BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 5 2009, pages 680–681
doi:10.1093/bioinformatics/btp045

Genetics and population analysis

QMSim: a large-scale genome simulator for livestock

Mehdi Sargolzaei* and Flavio S. Schenkel
Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario, Canada

Received on November 17, 2008; revised on January 15, 2009; accepted on January 18, 2009
Advance Access publication January 28, 2009
Associate Editor: Alex Bateman

ABSTRACT

Summary: QMSim was designed to simulate large-scale genotyping data in multiple and complex livestock pedigrees. The simulation is basically carried out in two steps. In the first step, a historical population is simulated to establish mutation-drift equilibrium, and in the second step, recent population structures are generated, which can be very complex. A wide variety of genome architectures, ranging from infinitesimal model to single-locus model, can be simulated. The program is efficient in terms of computing time and memory requirements.

Availability: Executable versions of QMSim for Windows and Linux are freely available at http://www.aps.uoguelph.ca/~msargol/qmsim/.

Contact: msargol@uoguelph.ca

1 INTRODUCTION

Linkage disequilibrium (LD) and linkage analyses have been extensively used to identify quantitative trait loci (QTL) in human and livestock species. Recently, interest in whole genome fine mapping and especially genome-wide selection has grown as a result of the dramatic increase in the number of known single nucleotide polymorphisms (SNP) and the decrease in genotyping costs. The access to dense marker maps has opened up the possibility for new approaches for fine mapping and genome-wide selection. However, even though genotyping costs have substantially decreased, large-scale genome-wide association studies are still costly.

Simulation is a highly valuable tool for assessing and validating proposed methods for QTL mapping and genome-wide selection at very low cost, allowing also for the prediction of future changes in genetic parameters. During the last few decades, simulation has played a major role in population genetics and genomics. Several software programs have been developed for simulating genomes as a means of validating new algorithms and methods especially in human research (e.g. Hoggart *et al*., 2008; Hudson, 2002; Li and Li, 2007; Schaffner *et al*., 2005). However, most of the developed software tools simulate non-overlapping generations and do not provide the functionality required for many studies in livestock. In addition, most of the existent genome simulation programs for livestock are not publicly available.

Simulating and analyzing genomic livestock data differ in several aspects from analyses carried out in humans. For instance, in livestock, a common strategy for detecting QTL is either to use multi-generational pedigrees, large family sizes, in which artificial insemination is practiced, or to design crossbreeding programs, such as F2 and back crosses (Andersson, 2001). Moreover, human populations have been experiencing an expansion in effective population size (Ne), while Ne in livestock populations has decreased. Consequently LD in livestock usually extends over longer distances than in humans (Farnir *et al*., 2000). Furthermore, combined LD and linkage QTL mapping has attracted more attention in livestock due to strong family structure (Meuwissen *et al*., 2002). Therefore, the population structure is crucial to identify and correctly interpret the associations between molecular and phenotypic diversity (Pritchard and Rosenberg, 1999).

The objective of this article is to introduce a forward-in-time software named QMSim that was designed to simulate large genomes and complex pedigree structures, mimicking livestock populations. QMSim is basically a family based simulator, which can also take into account predefined evolutionary features, such as LD, mutation, bottlenecks and expansions.

2 METHODS AND DISCRIPTION OF THE ALGORITHMS

Population structure: Step (1) in order to create initial LD and to establish mutation-drift equilibrium, a historical population is simulated based on forward-in-time approach by considering only two evolutionary forces: mutation and drift. The mating system is based on the union of gametes randomly sampled from the male and female gametic pools. Expansion and contraction of the historical population and unbalanced sex ratio are allowed. Step (2) after creating a historical population, one or multiple recent population structures are simulated. Animals from the last historical generation can be chosen as founders of the recent populations. However, for the case of multiple recent populations, founders can also come from previously defined recent populations. Selection and culling can be implemented based on different criteria such as phenotypes, true genetic values or estimated breeding values for a single trait with predefined heritability and phenotypic variance. Breeding value is a measure of the additive genetic value of an individual as a parent and, in QMSim, it can be estimated using three different approaches: (i) best linear unbiased prediction via an animal model; (ii) based on predefined accuracy; and (iii) approximated based on the number of offspring with record. The mating design can be random, assortative (positive or negative) or optimized to minimize or maximize inbreeding. The mating design that maximizes inbreeding allows one to quickly create an inbred line. Optimization of inbreeding is carried out using the simulated annealing method. The program can simulate sex limited traits, such as milk yield. Owing to the object-oriented programming, it is easy to simulate multiple populations with different structures and selection criteria. QMSim has flexibility for simulating wide range of population structures. For example, in livestock, some of QTL mapping designs involve line crosses produced from inbred

*To whom correspondence should be addressed.

lines with divergent phenotypes. In this case, inbred lines can be generated and subsequently crossed. Another example is the simulation of two lines coming from the same base population to assess accuracy of genomic breeding values for a particular genomic evaluation method. Here, one line can be treated as training set and the other one as validation set. These scenarios can be readily simulated by QMSim.

