A General Genomics Simulation Program

Hossein Jorjani
Interbull Centre, Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences
Box 7023, S-75007, Uppsala Sweden
Hossein.Jorjani@hgen.slu.se

Abstract

Stochastic simulation studies are an essential part on many quantitative genetics and animal breeding studies that have been especially useful in genomic studies. Here, the justification behind and the details of a general genetic/genomic stochastic simulation programs package is reported. The package is designed to be flexible enough to many purposes, especially international genomic evaluations. Further, the package can be used for educational purposes, as well as, for other kinds of genetic research including validation of genomic evaluation models and software.

Keywords: quantitative genetics, animal breeding, validation, finite locus model, genetic evaluation

Introduction

Stochastic simulation studies have been used in quantitative genetics and animal breeding for several decades (e.g. Bulmer, 1976). The initial skepticism about simulations has gradually decreased and now there seems to be a consensus about their usefulness. In other words, the role of stochastic simulations, alongside analytical work and real data analysis, in creation of new knowledge is commonly acknowledged.

Most of the stochastic simulations in the field of animal breeding have generated data, more notably breeding values, by sampling from some sort of continuous distribution, with normal distribution being the absolute favorite. However, simulation of individual loci in the frame work of "finite locus models, FLM" has always been the choice in quantitative genetic studies (e.g. Bulmer, 1976; de Boer & Van Arendonk,1992; Verrier, *et al.*, 1989; Jorjani *et al.*, 1994, 1997a,b,c, 1998; Jorjani, 2002).

With recent years' technological progress in the area of genomics and advent of high density SNP-chips, even animal breeding studies have turned into FLM for stochastic simulation studies. In a sense, the border between animal breeding and quantitative genetic studies has become more diffuse than before.

Personally, I have a sense of vindication when people who were advising me not to use FLM in animal breeding studies are now recommending the use of such models.

Why do we need a simulation software?

International genetic evaluation through any variant of MACE (Schaeffer, 1994) has an absolute dependence on the quality of data from participating organizations/countries. It is not only the unbiasedness of the national genetic evaluation results that affect MACE, but also a large number of other factors must be checked for quality assurance. In the past, in addition to the extensive field data analyses a number of special purpose simulation studies have served to make the check the system (e.g. Sigurdsson *et al.*, 1996; Klei & Weigel, 1998; Fikse & Banos, 2001; Jorjani *et al.*, 2005).

With regards to genomic evaluations some of the important issue are the validation of national genomic evaluations (through estimation of

Mendelian sampling variance or otherwise), taking into account the selection bias, and all sorts of residual covariances that might exist (among other things from bull genotypes used in multiple countries). Further, because of multiplicity of methods to estimate SNP effects, DGVs and GEBVs, it is also imperative to have a testing environment for validation of algorithms and software packages.

Therefore, it is prudent (and economically justifiable) to invest on adopting an existing general simulation software, or creating such a software, that can be used for a variety of purposes.

Short review of available simulation software

There has been an explosion of simulation studies to explore different aspects of using genomics data in animal breeding. A review of the simulation strategies used in these studies shows that:

a) The simulated genomic structures (e.g. number of chromosomes, number and nature of linkage groups, rates of recombination and mutation, number of markers and QTLs, allele frequencies, allele effects, etc.) are often small and simple. For example, there are only a few chromosomes, recombinations ignore linkage groups, non-additive effects are left out, there are few QTLs, only one or a few structurally similar traits, etc.;
b) Creation of linkage disequilibrium is done through the use of random mutation, drift, and recombination, and invariably ignore natural selection;
c) Population structures are also invariably simple, i.e. there are no structured cohorts, there exists only one population (and therefore, migration is left out), etc.;
d) The majority of simulation strategies are too specialized to study one or a few questions; and finally
e) They are not flexible enough and not user friendly.

Desirable properties of a general simulation package

General purpose simulation software should be able to:

a) Accommodate all major (random drift, migration, mutation and natural selection) and minor evolutionary forces (mating pattern, realistic linkage structure and recombination);
b) Accommodate similarity to domestic animal genetic resources (species, breed and local population (e.g. country) structures;
c) Accommodate cohort structures similar to farm animals (e.g. herd, systematic environmental factors, production systems);
d) Provide flexibility for maintenance, usage and development of programs.

