TESS3 reference manual
Command-line program

Kevin Caye (kevin.caye@imag.fr)

Olivier François (olivier.francois@imag.fr)

November 9, 2015

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

Summary

Geography is an important determinant of genetic variation in natural populations, and its effects are commonly investigated by analyzing population genetic structure using spatial ancestry estimation programs. A common issue is that classical spatial ancestry estimation programs do not scale with the dimension of the data sets generated from modern sequencing technologies, and more efficient algorithms are needed to analyze genome-wide patterns of population genetic variation in their geographic context.

The computer program TESS3 [1] implements admixture models. The program TESS3 has functionalities similar to the previous versions of TESS [2, 3], and has run-times several order faster than those of common Bayesian clustering programs. In addition, the program TESS3 can be used to perform genome scans for selection based on ancestral allele frequency differentiation statistics, and to separate non-adaptive genetic variation from adaptive genetic variation.

This documentation aims to help users to run the TESS3 command-line program on Linux, Mac and Windows operating systems.

1   Program installation

The installation of the TESS3 command-line software requires that CMake, a the cross-platform open-source build system, is installed on the computer OS (http://www.cmake.org/). The TESS3 program source code can be downloaded from the following Github address: https://github.com/cayek/TESS3. After downloading the files as a zipped archive, unzipping the archive will create a TESS3-master directory on the system. The installation of TESS3 can be performed by typing the following instructions in a terminal session from the TESS3-master directory.

```
mkdir build/
cd build/
cmake -DCMAKE_BUILD_TYPE=release ../
make TESS3
```

The executable program file `TESS3` will be found in the `build` directory. To run the program, copy and paste the executable program file to your working directory.

2   Data format

2.1   Input files

`TESS3` requires two input files, the first one recording individual genotype data and the second one containing the geographic coordinates of each sampled individual. For genomes of arbitrary ploidy, the standard data type for `TESS3` is the **single nucleotide polymorphism** (SNP) type. The genotype matrix must be formatted in the **geno** format and the coordinate file must be formatted in the **coord** format.

Users who want to process allelic data, such as microsatellite markers or AFLPs, and have their data in the `TESS 2.3` format can also use `TESS3`. They need to convert their data in the geno+coord data using the conversion tools in the R package `LEA` (`http://membres-timc.imag.fr/Olivier.Francois/LEA/tutorial.htm`).

- **geno** (example.geno)

  The **geno** format has one row for each SNP. Each row contains 1 character per individual. For diploid genomes, 0 means zero copies of the reference allele, 1 means one copy of the reference allele, 2 means two copies of the reference allele, and 9 codes for some missing data. Here is an example of a geno file for $n = 3$ individuals and $L = 4$ loci.

  ```
  112
  010
  091
  121
  ```

- **coord** (example.coord)

  The **coord** format has one row for each individual. Each row contains the `longitude` and `latitude` coordinates of each individual.

  ```
  2.5154 5.4390
  -8.4293 4.0197
  1.3536 5.5852
  ```

Users having their genotype data in the **ped**, **ancestrymap**, **vcf** or **lfmm** format can use the R package `LEA` to convert them in the **geno** format [4].

2.2 Output files

`TESS3` produces three main **output files**:

- The file with the extension the **.K.Q** contains the Q-matrix of individual ancestry coefficients. The Q-matrix has n rows (the number of individuals) and K columns (the number of ancestral populations).

- The file with the extension **K.G** contains the G-matrix of ancestral genotype frequencies. The G-matrix has $N_g \times L$ lines (the number of genotypes times the number of SNPs) and K columns (the number of ancestral populations). For diploid organisms, the first row represents the frequencies of genotype 0 in each ancestral population, the second row represents the frequencies of genotype 1, and the third row the ancestral frequencies of 2.

- The file with the extension **.Fst** contains the values of the ancestral allele frequency differentiation statistic F_{ST} computed at each locus.

There are additional **output files** that are useful to get information about the nearest-neighbor graph, the least-squares error and the value of the cross-entropy criterion computed by the program.

- The file with the extension **.W** contains the weight matrix for the nearest-neighbor graph.

- The file with the extension the **.sum** contains the least-squares error and the cross-entropy criterion computed with all the data and with a percentage of masked data.

