Simulating a Coalescent Process with Recombination and Ascertainment

Ying Wang and Bruce Rannala

University of Alberta, Edmonton AB T6G2H7, Canada,
brannala@ualberta.ca,
WWW home page: http://rannala.org

Abstract. A new method is presented for use in simulating samples of disease and normal chromosomes bearing multiple linked genetic markers under a neutral model of mutation, genetic drift, and recombination. The method accounts for the potential effects of investigator sampling bias by allowing for ascertainment of chromosomes according to disease status and of markers according to a pre-specified polymorphism cutoff level. The method was implemented in a computer program and applied to study the general effects of disease mutation age (or frequency), levels of marker polymorphism, and sample size, on pairwise LD between markers and a disease mutation. It is shown that the average pairwise LD between a marker and a disease mutation is lower for older, or more prevalent, disease mutations, as expected. The marker polymorphism cutoff level also has an important influence on LD. Potential applications of the method for predicting the power of genome-wide marker-disease association studies are discussed.

1 Introduction

With the advent of a human haplotype map initiative [1] and the emerging prospect of large amounts of data on haplotype frequencies and linkage disequilibrium (LD) in the human genome [2–5], as well as future large-scale projects aimed at mapping genes influencing complex diseases by marker-disease association analysis [6], it has become increasingly important to understand the processes determining human genetic variation. Of particular interest is the influence of underlying evolutionary forces [7, 8] and sample ascertainment strategies [9, 10] on variation observed at commonly used genetic markers, such as single nucleotide polymorphisms (SNPs) and microsatellites. A related question concerns the potential power of disease-marker association studies that rely on linkage disequilibrium (LD) to locate disease susceptibility genes [6, 11, 12]. Recent studies have used computer simulations to address these, and other, questions [7, 10, 13, 14].

Most simulation studies generate samples of chromosomes under a neutral coalescent model, taking account of the population processes (mutation, recombination, genetic drift, etc) that determine patterns of marker polymorphism and linkage disequilibrium [10, 15]. Such studies are useful for predicting the patterns of neutral variation one would expect to see in a random sample of chromosomes from a population, but the genetic variation will be different in a case-control study because individuals are ascertained based on disease status. Krugylak [7] attempted to take account of disease allele frequency by only accepting genealogies with a mutation present at a frequency in the sample that was equal to the required population frequency of the disease mutation. There are two problems with this design: (1) for small samples the sample frequency of a mutation will deviate considerably from the population frequency; (2) disease studies typically enrich the sample for a disease allele by ascertaining for affected individuals and the frequency of a disease allele in a sample will therefore be much greater than its frequency in the population. A case-control sampling strategy enriches the sample for chromosomes descended from one or more disease mutations and this has the effect of altering the shape of the underlying genealogy causing it to differ from that expected under a neutral coalescent model [14, 16]. There is also an ascertainment effect for markers because SNPs and/or microsatellite markers are chosen from panels of markers known to be polymorphic [9, 10].

To examine many of the above questions (e.g., the power of case-control association studies for mapping disease mutations, etc) via simulation studies, a new simulation method is needed that allows samples to be generated under a coalescent process with a case-control sampling strategy and polymorphic marker ascertainment. Here we outline a method to simulate samples under a coalescent process that allows for these sources of ascertainment bias. Our basic strategy is to simulate the sample path (over time) of the population frequency of a disease allele, using a diffusion approximation, and then to simulate the coalescent of a sample of chromosomes conditional on the sample path of the allele frequency (using theory for the coalescent process in a population of variable size). This improves upon an earlier method for simulating a coalescent process with disease ascertainment proposed by Zollner and von Haeseler [14], eliminating several key assumptions in their model which will often be violated in practice. In particular, they assumed that the frequencies of disease mutations remain constant over time; our diffusion simulation eliminates the need for this assumption.

