User Guide

ClonEstiMate

version 1.02, March 2018

Software to estimate rates of clonality in finite populations that mutate using two temporal genotyped samples.

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:

Becheler Ronan, Masson Jean-Pierre, Arnaud-Haond Sophie, Halkett Fabien, Mariette Stéphanie, Guillemin Maire-Laure, Valéro Myriam, Destombe Christophe, Stoeckel Solenn (2017). *ClonEstiMate, a Bayesian method for quantifying rates of clonality of populations genotyped at two-time steps.* **Molecular Ecology Resources**, 17, e251–e267. https://doi.org/10.1111/1755-0998.12698

I). Purpose of this software

This software aims at inferring rates of clonality in populations suspected to reproduce using partial and full clonality. This method uses as fuel at least two temporal genotyping of a same population which is performed in likelihood (computed from an explicit mathematical population genetics scenario) “motor” to produce the posteriori probabilities that each genotyped locus evolved as expected under a list of semi-quantitative rates of clonality. The population genetics models we developed assume that the evolution of genotype frequencies follows a classical Wright-Fisher model modified to include explicitly partial clonality as developed in Stoeckel & Masson 2014. In this method, we thus doesn’t rely on repeated genotypes or other proxy for evolution of genetic diversity under partial and full clonality, but use the full information of genetic diversity transitions between two generations. The method approximates the combined posterior probability over all loci as independent loci (classical “naïve Bayes” approximation), which can result to overtrust the maximum posterior when loci at tightly linked, especially in highly clonal populations.

As a theoretician and user of population genetics models, I recommend to cross validate all results you can obtain from one model/method/software with all available other methods and knowledge at your disposal, and to keep your naturalist knowledge and intuition widely awaken when interpreting outputs.

II). How to get and use ClonEstiMate

1) On GNU/Linux

(tested on Lubuntu 15.10, 16.04, 16.10, 17.10)

1- Download ClonEstiMate1.02.tar.gz at *https://www6.rennes.inra.fr/igepp_eng/Productions/Software*

2- Unpack the downloaded archive ClonEstiMate1.02.tar.gz. In a terminal, enter *tar xzvf ClonEstiMate1.02.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 **ClonEstiMate1.02** 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 ./ClonEstiMate1.02` 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 **input data 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 **plan of analyses**.

5- The program will run and 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 **4 different *.txt files** with the same date beginning their names. See Ouput section to help for reading those result files.

2) On Windows

(tested on windows 7 and windows 10)

1- Download ClonEstiMate1.02.zip at *https://www6.rennes.inra.fr/igepp_eng/Productions/Software*

2- Unzip the downloaded archive ClonEstiMate1.02.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 **ClonEstiMate1.02.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 ClonEstiMate1.0.exe to run it.

4- A first window will open. You will have to select your **input data 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, thus please take care). Select the file containing your **plan of analyses**.

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 **4 different *.txt files** with the same date beginning their 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 all the genotypes you want to analyse, Then one file containing the plan of inferences you want to perform.

1). DATA FILE

This file should be formatted in *.txt (tabulation-separated values). In line, we expect individuals and in column alleles. First line should contain comments (utf-8) and even tabulations if you need them. If comments annoy you, please leave at least one character on this line anyway like one space. Since the second line, you should put data. In each column, never use space character or tabulation and take care to never feed a column after midnight. One line of data should contain at least 4 columns.

- on the first column, the name of the population (should be the same for all individuals within the pop)
- on the second column, the generation/time of the sample (remember we expect at least two temporal sample of a same population)
- on the third column, the first allele (expected to be coded as an integer number) of the first locus you genotyped
- on the fourth column, the second allele (remember the current version only work for diploid) of the first locus you genotyped

then, the other columns should contain alleles of other loci you genotyped: eg

- on the fifth column, the first allele (remember the current version only work for diploid) of the second locus you genotyped
- on the sixth column, the second allele (remember the current version only work for diploid) of the second locus you genotyped
- …etc.

