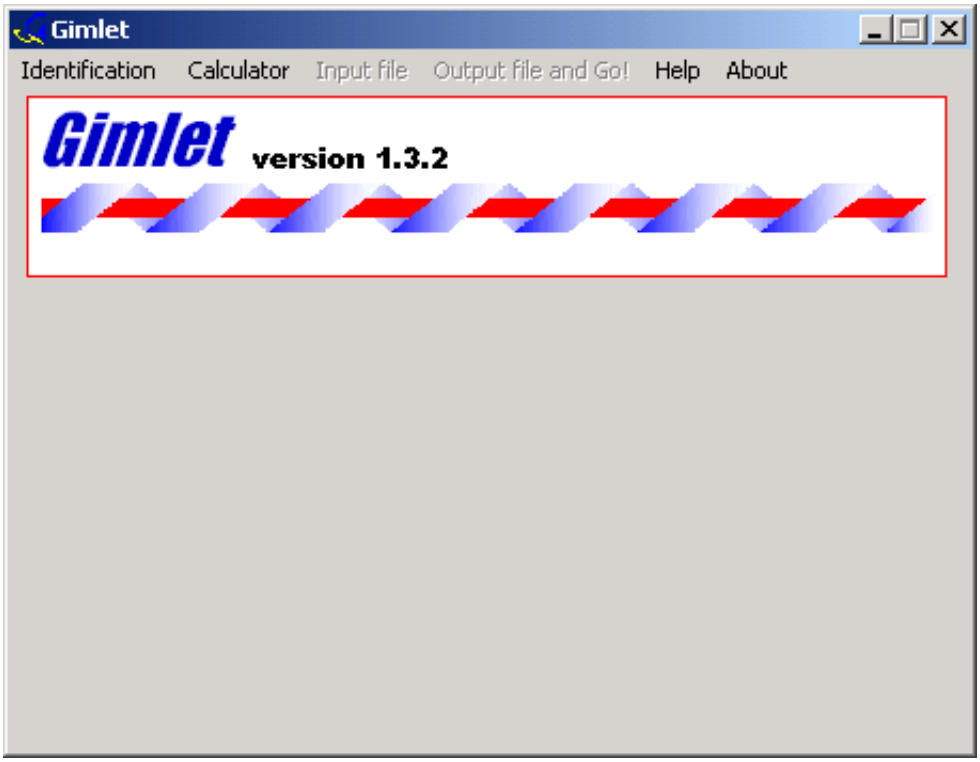
GIMLET v.1.3.2 Guide

Nathaniel Valière *

October 6, 2003



Document prepared with LaTeX 2$_\varepsilon$

*Laboratoire de Biométrie et Biologie Evolutive - UMR5558 - 43, boulevard du 11 novembre 1918 - F69622 Villeurbanne - FRANCE

Contents

1 Presentation

1.1 What is GIMLET and what can we do with?

GIMLET (Genetic Identification with MultiLocus Tags) is a software designed for individual identification using molecular tags in diploid species. In this context, molecular tags are composite genotypes at multiple loci, each composed by two alleles. Thus, N pairs of coding values compose the molecular tag for a sample. Coding values could be for example two-digits-numbers designing alleles (e.g. 01, 02, 03 for first, second and third alleles) or could be allele size (three-digits-number) for electrophoretic sized markers (e.g. 242, 244, 246 for first, second and third alleles).

GIMLET can perform following tasks:

- estimating error rates during genotyping and constructing consensus genotypes from repeated genotyping,
- pooling identical genotypes among a set of several genotypes,
- identifying one (or several) genotype(s) comparing it (them) with references,
- determining kinship between individuals
- estimating several parameters (allelic frequencies, Heterozygosity, probability of identity, population size) from genotyped samples.

1.2 Configuration and installation

1.2.1 Configuration

GIMLET is a $Windows^{TM}$ based software compiled with Visual Basic. You can install GIMLET on any $Windows^{TM}$ compatible operating system. The installation of GIMLET required about 4 Mo on hard disk (about 1 Mo in the folder GIMLET and 3 Mo for system files).
All input and output files are text files and input files must be under GENEPOP format. GIMLET accepts any file extension.

Warning: if you have many genotypes to analyse, some memory troubles can occurred. If a problem during the opening of a big file (in which no format incompatibilities are expected) occurred, check the memory performance of your computer (in this case you can increase the virtual memory for example).

1.2.2 Get it!

GIMLET Software is freely available at the following address:

`http://pbil.univ-lyon1.fr/software/Gimlet/gimlet.htm`

When get GIMLET Software, I strongly encourage users to register with me, sending an email entitled "GIMLET register" at `valiere@biomserv.univ-lyon1.fr`.
This permit to constitute a mailing list and so you will receive all the information and update about the software.
Questions, comments and bugs reports should also directed to me, indicating in the subject of your email "GIMLET bugs" and will be redirect in the mailing list.

1.2.3 Install it!

Package for GIMLET is composed by `Setup.exe` file installing GIMLET Software, and all dependency files needed to run GIMLET.
For installing GIMLET Software on your computer, open the `GimletvXXX.zip`. Then, simply double-click on `Setup.exe` (in the WINZIP window) and follow the instructions. Installation requires about 4 Mo space on your hard disk.
After installation, go on your *StartingMenu/Programs* to launch GIMLET Software (GIMLET icon), or go in the GIMLET folder and double click on the `GIMLET.exe` to run GIMLET Software.

To update a GIMLET version to the lastest released version, just download the `GIMLET.exe` file, available on the website and replace the old file in the program folder.

2 Getting started and Input file

GIMLET is composed by several main menus :

- Identification: for the pooling of genotypes, the identification and the determination of the kinship.
- Calculator: for the construction of consensus genotypes and estimation of error rates, the estimation of genetic parameters (allele frequencies, heterozygosity and probability of identity, matching probabilities) and population size.
- Input file: select the input file containing the genotypes to analyse (see below).
- Output file and Go!: create the output file(s) and launch the analysis.
- Help: if you need help.
- About: information about GIMLET.