Genome: A wide range of parameters can be specified for simulating the genome such as: mutation rate, crossover interference, length and number of chromosomes, number of markers, number of QTL, location of markers and QTL, number of alleles, allelic frequencies, allelic effects, missing genotype rate and genotyping error rate. This flexibility permits for wide variety of genetic architectures to be simulated.

No allelic effects are simulated for markers, so they are treated as neutral. For QTL, additive allelic effects can be sampled from gamma, normal or uniform distributions. Alternatively, predefined relative additive variance for each QTL can be supplied by the user.

Crossover is a key factor that gradually breakdowns the allelic associations or LD. The number of crossover events for each chromosome is sampled from a Poisson distribution with mean equal to the length of chromosomes in Morgan. Then locations of crossovers along chromosomes are assigned at random. Because QMSim requires centiMorgan positions for markers and QTL as input information, simulating recombination hot spots and cold spots is straightforward by adjusting the distance between markers or QTL. For example, for generating a hot spot one might increase the interval between loci and the size of the interval will determine the intensity of the hot spot. For recent genealogies, positions of crossover events can be stored in an output file, allowing one to trace the haplotype blocks. To establish mutation-drift equilibrium in the historical generations either infinite-allele mutation model or recurrent mutation model is used. The infinite-allele mutation model assumes that a mutation creates a new allele, while the recurrent mutation model assumes that a mutation alters an allelic state to another and, hence, does not create a new allele. In the recurrent model, transition probabilities from one allelic state to another are assumed equal. Different mutation rates for markers and QTL can be defined. The number of mutations is sampled from a Poisson distribution.

3 INPUTS AND OUTPUTS

The program requires a parameter file, in which various parameters for the simulation are specified. A simple internal lexer reads and translates the parameters from the input file and, in the case of incorrect inputs, an appropriate message is displayed. The parameter file consists of five main sections. Parameters for each section are explained in details in the user's guide.

The current version of the simulator can optionally produce several detailed output files in text format, such as pedigree information, population structure, phenotypes, true genetic values, crossover positions, LD statistics, linkage map, phased genotypes and allele frequencies and effects. In order to make the program as memory efficient as possible, output files are saved during simulation. However, when simulating large marker panels or large populations over many replicates, large genotype files might become an issue. In this situation, the genotype file can be stored for user-specified generations or in binary format by altering the output options. Additionally when simulating bi-allelic markers (i.e. SNP), the two alleles can be coded in one genotype for each locus to save hard disk space.

Initial seed numbers are backed up for each simulation run, thus allowing one to regenerate the same outputs whenever needed.

4 IMPLEMENTATION AND EFFICIENCY

The program is written in C++ and is portable to multiple operating systems. Executable files are currently available for Windows and Linux platforms. QMSim is equipped with the high-quality and fast Mersenne Twister random number generator (Matsumoto and Nishimura, 1998). Computing time for simulating 500 K SNP panel in a historical population of 1000 individuals for 20 discrete generations (i.e. a total of 21 000 individuals) and in a recent population with the same size and number of discrete generations on an AMD Opteron server, running at 2.6 GHz with 16 GB RAM was 5 and 12 min, respectively. RAM requirement was around 2 GB. The corresponding times for 50 K SNP panel were 14 and 70 s and for the 10 K SNP panel were 2 and 10 s, respectively.

5 CONCLUSIONS

As genomic applications continue to develop in livestock, there is an increasing demand for efficient and reliable tools to simulate genomic data in complex pedigree populations. QMSim is a user-friendly tool for simulating large-scale genomic data in livestock, which helps to validate new approaches for fine mapping and genomic selection. QMSim integrates efficient and fast algorithms, which lead to high computing performance.

ACKNOWLEDGEMENTS

The authors thank three anonymous reviewers for their very helpful comments.

Funding: L'Alliance Boviteq Inc. (SEMEX Alliance, Canada); Natural Sciences and Engineering Research Council of Canada (Collaborative Research and Development grant); Ontario Centre for Agriculture Genomics (Challenge Fund).

Conflict of Interest: none declared.

REFERENCES

Andersson,L. (2001) Genetic dissection of phenotypic diversity in farm animals. *Nat. Rev. Genet.*, **2**, 130–138.

Farnir,F. *et al.* (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.*, **10**, 220–227.

Hoggart,C.J. *et al.* (2007) Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725–1731.

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, **18**, 337–378.

Li,C. and Li.M. (2007) GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, **24**, 140–142.

Matsumoto,M. and Nishimura,T. (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Model. Comput. Simul.*, **8**, 3–30.

Meuwissen,T.H.E. *et al.* (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, **161**, 373–379.

Pritchard,J.K. and Rosenberg,N.A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, **65**, 220–228.

Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.