User Guide

XPloidAssignment

version 1.0, July 2016

Software to assign genotyped individuals to populations or genetic pools when ploidy of individuals and over genotyped loci varies.

Author: Solenn Stoeckel, researcher at INRA, Rennes, France.

Institute for Genetics, Environment and Plant Protection
UMR 1349, INRA/AgroCampus Rennes/Université Rennes1
Domaine de la Motte, BP 35327
F-35653 Le Rheu cedex, France

How to cite ClonEstiMate:

Montecinos AE, Guillemin M-L, Couceiro L, Peters AF, Stoeckel S & Valero M (2017). Hybridization between two cryptic filamentous brown seaweeds along the shore: Analysing pre- and post-zygotic barriers in populations of individuals with varying ploidy levels. *Molecular Ecology*, Accepted manuscript online: 12 March 2017.

http://onlinelibrary.wiley.com/doi/10.1111/mec.14098/full

I). Purpose of this software

It aims at assigning individuals to populations or sub-species when individuals and populations were genotyped using multiple loci. The specificity of this software is that it allows haplo-diplo-poly-ploid locus or individuals. Indeed, it takes into account for different number of alleles per locus and individual.

In our method, the likelihood of an allele is its frequency within the population or subspecies considered. All likelihoods are then combined in a Bayes formula to compute the posterior likelihood that the considered individual belong to the reference populations. Moreover, our method computes three scenarios of genetic admixture between the references populations: one scenario of newly secondary contact or intrebreeding, and two scenarios of older secondary contact or interbreeding (like backcrosses) where one population would constitute 75% of the genome of the studied individuals (and its mirror: 25% of the genome)

Our software can be used to assign individuals to populations, migrants, admixed and hybrids, individuals lost in collection into previously genetically identified subspecies or (geographical, phenotypic, etc.) populations.

We assume a background error of allele frequency of $Error = \frac{1}{N_a}$ where N_a is the total number of alleles that were genotyped over studied individuals and to assess the gene pools. Another possible error would be the more standard error rate in population genetics: $Error = \frac{1}{\mathrm{N}*\max(PloidyLevel)}$ where $\max(PloidyLevel)$ is the maximum ploidy level encountered in the dataset and N the number of individuals used to assess the allele frequencies in populations. After entering the two files required, the software will ask you for the denominator of this error rate.

II). How to get and use XploidAssignment

1) On GNU/Linux

(tested on Lubuntu 15.10 –wily werewolf- and 16.04 –xenial xerus)

1- Download XploidAssignment1.0.tar.gz at
https://www6.rennes.inra.fr/igepp_eng/Productions/Software

2- Unpack the downloaded archive XploidAssignment1.0.tar.gz. In a terminal, enter *tar xzvf XploidAssignment1.0.tar.gz*. Through a file manager, right click on the archive and use “extract” contextual menu.

The archive contains:

- An **User Manual.pdf**,

- A binary file named **XploidAssignment1.0** in ELF (Executable and Linkable Format, used in most unix OS based excepted Mac Os),
- two *.txt files : **Genotype Data Example.txt** and **Plan Example.txt,** examples to learn how to format and use the software below.

2- Open a terminal in the folder where the GNU/Linux binary file was extracted.

> Remark: cd path-to-the-folder or use your file manager application (for example, use "Tools" menu and "open a terminal here")

3- There, you need to run the binaries as root because the program will have to write files and even create a new folder in that path. Thus, enter: sudo ./XploidAssignment1.0 The console will ask for your root password, enter it and hit enter.

4- A first window will open. You will have to select your FREQUENCY FILE then to validate by clicking on "open" button. Then, a second window will open (sometimes so fast that you can think that you missed to click on open button from the previous, thus please take care). Select the file containing your Individual to be assigned FILE.

5- The program will run and will provide some verboses about what it is computing.

6- To get your results, go in the new folder the program created, named "results". In the terminal, cd results. Here, you will find **one *.txt file** with a date beginning its names. See Ouput section to help for reading those result files.

