LEA: An R package for Landscape and Ecological Association Studies

Olivier Francois
Ecole GENOMENV – AgroParisTech,
Paris, 2016

Outline

- Installing LEA – Formatting the data for LEA
- Basic principles
  - Analysis of population structure
  - Genome scans for association with environmental variables
- Main outputs
- Interpreting results
- Short tutorial

LEA

LEA: an R package for Landscape and Ecological Association Studies

Bioconductor version: Release (3.1)

LEA is an R package dedicated to landscape genomics and ecological association tests. LEA can run analyses of population structure and genome scans for local adaptation. It includes statistical methods for estimating ancestry coefficients from large genotypic matrices and evaluating the number of ancestral populations (snmf, pca); and identifying genetic polymorphisms that exhibit high correlation with some environmental gradient or with the variables used as proxies for ecological pressures (lfmm), and controlling the false discovery rate. LEA is mainly based on optimized C programs that can scale with the dimension of very large data sets.

Author: Eric Frichot <eric.frichot at gmail.com>, Olivier Francois <olivier.francois at imag.fr>

Installing LEA

- **Requires > R.3.1**

Installation

To install this package, start R and enter:

```
## try http if https is not available
source("https://bioconductor.org/biocLite.R")
biocLite("LEA")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("LEA")
```

Installing LEA

- **If the biocLite fails, an option is to**
  - **Download the source files from Bioconductor**
  - **Install LEA manually using the "install.packages" function by entering**

```
install.packages("LEA_1.2.0_tar.gz", repos = NULL, type ="source")
```

- **Note that the main functions of LEA have their own software that can run without R**

Data format

- LEA handles matrices of genotypes and vectors of environmental variables
- The data file extension for genotypic matrices is ".lfmm"
- The data file extension for environmental variables is ".env"

Genotypes

- The LEA programs can deal with individual genotypes or population genotypic frequencies

- Example of individual SNP data for 3 diploid organisms

	SNP 1	SNP 2	…	
Ind 1	0	1	1	1
Ind 2	1	2	0	2
Ind 3	0	1	1	0

Genotypes

- **Example of population genotypic frequency data for 2 populations**

	SNP1	SNP2	…	
Pop 1	1.23	.76	.54	.12
Pop 2	.12	0.98	1.10	1.98

Data conversion

- **LEA provides functions that convert data from standard formats**
  - **geno:** geno2lfmm / lfmm2geno
  - **STRUCTURE:** struct2geno
  - **ped:** ped2geno
  - **vcf:** vcf2geno
  - **ancestrymap:** ancestrymap2geno

Environmental data

- **Any type of continuous or discrete data corresponding to individual or population samples.**
- **Example: temperature and precipitation data**
- **Stored in vectors or matrices (.env files)**
- **It's better that the environmental data are uncorrelated.**

Example of climatic data

- Temperatures extracted from « Worldclim »

```
library(raster)
Climate = getData('worldclim', var = 'tmax', res = 2.5)
temp = extract(Climate, coordinates)
write.table('temp.env', temp, row.names=F, quote=F)
```

LEA's main functions

- **snmf{LEA}**

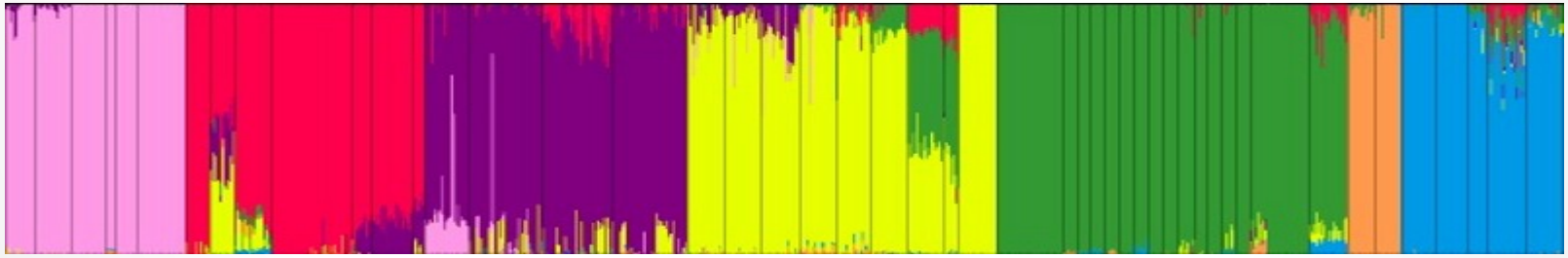
  Estimates individual ancestry coefficients and ancestral allele frequencies from the genotypic data