The input file required by GIMLET must be under GENEPOP format.

The GENEPOP format must follow these rules:

- first line: any character. Use this line to store information about your data.
- second line: the name of the first locus
- third line: the name of the second locus (if needed)
- etc.
- line N+1: the name of the Nth locus.
- line N+2: type the word "POP, "Pop" or "pop".
- line N+3: identifier (personal use) of the first individual for the first population followed by its genotype. A comma separates identifier and genotype. Identifier could contain any character except a comma.

Although GENEPOP program requires only 2-digits format for alleles, GIMLET accepts also the 3-digit format. Each genotype at each locus (a four or six-digits genotype) is indicating preceded by a space blank.

Example:
```
Ind1, 0101 0103 0306 0404
```

or
`Ind1, 134136 096098 215217 156160`

Indications for GENEPOP format files in GIMLET Software:

- If you use a Excel spreadsheet to construct file, prefer the "space separator" file (extension .prn) instead of the "Tabulation separator".
- You can use pop", Pop" or "POP" to use as separator for different populations (without any space before or after).
- **Replace all tabular by space especially for 'pop' separator**.
- **Do not leave space or free lines after the last genotypes of the last 'pop'**.

3 Identification menu

Identification menu propose four distinct tasks: identification of only one genotype, identification of several genotypes, the pooling of identical genotypes and finally the determination of the kinship.

3.1 Identification of only one genotype

Enter the genotype: Genotype to identify must be formatted as a genotype in a GENEPOP file: the alleles of a locus are collapsed and each locus is separated by a space. The code for an allele could be a number such as 01 or the allele size such as 244 for sized markers. **Be sure that the genotype format and the input file's genotypes have the same digit format**. Missing data must be indicated by value "00" or "000".

114116 149149 094096
or
0103 0404 0203

Input file: The input file (GENEPOP format) must contain the reference genotypes. Several "pop" can be present.

Results: The results are given in a window indicating the genotype's identifier that matches with the multilocus genotype to identify. If no individual match with multilocus genotype, program determines the closest genotype. The closest genotype is the genotype that shares the higher number of loci with the genotype. The associated score of the identification is the number of locus shared by the two genotypes divided by number of analysed loci. If no genotype match for any locus, result window returns that no match has been found.
If several genotypes are identified, results are not indicated in the window and a outfile can be produced, which indicates the details of the comparison between genotypes. Program also indicates the pairs of genotypes where only one allele (for only one or two loci) or two alleles (for only one locus) is (are) different between the genotypes.
User can update the file of reference genotypes with the analysed genotypes if this later does not match (over all locus or not) with any reference genotype.

3.2 Identification of several genotypes

Input files: only one GENEPOP file is required. It must contain the reference genotypes after the first 'pop' AND the genotypes to identify after the following 'pop' (several groups of genotypes to identify could be inserted).

Process: the identification of the genotypes is first conduct using multilocus genotypes. Each genotype is identified when its multilocus genotypes match completely with a reference genotype. However, a locus by locus procedure is performed. For each locus, a genotype is identified after the identification of genotype based on the single locus genotypes. The reference attributed to the multilocus genotype is the reference that have the higher number of single locus identification.

Output file: The output file indicates for each genotype to identify the result of the identification (the identifier of the closest genotype) and the associated score for the identification locus by locus. The score of the identification is the number of locus shared by the two genotypes divided by number of analysed loci.
Program indicates the pairs of genotypes where only one allele (for only one or two loci) or two alleles (for only one locus) is (are) different between the genotypes.

A file (GENEPOP format) could be constructed with the reference genotypes and the unambiguous unique genotypes. Unambiguous genotypes are genotypes for which no reference genotypes could be identified over all loci and for which the score of the locus by locus comparison is below a threshold set by user (if the score is below the threshold, it means that the comparison result has a too low confidence and thus that the genotypes has chance to be unique). Reference file could also be updated with these unambiguous unique genotypes to obtain a new reference file.
If you choose to save unambiguous genotypes in files, you may choose the identifier prefix of the genotypes. For example, if you choose "Gen" for identifier prefix, "Gen1", "Gen2", "Gen3", "Gen4", etc will reference genotypes. Digit format can be either 2-digit or 3-digit format in the case of the input file is 3-digit format.

3.3 Regrouping genotypes

This option is used for pooling several genotypes that match themselves.

Input file: only one input file under GENEPOP format is required. Several groups (separated by 'pop' in GENEPOP file) could be included in the file. Program considers either the different 'pop' as different groups to analyse and all 'pop' as a single group to analyse. Program gives the result for both cases (the output file for the second case is indicated with a 'TOT' suffix).

Process: the regrouping is conducted as an identification (*cf.* Identification section) where all genotypes are potential reference genotypes.

Output file: The output file indicates the different group and the new identifier attributed the analysed genotypes (results based on all loci).
Then the results table is composed by: the identifier of the analysed sample, the identifier of the sample with the best match score, and the score of this match (number of loci identified the match genotype / total number of loci). When there are several samples with same score, and program returns 'ambiguity'.
Program indicates the pairs of genotypes where only one allele (for only one or two loci) or two alleles (for only one locus) is (are) different between the genotypes.

Program constructs the matrix of the pairwise comparison of genotypes. The values in this matrix are the number of loci with same genotype between the two compared samples (below diagonal) and the score (%) of this match (above diagonal).

All distinct groups detected in the input file could be saved into a file in GENEPOP format. You can also choose the identifier prefix of the genotypes. Digit format can be either 2-digit or 3-digit format in the case of the input file is 3-digit format.

