Zhang et al. BMC Genetics (2015) 16:10
DOI 10.1186/s12863-015-0173-4

BMC Genetics

SOFTWARE Open Access

GPOPSIM: a simulation tool for whole-genome genetic data

Zhe Zhang[1†], Xiujin Li[2†], Xiangdong Ding[2*], Jiaqi Li[1] and Qin Zhang[2]

Abstract

Background: Population-wide genotypic and phenotypic data is frequently used to predict the disease risk or genetic/phenotypic values, or to localize genetic variations responsible for complex traits. GPOPSIM is a simulation tool for pedigree, phenotypes, and genomic data, with a variety of population and genome structures and trait genetic architectures. It provides flexible parameter settings for a wide discipline of users, especially can simulate multiple genetically correlated traits with desired genetic parameters and underlying genetic architectures.

Results: The model implemented in GPOPSIM is presented, and the code has been made freely available to the community. Data simulated by GPOPSIM is a good mimic to the real data in terms of genome structure and trait underlying genetic architecture.

Conclusions: GPOPSIM would be a useful tool for the methodological and theoretical studies in the population and quantitative genetics and breeding.

Keywords: Data simulation, SNP, Pedigree, Multiple traits, Mutation-drift equilibrium, Genetic correlation

Background

Single nuclear polymorphism (SNP) markers are widely implemented in the investigation of human genetics and animal/plant breeding, due to its high abundance and extensive coverage across the whole-genome. They were usually used to predict the disease risk in human [1,2], to localize genetic variations responsible for complex traits through genome wide association study (GWAS) [3], and to predict the genetic values of economically important traits in plant and animal breeding [4,5]. The techniques and methodologies related to this discipline are moving fast, and these new methods need to be evaluated before implemented to real data. The most efficient way for such kind evaluation is computer simulation.

Data simulation has been employed in genetic analysis for decades. Recently, many novel findings in genomic prediction using simulated whole-genome data were reported [6,7]. The most commonly used model for whole-genome genotypic data simulation is the mutation-drift equilibrium (MDE) model [8]. However, the rules applied in the MDE model vary in different studies, which made results from different studies incomparable. Meanwhile, only independent traits could be simulated by most programs, and function of simulating multiple correlated traits are seldom to be developed.

We present GPOPSIM: a simulation tool for population genetic data based on MDE. The mechanism to create polymorphic markers, population structure, and trait phenotypes were detailedly proposed. Moreover, simulating multiple genetically correlated traits were explored as well. In order to demonstrate the performance of our program, a series of implementation were carried out in this study.

Implementation

In this section, we describe the implementation of the method from [9] in the presented software GPOPSIM. The software can be compiled and executed in multiple platforms, and driven by a parameter file. The parameter setting is illustrated in Table 1 and more details could be found in the project home page (https://github.com/SCAU-AnimalGenetics/GPOPSIM).

The simulation of whole-genome genotypes is based on the MDE model [8]. It starts from an initial population,

* Correspondence: xding@cau.edu.cn
†Equal contributors
[2]Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China
Full list of author information is available at the end of the article

Table 1 Parameter setting

Category	Parameters
Overall	population stages, number of sub populations in the current population stages, chromosome number, chromosome length
Marker	marker number per chromosome, marker distribution, mutation rate for marker& QTL
QTL	QTL effect distribution, QTL number, QTL ratio for multiple trait simulation
Trait	trait number, trait type, heritability, correlations between traits
Population setting	population size, number of sires selected, number of dams selected, number of generations, selection rule, matting rule, mutation rule

through many generations of historical population, ends to the current population. In this process, the polymorphism of markers is increased by mutation, but decreased by genetic drift, and reaches equilibrium status throughout a number of historical population, which was named mutation-drift equilibrium [10]. The whole-genome data generated in the current population can be used for data analysis. Figure 1 illustrates the workflow and acting parameter categories in each population stage.

Population structure

The populations simulated by GPOPSIM include one historical population and one or more current population(s). The population structures can be very flexible in different population stages by assigning parameters such as different population sizes, number of selected breeding male and female, the selection rules and other parameters for each population stage (Table 1, Figure 1). The population/pedigree structure of the simulated data is decided by the parameter settings of the current populations. The parameter settings for the historical population mainly affect the genome structure of the current population.

