BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 18 2014, pages 2686–2688
doi:10.1093/bioinformatics/btu206

Genetics and population analysis

Advance Access publication May 28, 2014

AlphaMPSim: flexible simulation of multi-parent crosses

John M. Hickey[1,*], Gregor Gorjanc[1], Sarah Hearne[2] and Bevan E. Huang[3]

[1]The Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK, [2]Genetic Resources Program, International Maize and Wheat Improvement Center (CIMMYT), Apdo. 06600 México D.F. and [3]CSIRO Computational Informatics and Food Futures National Research Flagship, Dutton Park, QLD 4001, Australia

Associate Editor: Janet Kelso

ABSTRACT

Summary: Multi-parent crosses of recombinant inbred lines exist in many species for fine-scale analysis of genome structure and marker–trait association. These populations encompass a wide range of crossing designs with varying potential. AlphaMPSim is a flexible simulation program that is efficiently designed for comparison of alternative designs for traits with varying genetic architectures and biallelic markers with densities up to full sequence. A large pool of founder haplotypes can be supplied by the user, or generated via integration with external coalescent simulation programs such as MaCS. From these, diverse founders for multi-parent designs can be generated automatically, and users can compare designs generated from diverse pedigrees. Full tracking of identity by descent status of alleles within the pedigree is undertaken, and output files are compatible with commonly available analysis packages in R.

Availability and implementation: Executable versions of AlphaMPSim for Mac and Linux and a user manual are available at http://www.roslin.ed.ac.uk/john-hickey/software-packages/.

Contact: john.hickey@roslin.ed.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 10, 2013; revised on April 14, 2014; accepted on April 15, 2014

1 INTRODUCTION

Multi-parent recombinant inbred line (RIL) crosses offer the ability to capture a wide range of genetic and phenotypic diversity in a controlled, balanced family structure with increased mapping resolution. Their advantages have led to their implementation in several species, including mice (Complex Trait Consortium, 2004), maize (McMullen *et al.*, 2009), Arabidopsis (Kover *et al.*, 2009), wheat (Huang *et al.*, 2012), rice (Bandillo *et al.*, 2013), barley and oats.

Developing these populations takes time and is expensive. As differences in design will affect not only the resources required for creation but also the eventual power and resolution, optimization of the design is an important consideration in planning a multi-parent study. Simulation is a powerful approach for comparing alternative designs. Software packages in R for simulating multi-parent RIL crosses—qtl (Broman *et al.*, 2003) and mpMap (Huang and George, 2011)—are available, but have not been designed for simulation of sequence data and application to polygenic traits controlled by large numbers of quantitative trait loci (QTL). In contrast, we have developed AlphaMPSim as a flexible simulation software package for generating dense genotypes in large populations. Generated data can be easily linked to standard analysis software to answer questions about design and power of future studies. This versatile software helps meet the urgent need for computationally efficient, flexible and easy-to-use tools capable of handling the complexities associated with multi-parent inbred line crosses.

2 METHODS

The process of data generation in AlphaMPSim involves five steps:

(1) generation of a population of founder haplotypes; (2) selection of founders as parents of the experimental cross; (3) gene-dropping parental haplotypes through the pedigree; (4) generation of trait values; (5) reduction from sequence to single nucleotide polymorphism (SNP) data.

First, AlphaMPSim generates a population of founder haplotypes representing a diversity panel of accessions/strains. Two options exist for this. The default option is that AlphaMPSim calls MaCS (Chen *et al.*, 2009), which uses a coalescent model to simulate haplotypes. MaCS is fully flexible with regard to historical population size, mutation and recombination rates, genome size and number of founders generated. The user can also override the default, allowing external founder haplotypes from real data or simulated by other means to be read in and used.

Second, parents for the multi-parent design are selected from this population. Five options exist for this: (i) random selection; (ii) selection to maximize genetic diversity; (iii) selection on the basis of phenotypic value or true breeding value; (iv) any combination of these; (v) or any other user-specified criteria. Option (iv), for example, allows simulation of the scenario where selected parents are diverse elite lines, chosen to maximize genetic diversity among founders with high phenotypic values.

