JOINT BAYESIAN ESTIMATION OF MUTATION LOCATION AND AGE USING LINKAGE DISEQUILIBRIUM

B. RANNALA, J.P. REEVE
Department of Medical Genetics, University of Alberta, Edmonton Alberta T6G2H7, Canada

Associations between disease and marker alleles on chromosomes in populations can arise as a consequence of historical forces such as mutation, selection and genetic drift, and is referred to as "linkage disequilibrium" (LD). LD can be used to estimate the map position of a disease mutation relative to a set of linked markers, as well as to estimate other parameters of interest, such as mutation age. Parametric methods for estimating the location of a disease mutation using marker linkage disequilibrium in a sample of normal and affected individuals require a detailed knowledge of population demography, and in particular require users to specify the postulated age of a mutation and past population growth rates. A new Bayesian method is presented for jointly estimating the position of a disease mutation and its age. The method is illustrated using haplotype data for the cystic fibrosis $\Delta F508$ mutation in europe and the DTD mutation in Finland. It is shown that, for these datasets, the posterior probability distribution of disease mutation location is insensitive to the population growth rate when the model is averaged over possible mutation ages using a prior probability distribution for the mutation age based on the population frequency of the disease mutation. Fewer assumptions are therefore needed for parametric LD mapping.

1 Introduction

The term "linkage disequilibrium" (LD) describes a population distribution of alleles (at two or more loci) among chromosomes that is not independent [1]. In other words, alleles at different loci co-occur on chromosomes at a frequency that does not equal the product of the marginal frequencies of the alleles. One mechanism by which LD commonly arises is mutation. A variant allele at a locus arises by mutation on a particular chromosome (bearing specific alleles at other linked loci). Initially, the new allele is found exclusively on this chromosomal background, but over time the association of the new allele with alleles present at other loci on the ancestral chromosome breaks down due to recombination and mutation. Under certain conditions, the association among alleles at the linked loci may disappear and alleles will occur together on chromosomes in proportion to the product of the marginal frequency of each allele, a situation referred to as "linkage equilibrium."

The extent of linkage disequilibrium (LD) among alleles in a population is determined by many factors including the map distances among loci, the rates of mutation at the loci, natural selection, and genetic drift (population

demographic history, etc). Recently, human geneticists have begun to exploit linkage disequilibrium to map disease mutations [2], to estimate the ages of known mutations [3], and to reconstruct ancient demographic events [4].

1.1 Linkage disequilibrium gene mapping

One of the most important practical applications of linkage disequilibrium studies in humans is to map the positions of mutations that cause disease [2]. Linkage mapping methods for finding mutation locations using patterns of marker-disease co-segregation on pedigrees have limited resolution, usually less than 1 cM (roughly 1 Mb). To identify a disease mutation by positional cloning greater resolution is needed. LD mapping methods can have much greater resolution than linkage analysis because the methods exploit recombination events occurring in the extended genealogy relating a random sample of individuals from a population. The population genealogy will typically involve thousands of meioses versus a few hundred, at most, for linkage analysis using even very large pedigrees.

1.2 Estimating mutation ages

Several recent papers have proposed methods for estimating the age of a mutation with a known location using information from variation at linked genetic markers and the population frequency of the mutation [3]. These methods typically assume that the location of the mutation is known relative to a set of linked markers. Although it has been suggested that it might be possible to jointly estimate mutation ages and locations [5] the complexity of the analysis is such that this has so far not been achieved, except in simplified cases which are generally unrealistic for human populations [6].

1.3 Joint estimation of mutation age and location

In this paper, we explore the use of LD to map genes in the face of an unknown disease allele age and uncertain past population growth rates using Bayesian Markov chain Monte Carlo (MCMC) methods. Previous methods for LD mapping have assumed that one or more of the mutation age, location, or population growth rate parameters are known [7,8,9,10,11] or have made unrealistic assumptions about population demography [12,13,6]. Here, we show that by averaging over the possible age of a disease mutation, using a prior age distribution based on the present population frequency, estimates of mutation location may be obtained that are quite insensitive to the population growth

rate; this is encouraging since the past growth rates are usually poorly known and a less parameterized model is therefore desirable.

