Am. J. Hum. Genet. 67:155–169, 2000

Bayesian Fine-Scale Mapping of Disease Loci, by Hidden Markov Models

A. P. Morris, J. C. Whittaker, and D. J. Balding

Department of Applied Statistics, University of Reading, Reading, United Kingdom

We present a new multilocus method for the fine-scale mapping of genes contributing to human diseases. The method is designed for use with multiple biallelic markers—in particular, single-nucleotide polymorphisms for which high-density genetic maps will soon be available. We model disease-marker association in a candidate region via a hidden Markov process and allow for correlation between linked marker loci. Using Markov-chain–Monte Carlo simulation methods, we obtain posterior distributions of model parameter estimates including disease-gene location and the age of the disease-predisposing mutation. In addition, we allow for heterogeneity in recombination rates, across the candidate region, to account for recombination hot and cold spots. We also obtain, for the ancestral marker haplotype, a posterior distribution that is unique to our method and that, unlike maximum-likelihood estimation, can properly account for uncertainty. We apply the method to data for cystic fibrosis and Huntington disease, for which mutations in disease genes have already been identified. The new method performs well compared with existing multi-locus mapping methods.

Introduction

The problem in the localization of genes contributing to human diseases has been at the forefront of research in genetic epidemiology for many years now. Linkage-based analyses, often performed in candidate regions of the genome, have had success in locating, to within 1 cM, genes contributing major effects to human disease. However, for genes contributing less significant effects to polygenic disorders, linkage methods have been shown to be less powerful than population-based disease-marker–association studies (Risch and Merikangas 1996).

The key to population-based disease-gene mapping is the relationship between physical distance and the strength of disease-marker association. A higher level of association with the disease at marker *A* than at marker *B* suggests that, in previous generations, less recombination has occurred between the disease gene and marker *A* and thus that this marker is the closer of the two to the disease gene. The simplest approach toward identification of a likely location for a disease gene on a map of candidate marker loci is a *single-locus* approach. On the map, the marker with greatest evidence of association with the disease is taken as being most tightly linked to the predisposing gene.

Received January 27, 2000; accepted for publication April 20, 2000; electronically published June 1, 2000.

Address for correspondence and reprints: Dr. Andrew Morris, University of Reading, Department of Applied Statistics, P.O. Box 240, Earley Gate, Reading RG6 6FN, United Kingdom. E-mail: sns98am@reading.ac.uk

0002-9297/2000/6701-0019$02.00

Greater power and accuracy to locate a disease gene would be expected by taking account of information from all markers, simultaneously, in the region of the disease gene, in so called *multilocus* models. A number of these multilocus methods have been proposed recently (Terwilliger 1995; Xiong and Guo 1997; Collins and Morton 1998) and have had some success in locating the known mutations for cystic fibrosis (CF), Huntington disease (HD), Friedreich ataxia, and progressive myoclonus epilepsy. These methods rely on the assumption of independent marker loci in the region of the disease gene. Under this assumption, log likelihoods are calculated for each marker in turn and summed to form a *composite* log likelihood for the set of loci. This assumption is, of course, incorrect, since we would expect correlation between linked markers. Composite likelihoods are thus only an approximation for the full likelihood obtained by use of complete haplotypes.

In the present report, we present a new multilocus method for the mapping of disease genes, one that takes account of correlation between linked marker loci. The method is designed specifically for use with biallelic markers such as single-nucleotide polymorphisms (SNPs). Current research is likely to provide a highly dense map of SNPs in the near future, in which they are perhaps as frequent as one marker per kilobase of the human genome (Kruglyak 1999).

Consider a disease that, as a result of a single mutation at the disease locus a number of generations ago, was introduced into a population. All affected individuals today will be descended from this founder chromosome. Thus, in a sample of chromosomes ascertained today, the allele that we observe at an SNP linked to the disease gene will depend on whether, at that locus,

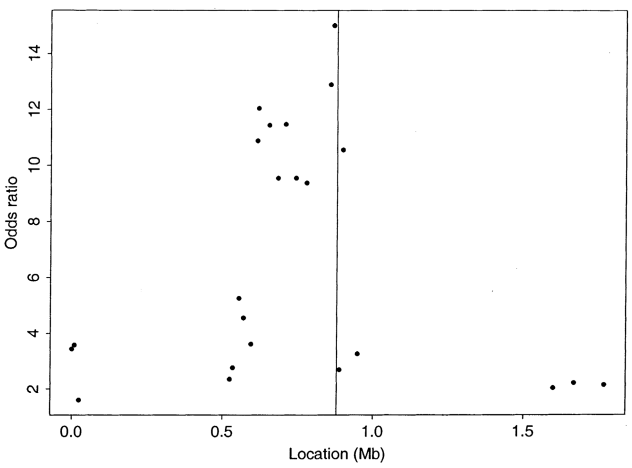

Figure 1 Odds ratios for association with CF, at 23 RFLPs in the region of the CFTR gene on chromosome 7q31 (Kerem et al. 1989).

the chromosome is *identical by descent* (IBD) with the founder chromosome. If, at the marker, the chromosome is IBD with the founder, we observe the ancestral allele. If the chromosome is not IBD with the founder, we may observe either the ancestral or the nonancestral allele, the probability of which will depend on the relative population frequencies of the two alleles. The probability that a chromosome is IBD with the founder will be greater for a marker in the proximity of the disease gene than for more-distant markers, since there will have been less opportunity for recombination. Thus, stronger disease-marker association will be expected at markers adjacent to the disease gene.

For a given location on a chromosome, IBD status itself is not directly observable and can be thought of as a *hidden state*. The probability of changing from one hidden state to another at adjacent markers depends only on previous recombination events between them and thus can be modeled as a function of the physical intermarker distance. For a fine-scale map of markers, this probability will be small, and if it assumed that there is no interference, will be independent of similar probabilities defined in any other interval between adjacent markers. Under these conditions, we can employ a hidden Markov model (Rabiner 1989) to describe marker haplotype frequencies in the vicinity of the disease gene, accounting for the correlation between linked marker loci.

The model that we present here is similar to that of McPeek and Strahs (1999), who also use a hidden Markov process to account for correlation between linked marker loci. We employ Markov-chain Monte Carlo (MCMC) stochastic simulation methods in a Bayesian framework, which has a number of advantages over the maximum-likelihood approach used by McPeek and Strahs (1999). We are able to properly account for the uncertainty in the ancestral marker haplotype—unlike the method of McPeek and Strahs (1999), which treats it as a nuisance parameter to be estimated. With this approach, we obtain posterior distributions for model parameter estimates, including disease-gene location and the age of the mutation (for a complete list of the model parameters used in the present study, see Appendix A). In addition, the flexibility of this framework allows us to incorporate heterogeneity in recombination rates in the region of the disease gene, to account for crossover hot and cold spots. We apply the method to data for CF (Kerem et al. 1989) and HD (MacDonald et al. 1991), for which mutations in disease genes already have been identified.

Models and Methods

In this section, we derive a model for disease-marker association in a candidate region, using hidden Markov processes. We begin with the simplest case—of a single founding mutation of a normal allele at the disease locus to a high-risk allele. Any chromosome in the current generation can be divided into regions, each corresponding to one of two possible ancestral states. A region may be IBD with the ancestral founder chromosome and is then labeled "F"; otherwise, the region does not descend from the founder and is labeled "N." The probability that, at any given locus, a chromosome in the current generation is IBD with the founder is denoted as "α."