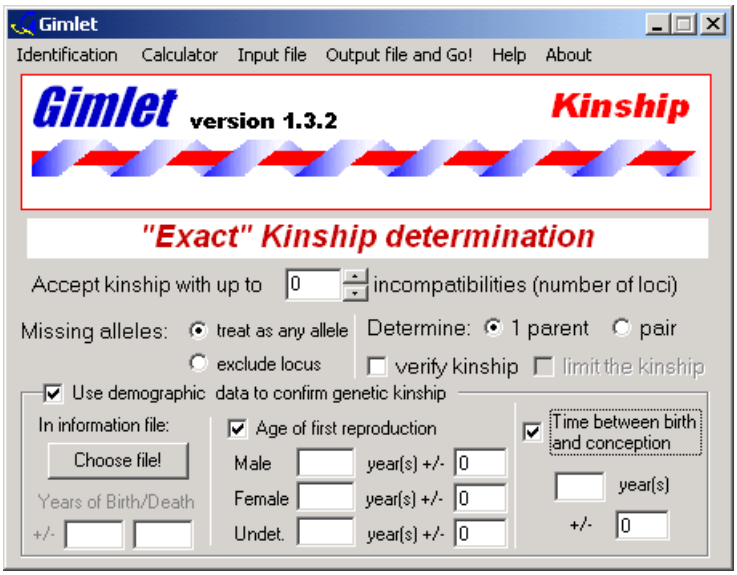


Figure 1: Kinship module

3.4 Kinship

This option is used for determining the kinship between individuals based on the mendelian transmission of alleles from parents to offspring. Other information (age,sex, mean age of first reproduction, gestation time) can be used to confirm or not the genetic determination of the kinship.

Input files: You need one, two or three input file(s). The first file must be under GENEPOP format and contains the genotypes of the offspring and the potential parents. All individuals are saved in the same 'pop' in the Genepop structure.

The second file (optional) contains information that could be used to confirm and precise the kinship between individuals : the class of the individuals (parent, offspring or both), the sex (male, female, undetermined) and the years of birth and death of individuals.
This information file is structured as follows.The first line is the title describing the file. The next lines contain the information about individuals. The individuals must have the same identifier and must be in the same order as in the GENEPOP file containing the genotypes. The data must be organized with one line per individual in this order: identifier, class, sex, year of birth, and year of death (see example next page).

The codes for the class and the sex are given below:

Class	Sex	Years
P=Parent	F=Female	1998= year 1998
O=Offspring	M=Male	0000= unknown year
B=Both	U=Unknown	9999= still alive (death year)

The mean age at first reproduction (or the age of sexual maturity) for male, female and undetermined individual, and the time between the conception and the birth of an individual [1] could also be set in the kinship menu.

[1] this time must include gestation or incubation time and also the possible delayed fecundation or gestation in some species

NB: user must follow several guidelines to correctly assign the years of birth and death, and the mean age of first reproduction and gestation time. For the years of birth, each integer value corresponds to the first birth season of the calendar year. For example, an individual born in 1998 was born during the first birth season of 1998. If there are several birth periods per year, decimal number must be used. For example, an individual born in 1998,5 was born during the birth season corresponding to six months after the first birth season of 1998. An individual born in 1998,25 was born at the birth season corresponding to 3 months (0,25=3/12) after the first birth period of the year 1998. Identically, decimal number could defined the gestation time and the mean age of first reproduction. For example, the individuals of a population with a mean age of first reproduction of 1,5 could copulate 1 year and 6 months after their birth. A gestation time of 2 months is defined by 0,17 (2/12).

For example, the information file could be:

File begins below this line

Example of a file containing information data
Ind1, P, M, 1998 2001
Ind2, O, F, 1996.25 9999
Ind3, B, U, 0000 2000.5

File ends above this line

Ind1 is a potential parent (P), is a male (M), was born in 1998 (first season of birth) and dead in 2001
Ind2 is an offspring (O), is a female (F), was born 3 months after the first birth season of 1996. She is still alive (9999).
Ind3 if both an offspring and a potential parent (B), its sex is undetermined (U). Its birth year is unknown and it dead in 6 months after the first birth season of 2000.

NB: decimal separator must be a point (.) and not a comma(,).

The figure3.4 shows the life cycle and the year system (with a species with several reproduction season per year) used in kinship determination.

The third file (optional) contains the list of individuals for which the determination of the kinship will be limited.
For example, the limitation file could be:

File begins below this line

Example of a file containing name of potential parents for each individual
Ind1, 1 Ind3
Ind2, 0
Ind3, 1 Ind2
Ind4, 2 Ind1, Ind2

File ends above this line

Each individual contained in this file has to be indicated and sorted in the same way than in the file containing the genotypes or information data. The number after the comma is the number of individuals that can be the parent of the individual ("0" if all individuals could be the parent). The identifier of each potential parent follows. Each identifier has to be separated by a