2 Theory

Following standard terminology [1] we define a genetic locus to be a specific physical position in the nucleotide sequence of a chromosome. An allele is defined to be a variant of a locus with one or more nucleotides altered by point mutation, nucleotide insertion or deletion, etc. Define $\mathbf{X}$ to be a matrix of haplotypes for a specified set of marker loci (i.e., phase-determined alleles for the markers) for a sample of chromosomes bearing a disease mutation of unknown location. Let $\mathbf{Z}$ be a matrix of the marker haplotypes from a random (ethnically matched) sample of normal individuals. Define $\mathbf{p}$ to be a matrix of the (unknown) marker allele frequencies in the population of normal chromosomes. Let τ be the unobserved ancestral genealogy underlying the sample of disease chromosomes, $\mathbf{Y}_{-0}$ be a matrix of the ancestral haplotypes in the genealogy, $\mathbf{Y}_0$ be the (unknown) ancestral haplotype on which the disease mutation first arose, and t_0 be the (unknown) age of the mutation. Define θ to be the position (in Morgans) of the mutation relative to marker locus 1, Θ to be a vector of genetic parameters, such as the map distances among marker loci, etc, and Λ to be a vector of the demographic parameters, including the fraction of the population of disease chromosomes sampled, f, and the population growth rate, r (assuming exponential growth). In this paper, we consider haplotype data but our method also can be used with genotype data under simple models of inheritance [14]

2.1 Likelihood and prior distributions of parameters

The method we present is an extension of the Bayesian LD mapping method of Rannala and Reeve [11,14]. Details of the likelihood and the priors for parameters other than t_0 can be found in the earlier papers. Here, we focus on the addition of a prior for t_0 and the steps involved in integrating over this prior. The likelihood of the sampled disease haplotypes and the (unobserved) ancestral haplotypes is

$$f(\mathbf{X}, \mathbf{Y}_{-0}|\theta, \Theta, \tau, \mathbf{p}, \mathbf{Y}_0, t_0).$$

The prior probability density that we use for t_0 is proportional to the likelihood of the observed sample frequency, i, of the disease allele. From Slatkin and Rannala [15] this is

$$f(t_0|i, r, N, q) \propto \left\{ \frac{1 - e^{-rt_0}}{1 - e^{-rt_0} + e^{-rt_0}(2Nqr/i)} \right\}^{i-1} \frac{1}{1 - i/(2Nqr) + e^{rt_0}/(2r)},$$

where we have substituted $f = i/(Nq)$, where i is the number of disease chromosomes in the sample, N is the population size and q is the relative population frequency of disease chromosomes (e.g., estimated based on disease incidence and mode of inheritance). The above equation is sometimes multiplied by an additional term, e^{-rt_0} to take account of the decreased influx of mutations at time t_0 in the past due to the smaller past population size under a model of exponential population growth [16] but this addition has little effect on estimates of disease location for the examples we present below.

2.2 Posterior distribution of parameters

The joint posterior density of the parameters is given by

$$f(\theta, \mathbf{Y}, \tau, \mathbf{p}, t_0 | \mathbf{X}, \mathbf{Z}, \Theta, N, q, r, i) = \frac{f(\mathbf{X}, \mathbf{Y}_{-0} | \theta, \Theta, \tau, \mathbf{p}, \mathbf{Y}_0, t_0) f(\tau | N, q, r) f(\mathbf{Y}_0 | \mathbf{p}) f(\mathbf{Z} | \mathbf{p}) f(\theta) f(\mathbf{p}) f(t_0 | i, N, q, r)}{f(\mathbf{X}, \mathbf{Z} | \Theta, N, q, r, i)},$$

where the priors for the variables other than t_0 are as given in Rannala and Reeve [11]. We used a uniform prior for the position of the disease mutation, θ, although a prior based on an annotated human genome sequence and a mutation database could also have been used [11].

Markov chain Monte Carlo (MCMC) methods were used to generate the joint posterior density of the parameters in the above equation based on a Metropolis-Hastings algorithm. The basic principle of MCMC analysis is to simulate observations from a Markov chain with a stationary distribution that is the joint posterior probability density of the parameters. The joint and marginal posterior densities can be estimated by running this chain on a computer until it converges and then sampling parameter values from the chain at equal intervals. The above method has been implemented in version 2.2 of the LD mapping program DMLE+ available for downloading from dmle.org.

3 Examples

To illustrate the method, we apply it to two data sets for which haplotypes are available and a disease mutation has been cloned. This will allow us to directly test the accuracy of the method using relevant empirical data. The first data set that we examine is for a common mutation causing cystic fibrosis (CF) in europeans, the $\Delta F508$ mutation; the second data set we examine is for a founder mutation in Finland that causes diastrophic dysplasia (DTD). All analyses were carried out using DMLE+ version 2.2.