The occurrence of the different ancestral states, F or N, along a chromosome is a result of recombination events in previous generations. Consider two particular loci on a chromosome selected at random from the current generation. Given the chromosome's ancestral state at locus 1, it is straightforward to calculate the probabilities of the two ancestral states at locus 2. Let "NR" denote the event "no recombination has occurred between the loci"; then, for example,

$$\begin{aligned}&\Pr(\text{locus } 2 = \text{F}|\text{locus } 1 = \text{F})\\ &\quad = \Pr(\text{NR}) + [1 - \Pr(\text{NR})]\Pr(\text{MRR} = \text{F}) ,\end{aligned} \qquad (1)$$

where MRR = F is used to denote the event "most recent recombination event occurred, at locus 2, with a chromosome IBD with the founder"; similarly,

$$\begin{aligned}&\Pr(\text{locus } 2 = \text{F}|\text{locus } 1 = \text{N})\\ &\quad = [1 - \Pr(\text{NR})]\Pr(\text{MRR} = \text{F}) ,\end{aligned} \qquad (2)$$

since a recombination event must have occurred between two loci of different ancestral states. We assume that the probability that, at locus 2, a chromosome is

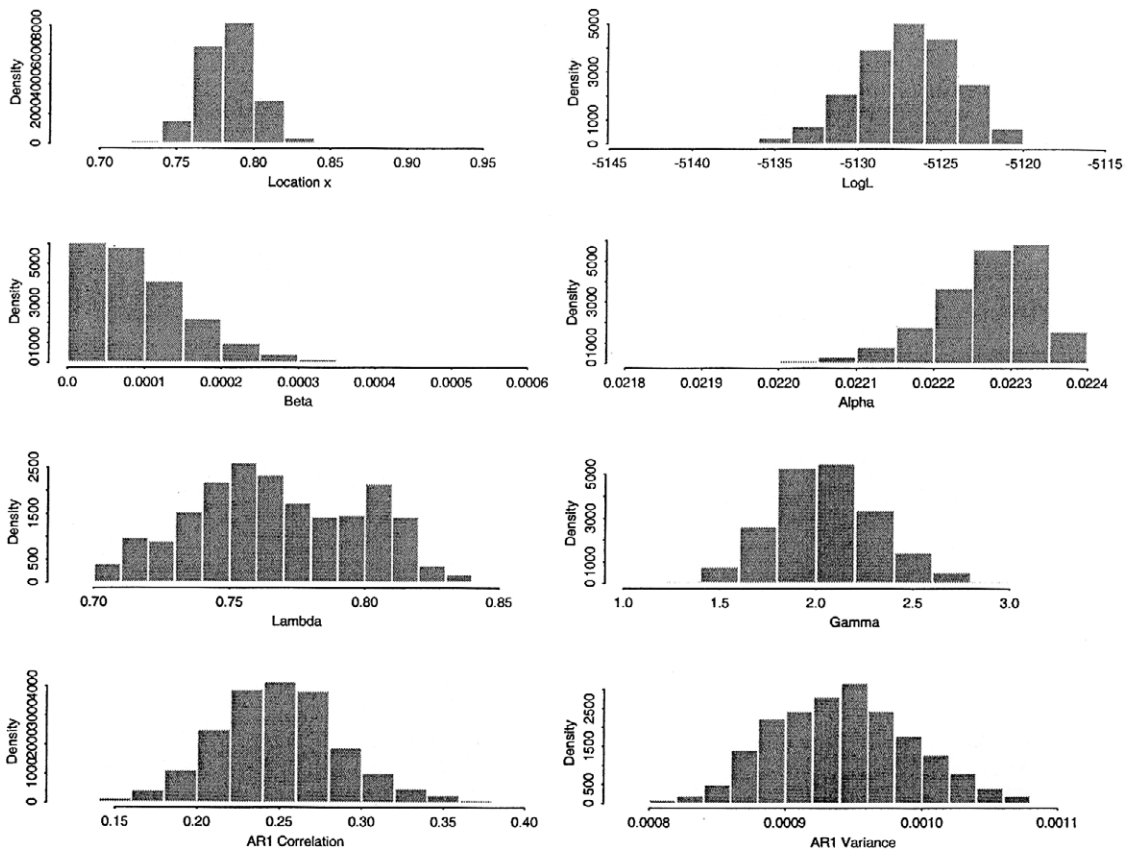

Figure 2 Posterior distributions of model parameter estimates for the CF data of Kerem et al. (1989), under the assumption of independent recombinational histories for case chromosomes. Estimates are obtained from every 100th of 1 million iterations of the Metropolis-Hastings rejection-sampling scheme, after an initial burn-in period.

IBD with the founder has remained constant over time, Pr(MRR = F) = α.

This principle can be generalized to more than two loci on the chromosome. The probabilities of the two ancestral states at any locus on the chromosome, given the ancestral state at an adjacent locus on the chromosome, depend only on recombination events between the loci and not on recombination elsewhere along the chromosome (under the assumption of no interference). Thus, given the ancestral state at some starting locus of a chromosome, we can calculate joint probabilities of ancestral states at any other loci, using two independent Markov chains, one acting on each side of the starting locus.

Consider a map of SNPs with known location in a candidate region of the chromosome and assume an arbitrary location x for the disease locus, 0, on this map. Given this location, the map is effectively divided into two regions with "L" markers present to the left of the disease locus and "R" markers present to the right. The marker loci to the left of the disease locus are denoted "$-1, -2, \ldots, -$L," where -1 is adjacent to the disease locus, -2 is adjacent to -1, and so on; similarly, the marker loci to the right of the disease locus are denoted "1, 2, …, R." The physical distances (in Mb) between the disease locus and marker loci -1 and 1 are denoted "d_{-1}" and d_1, respectively. The distance between any pair of adjacent marker loci to the left of the disease locus, $-i$ and $-(i+1)$, is denoted "$d_{-(i+1)}$"; similarly, $d_{(i+1)}$ denotes the distance between marker loci i and $(i+1)$ to the right of the disease locus. The choice of location of the disease locus, x, thus defines a unique set of interlocus distances.

A chromosome can be considered as two independent paths of ancestral states, conditional on the ancestral state at the disease locus, S_0. For the marker loci to the left of the disease locus, the path is denoted "$\mathcal{S}_L = \{S_0, S_{-1}, S_{-2}, \ldots, S_{-L}\}$," whereas, for marker loci to the right, the path is denoted "$\mathcal{S}_R = \{S_0, S_1, S_2, \ldots, S_R\}$."

Consider the marker loci to the right of the disease locus. The chromosome's ancestral state at locus $i+1$, S_{i+1}, depends only on both the chromosome's ancestral state at locus i, S_i, and the occurrence of previous generations of recombination events between the pair of

