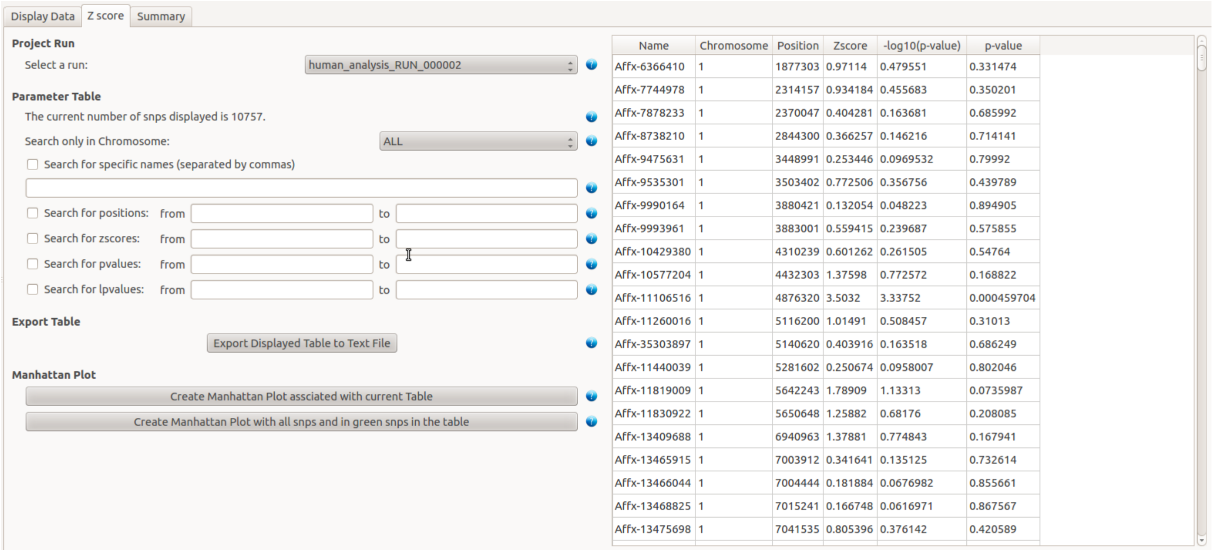
Number of Processes used: 1

We set one run per value of K, the number of latent factors. We choose to set $K = 7$ because it was the result returned by TESS and STRUCTURE. Several value should be tested. As we want to launch LFMM with the environmental variable 1, the temperature and then the environmental variable 2, the precipitation, we set the environmental variable to "1, 2". We set a total number of sweeps to 1200 and a burnin number of sweeps to 200. These values should be carefully chosen. It is dependent of the size of your data set for example. As we don't know the capacity of your computer, we set the number of processes to 1. We are ready to launch the runs. You can click on "OK". The software checks the consistency of your information and launch the first run. If anything goes wrong, a warning message indicates the detected inconsistency. Information about the stat of the current run is displayed in the information box at the bottom. On this data set, a run should be of a few minutes at most. On my laptop, it takes around 7 minutes per run with one process. So, you can get a cup of cofee or tea. You also notice that this new button appeared:

You can click on it if you want to abort the runs that you launched.

6.5 Analyze Run

Both runs that you launched should be over. When a run is over, the results are automatically displayed in the Zscore panel, as follows:


Display Data
Z score
Summary
Export Displayed Table to Text File
Create Manhattan Plot assciated with current Table
Create Manhattan Plot with all snps and in green snps in the table

Project Run

Select a run: human_analysis_RUN_000002

Parameter Table

The current number of snps displayed is 10757.

