```
////////////////////////////////////////////////////////////////////////////////////////////////
//                                                                                          ////
//          RandomMatingModel v1.3                                                          ////
//          Copyright© - INRA - 2009                                                        ////
//          Klein EK, Desassis N, Oddou-Muratorio S, Carpentier F, Bonnefon O               ////
//          Licence GPL                                                                     ////
//                                                                                          ////
//          Bayesian estimation of individual fecundities and pollen dispersal kernel       ////
//          from microsatellite data on seeds and adult trees.                              ////
//                                                                                          ////
//          This code was developed by Etienne Klein, INRA Avignon, BioSP                   ////
//          The reference to this method is                                                 ////
//          Klein EK, Desassis N, Oddou-Muratorio 2008. Molecular Ecology. 17: 3323-3336    ////
//                                                                                          ////
//          Improved: Simplified Burn-in and non-burn-in loops                              ////
//          Improved: Gamma and LN options provided as objects                              ////
//          Included: Option to read distances from an external files                       ////
//                     !!!!!! New Parameters.txt files !!!!!                                ////
//                                                                                          ////
//                                                                                          ////
////////////////////////////////////////////////////////////////////////////////////////////////
```

# Mixed Effects Mating Model (v1.2)

## Introduction

This software is an executable developed with C++ (MCMCGeneration) that runs the Bayesian approach presented in Klein et al. 2008. It is associated with a code developed in R (MCMCVisu.R) that provides four simple functions to visualize the results. However, any user familiar with R can make his/her own favourite figures using the text files containing the MCMC data generated by the C++ executable.

Here we present shortly the main information necessary to run the program. Both the code and this manual are early first versions that we hope to expand in a next future. Any suggestion or bug detected are welcome (etienne.klein@avignon.inra.fr).

Klein, E.K., Desassis, N., Oddou-Muratorio, S. (2008) Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. Whole inter-individual variance of male fecundity estimated jointly with dispersal kernel. **Molecular Ecology**. 17: 3323-3336.

## Installation

Pre-compiled versions are available for MacOSX (intel), Windows and Linux. All are available at http://MEMM.biosp.org.

To install the program, just copy the unzipped directory anywhere in the computer. To run the program, double-click on the executable file after having modified the `Parameters.txt` file and placed the input file in the appropriate directory (i.e. in accordance with the `Parameters.txt` file).

## The parameters file.

```
/Users/klein/Chercher/C/Genotypes1/AlisierPar.txt
/Users/klein/Chercher/C/Genotypes1/AlisierDesc.txt
6 6 2 1
file_dist /Users/klein/Chercher/C/Genotypes1/distMP.txt
file /Users/klein/Chercher/C/Genotypes1/freqallext.txt
LN
12432
2. 1. 1000.
100. 0. 10000.
1. 0.1 10.
0.5 0.1 1.
0.05 0. 0.1
5000 20000 50
/Users/klein/Chercher/C/Genotypes1/ParamFec.txt
/Users/klein/Chercher/C/Genotypes1/IndivFec.txt
/Users/klein/Chercher/C/Genotypes1/ParamDisp.txt
```

Name : `Parameters.txt`

Location : Most generally, this file should be in the same directory as the executable file. For some computers running Mac OSX, this file should be in the home directory of the user.

Function : This file contains all the parameters used to run the MCMC: location of the data, location of the outputs, tuning parameters of the MCMC, model choices…

Annotated example:

```
/Users/klein/Demo/AlisierPar.txt                        file containing information about parents[1]
/Users/klein/Demo/AlisierDesc.txt                      file containing information about offspring[1]
6 6 2 1                                                       numbers of locus, class covariates,
                                                      quantitative covariates, weighting variables[2]
file /Users/klein/Demo/distMP.txt         mode of computation of distances and file containing distances[7]
file /Users/klein/Demo/freqallext.txt                   mode of computation of AF and file containing AF[3]
LN                                                      distribution chosen for individual fecundities[4]
12432                                                                      seed for the random generator[5]
2. 1. 1000.                                                   initial, minimum and maximum value for dobs/de
100. 0. 10000.                                initial, minimum and maximum value for mean dispersal distance delta
1. 0.1 10.                                               initial, minimum and maximum value for shape parameter b
0.5 0.1 1.                                               initial, minimum and maximum value for migration rate m
0.05 0. 0.1                                                initial, minimum and maximum value for selfing rate s
5000 20000 50                          numbers of burn-in iterations, definitive iterations, thinning parameter[6]
/Users/klein/Demo/ParamFec.txt                                                       output file for dobs/de[1]
/Users/klein/Demo/IndivFec.txt                                          output file for individual fecundities[1]
/Users/klein/Demo/ParamDisp.txt                                          output file for dispersal parameters[1]
```