2 Theory

The ancestral processes of lineage coalescence and recombination (within a panmictic population) traced backwards in time can be represented as a graph, with recombination events splitting a chromosome to create two ancestors (each carrying only a segment of the descendent chromosome), and coalescence events uniting pairs of chromosomes to generate a common ancestral chromosome [17, 18]. As is usual, this will be referred to as the “ancestral recombination graph.” One can simulate population samples of chromosomes and genetic markers by simulating the ancestral recombination graph and then simulating independent mutations on the lineages of this graph according to a Poisson process. The standard model of a coalescent process with recombination and mutation assumes that chromosomes are a random sample from a population (i.e., each is equally likely to be sampled) and that the markers are a random sample of loci that may have any number of alleles segregating in the population, including a single allele (no polymorphism).

2.1 Coalescent Process with Disease and Marker Ascertainment

Here we propose a procedure for simulating genetic markers on chromosomes under a coalescent and recombination process taking account of the effects of both disease and marker ascertainment. Our simulation method is similar to that of [14] but uses fewer approximations and relaxes some assumptions. It is assumed that no heterogeneity exists at the disease locus (e.g., all chromosomes bearing the disease mutation descend from a single ancestral chromosome on which the disease mutation arose), although this assumption can be relaxed by simulating multiple disease allele genealogies. Let n_0^D be the number of disease chromosomes sampled and let n_t^D be the number of disease chromosomes ancestral to the sample at generation t in the past. Let n_0^N being the number of normal chromosomes sampled and let n_t^N be the number of normal chromosomes ancestral to the sample at generation t in the past.

Let N_0 be the present population size, let N_t be the population size at generation t in the past, and let p_t be the frequency of the disease mutation at generation t in the past. We define L to be the total number of marker loci examined and simulate either SNP markers with mutation rate $\mu_{nc} = 10^{-8}$ or microsatellite markers under a stepwise mutation model [19, 20] with mutation rate $\mu_{mt} = 10^{-3}$. Other models of mutation could be easily incorporated. The map distance between marker 1 and marker L is defined as ρ, the distance between markers i and j is defined as ρ_{ij}, and the location of the disease mutation, denoted as θ, is defined as the distance of the disease locus (in map units) from marker 1.

2.2 Ancestral Graph with Disease Ascertainment

The coalescent and recombination processes are simulated jointly. The process is illustrated in figure 1. Note that some time after the generation at which the disease mutation arose, the disease chromosomes share a most recent common ancestor (MRCA). After the time of origin of the disease mutation, disease chromosomes only coalesce with one another as do the normal chromosomes. Recombinations, on the other hand, can occur both within, and between, genealogies of the disease and normal chromosomes. If a disease chromosome recombines with a normal chromosome, this increases the rate of the coalescence-recombination process in the genealogy of normal chromosomes by one and vice versa. If a recombination occurs between two disease chromosomes, or two normal chromosomes, the effect is to increase the rate by one in the respective class in which the recombination event occurred (Figure 1). To obtain the waiting times between events, as well as the type of each event, we simulate three waiting times: (1) the time until a coalescent event occurs in the sample of disease chromosomes; (2) the time until a coalescence event occurs in the sample of normal chromosomes;

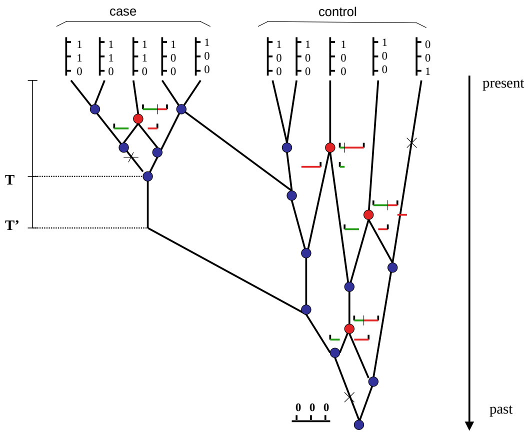

Fig. 1. An example genealogy of cases (inheriting a disease mutation) and controls. The age of the most recent common ancestor (MRCA) of the cases is denoted as T and the age of the disease mutation as T'. Marker mutations are indicated by Xs and locations of recombination events by vertical lines on a horizontal line representing the segment of chromosome spanned by markers. The states of 3 biallelic markers are indicated by 0s (ancestral) and 1s on chromosomes at the tips of the genealogy.