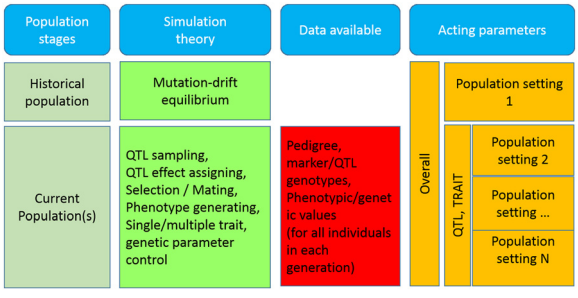

Figure 1 Workflow and parameter setting in GPOPSIM.

Genome structure

The genome structure could be clearly defined with the overall parameters and mutation rules applied in each current population. Generally, the number of chromosome and the lengths of different chromosomes are arbitrarily assigned [4,11,12], e.g. 1 Morgan for each of five chromosomes. The number of markers on each chromosome could vary, and each segment between two adjacent markers was considered to harbor a potential QTL. In GPOPSIM, the position of markers and potential QTLs were simply assumed a mixture of uniform and exponential distribution to mimic the real SNP data in currently available SNP chips [9], such as the Illumina BovineSNP50 BeadChip [13].

The polymorphic markers and the linkage disequilibrium (LD) among them are mainly created in the historical population. The expected marker heterozygosity (H_e) is $H_e = (4N_e u)(4N_e u + 1)^{-1}$ [10], where N_e is the effective population size and u is the mutation rate. And the expected LD is $r^2 \approx 1/(\alpha + kN_e c)$ [8], where α is an indicator of mutation, c is the genetic distance between markers.

Genetic and phenotypic values

Based on the genome structure generated in the historical population, the trait and QTL parameters, GPOPSIM simulates genetic and phenotypic values for each individual in the current population. The true QTLs are randomly sampled from all candidate QTLs. The true genetic effects of each true QTL are sampled from normal [1] or gamma distribution [4]. By setting different QTL number and effect distribution, a wide range of genetic architecture from simple disease traits to complex traits can be simulated easily. For each trait, the true genetic merit of one individual is defined as the cumulative effect across all true QTLs. For quantitative traits, the phenotypic value is generated by adding the true genetic merit with a random residual error, while the 0/1 phenotype is generated by setting an incidence for threshold traits.

The principles applied to single-trait data simulation can be easily extended to two or multiple genetically correlated traits simulation. For two traits simulation, more flexible parameters and rules can be applied. All true QTLs affecting both traits are divided into three groups: (1) Group1 is a group of true QTLs simultaneously affecting both traits, in which their effects are sampled from a multivariate normal distribution or a gamma distribution [14], (2) the true QTLs in Group2 and Group3 affect only one of the two traits, respectively, for which the effect of each causative locus in Group2 and Group3 is sampled from a normal or gamma distribution. The genetic correlation between two traits ranged from −0.88 to 0.88, which can basically cover the genetic correlated traits.

Random residual errors are sampled from a multivariate normal distribution. Similarly, the phenotypic value and genetic merit of one individual on both traits are generated as the single trait module does. Considering the sampling error of simulation, the expected genetic correlation (r_g) of two traits is evaluated and provided by GPOPSIM according to the formula [15]

$$r_{AB} = \sum 2p_i(1-p_i)\alpha_{Ai}\alpha_{Bi} / \left(\sqrt{\sum 2p_i(1-p_i)\alpha_{Ai}^2} * \sqrt{\sum 2p_i(1-p_i)\alpha_{Bi}^2} \right) \quad (1)$$

where p_i is the frequency of one of two alleles for the locus i; α_{Ai} is the effect of the locus i for Trait A; α_{Bi} is the effect of the locus i for Trait B. For multiple traits simulation, all true QTLs are assumed to affect all traits simultaneously for simplicity and their effects are sampled from a multivariate normal distribution with the restriction of assigned genetic correlations [16].