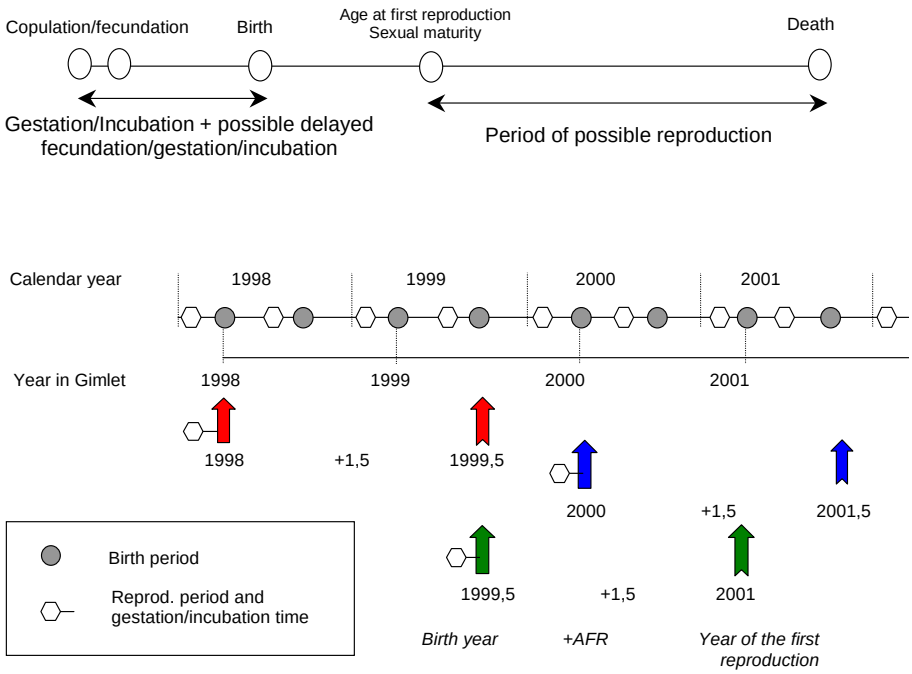


Figure 2: The system of life cycle used in GIMLET. Ex: first individual (red arrows) could be the mother or father of the second individual (blue arrows). However, first individual could not be the mother or the father of the third individual (green arrows) because the conception of this individual is prior to the date of the first reproduction for the first individual.

comma from the previous identifier. **Be carreful to delete all tabulation marks or spaces after the identifier. For example, "Ind1 " will be considered as different than "Ind1".**

If the option "Determine 1 parent" is selected (that is the program will determine the kinship parent after parent for each individual), the program will verify the potential kinship of Ind3 for Ind1 and Ind2 for Ind3.
If the option "Determine pair" is selected (that is the pairs of parent will be determined for each individual), the program will determine all pair of parent that includes Ind3 (for Ind1) and Ind2 (for Ind2).

Process:
Genetic determination of kinship: For each potential offspring, GIMLET inspects the mendelian inheritance of its alleles from two other individuals of the file. An individual is considered as potential parent when at least one allele per locus matches with alleles of the offspring. User can set a maximum number of incompatibilities. This number is the number of locus that could not match between an offspring and a potential parent. User can choose if a missing data (allele that are not determined, indicated by 0 in the input file) must be treat as any other allele (missing allele could take any state at the locus) or if the locus in which missing allele appeared must be exclude of the comparison (in this case, a missing data will be considered as a incompatibility).
When an allele missed in the potential parent or offspring, it is indicated in the output file. The user can choose between determine single parent alone or both single and pair of parents ('Determine 1 parent' and 'Determine pair' options).

Demographic information: demographic data are used to confirm and precise the kinship. An individual is considered as potential parent when the following relations are true:

$[Birth.year(Parent) + Age.first.reprod.] \leq [Birth.year(offsping) - Conception.time]$

AND

$[Birth.year(offsping) - Conception.time] < Death.year(Parent)$

Standard deviations for each parameters can be set. In this case the equations begin:

$[AFR(parent)_{min}] \leq Concept.Date(offspring)_{max}]$

AND

$[Concept.Date(offspring)_{min}] < Deathyear(Parent)_{max}$

where

$AFR(parent)_{min} = Birth.year(Parent) + Age.first.reprod. - sd_{birth} - sd_{AFR}$
$Concept.Date(offspring)_{max} = Birth.year(offsping) - Conception.time + sd_{birth} + sd_{Conception.time}$
$Concept.Date(offspring)_{min} = Birth.year(offsping) - Conception.time - sd_{birth} - sd_{Conception.time}$
$Deathyear(Parent)_{max} = Deathyear(Parent) + sd_{death}$

If the option for determining pair of parent ('Determine pair') was chosen, a pair of parents is validated when their sexes are different.

Output file: Two files are outputted. One file (with suffix "(1P)") indicates the results for the determination of parents (only one parent, not pair of parents) for each offspring. The other file (with suffix "(2P)") indicates the results for the determination of the possible pairs of parents for each offspring.
The level of genetic confirmation of the kinship is indicated by the number of locus that matched between the offspring and the potential parent. For example, "9/12" means that one allele matched between offspring and parent for 9 locus over the 12 analysed loci .
When missing alleles appear in the genotype of a offspring, `(ma o)` is indicated. If missing alleles appear in the genotype of a parent `(ma p)` is indicated. Finally, if missing alleles appear in the genotypes of both offspring and parent, `(ma p+o)` is indicated.

If demographic data are included in the analysis, the confirmation of kinship is indicated for each determination.

- `Confirmed level3`= when all data are set and confirmed the kinship.
- `Confirmed level2`= when death year of the parent was missing
- `Confirmed level1`= when birth year of the parent was missing
- `Not confirmed`= when the birth year of offspring, or both birth and death years of the parent were missing.
- `(*)`= when the sex information is available for both parent. The kinship is indicated by mother/father
- `(na)`= the sex of at least one parent is not available. The order of the individual in the pair does not reflect the sex of the parents.

4 Calculator menu

4.1 Construction of the consensus genotypes and estimation of error rates

When using the multitube approach (Navidi et al. 1992), it is useful to calculate easily the error rates. In our purpose, these errors could be principally allelic dropout (ADO) and the false allele (FA). The former error consists on the detection of only one allele in the case of heterozygous individual. The second type of error is due to the creation (eventually due to the error of the Taq polymerase,) of a new allele, which is amplified and revealed.

A way to avoid these errors is to conduct several PCR amplifications per samples and locus and to construct a consensus genotype (the most likely genotype based on the repeated results) for each sample.

GIMLET allows user (i) to construct consensus genotypes from a set of PCR repetitions for each samples, and (ii) to calculate the error rates comparing the repeated genotypes and the consensus.

GIMLET creates a consensus genotype for each sample and each locus using the genotypes from the PCR repetitions results. Two methods are available.

Threshold method The first one is based on the number of apparition of each allele. An allele is retained in the consensus genotype if its score is above a threshold (set by the user). Score correspond of the number of apparition of its peak in the electrophoregram; for an allele at the homozygote state, the score is 1 and for alleles at the heterozygous state, the score is 1 for each of the allele (see Figure).