and (3) the time until a recombination event occurs in the ancestry of either the normal chromosomes, or the disease chromosomes. To facilitate the simulation of the disease allele frequency over time, we use a discrete-time model, rather than a continuous time model as is usual for the coalescent process. The probability density function (pdf) of the time until a coalescence event occurs in the gene tree of the disease chromosomes, given that the previous event (coalescence or recombination) occurred at time t_0, is

$$f_D(t_D) = \frac{\binom{n^D_{t_0}}{2}}{2N_{t_D}p_{t_D}} \exp\left\{ -\binom{n^D_{t_0}}{2} \sum_{t=t_0}^{t_D-1} \frac{1}{2N_t p_t} \right\} . \tag{1}$$

In our simulations, we assumed that the population has grown at an exponential rate r [21, 22]. In that case equation 1 becomes

$$f_D(t_D) = \frac{\binom{n^D_{t_0}}{2} e^{rt_D}}{2N_0 p_{t_D}} \exp\left\{ -\frac{\binom{n^D_{t_0}}{2}}{2N_0} \sum_{t=t_0}^{t_D-1} \frac{e^{rt}}{p_t} \right\} . \tag{2}$$

The pdf of the time until a coalescence event occurs in the gene tree of the normal chromosomes, given that the last event (coalescence or recombination) occurred at time t_0, is

$$f_N(t_N) = \frac{\binom{n^N_{t_0}}{2} e^{rt_N}}{2N_0(1-p_{t_N})} \exp\left\{ -\frac{\binom{n^N_{t_0}}{2}}{2N_0} \sum_{t=t_0}^{t_N-1} \frac{e^{rt}}{(1-p_t)} \right\}. \tag{3}$$

The pdf of the time until a recombination event occurs is

$$f_R(t_R) = e^{-(n^D_{t_0}+n^N_{t_0})t_R\rho/2}. \tag{4}$$

A method is described below for generating random variables from each of the density functions of equations 2, 3 and 4 above using either the usual inverse transformation method or a recursive equation.

2.3 Method for Simulating Events in the Ancestral Graph

To simulate the waiting time, t_N*, until a coalescence occurs in the sample of normal chromosomes, for example, using the inverse transformation method, we would first simulate a uniform random variable $U \in [0,1]$ and then solve for t_N* in the equation

$$\begin{aligned} U &= 1 - F_N(t_N), \\ &= 1 - \sum_{j=t_0}^{t_N*} \left[\frac{\binom{n^N_{t_0}}{2} e^{rj}}{2N_0(1-p_j)} \exp\left\{ -\frac{\binom{n^N_{t_0}}{2}}{2N_0} \sum_{t=t_0}^{j-1} \frac{e^{rt}}{(1-p_t)} \right\} \right], \end{aligned} \tag{5}$$

where $F_N(t_N)$ is the cumulative density function of t_N. Waiting times until recombinations are simulated from the exponential density of equation 4 using the inverse transformation method. If $t_D < t_N$ and $t_D < t_R$, the next event is the coalescence of a random pair of chromosomes in the genealogy of the sampled disease chromosomes at time t_D. If $t_N < t_D$ and $t_N < t_R$, the next event is the coalescence of a random pair of chromosomes in the genealogy of the sampled normal chromosomes at time t_N. Otherwise, the next event is a recombination involving a randomly chosen chromosome at time t_R.

In order to speed up simulation of the time until coalescence, instead of faithfully following the CDF and using inverse transformation, we make use of the probability that a coalescence occurs at time t^*-1 to calculate the probability of a coalescence at time t^*. To simulate the genealogy of the disease chromosomes, for example, this is done by multiplying a value,

$$Q\left(t_D{}^*\right) = \left[\tfrac{p_{(t^*-1)} e^r}{p_{t^*}} \exp\left\{ -C \ \tfrac{e^{r(t^*-1)}}{p_{(t^*-1)}} \right\} \right] \times Q\left(t^*-1\right), \tag{6}$$

where C is equal to $n_{t^D_0}(n_{t^D_0}-1)/\left(4N_0\right)$.