- **lfmm{LEA}**

  Fits latent factor mixed models and tests association with environmental variables

The snmf() function

- **Provides output similar to STRUCTURE (Pritchard *et al*. 2000), but faster**
- **Estimates the number of cluster, *K*, using a cross-entropy criterion**



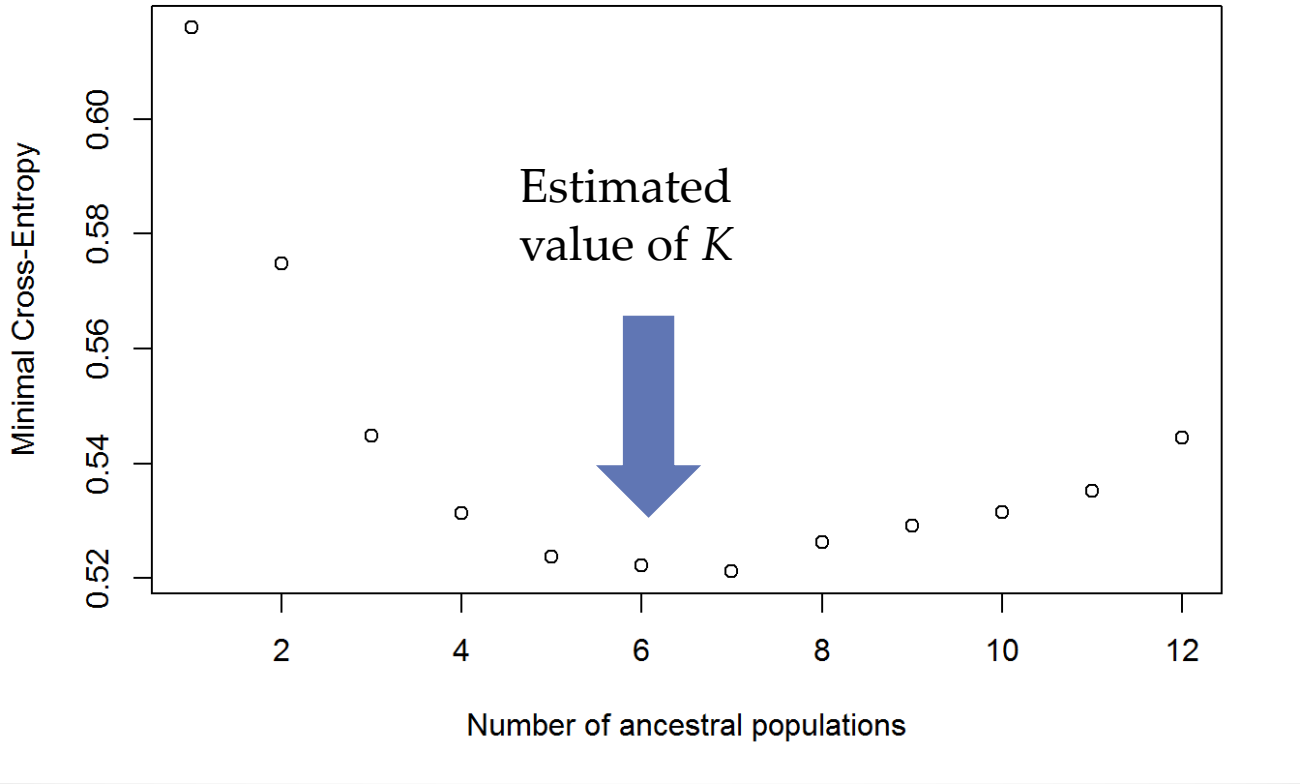
- **See:** Running Structure-like Population Genetic Analyses with R

Usage

Genotype = lfmm2geno('genotype.lfmm')

object = snmf(Genotype, K=1:12, entropy = T)

plot(object)