Because the existing simulation programs suffer from some shortcomings that renders them possessing the desired properties mentioned above, it was decided to start creation of a general purpose simulation program.

Programming philosophy

The simulation program is written in Fortran 90/95. The executive (pre-compiled) version of the program capable of running on Windows or Linux will be distributed for the end users. An “instruction text file” containing the user defined options should be the only concern of the end user. A “program parameter” file should read the user defined options and dynamically allocate the memory. The main program should be as simple as possible and just applying the user defined options. All user defined option, small and large, should be carried out by “special subroutines” included in a single module. “General subroutines, functions, random number generators, an alike should be included in separate modules.

In building up the new software package a set of arbitrarily chosen genetic and population structure parameters were used as examples.

These are as follows:

Organism: Any diploid, sexually reproducing organism;
Chromosome structure: 29 autosomal and 1 sexual chromosomes. Number of autosomals could be set to any value. However, it should be possible to have built-in genome structures for all agriculturally important species;
X/Y chromosomes: X chromosome is twice as large as the Y chromosome;
Linkage groups: each chromosome is composed of 1000 linkage groups, each with 32 loci. In future, the linkage groups could be made of variable size depending on the species;
Recombination sites: Recombination is now independent of the linkage groups and can occur anywhere on the chromosome. It should be, however, possible to assign specific positions as recombinational hot spots and/or make them related to the real linkage groups;
Recombination rate: Each chromosome is assumed to be 100 cM long. When more realistic chromosome structures are implemented, it would be possible to have variable chromosome sizes.
Number of SNPs: Given the number of chromosomes and linkage groups, there are 960,000 SNPs in females and 944,000 SNPs in males. This number can easily be changed to larger or smaller values;
Allele frequencies: Allele frequencies are 0.5 for all alleles in the base population. In future, the allele frequencies can be sampled from any distribution;
Allelic effects: Additive and dominance effects are envisaged for each locus. For the time being, and for the sake of simplicity in the programming, the dominance to additive ratio (V_D/V_A) will be equal to $1/h^2$. Dominance effect is initially considered to be equal for all loci;
Mutation rate: The preliminary mutation rates are 10^{-5} for the SNPs and 10^{-6} for the QTLs. These should be made flexible in the future;
QTL effects: Gamma distribution is used for QTL effects. Any distribution can be used;
Simulated traits: Thirty pairs of traits (60) are simulated. Ten pairs with heritability values 0.01 to 0.10 with interval of 0.01, 10 pairs with heritability values of 0.12 to 0.30 with interval of 0.02, and 10 pairs with heritability values of 0.35 to 0.85 with interval of 0.05.
Number of QTLs: Of the two traits of a pair of traits, one is controlled by a randomly chosen set of 5% of the all SNPs (i.e. 48000). The other trait is controlled by variable number of SNPs depending on the heritability of the traits (number of SNPs = 500 * heritability value);
Species generation: The closed based population undergoes 1000 generations of mating to establish a specific species of farm animals (species data). Species data is subjected to random drift, mutation, recombination and natural selection (i.e. selection for low heritability traits);
Breed generation: Species data at generation 1000 is sampled 10 times. Each sample is treated as a closed population and undergoes 200 generations of mating to establish a specific breed of farm animals (breed data). Breed data is subjected to random drift, mutation, recombination, natural selection (as described above) and artificial selection (i.e. selection for high heritability traits);
Country generation: Breed data at generation 200 is sampled 10 times. Each sample is treated as an open population and undergoes 50 years (~ 10 generations) mating to establish a specific population of farm animals (country data). Country data is subjected to random drift, mutation, recombination, natural selection (as described above) and artificial selection (i.e. selection for medium heritability traits). Country data is structured in cohorts affected by systematic environmental factors.

Results

In order to demonstrate how the program package works some results are presented here.

Table 1 shows the assignment of SNPs to QTLs. At generation 0 (base population) 48,000 SNPs are randomly chosen as QTLs. Then a proportion of these are randomly assigned to the first trait of each pair of traits. In the examples shown here, there are two pairs of traits with heritability values equal to 0.06 and 0.60, respectively.

Table 1. Assignment of SNPs as QTLs.

