Maternal Genealogy Lineage Analyser - MaGelLAn 1.0.

Authors: Ino Čurik and Strahil Ristov

Software for processing and analysis of the pedigree data with emphasis on the maternal lineage and mitochondrial DNA information. In version 1.0 MaGelLAn is a suite of four Python scripts (modules): *mag_verif*, *mag_stat*, *mag_calc* and *mag_sampl*. A detailed description of MaGelLAn modules functionality can be found in the paper:

MaGelLAn 1.0: A software to facilitate quantitative and population genetic analysis of maternal inheritance combining molecular and pedigree information
S. Ristov, V. Brajković, V. Cubric Curik, I. Michieli, I. Curik,
Genetics Selection Evolution.2016, 48:65
DOI: 10.1186/s12711-016-0242-9
URL: http://www.gsejournal.org/content/48/1/65

General notes

Each module performs the initial verification of pedigree correctness. In case of errors in the pedigree, module *mag_verif* outputs the description of errors, while other modules abort the execution and direct the user to module *mag_verif*.

Input format

Magellan modules accept input file name as an argument on a command line. If this argument is not present the default input file name is **pdg_in.csv**. The input file must be in CSV (comma-separated values) format. The first line is a header that includes five mandatory keywords:

ID	(name tag of an individual in the pedigree)
father	
mother	
YOB	(year of birth in single number format)
gender	('1' for male, '2' for female)

and two optional keywords:

haplotype	(name tag of a haplotype; if present, haplotype data enables full functionality of all modules, however, modules are partially functional without it)
available	('1' for available, anything else for not available; exclusive use in module *mag_sampl*)

The place of keywords in the header line must match the column with the corresponding information in the pedigree file. If a keyword is present but the corresponding column is missing the program crashes.

An example of valid first lines in the input pedigree file is:

```
anything,ID,mother,father,gender,anything,YOB,haplotype,available,anything,
1,Id100,0,0,1,anything,2000,,,anything,
2,Id101,Id102,Id100,2,anything,2005,hap1,1,,
3,Id102,0,0,2,,2001,hap1,0,anything,
```

The same example in a "clean" form (without unused columns and with explicit '0' for non-available individuals):

```
ID,father,mother,YOB,gender,haplotype,available,
Id100,0,0,2000,1,,0,
Id101,Id100,Id102,2005,2,hap1,1,
Id102,0,0,2001,2,hap1,0,
```

Additional input files are:

reference_years.txt - stores the first and the last year of birth for the individuals included in the reference population; formatted as two numbers in two lines; used in modules *mag_ stat, mag_calc* and *mag_sampl*

planned_number_of_sequencings.txt - stores the number of planned sequencings; formatted as one number in a single line; exclusive use in module *mag_sampl*

If any of the additional files is missing, the default values (coded in the scripts) are used.

Optional output file:

autocorrection_log.txt

This file is created in all modules in case when one or more individuals in the pedigree have a non-empty mother or a father field, but when that ancestor is not present with an individual record in the pedigree. In cases when the same ancestor occurs twice or more in such a situation, a new record is created. If she or he is mentioned only once, the ancestor data is treated as an empty field. This action is performed automatically and does not affect the operation of the modules. If this occurs, the corrected records are documented in the above file.

Module mag_verif

Finds inconsistencies in female haplotype line. Calculates haplotype error rates. Performs the verification of a pedigree and lists found errors. Initially performs the search for standard pedigree errors and, if found, reports the errors one at a time in ERROR_ALERT.TXT file. After the first error is reported, the program stops.

Output files are:

OutputVerif_Summary.txt - displays the statistics related with pedigree conflicts and HC, IC and MISPLACED indices

OutputVerif_ConflictingIndividuals.txt - lists all conflicting individuals with their haplotype and number of conflicts

OutputVerif_MisplacedBranches.txt - lists haplotyped individuals belonging to misplaced branches, if the misplaced branches exist in a pedigree

and, optionally, if a fatal error is found:

ERROR_ALERT.TXT

This output file reports the first encountered fatal error in the pedigree file. Fatal errors are circles and gender inconsistencies. Since it is neither possible to calculate the effective population sizes, nor to suggest the proper individuals for sampling if such errors exist in a pedigre, they must be corrected before using other modules. If a fatal error is encountered in any of the other modules the program aborts execution.