3.1 Cystic fibrosis $\Delta F508$ *mutation in Europe*

The $\Delta F508$ mutation is the most common cause of CF in european populations, accounting for roughly 70 percent of CF mutations. The data set we analyze was originally used to map the CF gene [17]. In total, 63 chromosomes from this data set carry the $\Delta F508$ mutation. We excluded one of these chromosomes from our analysis as it appears to belong to a very different haplogroup and possibly represents a recurrent mutation. The chromosomes were typed for 23 biallelic markers (RFLPs) that span 1.8 Mb, with the mutation located 880 kb from marker 1. The two closest markers to the mutation are at 869.8 and 889.8 kb, respectively. Haplotype phase was inferred via linkage analysis of relatives of probands. Approximately 6 percent of the markers in the disease chromosomes are missing data; we integrate over the missing markers using data augmentation in our algorithm [14].

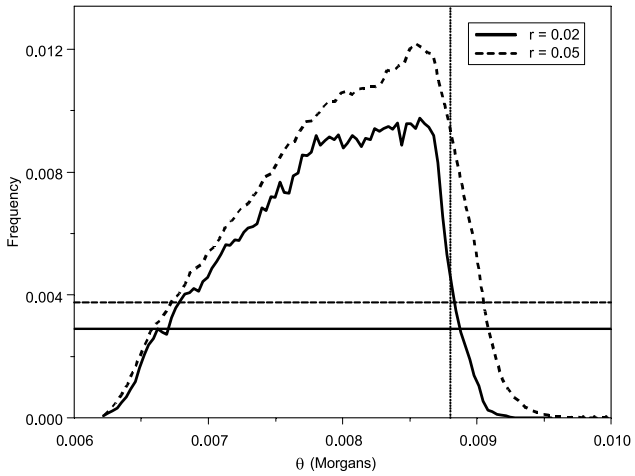

Figure 1: The posterior probability density of the position, θ, of the CF mutation $\Delta F508$ relative to marker 1. A total of 23 markers were typed for 62 disease chromosomes carrying the $\Delta F508$ mutation. Separate analyses were carried out using two different population growth rates $r = 0.05$ (dashed line) and $r = 0.02$ (solid line). The 95 percent credible set of values for each posterior density is indicated by the (dashed and solid) horizontal lines at the bottom of the figure. The true position of the mutation (assuming 1 cM = 1 Mb) is indicated by the vertical line. The posterior density is little affected by the growth rate. The program is simultaneously estimating the age of the mutation, t_0.

The results of the analysis are shown in Figures 1 and 2. Using either a growth rate of $r = 0.02$ or $r = 0.05$ ($r = 0.05$ is often assumed to be the recent growth rate for european populations) the 95 percent credible set for θ, the

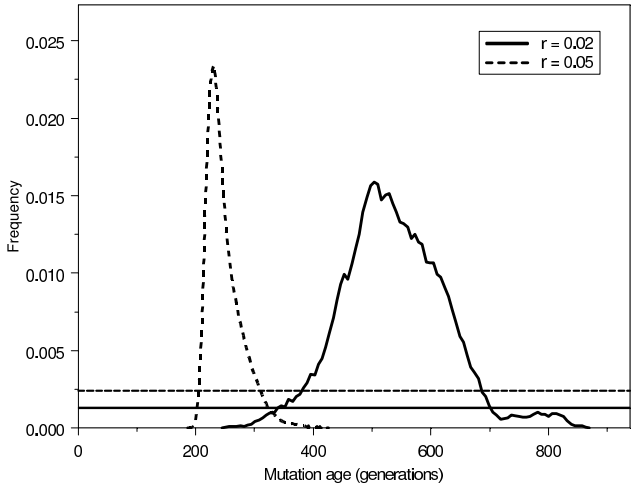

Figure 2: The posterior probability density of the age, t_0, of the CF mutation $\Delta F508$. A total of 23 markers were typed for 62 disease chromosomes carrying the $\Delta F508$ mutation. Separate analyses were carried out using two different population growth rates $r = 0.05$ (dashed line) and $r = 0.02$ (solid line). The 95 percent credible set of values for each posterior density is indicated by the (dashed and solid) horizontal lines at the bottom of the figure. The posterior density is little affected by the growth rate. The program is simultaneously estimating the position of the mutation, θ.