Probability method The probability method is based on the calculation of "relative likelihood" to obtain the result overall PCR repetition from each possible consensus genotype. The consensus genotype will be designated as the genotype which has the best probability to generate the serie E.
Probabilities (or likelihoods) for each potential consensus genotypes (G) are estimated using error rates (see types of errors below) and the following formulae for the general case:

$$Pr(E|G) \quad = \quad Pr([N_{A_1A_2} = k_{12}] \cap [N_{A_1A_1} = k_1] \cap [N_{A_2A_2} = k_2] \,|G) \tag{1}$$

where $P(E|G)$ is the probability of obtaining E from a genotype (G) ; $N_{A_1A_2}$ is the number of apparition of the genotype A_1A_2 overall PCR repetition E (respectively for A_1A_1 et A_2A_2).

User have to set several error rates.

GIMLET could be used to estimate these error rates. Detected errors are:

- Allelic dropout: when a heterozygote (from the consensus genotype) is typed as a homozygote (from repeated genotypes).
- False allele: when a homozygote (from the consensus genotype) is typed as a heterozygote (from repeated genotypes).
- Five types of errors (true genotype in the first line; erroneous genotype in the bottom line):

Type I	Type II	Type III	Type IV	Type V
--|-----	--|------	--|--|---	--|--|---	--|---|--
------|-	----|--|-	--|----|-	----|--|-	----|----

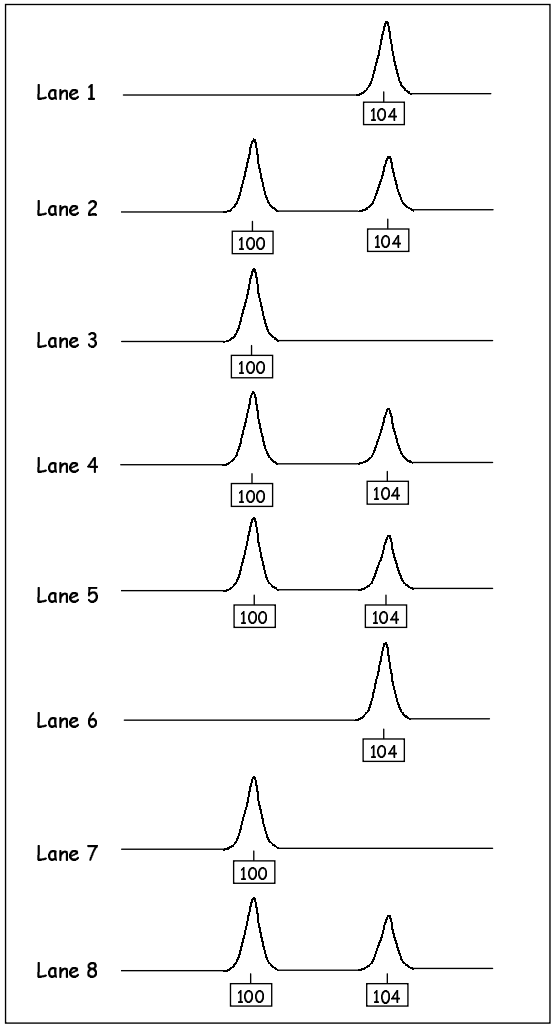


Figure 3: Genotyping is repeated eight times for the same sample and same locus. Lanes 1, 3, 6, 7 illustrate Allelic DropOut. Lane 4 illustrates both Allelic DropOut and False Allele. The score of the smaller allele (100) is 6. The score of the medium (104) is 5, and the score for the higher allele (105) is 1. So, if the threshold is not set (default value = 1), we retain the alleles 100,104 and 105, and there is an ambiguity. If the threshold is upper than 1 (for example 5), we retain the alleles 100 and 104.

Input file: input file must be a GENEPOP format file. Each 'pop' must correspond to a sample. Each genotype for a 'pop' (sample) must correspond to the multilocus genotypes from a single PCR amplification. All items of a 'pop' constitute the repeated genotypes for a sample. User could add a consensus genotype (placed at the last position in the 'pop'). Then user checks the case `"Consensus contained in input file"`. If user wants GIMLET to construct the consensus genotypes for each sample ('pop'), he must check the case `"Construct consensus genotypes"` (in this case the last item will be considered as a repeated genotype).

Example of input file:

3 samples; 3 loci; 5, 7 and 6 repetitions for each; construct consensus option

File begin below this line

```
GENEPOP file format for error estimation
Locus1
Locus2
Locus3
Pop
Spl1, 0303 0407 0205
Spl1, 0303 0707 0202
Spl1, 0403 0707 0205
Spl1, 0303 0707 0505
Spl1, 0303 0707 0205
pop
Spl2, 0303 0101 0106
Spl2, 0000 0101 0101
Spl2, 0304 0102 0106
Spl2, 0304 0101 0101
Spl2, 0000 0101 0101
Spl2, 0304 0404 0106
Spl2, 0304 0101 0106
pop
Spl3, 0101 0507 0106
Spl3, 0000 0507 0106
Spl3, 0506 0407 0205
Spl3, 0000 0507 0106
Spl3, 0101 0507 0106
Spl3, 0101 0507 0106
```

File ends above this line

Output file: output file gives results for each of the type of error. First it gives the mean across loci and the mean across samples. Then it gives the number of missing data ("00" in genotypes) and the percent of positive PCR for each locus (that is the PCR that give a genotype different to "00"). Then file gives the mean percent for each locus (mean for each locus across samples) and mean percent for each sample (mean for each samples over loci). Finally, it gives the number of each type of errors for each pair samples/locus (indicating the homozygote sample at locus by "Hmz" and heterozygote ones by "Htz").

Consensus genotypes determined by any of the two available methods can be saved in a file.

4.2 Estimating genetic parameters

GIMLET could be useful to calculate and estimate several parameters from a set of genotypes: allele frequencies, heterozygosity, probability of identity and matching probability.
Input file: a GENEPOP format file is required. Program calculates parameters for each group in the file ('pop' separator), and also produces results for all 'pop' considered as a single item.

