GenSim: Simulation of Descendants from Sequenced Ancestors Data

A.S. Leaflet R2955

Hao Cheng, Research Assistant; Rohan Fernando, Professor; Dorian Garrick, Professor, Department of Animal Science

Summary and Implications

High-density real or imputed SNP genotypes are now routinely used for genomic prediction and genome-wide association studies. This is extending to the use of actual or imputed next generation sequence data in these activities. Simulation studies are useful to mimic these complex scenarios and test different analytical methods. We have developed a software tool GenSim to simulate sequence data in descendants. In this software, a crossover position and origin simulation (CPOS) algorithm is implemented to efficiently drop down sequence data from founders in complex pedigrees. Parallel programming techniques are used to reduce computing time. Now C++ and Julia versions of GenSim are available on request (haocheng@iastate.edu).

Introduction

Analysis of real or imputed genotypes for genomic prediction and genome-wide association studies can result in findings that are difficult to validate. Simulated data have advantages in that the underlying causal mutations and simulated breeding values are available for direct validation. In general, there are two types of simulation methods: coalescent methods and forward-in-time (drop down) methods. Compared to coalescent-based simulations, forward-in-time simulations are very flexible, which allows modeling large numbers of recombination events in concert with complex life-like selection scenarios. On the other hand, forward-in-time methods, which drop down the pedigree to simulate and record genomic information for every individual in the entire population, are computationally intensive. Here we propose a crossover position and origin simulation (CPOS) algorithm to efficiently simulate sequence data and complicated pedigree structures across multiple generations.

A software tool GenSim incorporating the CPOS algorithm has been developed to use founders characterized by real genome sequence data, and complicated pedigree structures among descendants. Parallel programming techniques are used to reduce computing times.

Materials and Methods

The basic idea of CPOS is to record the crossover locations and founder origins of chromosome segments upstream and downstream of the crossover site. It is not necessary to store all information for the whole genome. At first, founders are labeled with origin chromosome identifiers. These founders can be generated with user-defined map positions and allele frequencies or from real genotypes or sequence data. During meiosis, only the positions of the crossover sites and corresponding origin chromosomes need to be recorded in order to reconstruct entire genomes to the density of the founder genomes. Each individual can be represented using two vectors: a vector of crossover locations and a vector of founder chromosome identifiers. Two ways to deal with mutations are implemented. One way is to make a new founder every time a de novo mutation occurs. Another way is by recording in an additional vector the positions of inherited de novo mutation sites for every individual.

Three hierarchical C++ classes referred to as LocusInfo, ChromosomeInfo and GenomeInfo were created to define genetic characteristics at locus, chromosome and genome levels. These classes allow user-defined parameters such as allele frequencies, map positions, number of loci, chromosome lengths, numbers of chromosomes and mutation rates. Values for these parameters can also be generated randomly.

To allow complex pedigree structures, three classes Animal, Cohort and Population are created to store and pass information on collections of individuals. Complex mating structures such as cross breeding, overlapping generations and arbitrary user-defined pedigrees are straightforward. Real founder genotypic data instead of limited user-defined parameters can be used in GenSim to mimic complex genomes.

Results and Discussion

Both C++ and Julia versions of GenSim are available on request (haocheng@iastate.edu).

Acknowledgments

Hao Cheng is funded by Endowment of the Lush Chair.