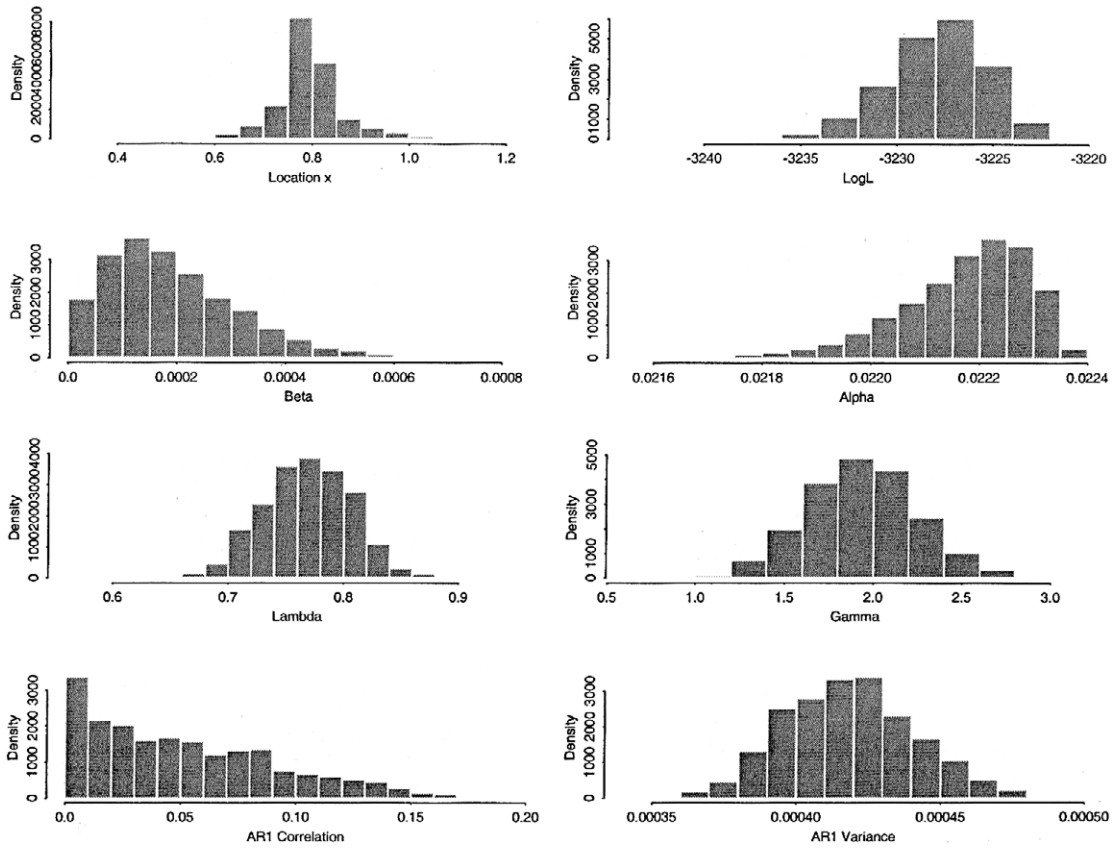

Figure 3 Posterior distributions of model parameter estimates for the CF data of Kerem et al. (1989), under the assumption of a conditional coalescent model for dependence between case chromosomes. Estimates are obtained from every 100th of 1 million iterations of the Metropolis-Hastings rejection-sampling scheme, after an initial burn-in period.

adjacent loci. In the same way as for equations (1) and (2), we define *transition probabilities* $\tau_{S_i S_{i+1}}^{i+1}$ of ancestral state S_{i+1}, at locus $i+1$, given the chromosome's ancestral state, S_i:

$$\begin{cases} \tau_{\mathrm{FF}}^{i+1} &= \exp(-\gamma d_{i+1}) + [1-\exp(-\gamma d_{i+1})]\alpha \\ \tau_{\mathrm{FN}}^{i+1} &= [1-\exp(-\gamma d_{i+1})](1-\alpha) \\ \tau_{\mathrm{NF}}^{i+1} &= [1-\exp(-\gamma d_{i+1})]\alpha \\ \tau_{\mathrm{NN}}^{i+1} &= \exp(-\gamma d_{i+1}) + [1-\exp(-\gamma d_{i+1})](1-\alpha) \end{cases}$$

Here, $\exp[-\gamma d_{i+1}]$ is the probability of no recombination events in generations since the founding mutation in the interval between marker loci i and $i+1$. The parameter $\gamma > 0$ represents the expected frequency, since the founding mutation, of recombination events per 1 Mb of a chromosome in the candidate region. If we assume that the physical distance of 1 Mb corresponds to a genetic distance of 1 cM, then 100γ can be interpreted as the number of generations since the founding mutation.

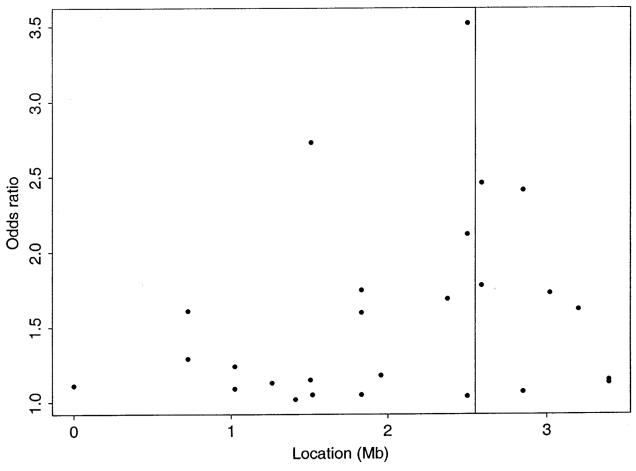

Figure 4 Odds ratios for association with HD, at 27 RFLPs in the region of the IT15 gene on chromosome 4p16 (MacDonald et al. 1991).

Given values for the transition parameters α and γ, we can calculate the probability, $\rho[S_i|S_0]$, that a chromosome is of ancestral state S_i at marker locus i, conditional on the chromosome's ancestral state at the disease locus, S_0, using the recursive formula

$$\rho[S_i|S_0] = \sum_{S_{i-1}=\mathrm{F,N}} \rho[S_{i-1}|S_0]\tau^i_{S_{i-1}S_i} ,$$

for all $i > 1$, and $\rho[S_1|S_0] = \tau^1_{S_0S_1}$. Ancestral-state frequencies at loci to the left of the disease locus are calculated in the same way, on the basis of an independent Markov process defined in terms of the same model parameters α and γ.

The model described thus far can be used to calculate the probability that, at any marker locus in the candidate region, a chromosome is IBD with the founder, given the chromosome's ancestral state at the disease locus, S_0. Of course, ancestral states are hidden and cannot be observed. At each SNP in the candidate region, one of two possible alleles, denoted as "M_{i1}" and "M_{i2}," can occur at marker locus i. The marker allele present is dependent only on the chromosome's ancestral state at marker locus i and not on that elsewhere in the region. If, at marker locus i, the chromosome is IBD with the founder chromosome, then the allele present will be the same allele that is present on the founder chromosome, if it is assumed that no mutations have occurred at the marker locus; if the chromosome is not IBD with the founder, then either allele may be present, with probability p_i denoting the frequency of allele M_{i1} on such chromosomes. Thus, given that a chromosome

Table 1

Expected Frequencies of Alleles M_{i1} and M_{i2} at Marker Locus i, for Known Disease Frequency Q, Sample-Enrichment Factor κ, and $T = 1 + Q(\kappa - 1)$

	Frequency		
Sample	M_{i1}	M_{i2}	Total
Cases	$\kappa\phi^{-f}_{i1}/T$	$\kappa\phi^{-f}_{i2}/T$	$Q\kappa/T$
Controls	ϕ^{u}_{i1}/T	ϕ^{u}_{i2}/T	$(1 - Q)/T$
Total			1