Using the above recursion to calculate the probability for each generation, the simulation procedure can be expressed as follows:

1. Generate a random number U, between 0 and 1.
2. Let $j = t_0$, $Q = \frac{C\, e^{rt_0}}{p_{t_0}}$, $F = Q$.
3. If $U < F$, return $t^* = j$. Otherwise, go to step 4.
4. According to equation 6, $Q = Q \times [\frac{p_{(t_D{}^*-1)} e^r}{p_{t_D{}^*}} \exp\left\{-C\ \frac{e^{r(t_D{}^*-1)}}{p_{(t_D{}^*-1)}}\right\}]$, $F = F + Q$, $j = j + 1$.
5. go to Step 3.

Once the disease chromosomes coalesce to a MRCA, the rate of coalescence in the genealogy of disease chromosomes is zero until either the time is reached at which the disease mutation arose, or a recombination occurs. If the sample of normal chromosomes coalesce to a MRCA before the disease mutation arose the rate of coalescence in the genealogy of normal chromosomes is also zero until either the time is reached at which the disease mutation arose, or a recombination occurs. At times, in the past, greater than the age of the disease mutation the rate of the coalescence process is scaled by $n_t^N + 1$ until the common ancestor of the disease chromosomes coalesces with a normal chromosome; the rate is then scaled by n_t^N. If a recombination event occurs during the simulations the position of the recombination event is chosen uniformly on the interval of length ρ spanning the markers; this assumes that if recombination hotspots are present they are accounted for by the map lengths of the intervals (other more complex models of recombination with explicit hotspots could also be used).

2.4 Diffusion Model of Disease Allele Frequency

In our simulations, we fix the age of the disease mutation, and then simulate the change of frequency of the mutation over time, conditional on non-extinction. To simulate the change of frequency of the disease mutation we assume that the process can be modeled using a diffusion approximation [24]. We simulate the diffusion process using a procedure suggested by Kimura and Takahata [25]. The basic idea is that the allele frequency at the next generation, given the current frequency, is normally distributed with expectation and variance determined by the diffusion model. If population size is constant, N, then $\mathtt{E}[p_{t+1}] = p_t$ and $\sigma_{p_{t+1}} = p_t(1-p_t)/2N$. If the population size is growing exponentially with rate r and current population size N_0 then $\mathtt{E}[p_{t+1}] = p_t$ and $\sigma_{p_{t+1}} = p_t(1-p_t)/(2N_0 e^{rt})$. If an allele has fewer than 4 copies in the population, then the number of alleles in the next generation is instead simulated as a Poisson random variable, using the branching process approximation for a rare allele [24], with parameter xe^r. This is because the diffusion approximation is no longer accurate in this case. In our simulation method, the initial copy number of the disease mutation when it arises is $x = 1$.

2.5 Models of Mutation and Marker Ascertainment

Mutations are simulated on the ancestral recombination graph generated according to the procedure outlined above. Conditional on the graph (genealogy), the distribution of the positions of mutations on branches is uniform under a Poisson process model of mutation. However, in a coalescent simulation with recombination, the only relevant mutations are those that occurred on portions of ancestral chromosomes that were transmitted to one or more chromosomes in the sample (recall that with recombination only certain segments of recombinant chromosomes are ancestral to the sample). To improve the efficiency of our simulation procedure, we therefore used a simple algorithm to label markers on those chromosomal segments (branches of the ancestral recombination graph) that leave descendents in the sample.

The basic algorithm proceeds as follows: (1) all markers on all sampled chromosomes are initially labeled with integer 1; (2) the network is traversed from the tips to the final MRCA; (3) at each recombination event markers peripheral to the recombination point are labeled with integer 0 and if a marker already is labeled zero it retains its label regardless of its position relative to the current recombination point; (4) at each coalescence event if markers at a locus on the two coalescing chromosomes are 0/1, 1/0 or 1/1 then the locus is labeled 1 on the branch preceding the coalescence event and if they are 0/0 it is labeled 0. The algorithm proceeds until all loci on all branches have a binary integer index.

