A large scale genome simulation model incorporating patterns of linkage disequilibrium

J. Song[1] S. R. Wilson[2]

(Received 27 January 2011; revised 5 November 2011)

Abstract

Evaluation of multiple hypotheses is a common problem, especially in genome wide association studies. False Discovery Rate control is often used to correct for multiple comparisons. However, this approach is influenced by linkage disequilibrium and needs to be evaluated. For such evaluation, first a model needs to be developed with a simulated linkage disequilibrium pattern that matches as closely as possible the observed linkage disequilibrium structure. This paper outlines the steps involved in the development of such a model, using mouse genome data as an example.

http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/3926 gives this article, © Austral. Mathematical Soc. 2011. Published November 13, 2011. ISSN 1446-8735. (Print two pages per sheet of paper.)

Contents

1 Introduction **C933**

2 Method **C934**
2.1 Model and its parameters C935
2.2 Simulated population structure C936
2.3 Recombination C936
2.4 Mutation C939
2.5 Base gamete population C940
2.6 The number of generations C940
2.7 Linkage disequilibrium measures C941
2.8 Mean square error C941

3 Results **C942**

4 Discussion **C946**

References **C946**

1 Introduction

Due to the rapid improvement in high-throughput genotyping technology, there is now a proliferation of genome wide association studies (GWAS), where a dense set of single nucleotide polymorphisms (SNPs) across the whole genome is genotyped to survey the associations between the common genetic variation and disease or quantitative traits. This method relies on either genotyping the causative polymorphism directly, or on genotyping markers that are in strong linkage disequilibrium (LD) with the causative site. Commonly, in such studies large numbers of hypotheses are evaluated simultaneously, and this multiple testing problem is frequently addressed by controlling the False Discovery Rate (FDR) [4]. Linkage disequilibrium (LD) is the non-random

association of alleles at different loci in a population and is one factor that influences FDR control in GWAS. To evaluate the performance of FDR control, first a model with relevant LD structure needs to be developed. So the aim here is to outline the computational steps used in a model with LD structure that matches real data as closely as possible, and to compare this model to the data. Here, we focus on the mouse genome because in GWAS the inbred strains of laboratory mice provide a powerful way to identify the variants that affect a variety of complex traits, including many related to human diseases such as atherosclerosis, diabetes, and obesity.

Ardlie et al. [2] reported factors that influence LD pattern, including mutation rate, variation in recombination rate and population structure. LD structure is population specific. The level of LD in humans is low due to the large effective population size [6]. Conversely, the degree of LD is substantial in laboratory mice because of inbreeding problems from multigenerational crosses [7].

In this study, a simulation model is developed to evaluate the impact and importance of parameters affecting LD in a population, with the aim of matching the simulated population as closely as possible to the LD structure of the inbred mouse population. Both D' and r^2 are used to measure LD for the simulated data, and then the LD patterns associated with these two measures are investigated. Here, D is the deviation of the observed frequency of a haplotype (a combination of alleles at different loci on the chromosome that are transmitted together) from the expected, and D' is determined by dividing D by its maximum possible value, given the allele frequencies. Sometimes r^2 is denoted by Δ^2 and is the square of a correlation measure between pairs of loci [2].

2 Method

The simulation model is described first, then the LD measures, D' and r^2, are obtained for the simulated data. In order to build the model best

fitting the real data, many simulation scenarios are evaluated using different parameter combinations. Finally, the mean square error (MSE) is used as the criterion to find the best simulation scenario for the real mouse genome data. Data on 2202 mice genotyped for 13,459 SNPs were downloaded from the Wellcome Trust Centre for Human Genetics web page [1] and analysed for all 19 autosomes. Pedigree structure is ignored and the LD measures (D' and r^2) are pooled over all autosomes. This data set is not from natural populations: it is the result of intercrossing eight inbred lines to create a population which is maintained for 50 generations by pseudo-random mating and is referred to as heterogeneous stock mice. It is of interest because of its central use in Quantitative Trait Loci (QTL) mapping [8].

2.1 Model and its parameters

The population structure is one of the factors that affects the LD pattern, and this is described first. The model outlined here uses forward simulation of a population, where the entire population is simulated from past to present. This model allows for mutation, varied recombination rate along the chromosome, variation in the distribution of the minor allele frequency (MAF) of the SNPs, varying the size of the base gamete population and genetic drift. Here, mutation is a change in a genomic sequence, and genetic recombination is a process by which the combinations of alleles observed at different loci in the two parents become shuffled in their offspring.