position of the mutation relative to marker 1, ranges from roughly .65 cM to either 0.9 cM (for $r = 0.02$) or 0.925 cM (for $r = 0.05$) as shown in Figure 1. In both cases, the true position of the mutation (0.88 cM assuming 1 cM = 1 Mb) is bracketed by the 95 percent credible set. The growth rate parameters have a much larger effect on the posterior density of the allele age parameter, t_0 as shown in Figure 2. In that case, the 95 percent credible set is (200,310), for $r = 0.05$, and is (350,700) for $r = 0.02$ (in units of generations). This indicates that estimates of allele ages are highly sensitive to inferred population growth rates but estimates of mutation location are much less sensitive to the population growth rate if one integrates over the mutation age.

3.2 Diastrophic dysplasia mutation in Finland

DTD is an autosomal recessive disease with an unusually high frequency (roughly 2 percent are carriers) in the Finnish population, probably due to a founder event that occurred about 2,000 years ago (100 generations). We used 5 markers (2 RFLPs and 3 microsatellites) from an unpublished data set provided by

J. Hästbacka that was used to clone the DTD gene [18]. This data set was previously analyzed using our DMLE+ program and fixing the age of t_0 to be 100 generations [11]. The data set is comprised of 148 disease chromosomes typed for the 5 markers and 126 controls. There are no missing marker genotypes. The five markers span 20 kb with the mutation located roughly 86 kb centromeric to marker 1. Haplotype phase was inferred via linkage analysis of relatives of probands. A growth rate of $r = 0.085$ was used based on records of the demographic expansion in Finland [11].

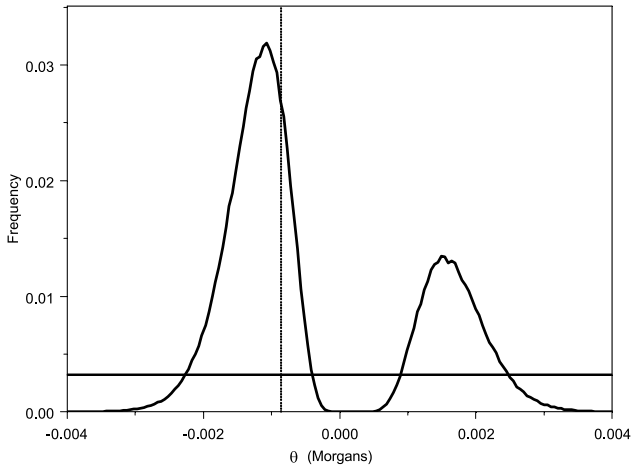

Figure 3: The posterior probability density of the position, θ, of the DTD mutation relative to marker 1. A total of 5 markers were typed for 148 disease chromosomes carrying the DTD mutation. Analyses were carried out using a population growth rate of $r = 0.085$. The 95 percent credible set of values is indicated by the horizontal line at the bottom of the figure. The true position of the mutation (assuming 1 cM = 1 Mb) is indicated by the vertical line. the program is simultaneously estimating the age of the mutation, t_0.

The results of the analysis are shown in Figures 3 and 4. Integrating over the allele age has little effect on the posterior density of θ in this case. Figure 3 shows the posterior density of θ. This density is nearly identical to the posterior density that is obtained by assuming $t_0 = 100$ (see e.g., Figure 4 of Rannala and Reeve [11]). As in the earlier analysis, the 95 percent credible set for the location of the DTD mutation includes the true location (assuming 1 cM = 1 Mb). Figure 4 shows the posterior density of the mutation age. This density has a mode of roughly 80 generations with a 95 percent credible set of ages ranging from 65 generations to 105 generations; this range includes the postulated time of the founding of the Finnish population roughly 100