Input and output files

Only one input file, also being the parameter file is needed to run GPOPSIM (Table 1). Generally, GPOPSIM generates four types of output files: (1) a data file including pedigree information, the individuals and their parents identities, and the simulated true genetic value and phenotypic value for each trait and each individual; (2) marker genotype file providing the marker genotypes in phased format; (3) QTL genotype file providing the true QTL genotypes; and (4) several separate parameter files include a marker map file, a true QTL map file including their simulated true QTL effects, and a genetic parameters file. All these files are in text format with the file extension of '.txt'. And the first three types of files are generated for each generation with a filename including the number of generation.

Source code and software availability

Based on the methods described above and in [9], we developed a whole-genome data simulation software GPOPSIM in Fortran 90 and tested on Microsoft Windows (version XP/7/8), and Linux (Red Hat Enterprise, Ubuntu, Fedora). It can simulate population with various population structure, genomic data, one or more independent/correlated continuous trait(s). The volume of simulated dataset depends on the running environment of the user's PC or server.

A series of simulations were carried out to investigate the quality of the simulated data using GPOPSIM, and the Haploview software [17] was used for data quality control and linkage disequilibrium analysis. The variance components and genetic correlations were estimated by DMU [18].

Results and discussion

We describe the quality of data simulated by GPOPSIM first, and followed by a general discussion of the implementation of GPOPSIM.

Besides the features predefined by users, e.g. marker density, minor allele frequency (MAF) and LD can typically reflect the characteristic of the simulated genotypic data. Usually, MAF in the current population generated

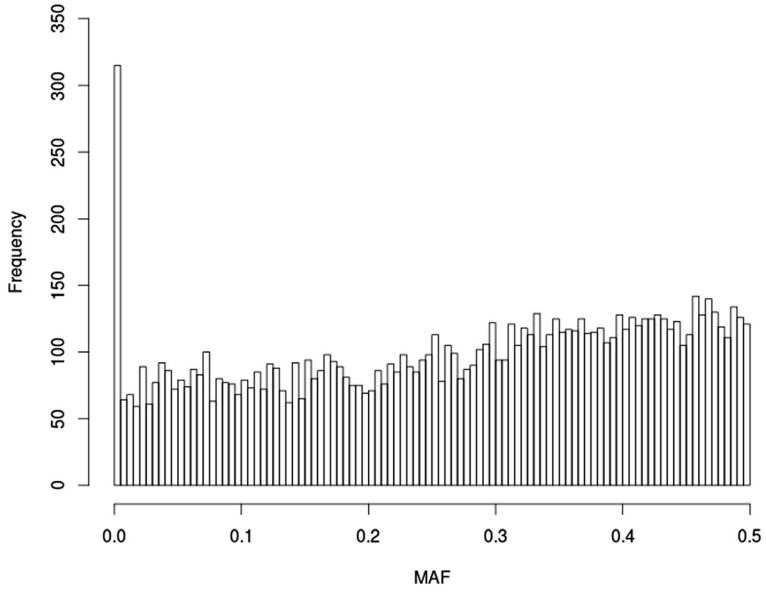

Figure 2 Distribution of the minor allele frequency (MAF) of genotypes simulated by GPOPSIM. Parameter setting for this simulation is $Ne = 100$, mutation rate $u = 2.5 \times 10^{-3}$, number of markers = 10,000.

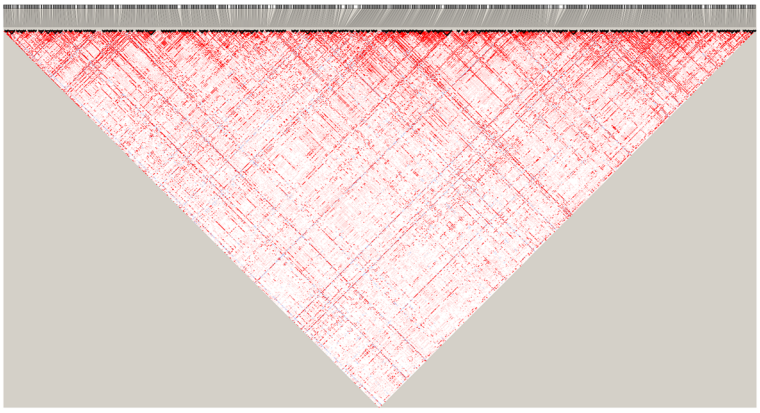

Figure 3 Pattern of linkage disequilibrium (LD) of the genotypes simulated by GPOPSIM. Parameter setting for this simulation is $Ne = 100$, mutation rate $u = 2.5 \times 10^{-3}$, number of markers = 10,000. The pairwise LD among the first 1000 markers were shown in this figure.