Mutations are simulated at each locus on all branches labeled 1 according to a Poisson process (with rate $\mu \times t$ on a branch of length t). This mimics a sampling process in which the investigator continues typing SNPs or microsatellite markers for a particular sample of chromosomes until a total of L polymorphic markers are obtained. A Jukes-Cantor model [23] is used to simulate SNP mutations (this model assumes that nucleotides `A`, `T`, `G` and `C` occur in equal frequencies), although more complex models with transition/transversion bias, unequal base frequencies, etc could be easily incorporated. In practice, with the low rate of nuclear mutation only one substitution will occur on a genealogy in most cases, so the substitution model is rather unimportant. A stepwise mutation model is used for modeling microsatellite mutation [19] such that each mutation (with equal probability) either increases, or decreases, the numbers of repeats by one unit (e.g., changes the length by four in a tetranucleotide repeat, etc). The number of (potentially polymorphic) microsatellite markers is specified by the user as a function of the size of the interval. The simulation results that we present here are for SNP markers only. Note that under our simulation procedure both the number (and positions) of polymorphic SNPs for a given simulation are random variables.

3 Simulation Results

In this section, we describe a simulation study conducted to examine the properties of our method as well as the general features of the distribution of LD in ascertained samples of disease and normal chromosomes. The simulations examine the effects of disease mutation age and frequency, levels of marker polymorphism, and sample size, on average levels of pairwise LD between markers and a disease mutation.

3.1 Pairwise patterns of LD in a growing population

We assumed a population size of 10^6 with an exponential growth rate of 0.01. SNPs are distributed along a region of 1 cM (1000 kb). The disease mutation is to the left of all markers. The sample size is set to be 50 disease and 50 normal chromosomes in both figures 2 and 3. Figure 2 shows the influence of disease mutation age on LD. Disease mutation ages were chosen so that expected population frequencies of 0.0001, 0.001, 0.03, 0.1, 0.2 would be obtained. These correspond to disease mutation ages of 164, 600, 710, 840 and 910 generations. The polymorphism cutoff level was 0.05. In figure 3, the disease age was set to be 840 generations and the polymorphism cutoff levels were 0.05, 0.1, 0.2, and 0.3. In figure 4, the disease ages and polymorphism cutoff level are the same as those used in figures 2 and 3. The sample sizes are 10/10, 20/20, 50/50, 100/100, 250/250 disease/normal chromosomes. Linkage disequilibrium is measured using D and r^2 [26]. In total, 100 replicate simulations were carried out for each set of conditions and the average values of pairwise LD are plotted. We applied the

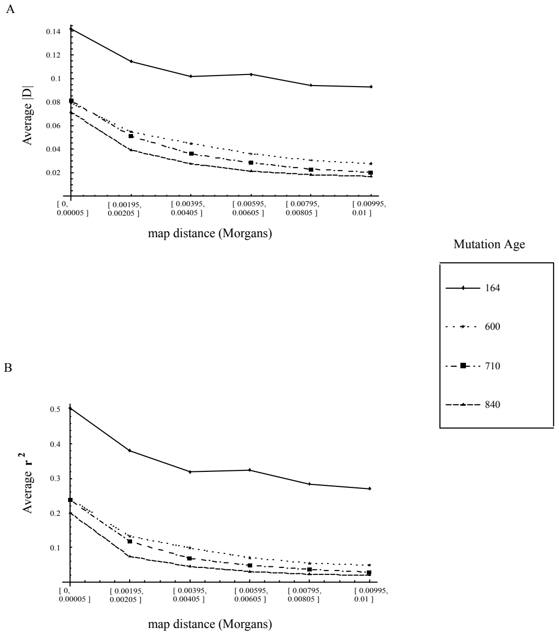

Fig. 2. Simulation results showing average pairwise linkage disequilibrium (LD) between markers and a disease mutation as a function of map distance and disease mutation age. LD is measured using either $|D|$ (panel A) or r^2 (panel B). Younger mutations tend to show greater levels of LD with adjacent markers and average LD decreases monotonically with map distance. Because the locations of polymorphic markers are random under our simulation method map distances are given as intervals.

simulation method to investigate the effects on pairwise LD of disease mutation age (frequency), marker allele polymorphism cutoff level, and sample size. The results suggest that the average LD between a disease allele and markers is lower for older (more prevalent) disease mutations and higher for younger (less prevalent) mutations (Figure 2). The marker polymorphism level also has an important effect on LD. Figure 3 suggests that selecting marker loci with higher polymorphism levels increases the average LD and could potentially increase the power of an LD mapping study. This effect decreases with an increasing the map distance of the marker from the disease mutation. From figure 4, we see that the LD may be biased with small sample sizes. For the simulation conditions used in our study, there is a positive bias in LD for markers near the disease locus (at a distance of less than about .004 Morgans) and a slight negative bias for markers beyond this distance. If more chromosomes are sampled (more than