Third, the haplotypes of the selected parents are dropped through the user-supplied pedigree. Any number, type and size of pedigree can be supplied. The ability to supply multiple pedigrees and select the founders in multiple ways can be invoked within the same analysis. Hence, different methods of founder selection and different designs can be compared on the basis of identical underlying genetic architecture in terms of traits.

Fourth, QTL are selected at random to generate phenotypes and true breeding values. Loci segregating in the population generated in Step 1 are selected from the simulated sequence data. Effects for these loci are sampled from either a Gaussian or Gamma distribution. Phenotypic values are generated by combining the true breeding values with a residual Gaussian error. The heritability of the phenotypes is calculated relative to the breeding values of the founders selected to be the parents of the multi-parent cross. The user has full flexibility in specifying the

*To whom correspondence should be addressed.

number of QTL generated, the parameters of the effect distributions and the overall heritability.

Fifth, biallelic loci segregating in the population generated in Step 1 are selected at random from the simulated sequence data. Note that before this, all steps involve the full sequence data. Any number of SNP genotyping platforms can be generated with any level of SNP density. Full sequence data can also be generated. The identity by descent (IBD) status along the whole genome is tracked for all individuals within the pedigree.

3 DATA INPUT AND OUTPUT

Default data inputs to AlphaMPSim include a pedigree file and parameter file. Functions available in R/mpMap (Huang and George, 2011) can be used to generate pedigrees for a variety of multi-parental designs. A file specifying recombination hotspots can also be added if desired.

The simulated data are output to several files that specify the SNP genotypes for each observed individual, genotypes for each observed individual at each QTL locus, QTL effect sizes and positions, the true breeding values and phenotypes for each trait and files with physical and genetic map. Additionally, the user can optionally output files with the IBD status along the whole genome and whole-genome sequence information. Files that track the IBD using sparse representation are always produced.

We anticipate that typical uses of the package will be to generate multiple datasets to be analysed using standard tools for multi-parental sequence (e.g., http://mus.well.ox.ac.uk/19genomes/magic.html). Although this is straightforward for advanced users, we have included scripts to convert output to formats used in common analysis software for these populations.

4 IMPLEMENTATION

AlphaMPSim was primarily written in Fortran 95 as a standalone program. Because the program makes use of some Unix system and shell script calls, executable versions are only available for Linux and Mac. Full simulation of a multi-parent population requires the program to be run twice. The first run performs Step 1; after this, an easy-to-use R script performs Step 2; and finally the second run performs Step 3. To enable AlphaMPSim to be run in a pipeline involving several alternative design scenarios, a flexible shell script that is easily used and modified is also supplied.

AlphaMPSim can simulate data (e.g. sequence data for pedigrees with 100 000 individuals) because sequence data are efficiently stored when they are needed in memory and otherwise efficiently stored in binary files. Efficient storage of the full sequence data capitalizes on the fact that the sequence data generated contains only biallelic SNP that are represented as 0/1, and strings of 0 s and 1 s are binary numbers that represent integers. Therefore, each founder haplotype can be compactly represented by a string of integers. This has the benefit of packing the utilized memory more fully and reducing the number of operations required to process each locus because we process blocks of alleles rather than individual alleles. Housekeeping subroutines required for packing and unpacking each block of alleles and for modelling the recombination process were taken from AlphaDrop (Hickey and Gorjanc, 2012), a simulation program for livestock populations. Additionally, only ever one chromosome at a time is held in memory.

