Task 1.4.

Gene2farm Simulation

GenoSim software: simulation multiple populations

Latest update:

Jan 2014

Original version:
Hossein Jorjani. [Jorjani, 2009 - A general genomics simulation program. Interbull: http://www-interbull.slu.se/ojs/index.php/ib/article/view/1113/1104)

Modified (for Gene2Farm purposes):
Ezequiel L. Nicolazzi (FPTP)
Luis Varona (UNIZAR)
Clara Diaz (INIA)

For feedback/comments/questions/bug-report, please write to:
ezequiel [dot] nicolazzi [at] tecnoparco [dot] org

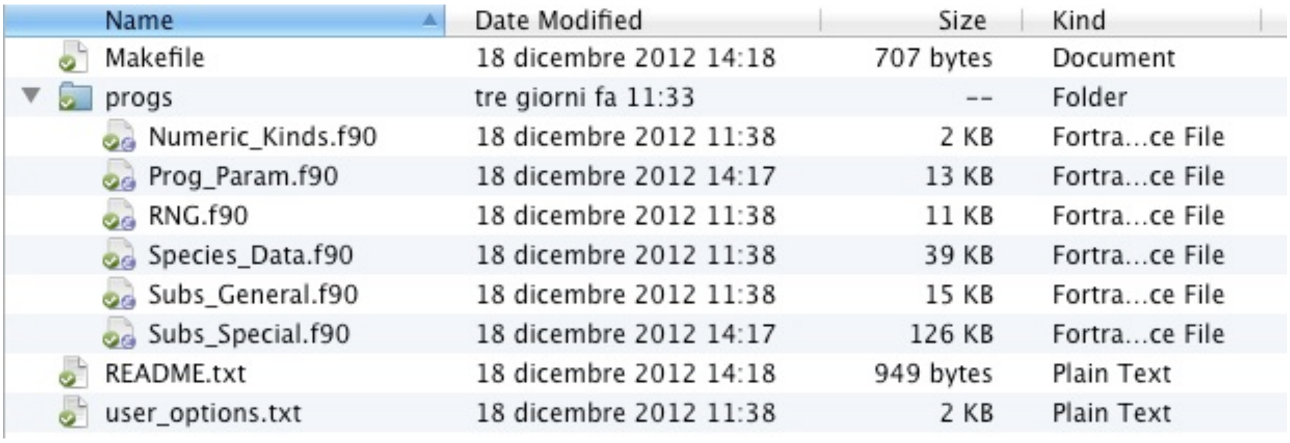
1) How to install/compile/run the program

Install:

GenoSim is a fortran90 set of programs to simulate multiple populations. It does not need to be installed, but compiled. To compile GenoSim you need to have installed a fortran compiler. Any compiler will (should) work, even GNU compilers. This software was tested with Absoft's *f90* and GNU(s) *gfortran* and *g95*. It is not OS dependent so it will (should) work on Linux, Mac and Windows (only tested on the former two).

Compile:

Once you have a compiler that fits you, you will need to create a folder with the files stored in the following way:



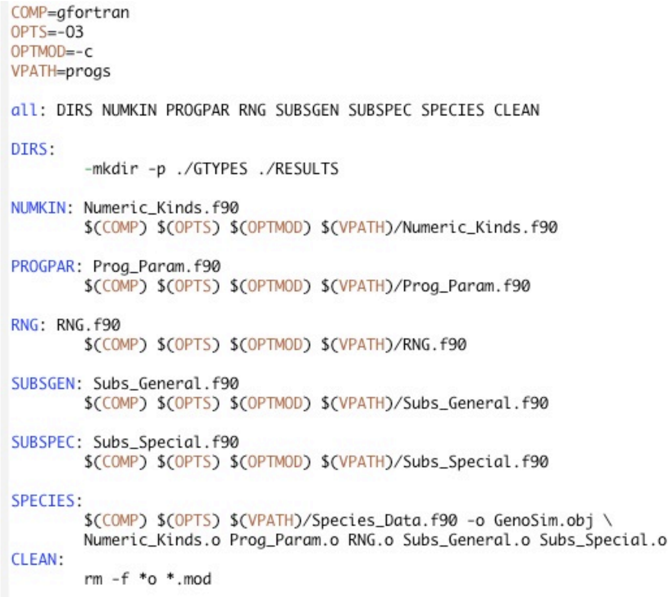
Name	Date Modified	Size	Kind
Makefile	18 dicembre 2012 14:18	707 bytes	Document
▼ progs	tre giorni fa 11:33	--	Folder
Numeric_Kinds.f90	18 dicembre 2012 11:38	2 KB	Fortra...ce File
Prog_Param.f90	18 dicembre 2012 14:17	13 KB	Fortra...ce File
RNG.f90	18 dicembre 2012 11:38	11 KB	Fortra...ce File
Species_Data.f90	18 dicembre 2012 11:38	39 KB	Fortra...ce File
Subs_General.f90	18 dicembre 2012 11:38	15 KB	Fortra...ce File
Subs_Special.f90	18 dicembre 2012 14:17	126 KB	Fortra...ce File
README.txt	18 dicembre 2012 14:18	949 bytes	Plain Text
user_options.txt	18 dicembre 2012 11:38	2 KB	Plain Text

I wrote a *Makefile* to help you compiling the programs. However, Makefiles work on linux/macs. If you use a Windows computer please contact me at Ezequiel [dot] Nicolazzi [at] tecnoparco.org for further instructions.

In any case, the original folder should have a *Makefile* file (to help you compile all the programs), a *progs* folder (with the original programs inside), and a *user_options.txt* file (is the user-defined parameter file for the program).

LINUX/MAC:

To compile the programs, first edit the *COMP=* variable in your Makefile. If you have set your compiler in your /usr/local/bin (linux/mac), you should be able to make it work just writing the name of the compiler. The *Makefile* looks like this:



```
COMP=gfortran
OPTS=-O3
OPTMOD=-c
VPATH=progs

all: DIRS NUMKIN PROGPAR RNG SUBSGEN SUBSPEC SPECIES CLEAN

DIRS:
        -mkdir -p ./GTYPES ./RESULTS

NUMKIN: Numeric_Kinds.f90
        $(COMP) $(OPTS) $(OPTMOD) $(VPATH)/Numeric_Kinds.f90

PROGPAR: Prog_Param.f90
        $(COMP) $(OPTS) $(OPTMOD) $(VPATH)/Prog_Param.f90

RNG: RNG.f90
        $(COMP) $(OPTS) $(OPTMOD) $(VPATH)/RNG.f90

SUBSGEN: Subs_General.f90
        $(COMP) $(OPTS) $(OPTMOD) $(VPATH)/Subs_General.f90

SUBSPEC: Subs_Special.f90
        $(COMP) $(OPTS) $(OPTMOD) $(VPATH)/Subs_Special.f90

SPECIES:
        $(COMP) $(OPTS) $(VPATH)/Species_Data.f90 -o GenoSim.obj \
        Numeric_Kinds.o Prog_Param.o RNG.o Subs_General.o Subs_Special.o
CLEAN:
        rm -f *o *.mod
```

To compile GenoSim, just type:

- *make*

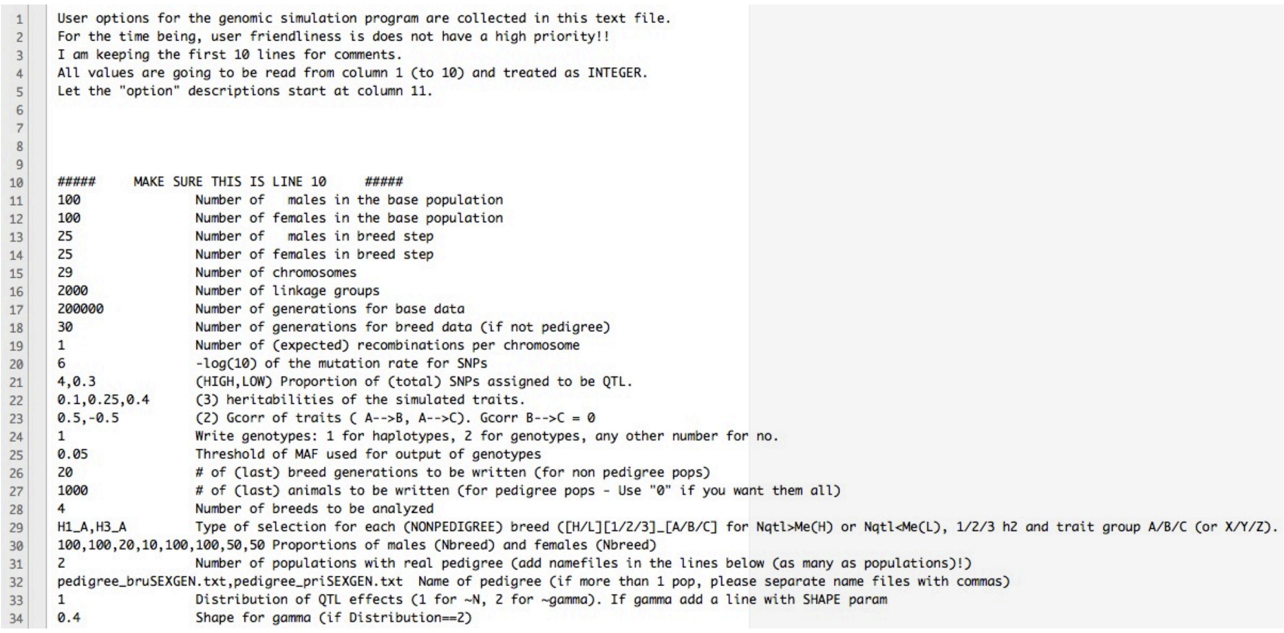
WINDOWS:
As said before, Makefile shouldn't work on Windows.
In this case you have to:
first make sure you have GTYPES and RESULTS folders created in the directory where the "Makefile" is.
Run the following commands sequentially (I use gfortran, but you use your own compiler!):

- gfortran -O3 -c progs/Numeric_Kinds.f90
- gfortran -O3 -c progs/Prog_Param.f90
- gfortran -O3 -c progs/RNG.f90
- gfortran -O3 -c progs/Subs_General.f90
- gfortran -O3 -c progs/Subs_Special.f90
- gfortran -O3 progs/Species_Data.f90 -o GenoSim.obj Numeric_Kinds.o Prog_Param.o RNG.o Subs_General.o Subs_Special.o [ALL THE RED TOGETHER!]

This will produce a file called "GenoSim.obj", other than a lot of files .mod and .o that you can now delete.

NOTE: If you do not change the simulation (at your own risk! ☺), this should be done just once..

Run the program:
Before running the program, you should modify the *user_options.txt* file, according to your preferences.
user_options.txt file looks like this:



```
 1  User options for the genomic simulation program are collected in this text file.
 2  For the time being, user friendliness is does not have a high priority!!
 3  I am keeping the first 10 lines for comments.
 4  All values are going to be read from column 1 (to 10) and treated as INTEGER.
 5  Let the "option" descriptions start at column 11.
 6
 7
 8
 9
10  #####     MAKE SURE THIS IS LINE 10     #####
11  100               Number of   males in the base population
12  100               Number of females in the base population
13  25                Number of   males in breed step
14  25                Number of females in breed step
15  29                Number of chromosomes
16  2000              Number of linkage groups
17  200000            Number of generations for base data
18  30                Number of generations for breed data (if not pedigree)
19  1                 Number of (expected) recombinations per chromosome
20  6                 -log(10) of the mutation rate for SNPs
21  4,0.3             (HIGH,LOW) Proportion of (total) SNPs assigned to be QTL.
22  0.1,0.25,0.4      (3) heritabilities of the simulated traits.
23  0.5,-0.5          (2) Gcorr of traits ( A-->B, A-->C). Gcorr B-->C = 0
24  1                 Write genotypes: 1 for haplotypes, 2 for genotypes, any other number for no.
25  0.05              Threshold of MAF used for output of genotypes
26  20                # of (last) breed generations to be written (for non pedigree pops)
27  1000              # of (last) animals to be written (for pedigree pops - Use "0" if you want them all)
28  4                 Number of breeds to be analyzed
29  H1_A,H3_A         Type of selection for each (NONPEDIGREE) breed ([H/L][1/2/3]_[A/B/C] for Nqtl>Me(H) or Nqtl<Me(L), 1/2/3 h2 and trait group A/B/C (or X/Y/Z).
30  100,100,20,10,100,100,50,50 Proportions of males (Nbreed) and females (Nbreed)
31  2                 Number of populations with real pedigree (add namefiles in the lines below (as many as populations)!)
32  pedigree_bruSEXGEN.txt,pedigree_priSEXGEN.txt  Name of pedigree (if more than 1 pop, please separate name files with commas)
33  1                 Distribution of QTL effects (1 for ~N, 2 for ~gamma). If gamma add a line with SHAPE param
34  0.4               Shape for gamma (if Distribution==2)
```