Examples

- **Download the data**

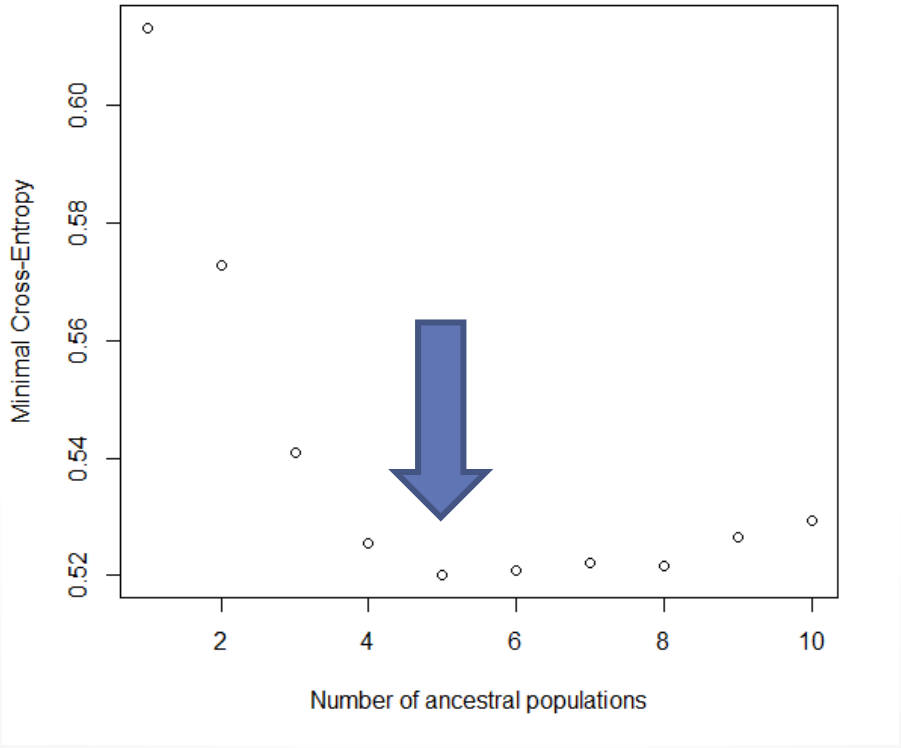
```
url = "http://membres-timc.imag.fr/Olivier.Francois/datasets_genomenv.zip"
download.file(url = url, destfile = "./datasets.zip")
```

- **Uncompress the zip file in your working directory**

Example

genotype = lfmm2geno('./datasets_genomenv/example/example.genotype_2.lfmm')

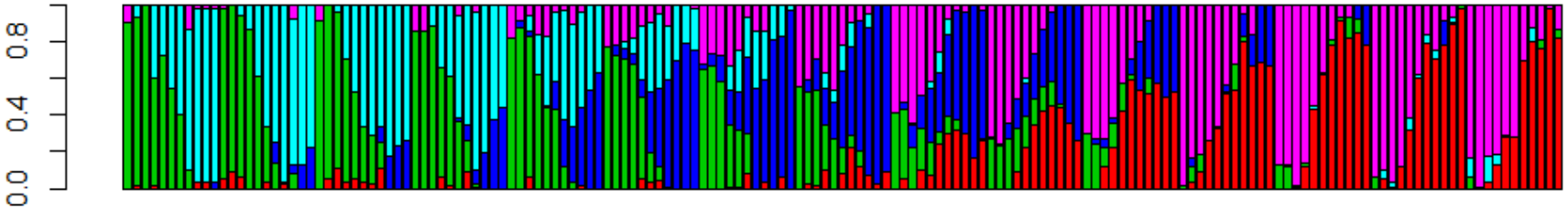
object = snmf(genotype, K=1:6, ploidy = 2, entropy = T,project = 'new')