by GPOPSIM nearly follows an uniform distribution with a long tail near MAF = 0, which is also called "L" shape distribution of MAF, or "U" shape distribution on the entire frequency spectrum. Figure 2 shows the distribution of MAF in the scenario with $N_e = 100$ and $u = 2.5 \times 10^{-3}$, nearly 50% loci's MAF were lower than 0.3, and the average MAF was 0.28, which is similar to the average MAF in Holstein detected with Illumina Bovine50SNP BeadChip [13,19]. The average MAF and heterozygosity could be altered by increasing or decreasing the value of mutation rate u in the historical population [9].

Linkage disequilibrium is another indicator for the quality of simulated genotypic data. Figure 3 illustrates the LD pattern of simulated data in the same scenario as in Figure 2, the average LD between adjacent markers is 0.24. High LD can be observed in both long range and short range (Figure 3), additionally, haplotype blocks can be found as well, these fit the real data very well [19].

We assessed the two-trait phenotypic data simulated by GPOPSIM by comparing the assigned and estimated genetic parameters on condition that partial common QTLs affect both traits. We set two genetically correlated traits (denoted as Trait A and Trait B) with heritability of 0.1 and 0.3, the genetic correlation between trait A and B was assigned 0, 0.2, 0.5 and 0.8, and the environmental correlation was set 0. From the results of 20 replicates of simulation (10,000 individuals in each replicate) (Table 2), we can see that the heritability estimated by DMU are very close to the assigned values in different levels of genetic correlations and the estimation vary in a very small range among replicates. Likewise, the estimations of genetic correlation from DMU are acceptable and close to those assigned, in addition, these estimations are also nearly same as those expected genetic correlations, which are calculated from equation 1. This indicates that GPOPSIM can be an ideal tool for simulating multiple traits with/without genetic correlation. The bias with the preset genetic correlations is acceptable. Besides, the estimates of variance components at all levels of genetic correlation fit the assigned values very well (Table 2).

GPOPSIM is distributed both as Fortran 90 source code and as executable procedure on Windows and Linux platform (https://github.com/SCAU-AnimalGenetics/GPOPSIM or Additional file 1). It is free of charge for all purpose users and no license is required. The computing time and RAM demanding on PC, with 3.0 GHz CPU, 2 GB RAM is 4.4 minutes and 8 Mb, respectively, for simulating 10000 markers, 1000 historical generations with

Table 2 The assigned and estimated heritability (h^2), genetic correlation (r_g) and residual correlation (r_e) for two trait phenotypes simulated by GPOPSIM

h^2		r_g			r_e
Estimated A	**Estimated B**	**Assigned**	**Expected**	**Estimated**	**Estimated**
0.101(0.011)	0.289(0.022)	0.0	0.000(0.000)	−0.004(0.092)	0.003(0.016)
0.100(0.009)	0.302(0.025)	0.2	0.180(0.046)	0.159(0.103)	−0.002(0.010)
0.100(0.012)	0.299(0.022)	0.5	0.506(0.045)	0.493(0.079)	0.004(0.011)
0.104(0.012)	0.290(0.024)	0.8	0.805(0.042)	0.805(0.078)	0.000(0.015)

The assigned heritability is 0.1 and 0.3 for trait A and B, respectively; the assigned residual correlation is 0; the mean (S.D.) of estimated genetic parameters were obtained from DMU and calculated from 20 replicate of simulations.

Ne = 100. The time demanding increased nearly linearly with the effective population size N_e, number of markers N_m and number of generations N_g.

GPOPSIM is designed for, but not limited to, data simulation in genetic or breeding researches that needs genomic and phenotypic data from a population, such as genome wide association study, whole genome prediction, population genomics studies, and genomic selection breeding program. Though GPOPSIM has been successfully implemented in our previous studies [11,20,21], there is still rooms for further extension, such as sequences data simulation.

