2012-05-09

MLGsim2.0
User manual

Aniek B.F Ivens, Marten Van de Sanden, Joke Bakker
Theoretical Biology Group,
University of Groningen

A.B.F.Ivens@rug.nl

Dear user,

Thank you for your interest in using MLGsim 2.0. MLGsim 2.0 is an updated version of MLGsim (Stenberg *et al.* 2003). Briefly, compared to MLGsim, MLGsim 2.0 now includes:

- direct reading of (microsatellite) data files
- possibility to simulate populations with more than 200 samples
- fix of the bug encountered during longer simulations (memory leak)
- extended output file with several new estimates and data sorted into MLGs

The program is currently still under construction and, despite being extensively tested, we cannot completely ensure that it works yet with every data set. Every user feedback is welcome. All suggestions and datasets provided for further testing will help improving MLGsim even further. Also, this manual is just a first draft. For more detailed information on the simulations and implemented estimates, we refer to the user manual belonging to Stenberg *et al.* 2003.

In any publications resulting from analyses with MLGsim2.0, please refer to this our url where you downloaded this manual.

Thank you very much for using MLGsim 2.0 and thank you in advance for your comments,

Kind regard,
On behalf of all authors,
Aniek Ivens

1. DATA ENTRY

Data can be entered by providing a .txt file containing allele sizes per locus for every individual. An example file is provided (EXAMPLE_DATA.txt).
The first seven lines give the following parameters:

TITLE=	type here the name of the dataset. This will be included in the name of the output file
SIMULATIONS=	number of simulations to run (default = 1000)
PLOIDY=	can be DIPLOID or HAPLOID (default = DIPLOID)
FREQUENCY=	can be SAMPLE or MLG (default = SAMPLE). This parameter lets the user decide how allele frequencies are calculated. When SAMPLE is chosen, population wide allele frequencies are calculated, when MLG is chosen, allele frequencies are based on a subset of the data, with every MLG only included once. This method avoids over-estimation of rare alleles/MLGs.
MODEL=	can be HWE or FIS (default = HWE). This parameter sets the assumptions under which P_{sex} is calculated. HWE assumes a random mating population in Hardy-Weinberg equilibrium. FIS gives a more conservative P_{sex} estimation, taking into account departures from Hardy-Weinberg equilibrium in the population following Arnaud-Haond et al., 2007.
LOCUSCOUNT=	number of markers used
LOCUSNAMES=	all marker names as given as in the column headers in the file, separated by commas, not followed by a space.

After these seven lines follows a blank line, followed by the table with the alleles given for each sampled individual, just like this example:

TITLE=EXAMPLE
SIMULATIONS=1000
PLOIDY=DIPLOID
FREQUENCY=SAMPLE
MODEL=HWE
LOCUSCOUNT=6
LOCUSNAMES=tu10,tu2,tu4,tu1,tu3

Sample	tu10-1	tu10-2	tu2-1	tu2-2	tu4-1	tu4-2	tu1-1	tu1-2	tu3-1	tu3-2
S28Q1	233	233	159	159	186	198	241	241	194	122
28Q3	233	233	159	159	186	198	241	241	194	194
S42B1	233	239	157	157	182	194	218	218	194	198

Remember to add -1 and -2 behind the two column headers belonging to one locus in case of a diploid species.
NB make sure there is no 'enter' at the end of the table and all entries are separated by tabbed spaces only.

2. PROGRAM RUN

The MLGsim.exe file can be run together with the input table, as long as the two files are in the same folder. Make sure that this folder is on a 'real' drive on your computer and not a 'virtual' drive.

The MLGsim.exe file *cannot* be run by double-clicking on it. Instead, you can run it using **command prompt** (open this by typing 'cmd' in the search/start field below the windows start menu. You should now see the logo of command prompt (black rectangle with C:\ on it), which you can double click).

To run MLGsim in command prompt:

(1) Go to the correct folder by typing 'cd' followed by the path of the folder the .exe file and the datafile are in and press enter. For example:
cd myfolder/analysis/MLG/
(2) Type: MLGsim.exe followed by the name of the datafile. For example:
MLGsim.exe EXAMPLE_DATA.txt

The program should run now (it outputs: *'reading file ….'*) and the result files can afterwards be found in the same folder as your data and program file.

In contrast to the MLGsim, MLGsim2.0 will now run simulations automatically. It will calculate allele frequencies from the dataset provided, count the number of individuals (samples) and then run simulations accordingly, just like you were used to in the previous version of MLGsim.

Briefly, the program sorts the individuals in identical Multilocus Genotypes (MLGs) and counts the number of times these MLGs occurred in the dataset. Based on this, for every MLG the *Psex* value is calculated (see Stenberg *et al.* 2003 for more detail).
Then, the program simulates 1000 populations (1000 is the default) based on the allele frequencies in the given dataset and based on the sample size of the dataset. These simulations provide a distribution of simulated *Psex* values against which the actual *Psex* values can be statistically tested. The result of this test is a Pvalue that gives information on how statistically significantly low the found *Psex* value

is (and thus, how high the chance is we are dealing with an individual that was asexually produced).

3. RESULTS OUTPUT

The program provides two output file: the 'fulltable_NAME.csv' and 'simresults_NAME.csv' files. The two files are comma-separated and can be read into Microsoft Excel.

The 'fulltable'-file, provides the input table, but now re-ordered by MLG. In addition, three more columns have been added: a column that gives the correct MLGname for each individual and the *Psex* and its statistical Pvalue.

The 'simresults'-file gives more detailed outcomes of the simulations and further relevant estimates, listed below:

- MLGtable: this table includes information per MLG:
    - MLGname (MLG)
    - Number of occurrences (n)
    - *Psex*, its Pvalue, and the significance level (ranging from -, *, ** or ***)
    - *Pgen* (part of the *Psex* calculation)
    - The exact allelic composition per locus
- Allelic richness table: estimates per locus
    - Number of alleles (Na)
    - Expected heterozygosity (He)
    - Observed heterozygosity (Ho)
    - Fis value for each locus (Fis)
- Clonal richness table: several estimates of clonal richness
    - Number of clones (G)
    - Number of individuals (N)
    - Genotypic diversity (R=(G-1)/N-1)) (reviewed in Arnaud-Haond *et al.* 2007)
    - Clonal diversity index (*Pd*= G/N) (reviewed in Arnaud-Haond *et al.* 2007)
    - Number of MLGs that occurred more than once (clonal MLGs)
    - Number of MLGs that occurred only once (single MLGs)
- Simulation table
    - The *Psex* values that were closest to significance levels 0.1, 0.05, 0.001 and the exact Pvalues they had. (see also the previous MLGsim manual for more detail)
    - Total number of individuals (Samplesize)

    - Total number of simulated Pvalues (PValues)
- Allele frequency table: alleles and their frequencies per locus
- *Psex* values: all simulated *Psex* values that make up the distribution based on which the statistical Pvalues are calculated.

5. REFERENCES

Arnaud-Haond *et al.*, Standardizing methods to address clonality in population studies, Molecular Ecology, 2007, **16**, 5115-5139.
Stenberg P. *et al.*, MLGsim: a program for detecting clones using a simulation approach, Molecular Ecology Notes, 2003, **3**, 329-331.

MLGsim2.0 is currently under development and will be soon available for download. When publishing results obtained with MLGsim2.0, please contact the authors, so we can provide you with the website details and/or the publication.