2012 4th International Conference on Computer Modeling and Simulation (ICCMS 2012)
IPCSIT vol.22 (2012) © (2012) IACSIT Press, Singapore

Whole-Genome Genetic Data Simulation Based on Mutation-Drift Equilibrium Model

Zhang Zhe[1,2], Ding Xiangdong[1]*, Liu Jianfeng[1], Ni Guiyan[1], Li Jiaqi[2], Zhang Qin[1]

[1] Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

[2] Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

Abstract. Whole-genome data simulation is popular in genomic research, especially in methodology development for whole-genome data analysis. In this research, the frequently used mutation-drift equilibrium (MDE) model was employed to simulate whole-genome data. Based on the MDE model, we proposed a mixture distribution to sample the positions of loci and a rule to recode multi-allelic markers to bi-allelic markers. The effects of effective population size and mutation rate in historical generations were assessed. GPOPSIM, a program using the rules proposed in this paper, was developed to simulate whole-genome data. Data simulated via GPOPSIM was compared with data from the Illumina Bovine SNP chip. Results show that the MDE model is a recommendable model for the whole-genome data simulation. The proposed strategy effectively increased the efficiency of simulation program.

Keywords :Monte carlo; Whole-genome approaches; Mutation-drift equilibrium; Genomic selection; SNP

1. Introduction

With the advances of molecular bio-technology, genetic markers have been widely implemented in investigation of human genetic diseases and animal and plant breeding. The frequently used genetic markers are restrictive fragment length polymorphism (RFLP), microsatellite (short tandem repeat, STR), single nuclear polymorphism (SNP), and copy number variation (CNV). Among them, the SNP markers are most frequently used because of its high abundance and extensive coverage across the whole-genome. It was reported that there are over 3.1 million SNPs in human genome [1] and 2.8 million SNPs in chicken genome [2]. These high density SNP markers have been widely used to predict the disease risk [3, 4], to localize genetic variations responsible for complex traits through genome wide association study (GWAS) [5], and to predict the genetic values [6, 7] of economic-important traits in plant and animal breeding. However, the related techniques and methodologies are still under development and any new techniques or methods need to be evaluated before they are implemented to analyze real data. The most popular way for such kind evaluation is using 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 [8, 9]. However, a detailed description for genomic data simulation is not available. The most commonly used model for whole-genome genotypic data simulation is the mutation-drift equilibrium (MDE) model [10]. However, the rules applied in the MDE model vary in different studies, which made results from different studies incomparable. Nevertheless, the differences between the strategies currently used in the MDE model and key factors affecting the simulated data are yet to be investigated.

The rules for genomic data simulation were detailedly proposed in the present study. As data simulation is usually a part of 'Materials and Methods' in most research paper, the rules and strategies for genomic data simulation are not described and compared in such detailedly. The investigation of these key parameters in genomic data simulation based on MDE model could help researchers to comprehensively understand the results from different simulation studies. Therefore, this study might increase the easiness of interpreting and understanding results from simulated genomic data.

The objectives of this study were developing a novel genomic data simulation approach based on the mutation-drift equilibrium model, and investigating the key factors affecting the simulated data. The rules to create polymorphic markers and a mixture distribution to sample marker position were detailedly proposed. Furthermore, the simulated genotypic data was compared with data from the Illumina Bovine SNP chip.

2. Materials and methods

Simulation of whole-genome data based on the MDE model usually starts from an initial population, through many historical generations, ends to recent generations. In this process, the polymorphism of markers is increased by mutation, but decreased by drift, and will reach its equilibrium status throughout the historical generations, which was named mutation-drift equilibrium [11]. The whole-genome data generated for the recent generations can be used for data analysis. For whole-genome data simulation, the following issues should be taken into consideration.

Population structure: Generally, the simulation of historical generations is relatively simple. The population across all generations is assumed practicing random mating, and mutations constantly introduce new variation and genetic drift shifts the variation to fixation. After many historical generations, recent generations are simulated, where mutation will be ignored. The simulation of recent generations could be very complicated in order to produce appropriate data, which depends on the population structure parameters.