Population structure

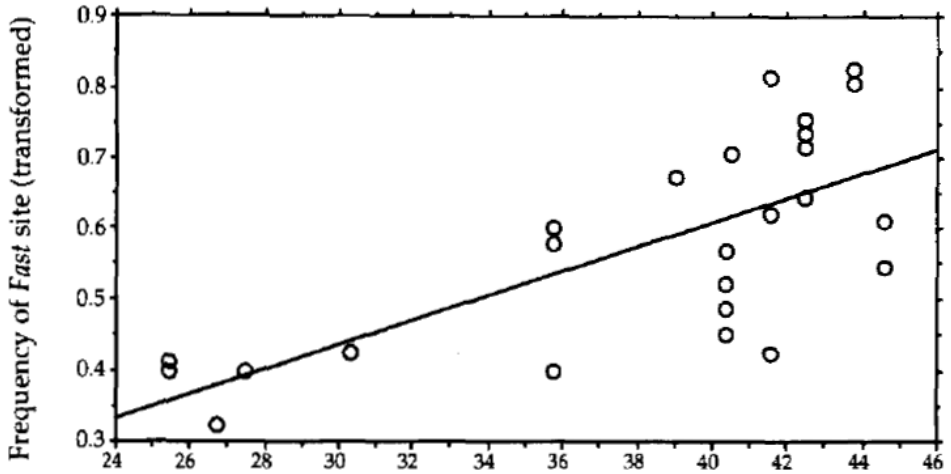
- **Individual ancestry coefficients for *K*=5 ancestral populations**

barplot(t(**Q**(object, K = 5)), col = 2:6)



Natural selection and clines

- **Selection along environmental gradients often results in the observation of clines in spatially distributed populations.**



frequency of *Adh-F* (square-root, arcsine transformed) on the latitude of each sample

Association methods

- For G, a matrix of genotypes and X a set of environmental variables, EA tests are based on regression models

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \epsilon_{i\ell},$$

where $G_{i\ell}$ is the genotype at locus ℓ, and X_i is the enviromnental variable for individual i.

Association methods

The significance of an environmental effect is measured by a *z*-score statistic computed at each locus.

Issue: **Inflation of the test statistic** due to population structure and other **confounding factors**.

LFMM

- Latent factor models

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \mathbf{U_i^T V}_\ell + \epsilon_{i\ell},$$

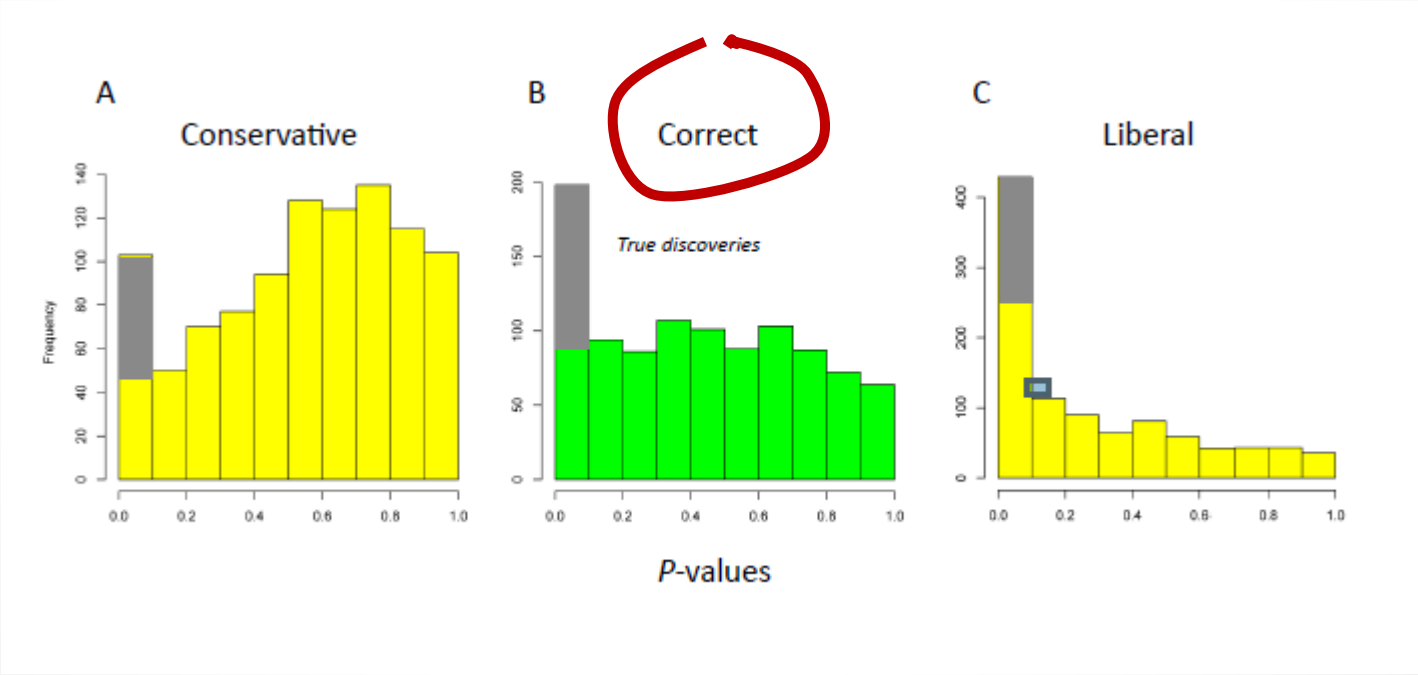
  where β_ℓ is a vector of regression coefficients, U_i are latent factors, and V_ℓ contains their corresponding loadings.
- The model assumes that there are K latent factors (Frichot et al. 2013).