1. Allele frequencies

   Allele frequencies are calculated by counting the number of apparitions for each allele at each locus across all individual genotypes contained in data set.

2. Heterozygotity

   This module calculates also both expected and observed heterozygosity rates from the same files. The expected heterozygosity is calculated from the formula considering HW proportions: $H_{exp} = 1 - \sum p_i^2$. The observed heterozygosity is computed simply by counting the number of heterozygotes in the population or sample.

   The mean of the heterozygosity rates over all loci is calculating as the mean of the rates for each locus (considering independence of loci).

3. Probability of identity

   Probabilty of identity is computed using the allele frequencies in the following formula:

$$PI_{theoric} = \sum_i p_i^4 + \sum_i \sum_{j \neq i} 2 p_i p_j$$

$$PI_{unbiased} = \frac{n^3(2a_2^2 - a_4) - (2n^2(a_3 + 2a_2) + n(9a_2 + 2) - 6}{(n-1)(n-2)(n-3)}$$

$$PI_{sibs} = \frac{1}{4} + \frac{1}{2}\sum_i p_i^2 + \frac{1}{2}\left(\sum_i p_i^2\right)^2 - \frac{1}{4}\sum_i p_i^4$$

   where p_i and p_j are the frequencies of the i^{th} and j^{th} alleles and $a_n = \sum_i p_i^n$.

   The first equation is the equation equation for a population where individuals randomly mate and is given in Paetkau et Strobeck (1994) . The second one is a less biased equation for correcting for small samples of individuals (Kendall et Stewart 1977). The third equation is for a population composed only by sisters-brothers (*sibs*) and is given in Evett et Weir (1998) and Taberlet et Luikart (1999).

   The PI is calculated for each locus (PIrnd/locus, PIunb/locus or PIsibs/locus) and over several loci [Prod(Pirnd), Prod(Piunb) or Prod(Pisib)] by multiplying sequentially the PI value over loci (considering that loci are independant). As PI is the most common statistic used to quantify the power or ability of molecular marker to resolve between two individuals, the locus could be sorted by their PI values. The output file gives the rank for each locus based on PI values.

4. Matching probabilities

   The matching probabilities are those used in Woods et al. (1999). The matching probability is the probability of drawing a given genotype from the population, multiplyied across all loci, considering independance of loci. Three probabilities are calculated: P(random) for radom population, P(par-off) and P(sibs) for probability that the parent or offspring of, respectively, a particular individual or their sibling would have the same observed genotype.

4.3 Estimating population size

Population size could be estimated from genotypes using two methods.

4.3.1 CAPTURE method

When several sampling occasions were done, we could estimate population size using mark-recapture approach where a marked individual is an individual sampled in a previously occasion and defined by a multilocus genotype. In this case, GIMLET builds output files to use as input data in the CAPTURE program. User can choose among the models for estimation of population size proposed by program CAPTURE (Otis et al. 1978).

Input file: one GENEPOP file containing the different capture occasion. Capture occasions are designated by a 'pop' separator. A 'pop'-item contain the genotypes found for samples collected during the sampling occasion.

User must select the models (all models available in CAPTURE program) and the tests to use for the estimation of the population size from the input data.

Process: GIMLET detects each unique genotypes and constructs the capture history of each unique genotype through the different capture occasion.

Output file: output generated by Gimlet contains

- the code lines calling the reading of input data (the capture history for each unique genotype),
- the matrix that records the capture history of each unique genotype
- the code lines calling the estimation of population size using models chosen by the user.

To analyse this output file in CAPTURE, regroup the file to analyse and the `capture.exe` in the same folder. Then, open a DOS windows and change directory until this folder (for example type `cd c:/Progra~1/Capture` to go in the folder `c:/Program Files/Capture/`). Then type `capture i=input o=output`, where `input` is the name of the input data file (created by GEMINI) and `output` is the output where the result will be saved.

4.3.2 The rarefaction curve method

The second method to estimate population size is the rarefaction curve method used by Kohn et al. (1999). These authors estimated the population size as the asymptote of the function between the cumulative number of unique genotype and the number of samples typed (rarefaction curve). In their paper, Kohn et al. (1999) used the equation $y = \frac{ax}{(b+x)}$ to fit the rarefaction curve.
D. Chessel proposed to use another equation $y = a - a\left(1 - \left(\frac{1}{a}\right)\right)^x$ which represents the expectation of the number of full box when we distribute x balls in a boxes. Preliminary simulations data showed that the second equation seems to give lower estimates of the population size than the equation used by Kohn et al. (1999). The second equation gives better results in the case where

the probability of capture is uniform among individuals (NB: this is one assumption of the model) than the equation of KOHN ET AL. (1999), which over-estimate the population size (this bias seems to be reduced by increasing the sampling effort). In the case of heterogeneity of capture probability, the equation proposed by D.Chessel seems to under-estimate the population size (bias reduced by increasing sampling effort), whereas equation used in KOHN ET AL. (1999) give unbiased estimation for medium sampling effort (about two third of the population size) and over-estimation when the sampling effort is high (number of samples >> number of individuals in the population).

Recently, EGGERT ET AL. (2003) have used another equation to estimate the population of elephant. This equation is :
$y = a(1 - e^{bx})$.

Input file: Input file must be formatted as GENEPOP file. In this file, each line is a sample and all groups (separated by 'pop') are either considered as single data sets or grouped and combined as a data set. GIMLET asks you if you want to consider separated 'pop' or regrouped 'pop' or both.
When you choose the rarefaction curve method, a panel is opened for the options of the analysis (Figure 4).