Population structure parameters include number of generations (*Ng*), population size of each generation, ratio of male to female, mating design, reproduction rate and selection design, etc. The choice of these parameters is flexible. Different parameter combinations can be made to meet the requirement of study. Recent generations normally start from a base population as the conventional data simulation [12]. In recent generations, individuals and their corresponding information are produced according to specific population structure parameters. The pedigree, marker genotypes, genetic values and phenotypic values can be recorded for each individual in these generations. This is different from historical population, where only the polymorphic markers are created and not necessary to be recorded.

Genomic structure: The genomic structure should be clearly defined in the whole-genome data simulation. The parameters related to the genomic structure are number of chromosomes (N_c), length (in Morgan) of each chromosome (L_c), total number of markers (N_m), distribution of markers on the chromosomes and number of quantitative trait loci (N_{qtl}). Generally, the lengths of different chromosomes are set to be identical, e.g. 1 Morgan for each chromosome. The number of markers on each chromosome could vary.

So far, there is no a general rule for marker distribution in whole-genome data simulation studies. The markers were simply assumed either evenly or randomly distributed on chromosomes in different studies. Actually, the even distribution assumption is too simple to mimic the real data because markers in all currently available SNP chips are not actually evenly spaced. Based on the investigation of maker distribution of Illumina BovineSNP50 BeadChip [13], we suggest to use the following mixture distribution to sample the marker intervals.

$$d = wd_mR_{exp}(1) + (1\text{-}w)U(0, 2d_m), \qquad [1]$$

where w is a weighting factor, dm = LcNc /Nm is the average marker interval, Rexp(1) is a random number sampled from an exponential distribution with parameter 1, and $U(0,\ 2d_m)$ is a random number sampled from an uniform distribution. The expectation and variance of this mixture distribution are d_m and $d_m^2(1 - 2w + 4w^2)/3$, respectively.

Two strategies can be used to decide the number of chromosomes. One is to simulate the total length of all chromosomes as nearly long as a real one. For example, a length of 30 Morgans was simulated by VanRaden et al. [14], which is very close to the actual bovine genome length reported [15]. However, such kind of simulation will take long time and is computationally too demanding in further analysis. Therefore, the alternative strategy is that the number of chromosomes can be arbitrarily assigned according to the needs of computation. The length of whole genome was often simulated to be 3 to 10 Morgans in most studies [6, 16].

Marker genotype: The rule used to create polymorphic markers is one of the most critical points in whole-genome genotypic data simulation. Based on the MDE model, two rules could be used in the simulation of historical generations: simulating bi-allelic markers directly or simulating multi-allelic markers in the historical generations first, then recode them to bi-allelic markers in the recent generations. The former is simple and easy but low in efficiency. Therefore, we employed the second rule in this study, and there are three related issues need to be taken into consideration:

Rule to initialize the genotypes: The initial population is the starting point of the simulation. Following the argument of Kimura and Crow [11], the expected heterozygosity (H_e) of a marker in a population in equilibrium status is

$$H_e = (4N_eu)(4N_eu+1)^{-1}, \qquad [2]$$

where N_e is the effective population size and u is the mutation rate. According to equation 2, the heterozygosity in a population in equilibrium is not influenced by the heterozygosity status in the initial population. In other words, the marker genotypes in the initial population could be assigned to be either monomorphic [16] or polymorphic [6].

Rule to create mutant: In the historical generations, the polymorphism of markers mainly arises from mutation, which, together with genetic drift, Mendel's law and recombination, determines the genotypes of an offspring. The genetic drift and recombination depend on the effective population size (N_e) and marker interval (d), respectively. Hence, the only parameter needs to be set here is the mutation rate u. The empirical mutation rate can be calculated from equation [2] for a desired heterozygosity. For example, if the heterozygosity of 0.5 is desired with N_e of 100, the empirical mutation rate could be 2.5×10^{-3}. In order to create multi-allelic markers, each mutant on each locus in each generation should be coded as a new allele.

