tess3r : An R Package for Population Genetics Study

Kevin Caye and Olivier François

2015-10-14

Contents

This R package implements the **TESS3** [REFERENCE à l'article de TESS3] program and tools useful to plot program outputs. The program has functionalities similar to the previous versions of **TESS** (Chen et al. 2007; Durand et al. 2009), has run-times several order faster than those of common Bayesian clustering programs. In addition, the program can be used to perform genome scans for selection based on ancestral allele frequency differentiation statistic, and to separate non-adaptive and adaptive genetic variation.

This documentation aims to present main functions of this package through two exemples. In the first exemple, we present how to use **TESS3** to find population structure and plot this structure on a map. The second exemple shows a way to use the ancestral allele frequency differentiation statistic to perform genome scan for selection. The main features of **TESS3** are illustrated using an example data set, simulated from European lines of the plant species *Arabidopsis thaliana*. More details on **tess3r** package can be found in the package documentation.

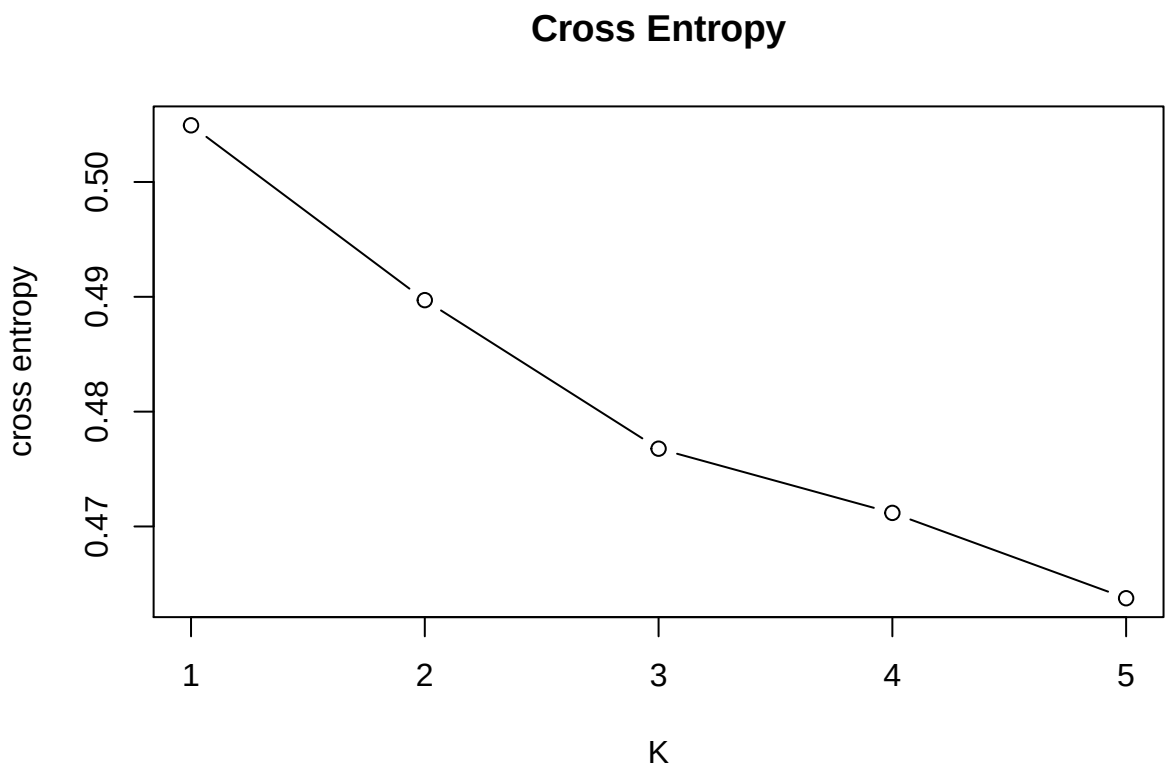
0.1 Visualization of Ancestry Coefficients Using TESS3

This example describes how to use **tess3r** R package to run **TESS3** algorithm for several values of the number of ancestral populations.

```
library(tess3r)

# Retrieve data file name
genotype.file <- system.file("extdata/Athaliana","Athaliana.geno",package = "tess3r")
coord.file <- system.file("extdata/Athaliana","Athaliana.coord",package = "tess3r")
# Read coordinate file
coord <- read.coord(coord.file)
n <- nrow(coord)

project <- tess3(input.file = genotype.file,
                 input.coord = coord.file,
                 K = 1:5,
                 ploidy = 1,
                 repetitions = 1,
                 entropy = TRUE,
                 percentage = 0.2,
                 project = "new")
```

