A short manual for TESS3: a program to estimate spatial population structure
(command-line and R wrapper)

Kevin Caye (kevin.caye@imag.fr)
Olivier François (olivier.francois@imag.fr)

June 11, 2015

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

This short manual aims to help users to run TESS3 command-line engine and R wrapper function on Mac, Linux and Windows.

1 Description

Inference of spatial population structure are commonly perform with computer program based on intensive stochastic simulation. These methods do not scale with the dimension of the data sets generated from next-generation sequencing technologies. The computer program TESS3 has functionalities similar to the spatial bayesian clustering program TESS [2], but has run-times several order faster than those of TESS 2.3. The method is based on geographically constrained non-negative matrix factorization, and provides similar results that with TESS 2.3. In addition TESS3 computes an ancestral allele frequency differentiation statistic that can be used to perform selection scans.

2 Installation

The source code of TESS3 can be download on github: `https://github.com/cayek/TESS3/archive/master.zip`. Then you just need to follow instructions on the github project page `https://github.com/cayek/TESS3`.

3 Data format

3.1 input file

The sNMF input file consists of a genotype file in the **geno** format and coordinate file in the **coord** format.

- **geno** (example.geno)

  The **geno** format has one row for each SNP. Each row contains 1 character per individual: 0 means zero copies of the reference allele. 1 means one copy of the reference allele. 2 means two copies of the reference allele. 9 means missing data.

  Below, 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 latitude` information of the individual.

  Below, an example of a geno file for $n = 3$ individuals and $L = 4$ loci.

```
2.5154552002036901 11e+01 5.439077948590419709e+01
-8.429345306090160861e+00 4.019713388759510053e+01
1.351129055178800087e+01 5.585331731335860184e+01
```

Other formats for genotype data sets can be used thanks to the `LEA` R package [3]. Indeed, `LEA` enable to convert into **geno** format the following usual formats : **ped**, **ancestrymap**, **vcf** and **lfmm**.

3.2 output files

There are three main **output files**.

- The file with the extension the **.Q** contains individual admixture coefficients. It contains a matrix with n rows (the number of individuals) and K columns (the number of ancestral populations).
- The file with the extension **.G** contains the ancestral genotypic frequencies. It contains a matrix with $n_a \times L$ lines (the number of alleles times the number of SNPs) and K columns (the number of ancestral populations). For a diploid SNP, the first line contains the ancestral frequencies for the number of allele equals to 0, the second line contains the ancestral frequencies for the number allele equals to 1, the third line contains the ancestral frequencies for the number of alleles equal to 2.
- The file with the extension **.Fst** contains ancestral allele frequency differentiation statistic. In this file each row is the statistic estimate for each loci.

There also are less important **output files**, but that can be useful to have information on the graph computed, the least-squared criterion and the cross-entropy criterion.

- The file with the extension the **.W** contains the edge weight matrix of the nearest-neighbor graph computed from coordinates data.
- The file with the extension the **.sum** contains the value of least-squared criterion, the cross entropy criteria on all data and only on masked data if the user asked it.

4 Run the program

4.1 Command-line

The `TESS3` program can be executed from a command line. The format is:

```
./TESS3 -x genotype_file.geno -K number_of_ancestral_populations  -r coordinates_file
```

Only these three options are mandatory. There is no ordering for the options in the command line. Here is a description of these options:

- `-x genotype_file.geno` is the path to the genotype file (in .geno format).
- `-K number_of_ancestral_populations` is the number of ancestral populations.
- `-r coordinates_file` is the path to the coordinate file (in .coord format).

Additional options are available:

- `-a alpha` is the value of the normalized regularization parameter (by default: 0.001). This parameter control the spatial regularity of the ancestry estimates.
- `-W edge_weight_input` is the path to the file that contains a edge weight matrix of the spatial graph. The file has to contain each element of the matrix separates by a space row by row. The program use this graph in place of the one computed if no edge weight input file is given.
- `-q output_Q` is the path for the output file containing the ancestry coefficients. By default, the name of the output file is the same name as the input file with the extension .K.Q.
- `-g output_G` is the path to the output file containing the ancestral genotype frequencies. By default, the name of the output file is the same name as the input genotype file with the extension .K.G.
- `-f output_FST` is the path to the output file containing the ancestral allele frequency differentiation statistic. By default, the name of the output file is the same name as the input genotype file with the extension .K.Fst.
- `-y output_FST` is the path to the output file containing the value of least-squared criterion, the cross entropy criteria on all data and only on masked data. 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 calculated (see [4] for more details on the cross-entropy criterion). The default percentage is 5%.
- `-e tolerance` is the tolerance error in the `sNMF` optimization algorithm (by default: 0.0000001).
- `-i iteration_number` is the max number of iterations of the algorithm (default: 200).
- `-I nb_SNPs` starts the algorithm with a run of sNMF using a subset of nb_SNPs random SNPs. This option can speed up sNMF estimation for very large data sets.
- `-Q input_Q` is the path to an initial file for the Q matrix containing individual admixture coefficients. If both `-I` and `-Q` are set, `-Q` is chosen.
- `-s seed` is a seed to initialize the random number generator.
- `-m ploidy` 1 if haploid, 2 if diploid (default: 2).
- `-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).

If you need a summary of options, you can use the `-h` option by typing the following command

```
./bin/sNMF -h
```

4.2 R wrapper

The `TESS3` program can be executed using a wrapper in R software environment. The wrapper and helper functions are defined in the R script `src/Rwrapper/TESS3.R`, the user can directly source this script in a R session. We now present function define in this script:

- `TESS3`

  Description The wrapper function that call the command-line program. This function create a directory `TESS3_workingDirectory` to store input and output file.