Rule to recode genotypes: In the last generation of the historical generations, all loci are multi-allelic. To obtain the bi-allelic markers like SNPs, one common procedure is applied to recode the genotypes at each locus, i.e., the allele with the highest frequency is recoded as mutant allele, and all other alleles are recoded as the original allele. Once recoding is finished, mutation is not allowed in the successive generations. Because recoding all other mutants to be the original type equals reverse mutation, reverse mutation could be avoided during the historical generations.

Genetic values and phenotypic values: The genotypic data with all kinds of populations or genomic structures can be simulated according to the rules mentioned above. In addition to the genotypic data and pedigree, genetic merit and/or phenotypic values are also required for most whole-genome researches. The rules to generate these values usually vary due to the assumption of the genetic architecture underling the trait of interest. In case of simple disease traits, one or few loci can be treated as the causative loci, and the disease status of each individual can be defined by the genotype of these causative loci. For complex traits, which are controlled by tens to hundreds of causative loci with minor effect, the effect of each causative locus conforms a distribution, e.g. normal distribution [3] or gamma distribution [6]. The true genetic merit of one individual is the cumulative effect across all causative loci, and the phenotypic value is generated by adding the true genetic merit to a random residual error.

Program: Based on all the rules described above, we developed a whole-genome data simulation software in Fortran, named GPOPSIM. A series of simulations were carried out to investigate the quality of the simulated data using GPOPSIM, and the Haploview software [17] as used to do the data quality control and linkage disequilibrium analysis.

3. Results and discussion

Using the rules described above and the program GPOPSIM, we simulated genomic data with total genome length = 1000 Mb, total number of markers = 20 000 (leading to a average length of marker intervals d_m = 50kb), variable effective population size (N_e = 5, 50, 100, 200 and 500), variable mutation rate (u = 2.5×10^{-1}, 2.5×10^{-2}, 2.5×10^{-3}, 2.5×10^{-4} and 2.5×10^{-5}), and variable weighting factors for the mixture distribution of marker positions (w = 0.2, 0.5, 0.8 and 1.0, see Equation [1]). The population undergoes a total of 2001 generations, of which the first 2000 generations are historical generations and the last one generation is recent generation. The features of the simulated data are presented below and compared with some real data.

Marker intervals: We proposed a mixture distribution of uniform and exponential distribution for marker positions in our simulation program, where the weights of these two distributions are adjusted by a weighting factor *w*. Figure 1 showed the distribution of marker intervals given different *w*. Taking 50kb as the average length of marker intervals, the lengths of marker intervals follow a ‘L’ shape distribution. In case of low *w*, e.g. 0.2, the lengths of marker intervals more centralize on the assigned average length of 50kb with very small variation and no intervals shorter than 50kb. It is not realistic to the marker distributions of currently available SNP panels, where the lengths of marker intervals vary greatly. With the increase of *w*, the variation of marker interval increases and more shorter marker intervals show up. In particular, with *w* equal to 0.8, the distribution of marker intervals is very similar to the SNP distribution of the Illumina BovineSNP50 BeadChip as plotted in Figure 2 and as illustrated by Matukumalli et al. [13]. It is not clear whether the distributions of marker intervals in different species follow a universal distribution and how the distribution of marker intervals affects the final analysis. However, employing a mixture distribution other than an exponential or a uniform distribution to sample the positions of loci is a significant step forward to mimic real life genotypic data.