In detail:

- *Rows 1-10* are used for user general comments. Anything written in here won't be taken into account. Make sure the last line before "Number of males…" is number 10 ("#### MAKE SURE THIS LINE IS 10 ####" should help you!)
- *Rows 11-12: Number of males/females in base population*. Same as number of animals in the <u>species step</u> (see Jorjani, 2009 for more details). This is the step where common origin of multiple populations is generated. In this case 200 animals (100 male and 100 female) will be the base population.
- *Rows 13-14: Number of males/females in the breed population.* Number of animals in the <u>breed step</u> (see Jorjani, 2009 for more details). This is for EACH BREED that you'll be simulating. The population will be expanded/reduced (more/less copies of animals) if needed. Note that, if you choose to use real pedigree this number will be automatically set on the number of base male/female animals needed, but expanding it from this number of animals. For example: In this example, we'll generate 4 breeds: 2 pedigree and 2 non-pedigree breeds. The number of breed males and females is set to 25. This means that each breed will have 50 (all different) animals of the base population (note that 25+25 is the maximum number of animals allowed for 4 populations, if the base population is built out of 200 animals!!). In case one pedigree population needs more that 25 males(or females), let's say 100, these 25 will be copied during "expansion" (e.g. 4 copies of the same animal).
- *Row 15: Number of chromosomes*. Number of simulated cromosomes (note that the higher the number, the higher the number of SNPs and the higher the computational burden required).
- *Row 16: Number of linkage groups.* Total number of SNPs simulated will be:

  SNP= 32 * N_linkage_groups * N_chromosomes.

  If you're asking yourself: "why 32?" Its because SNPs are "compressed"/stored as bits within a default integer number, which contains 4 bytes. In other words: 4 (integer of 4 bytes) * 8 (bits on each byte). This allows to A) be fast and B) simulate millions of SNPs using VERY little memory!

Note that “linkage groups” is a way of calling a method to compress data, but **recombination and mutation can happen anywhere in the genome.**

- *Row 17: Number of generations for base data*. Number of (random_mating) generations to obtain the base population data. The number of generations should be chosen in terms of the mutation-drift equilibrium. Please check “*runtime_stats.txt*” file.
- *Row 18*: *Number of generations for breed data.* Number of selection generations in the breed data. Selection will be performed when no pedigree is assigned and on user defined trait group (please see *Row 29: Type of selection for each breed*). To avoid selection you should specify “100%” in row 31 (*proportion of animals to be selected*).
- *Row 19: Number of (expected) recombination events per chromosome.* As title says.
- *Row 20: -log(10) of the mutation rate for SNPs.* Mutation rate will be obtained from:

$$Mutation\_rate = 1.0 \times 10^{-(value)}$$

- *Row 21: (HIGH,LOW) proportion of (total) SNPs assigned to be QTL*. Depending on initial Ne (e.g. user defined), Me will be obtained using Goddard (2009) formula and final number of QTLs will be defined as follows:

  Traits 1-3 : HIGH * Me (Group “A1/B1/C1” HIGH # QTLs, first h2 defined)
  Traits 4-6 : HIGH * Me (Group “A2/B2/C2” HIGH # QTLs, second h2 defined)
  Traits 7-9 : HIGH * Me (Group “A3/B3/C3” HIGH # QTLs, third h2 defined)

  Traits 10-12: LOW * Me (Group “X1/Y1/Z1” … as before)
  Traits 13-15: LOW * Me (Group “X2/Y2/Z2” … as before)
  Traits 16-18: LOW * Me (Group “X3/Y3/Z3” … as before)

  Ideally, HIGH should be an integer number from 1 to N and LOW a real number from 0 to 1. Heritabilities of trait groups 1/2/3 are specified in row 22. Group A and B, and A and C are correlated following user specifications in row 24.

- *Row 22: Heritabilities of the simulated traits:* as title says. It is compulsory that 3 (comma or space) separated values are provided.
- *Row 23: Gcorr of traits.* Genetic correlation between traits A->B (first value) and A->C (second value). E.g. if 0.5 and -0.5 values are set, the Gcorr(A,B)=0.5 and Gcorr(A,C)=-0.5.
- *Row 24: Write genotypes*. If and how genotypes should be written out. 3 possible values can be input: 0 (don’t print any genotypes), 1 (haplotypes – 2 rows x animal), any other number (genotypes coded as 0,1,2, so 1 row x animal). Please consider disk space before deciding this option. For example: 960k genotypes might take HUGE amount of memory in no time. Considering that each animal using option 1 uses a bit less than 2Mb/animal, 500 animals are ~ 1Gb. You probably want to choose this carefully considering the number of generations that will be written into the file!!!
- *Row 25: Threshold of MAF.* As title says. Genotypes will be output if their frequency is > than this vale. While (very!) useful for running tests, use with caution when you are running your final iterations as the number and location of your SNPs will be different for each one of your breeds!!! (its breed specific, off course!)
- *Row 26*: *Number of (last) breed generations to be written (for non pedigree pops)*. As title says. For example a value of 3 will write out all genotypes and phenotypes for all animals in the last 3 generations of the simulation (not applicable when real pedigree is used).

- *Row 27: Number of (last) animals to be written (for pedigree pops).* Same as before but you must indicate the number of animals you want to have as output (following PEDIGREE order, starting from youngest animals).
- *Row 28*: *Number of breeds*. How many breeds will be simulated (ped + no ped)
- *Row 29*: *Type of selection for each breed*. Although (so far) this does not involve traits used for divergence of breeds, this option allows the user to choose the type of trait for which, once diverged, breeds are selected. There are a number of options available, and these options must be coded as: **1 letter, 1 number, 1 underscore and 1 letter**. This is what they mean:
  One letter: **H** or **L:** High or Low number of QTLs (underlying genetic structure of the trait, with high and low defined by the combination of Ne (and consequently Me) multiplied by the coefficient user defined in row 21.
  One number: **1, 2 or 3:** One is for h2 (as first, second or third, as assigned in row 23)
  One underscore ("_"): this is fixed.
  One letter: **A, B** or **C:** meaning trait group A, B or C (see row 21 for further details).
  Note that:
  H[1/2/3]_[A/B/C] correspond to groups **A**[h1/h2/h3], **B**[h1/h2/h3], and **C**[h1/h2/h3],