Interpretation

- The latent factors, U_j, account for confounding effects due to correlation among individuals.
- The loadings, V_ℓ, account for confounding effects due to interactions of genes and linkage disequilibrium.
- z-scores can be computed using the `R` function `lfmm` (package LEA, Frichot et al. 2015)

Objective of LFMM

- **Provide correct null-hypothesis testing procedure and p-values for ecological association tests**



Usage of lfmm()

- **We use individual genotypes**

```
Genotype =
"./datasets_genomenv/example/example.genotype_2.lfmm"
Gradient  =
"./datasets_genomenv/example/ecological.gradient.env"
```

- **and run the lfmm model with *K*=5 latent factors (3 runs!)**

```
project.lfmm = lfmm(Genotype, Gradient, K = 5,
   iterations = 3000, burnin = 2000,  rep = 3)
```

Getting the results

- **Get the scores from each run**

  ```
  zs = z.scores(project.lfmm, K = 5)
  ```

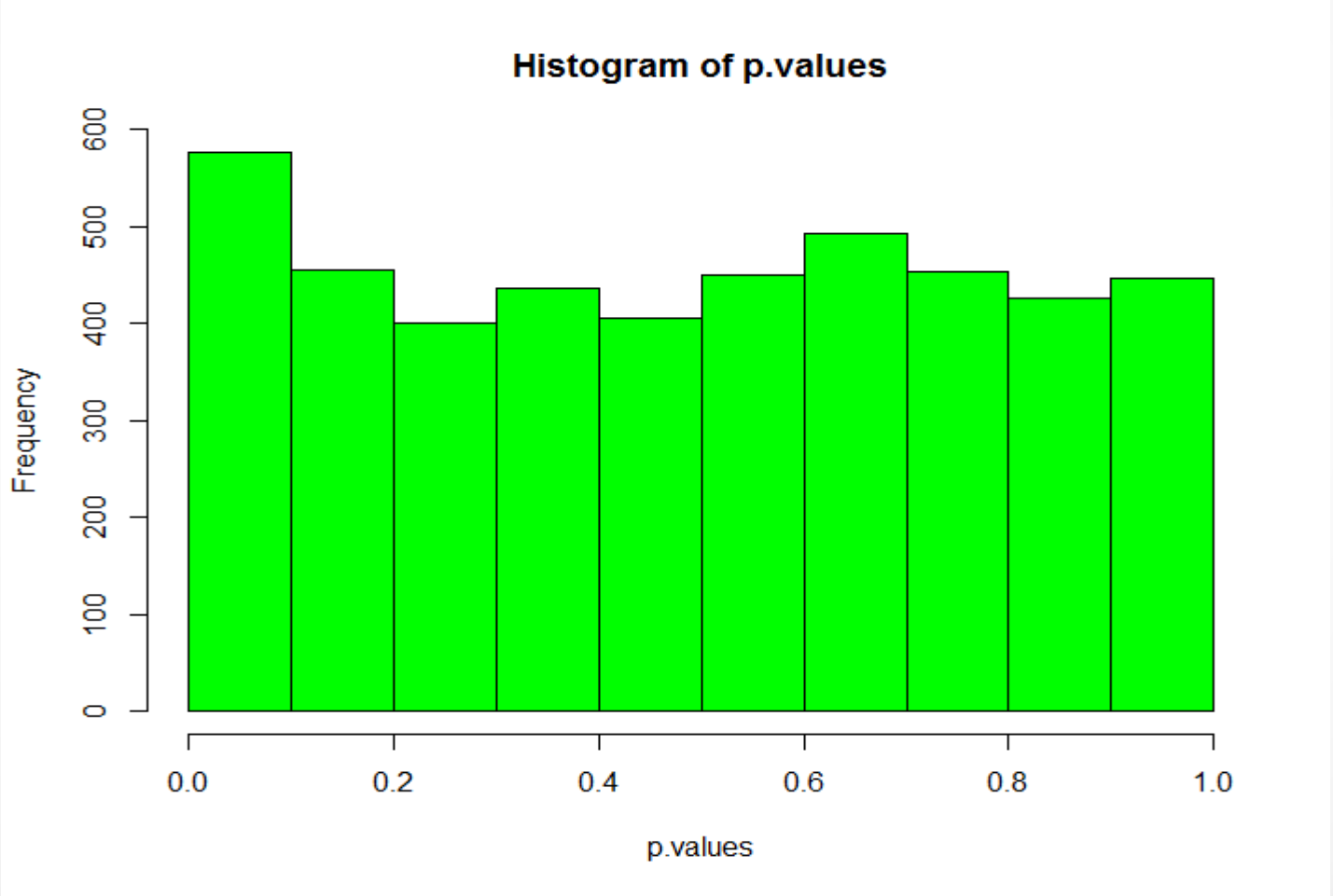
- **Combine the 3 run results using the Stouffer method (median value)**

  ```
  zs.stouffer = apply(zs, MARGIN = 1, median)
  gc = median(zs.stouffer^2)/.456
  ```