  Usage

```
project = TESS3( genotype,
                 spatialData,
                 K,
                 ploidy=1,
                 seed=-1,
                 rep = 1,
                 maskedProportion = 0.0,
                 alpha = 0.001 )
```

  Arguments

  - `genotype`: genotype R matrix of size n individual by L loci or the .geno format file name.
  - `spatialData`: coordinate R matrix of size n individual by 2 or the .coord format file name.
  - `K`: vector of number of ancestral cluster. The `TESS3` program is run for each element of this vector.
  - `ploidy`: 1 if haploid, 2 if diploid.
  - `rep`: number of run for each number of ancestral population.
  - `maskedProportion`: if `maskedProportion > 0`, the cross-entropy criterion is calculated for this percentage of masked genotypes.
  - `alpha`: value of the normalized regularization parameter. This parameter control the spatial regularity of the ancestry estimates.

- `Getter`

  Description Functions useful to fetch results of `TESS3` R function.

Usage

```
getQ( project, K, run = "best" )

getG( project, K, run = "best" )

getFst( project, K, run = "best" )

getCrossEntropy( project, func = mean )

getLeastSquared( project, func = mean )
```

Arguments

- `project`: object returned by the `TESS3` R function.
- `K`: number of ancestral cluster.
- `run`: number of the run, or `"best"` if you want the best result with respect to the least-squared criterion.
- `func`: function used to summarize data over all run.

- `Reader`

Description Function useful to read data from file.

Usage

```
read.coord( file )
```

Arguments

- `file` name of the file to read.

A full example in R is available at the end of this note.

5 Tutorial

5.1 Data set

The data set that we analyse in this tutorial is a simulated dataset from a *Arabidopsis thaliana* Data set used in [1]. We obtain a haploid data set of $n = 170$ individual and $L = 26943$ loci.

5.2 Example of analysis using `TESS3`

This tutorial example describe how to run `TESS3` in R for different values of K the number of ancestral populations. Then we plot the cross entropy criterion with respect to K. Finally we use R script available on `POPS` website (see http://membres-timc.imag.fr/Olivier.Francois/pops.html) to plot the ancestry coefficient with the K found with the cross entropy criteria [5].

```
source("TESS3_directory/src/Rwrapper/TESS3.R")
library(LEA)

setwd( "TESS3_directory/data/simulated/Athaliana" )
#############################################################################
# Run TESS3 on a data set simualted from an Arabidopsis Athalina data set #
#############################################################################

#read data
spatialData = read.coord("Athaliana.coord")
n = nrow(spatialData)
```

```
project = TESS3( genotype = "Athaliana.geno",
                 spatialData = "Athaliana.coord",
                 K = 1:5,
                 ploidy = 1,
                 rep = 1,
                 maskedProportion = 0.2)


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

plot( 1:5,
      getCrossEntropy( project ),
      main  = "Cross entropy",
      type="b",
      xlab = "K",
      ylab = "corss entropy" )

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

# R script available on http://membres-timc.imag.fr/Olivier.Francois/pops.html
source("POPS_direction/R/POPSutilities.r")

asciiFile="/home/cayek/Projects/TESS3/data/simulated/Athaliana/down_etopo1.asc"
grid=createGridFromAsciiRaster(asciiFile)
# To display only altitudes above 0:
constraints=getConstraintsFromAsciiRaster(asciiFile,cell_value_min=0)

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

5.3 Example of genome-scan for selection using TESS3

This tutorial show an example in R of use of the ancestral allele frequency differentiation statistic computed by TESS3. We first run TESS3 with $K = 3$ ancestral populations. Then using standard population genetic theory the F_{ST}-statistic is transformed into squared t-score and p-value are computed using a fisher distribution. The test inflation due to neutral population structure is corrected by recalibrating the t^2-score to have uniform distribution under the null hypothesis. Finally, we control the false discovery rate using a Benjamini Hochberg procedure and plot the Manhattan plot.

```
source("/home/cayek/Projects/TESS3/src/Rwrapper/TESS3.R")
library(LEA)

setwd( "/home/cayek/Projects/TESS3/data/simulated/Athaliana" )
#############################################################################
# Run TESS3 on a data set simualted from an Arabidopsis Athalina data set #
#############################################################################

#read data
spatialData = read.coord("Athaliana.coord")
genotype = read.geno("Athaliana.geno")
n = nrow(spatialData)

project = TESS3( genotype = genotype,
                 spatialData = spatialData,
```

```
                K = 3,
                ploidy = 1,
                rep = 5 )


############################
# Genome scan for selection #
############################

#### Fst with TESS3
Fst = getFst( project, K = 3 )
Fst[Fst < 0.0] = 0.0

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

#### recalibrated p-values
gif = 25
adj.p.values = pf( squared.t.scores/gif , df1 = 2, df2 = n-3, lower = FALSE )

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

#### 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])

#### 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 )
```

6 Contact

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

References

[1] Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

[2] 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.

[3] Eric Frichot and Olivier François. Lea: an r package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 2015.

[4] 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.

[5] Flora Jay, Stéphanie Manel, Nadir Alvarez, Eric Y Durand, Wilfried Thuiller, Rolf Holderegger, Pierre Taberlet, and Olivier François. Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, 21(10):2354–2368, 2012.