BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 6 2002
Pages 894–895

DMLE+: Bayesian linkage disequilibrium gene mapping

*Jeff P. Reeve and Bruce Rannala**

Department of Medical Genetics, University of Alberta, 8–39 Medical Sciences Building, Edmonton, Alberta T6G 2H7, Canada

Received on November 30, 2001; revised on February 1, 2002; accepted on February 13, 2002

ABSTRACT

Summary: The program DMLE+ allows Bayesian inference of the location of a gene carrying a mutation influencing a discrete trait (such as a disease) and/or other parameters of interest (such as mutation age) based on the observed linkage disequilibrium at multiple genetic markers. DMLE+ uses either individual marker genotypes, or haplotypes, integrates over uncertain population allele frequencies, and can incorporate prior information about gene location from an annotated human genome sequence.

Availabilty: DMLE+ is available in both Windows GUI and portable UNIX command line versions at http://dmle.org

Contact: queries@dmle.org; brannala@ualberta.ca

Pedigree-based linkage mapping methods for inferring the position of a disease mutation relative to a set of linked genetic markers (using the inferred frequency of recombination) have quite low resolution (usually less than 1 cM–1 Mb). To carry out positional cloning, or sequencing of candidate genes, greater resolution is needed. A promising approach for high-resolution mapping is linkage disequilibrium (LD) mapping; this technique can map mutations at resolutions of greater than 0.01 cM (roughly 10 kb). The method relies on LD between disease mutations and linked markers in samples of unrelated normal and affected individuals to fine-map. Early methods for LD mapping used a method of moments estimator developed for a single linked marker (Hästbacka *et al.*, 1992). Methods have now been developed for maximum likelihood LD mapping using one or more linked markers (reviewed in Rannala and Slatkin, 2000), providing more accurate estimates of map position. Recently, Bayesian methods for multipoint LD mapping have been proposed using Markov chain Monte Carlo (MCMC) methods (Morris *et al.*, 2000; Rannala and Reeve, 2001) that offer computational advantages and allow more realistic models and prior information to be used.

The Bayesian LD mapping method implemented in DMLE+ version 1.0 (Rannala and Reeve, 2001) calculates the posterior probability density of the parameters as

$$f(\theta, \boldsymbol{Y}, \tau, \boldsymbol{p}|\boldsymbol{X}, \boldsymbol{Z}, \Theta, \Omega, \Lambda) = f(\boldsymbol{X}, \boldsymbol{Y}_{-0}|\theta, \Theta, \tau, \boldsymbol{p}, \boldsymbol{Y}_0) g(\tau|\Lambda) f(\boldsymbol{Y}_0|\boldsymbol{p}) \times f(\boldsymbol{Z}|\boldsymbol{p}) f(\theta|\Omega)\pi(\boldsymbol{p}) / f(\boldsymbol{X}, \boldsymbol{Z}|\Theta, \Omega, \Lambda), \quad (1)$$

where $\boldsymbol{X}$ is a matrix of genotypes (or haplotypes) from individuals displaying a particular phenotype for which a susceptibility mutation is being mapped, $\boldsymbol{Z}$ is a matrix of genotypes from a random (ethnically matched) sample of normal individuals, $\boldsymbol{p}$ is a matrix of the (unobserved) gene frequencies in the population of normal chromosomes, $\boldsymbol{Y}_{-0}$ is a matrix of the ancestral haplotypes in the genealogy relating individuals displaying the phenotype (unobserved random variables), $\boldsymbol{Y}_0$ is the (unobserved) ancestral haplotype on which the mutation first arose, θ is the position of the mutation relative to marker 1, Θ is a vector of genetic parameters such as the map distances among the marker loci, etc, Λ is a vector of demographic parameters such as the population growth rate, population frequency of the mutation, etc, τ is the (unobserved) gene tree underlying the sample of mutation-bearing chromosomes, and Ω is the prior information about the position of the disease mutation available from the positions of introns, exons and non-genic regions specified in an annotated human genome sequence and from a mutation database specifying the observed frequencies of disease mutations in introns, exons, etc. A uniform (Dirichlet) prior is used for $\pi(\boldsymbol{p})$.

The marginal posterior density of any parameter of interest may be obtained by integrating over equation (1) above with respect to the remaining parameters. For example, the posterior density of θ can be used to obtain point estimates and confidence intervals for the position of the disease mutation:

$$f(\theta|\boldsymbol{X}, \boldsymbol{Z}, \Theta, \Omega, \Lambda) = \int_{\boldsymbol{p}} \int_{\tau} \sum_{\boldsymbol{Y}} f(\theta, \boldsymbol{Y}, \tau, \boldsymbol{p}|\boldsymbol{X}, \boldsymbol{Z}, \Theta, \Omega, \Lambda) \mathrm{d}\tau \mathrm{d}\boldsymbol{p}. \quad (2)$$

*To whom correspondence should be addressed.