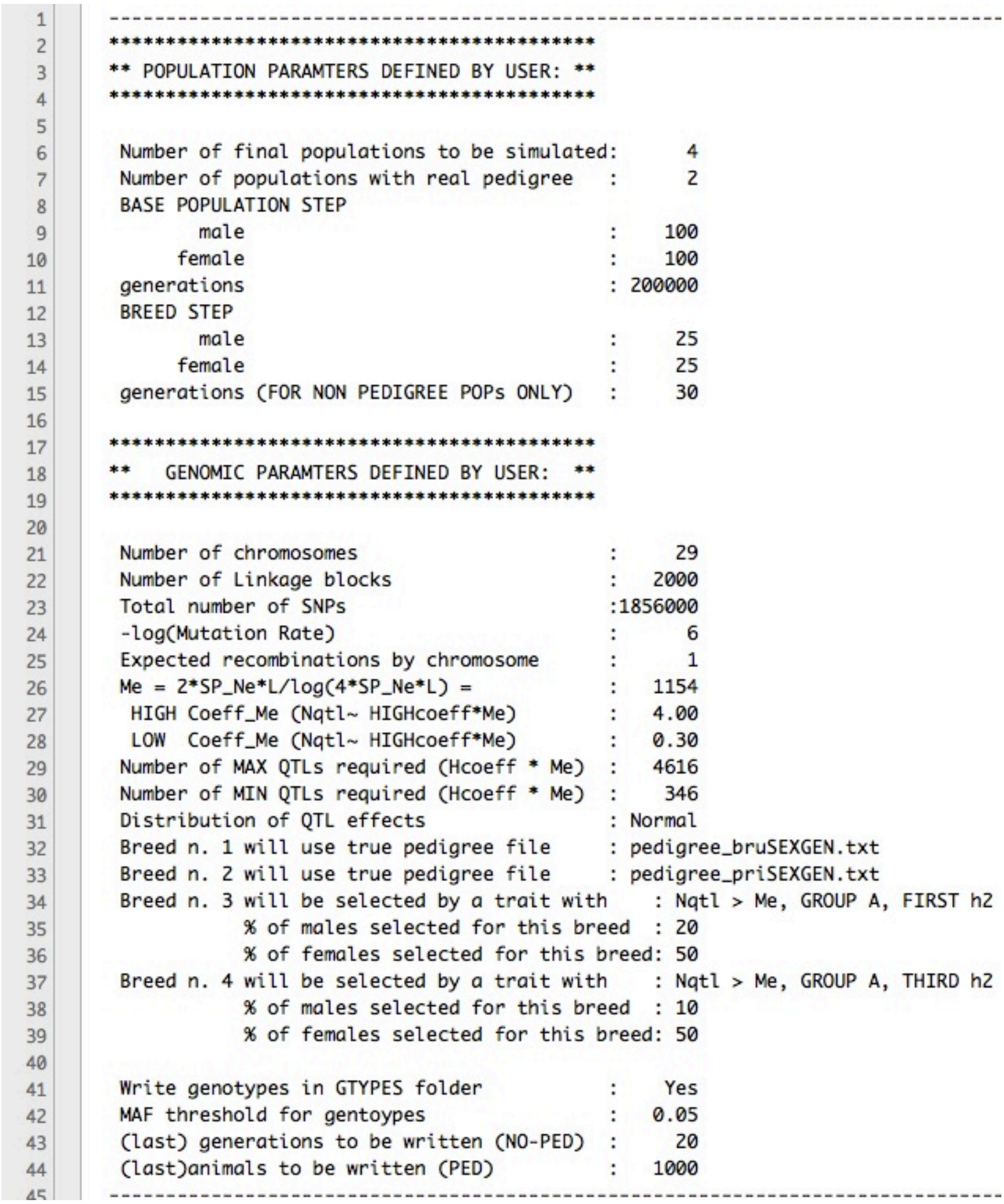
whereas

  L[1/2/3]_[A/B/C] correspond to groups **X**[h1/h2/h3], **Y**[h1/h2/h3], and **Z**[h1/h2/h3]