3 Run the command-line program

3.1 Command-line

The `TESS3` program can be executed from a command line. The basic format is as follows

```
./TESS3 -x genotype_file.geno  -r coordinates_file.coord -K number_of_ancestral_populat
```

The three options `-x` `-r` `-K` are mandatory. The ordering of options in the command line is unimportant. Here is a description of these options:

- `-x genotype_file.geno` is the path to the genotype matrix file (in the .geno format),

- `-r coordinates_file.coord` is the path to the coordinate file (in the .coord format),

- `-K number_of_ancestral_pop` is the number of ancestral populations.

Some additional options are available:

- `-m ploidy` 1 if haploid, 2 if diploid (default: 2).

- `-a alpha` is the value of the normalized regularization parameter (default: 0.001). This parameter controls the geographic regularity of the ancestry estimates.

- `-W edge_weight_input` is the path to a file that contains a pre-specified edge weight matrix for the nearest-neighbor graph (separator = space character). The program uses this matrix when the option is set, otherwise it uses the default values. The default nearest-neighbor graph is described in [1].

- `-q output_Q` is a path to the output file containing the Q-matrix of ancestry coefficients. The default name of the output file is the same name as the input file with the extension `.K.Q`.

- `-g output_G` is a path to the output file containing the ancestral genotype frequencies. The default name of the output file is the same name as the input genotype file with the extension `.K.G`.

- `-f output_FST` is a path to the output file containing the values of the ancestral allele frequency differentiation statistic for each loci. The default name of the output file is the same name as the input genotype file with the extension `.K.Fst`.

- `-y output_sum` is a path to the output file containing the value of the least-squares criterion, the cross-entropy criteria on all data and on masked data only. By default, the name of the output file is the same name as the input genotype file with the extension `.K.sum`.

- `-c perc` is the percentage of masked genotypes. If this option is set, the cross-entropy criterion is computed (see [5] for more details on the cross-entropy criterion). The default percentage is 0.05.

- `-e tolerance` is the tolerance error in the optimization algorithm (by default: 0.0000001).

- `-i iteration_number` is the maximum number of iterations of the algorithm (default: 200).

- `-I nb_SNPs` starts the algorithm with a run using a subset of nb_SNPs random SNPs. This option speeds up the estimation algorithm for very large data sets.

- `-Q input_Q` is the path to an initial file for the Q matrix containing individual ancestry coefficients. If both `-I` and `-Q` are set, `-Q` is chosen.

- `-s seed` is a seed to initialize the random number generator.

- `-p p` is the number of CPUs to use when the algorithm is run on a multiprocessor system. Be aware that the number of processes has to be lower or equal to the number of CPU units available on your computer (default: 1).

A summary of options can be obtained by typing the following command

```
./TESS3 -h
```

3.2 Example

To run a very simple example of the use of TESS3, consider data simulated from real data of the model plant species *A. thaliana*. The genotype matrix contains 170 genotypes from European lines of the plant (26943 SNPs). Since *A. thaliana* is a selfing species, the heterozygote frequency was equal to zero, and the diploid genotypes were recoded as haploid. Copy the program executable file TESS3 in the ./data/simulated/Athaliana directory. Then, move to the directory ./data/simulated/Athaliana and type

```
./TESS3 -x Athaliana.geno  -r Athaliana.coord -K 3 -m 1
```

Then, open a R session and type

```
par( mfrow = c(2,1) )
Qmatrix = t( read.table("Athaliana.3.Q") ) # get the Q_matrix
barplot(Qmatrix, col = 2:4)                 # display a barplot for the Q-matrix
Fst = read.table("Athaliana.3.Fst")[,1]    # get the Fst
plot(Fst, main = "Manhattan Plot", cex = .4, pch = 19, col = "blue")
```

These commands visualize the ancestry coefficient matrix computed by TESS3, and display a Manhattan plot of F_{ST} values along the fifth chromosome of the plant species.

For visualizing the Q-matrix using a geographic map see the explanations given in the TESS webpage: http://membres-timc.imag.fr/Olivier.Francois/tess.html (How to display the Q matrix from a unique geographic map). The Q-matrix obtained with $K = 3$ is displayed in Figure 1.