Module mag_stat

Bookkeeping module. Outputs various useful distributions of individuals over maternal pedigree lines. In particular, lists all individuals belonging to a given founder dam line with the corresponding haplotype. Requires an input pedigree without conflicts in maternal haplotype lines.

Output files are:

OutputStat_DamLineMembership_1.txt - lists the assignment to founder female line for all individuals in a pedigree, effectively imputing the known haplotypes to each individual in pedigree; the format is explicit and easily readable

OutputStat_DamLineMembership_2.txt - the same as above, but the format is adjusted to facilitate further processing with spreadsheet programs; less readable

OutputStat_DamLineMembershipAllInRefPop.txt – lists the assignment to founder female line for all individuals in reference population; these individuals are included in haplotype line effective population size calculation

OutputStat_DamLineMembershipFemaleOnlyInRefPop.txt - lists the assignment to founder female line for only female individuals in reference population; these individuals are included in maternal line effective population size calculation

OutputStat_DamLinesWithFemalesInRefPop.txt – lists the details of a female line representation for lines that have female descendants in reference population

OutputStat_DamLinesWithOnlyMalesInRefPop.txt - lists the details of a female line representation for lines that have only male descendants in reference population; such lines carry the information about haplotype but are lost in the maternal lineage

Module mag_calc

Calculates the effective population size separately for founder dam lines, founder haplotype lines, and founder sire lines. Requires an input pedigree without conflicts in maternal haplotype lines. A meaningfull haplotype line N_e can be computed only if enough of the individuals are sampled.

Output file is:

OutputCalc_InputAndResults.txt

Together with the results of calculations, this file lists the details of pedigree maternal lineage statistics, including all quantities used in the calculations. The information about the number of founder dam lines in reference population with only one sample is important since there is no possibility of verification of the obtained haplotype. Such lines may be treated differently in the further analysis.

Module mag_sampl

Accepts the optional "available" column with ('1' / anything) values. Calculates the target number of the individuals per dam line for sequencing within given planned number of sequencing. Restricts the target numbers to available individuals if *available* data is present. Selects the candidates that provide the highest potential for haplotype diversity within dam line. Requires an input pedigree without conflicts in maternal haplotype lines.

Output files are:

OutputSampl_IndividualsForSampling.txt – for each founder dam line lists the selection of individuals for sequencing that would cover the highest diversity in pedigree; if the intended number of sequencings *N* is less than the number of lines only one (central***) individual is selected from the first *N* lines with the largest representation in reference population; if *available* data is present the choice of individuals is restricted to the available individuals

OutputSampl_DetailedInfo.txt – presents the details of calculation; for each dam line lists the number of individuals in the reference population, and the breakdown of the target number of samples in the numbers of the previously sampled and the samplings that remain to be done; the target number per line is obtained from the planned number of samplings as a factor proportional to the line size

and, optionally, if *available* data is present:

OutputSampl_AvailabilityRestrictions.txt

This file lists the restrictions imposed if the available number of individuals per line is smaller than the target number. AvailableAllCount inncludes available and already sampled individuals. AvailableRealCount is the number of available non-sampled individuals. DIFF is the difference between the target number of samplings and the number of available individuals per line. If there are more available individuals than the target number per line, the reamaining number of available individuals is reported.

Software availability

MaGelLAn 1.0. source codes, pedigree examples and this help file can be downloaded from:

http://lissp.irb.hr/software/magellan-1-0/ and https://github.com/sristov/magellan

Changing the *haplotype* and *available* keywords in header lines of the example pedigree files into anything else will exclude the data in the corresponding pedigree columns. This can be used to test different functionalities of the software.