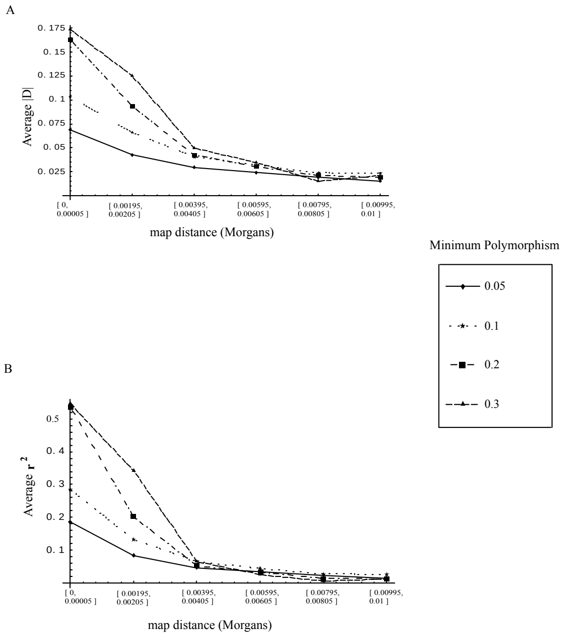

Fig. 3. Simulation results showing average pairwise linkage disequilibrium (LD) between markers and a disease mutation as a function of map distance and the cutoff for the minimum level of polymorphism for markers. LD is measured using either $|D|$ (panel A) or r^2 (panel B). More polymorphic markers tend to show greater LD with the disease mutation. Average LD decreases monotonically with map distance. Because the locations of polymorphic markers are random under our simulation method map distances are given as intervals.

100), the average LD changes little. More extensive simulations are needed to fully understand the extent and importance of such bias.

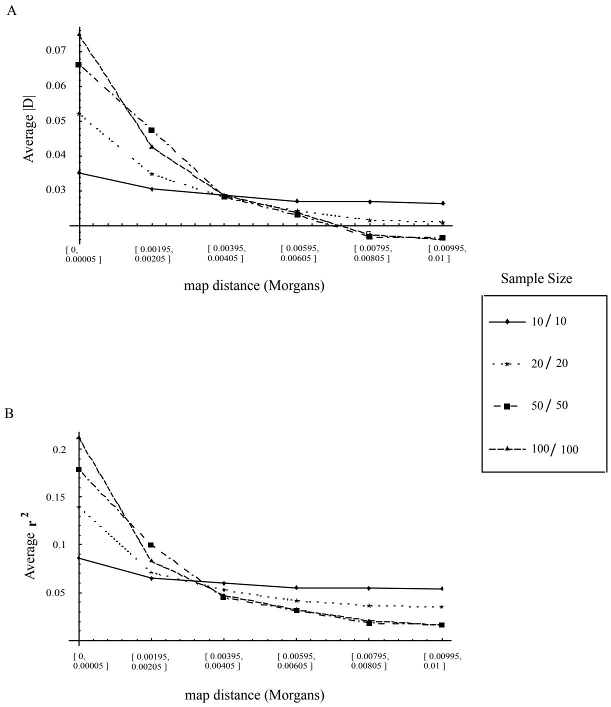

Fig. 4. Simulation results showing average pairwise linkage disequilibrium (LD) between markers and a disease mutation as a function of the number of control and normal chromosomes sampled. LD is measured using either $|D|$ (panel A) or r^2 (panel B). Smaller samples tend to show greater LD between markers and the disease mutation when markers are nearby (less than .004 Morgans away) and larger samples show greater LD when markers are further away (more than .004 Morgans distant). Because the locations of polymorphic markers are random under our simulation method map distances are given as intervals.