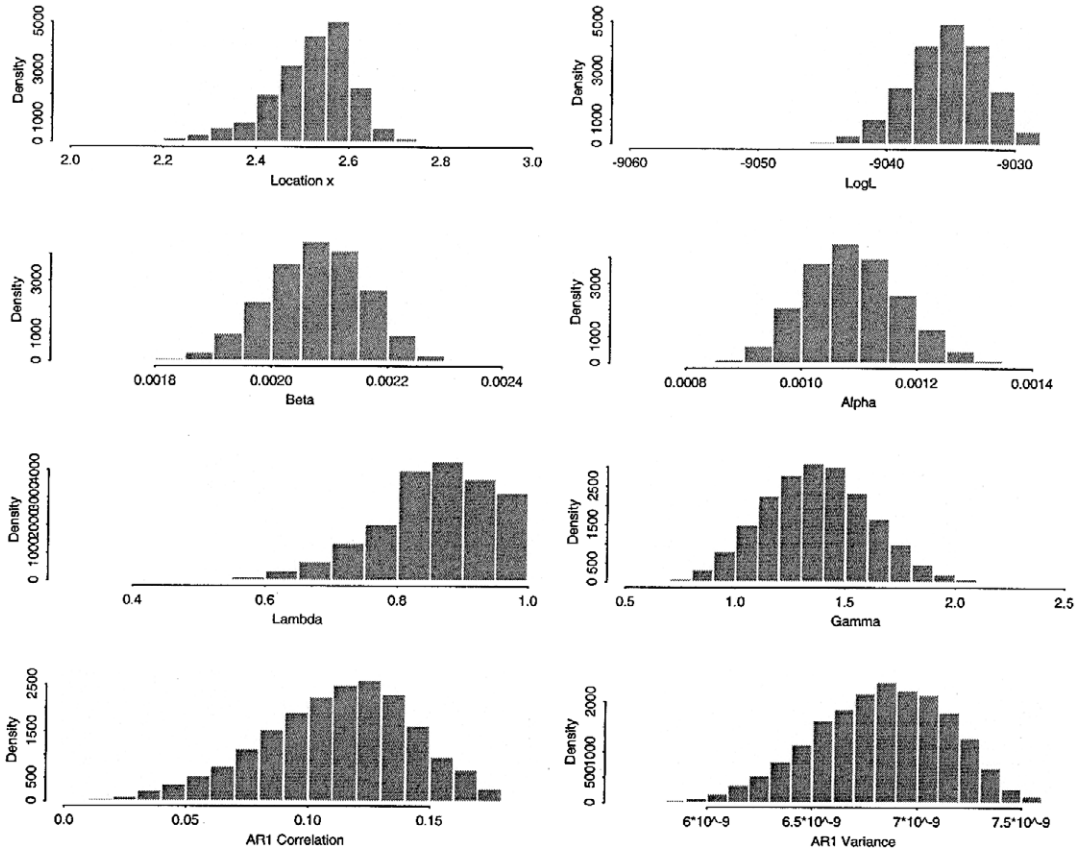

Figure 5 Posterior distributions of model parameter estimates for HD data of MacDonald et al. (1991). Estimates are obtained from every 100th of 1 million iterations of the Metropolis-Hastings rejection-sampling scheme, after an initial burn-in period.

Table 2

Model Parameter Estimates for Complete CF Data of Kerem et al. (1989)

Parameter	True Value	Initial Value	Estimate	99% Credibility Interval
Location x	.88	.6	.784	.731–.838
β	0	.1	9.59×10^{-5}	8.20×10^{-7}–3.47×10^{-4}
α	2.24×10^{-2}	...	2.23×10^{-2}	2.20×10^{-2}–2.24×10^{-2}
λ	.7	.9	.768	.703–.836
γ	...	1	2.05	1.37–2.86
ϱ	...	0	.250	.151–.361
σ^2	...	.1	9.42×10^{-4}	8.18×10^{-4}–1.07×10^{-3}

NOTE.—Estimates are obtained from every 100th of 1 million iterations of the Metropolis-Hastings rejection-sampling scheme after an initial burn-in period.

is of ancestral state S_0, the expected frequency of allele M_{i1} is given by

$$m_{i1}^{S_0} = \omega_i \rho[S_i = \mathrm{F}|S_0] + p_i \rho[S_i = \mathrm{N}|S_0] \ . \qquad (6)$$

The parameter ω_i is an indicator variable taking the value 1 if allele M_{i1} is present on the founder chromosome and taking the value 0 otherwise. Clearly,

$$m_{i2}^{S_0} = 1 - m_{i1}^{S_0} = (1 - \omega_i)\rho[S_i = \mathrm{F}|S_0] + (1 - p_i)\rho[S_i = \mathrm{N}|S_0] \ . \qquad (7)$$

We have assumed here that, conditional on a set of adjacent marker loci being in state N, the probability of the observed haplotype is simply the product of the allele frequencies p_i or $1 - p_i$, at marker i, from which it is constructed. McPeek and Strahs (1999) have suggested the use of a kth-order (in practice, $k = 1$) Markov-chain model for haplotype frequencies across loci in state N. Such an approach could be easily incorporated into the model presented here.

Consider a sample of n_{A} chromosomes obtained from affected cases and n_{U} chromosomes obtained from unaffected controls. We do not assume here that we can identify homologous pairs of chromosomes occurring together in the same individual in the sample. If this information is known, it can be easily incorporated in the analysis.

We cannot directly identify the chromosome's ancestral state at the disease locus. Instead, we observe the disease phenotype $\mathcal{P}$ of the individual from whom it is obtained, assumed here to be either affected ($\mathcal{P} = \mathcal{A}$) or unaffected ($\mathcal{P} = \mathcal{U}$). The disease phenotype of an individual depends on the ancestral state at the disease locus on their pair of homologous chromosomes. Since we do not assume that we can identify homologous pairs of chromosomes occurring together in the same individual in the sample, we average over the possible ancestral states, S_0', at the disease locus for the second chromosome, weighting by their relative frequencies:

$$\Pr(\mathcal{P}|S_0) = \Pr(\mathcal{P}|S_0, S_0' = \mathrm{F})\alpha + \Pr(\mathcal{P}|S_0, S_0' = \mathrm{N})(1 - \alpha) \ . \qquad (8)$$

We assume a multiplicative model for the disease, with parameters β_{F} and β_{N} for the ancestral states $S_0 = \mathrm{F}$ and $S_0 = \mathrm{N}$, respectively, at the disease locus. Thus, the penetrance of genotype $S_0 S_0'$ is given by $\Pr(\mathcal{P} = \mathcal{A}|S_0 S_0') = \beta_{S_0}\beta_{S_0'}$. Hence,

$$\begin{cases} \Pr(\mathcal{P} = \mathcal{A}|S_0 = \mathrm{N}) = \beta_F\beta_{\mathrm{N}}\alpha + \beta_{\mathrm{N}}^2(1 - \alpha) = f_{\mathrm{N}} \\ \Pr(\mathcal{P} = \mathcal{A}|S_0 = \mathrm{F}) = \beta_F^2\alpha + \beta_F\beta_{\mathrm{N}}(1 - \alpha) = f_F \\ \Pr(\mathcal{P} = \mathcal{U}|S_0 = \mathrm{N}) = (1 - \beta_F\beta_{\mathrm{N}})\alpha + (1 - \beta_{\mathrm{N}}^2)(1 - \alpha) = 1 - f_{\mathrm{N}} \\ \Pr(\mathcal{P} = \mathcal{U}|S_0 = \mathrm{F}) = (1 - \beta_F^2)\alpha + (1 - \beta_F\beta_{\mathrm{N}})(1 - \alpha) = 1 - f_F \end{cases} . \qquad (9)$$

Under this model, we can calculate expected SNP frequencies in affected and unaffected individuals in the population. As an example, consider marker locus i. The probability that a chromosome is obtained from an individual of disease phenotype $\mathcal{P}$ and bears allele M_{ij} at marker locus i is denoted by "$\phi_{ij}^{\mathcal{P}}$." Then,

Table 3

Posterior Probabilities of Ancestral-State Haplotypes at 23 RFLPs in the Region of the Δ*F*508 Mutation

Ancestral Haplotype	Posterior Probability
11112221112212121121111	.980
11212221112212121121111	.005
11112221112212121121211	.005
11112221112212121121112	.005
11112221112212121121121	.003
Other	.002

NOTE.—Dependence between case chromosomes is modeled by the conditional coalescent (McPeek and Strahs 1999).