Search only in Chromosome: ALL

□ Search for specific names (separated by commas)

□ Search for positions: from to

□ Search for zscores: from to

□ Search for pvalues: from to

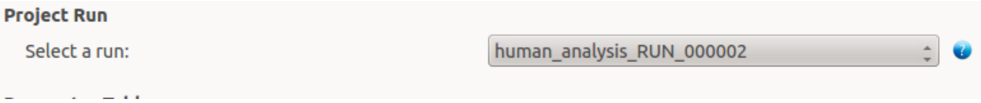
□ Search for lpvalues: from to

Export Table

Manhattan Plot

Name	Chromosome	Position	Zscore	-log10(p-value)	p-value
Affx-6366410	1	1877303	0.97114	0.479551	0.331474
Affx-7744978	1	2314157	0.934184	0.455683	0.350201
Affx-7878233	1	2370047	0.404281	0.163681	0.685992
Affx-8738210	1	2844300	0.366257	0.146216	0.714141
Affx-9475631	1	3448991	0.253446	0.0969532	0.79992
Affx-9535301	1	3503402	0.772506	0.356756	0.439789
Affx-9990164	1	3880421	0.132054	0.048223	0.894905
Affx-9993961	1	3883001	0.559415	0.239687	0.575855
Affx-10429380	1	4310239	0.601262	0.261505	0.54764
Affx-10577204	1	4432303	1.37598	0.772572	0.168822
Affx-11106516	1	4876320	3.5032	3.33752	0.000459704
Affx-11260016	1	5116200	1.01491	0.508457	0.31013
Affx-35303897	1	5140620	0.403916	0.163518	0.686249
Affx-11440039	1	5281602	0.250674	0.0958007	0.802046
Affx-11819009	1	5642243	1.78909	1.13313	0.0735987
Affx-11830922	1	5650648	1.25882	0.68176	0.208085
Affx-13409688	1	6940963	1.37881	0.774843	0.167941
Affx-13465915	1	7003912	0.341641	0.135125	0.732614
Affx-13466044	1	7004444	0.181884	0.0676982	0.855661
Affx-13468825	1	7015241	0.166748	0.0616971	0.867567
Affx-13475698	1	7041535	0.805396	0.376142	0.420589

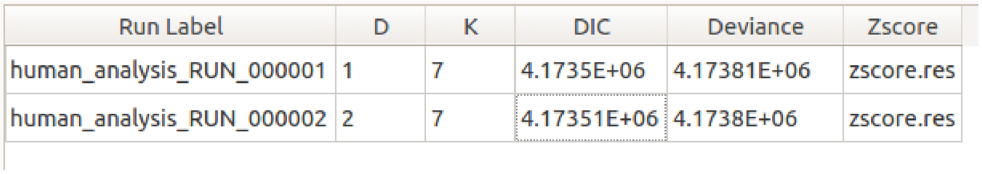
The results can be analyzed. As you can see, in the table at the right of the Zscore panel, the SNPs are displayed with their information and their results. For example, we analyze results for environmental variable 2. We select at the top, `human_analysis_run_2`.