	Trait			
	11	12	51	52
QTL position	h^2 =0.06	h^2 =0.06	h^2 =0.60	h^2 =0.60
6633	0.0000	**0.0269**	**2.0041**	**0.6868**
6638	**-0.8990**	**0.4346**	**-0.4749**	**0.0113**
6658	0.0000	**0.5877**	**0.0080**	**0.4185**
6664	0.0000	**0.1984**	**-1.1459**	**0.0264**
6705	0.0000	**0.5580**	**-0.4968**	**0.4736**
6728	0.0000	**0.4166**	**-1.2543**	**0.5345**
6749	0.0000	**0.9102**	0.0000	**0.0585**
6753	**1.1334**	**0.0841**	0.0000	**0.1216**
6776	0.0000	**0.2162**	0.0000	**0.1668**
6810	0.0000	**0.6267**	**0.8628**	**0.2026**
6817	0.0000	**0.7705**	**-0.0854**	**0.6890**
6873	**-0.3692**	**0.9759**	**2.2412**	**0.9369**

For Traits 11 and 51 the expected number of QTLs are 3000 and 30,000, respectively (6 x 500 and 60x 500). For Traits 12 and 52 there are 48,000 QTLs. Because there is an overlap between the QTLs affecting different traits, an automatic way of introducing pleiotropy (and therefore, correlation) is built-in the simulations. The additive values assigned to each QTL are orthogonal. However, it is possible to intentionally introduce a covariance structure among the traits in the base population.

Because of random assignment of SNPs to QTLs the actual number of QTLs might be different from the expectation. Table 2 shows the expected and realized number of QTLs for eight traits.

Table 2. Expected and realized number of QTLs.

	QTL Number	
h2	Expected	Realized
1	500	482
2	1000	997
3	1500	1502
4	2000	2005
65	32500	32513
70	35000	35194
75	37500	37565
80	40000	39982

Table 3. In the example run whose results are shown here the average distance between QTLs in terms of intervening SNPs and some descriptive statistics are shown.

Table 3. Statistics on QTL positions.

Average	19.92
Standard deviation	19.30
Minimum	1
Median	14
Maximum	227

Conclusions

Such a general simulation program package can be useful for many purposes. Some examples areas are education, research, validation of national and international genetic/genomic evaluation models, and software validation.

References

Bulmer, M.G. 1976. *Genet. Res. Camb. 28,* 101-117.

de Boer, I.J.M. & van Arendonk, J.A.M. 1992. *Theor. Appl. Genet. 84,* 451-459.

Fikse, W. F. & Banos, G. 2001. *J. Dairy Sci. 84*, 1759-1767.

Jorjani, H. 2002. *Proc. 7th WCGALP* 19-23, August 2002, Montpellier, France. *33*, 95-98.

Jorjani, H., Emanuelson, U. & Fikse, W.F. 2005. *J. Dairy Sci. 88*, 1214-1224.

Jorjani, H., Engström, G. & Liljedahl, L.-E. 1994. *Proc. 5th WCGALP*, University of Guelph, Guelph, Canada. *19*, 163-166.

Jorjani, H., Engström, G., Strandberg, E. & Liljedahl, L.-E. 1997a. *Acta Agric. Scand. Sec. A, Animal Sci. 47*, 65-73.

Jorjani, H., Engström, G., Strandberg, E. & Liljedahl, L.-E. 1997b. *Acta Agric Scand. Sec. A, Animal Sci. 47*, 74-81.

Jorjani, H., Engström, G., Strandberg, E. & Liljedahl, L.-E. 1997c. *Acta Agric Scand, Sect A, Animal Sci. 47*, 129-137.

Jorjani, H., Strandberg, E., Engström, G. & Liljedahl, L.-E. 1998. *Proc. 6th WCGALP*, University of New England, Armidale, Australia, *26*, 45-48.

Klei, L. & Weigel, K.A. 1998. Proc. of the Interbull Open Meeting, Rotorua, New Zealand. January 18-19, 1998. *Interbull Bulletin 17*, 8-14.

Schaeffer, L.R. 1994. *J. Dairy Sci. 77*, 2671-2678.

Sigurdsson, A., Banos, G. & Philipsson, J. 1996. *Acta Agric. Scand., Sect. A, Anim. Sci. 46*, 129-136.

Verrier, E., Colleau, J.J. & Foulley, J.L. 1989. *Theor. Appl. Genet. 77*, 142-148.