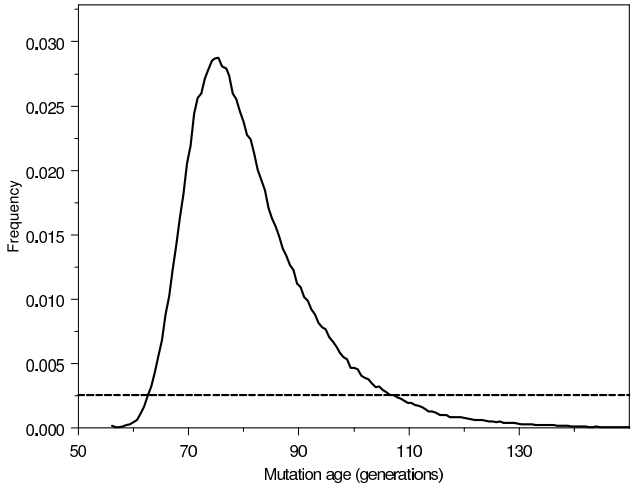

Figure 4: The posterior probability density of the age, t_0, of the DTD mutation . A total of 5 markers were typed for 148 disease chromosomes carrying the DTD mutation. Analyses were carried out using a population growth rate of $r = 0.085$. The 95 percent credible set of values for the posterior density is indicated by the horizontal line at the bottom of the figure. The program is simultaneously estimating the position of the mutation, θ.

generations ago supporting the idea that the mutation was introduced at the founding event. Thus, the analysis of the DTD data set supports the idea that it is possible to jointly estimate mutation location and age as both the estimates obtained in this case appear quite reasonable.

4 Discussion

In this paper, we have explored the feasibility of jointly estimating the age of a disease mutation and its location using multilocus marker haplotypes for a sample of chromosomes bearing a disease mutation and a sample of normal chromosomes. We have used MCMC methods, extending our previously developed LD mapping program DMLE+. In the case of the two data sets analyzed, one for the DTD mutation in Finland and another for the CF $\Delta F508$ mutation, the method performs well, allowing joint estimates of the two parameters and leading to reasonable results for mutation ages and successfully locating the disease mutation with 95 percent probability, One new finding that is particularly encouraging is that the posterior density of the mutation location appears insensitive to the (unknown) population growth rate when

the method integrates over possible allele ages. Thus, fewer assumptions may be needed for parametric LD-based disease mutation mapping than was previously thought. The estimate of mutation age, on the other hand, appears quite sensitive to the assumed growth rate and should therefore be interpreted with greater caution in future, perhaps instead considering values for the estimated mutation age based on a range of plausible population growth rates.

Acknowledgments

This research was supported by grants to B. Rannala from the Alberta Heritage Foundation for Medical Research, the CIHR (MOP 44064), the Peter Lougheed Foundation (CIHR-PLS 47851), and the NIH (R01-HG01988). J. Reeve was supported by a Killam Memorial Postdoctoral Fellowship.

References

1. J.H. Gillespie *Population Genetics: A Concise Guide* (John Hopkins Press, Baltimore, 1998).
2. E.S. Lander and D. Botstein, *Cold Spring Harb. Symp. Quant. Biol.* **51**, 49 (1986).
3. M. Slatkin and B. Rannala, *Annu. Rev. Genomics Hum. Genet.* **1**, 225 (2000).
4. S.A. Tishkoff, et al., *Science* **271**, 1380 (1996).
5. S-W. Guo and M. Xiong, *Hum. Hered.* **47**, 315 (1997).
6. A.P. Morris, J.C. Whittaker and D.J. Balding, *Am. J. Hum. Genet.* **70**, 686 (2002).
7. J. Hastbacka et al., *Nat. Genet.* **2**, 204-211 (1992).
8. N.L. Kaplan, W.G. Hill and B.S. Weir, *Am. J. Hum. Genet.* **56**, 18 (1995).
9. B. Rannala and M. Slatkin, *Am. J. Hum. Genet.* **62**, 459 (1998).
10. J. Graham and E.A. Thompson, *Am. J. Hum. Genet.* **63**, 1517 (1998).
11. B. Rannala and J.P. Reeve, *Am. J. Hum. Genet.* **69**, 159 (2001).
12. S.K. Service et. al., *Am. J. Hum. Genet.* **64**, 1728 (1999).
13. A.P. Morris, J.C. Whittaker and D.J. Balding, *Am. J. Hum. Genet.* **67**, 155 (2000).
14. J.P. Reeve and B. Rannala, *Bioinformatics* **18**, 894 (2002).
15. M. Slatkin and B. Rannala, *Am. J. Hum. Genet.* **60**, 447 (1997).
16. M. Slatkin and B. Rannala, *Genetics* **147**, 1855 (1997).
17. B. Kerem et al., *Science* **245**, 1073 (1989).
18. J. Hästbacka et al., *Cell* **78**, 1073 (1994).