As a starting point, the following parameters and assumptions are made: the mutation rate is set at 10^{-5} per bp (base pair) per generation in the model; we assume the first generation of the simulated population is from the base gamete population of size 100, and the distributions of the MAFs of the SNPs follow the uniform distribution. Step 1 is designed to find the simulation scenarios which best fit the initial parameters of the mouse data. Next, the initial parameters are changed sequentially in Steps 2, 3, 4 and 5 to determine better simulation scenarios from the previous step, and then the

final simulation model is built. The details of each step are described in the following subsections. Figure 1 outlines the building process of this model. The procedure described here will not necessarily find the local minimum in the parameter space, but it should find at least a good approximation.

For convenience, one chromosome, length approximately 100 Megabases (Mb), is used for this study. The number of SNPs is set to $2,000$ and the SNPs are uniformly allocated along the chromosome.

2.2 Simulated population structure

Initially 100 generations were simulated, and then the number of generations was altered (details given later). In each generation of pedigree, the number of individuals was kept at 20, which includes ten males and ten females. Parents were randomly mated and each of the ten mating pairs produces six offspring, but only two (one male and one female) were kept in the pedigree due to the high mortality rate. Although the genotypes of SNPs for all the individuals in the population were simulated, only those of the last 600 individuals were used for further analyses.

2.3 Recombination

Recombination plays an important role in generating LD in populations that are typically of limited size. This simulation model allows a variable recombination rate across the simulated chromosome in the physical scale which shows the physical locations of genes and other DNA sequences of interest. We assumed there are four types of segments along the chromosome, namely: (A) block (without recombination), (B) partial block (with a small recombination rate, < 1 cross-over per 100 Mb), (C) normal segment (with a moderate recombination rate, one cross-over per 100 Mb, and (D) hot spot (with a high recombination rate, > 1 cross-over per 100 Mb). A genetic