Table 4

Model Parameter Estimates for HD Data of MacDonald et al. (1991)

Parameter	True Value	Initial Value	Estimate	99% Credibility Interval
Location x	2.5–2.6	2.2	2.52	2.20–2.75
β	>0	.1	2.08×10^{-3}	1.84×10^{-3}–2.28×10^{-3}
α	…	…	1.09×10^{-3}	8.87×10^{-4}–1.32×10^{-3}
λ	…	.9	.854	.564–.998
γ	…	1	1.37	.772–2.06
ϱ	…	0	.111	.024–.177
σ^2	…	.1	6.88×10^{-9}	6.00×10^{-9}–7.60×10^{-8}

NOTE.—See Note to table 3.

$$\begin{aligned}\phi_{ij}^{\mathcal{P}} &= \Pr(\mathcal{P} \cap M_{ij}) \\ &= \Pr(\mathcal{P} \cap M_{ij} | S_0 = \mathrm{F})\alpha + \Pr(\mathcal{P} \cap M_{ij} | S_0 = \mathrm{N})(1-\alpha) \\ &= \Pr(\mathcal{P} | S_0 = \mathrm{F}) m_{ij}^{F} \alpha + \Pr(\mathcal{P} | S_0 = \mathrm{N}) m_{ij}^{\mathrm{N}} (1-\alpha) \ , \end{aligned} \tag{10}$$

since disease status and SNP type are independent, conditional on the chromosome's ancestral state at the disease locus. Thus, when we substitute for the appropriate $\Pr(\mathcal{P}|S_0)$ from equation (9) and for $m_{ij}^{S_0}$ from equations (6) and (7),

$$\begin{aligned}\phi_{i1}^{\mathcal{A}} &= m_{i1}^{F} f_F \alpha + m_{i1}^{\mathrm{N}} f_{\mathrm{N}} (1-\alpha) \ , \\ \phi_{i2}^{\mathcal{A}} &= m_{i2}^{F} f_F \alpha + m_{i2}^{\mathrm{N}} f_{\mathrm{N}} (1-\alpha) \ , \\ \phi_{i1}^{U} &= m_{i1}^{F} (1-f_F) \alpha + m_{i1}^{\mathrm{N}} (1-f_{\mathrm{N}})(1-\alpha) \ , \\ \phi_{i2}^{U} &= m_{i2}^{F} (1-f_F) \alpha + m_{i2}^{\mathrm{N}} (1-f_{\mathrm{N}})(1-\alpha) \ . \end{aligned}$$

In a case-control study, affected individuals are ascertained with greater probability than is their population frequency, so that a sample will be enriched with case chromosomes. We denote by $n_{ij}^{\mathcal{P}}$ the observed frequencies of allele M_{ij} in the sample of chromosomes obtained from individuals of disease phenotype $\mathcal{P}$. Table 1 presents the expected case-control frequencies of SNP alleles at marker locus i. The parameter Q is the population frequency of the disease, which is assumed to be known, and κ is a sample-enrichment factor: $\kappa = \{[(1-Q)n_A]/Qn_U\}$. The expected frequencies are scaled by the parameter $T = 1 + Q(\kappa - 1)$ to sum to 1.

Allowing for Mutations at Marker Loci

In deriving the model thus far, we have assumed that no mutations at marker loci have occurred since the founding disease mutation on the ancestral chromosome. The method is designed for use with SNPs, which are thought to have low mutation rates in humans, $\sim 10^{-8}$–10^{-9}/locus/generation (Nielsen 2000). For recent disease mutations, the effects of such a low rate of mutation will be negligible. Nevertheless, we may wish for the model to account for marker mutation.

Under the assumption of no marker mutation, we observe, at that locus, only the ancestral allele at marker i on a chromosome IBD with the founder. However, if we allow for marker mutation, we may observe the nonancestral allele at a locus in state F. In terms of the indicator parameter for marker i,

$$\begin{cases} (1-m)^{100\gamma} & \text{if allele } M_{i1} \text{ present on founder chromosome} \\ 1-(1-m)^{100\gamma} & \text{otherwise} \end{cases} ,$$

where m is the mutation rate per locus per generation and 100γ is the number of generations since the founding mutation.

Allowing for Phenocopies

The model described has assumed, thus far, that all mutant chromosomes have descended from a single ancestral founder. This assumption is unlikely to be realistic for most human diseases (Penisi 1998). Phenocopies may occur either as a result of multiple mutations in the same gene or, especially for complex diseases, as a result of the effects of multiple susceptibility loci and the environment. In this section, we develop the model to allow for phenocopies, under the assumption that there is a single major mutation that accounts for a substantial proportion of affected individuals in the current generation. This is true, for example, of CF, for which the major ΔF508 mutation in the CFTR gene accounts for almost 70% of all chromosomes in affected individuals, with many other mutations in the same gene accounting for the remaining 30% (Kerem et al. 1989). Previous approaches, with the exception of that of McPeek and Strahs (1999), fail to explicitly allow for this in their association models.

Assume that the major mutation (F) accounts for a proportion λ of all mutant chromosomes in the current generation and that the major mutation and all other mutations ($\bar{F}$) have the same penetrance, β_F. If equation

(8) is generalized to allow for three possible ancestral states,

$$\begin{aligned}\Pr(\mathcal{P} = \mathcal{A}|S_0) = \Pr(\mathcal{P} = \mathcal{A}|S_0,S_0' = \mathrm{F})\alpha\lambda \\ +\Pr(\mathcal{P} = \mathcal{A}|S_0,S_0' = \overline{F})\alpha(1-\lambda) \\ +\Pr(\mathcal{P} = \mathcal{A}|S_0,S_0' = \mathrm{N})(1-\alpha)\ .\end{aligned}$$

Hence, as defined in equation (9),

$$\begin{cases}\Pr(\mathcal{P} = \mathcal{A}|S_0 = \mathrm{N}) = f_{\mathrm{N}} \\ \Pr(\mathcal{P} = \mathcal{A}|S_0 = \mathrm{F}) = \Pr(\mathcal{P} = \mathcal{A}|S_0 = \overline{F}) \ = f_F\end{cases}. \quad (11)$$

Then, in the same way as for equation (10),

$$\begin{aligned}\phi_{ij}^{\mathcal{P}} = \Pr(\mathcal{P}|S_0 = \mathrm{F})m_{ij}^{F}\alpha\lambda + \Pr(\mathcal{P}|S_0 = \overline{F})m_{ij}^{\overline{F}}\alpha(1-\lambda) \\ +\Pr(\mathcal{P}|S_0 = \mathrm{N})m_{ij}^{\mathrm{N}}(1-\alpha)\ ,\end{aligned}$$

where the appropriate $\Pr(\mathcal{P}|S_0)$ is obtained from equation (11). However, since the phenocopies may be spurious or will have descended from many different ancestral founding chromosomes, we assume that, in terms of the occurrence of marker alleles, they are indistinguishable from any chromosome not bearing the major mutation at the disease locus; in other words, $m_{ij}^{\overline{F}} = m_{ij}^{\mathrm{N}}$ as defined in equations (6) and (7).