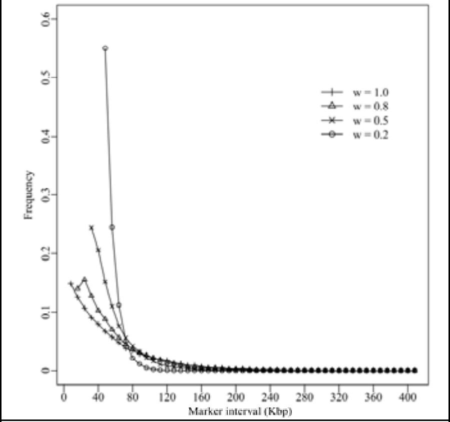

Fig 1. Distribution of marker intervals with different weighting factors (*w*) applied in the mixture distribution for sampling marker positions (total length of genome = 1 000 Mb, average length of marker interval = 50 Kb)

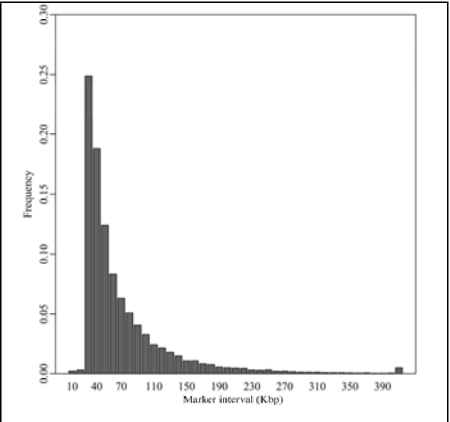

Fig 2. Distribution of marker intervals with the Illumina Bovine50SNP BeadChip.

Heterozygosity (H): The heterozygosity of markers is an important measure for polymorphism in a population. It depends on the effective population size N_e and mutation rate *u*. Figure 3 shows the observed heterozygosity (H_o) in all generations with different mutation rate *u* at N_e = 100. As expected, the observed heterozygosity H_o decreases with the decrease of *u*. Also, with the decrease of *u*, more generations are needed to reach MDE, e.g., the population will be in equilibrium after 10 generations when $u = 2.5\times10^{-1}$, while it will take nearly 1000 generations to reach equilibrium when $u = 2.5\times10^{-4}$.

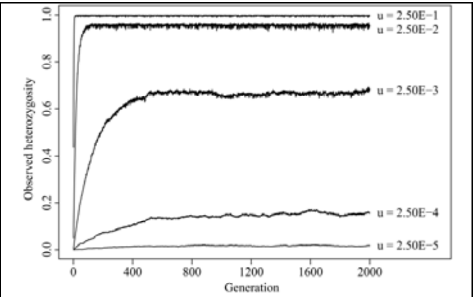

Fig 3. Observed heterozygosity (H_o) under different mutation rates (*u*) given the effective population size N_e = 100

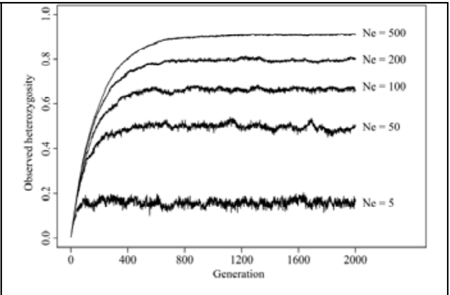

Fig 4. Observed heterozygosity under different effective population sizes (N_e) given the mutation rate $u = 2.5\times10^{-3}$

The effect of effective population size N_e on H_o given $u = 2.5\times10^{-3}$ is shown in Figure 4. The smaller N_e, the lower H_o. When N_e is increased from 5 to 500, H_o (the average from generation 1001 to generation 2000) is improved from 0.16 to 0.91. In the meanwhile, the variance of H_o decreased from 2.08×10^{-4} to 5.83×10^{-6}. This is because when N_e is small, the markers could be easily fixed due to genetic drift, resulting in low heterozygosity in the population.

Minor allele frequency (MAF): Figure 5 shows the distribution of minor allele frequency (MAF) in the base population at N_e =100 and $u = 2.5\times10^{-3}$. The MAFs of nearly 50% loci were lower than 0.3. The average MAF was 0.28 in this scenario, which is very close to the average MAF in Holstein detected with Illumina Bovine50SNP BeadChip [13, 15]. By increasing or decreasing the mutation rate u slightly, the average MAF could become higher or lower slightly (results not shown), because the heterozygosity increased or decreased as shown in Figure 3 and 4. Therefore, adjusting the mutation rate properly can create a population with desired polymorphic markers.