  This should be written as [H/L][1/2/3]_[A/B/C].
  For example: "H1_A" and "H2_C" are both traits with higher number of QTLs (> Me). The first one (H1_A) has the first heritability (assigned in row 23), and belongs to group A. The second one has the second heritability (assigned in row 23), and belongs to group C. These two traits are (genetically) correlated as explained in row 24.
  "L3_C" corresponds to a trait with low number of QTLs(<Me), using the third h2 (assigned in row 23), and the third group (Z, that would be correlated to X, with a genetic correlation assigned in row 24).
  <u>IMPORTANT NOTE: You must choose one option for each NON PEDIGREE breed you simulate (separated by comma or a blank space)</u>
- *Row 30: Proportion of (MALE/FEMALE) animals selected*. This is a proportion of animals kept when selection is performed. Two values (one for male and one for female) must be entered. For example, entering "90,90" means that 90% of the male and female animals will be selected for next generation (and consequently, 10% will be culled).
  PLEASE NOTE THIS HAS CHANGED FROM LATEST VERSION!: These values should be sorted by population within SEX, and DO NOT consider pedigree populations <u>anymore!</u> In this example we consider 4 populations: 2 pedigree and 2 non-pedigree = 4 values have to be entered (only for 2 non-pedigree pops):
  Value 1: Population 3 (non pedigree – male) 20 → this is actually used!
  Value 2: Population 4 (non pedigree – male) 10 → this is actually used!
  Value 3: Population 3 (non pedigree – female) 50 → this is actually used!
  Value 4: Population 4 (non pedigree – female) 50 → this is actually used!
- *Row 31: Number of populations with real pedigree*. As title states (note that populations with real pedigree will be analyzed first!!)
- *Row 32: Name of pedigree file* (if any). In case of more than one pedigree file, the names should be comma- or blank space- separated. In case NO pedigree is considered, you can leave this row blank.

- *Row 32: Distribution of QTL effects*: "1" for "normal distribution" and 2 for "Gamma distribution". Note that our preliminary tests showed no substantial differences in terms of results. This has, however, to be tested more thoroughly.
- *Row 33: Shape for gamma distribution*. In case a gamma distribution is chosen, otherwise this won't be read.

Be aware that when you run the program the options used will be displayed on screen. If you have doubts about the options you chose, check the screen! The following is for our example:



```
------------------------------------------------------------------------------
*******************************************
** POPULATION PARAMTERS DEFINED BY USER: **
*******************************************

 Number of final populations to be simulated:      4
 Number of populations with real pedigree   :      2
 BASE POPULATION STEP
        male                                :    100
      female                                :    100
 generations                                : 200000
 BREED STEP
        male                                :     25
      female                                :     25
 generations (FOR NON PEDIGREE POPs ONLY)   :     30

*******************************************
**   GENOMIC PARAMTERS DEFINED BY USER:  **
*******************************************

 Number of chromosomes                      :     29
 Number of Linkage blocks                   :   2000
 Total number of SNPs                       :1856000
 -log(Mutation Rate)                        :      6
 Expected recombinations by chromosome      :      1
 Me = 2*SP_Ne*L/log(4*SP_Ne*L) =            :   1154
  HIGH Coeff_Me (Nqtl~ HIGHcoeff*Me)        :   4.00
  LOW  Coeff_Me (Nqtl~ HIGHcoeff*Me)        :   0.30
 Number of MAX QTLs required (Hcoeff * Me)  :   4616
 Number of MIN QTLs required (Hcoeff * Me)  :    346
 Distribution of QTL effects                : Normal
 Breed n. 1 will use true pedigree file     : pedigree_bruSEXGEN.txt
 Breed n. 2 will use true pedigree file     : pedigree_priSEXGEN.txt
 Breed n. 3 will be selected by a trait with    : Nqtl > Me, GROUP A, FIRST h2
           % of males selected for this breed  : 20
           % of females selected for this breed: 50
 Breed n. 4 will be selected by a trait with    : Nqtl > Me, GROUP A, THIRD h2
           % of males selected for this breed  : 10
           % of females selected for this breed: 50

 Write genotypes in GTYPES folder           :    Yes
 MAF threshold for gentoypes                :   0.05
 (last) generations to be written (NO-PED)  :     20
 (last)animals to be written (PED)          :   1000
------------------------------------------------------------------------------
```

Output files:
Each time that you run the simulation, you'll obtain a number of output files in two folders. Please note that [PEDIGREE/NOPED] will follow your choices in the user_options.txt file, and the [#] means the number of the population as simulated.