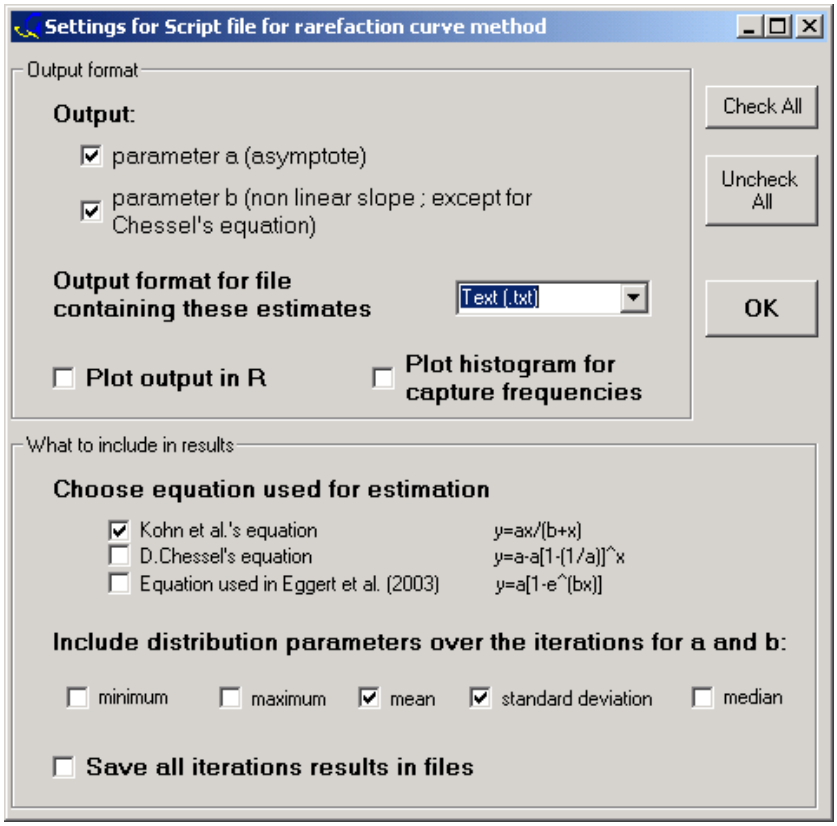


Figure 4: Panel of the options of the rarefaction curve method.

You can choose to estimate either the asymptote a (the estimate of the population size) or the parameter b (non linear slope coefficient; not available for the Chessel's equation). You can choose one, two or the three available equations. You have also to choose what parameter(s) (minimum, maximum, mean, median or standard deviation) you want to describe the distribution of a and b over iterations. You have the possibility to print the results for all the iterations in files (check

`"Save all iterations results in files"`). For convenience, you can choose to add `.txt` extension or `.xls` extension (if you want to automatically open files with Excel software).
If you choose to check `"Plot output in R"`, R will produce graphic showing the relationship between number of analysed feces and number of unique genotypes (see Figure 5; observed relationship for all iterations, mean of the predict rarefaction curves for the chosen equations). If you choose to check `"Plot histogram for capture frequencies"`, R will produce a graphic showing the distribution of the number of capture per samples (number of samples captured 1, 2, 3, ... times). This could be useful to detect heterogeneity of capture probability between samples and thus to choose the best equation (for example, Chessel's equation is more appropriate for data with Poisson distribution of capture probabilities between samples).

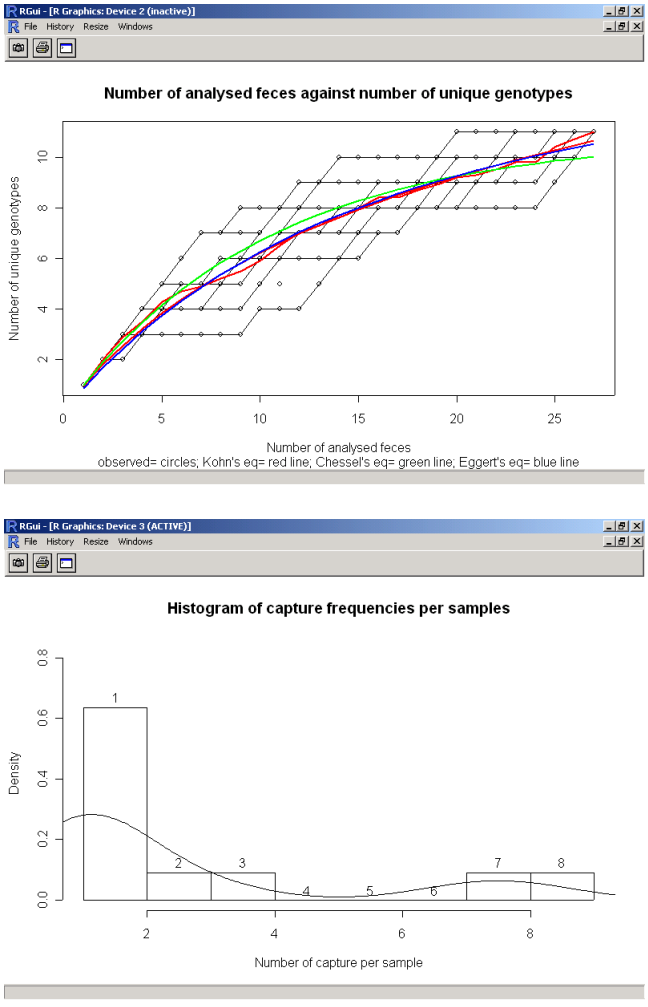


Figure 5: Graphics produced in R when checking `"Plot output in R"` (left) or `"Plot histogram for capture frequencies"` (right)

Output file: Files build by GIMLET regroup the number of samples having the same genotype. These files will be analysed in R program, using a script file automatically wrote by GIMLET. For this, you must transfer (if not already done when you set the pathname of the output files) the script file and all files produced by GIMLET for rarefaction method (default name: `Rarefaction*.txt + script.R`) in the folder of R program (for example `c:/rw1021/bin`). Open

R program by double clicking on the `Rgui.exe` file. In the console, type `source("infile")` where *infile* is the name of the script file (default name: `Script.R`).
This creates several files containing the estimation of asymptote using the equations described above.
As using the rarefaction procedure, the order in which the samples were analysed has an effect on the estimation of the population size (Kohn et al. 1999), the user has to choose (on the request in R program) how many iterations will be run to do the estimation.
As some data will not or will be difficult to analysed (for example when all samples were sampled once or twice each), the estimation will be tried up to 5 times. If no numeric result will returned, "NA" in the output files will be printed. In the result file, the number of iteration, for which a numeric results as been returned, is printed in the first column of the results files.