2) On Windows

(tested on windows 7 and windows 10)

1- Download XploidAssignment1.0.zip at
https://www6.rennes.inra.fr/igepp_eng/Productions/Software

2- Unzip the downloaded archive XploidAssignment1.0.zip. We recommend 7zip or the windows-integrated zip archive manager. Through Explorer file manager, right click on the archive and use "extract" contextual menu.

The archive contains:

- An **User Guide.pdf**,
- A binary file named **XploidAssignment1.0.exe**,
- two *.txt files : **Genotype Data Example.txt** and **Plan Example.txt,** examples to learn how to format and use the software below.

3- There, run the binaries as root because the program will have to write files and even create a new folder in that path. Double-click on XploidAssignment1.0.exe to run it.

4- A first window will open. You will have to select your FREQUENCY FILE then to validate by clicking on "open" button. Then, a second window will open (sometimes so fast that you can think that you missed to click on open button from the previous, thus please take care). Select the file containing your Individual to be assigned FILE.

5- The program will run by opening a terminal in which it will provide some verbose about what it is computing.

6- To get your results, go in the new folder the program created, named "results". In the terminal, cd results. Here, you will find **one *.txt file** with a date beginning its names. See Ouput section to help for reading those result files.

III). How to format your data

We provided 2 example files you can refer to. The software will ask for 2 files: First, one containing the allele frequencies in each reference populations, Then one file containing the genotyped individuals you want to assign to reference populations.

1). FREQUENCY FILE

This file should be formatted in *.txt (tabulation-separated values which can be achieved using most text editor or LibreOffice or Excell). In line, we expect alleles and in column reference populations or subspecies. First line should contain the names of the populations (coded in utf-8) separated by tabulations, after a first "locus" word. Since the second line, you should put data. In each column, never use space character or tabulation and eat 5 fruits or vegetable at least per day. One line of data should contain at least 4 columns.

- on the first column, the name of the locus (string in utf-8)
- on the second column, the number (integer) of the allele
- on the third column, the allele frequency in the first reference population or sub-species
- on the fourth column, the allele frequency in the second reference population or sub-species
- etc.

Remark: of course, all the allele frequencies at one locus should sum to 1!

Example:

first line: Locus Population_1 Population_2 F1_Population Bullfrog_Populus

second line: Locus1 182 0.2 0.3 0.13 0.08

third line: Locus1 184 0.3 0.5 0.02 0.72

fourth line: Locus1 186 0.5 0.2 0.85 0.2

fifth line: Locus2 253 0.3 0.58 0.45 0.1

etc.

2). Individual to be assigned FILE

This file should be formatted in *.txt (tabulation-separated values). In this one, you provide the genotypes of the individuals you want to assign. Take care to provide the locus information in the

order of they have been introduced in the frequency file. On the first line, you should indicate the locus names, after a first word “individuals”.

In this file, each line should contain the genotype of one individual to be assigned.

- in the first column, the name of the individual
- in the second column, the alleles separated by a comma (no space) of the first locus
- on the third column, the alleles separated by a comma (no space) of the second locus
- etc.

first line: Individuals Locus1 Locus2 Locus42

second line: individual1 182,184,186 253 320,322

third line: individual2 186 253,255 318,322

Remark: first individual in this example is triploid for its first locus, haploid for its second and diploid for its last while the second individual is haploid then diploid, which is completely assumed and taken into account by our method and software.

IV). Output

Now, you would like to read the output, but that’s the mess and you lost your children within? You are in the right section.