Likelihood Calculations

Expected SNP allele frequencies to the left and right of the disease locus are determined by independent Markov processes. Thus, over the whole candidate region, the log-likelihood of a sample of data for a fixed location of the disease gene, x, and a given set of hidden Markov–model parameters is given by

$$\ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w)_{\mathrm{TOT}} = \ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w)_L + \ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w)_R\ ,$$

where $\boldsymbol{\Gamma}$ is a vector of model parameters, $\boldsymbol{\Gamma} = (\alpha,\beta_F,\beta_{\mathrm{N}},\gamma,\lambda)^T$, and $\mathbf{p}$ and w are vectors of allele frequencies and ancestral indicators, respectively:

$$\mathbf{p} = (p_{-L},p_{-(L-1)},\ldots,p_{-2},p_{-1},p_1,p_2,\ldots,p_{R-1},p_R)^T\ ,$$

and

$$\omega = (\omega_{-L},\omega_{-(L-1)},\ldots,\omega_{-2},\omega_{-1},\omega_1\omega_2,\ldots,\omega_{R-1},\omega_R)^T\ .$$

The log-likelihood to the right of the disease locus is given by

$$\ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w)_R = \sum_{i=1}^{R}\sum_{j=1}^{2}\sum_{\mathcal{P}=\mathcal{A}}^{U} n_{ij}^{\mathcal{P}}\ln[\phi_{ij}^{\mathcal{P}}] + C_R\ ,$$

where C_R is constant for a known population disease frequency Q: $C_R = R\kappa n_A - RT(n_A + n_U)$. The independent log-likelihood to the left of the disease locus is calculated similarly.

Parameter Estimation

The hidden Markov model described here is overparameterized. We reduce the number of free parameters by noticing the following relationships.

First, the population frequency of the disease is given by

$$Q = \alpha^2\beta_F^2 + 2\alpha(1-\alpha)\beta_F\beta_{\mathrm{N}} + (1-\alpha)^2\beta_{\mathrm{N}}^2\ .$$

The frequency of the disease is generally known so that we can eliminate α from the likelihood calculation:

$$\begin{aligned}\alpha &= \frac{-2\beta_{\mathrm{N}}(\beta_F-\beta_{\mathrm{N}}) \pm \sqrt{4\beta_{\mathrm{N}}^2(\beta_F-\beta_{\mathrm{N}})^2 - 4(\beta_F-\beta_{\mathrm{N}})^2(\beta_{\mathrm{N}}^2-Q)}}{2(\beta_F-\beta_{\mathrm{N}})^2} \\ &= \frac{\sqrt{Q}-\beta_{\mathrm{N}}}{\beta_F-\beta_{\mathrm{N}}}\ ,\end{aligned}$$

since $\alpha > 0$.

Second, the likelihood is constant for a fixed ratio of disease-model parameters $\beta = \beta_{\mathrm{N}}/\beta_F$. Thus, the two parameters can be eliminated from the likelihood and can be replaced by a single penetrance parameter for which $\beta \leqslant 1$, since it is assumed that, at the disease locus, the mutation has greater propensity for the development of the disease than does the normal allele. Overall, for a known disease frequency Q, the vector of model parameters to be estimated reduces to $\boldsymbol{\Gamma} = (\beta,\gamma,\lambda)^T$, together with the allele frequencies and ancestral indicators.

We use MCMC methods to obtain posterior distributions for the model parameter estimates. The advantage of this approach is that we can incorporate prior information for the model parameters—which may be useful if we have reliable values for the age of the mutation or disease model, for example. We employ a Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) to obtain realizations of each model parameter by sampling from the full conditional distribution, using a rejection-sampling scheme. Each iteration of the sampling scheme consists of a six-step procedure summarized in Appendix B. From initial parameter values, the algorithm is run for a substantial burn-in period, to allow convergence. During the subsequent sampling period, realizations of the parameter set are recorded every 100th iteration. Over many iterations, posterior distributions of parameter estimates are obtained from these realizations.

Allowing for Heterogeneity in Recombination Rates

In developing the hidden Markov model for fine-scale mapping, we have assumed a constant ratio of recombination fraction to physical distance across the whole of the candidate region. However, it is thought that recombination hot spots and cold spots occur along the genome. Lazzeroni (1998) accounts for heterogeneity in recombination rates in a generalized least-squares approach to fine-scale mapping, by allowing the ratio of physical distance to genetic distance to be different to the left and to the right of the disease locus. However, this may not be sufficient to allow for the variability in recombination rates, particularly in larger candidate regions.

As an alternative, we propose that the ratio of physical distance to genetic distance across the candidate region can be described by a first-order Gaussian autoregressive process. We divide the candidate region into K equal intervals so that the rate of recombination, y_r, in the rth interval is given by $y_1 = \mu + \epsilon_1^*$ and $y_r = \mu + \varrho(y_{r-1} - \mu) + \epsilon_r$, where μ is the mean recombination rate across the region and ϱ is the first-order correlation coefficient. The errors are assumed to be independently distributed, so that $\epsilon_1^* \sim \mathrm{N}(0,\sigma^2/(1-\varrho)^2)$ and $\epsilon_r \sim \mathrm{N}(0,\sigma^2)$ for $r = 2,3,\ldots,K$. The log-likelihood of a sample of K recombination rates from this process is then given by

$$\ell(\mathbf{y}|\mu,\varrho,\sigma^2)_{AR1} = -\frac{K}{2}\ln(2\pi\sigma^2) + \ln(1-\varrho^2) - \frac{(y_1-\mu)^2(1-\varrho^2)}{\sigma^2} - \frac{1}{\sigma^2}\sum_{r=2}^{K}[(y_r-\mu)-\varrho(y_{r-1}-\mu)]^2 \ .$$

We assume that μ is known from existing physical and genetic maps. Uncertainty for this parameter can be incorporated by assuming a tight prior distribution for μ, centered about the estimated ratio. For example, in the region of the CFTR gene for CF, a physical distance of 1.6 Mb corresponds to a genetic distance of ~0.8 cM (Collins et al. 1996)—in other words, $\mu = .5$.

The recombination rates across the region can then be used in calculating the probability of no recombination events in the interval between any pair of adjacent marker loci; for example, the probability of no recombination events in the interval between marker i and $i+1$ is given by $\exp[-\gamma d_{i+1}\theta_{i+1}]$, where $\theta_{i+1} = \Sigma_{r=1}^{K} y_r\pi_r/\Sigma_{r=1}^{K}\pi_r$ and π_r denotes the proportion of the rth recombination-rate interval contained in the interval between the two marker loci. The log-likelihood of the sample of data for a given set of recombination rates $\mathbf{y}$ and hidden Markov–model parameters is expressed by $\ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w,\mathbf{y})_{\mathrm{TOT}}$. The recombination rates are, in effect, nuisance parameters, so that

$$\ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w)_{\mathrm{TOT}} = \int_{\mathbf{y}} \ell(\mathrm{data}|x,\boldsymbol{\Gamma},\mathbf{p},w,\mathbf{y})_{\mathrm{TOT}}\,\ell(\mathbf{y}|\mu,\varrho,\sigma^2)_{AR1} \ .$$

In this way, we can then incorporate heterogeneous recombination rates into the Metropolis-Hastings rejection-sampling scheme for the model parameters as described in Appendix B.