[1] File names can include full addresses (as shown here) or relative adresses. In particular, using simply `AlisierPar.txt` indicates that the file is in the same directory as the executable file (except for Mac OS X 10.5 users for which it indicates that the file is in the user home directory).

[2] These numbers determine how to read the input files and must thus be in accordance with the data contained in the files (see below). If the input is $l$, $kc$, $kq$, $kw$, the program expect to find $kc + kq$ covariates in the data file and will use the $kw$ first quantitative covariates (among the $kq$ available) to compute the weights of the pollen donors by multiplying them (thus they must be positive). If $kw$=0 then all pollen donors have an equal weight.

Presently, the covariates can be used (i) for visualization of the results (see below) or (ii) to weight the pollen donors according to the weighting variables. The statistical analysis implemented yet does not estimate the effects of covariates on fecundity. We will develop this aspect in a further version.

[7] Using the instruction `file_dist` indicates that the distances between mother-trees and candidate fathers are loaded from the file whose name follows. In that case the x-y coordinates loaded from the parental file (e.g. `AlisierPar.txt`) are useless. Using any other instruction (e.g. `nofile_dist`) indicates that distances are computed from the adult x-y coordinates loaded from the parental file. In that case, **NO** file name should follow.

[3] Using the instruction `file` indicates that the allelic frequencies are loaded from the file whose name follows. Using any other instruction (e.g. `nofile`) indicates that allelic frequencies are computed from the adult genotypes. In that case, **NO** file name should follow.

[4] Using the instruction `LN` indicates that a log-normal distribution is used. Using any other instruction (e.g. `Gamma`) indicates that a gamma distribution is used.

[5] Any integer to initialize the random number generator. Two repetitions with the same seed will provide exactly the same results.

[6] The number of burn-in iterations is the number of initial iterations whose results are not stored. After these, ALL the definitive iterations are stored concerning the dispersal parameters and dobs/de. Because there are numerous individual fecundities it is better to store them only once every *S* iterations. *S* is the thinning parameter. Usually, 1000 stored iterations are enough for a good estimation of the individual fecundities.

## The input files.

### *File defining the adult trees*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains all the information about the adult trees: names, genotypes, positions, covariates explaining fecundity (not used yet…), and weighting variables

Example:

```
III-24.01   8  11   1  1  1  6  4  7  4  5  15 15   558476.0   2417501.0    2  1  2  1  1  2  1  8
III-26.01   8  8    3  9  1  3  7  16 4  4  1  15   558494.5   2417532.2    3  2  4  2  1  4  1  19
III-26.02   8  9    1  3  1  17 7  16 1  5  1  9    558493.6   2417529.9    2  2  4  2  1  4  1  13
III-26.03   8  8    1  1  2  6  11 11 5  5  9  15   558512.0   2417405.2    1  2  2  1  1  2  1  15
III-26.04   8  11   3  9  1  6  11 12 5  5  1  20   558489.6   2417358.6    2  2  3  1  1  3  1  11
III-26.05   8  8    3  9  1  6  7  16 4  4  1  15   558486.7   2417360.8    2  2  3  1  1  3  1  14
III-26.06   11 11   3  3  1  6  4  7  1  5  1  7    558484.5   2417359.7    2  2  3  1  1  3  1  18
```

The first column contains the names. Next columns contain the genotypes: 2 successive columns for each locus. The number of locus has been defined in the parameters file (both must be compatible). The 2 next columns contain xy coordinates. The next columns contain class covariates (here 6). The number of class covariates has been defined in the parameters file (both must be compatible). The next columns contain quantitative covariates (here 2). The number of quantitative covariates has been defined in the parameters file (both must be compatible). AMONG these quantitative covariates, the first are used to weight the observations: The *kw* first columns are multiplied to provide a weight to each adult (fixed during the analysis). *kw* has been defined in the parameters file (here 1).