4.3.3 Other population parameters and tests

Other population parameters can be estimated choosing to create file into different format. You can choose the option `"MARK program"` or `"SURGE program"` which allows creating files containing capture histories of the genotypes into MARK and BIOMECO format. For analysing SURGE files, use the package CR to translate BIOMECO file into SURGE format. Various tests could be applied using GENEPOP program. GIMLET can produce GENEPOP format files in the Identification module.

5 Miscellaneous

Several programs could be used to apply some methods on files created by GEMINI. These programs are described below.

5.1 R package

- Description: R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It can be regarded as an implementation of the S language developed at AT&T Bell Laboratories. Development is now overseen by a 'core team'.
- Availability: free under GNU Public License terms
  (see http:// www.gnu.org/)
- Source: 1 file (24 Mo) at
  http://cran.r-project.org/
- Reference: Ihaka and Gentleman, 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5:299-314.*
- e-Manual: http://cran.r-project.org/

5.2 CAPTURE

- Description: CAPTURE computes estimates of capture probability and population size for closed population capture-recapture data.
- Availability: free
- Source: 1 file of 264 Ko at
  ftp://ftp.cnr.colostate.edu/pub/capture/capture.zip
- Reference: Otis, Burnham, White and Anderson, 1978. Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs 62*

- Manual: White, Burnham, Otis and Anderson, 1978, User's manual for Program CAPTURE, Utah Univ. Press, Logan, Utah ; Rexstad and Burnham, 1991, User's guide for interactive program CAPTURE.Colorado Cooperative Fish & Wildlife Research Unit, Colorado State University, Fort Collins, Colorado.

5.3 SURGE and CR

- Description: SURGE (SURvival Generalized Estimation) is a flexible program for modeling survival and capture rates from capture-mark-recapture or resighting data. A large variety of models can be fitted using SURGE. CR is a menu-based system for treating capture-recapture data on PC-compatible computers.
- Availability: free
- Source (CR including SURGE): 3 files (1,9 Mo) at

  ftp://ftp.cefe.cnrs-mop.fr/biom/CR1.5
- Manual: Pradel et Lebreton (1993) User Manual for Program SURGE, version 4.3, CEFE/CNRS, Montpellier. Cezilly, Pradel, Viallefont et Lebreton (1992). Working with CR: a guide with examples, C.E.F.E, C.N.R.S., Montpellier. 52 pages.

5.4 MARK

- Description: MARK computes parameter estimates (via Maximum Likelihood techniques) from marked animals when they are re-encountered at a later time. Sixteen different parameterisations of encounter data are provided in Program MARK.
- Availability: free
- Source: 1 file of 8 881 Ko at

  ftp://ftp.cnr.colostate.edu/pub/mark/cdrom/disk1/setup.exe
- Reference: White and Burhnam, 1999. Program MARK: survival estimation from populations of marked animals. *Bird Study 46 Supplement, 120-138*
- e-Manual:

  http://canuck.dnr.cornell.edu/mark/

  http://www.cnr.colostate.edu/class_info/fw663/Mark.html

5.5 GENEPOP

- Description: GENEPOP is a population genetic software package, able to perform three major tasks: exact tests (HW equilibrium, population differentiation, genotypic disequilibrium among pair of loci), computation of classical population parameters (Fst and correlations, allele frequencies,) and conversion to formats used by other programs (Biosys, Fstat, Linkdos).
- Availability: free
- Source: 1 file of 1 884 Ko at:

  Anonymous login to ftp.cefe.cnrs-mop.fr (pub/pc/msdos/genepop)

  Anonymous login to isem.isem.univ-montp2.fr (pub/pc/genepop)

  http://cefe.cnrs-mop.fr/
- Reference: Raymond and Rousset, 1995. GENEPOP (version 1.2): population genetics software for exacts tests and ecumenicism. *Journal of Heredity, 86:246-249*

5.6 WINZIP

- Description: WINZIP is a useful compression/decompression tool for Windows systems. WINZIP is the way to easily handle compressed ZIP files.
- Availability : free for the evaluation version
- Source: 1 file of 950 Ko (for Version 7.0) at:

  http://winzip.com

6 How to cite GIMLET?

Paper about GIMLET has been published in *Molecular Ecology Notes* .

Valière, N. GIMLET: a computer program for analysing genetic individual identification data. Molecular Ecology Notes (2002) 2:377-379.

References

Eggert, L., Eggert, J., et Woodruff, D. (2003). Estimating population sizes for elusive animals: the forest elephants of kakum national park, ghana. *Molecular Ecology*, 12: 1389–1402.

Evett, I. et Weir, B. (1998). *Interpreting DNA evidence: Statistical Genetics for Forensic Scientists.* Sinauer Associates, Sunderland.

Kendall, M. et Stewart, A. (1977). *The advanced Theory of statistics*, volume 1. Macmillan, New York.

Kohn, M., York, E., Kamradt, D., Haught, G., Sauvajot, R., et Wayne, R. (1999). Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London, Serie B*, 266: 657–663.

Navidi, W., Arnheim, N., et Waterman, M. (1992). A multitubes approach for accurate genotyping of very small dna sample using pcr: statistical considerations. *American Journal of Human Genetics*, 50: 347–359.

Otis, D., Burnham, K., White, G., et Anderson, D. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62.

Paetkau, D. et Strobeck, C. (1994). Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology*, 3: 489–495.

Taberlet, P. et Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society*, 68: 41–55.

Woods, J., Paetkau, D., Lewis, D., McLellan, B., Proctor, M., et Strobeck, C. (1999). Genetic tagging of free-ranging black and brown bears. *Wildlife Society Bulletin*, 27(3): 616–627.