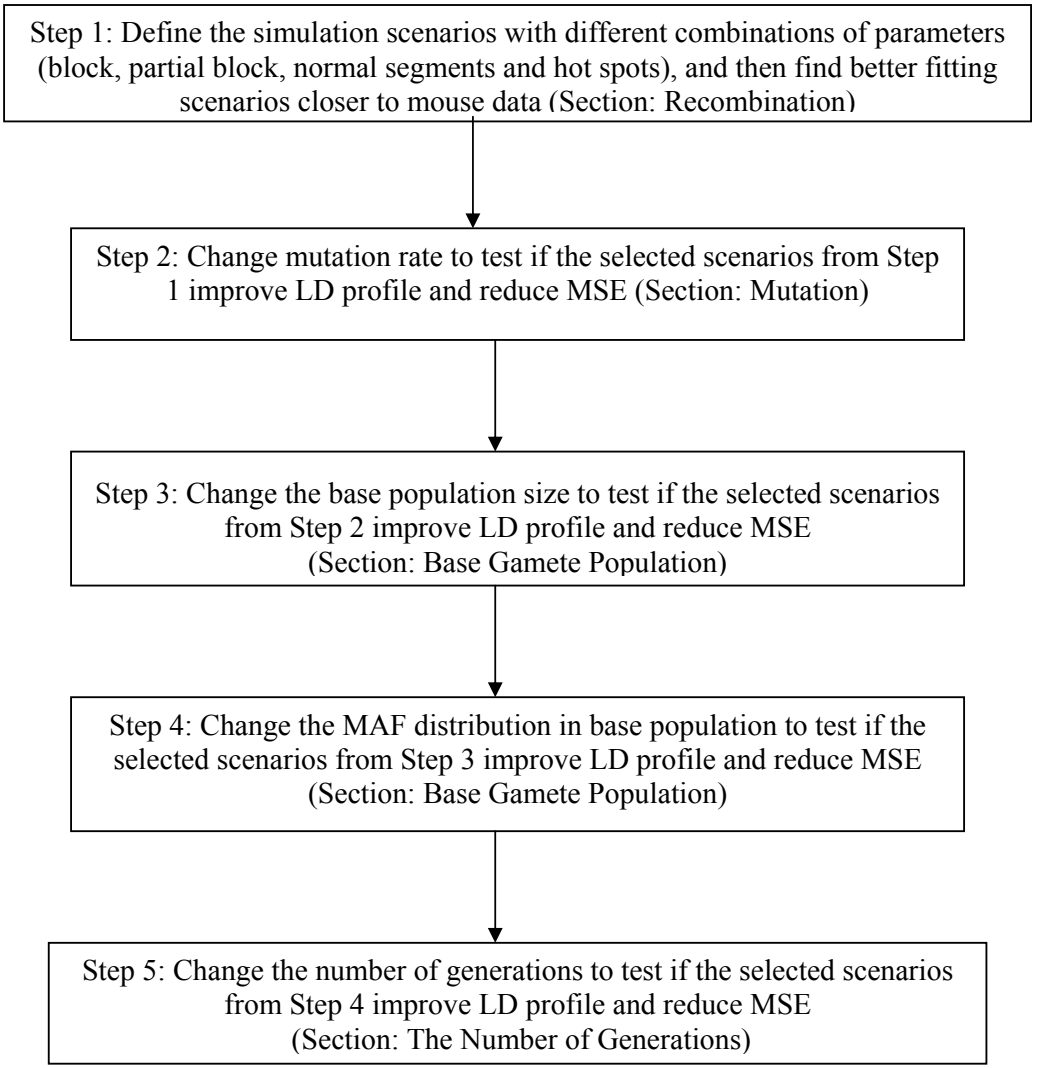

Figure 1: Outline of the procedure to build the simulation model.

map is built to correspond to this simulated chromosome. For the genetic map, there is by definition a constant average recombination rate of one cross-over per Morgan (one Morgan ≈ 100 Mb), so the concept of blocks, partial block and hot-spots is not relevant on this scale, but is on the physical scale. Here, Morgan is a genetic map unit; 1 Morgan = 100 centimorgans (cM). A centimorgan is defined as the distance between loci in which one product of meiosis (a process by which one cell divides into four different cells) in 100 is recombinant. To simulate the gene flow across each pedigree, the number of recombination events on the genetic scale is assumed to follow the Poisson distribution with parameter λ equal to the length of the genetic map, and with the location of recombination on the genetic scale following a uniform distribution. Each recombination location on the genetic scale corresponds to its own location on the physical scale. There are four types of segments (A, B, C and D) on the physical map. By multiplying the segment length with the different weights, a, b, c and d, each corresponding to A, B, C, and D type segments respectively, the genetic map is built. Figure 2 gives an example of building the genetic map. In Figure 2, the weights are $a = 0$, $b = 0.5$, $c = 1.0$ and $d = 2.0$, so there is no block (A) appearing on the genetic map, and hence in the simulation there was no recombination in this block.

The four segments types on the chromosome were simulated from a multinomial distribution with four associated probabilities (namely the percentage of the total simulated chromosome comprising that segment type), $\mathbf{p} = (p(A), p(B), p(C), p(D))'$ with $p(A) + p(B) + p(C) + p(D) = 1$, that were specified initially. The length of each piece of segment on the chromosome was simulated from an exponential distribution. The mean length for each segment type ($\boldsymbol{\mu} = (\mu_A, \mu_B, \mu_C, \mu_D)'$) was set in advance. For each segment type, there are three ranges of probabilities that were evaluated, small, medium and large, and similarly three different ranges of mean of segment length were simulated. Therefore, there are nine simulation scenarios for each segment type, and there are 36 simulation scenarios for the four segment types.