#first line: annoying comments about your data that nobody cares else you (as usual, welcome into research)

# second line: Population_name   Generation_code   allele_1_locus1   allele_2_locus1   allele_1_locus2   allele_2_locus2 ... allele_1_locusk   allele_2_locusk (of individual 1 sampled in the population)

# second line: Population_name   Generation_code   allele_1_locus1   allele_2_locus1   allele_1_locus2   allele_2_locus2 ... allele_1_locusk   allele_2_locusk (of individual 2 sampled in the population)

With k, the number of loci you genotyped… Frankly, go on. Target a lot of loci and spend lavishly… if you read that you are probably a PhD student or a Postdoc, and your supervisor is probably too busy to take care of his/her impact factor and ego to spare no expense. Moreover, funders are still behaving like in Deadalus time: their desires and hubris are so inflated that they ask you for the moon and complain when you succeeded and by a side effect, fall a massive meteor shower on humanity.

2). PLAN FILE

This file should be formatted in *.txt (tabulation-separated values). In this one, you provide the plan of the inferences you want to perform over all the data you included in the first file. Yes, you can do multiple inferences between populations sampled at different times without relaunching the software and enter again masses of inputs nobody cares out of those f***ing software authors!

In this file, each line will result to perform one inference on one population sampled at two generations.

- in the first column, the name of the population (should be the same used in column 1 of DATA FILE)
- in the second column, code of the first generation/time you sampled (coded as in the second column of DATA FILE)
- in the third column, code of the second generation/time you sampled (coded as in the second column of DATA FILE)
- in the fourth column, put here all the prior mutation rate values you want to use to infer rates of clonality. Separate multiple prior values by a coma without space. We recommend by default 0.001,0.000001,0.000000001 as mutation prior seem to have low impact on results out of some precise situations.
- in the fifth column, put here all semi-quantitative selfing rates you want to assess as prior. Separate multiple values by a coma without space. Enter 0 in this column if no selfing in the population.
- in the sixth column, put here all semi-quantitative inbreeding rates you want to assess as prior. Separate multiple values by a coma without space. Enter 0 in this column if no inbreeding in the population.
- in the seventh column, put here all semi-quantitative rates of clonality you want to assess. Separate multiple values by a coma without space. We recommend 0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95,0.98,0.99,0.999,0.9999,1 but if you want to be more precise or you like Marquis de Sade fantasies, knock yourself out. If covering the full semi-quantitative range of rates of clonality is too mainstream for you, you can also provide at least 2 values between 0 and 1 included.

```
# Population_name   Generation_code1   Generation_code2     list_of_mutation_priors
list_of_rates_of_clonality
```

IV). Computational complexity

The more mutation, selfing, inbreeding rates and rates of clonality you will ask the software to be estimated and you provide in the plan file as prior, the longer the computation.

V). Outputs

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

1). Synthetic output file of posterior probabilities

*year-month-day-hour min*ProbPostSynt.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 Becheler *et al.* (2017).

The file has 4 columns to identify one line of estimates then multiple numerical (floats) columns of results.

- 1st column: the name of the population, (header: *Pop_Name*)
- 2nd and 3rd columns: the coded date of temporal samples. (header: *Time_t* and *Time_t+1*)
- 4th column: the prior value of mutation rate that configured the population genetic model to compute the log-likelihood. (header: *Mutation_Rate(prior)*)
- 5th column: the prior value of selfing rate that configured the population genetic model to compute the log-likelihood. (header: *Selfing_Rate(prior)*)
- 6th column: the prior value of inbreeding rate that configured the population genetic model to compute the log-likelihood. (header: *Inbreeding_Rate(prior)*)
- The next columns contain the discrete distribution of posterior probabilities per classes of assessed rates of clonality (entered in file "plan"). Together, those values shape the posterior distribution of the rates of clonality (parameter of our model) after taking into account your observed data (transition of genotypic frequencies per locus per population between two temporal sampling). Remember in the method that the data are the transitions of genotypic frequencies per locus per population, not genotypes nor genetic proxies nor identities of genotypes. Those transitions come from a temporal-explicit population genetics model (Wright-Fisher like) taking into account for population size (here the filial sample size as transition are the change from parental population of the sample of offspring you genotyped, mutation rate and rate of clonality). This population genetics model is extensively detailed into Stoeckel & Masson (2014, Plos One). In header, the value of the rate of clonality (eg *0.45*) Underneath, the posterior probability of the class (*ie* the rate of clonality assessed as prior, *c*). Each line since the 5th column should sum to 1.

