Laboratory of Recombination and Segregation analysis
Institute of Cytology and Genetics SD RAS
Novosibirsk, 630090, Russia

MGA-SIMULATE

documentation

Please address correspondance to: Yurii Aulchenko, Institute of Cytotlogy and Genetics, Novosibirsk, 630090, Russia. e-mail: yurii@bionet.nsc.ru

1 GENERAL INFORMATION:

MGA-SIMULATE Copyright 2000 Institute of Cytology and Genetics. Portions may be copirighted by other sources and used by permission.

2 PURPOSE:

Simulation of genotypes and quantitative phenotypes in pedigrees given the structure of pedigree, genetic map for loci, populational allelic frequencies, and correspondance genotype — quantitative trait (QT). Pedigrees' founders may come from 2 populations (breeds) differing in distribution of allelic frequencies.

3 AUTHORS:

Y. Aulchenko. The package of programs uses portions of the code developed by T. I. Axenovich and I. V. Zorkoltseva. MGA-UTIL procedures (http://mga.bionet.nsc.ru/) are also used. `randoml.f` is

```
C         Subtract-and-borrow random number generator proposed by
C         Marsaglia and Zaman, implemented by F. James with the name
C         RCARRY in 1991, and later improved by Martin Luescher
C         in 1993 to produce "Luxury Pseudorandom Numbers".
C     Fortran 77 coded by F. James, 1993
C
C       references:
C  M. Luscher, Computer Physics Communications  79 (1994) 100
C  F. James, Computer Physics Communications 79 (1994) 111
```

4 ORIGINAL COMPILERS:

The package of programs was originally compiled using GNU Fortran, C and C++ compilers. For any UNIX-like platforms, use GNU products. For MS-DOS, use products from DJGPP project. Programs are written in standart Fortran/C, so many other compilers would compile them well. For example, Borland C 3.11 compiles all modules quite well. However, some fortran compilers for DOS/Windows might have some problems due to 'iargc' procedure and 'getarg' subroutine. Just modify the source.

5 ORIGINAL PLATFORM:

The package was origionally compiled and tested on Red Hat Linux 6.0. The binaries for this platform are provided in bin/ directory of the distribution.

6 USAGE:

MGA-SIMULATE is a set of about 9 programs, redirecting output from one to another. Therefore the best way to use it is involving a script. The basic script is mga-simulate.sh (bash script). The structure of the script is:

```
# INPUT FILES
```

```
PEDNAME0="testped.dat"     # Pedigree structure data file
MAP="chrom.cnt" # genetic map and allelic frequencies data file
MODEL="qt-model.dat" # mode of inheritance for QT


#OUTPUT FILES (SOLAR FORMAT)

PEDOUT="pedigree.sol" # Pedigree structure.
MAPOUT="map.sol" # Genetic map
PHENOUT="phenotypes.sol" # Simulated phenotypes
MARKOUT="markers.sol" # Simulated marker data

# Main module

BD="./bin"
$BD/renum $PEDNAME0 ped1.wrk
$BD/checkped ped1.wrk
$BD/modelgty $MAP ped1.wrk gnt.wrk
$BD/extract gnt.wrk mark.wrk qtl.wrk
$BD/modelpht qtl.wrk $MODEL phen.wrk
$BD/Sc-ped $PEDNAME0 $PEDOUT
$BD/Sc-mark ped1.wrk mark.wrk $MARKOUT
$BD/Sc-phen ped1.wrk phen.wrk $PHENOUT
$BD/Sc-map $MAP $MAPOUT

# Removing working files

rm *.wrk
rm *.wrk_tmp
```

The 'main module' section involves 9 programs which MGA-SIMULATE consist of. For each personal program, cunsult documentation included in src/¡program-name¿/¡program-name¿-doc.txt

For this script, files with 'wrk' extension are files intermediate between input and output.