### *File defining the offspring and mother trees*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains all the information about the sampled offspring: names, mother identity, genotypes.

Example:

```
III-26.01.01    III-26.01     8  9  9  1  1  1  7  19  4  5  1   15
III-26.01.02    III-26.01     8  8  3  1  3  17 7  7   4  1  1   1
III-26.01.03    III-26.01     8  9  9  1  1  17 7  7   4  1  1   15
III-26.01.04    III-26.01     8  9  3  9  3  17 7  16  4  5  15  9
...
III-26.02.01    III-26.02     8  8  3  9  1  3  16 8   5  4  1   9
III-26.02.02    III-26.02     8  8  1  3  1  3  7  16  1  4  1   9
...
III-26.07.01    III-26.07     8  4  9  1  1  5  11 -1  1  4  15  6
III-26.07.02    III-26.07     8  11 3  3  3  3  12 -1  1  4  15  20
III-26.07.03    III-26.07     8  11 3  1  3  3  12 3   1  5  11  6
```

The first column contains the name of the offspring. The second column contains the name of the corresponding sampled mother tree. All the sampled mother trees must exist in the parental file defined below. Then pairs of columns contain the genotypes at the diploid locus (same number as defined in the parameters file and same number as parental genotypes).

### *File defining the mother-to-candidate fathers distances (optional)*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains all the pairwise distance between mothers and candidate fathers. It is only necessary if the `file_dist` instruction has been chosen in the parameters file.

Example:

```
JV0786          JV0786       0.000000
JV0786          JV0787       26.236284
JV0786          JV0788       30.608998
JV0786          JV0789       30.608998
JV0786          JV0791       36.792949
```

One row is used for each mother-candidate father pair. First column contains the name of the mother (matches with name from the previous files), second column contains the name of the father (matches with name from the previous files), third column contains the distance. Take care, pairs that are omitted in the file will receive a distance = 0.

### File defining the allelic frequencies *(optional)*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains all the information about the genotyped locus. It is only necessary if the `file` instruction has been chosen in the parameters file.

Example:

```
15
1               0.00483871
2               0.021774194
3               0.068548387
4               0.100806452
5               0.169354839
...
10
1               0.347020934
2               0.153784219
...
```

For each locus, the first row contains the number of alleles. Each next row contains the allele (here 1…15 but it could have been observed numbers of base pairs, e.g. 104, 106, 110, 118…) and the corresponding allelic frequency.

## The output files.

Three output files are generated by the `MCMCGeneration` executable: one that stores the MCMC concerning the dispersal parameters, one for the parameter of the inter-individual variance of fecundity and one for the individual values of the fecundity.

### *File storing the dispersal parameters*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains all the information about the dispersal and mating parameters: log-likelihood, scale and shape parameters of the dispersal kernel, migration and selfing rate

Example:

```
0 -14380.9 100 1 0.5 0.0346765
1 -14326 100 1 0.5 0.0284153
2 -14206.7 142.124 1 0.514403 0.0181131
3 -14142.4 190.129 1 0.514403 0.0175781
4 -14107.8 190.129 1 0.514403 0.0175781
5 -14005.6 252.398 0.785027 0.484737 0.0196833
6 -13986 253.026 0.698175 0.484737 0.0196833
...
```

Each row stores one iteration. The first column contains the iteration number. The second column contains the log-likelihood (eqn 6 in Klein et al. 2008). The third column contains the mean dispersal distance, the 4th contains the shape parameter *b*. The last two columns contain the immigration rate and the selfing rate. The values stored in these columns can be used to compute the posterior distribution for the parameters (posterior mean, median, mode, CI...).

### *File storing the fecundity parameters*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains the information about the variance of fecundity

Example:

```
0 8.87757 2
1 -32.5814 1.94342
2 -39.4542 1.84538
3 -63.4262 1.91922
4 -74.5459 2.02083
5 -66.114 2.02984
6 -69.0617 2.06704
...
```

Each row stores one iteration. The first column contains the iteration number. The second column contains the log-likelihood of the fecundities (log of eqn 2 in Klein et al. 2008). The third column contains the (theoretical) value of $d_{obs}/d_{ep}$ at the given iteration.