- **Compute p-values**

  ```
  p.values = pchisq(zs.stouffer^2/gc, df = 1, lower = FALSE)
  ```

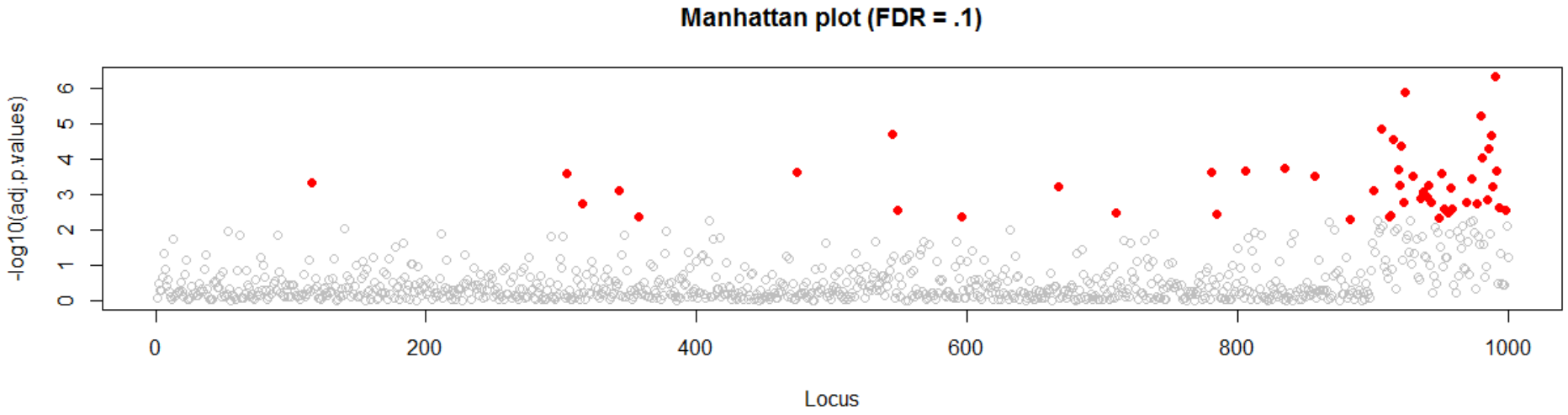
Checking the model



Providing a candidate list

- **Use the BH FDR control method**

```
alpha = .10
L = length(p.values)
w = which(sort(p.values) < alpha * (1:L) / L)
candidates = order(p.values)[w]
```



Ready for the practicals!

- **Artificial genotypes (example data)**

  Find which K and run lengths provide the best candidate list (highest power, lowest FDR)

- **Real data (*A. thaliana*, Chr1). Find loci with association with climatic data (PC1 of climatic variables)**

- **Use the Gbrowser of TAIR to check your top 10 list.**

https://gbrowse.arabidopsis.org/cgi-bin/gb2/gbrowse/arabidopsis/