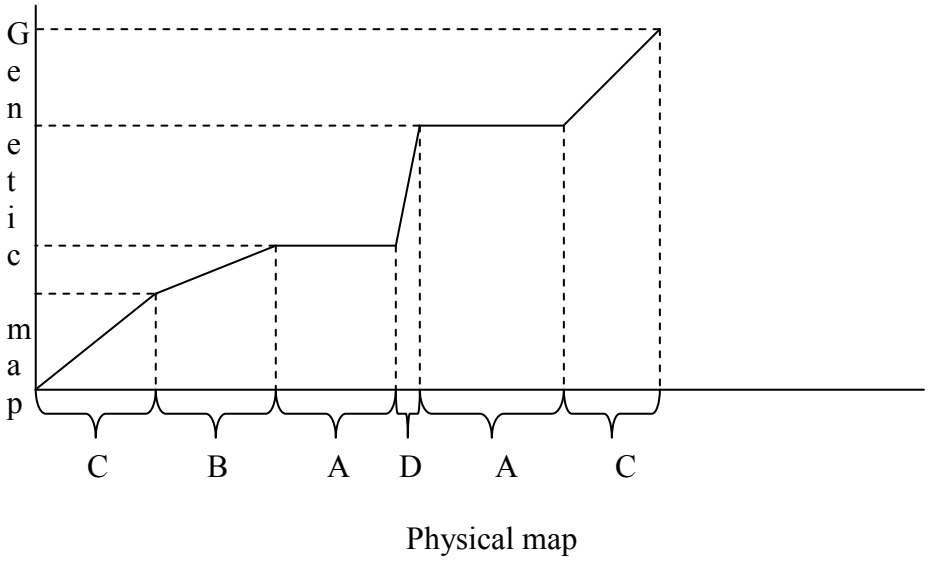

Figure 2: The physical map versus genetic map.

Each simulation scenario was replicated five times for the same LD pattern. To find better fitting scenarios among all 36 scenarios produced at Step 1, each was compared with simulation Scenario 1: constant recombination along the chromosome. These 36 scenarios are labelled Scenario 2, Scenario 3, ..., Scenario 37.

2.4 Mutation

Mutations were randomly placed on the chromosome according to the usual assumption that the number of mutations is Poisson distributed with mean given by the product of the mutation rate and chromosome length (L) and the mutation locations are drawn from a uniform $U(0, L)$ distribution. Initially the mutation rate was set at 10^{-5} per bp per generation for the above simulation scenarios. However, to find the scenarios which best match the observed

mouse data, two other mutation rates, 4×10^{-5} and 10^{-4}, were used for the better fitting scenarios of Step 1. The change of mutation rate happens at Step 2.

2.5 Base gamete population

We assume that there is an initial pool of (base) gametes and we call this a base gamete population. In the pool, small segments from different base gametes were randomly chosen to form one pair of homologous chromosomes for each (base) individual in the first generation that has ten females and ten males. Initially, we set the number of base gametes in the pool to be 100 (that is, the base population size was set at 100), and then the size was changed to 50 and to 1,000 to find the simulation output that best matches the observed mouse data using the better simulation scenarios from Step 2; this is Step 3.

Two different distributions were evaluated for the MAF of the SNPs in the base gamete population. Initially, the distribution of the MAF was set to follow a uniform distribution. In order to simulate more SNPs having higher MAFs, a beta distribution with both parameters set to ten was used. That is, $\text{MAF} = \min(X, 1 - X)$ where $X \sim \text{Beta}(10, 10)$. The change in the MAF distribution is in Step 4 using the best simulation scenarios from Step 3.

2.6 The number of generations

The final parameter evaluated was the number of generations in the simulated population. The number of generations was changed to 50 and 200 in Step 5 for the best simulations in Step 4, to evaluate whether the MSE decreases.

2.7 Linkage disequilibrium measures

Two of most commonly used LD measures, D' and r^2, were obtained via Haploview software [3] for each pair-wise combination of SNPs on this chromosome based on the genotype information of the last three generations (600 individuals). The mean D' and r^2 were calculated for each of 15 intervals: 0–1 kb, 1–10 kb, 10–20 kb, 20–40 kb, 40–60 kb, 60–100 kb,100–200 kb, 200–500 kb, 0.5–1 Mb, 1–2 Mb, 2–5 Mb, 5–10 Mb, 10–20 Mb, 20–50 Mb and $>$ 50 Mb. The choice of intervals is from a study reported by Khatkar et al. [5].

SNPs showing significant deviations from Hardy–Weinberg equilibrium (HWE) ($P < 0.0001$) were excluded from analysis, as were SNPs with minor allelic frequency (MAF) < 0.05. The number of SNPs is $2,000$ at the beginning of the simulation, but there were usually less than half this number in the final LD analysis because some SNPs become fixed during the simulation procedure.