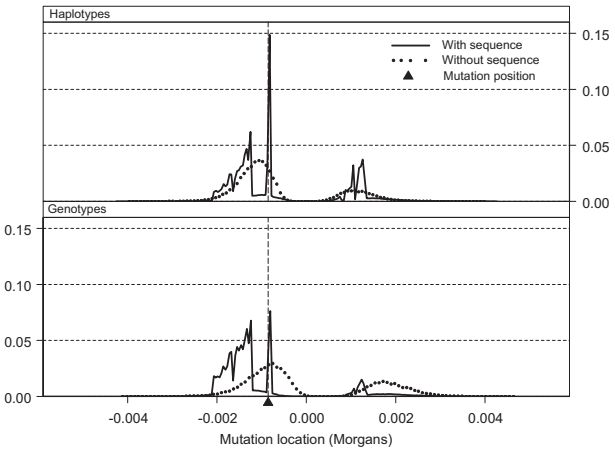

Fig. 1. Summary of the estimated posterior probability distributions for the position of the DTD mutation using five linked markers and either haplotypes (top) or genotypes (bottom). The posterior density of the position of the DTD mutation obtained using a prior distribution from an annotated human genome sequence and the Human Gene Mutation Database (HGMD; http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html) is indicated by a solid line, the posterior distribution without a prior is indicated by a dotted line. The actual position of the DTD mutation is indicated by a dashed vertical line. Haplotype phase information reduces the size of the 95% credible set, in this case, but only by a small amount.

The Metropolis–Hastings algorithm is used to numerically estimate the posterior probability densities in equations (1) and (2). The basic idea is to construct a Markov chain with a stationary distribution that is the joint posterior density of the parameters and implement this chain in a computer program, sampling from the stationary chain to estimate the posterior densities.

DMLE+ release 2.0 incorporates several new features: direct use of genotypes as data (by data augmentation) under simple Mendelian models of inheritance (e.g. dominant and recessive with full penetrance); multiple marker alleles are allowed; missing marker genotypes are allowed (by data augmentation); integration over (unobserved) marker frequencies in normal individuals; and joint estimation of mutation age and map position. DMLE+ release 2.0 also includes automated tools for constructing the prior probability distribution of disease mutation location using introns, exons, etc, from an annotated contig of human genome sequence.

Figure 1 shows the posterior probability density of the position of the DTD mutation that causes diastrophic dysplasia (Hästbacka *et al.*, 1992) inferred (using DMLE+ 2.0) with five linked markers using either complete haplotype phase information (from family pedigrees), or using only individual multilocus genotypes (with, or without, an annotated human genome sequence). The information gained by using haplotypes rather than genotypes narrows the 95% credible set only slightly, with even less effect when human genome sequence data for the region is used. Given the difficulty of collecting relatives of probands to infer haplotype phase this is an encouraging result. The rate of convergence of the MCMC algorithm does not increase substantially when using genotypes rather than haplotypes (for datasets we have analysed). DMLE+ has two parallel distributions: the Windows GUI version includes additional graphical features such as automatic plotting of the log-likelihood, or the parameter values, while the chain runs, and histogram plots estimating the posterior densities of specified parameters; the portable command line version, available for UNIX and other operating systems, instead summarizes the output in a text file.

ACKNOWLEDGMENTS

Support for this research was provided by grants from the Alberta Heritage Foundation for Medical Research, the Canadian Institutes of Health Research (MOP 44064), the Peter Lougheed Foundation (CIHR-PLS 47851), and the National Human Genome Research Institute (NIH, HG01988). Salary support for Jeff Reeve was provided by a Killam Memorial Postdoctoral Fellowship.

REFERENCES

Hästbacka,J., de la Chapelle,A., Kaitila,I., Sistonen,P., Weaver,A. and Lander,E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.*, **2**, 204–211.

Morris,A.P., Whittaker,J.C. and Balding,D.J. (2000) Bayesian fine-scale mapping of disease loci, by hidden Markov Models. *Am. J. Hum. Genet.*, **67**, 155–169.

Rannala,B. and Slatkin,M. (2000) Methods for multipoint disease mapping using linkage disequilibrium. *Genet. Epid.*, **19**, S71–S77.

Rannala,B. and Reeve,J.P. (2001) High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.*, **69**, 159–178.