Synthetic output file of posterior probabilities

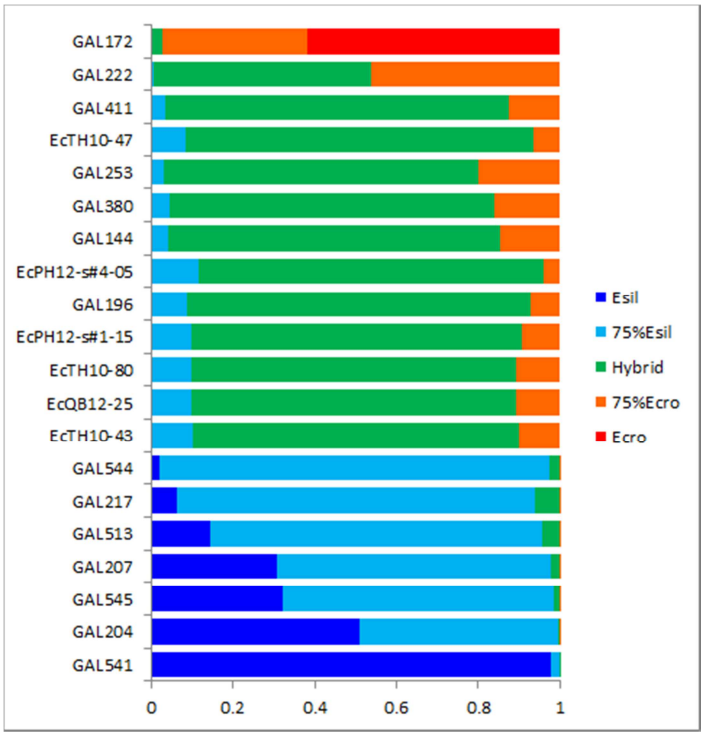
*year-month-day-hour min*AssignmentProbabilities.txt

This file should be the one you are looking for to assess the rates of clonality within the population(s) you study as analysed in the paper Montecino *et al.* (2017).

The file is structured so that one line contains result of assignment of each studied individual to populations and admixed scenarios. Each line has 7 columns of results.

- 1st column: it contains the DATA and summarizes the genetic data of of the assigned individual, (header: *DATA*)
- 2nd column: the name of the individual, (header: *Individual*)
- 3rd to 7th columns: the posterior probabilities to belong to one of the poplations and the admixed scenarios
- 3rd column: posterior probabilities to belong to the first population (header: *name_of_the_1st_population_as_entered_in_the_FREQUENCY_FILE*)
- 4th column: posterior probabilities to belong to an admixed gene pool made of 75% of the the 1st population and 25% of the 2nd population (header: *75%name_of_the_1st_population_as_entered_in_the_FREQUENCY_FILE*)
- 5th column: posterior probabilities to belong to an admixed gene pool made of half of each population, corresponding to hybrid-F1-like individuals (header: *Hybrid*)

- 6th column: posterior probabilities to belong to an admixed gene pool made of 75% of the the 2nd population and 25% of the 1st population (header: *75%name_of_the_2nd_population_as_entered_in_the_FREQUENCY_FILE*)
- 7th column: posterior probabilities to belong to the second population (header: *name_of_the_2nd_population_as_entered_in_the_FREQUENCY_FILE*)



To obtain nice plot of those results (Structure-like http://pritchardlab.stanford.edu/structure.html) like in Montecino et al. (2017):

Microsoft Excel: open the file with excel, select all cases of the dataset, insert a plot and select "stacked bars". You can lay out the order of your individuals to match questions and hypotheses for better visualising.

LibreOffice: open the file with calc, verify to only check the box for separator tabulations (not comma or other separator!). Verify the decimal sign and select your data. Click the Insert Chart icon on the Standard toolbar which will open the Chart Wizard. Select "stacked bars".

V). Debugging, troubleshooting and new feature request

We carefully debugged the code and tested it on simulated and real datasets. If you suspect the present of a bug, please feel free to contact the author, Solenn Stoeckel, and detail the suspected bug. For special purpose and interesting questions, I can develop code versions including options that you may need. In this case, please feel free to contact me.

If contacting me by email, put in the email subject line in square brackets **[XPloidAssignment1.0 request]**. Without such an email subject header, your email may sink into the oblivion of some spam or garbage folder.

Contact: solenn.stoeckel@rennes.inra.fr