2.8 Mean square error

The MSE is used here as the criterion for determining which simulation scenario best fits the mouse data. The best simulation scenario is the one with the smallest MSE, and this was estimated for both D' and r^2. The simulation was repeated five times for each scenario to obtain the mean and standard deviation of the estimated D' and r^2 for each interval ($i = 1, 2, \ldots, I = 15$), say μ_i and σ_i. The mean D' and r^2 were obtained for each inter-SNP interval from the observed mouse data, denoted $\hat{\mu}_i$. Therefore, the MSE for each interval is $\text{MSE}_i = [\mu_i - \hat{\mu}_i]^2 + [\sigma_i]^2$. Because the value of D' (or r^2) is different for each interval, and also different for the simulation versus the observed mouse data, 'average' weights are needed for each interval to calculate an overall MSE. The 'average' weight

$$p_i = (n_i^{(\text{sim})}/N^{(\text{sim})} + n_i^{(\text{mouse})}/N^{(\text{mouse})})/2 = (p_i^{(\text{sim})} + p_i^{(\text{mouse})})/2$$

where

- $n_i^{(sim)}$ = average number of simulated values in the ith interval;
- $n_i^{(mouse)}$ = the number of mouse-derived values in the ith interval;
- $N^{(sim)} = n_i^{(sim)} + \cdots + n_I^{(sim)}$ = the sum of the average number of simulated observations in each interval;
- $N^{(mouse)} = n_i^{(mouse)} + \cdots + n_I^{(mouse)}$ = the total number of mouse-based observations.

The 'average' in $n_i^{(sim)}$ is over the series of five simulations. Next the weighted average MSE

$$\overline{\text{MSE}} = \sum_{i=1}^{I} p_i \text{MSE}_i \, .$$

Because the MSE is different for r^2 and D', there are three types of MSE: r-squared.mse, D'.mse, and mean.mse which is the mean of r-squared.mse and D'.mse.

3 Results

In Step 1 of this simulation study, the three MSE values for the simplest scenario, namely Scenario 1, ware compared with the corresponding values for the other 36 scenarios. A subset of the data (namely Scenario 1, ..., Scenario 10) is given in Figure 3(a). Considering all scenarios (the remaining 27 are not shown here), Scenario 3 is found to be the best as it has overall smaller MSE values than Scenario 1, and so is selected. For Scenario 3, the empirical mean length (kb) and associated probability (%) of block, partial block, normal segment and hot spot are 55.9 kb and 3.3%, 190.3 kb and 22.5%, 203.5 kb and 72.4% and 81.9 kb and 1.8%, respectively.

For the selected scenario identified above (Scenario 3), the mutation rate was changed to 4×10^{-5} (Scenario 3A) or 10^{-4} (Scenario 3B) to assess if this improves the fit to the observed mouse data. We found that Scenario 3 with

the initial mutation rate (10^{-5}) gives a better fit than the scenarios with increased rates because the three types of MSE of Scenario 3 with mutation rate 10^{-5} are smaller than the other corresponding MSEs.

In Step 3, the same candidate Scenario 3 was further examined by changing the base population size to 50 (Scenario 3C) or 1,000 (Scenario 3D) from 100. We found that the MSEs for the model with base population size 50 or 1,000 generally become larger, so the base gamete population size is kept at 100.

In Step 4, the distribution of the MAF in the base population was changed to the beta distribution with both parameters equal to ten (Scenario 3E) for the scenario identified above (Scenario 3). It appears that the beta distribution is the preferred choice because the three types of MSE decrease when the distribution of the MAF was changed to a beta distribution.

Until now, the best scenario is Scenario 3E. In Step 5 we changed the number of generations in the simulated population to 50 (Scenario 3F) and 200 (Scenario 3G) rather than 100. We found that the three types of MSE of Scenario 3E are the smallest ones among these three scenarios, so the number of generations was kept at 100.

Now, the model building is complete. Scenario 3E gives the closest fit to the mouse data with the associated model having mutation rate 10^{-5}, base population size 100, beta distribution for the MAF of the SNPs in the base population and the number of generations is 100. Figure 3(b) shows the MSEs for Scenarios 3 and 3E.

For all the scenarios evaluated, Scenario 3E best matches the LD pattern of the observed mouse data. Next the LD patterns for Scenarios 3E and the mouse data are compared. Figure 4 shows the mean LD estimates (D' and r^2) at different intervals for Scenario 3E and the mouse data. From this figure, it can be seen that the fit is good, especially for D'.