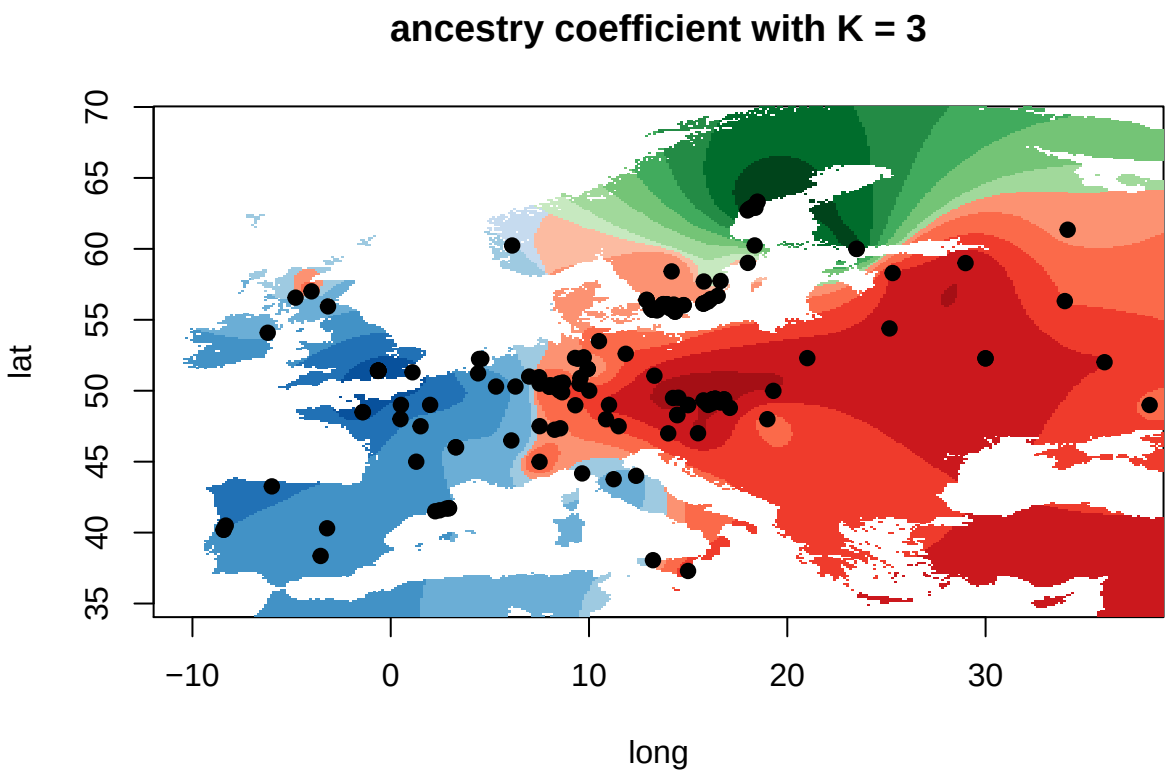
Then we plot a graph of the cross-entropy criterion as function of K.

```
##########################################
# Chose of K with cross-entropy criterion #
##########################################

cross <- sapply(1:5, function(k) { cross.entropy(project, run = 1, K = k) })
plot( 1:5, cross, main  = "Cross Entropy",type="b", xlab = "K", ylab = "cross entropy" )
```



Finally, geographic representations of ancestry coefficient maps is displayed using raster files.

```
################################
# Plot result on map for K = 3 #
################################

asciiFile=system.file("extdata/","lowResEurope.asc",package = "tess3r")
grid=createGridFromAsciiRaster(asciiFile)
# To display only altitudes above 0:
constraints=getConstraintsFromAsciiRaster(asciiFile,cell_value_min=0)

maps(matrix = Q( project, K = 3, run = 1 ),
     coord = coord,
     grid=grid,constraints=constraints,method="max",main="ancestry coefficient with K = 3")
```