Allowing for Nonindependent Recombinational Histories

In developing the hidden Markov model for disease-marker association in the region of a disease gene, we have assumed independent recombinational histories for each chromosome in the sample. However, the key to this approach to disease-gene mapping is that all—or at least a majority of—affected individuals share a recent single common ancestor bearing the disease-predisposing mutation. Treating the recombinational histories as independent is equivalent to assessing a star-shaped genealogy, which is not consistent with likely demographic scenarios for the development of a disease mutation in a finite population. Instead, we expect particular pairs of chromosomes to have a more recent common ancestor than do other pairs of chromosomes—and, consequently, to share a greater proportion of their recombinational history. The effect of this shared ancestry is to down-weight the contribution of each case chromosome to the total log-likelihood by a factor $[1 + (n_A - 1)c]^{-1}$, where c is given by

$$\frac{2(n_A-2)!(n_A+1)}{n_A-1}\sum_{k=1}^{n_A-2}\frac{\left[\sum_{i=1}^{\infty}(-1)^{i-1}\binom{n_A+i-1}{n_A-k}^{-1}\right]}{(n_A-k+1)(n_A-k+2)(n_A-k)(k-1)!(n_A-k-2)!} \ .$$

Since the correction factor is <1, we effectively down-weight the contribution of each case chromosome to the total log-likelihood, to account for the dependence between them. We emphasize here that a quasi-likelihood approach is not applicable in a Bayesian framework; but it does suggest the use of a likelihood approximation. We propose to multiply the log-likelihood calculated under a star-shaped genealogy by the same correction factor. This has the effect of increasing the variance of the posterior distribution, to account for the shared ancestry of the case chromosomes.

Examples

To illustrate our proposed method, we consider two diseases: CF and HD. Mutations responsible for the occurrence of these two diseases have been located on the

genome and are thus ideal for testing the accuracy and precision of the new method. In this section, we apply the proposed method to marker-haplotype data collected in candidate regions for the two disease genes (Kerem et al. 1989; MacDonald et al. 1991). In both samples, cases and controls have been typed by RFLPs. Since these markers have low rates of mutation, we have assumed that $m = 0$, corresponding to no marker mutation in the period since the founding disease mutation.

CF

CF is one of the most common autosomal recessive disorders affecting whites, occurring with an incidence of 1 case/2,000 births. Initial scans of the genome in the 1980s provided evidence of a single CF gene on chromosome 7q31 (Kerem et al. 1989). More recently, a 3-bp deletion (ΔF508) has been identified within this region in the CFTR gene. It is now known that ΔF508 accounts for ~68% of all chromosomes in affected individuals today, with the remainder consisting of several other, rarer mutations in the same gene. Kerem et al. (1989) collected marker data from affected cases and healthy controls, using 23 RFLPs in a 1.8-Mb candidate region of chromosome 7q31, from the MET locus to marker D7S426.

Figure 1 presents odds ratios for each of the RFLPs in the candidate region. There is strongest evidence of disease-marker association in a region of 0.6–0.9 Mb from the MET locus, with a peak observed at 0.869 Mb. Within this region, however, there is a single marker, 0.889 Mb from the MET locus, at which disease-marker association is much lower. This marker is, in fact, closest to the ΔF508 mutation in the CFTR gene, at ~0.880 Mb from the MET locus.

Previous analyses of these data by published methods have yielded a variety of results. Terwilliger (1995) places the mutation 0.77 Mb from the MET locus, with a 99.9% support interval of 0.69–0.87 Mb. Although this interval overlaps part of the CFTR gene, it does not include the ΔF508 mutation. Xiong and Guo (1997) obtained an improved estimate of the location of ΔF508, at 0.80 Mb, although this was derived from only a selected subset of the CF data, a subset for which any case chromosomes not bearing the ΔF508 mutation were excluded. With additional information for the region of the mutation (Morral et al. 1994), Collins and Morton (1998) analyzed the same subset and obtained an estimate of 0.83 Mb.

We applied the hidden Markov model–based mapping method proposed here to the complete CF data set of Kerem et al. (1989). We assumed a disease frequency of $Q = .0005$, on the basis of estimates for the population from which the sample was ascertained. We also assumed a mean recombination rate of .5, since, in the candidate region around the CFTR gene, the physical distance of 1.6 Mb corresponds to a genetic distance of 0.8 cM (Collins et al. 1996). A number of sets of initial values for the model parameters were considered, all resulting in similar posterior distributions and parameter estimates after an initial burn-in period of the Metropolis-Hastings rejection-sampling scheme followed by a sampling period of a further 1 million iterations for which every 100th iteration was recorded. Regardless of the starting values for the model parameters, there is rapid convergence to parameter estimates, which also appear to mix well (data not shown).

Figure 2 presents the posterior distributions of the location of the mutation, the hidden Markov–model parameters β, α, λ, and γ, and the first-order autoregressive parameters ϱ and σ^2 for recombination-rate heterogeneity across the candidate region, when independent recombinational histories for the case and control chromosomes are assumed. Also presented is the distribution of the hidden Markov–model log-likelihood obtained throughout the sampling period. Table 2 presents the initial parameter values for this run, together with the true parameter values (where known) and summary statistics from the posterior distributions.

The mean estimate of the location of the mutation is $\hat{x} = 0.784$Mb from the MET locus, with a 99% credibility interval of 0.731–0.838 Mb. Although there is substantial error in this estimate, the results are consistent with estimates obtained by other case-control–based mapping methods, which have been described above. The frequency of the mutation is estimated as $\hat{\alpha} = .223$. This is in agreement with a mutation-frequency estimate of .224 based on a fully penetrant recessive disease with frequency .0005 (Kerem et al. 1989). The estimate of the disease-model parameter $\hat{\beta}$ approaches 0, which is as would be expected for a fully penetrant recessive disease for which $\beta_F = 1$ and $\beta_N = 0$. The estimated major-mutation proportion is $\hat{\lambda} = .768$, which is close to the estimate that 70% of existing CF chromosomes bear the ΔF508 mutation. The estimated age of the mutation is $\hat{\gamma} = 2.05$, corresponding to 205 generations. Again, this is not inconsistent with other, independent estimates of the age of ΔF508, which suggest that it is ~200 generations old (Serre et al. 1990). Credibility intervals for the first-order autoregressive parameters do not include 0, suggesting that there is recombination-rate heterogeneity across the candidate region.

For comparison, we have also applied the hidden Markov model–based mapping method to the same set of data but have modeled dependence between case chromosomes by using the conditional coalescent as proposed by McPeek and Strahs (1999). Figure 3 presents the posterior distribution of the model parameters that is based on every 100th of 1 million iterations of the Metropolis-Hastings rejection-sampling scheme, with

the log-likelihood being corrected for between-chromosome correlations.

We obtain an improved estimate of the location of ΔF508: 0.798 Mb from the MET locus, with a 99% credibility interval of 0.610–1.069 Mb, this time with the true location of the mutation being included. The other model parameter estimates remain relatively unchanged, but with noticeably wider posterior credibility intervals (data not shown). The exception is in the first-order autoregressive-process–correlation parameter, the mean estimate of which is considerably closer to 0 under the coalescent model (.053) than under independence (.250). This would suggest that much of the correlation between marker loci is accounted for by the correlation between related chromosomes.

With the same correction, McPeek and Strahs (1999) estimate the location of the mutation to be 0.95 Mb from the MET locus, with a 99% confidence interval of 0.28–1.62 Mb (calculated on the basis of their presented 95% confidence interval). The difference, in estimated location, between the two methods is likely a result of McPeek and Strahs's (1999) assumption of a homogeneous recombination rate of 1 cM–1 Mb across the map of marker loci.

Table 3 presents the posterior ancestral-haplotypes probabilities realized over the 1 million iterations of the Metropolis-Hastings rejection-sampling scheme. There is complete agreement over all but the markers most distant from the ΔF508 mutation at which levels of disease-marker association are weakest. For this particular sample, maximizing the model likelihood over the ancestral haplotype, as in the method of McPeek and Strahs (1999), would be expected to yield results similar to those of our proposed method, since the maximum-likelihood estimate has such high posterior probability. With less certainty with regard to ancestral haplotypes, maximum-likelihood–based approaches may suffer bias and warrant further investigation.