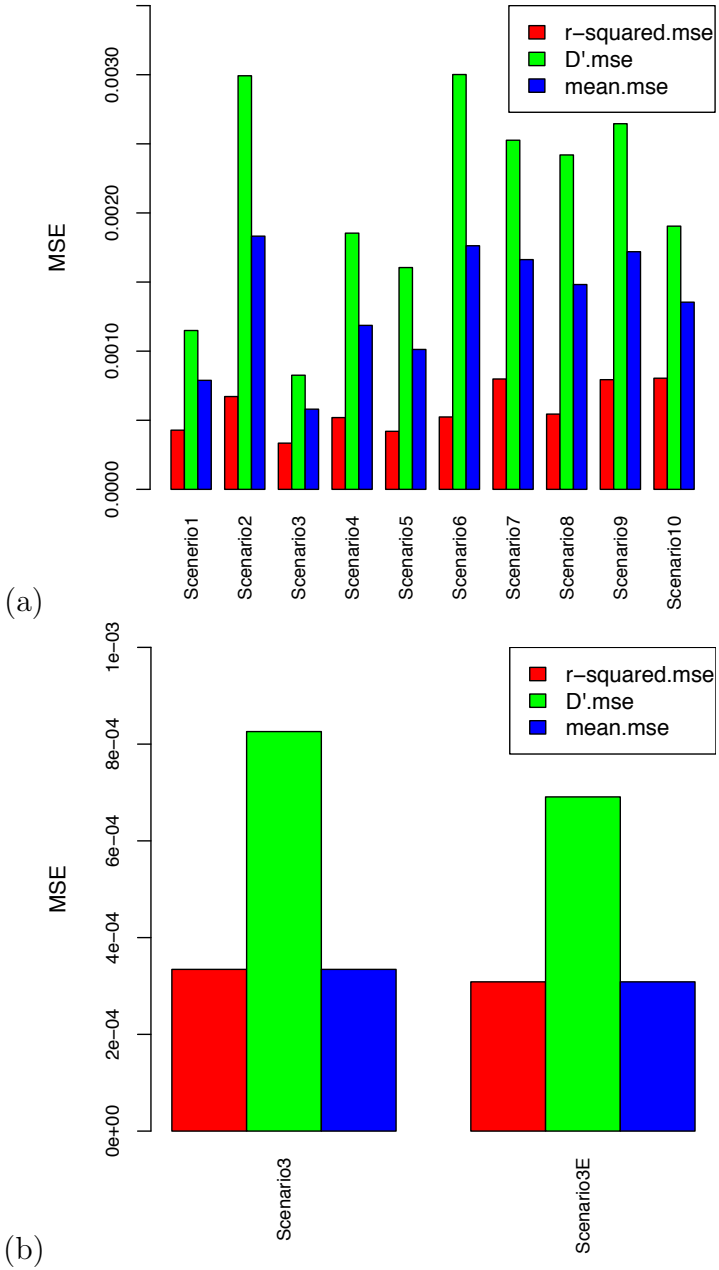

Figure 3: MSE results for different scenarios where Figure 3(a) is for Scenarios 1–10 and Figure 3(b) is for Scenarios 3 and 3E.