Project Run

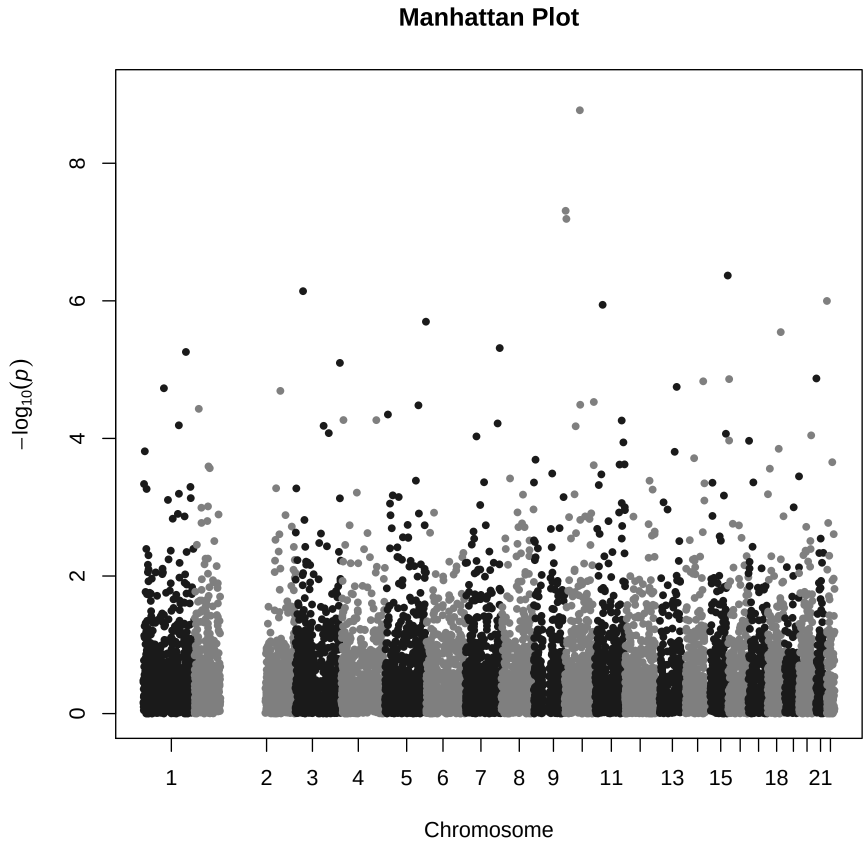
Select a run: human_analysis_RUN_000002

If you do not remember which run used variable 2, you can check it in the summary table panel.

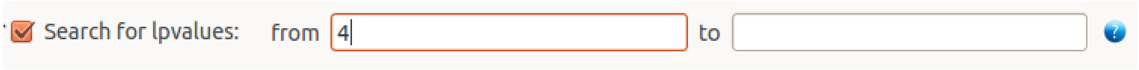


Run Label	D	K	DIC	Deviance	Zscore
human_analysis_RUN_000001	1	7	4.1735E+06	4.17381E+06	zscore.res
human_analysis_RUN_000002	2	7	4.17351E+06	4.1738E+06	zscore.res

We can create a global manhattan plot to have a visual idea of the results. Just click on the button, "create Manhattan plot associated with the current table". A manhattan plot, similar to the one below is displayed:

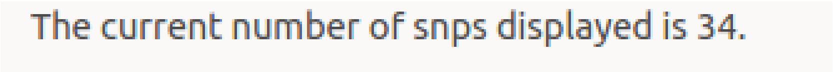


We can observe that some SNPs are significantly associated with this variable 2 because they have a high $-log_{10}(P-value)$. More precisely, for example, we would like to have a list of SNPs with $-log_{10}(P-value)$ higher than 4. To do this, we just have to fill the range for the corresponding box and check the checkbox as follows:



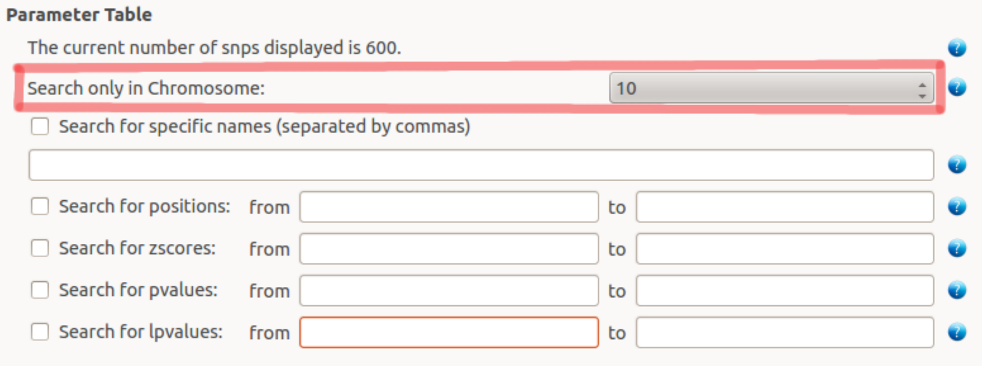
☑ Search for lpvalues: from 4 to

The number of SNPs is displayed as follows:



The current number of snps displayed is 34.