GTYPES:

[PEDIGREE/NOPED]_final_[#].end. This is where genotypes or haplotypes are stored. One (genotypes coded as 0,1,2) or two rows per individual (both paternal and maternal haplotypes) are stored. Trace of this file is: FOR PEDIGREE files: animal code and haplotype. FOR NOPED files: animal_code, haplotypes.

Examples (haplotype data here):
1 110101.. -> animalID 1, sireID 11, damID 200, sex 1, paternal haplotype, haplotype: 1 1 ...
1 011101.. -> animal ID1, sireID 11, damID 200, sex 1, maternal haplotype, haplotype: 0 1...
So, first snp is heterozygous(1/0), the second homozygous (1/1) and so on..

1 01110.. -> animalID 1 generation 20, sireID 22 generation 19, damID 220 generation 19, , sex 1, paternal haplotype, haplotype: 0 1 ...
1 01111..-> animalID 1 generation 20, sireID 22 generation 19, damID 220 generation 19, , sex 1, maternal haplotype, haplotype: 0 1 ...
So, first snp is homozygous (0/0), the second homozygous (1/1) and so on..

ID_Geneal[PEDIGREE/NOPED]_[#].end. In this file general information for the animals in the genotype file are printed. This file is comma separated, and contains:

- (sequential) Animal_code (same as in the genotype file)
- Animal id: Original id in *PEDIGREE*, id_[Generation_number] in *NOPED*
- Sire: Original sire (or phantom sire number/ID) in *PEDIGREE*, Sire id_[generation_number] in *NOPED*.
- Dam: Same as for sire, but for female population
- Sex of the animal (1 for male, 2 for female)
- Type of genomic information: HT (Haplotype), GT (Genotype)

Note that animals in generation 1 will be forced to have 0 sire/dam (base population animals).

RESULTS:

- **gtl_positions.txt.** [PRESENT IF GENOTYPES ARE PRINTED] File with positions (first column) and effects for all 18 traits (columns 2-19) and type of QTL (column 20, for ABCXYZ QTLs)
- **freq_final[PEDIGREE/NOPED]_[#]_end.txt.** [PRESENT IF GENOTYPES ARE PRINTED] First row indicates the total number of fixed SNPs and QTLs. All other rows display a sequential number, the SNP "name" (useful if MAF is applied, otherwise these 2 columns are =), their p, q and if it is a marker (M) or a QTL (Q), and in this case, what kind of QTL it is (ABCXYZ).
- **g_var_pop[#].txt.** In this file, genetic variances are displayed. First and second column are for the genetic variance calculated over the first and last generation of base population (e.g. using their p's and q's), and the third is the genetic variance for the population at the last generation of selection (for NOPED populations) or the whole pedigree (for PEDIGREE populations).
- **phen_final[PEDIGREE/NOPED]_[#].txt.** This file contains a header. It simply contains the phenotypic values for each animal on the 18 simulated traits, comma separated. In case of NOPED populations (e.g. populations without real pedigree), an asterisk will indicate the trait on which selection has been performed. This file will contain the phenotypes for all the *genotypes* produced. NOPED animals will be identified with their generation number. In case

of PEDIGREE populations (e.g. populations with real pedigree), all animals in the pedigree will be included in this file.

- **tbv_final[PEDIGREE/NOPED]_[#].txt.** Same as for phen_final[PEDIGREE/NOPED] files, but for true breeding values.
- **corr_phen—tbv_final[PEDIGREE/NOPED]_[#].txt.** Correlation between phenotypes and true breeding values. This is separated for males/females in the NOPED populations (eg. populations without a real pedigree), and considering all animals (not separated for male/female animals) in the PEDIGREE populations.
- **Runtime_stats.txt.** A file that describes runtime statistics, each 1000 generations. This is useful to check equilibrium is reached. Stats displayed are: Generation, Number of homozygous SNPs, Observed heterozygosity, MAF distribution (from 0 to 0.5, in bins of 0.1)