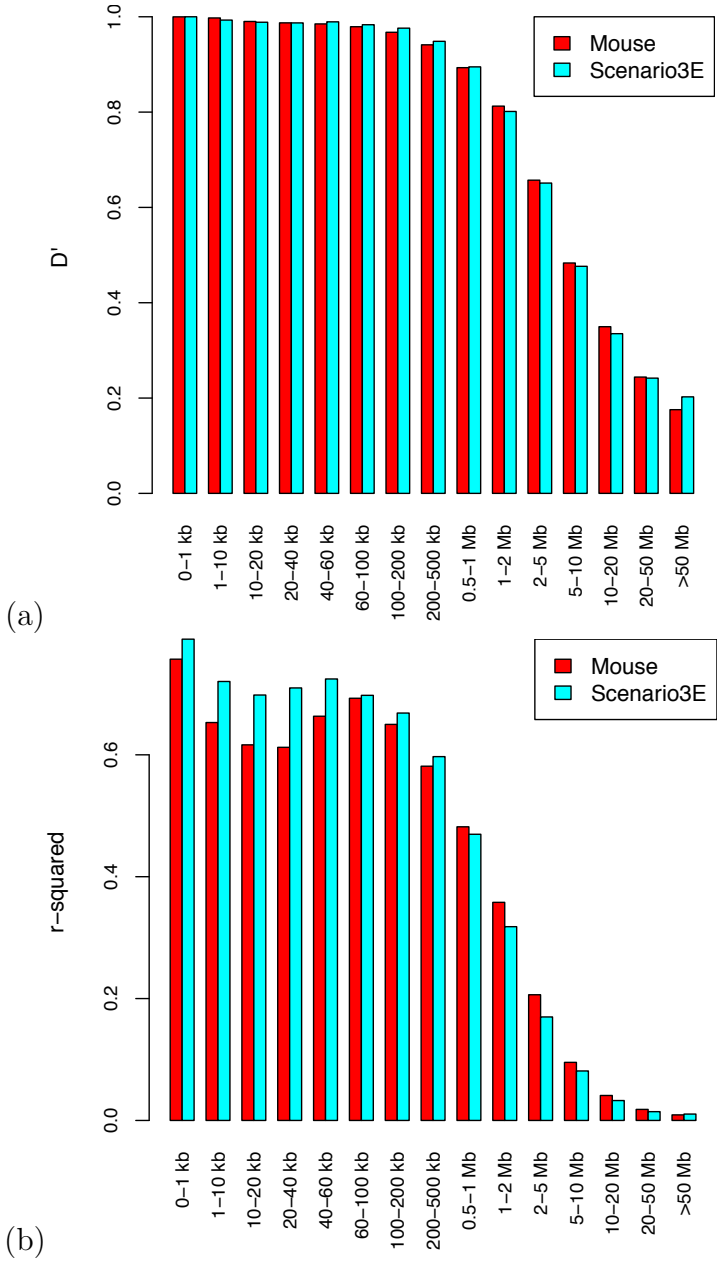

Figure 4: Mean LD measures at different physical distances for mouse and Scenario 3E data where the top figure is for D' and the bottom figure is for r^2.

4 Discussion

This is a comprehensive simulation study with the aim of developing a model to match the LD pattern found in mouse data. With the establishment of this tool, it will be possible to investigate the effects of LD on the evaluation of false positives and the estimation of the False Discovery Rate for the multiple testing problems that commonly arise in SNP-trait association studies. The construction of haplotype blocks and identification of tag SNPs have been found to be informative in detecting the variation in LD across the genome, and using this information could improve the efficiency of gene mapping.

Comparing the simulated LD pattern with the LD structure of the observed mouse data, shown in Figure 4, the fit is very good, so this tool can be applied to other species, as long as an appropriate LD structure is built into the model for the specific population by changing the parameters appropriately. However, we need to be aware of the practical limitation of this tool for the simulation of large populations that do not have a high degree of inbreeding. Then the computational demand will be very high or it may not even be possible to simulate an analogous population because the size of the first generation will need to be very large; ten females and ten males are not enough, and a few thousand may be needed.

Acknowledgements This work was funded by the Australian National Health and Medical Research Council grant number 525453. We thank the reviewer for several helpful comments.

References

[1] http://gscan.well.ox.ac.uk/ C935

[2] K. G. Ardlie, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309, 2002. doi:10.1038/nrg777 C934

[3] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263, 2005. doi:10.1093/bioinformatics/bth457 C941

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. http://www.jstor.org/pss/2346101 C933

[5] M. S. Khatkar, F. W. Nicholas, A. R. Collins, K. R. Zenger, J. A. L. Cavanagh, W. Barris, R. D. Schnabel, J. F. Taylor, and H. W. Raadsma. Extent of genome-wide linkage disequilibrium in Australian Holstein–Friesian cattle based on a high-density SNP panel. *BMC genomics*, 9(1):187, 2008. doi:10.1186/1471-2164-9-187 C941

[6] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature genetics*, 22:139–144, 1999. doi:10.1038/9642 C934

[7] C. C. Laurie, D. A. Nickerson, A. D. Anderson, B. S. Weir, R. J. Livingston, M. D. Dean, K. L. Smith, E. E. Schadt, and M. W. Nachman. Linkage disequilibrium in wild mice. *PLoS Genetics*, 3(8):e144, 2007. doi:10.1371/journal.pgen.0030144 C934

[8] W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. P. Rawlins, R. Mott, and J. Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature genetics*, 38(8):879–887, 2006. doi:10.1038/ng1840 C935

Author addresses

1. **J. Song**, Prince of Wales Clinical School, University of New South Wales, Kensington, AUSTRALIA.
mailto:json7944@uni.sydney.edu.au

2. **S. R. Wilson**, University of New South Wales; Australian National University, ACT, AUSTRALIA
mailto:sue.wilson@anu.edu.au