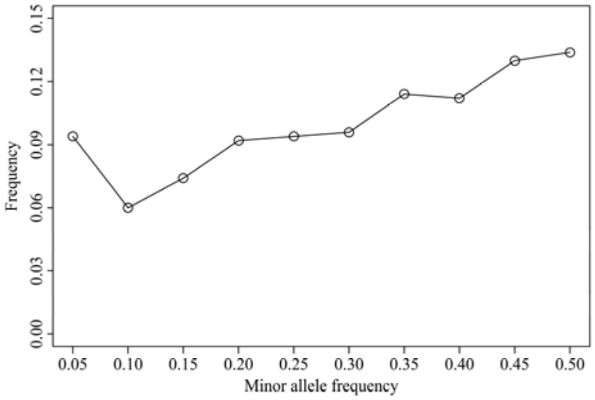

Fig 5. Distribution of minor allele frequency in the base population given the effective population size N_e = 100 and the mutation rate $u = 2.5\times10^{-3}$

Linkage disequilibrium (LD): Linkage disequilibrium can be treated as one criterion for the quality of simulated genotypic data. We employed Program Haploview [17] to analyze the linkage disequilibrium level of the simulated data for one chromosome. Figure 6 and 7 show the LD level (measure with r^2) between all adjacent marker pairs in our simulated dataset (Figure 6) and in Chinese Holstein population (Figure 7). The calculation of LD in Chinese Holstein population was based on 2093 individuals which came from Beijing area and were genotyped with Illumima Bovine50SNP BeadChip. The average LD between adjacent markers are 0.24 in both dataset. High LD are observed between markers in both short and long distance in real data and in our simulated data as well. In addition, haplotype blocks can be found as well.

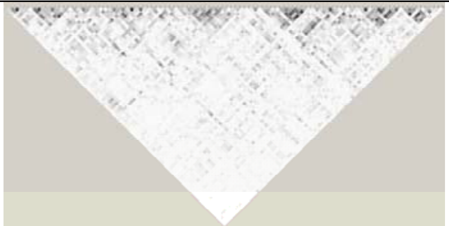

Fig 6. Linkage disequilibrium (LD) between all marker pairs on one simulated chromosomal segment (the average LD between adjacent markers r^2 = 0.24, number of markers = 100, effective population size N_e = 100, mutation rate $u = 2.5\times10^{-3}$)

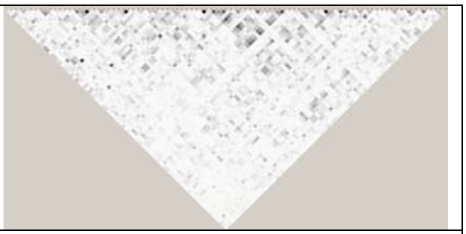

Fig 7. Linkage disequilibrium (LD) between all marker pairs on a chromosomal segment in Chinese Holstein population.

Implementation and efficiency of GPOPSIM: The program is distributed both as Fortran 90 source code and as a Windows executable (http://animalgenetics.cau.edu.cn/gpopsim/) and is free of charge for research purpose. The program is portable to multiple operation systems. The computing time and RAM demanding for simulating 5000 markers and 1000 historical generations with 100 individuals were 8 minutes (3.0 GHz

CPU, 2 GB RAM) and 5 Mb, respectively. The time demanding increased nearly linearly with the effective population size N_e, number of markers N_m and number of generations N_g.

4. Conclusions

We presented a novel whole-genome data simulation approach based on the MDE model, in particular we proposed the mixture distribution for sampling marker positions and a series of rules for generating marker genotypes. We developed a program, named GPOPSIM, based on the proposed approach. In comparison with real data, the simulated data has good quality in terms of distributions of marker intervals, marker heterozygosity, distribution of minor allele frequencies, and linkage disequilibrium pattern. The program can be used for further whole-genome simulation and methodology study.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 30800776), the National '948' Project (2011-G2A, 2010-C14) and the earmarked fund for China Agriculture Research System (CARS-36).