2). Synthetic output file of Maximum a posteriori probabilities (MAP)

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

In this file, you find the number of loci that estimated one rate of clonality as a maximum a posteriori (see https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation for example for more details).

The file has 4 columns to identify one line of estimates then multiple numerical (integers) columns of results.

- 1st column: the name of the population, (header: *Pop_Name*)
- 2nd and 3rd columns: the coded date of temporal samples. (header: *Time_t* and *Time_t+1*)

- 4,5&6th column: the prior values of mutation, selfing and inbreeding rates that configured the population genetic model to compute the log-likelihood. (header: *Mutation_Rate(prior), Selfing_Rate(prior), Inbreeding_Rate(prior)*)
- The next columns are classes of assessed rates of clonality (entered in file "plan"). In header, the value of the rate of clonality (eg *0.45*) Underneath, the number of locus for which maximum a posteriori pointed out this class of *c* value. The sum of each line since the 5th column should be the number of loci you genotyped the population.

Remark: Maximum A Posteriori by loci can estimate different *c* value than using the Naives Bayes classifier approach. Our Bayesian classifier integrates all likelihood of all loci into a posterior distribution while the MAP provides the sum of its maximum parts. As usual, looking at different method is a plus to support or doubt a result.

3). Posterior probabilities by locus

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

In this file, you find the log-likelihood of estimated rates of clonality given the input data at each locus. This file should have x lines and 7 columns, where x equals the number of populations times the number of mutation prior times the number of rates of clonality assessed times the number of couple of sampling dates compared.

- 1st column: the name of the population, (header: *Pop_Name*)
- 2nd and 3rd columns: the coded date of temporal samples. (header: *Time_t* and *Time_t+1*)
- 4,5&6th column: the prior values of mutation, selfing and inbreeding rates that configured the population genetic model to compute the log-likelihood. (header: *Mutation_Rate(prior), Selfing_Rate(prior), Inbreeding_Rate(prior)*). You can theoretically assess best mutation rate using the best likelihood, or, if you know it, use this proxy to evaluate the credibility of the estimates.
- 5th column, the value of the parameter "rate of clonality" that is estimated in the likelihood function. (header: *Rate_of_Clonality(prior)*)
- 6th column: the locus number (it starts from zero), sets in the order of the loci in your input file. (header: *Locus_Number*)
- 7th Column: the log-likelihood of the set of parameters (mutation rate and rate of clonality) considering our population genetics model. (header: *Log(Likelihood)*)

4) Synthetic output file of the sum of log-likelihood

This file is only for calculus and computing purpose. You can dig into it to compute by hand some likelihood ratio or Bayesian posterior.

The file has 4 columns to identify one line of estimates then multiple numerical (floats) columns of results.

- 1st column: the name of the population, (header: *Pop_Name*)
- 2nd and 3rd columns: the coded date of temporal samples. (header: *Time_t* and *Time_t+1*)
- 4,5&6th column: the prior values of mutation, selfing and inbreeding rates that configured the population genetic model to compute the log-likelihood. (header: *Mutation_Rate(prior), Selfing_Rate(prior), Inbreeding_Rate(prior)*)

- The next columns are classes of assessed rates of clonality (entered in file “plan”). No header there, but they are the same as in MAP and Posterior probagbility files. By column (value of the rate of clonality, eg *0.45*), you find the sum of the log-likelihoods overall loci

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 **[ClonestiMate1.02 request]**. Without such an email subject header, your email may sink into the oblivion of some spam or garbage folder.

Contact: solenn.stoeckel@ inra.fr

VI). Release notes

09/04/2018 - ClonEstiMate **1.02** – *user requests*
New features: added an automated analysis of input files with verbose to point input file error(s), and an intern debugger that output errors into a “ErrorFileLOG”. Such file should be sent to the author when facing an incomprehensible error that cannot be remedied.
Added a “data file” to be filled and a “analysis plan file” to be filled to ease future analyses.
Unit tests succeeded.

20/02/2017 – ClonEstiMate **1.01** – *reviewer requests*
New features: added priors on selfing rates and inbreeding rates. Best combinations can be weighed using likelihoods.
Changes: Popping windows for entering the two input files are now by default opening in the directory where ClonEstiMate software is located.

17/05/2016 - ClonEstiMate **1.0**
Initial release.