7 INPUT FILES

The input for MGA-SIMULATE consist of 3 files: pedigree structure data, genetic map/allelic frequencies data and mode of inheritance of QT.

7.1 PEDIGREE FILE:

First line — any

Second line — number of data lines

Third line — 1st line of pedigree data

pedigree data: id fa mo pop sex

id — personal identification number (id). Must be integer and unique in the file.

fa — father's id

mo — mather's id

pop — ID for population a person comes from. If the parents are available, set to 0. If the person is a founder, "pop" equals to 1 or 2. For pedigrees coming from a single population, just set "pop" to 1 for all founders.

sex — gender. 1 if male, 2 if female.

The file is read by C program, and fields should be separated by at least one space.

7.2 MAP/ALLELIC FREQUENCIES FILE:

lines 1-8 — a comment

line 9 — fortran format for reading "number of loci" variable

line 10 — fortran format for reading periodic data line 1

line 11 — fortran format for reading periodic data line 2

line 12 — a comment

line 13 — number of loci

line 14 and 15 — record for locus 1

Record for a locus:

```
locus name         _ N ofalleles
  |      position |    / --- freq. of allele 1 in pop. 1
  |      |        |   |   / --- freq. of allele 2 in pop. 1
d1mar17  33.2     4   0.01 0.00 0.98 --- freq. of allele 3 in pop. 1
                      0.50 0.44 0.02 --- freq. of allele 3 in pop. 2
                        |
                         \ --- freq. of allele 1 in pop. 2
```

The file uses fortran fixed width format, so the positions and format of every field is very restrictive.

Note, that if a name for a loci contain 'qtl' substring, this locus is considered as QTL. These names should be mentioned in the mode of inheritance file also, providing a rule to construct quantitative phenotype.

7.3 MODE OF INHERITANCE FILE:

Lines 1-8 — a comment

Line 9 — number of QT loci [(i5) fortran fixed width format]

Line 10 — additive effect of 1st locus.

Line 11 — additive effect of 2nd locus...

...

Last line — square root of enviromental variance

Effect of some locus is written in (a12,i5,4f9.3) format fields are:

```
 / --- locus name
|              / --- number if alleles
minor-qtl1     2   -0.100    0.000    0.000    0.100
                     |        |        |         \ --- mean for genotype 2 2
                     |        |         \ --- mean for genotype 2 1
```

```
|          \ --- mean for genotype 1 2
\ --- mean for genotype 1 1
```

The file uses fortran fixed width format, so the positions and format of every field is very restrictive.

8 OUTPUT FILES:

These are 4 files containing genetic map, pedigree structure, marker phenotypes and quantitative phenotypes. The output files have a format ready to be analysied by SOLAR package of programs. See SOLAR documentation for more details.

PEDIGREE FILE: first line — fields in pedigree, separated by comma: id,fa,mo,sex. id — personal identification number (id). Unique intger. fa — father's id mo — mather's id sex — gender. 1 if male, 2 if female.

MAP: first line: "10" (chromosome number). Then each line consist of

(marker-name) (position from the beginning of chromosome)

MARKER PHENOTYPES: first line: id,locus-name-1, locus-name-2, ... second line: marker genotypes for 1st person. Genotypes on different loci are separated by comma, while alleles of the same genotype are separated by slash.

QT PHENOTYPES: Essentially the same as MARKER PHENOTYPES file format.

9 STRUCTURE OF THE PACKAGE:

The package consist of 9 programs: Sc-map, Sc-mark, Sc-ped, Sc-phen, calcbl, checkped, extract, modelgty, modelpht and renum. Programs renum and checkped are coming from MGA-UTIL package, while others were specially developed for MGA-SIMULATE. modelgty and modelpht are core programs, simulating genotypes and phenotypes. Extract and Sc-util are utility programs, the latter converting files drom MGA-format to SOLAR format.

The structure of the package is summarised on the Figure. Read documentation for individual programs for more details.