Conclusions

We presented GPOPSIM, a simulation tool for pedigree, phenotypes, and genomic data, with a variety of population and genome structures and trait genetic architectures. It enables users to simulate (1) various genome structures via mutation drift equilibrium model with user defined historical population parameters; (2) pedigree from one or more current population(s) with flexible user assigned population structure parameters; (3) phenotypes on single or multiple traits with/without desired genetic correlation and genetic architectures. GPOPSIM is designed for, but not limited to, data simulation in genetic or breeding researches that needs genomic and phenotypic data from a population, such as genome wide association study, whole genome prediction, population genomics studies, and genomic selection breeding program. The software can run on multiple platforms and the code has been made freely available to the community. We speculated that this software could promote the methodological and theoretical studies in the discipline of population and quantitative genetics and breeding.

Availability and requirements

Project name: GPOPSIM
Project home page: https://github.com/SCAU-Animal-Genetics/GPOPSIM
Operating system(s): Compiled for Windows and Linux
Programming language: Fortran 90
Other requirements: None
License: None
Any restrictions to use by non-academics: None

Additional file

Additional file 1: A compressed file includes the GPOPSIM resource code (GPOPSIM.f90) and the parameter file for GPOPSIM (para.txt).

Abbreviations

GPOPSIM: Genome-wide population simulation; SNP: Single nuclear polymorphism; GWAS: Genome wide association study; GS: Genomic selection; MDE: Mutation-drift equilibrium; MAF: Minor allele frequency; LD: Linkage disequilibrium.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZZ, LXJ and XDD designed and developed the software, contributed with software specification, have been expert test users throughout the development phase, and drafted the manuscript. QZ and JQL initiated and led the project. All authors have read and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31200925, 31272418, 31371258), the Program for Changjiang Scholar and Innovation Research Team in University (Grant No. IRT1191), the earmarked fund for China Agriculture Research System (CARS-36), Beijing City Committee of Science and Technology Key Project, and the Ph.D. Programs Foundation (the Doctoral Fund) of Ministry of Education of China (20124404120001).

Author details

[1]Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China. [2]Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China.

Received: 23 October 2014 Accepted: 22 January 2015
Published online: 05 February 2015

References

1. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3(10):e3395.
2. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev. 2008;18(3):257–63.
3. Wellcome-Trust-Case–control-Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
4. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
5. Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. Crop Sci. 2009;49(1):1–12.
6. Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics. 2010;185(2):623–31.
7. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185(3):1021–31.
8. Sved JA. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol. 1971;2(2):125–41.
9. Zhang Z, Ding X, Liu J, Guiyan N, Li J, Qin Z. Whole-Genome Genetic Data Simulation Based on Mutation-Drift Equilibrium Model. In: The proceedings of the Proceedings of the 2012 4th International Conference on Computer Modeling and Simulation: 17–18 February 2012; Hong Kong. 2012. p. 87–93.
10. Kimura M, Crow JF. The number of alleles that can be maintained in a finite population. Genetics. 1964;49:725–38.
11. Zhang Z, Liu JF, Ding XD, Bijma P, de Koning DJ, Zhang Q. Best linear unbiased prediction of genomic breeding values using trait-specific marker-derived relationship matrix. PLoS One. 2010;5(9):e12648.
12. Calus MP, Meuwissen TH, de Roos AP, Veerkamp RF. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 2008;178(1):553–61.
13. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009;4(4):e5350.
14. Hayashi T, Iwata H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics: 2013;14: 34.
15. Falconer DS, Mackay TFC. Introduction to Quantitative Genetics. 4th ed. New York: Longman; 1996.
16. Calus MPL, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 2011;43:26.

17. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263–5.
18. Madsen P, Sørensen P, Su G, Damgaard LH, Thomsen H, Labouriau R. DMU - a package for analyzing multivariate mixed models. In: the proceedings of the 8th World Congress on Genetics Applied to Livestock Production; Brasil. 2006.
19. Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, et al. The pattern of linkage disequilibrium in German Holstein cattle. Anim Genet. 2010;41:346–56.
20. Zhang Z, Ding XD, Liu JF, Zhang Q, de Koning D-J. Accuracy of genomic prediction using low density marker panels. J Dairy Sci. 2011;94(7):3642–50.
21. Wang CL, Ding XD, Wang JY, Liu JF, Fu WX, Zhang Z, et al. Bayesian methods for estimating GEBVs of threshold traits. Heredity. 2013;110(3):213–9.