6. References

[1] International-HapMap-Consortium: 'A second generation human haplotype map of over 3.1 million SNPs', Nature,[J], 2007, 449, PP. 851-861

[2] International-Chicken-Polymorphism-Map-Consortium: 'A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms', Nature,[J], 2004, 432, PP. 717–722

[3] Daetwyler, H.D.; Villanueva, B.; and Woolliams, J.A.: 'Accuracy of predicting the genetic risk of disease using a genome-wide approach', PLoS ONE,[J], 2008, 3, (10), PP. e3395

[4] Wray, N.R.; Goddard, M.E.; and Visscher, P.M.: 'Prediction of individual genetic risk of complex disease', Curr. Opin. Genet. Dev.,[J], 2008, 18, (3), PP. 257-263

[5] Wellcome-Trust-Case-Control-Consortium: 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', Nature,[J], 2007, 447, (7145), PP. 661-678

[6] Meuwissen, T.H.E.; Hayes, B.J.; and Goddard, M.E.: 'Prediction of total genetic value using genome-wide dense marker maps', Genetics,[J], 2001, 157, (4), PP. 1819-1829

[7] Heffner, E.L.; Sorrells, M.E.; and Jannink, J.-L.: 'Genomic selection for crop improvement', Crop Sci.,[J], 2009, 49, (1), PP. 1-12

[8] Meuwissen, T.; and Goddard, M.: 'Accurate prediction of genetic values for complex traits by whole-genome resequencing', Genetics,[J], 2010, 185, (2), PP. 623-631

[9] Daetwyler, H.D.; Pong-Wong, R.; Villanueva, B.; and Woolliams, J.A.: 'The impact of genetic architecture on genome-wide evaluation methods', Genetics,[J], 2010, 185, (3), PP. 1021-1031

[10] Sved, J.A.: 'Linkage disequilibrium and homozygosity of chromosome segments in finite populations', Theor. Popul. Biol.,[J], 1971, 2, (2), PP. 125-141

[11] Kimura, M.; and Crow, J.F.: 'The Number of Alleles That Can Be Maintained in a Finite Population', Genetics,[J], 1964, 49, PP. 725-738

[12] Zhang, Q.: 'Computational methods in animal breeding' , Science Press, [B], 2007

[13] Matukumalli, L.K.; Lawley, C.T.; Schnabel, R.D.; Taylor, J.F.; Allan, M.F.; Heaton, M.P.; O'Connell, J.; Moore, S.S.; Smith, T.P.; Sonstegard, T.S.; and Van Tassell, C.P.: 'Development and characterization of a high density SNP genotyping assay for cattle', PLoS ONE,[J], 2009, 4, (4), PP. e5350

[14] VanRaden, P.M.: 'Efficient methods to compute genomic predictions', J. Dairy Sci.,[J], 2008, 91, (11), PP. 4414-4423

[15] Qanbari, S.; Pimentel, E.C.G.; Tetens, J.; Thaller, G.; Lichtner, P.; Sharifi, A.R.; and Simianer, H.: 'The pattern of linkage disequilibrium in German Holstein cattle', Anim. Genet.,[J], 2010, 41, PP. 346-356

[16] Zhang, Z.; Liu, J.F.; Ding, X.D.; Bijma, P.; de Koning, D.J.; and Zhang, Q.: ‘Best linear unbiased prediction of genomic breeding values using trait-specific marker-derived relationship matrix’, PLoS ONE,[J], 2010, 5, (9), PP. e12648

[17] Barrett, J.C.; Fry, B.; Maller, J.; and Daly, M.J.: ‘Haploview: analysis and visualization of LD and haplotype maps’, Bioinformatics,[J], 2005, 21, (2), PP. 263-265