0.2 Genome Scan for Selection Using TESS3

We show here how to use **tess3r** R package to perform a genome scan for selection based on the computation of ancestral allele frequency differentiation statistics. Here **TESS3** is run with $K = 3$ ancestral populations.

```
library(tess3r)

# Retrieve data file name
genotype.file <- system.file("extdata/Athaliana","Athaliana.geno",package = "tess3r")
coord.file <- system.file("extdata/Athaliana","Athaliana.coord",package = "tess3r")
# Read coordinate file
coord <- read.coord(coord.file)
n <- nrow(coord)

project = tess3(input.file = genotype.file,
                input.coord = coord.file,
                K = 3,
                ploidy = 1,
                repetitions = 5,
                project="new")
```

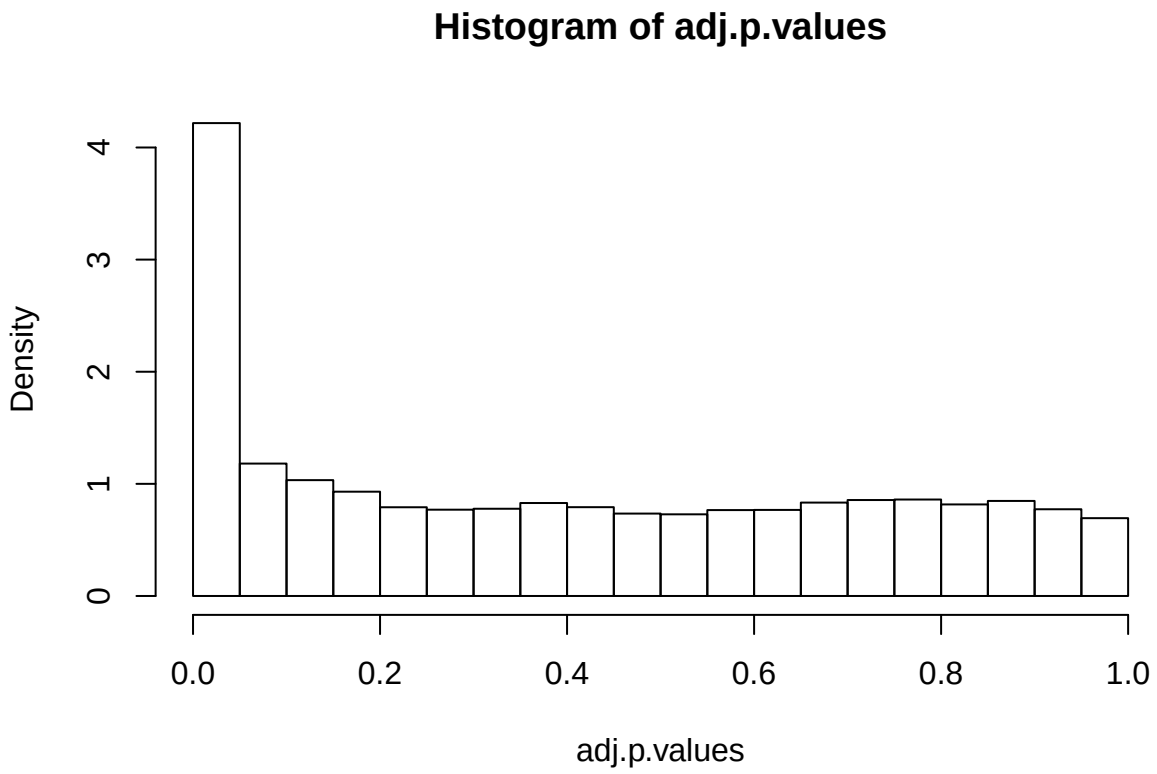
The F_{ST} statistics were transformed into squared t-scores and p-values using a Fisher distribution.

```
#### Fst with TESS3
Fst = FST( project, 3, 1 )
Fst[Fst < 0.0] = 0.0
```

```
#### Convert Fst into t score
squared.t.scores = Fst*(n-3) / 2 / (1-Fst)
```

Inflation of the test statistic caused by population structure was corrected by recalibrating the p-values to follow a uniform distribution under the null hypothesis.

```
#### recalibrated p-values
gif = 12.5
adj.p.values = pf( squared.t.scores/gif , df1 = 2, df2 = n-3, lower = FALSE )

hist(adj.p.values,prob=TRUE)
```