### *Storing the individual fecundities*

Name : any name. But it has to be provided in the parameters file.
Location : any location. But it has to be provided in the parameters file.
Function : This file contains the values of the MCMC for all individual fecundities

Example:

```
0 0.405777 1 1.25016 0.766808 0.946726 0.212047 0.164272 3.58104 1.21859 0.510965 0.286309
0.192152 1.48878 0.746375 0.283841 1 1 0.661728 0.198554 0.788516 0.490574 0.595992 1.01406
0.233984 1 1.60853 0.632318 1 1 0.161166 0.654654 1.13373 0.47031 0.846849 1 1 0.420055 0.752797
2.20459 0.475376 0.385241 1.08904 0.696016 0.239139 0.192906 0.589105 0.628054 1 1 0.352205 1 1
3.22132 ...
10 0.243109 4.78112 2.51394 0.38589 0.532809 0.27572 0.298084 3.52793 2.56598 1.3433 0.853494
0.529221 1.86258 1.76757 1.47735 1.92819 0.157908 1.74045 0.450413 0.158796 0.511134 0.726239
0.712561 0.0869671 1.03106 0.947242 1.27981 0.516835 0.500099 0.162142 1.98327 0.328165 0.511691
0.399872 0.0902774 0.248639 1.22777 0.240626 1.44101 0.520809 0.340008 0.108832 0.782097 0.225085
1.61395 1.344 ...
20 0.762449 4.78112 2.29024 0.827465 0.461725 0.813035 0.25065 1.94589 1.47179 0.152508 0.283444
0.484052 0.158014 0.845448 1.69182 1.92819 0.771052 3.03505 1.28364 0.249887 0.341934 0.404059
1.37282 1.20611 1.62568 1.31785 1.03199 0.483658 0.432162 0.575914 0.262516 0.315116 0.723904
0.169238 0.178186 0.419168 0.786211 0.434897 0.233016 0.327239 2.29136 0.35847 1.32304 0.655079
0.154158 0.314548 ...
...
```

Each row stores one iteration after thinning. The first column contains the iteration number. Then, each column corresponds to a given adult tree and the value stored is the fecundity of this tree at that iteration. Because this file contains numerous columns, it is suggested not to store all iterations. This is possible by choosing a thinning parameter that controls the number of steps between two stored vectors of individual fecundities.
This file is used to compute the posterior distributions of the individual fecundities and also to compute the posterior distribution of the empirical ratio dobs/de, based on the empirical variance among individual fecundities.

## The plotting file.

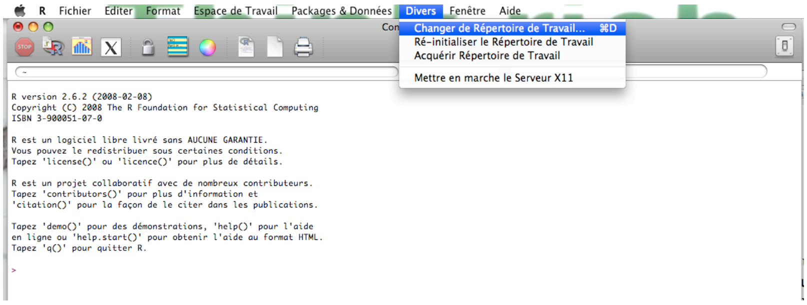
Name : MCMCVisu.R

Location : For an easy use, this file should be in the same directory as the executable file. Users familiar with R can copy it anywhere and use it like any R source file.

