TESS3 reference manual
R package

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

October 19, 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 has functionalities similar to the previous versions of TESS [2,3], 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 help users to run the R package tess3r which implements TESS3 with some R functions that facilitate the post-processing of the program outputs. The main features of TESS3 are illustrated using an example data set, simulated from European lines of the plant species *Arabidopsis thaliana*.

1 Program installation

The installation of tess3r R package requires that R is install on your computer (https://www.r-project.org/). You can install the R package directly from the github repository thanks to the package devtools. If you don't already have devtools R package you can install it from CRAN. In a R session paste this command:

```
install.packages("devtools")
```

Then, you can install the R package. In a R session paste this command:

```
devtools::install_github("cayek/TESS3/tess3r")
```

2 Data format

2.1 Input files

The R package tess3r handles two kind of input files, the first one recording individual genotype data and the second one containing the geographic coordinates of each sampled individual. For organism genomes of arbitrary ploidy, the standard data type for tess3r 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 tess3r. They need to convert their data in the geno+coord data format, and can do this using the tess2tess3 function implemented in the package.

- **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].

3 Run TESS3

In a `R` session the `TESS3` program can be run by typing:

```
> # Main parameters:
> # input.file  is the genotype data file at .geno format
> # input.coord is the coordinates data file at .coord format
> # ploidy      is the ploidy of the species,
> #             here we assume that the data come from haploide species
> # K           is the number of ancestral population,
> #             here we run the algorithm for K equals from 2 to 4
> # repetition  is the number of algorithm run per parameter set.
> tess3.obj = TESS3( input.file = "genotype.geno",
>                    input.coord = "coordinates.coord",
>                    ploidy = 1
>                    K = 2:4,
>                    repetition = 5)
```

Then, ancestry coefficients Q matrix, ancestral population allele frequencies G matrix and the F_{ST} vector can be retrieve:

```
> Q = Q(tess3.obj, K = 3, repetition = 1)
> G = G(tess3.obj, K = 3, repetition = 1)
> Fst = FST(project, K = 3, repetition = 1)
```

Other features of `tess3r` package are illustrated in more complete tutorial: `https://github.com/cayek/TESS3/raw/master/doc/tess3r_tutorial.pdf`

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 comparison 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.