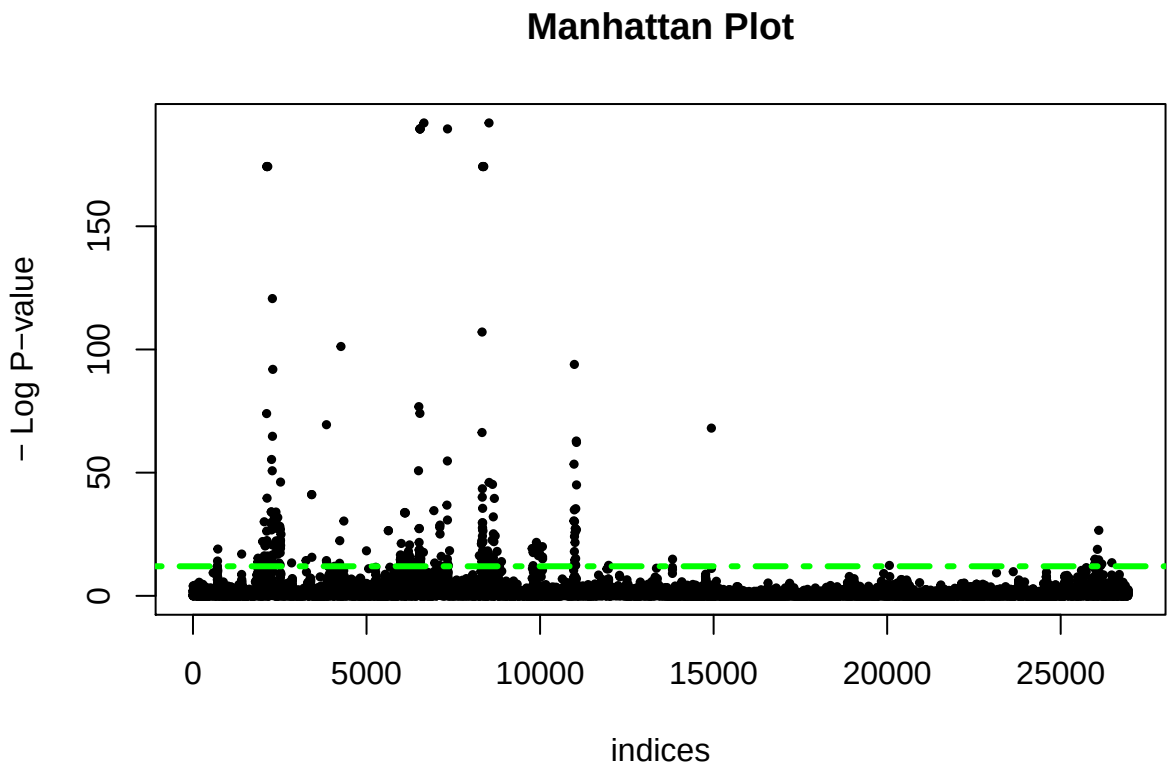
The false discovery rate was controlled using the Benjamini-Hochberg procedure.

```
#### Benjamini Hochberg procedure
alpha = 1e-10
L = length(adj.p.values)
# return a list of candidates with an expected FDR of alpha.
w = which(sort(adj.p.values) < alpha * (1:L) / L)
candidates = order(adj.p.values)[w]
limite = max(adj.p.values[candidates])
```

Finally, candidate loci are shown above the green dashed line threshold in the Manhattan Plot.

```
#### Manhattan plot
plot( 1:length(adj.p.values),-log10(adj.p.values) ,
      main = "Manhattan Plot" ,
      xlab = "indices",
```

```
      ylab="- Log P-value",
      pch=19, cex = .5)
#add limite
abline( -log10(limite), 0, col = "green", lty = 6, lwd = 3 )
```



References

Chen, Chibiao, Eric Durand, Florence Forbes, and Olivier François. 2007. "Bayesian Clustering Algorithms Ascertaining Spatial Population Structure: A New Computer Program and a Comparison Study." *Molecular Ecology Notes* 7 (5). Wiley Online Library: 747–56.

Durand, Eric, Flora Jay, Oscar E Gaggiotti, and Olivier François. 2009. "Spatial Inference of Admixture Proportions and Secondary Contact Zones." *Molecular Biology and Evolution* 26 (9). SMBE: 1963–73.