For generating a Manhattan plot for significance values, use the following R commands.

```
n = dim(Qmatrix )[2]
K = 3
f.scores = Fst*(n-K)/(1-Fst)/(K-1)
lambda = median(f.scores)/qf(0.5, K-1, n-K)
p.values = pf(f.scores/lambda, df1 = K-1, df2 = n-K, lower = F)
plot(-log10(p.values),
     xlab = "SNP number",
     main = "Manhattan plot",
     pch = 19,
     cex = .3,
     col = "blue")
```

The result of the genome scan for selection is displayed in Figure 2. Lists of outlier loci can be proposed using Storey's method as implemented in the R package "qvalue".
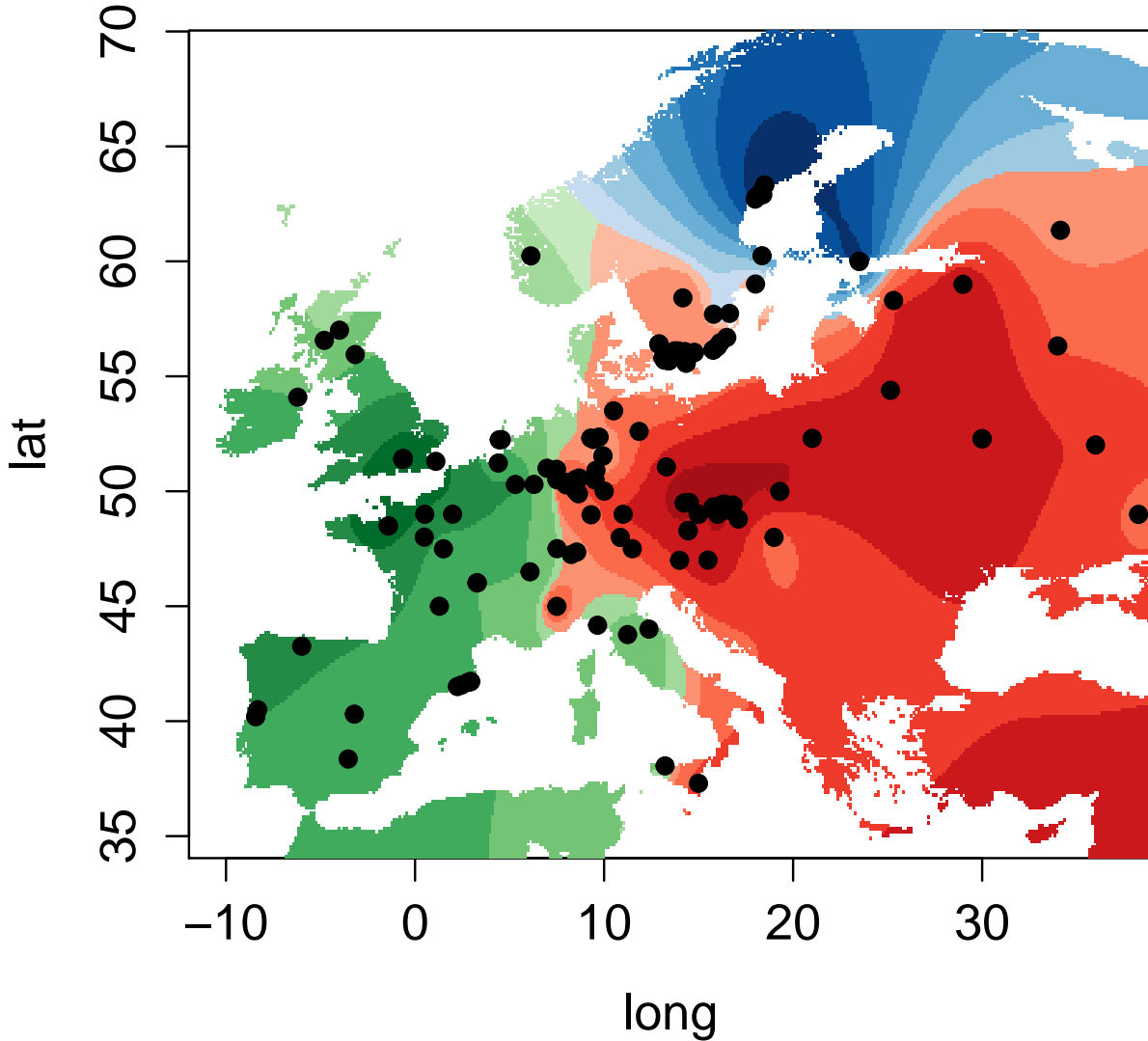
Ancestry Coefficient with K = 3

Figure 1: *Ancestry map for European lines of A. thaliana (K=3 ancestral populations).*

4   Contact

If you need assistance, do not hesitate to send us an email (kevin.caye@imag.fr or olivier.francois@imag.fr).

References

[1] Kevin Caye, Timo M. Deist, Helena Martins, Olivier Michel, and Olivier Francois. Tess3: Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, pages n/a–n/a, 2015.