Another example could be that we would like to analyze SNPs in chromosome 10. To do this, we just have to select chr10 in the list as follows:



Parameter Table

The current number of snps displayed is 600.

Search only in Chromosome: 10

□ Search for specific names (separated by commas)

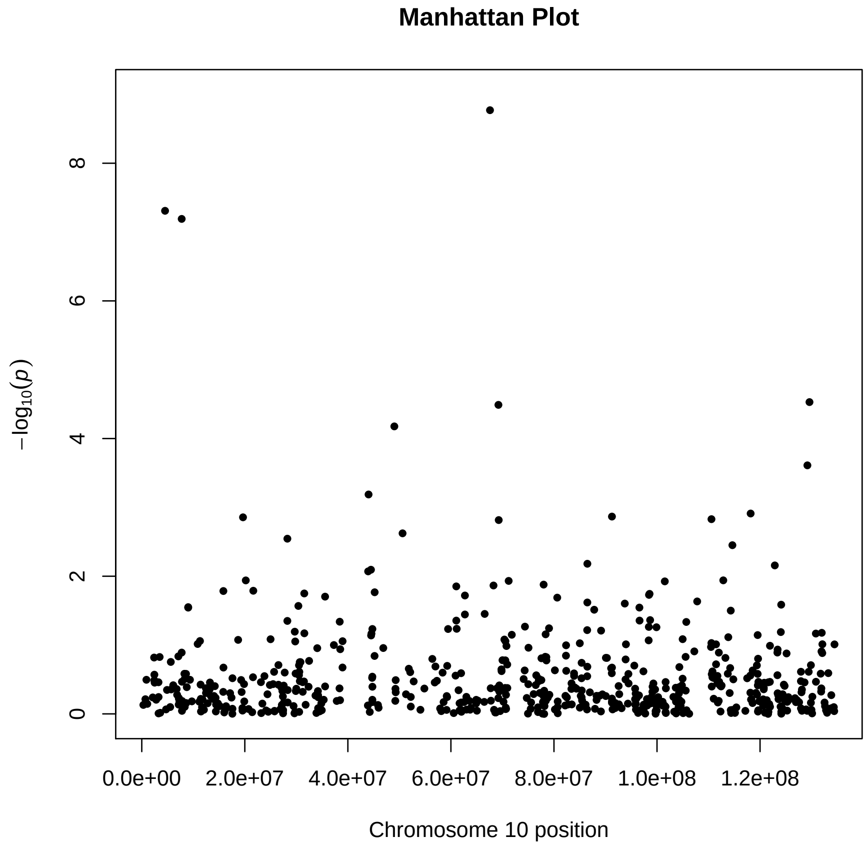
□ Search for positions: from to

□ Search for zscores: from to

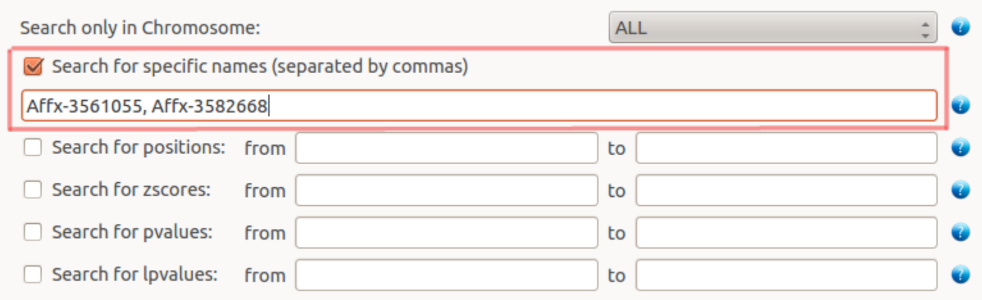
□ Search for pvalues: from to

□ Search for lpvalues: from to

We can also display the manhattan plot for this specific chromosome by clicking on the button "create Manhattan plot associated with the current table" after we selected chromosome 10. A manhattan plot, similar to the one below is displayed:



Finally, for example, we would like to know if SNPs Affx-3561055 and Affx-3582668 are significantly correlated with variable 2. We can look for these SNPs by filling these names as follows:



Search only in Chromosome: ALL

☑ Search for specific names (separated by commas)

Affx-3561055, Affx-3582668

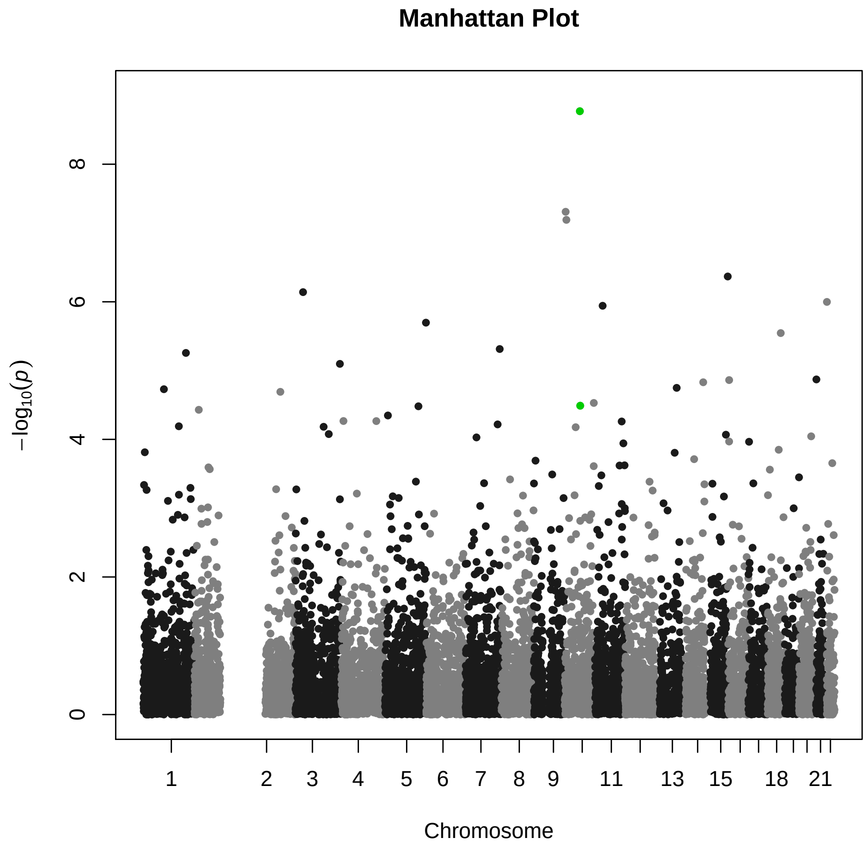
□ Search for positions: from ______ to ______

□ Search for zscores: from ______ to ______

□ Search for pvalues: from ______ to ______

□ Search for lpvalues: from ______ to ______

We can also display the manhattan plot with all SNPs and these two SNPs underlined in green by clicking on the button "create Manhattan plot with all SNPs and in green SNPs in the table table" after we looked for these SNPs in the zscore table. A manhattan plot, similar to the one below is displayed:



References

[1] Chibiao Chen, Eric Durand, Florence Forbes, and Olivier François. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, 7(5):747–756, 2007.

[2] E Durand, F Jay, O E Gaggiotti, and O François. Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, 26(9):1963–1973, 2009.

[3] N Patterson, A L Price, and D Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2:20, 2006.

[4] Nick J. Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics, doi:10.1534/genetics.112.145037*, 2012.

[5] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.