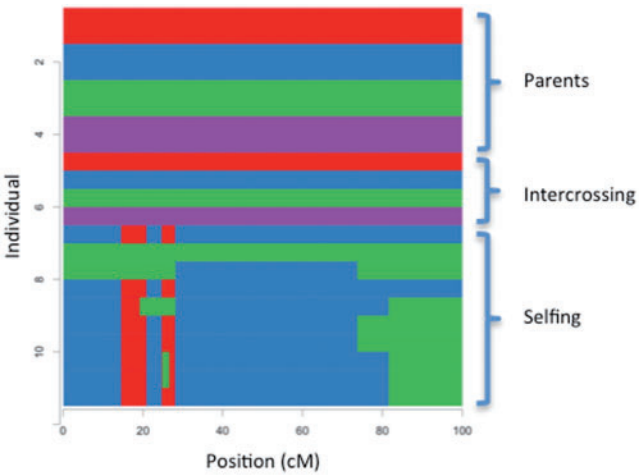

Fig. 1. Visualization of the IBD status of a single chromosome in 11 individuals from a pedigree supplied in the distribution representing the intercrossing of four inbred parents, one generation of random intermating and four generations of selfing. A recombination hotspot was simulated in the first third of a chromosome

An efficient system to track the IBD status along the whole genome for all individuals within the pedigree was developed. This involves recording an indicator variable for each gamete of an individual that identifies whether an individual inherited its parent top or bottom gamete at the first position of each chromosome, the number of recombinations on each gamete and the positions of these recombinations. Combined with the pedigree, this information suffices to recursively track complete IBD status for any individual. An easy to use R-script IBDPlot.R to visualize this status as in Figure 1 is supplied with the program, but this requires a memory-intensive processing of the efficient IBD tracking.

5 COMPUTATIONAL REQUIREMENTS

To benchmark AlphaMPSim, data were simulated for 5040 multi-parent advanced generation intercross (MAGIC) lines with eight parents. The parental genomes were combined through three generations of intercrossing in a single funnel before progressing through three generations of random intermating and 14 generations of selfing to produce RILs. In total, this resulted in a pedigree of 100 812 individuals. The genome comprised 30 chromosomes, each one Morgan in length. The MaCS parameters relating to the historical effective population size, mutation and recombination rates resulted in a total of 1 548 600 segregating sites across the genome, of which 60 000 were chosen to be SNP and 3000 to be QTL. Computations were performed on a single core of a Linux server on which each node had a dual core Intel Xeon Westmere processor with six threads running at 2.93 GHz. For most practical applications, full sequence information and IBD status are not

required, and even for this large population, the program requires just <10 h running time and 3GB RAM. Writing out all information increases the running time to 14 h. Although this would be computationally onerous for simulation of many datasets, we expect most pedigrees of interest, and hence running times, to be substantially smaller because of fewer generations of selfing and fewer individuals. With deep pedigrees, the memory requirement is large for parsing IBD status using the efficient tracking method. Options to reduce this are currently explored.

6 DISCUSSION

We have described a new computer package for the simulation of marker, sequence and trait data from complex multi-parent RIL crosses. The program is computationally efficient, flexible and easy to use. It makes the simulation of large datasets feasible for such populations. Genomic sequence data, and in particular low-coverage sequencing data, will be increasingly used in multi-parent crosses. The computational efficiency of AlphaMPSim will enable researchers to evaluate the power of alternative low-coverage sequencing strategies in advance of investing in such data.

Funding: This work was supported by the Australian Research Council [DE120101127 to B.E.H.]. J.H. would like to acknowledge the support of the CGIAR Generation Challenge Program and ICRISAT for funding this work in part. J.H., G.G., and S.H. would like to acknowledge the support of the Seeds of Discovery and CIMMYT.

Conflict of Interest: none declared.

REFERENCES

Bandillo,N. *et al.* (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice*, **6**, 11.

Broman,K.W. *et al.* (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.

Chen,G.K. *et al.* (2009) Fast and flexible simulation of DNA sequence data. *Genome Res.*, **19**, 136–142.

Complex Trait Consortium. (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.*, **36**, 1133–1137.

Hickey,J.M. and Gorjanc,G. (2012) Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. *G3 (Bethesda)*, **2**, 425–427.

Huang,B.E. and George,A.W. (2011) R/mpMap: a computational platform for analysis of multi-parent recombinant inbred lines. *Bioinformatics*, **27**, 727–729.

Huang,B.E. *et al.* (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnol. J.*, **10**, 826–839.

Kover,P.X. *et al.* (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.*, **5**, e1000551.

McMullen,M.D. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.