[2] Chibiao Chen, Eric Durand, Florence Forbes, and Olivier François. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a com-
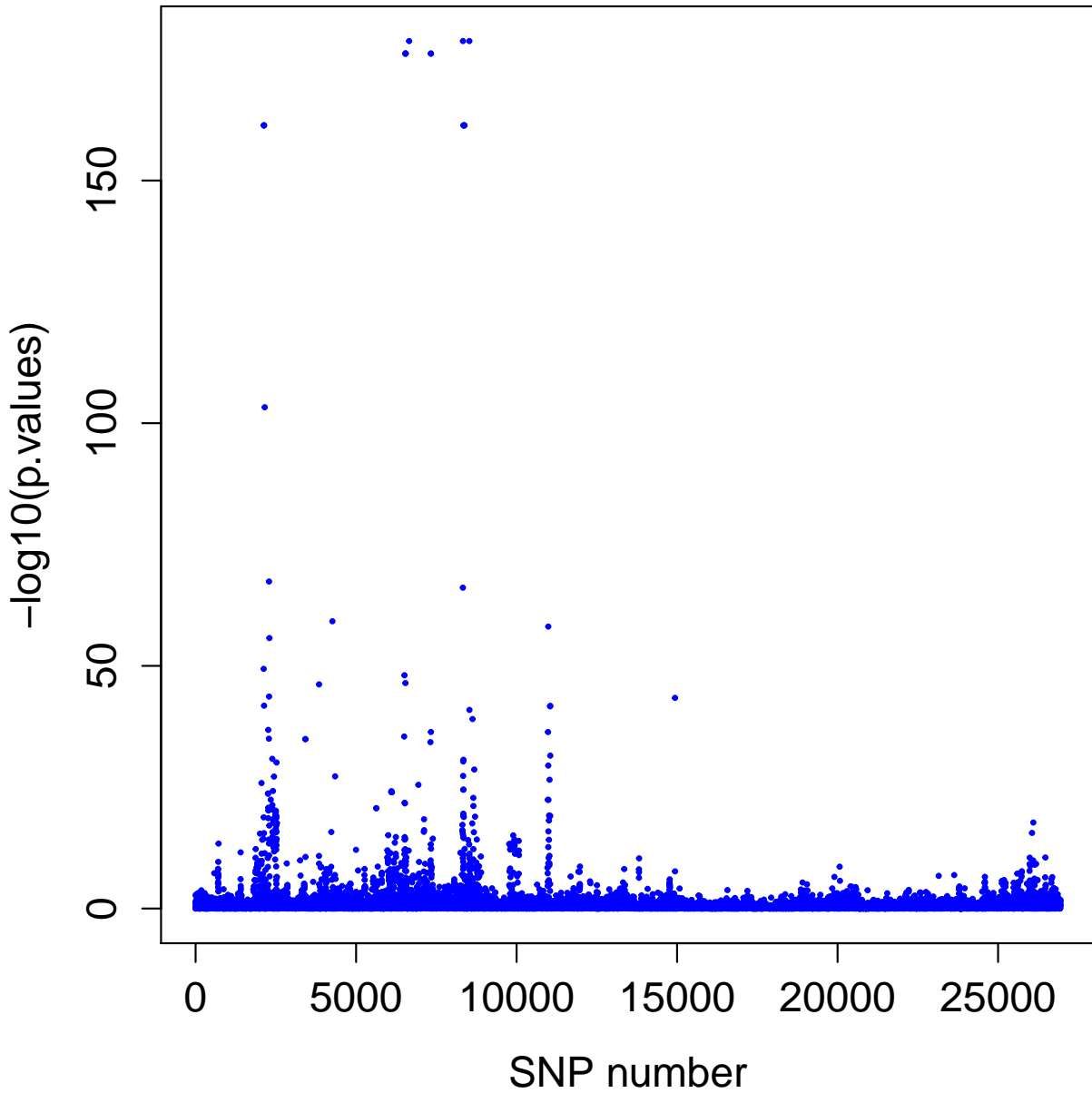
Manhattan plot

Figure 2: *Manhattan plot for European lines of A. thaliana (K=3 ancestral populations).*

parison study. *Molecular Ecology Notes*, 7(5):747–756, 2007.

[3] Eric Durand, Flora Jay, Oscar E Gaggiotti, and Olivier François. Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, 26(9):1963–1973, 2009.

[4] Eric Frichot and Olivier François. LEA: an R package for Landscape and Ecological Association studies. *Methods in Ecology and Evolution, in press*, 2015.

[5] Eric Frichot, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François.

Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4):973–983, 2014.