4 Discussion

There is currently a great deal of interest, and optimism, concerning the prospect of using population level linkage disequilibrium to detect markers that are closely linked to a disease locus. Whether such genome-wide association studies will have the power to detect common genetic variants associated with complex diseases remains uncertain. It appears likely that the potential power of such studies minimally depend on factors such as the demographic history, etc of the population(s) under consideration, population sampling strategies, and population frequencies of underlying disease loci. The goal of this paper has been to develop a more realistic simulation algorithm to generate population samples for the purpose of studying the influences of such factors.

There are, of course, many additional factors that we have not considered that can be expected to influence the feasibility of studies aimed at identifying disease loci using population-level marker-disease associations. For example, the relationship between genotype and phenotype can be very complex and factors such as the degree of penetrance, phenocopy rate, etc can be expected to greatly influence the power. Fortunately, it is straightforward to simulate phenotypes conditional on genotypes at a disease locus under arbitrarily complex models (including multilocus quantitative genetic models) and therefore the simulation methodology developed here could be used in a two-step modeling approach whereby the genotypes are simulated using our algorithm and the phenotypes are simulated conditional on the genotypes under arbitrary models of the phenotype-genotype relationship.

5 Acknowledgements

Support was provided for this research by Canadian Institutes of Health Research (CIHR) grant MOP 44064 and National Institutes of Health (NIH) grant HG01988 to B.R. The authors thank Brian Golding for pointing out to B.R. (more than a decade ago) the usefulness of Kimura and Takahata's procedure for simulating a diffusion process.

References

1. Dove A (2002). Mapping project moves forward despite controversy. Nature Medicine 8(12):1337.
2. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, et al (2001). Linkage disequilibrium in the human genome. Nature 411(6834):199–204.
3. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, et al (2002). The structure of haplotype blocks in the human genome. Science 296:2225–9.
4. Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33(3):382–7.
5. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A (2003). Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet 12(7):771–776.
6. Cardon LR, Abecasis GR (2003). Using haplotype blocks to map human complex trait loci. Trends Genet 19(3):135–140.
7. Krugylak L (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22(2):139–144.

8. Nordborg M, Tavare S (2002). Linkage disequilibrium: what history has to tell us. Trends Genet 18(2):83–90.

9. Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001). The discovery of single-nucleotide polymorphisms–and inferences about human demographic history. Am J Hum Genet 69(6):1332–47.

10. Akey J (2003). The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. Mol Biol Evol 20(2):232–242.

11. Ohashi J, Tokunaga K (2002). The expected power of genome-wide linkage disequilibrium testing using single nucleotide polymorphism markers for detecting a low-frequency disease variant. Ann Hum Genet 66(Pt 4):297–306.

12. Zhang K, Calabrese P, Nordborg M, Sun F (2002). Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71(6):1386–94.

13. Long AD, Langley CH (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9(8):720–31.

14. Zollner S, von Haeseler A (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am J Hum Genet 66(2):615–28.

15. Stumpf MP, Goldstein DB (2003). Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol 13(1):1–8.

16. Slatkin M (1996). Gene genealogies within mutant allelic classes. Genetics 143(1):579–87.

17. Hudson RR, Kaplan NL (1988). The coalescent process in models with selection and recombination. Genetics 120(3):831–40.

18. Griffiths RC, Marjoram P (1996). Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3:479–502.

19. Valdes AM, Slatkin M, Freimer NB (1993). Allele frequencies at microsatellite loci: The stepwise mutation model revisited. Genetics 133:737–749.

20. Ohta T, Kimura M (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetical Research 22:201–204.

21. Slatkin M, Hudson RR (1991). Pairwise comparisons of mitochondrial-DNA sequences in stable and exponentially growing populations. Genetics 129:555–562.

22. Griffiths RC, Tavaré S (1994). Sampling theory for neutral alleles in a varying environment. Phil Trans Roy Soc London Ser B 344:403–410.

23. Jukes TH, Cantor CR (1969). Evolution in protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism III*, pp. 21–132. New York: Academic Press.

24. Ewens WJ (1979). Mathematical Population Genetics. New York: Springer-Verlag.

25. Kimura M, Takahata N (1983). Selective constrain in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc Natl Acad Sci USA 80:1048–1052.

26. Hill WG, Robertson AR (1968). Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231.