



# GenetSim: Software for Simulation of Familial Data in Genetics and Epidemiology

Michael B. Miller & Na (Michael) Li  
University of Minnesota, School of Public Health



## Abstract

GenetSim is a collection of functions and scripts written in the Octave language ([www.octave.org](http://www.octave.org)), but it will be ported to the R language ([www.r-project.org](http://www.r-project.org)) with a graphical user interface (GUI) in coming months. GenetSim allows users to generate data under a wide variety of genetic models with no limits on pedigree complexity, sample size, number of trait or marker loci (limits are imposed only by machine architecture). We used GenetSim to generate data using the model of GAW 10 Problem 2 as a proof of concept for our new software. We could identify no other available software that is as flexible as GenetSim. GenetSim is freely available under the GNU General Public License.

## GAW 10 Resimulation

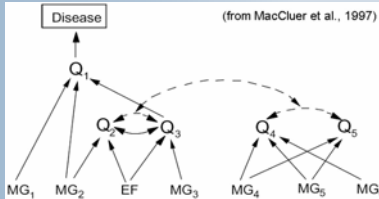
We repeated the GAW 10 Problem 2 genetic data simulation using the same number and structure of families and the complex multivariate model described below and in the figure below (taken from MacCluer et al. 1997).

- 6 major genes (MG1 – MG5 in the figure below)
  - epistasis
  - pleiotropy
  - genotype-by-sex interaction
- 5 quantitative traits (Q1 – Q5 in the figure below)
  - environmental factor (EF in the figure below)
  - age effects
  - one quantitative trait directly affects another
  - affection status (using a threshold on a quantitative trait)
  - missing data
  - 367 markers on 10 chromosomes

200 replicates of 239 nuclear families (N = 1,497 per rep.)  
200 replicates of 23 extended families (N = 1,164 per rep.)

We generated data under the GAW 10 model producing a total of

- 197,829,600 marker genotypes
- 3,190,800 trait locus genotypes
- 4,254,400 phenotypes



## Feature Comparison

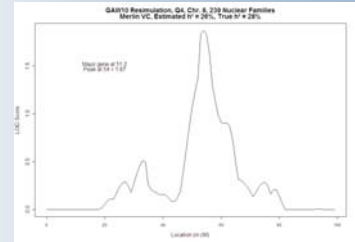
Because GenetSim uses an interpreted language that includes a vast number of ready-made mathematical and statistical functions, it is extremely flexible. Limits on array sizes and dimensionalities are imposed only by computer memory because a 64-bit version can access, theoretically, thousands of terabytes of memory. Thus, numbers of subjects, families, alleles, phenotypes etc., are essentially unlimited. GenetSim can run under UNIX, GNU/Linux, MS-Windows with Cygwin, and Macintosh OSX.

	GenetSim	GASP	Morgan	SIMLA	SIMLINK	SIBSIM	SIMULATE2
number of families	Any	5000	?	5000	20	?	75
size/structure family	Any	5000	?	109	200	?	65
handles inbreeding loops	Yes		Yes		?	?	?
handles MZ twinning	Yes		?			?	
number of marker loci	Any	400	500	12k	2	?	1000
numbers of alleles	Any	12	100	31	?	?	15
number of trait loci	Any	400	?	2	1	?	?
number of phenotypes	Any	5	?	11	1	?	?
number of chromosomes	Any	Any	22	?	2	1	Any
epistasis	Yes	Yes		Yes			?
pleiotropy	Yes	Yes					No
non-normal traits	Yes						?
sex effects	Yes		?	Yes	?	?	Yes
G x Sex interactions	Yes		?	Yes	?	?	?
G x E interactions	Yes		?	Yes		?	?
interference models	Many						?
specify inheritance	Yes		Yes				?
UNIX, Windows, Mac	All	U	U	U	UW	U	UW
GPL-compatible	Yes		?			Yes	?

## Results

The entire GAW 10 resimulation ran in less than 30 minutes on a dual Intel Xeon 2.8 GHz machine running Linux. This is an acceptable speed for data of this complexity.

We performed a model-free variance-components analysis in MERLIN for trait Q4 on chromosome 8 using the first replicate of 239 nuclear families. The MERLIN result (LOD plot shown in the figure below) provided estimates of heritability and major locus location that were extremely close to the true parameter values used in the simulation. This partially confirms the validity of our approach.



## Conclusions

GenetSim algorithms provide a very flexible method for producing genetic data for simulation studies within a high-level language like Octave or R. Because GenetSim functions can be called by other functions or scripts, it is possible to design very easy-to-use programs for users who have little knowledge of genetics, statistics or programming, e.g., students in an introductory course. A GUI will also be available in future versions. On the other hand, GenetSim users who would like to dig deeper can generate data sets under extremely complex genetic models. In fact, GenetSim is far more flexible and complete than any extant genetic simulation program. GenetSim functions are freely available under the GNU General Public License. To be posted soon at the following URL:

<http://taxa.epi.umn.edu/GenetSim/>

## Acknowledgments

Soon-Young Jang and Gregg Lind wrote much of the Octave code for this project. This research was supported by 5RO1-HL09609-12, 1R01-AG021917-01A1 and by the University of Minnesota.

**GAW10 Resimulation, Q4, Chr. 8, 239 Nuclear Families**  
**Merlin VC, Estimated  $h^2 = 26\%$ , True  $h^2 = 28\%$**