Function : This file contains 4 functions that enable to plot the results that are stored in the output file…

Annotated example:

In R, first select the working directory that contains the output files generated by MCMCGeneration and the code file MCMCVisu.R


R
Fichier
Editer
Format
Espace de Travail
Packages & Données
Divers
Fenêtre
Aide
Con
Changer de Répertoire de Travail… ⌘D
Ré-initialiser le Répertoire de Travail
Acquérir Répertoire de Travail
Mettre en marche le Serveur X11
STOP

```
R version 2.6.2 (2008-02-08)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

>
```

```
> source("MCMCVisu.R")
> figure1("ParamFec.txt","ParamDisp.txt")
```

This function represents the MCMC for the different parameters, providing insights about the convergence and mixture of the chains. The two arguments of the function are the names of the output files of the MCMCGeneration step. The 6 sub-plots represent the MCMC for the log-likelihood, the $d_{obs}/d_{ep}$ parameter and the δ, $b$, $m$ and $s$ parameters. x-axis represent the iterations. The mean, median of the posterior distribution and the 95%-credibility intervals are also provided on this figure.

```
> figure2("ParamFec.txt","ParamDisp.txt","IndivFec.txt",184,1,30,1,10,100,1500,0.0,0.8
,0.2,0.8,0,0.05)
```

This function represents the posterior distribution for the different parameters. The arguments of the function are the names of the output files of the MCMCGeneration step, the number of father-plants for which a individual fecundity was computed (here 184), the range of values to plot for the theoretical $d_{obs}/d_{ep}$ (here 1, 30), for the empirical $d_{obs}/d_{ep}$ (here 1, 10), for δ (here 100, 1500), $b$ (here 0.0, 0.8), $m$ (here 0.2, 0.8)and $s$ (here 0, 0.05).

```
> figure3("IndivFec.txt",184)
                   V2         V3         V4           V5          V6           V7           V8 ...
MeanIndiv 0.30054792 4.896937 3.653972 0.20167900 0.2246963 0.19427061 0.09444644
q1MCMC    1.06628000 8.303130 6.532120 0.64071700 0.5913910 0.46567400 0.28607000
q2MCMC    0.00306119 0.747969 0.603115 0.00293561 0.0110131 0.00795552 0.00125505
```

This function computes and represents the posterior means and 95% credibility intervals for all individual fecundities. The minimum arguments of the function are the names of the output file of the MCMCGeneration step and the number of father-plants for which a individual fecundity was computed (here 184).

With an additional argument containing a covariate value for each father-plant (a vector containing integer values), the same function represents the posterior means with different colours (see below).

With an additional argument containing the positions of the father-plants (a two-column matrix), the same function represents the posterior means in a spatial way (see below).

With three additional arguments containing the positions of the father-plants (a two-column matrix), the size of the circles and a vector of covariate values for all father-plants, the same function represents the posterior means in a spatial way with colours (see below).

```
> cov=read.delim("alisierpar.txt",h=FALSE,sep="\t")[,16]
> cov
  [1] 2 3 2 1 2 2 2 2 2 1 2 1 1 2 3 2 4 1 4 1 1 1 3 4 4 ...
> figure3("IndivFec.txt",184,cov)

> xy=read.delim("alisierpar.txt",h=FALSE,sep="\t")[,14:15]
> xy
        V14     V15
1   558476.0 2417501
2   558494.5 2417532
3   558493.6 2417530
4   558512.0 2417405
5   558489.6 2417359
...

> figure4("IndivFec.txt",184,xy)
> figure4("IndivFec.txt",184,xy,2,cov)
```

## What's new ?

### *Version V1.1*

- A bug in version V1.0 was fixed. This bug was leading to excessively low values of the shape parameter *b* and excessively high values of the dispersal distance δ and of the variance of fecundity dobs/de.
- The gamma distribution is now implemented
- It is now possibile to define weighting variables (Be careful, Parameters.txt needs one more information **even when not using weighting variables**)
- It is no more necessary to define a useless file name when allelic frequencies are computed from adult individuals.
- The CI for the parameter 1 was corrected.

### *Version V1.2*

- A bug in the use of weighting variables was fixed. In version V1.0, the weighting variables were not actually modifying the fecundity of individuals.
- Few small bugs in the R output codes were fixed. It includes in particular the wrong confidence interval that was provided for m in Figure 2.
- It is now possible to slightly modify the appearance of the output graphics windows in R (size of the window and size of the axis fonts are now two optional parameters in functions figure1-figure4).

### *Version V1.3*

- The possibility to use non-Euclidian distances, loaded from an additional file has been included. Take care, this modifies the shape of the "Parameters.txt" file, with an extra-line in fourth position! Even when not using this possibility…