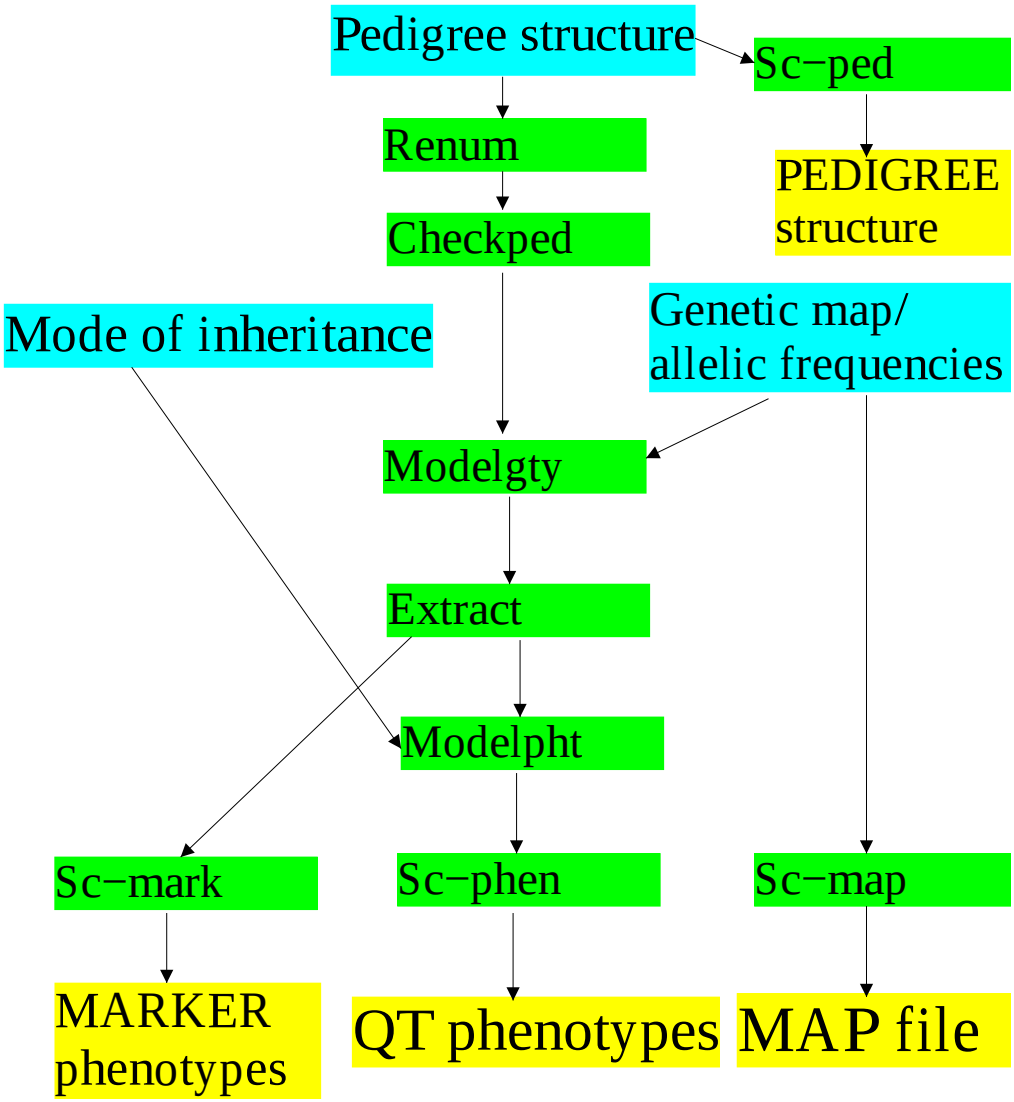

Figure 1: Structure of MGA-SIMULATE package of programms. Green: program modules. Cyan: input data files. Yellow: output files.