HD

HD is a midlife-onset autosomal dominant neurodegenerative disorder occurring at an incidence of ~1 case/10,000. The HD gene was first mapped to chromosome 4p16, in the region of marker D4S10, by Gusella et al. (1983, 1984). More recently, the Huntington's Disease Collaborative Research Group (1993) has identified within this region a large gene (IT15) with an expandable unstable trinucleotide-repeat sequence. It is now known that IT15 genes with many repeats of the trinucleotide sequence are responsible for the development of the disease. MacDonald et al. (1991) collected marker data from HD and normal chromosomes in a 2.5-Mb region of chromosome 4p16, from marker D4S90 to D4S10, using 27 RFLPs.

Figure 4 presents odds ratios for each of the RFLPs in the candidate region. The strongest evidence of disease-marker association lies in the interval between markers D4S182 and D4S180, at 2.38 Mb and 2.85 Mb, respectively, from marker D4S90. This is in agreement with the location of IT15 at ~2.5–2.6 Mb from marker D4S90. As in the CF data of Kerem et al. (1989), there are RFLPs with low levels of disease-marker association within this interval. Despite this apparent inconsistency, Xiong and Guo (1997), using their case-control–based mapping method, obtained 2.62 Mb from marker D4S90 as the estimated location of the disease gene.

We also have applied the hidden Markov model–based mapping method to the HD data of MacDonald et al. (1991). We assumed independent recombinational histories for the case chromosomes and a disease frequency of $Q = 10^{-4}$, in line with published estimates for populations of European descent. We also assumed a mean recombination rate of 1 in the candidate region, so that the usual 1 Mb– to–1 cM correspondence holds. We considered various sets of initial values for the model parameters, all resulting in similar posterior distributions and parameter estimates after the same burn-in period and sampling period that were employed in the analysis of the CF data.

Figure 5 presents, for the HD data, the posterior distributions for the hidden Markov model and autoregressive parameters. Summary statistics from the posterior distributions of model parameters are presented in table 4, together with true values (where known). The mean estimate of the location of the mutation is 2.52 Mb from marker D4S90, with a 99% credibility interval of 2.20–2.75 Mb. The mean estimate is accurate, being contained within the IT15 gene for HD. The wide credibility interval reflects the considerable variation in the strength of disease-marker association in the IT15 gene (fig. 4).

The estimate of the disease model parameter $\hat{\beta} = 2.1 \times 10^{-3}$ is >0, which we would expect for a dominant disease. The estimated age of the mutation is $\hat{\gamma} = 1.37$, corresponding to 137 generations, and is not inconsistent with other estimates of the age of HD (Kaplan et al. 1995; Xiong and Guo 1997). Credibility intervals for the autoregressive parameters do not include 0, suggesting recombination-rate heterogeneity across the candidate region.

Discussion

We have presented a new multilocus method for the fine-scale mapping of disease genes. We model disease-marker association in the vicinity of a disease gene by means of a hidden Markov process used in a way similar to that employed by McPeek and Strahs (1999). In this way, both models account for correlation between the

markers, a clear advantage over many existing multilocus composite-likelihood methods that assume independence (Terwilliger 1995; Xiong and Guo 1997; Collins and Morton 1998). In addition, both models allow for mutation at marker loci.

We employ MCMC methods in a Bayesian framework, to obtain posterior distributions for model parameter estimates including those for disease-gene location and the age of the disease-predisposing mutation. A potential advantage of this approach, over both the maximum-likelihood estimation used by McPeek and Strahs (1999) and other existing multipoint methods, is that, where appropriate, we are able to incorporate prior information for model parameters. In addition, by integrating over the marker haplotype present on the founding chromosome, we allow for the uncertainty in its makeup, in contrast to McPeek and Strahs (1999), who consider only the maximum-likelihood estimate.

Our model is more sophisticated than previous models in that we allow for recombination-rate heterogeneity across the candidate region, using a first-order Gaussian autoregressive process. In this way, we can allow for recombination hot spots and cold spots that may lead to bias in existing models. However, it would be relatively straightforward to incorporate variable recombination rates in the model proposed by McPeek and Strahs (1999).

We have used our method to identify the location of two known mutations—one for CF and one for HD. For HD, we obtain an accurate estimate of the location of the mutation within the IT15 gene, which is known to be responsible for the development of the disorder. In deriving the hidden Markov model for disease-marker association, we have assumed a multiplicative model for the disease. HD is a dominant (i.e., non-multiplicative) disorder, suggesting that our method is robust to deviations from a multiplicative-disease model.

For CF, we have presented two sets of simulation results, corresponding to two possible models of dependence in the recombinational histories of chromosomes in affected individuals. First, we have assumed independence, implying a star-shaped genealogy, which yields, for the location of the mutation, a 99% credibility interval that does not contain the true location of ΔF508. This result is consistent with analyses of the same data set by other multilocus models that assume independence between case chromosomes (Terwilliger 1995; Xiong and Guo 1997; Collins and Morton 1998). This clearly suggests deficiency in the star-shaped genealogical model of case-chromosome ancestry. For the second set of simulations, we correct for correlation between case chromosomes by means of a conditional coalescent model of dependence, proposed by McPeek and Strahs (1999). They justify the correction by means of quasi-likelihood arguments (Wedderburn 1974) that do not hold in a Bayesian framework. However, the same arguments suggest the use of an approximate log-likelihood, calculated by multiplication, by a correction factor, of the log-likelihood under independent recombinational histories (McPeek and Strahs 1999). This has the effect of increasing the variance of the posterior distribution, to account for the shared ancestry of case chromosomes.

An alternative approach to take account of the dependence between chromosomes is to model their ancestry directly, by means of a genealogical tree. In such a model, we can explicitly allow for multiple disease mutations, mutations at marker loci within the candidate region, and recombination events in the ancestry of the case sample. Lam et al. (2000) have constructed a genealogical tree for case chromosomes by using a combination of parsimony and likelihood methods, in which each chromosome in the tree is separated from its parent by a single marker mutation or recombination event. They then proceed to map the disease mutation as if the tree were known with certainty. A more appropriate approach would be to integrate over all possible genealogies, an approach that can be approximated by simulation. Graham and Thomson (1998) used such an approach to generate genealogical trees that are consistent with an observed sample of chromosomes, using a Moran (1962) model with known demographic parameters. However, their model assumes knowledge of the ancestral marker haplotype, the number of generations since the common ancestor, and the development of the population during this period. It is currently restricted to interval mapping using pairs of marker loci. Generalization of this approach to a full multilocus analysis with less stringent assumptions remains a challenge that will require considerable work in the future.

Acknowledgments

A.P.M. acknowledges financial support from Pfizer Limited. We thank the referees for their helpful comments on the submitted version of this article.

Appendix A

Summary of Model Parameters

Parameter	Range	Definition
α	[0,1]	Probability that chromosome is IBD with founder, at any given locus
β_F	[0,1]	Disease model parameter associated with mutation at disease locus
β_N	[0,1]	Disease model parameter associated with normal allele at disease locus
γ	[0,∞]	Recombination-rate parameter per 1 Mb of DNA in candidate region
p_i	[0,1]	Frequency of allele M_{i1} on chromosomes not IBD with founder, at locus i
ω_i	0 or 1	Indicator variable denoting presence/absence of allele M_{i1}
Q	[0,1]	Population frequency of disease
κ	[0,∞]	Sample-enrichment factor

Appendix B

Metropolis-Hastings Rejection-Sampling Scheme

Each iteration of the Metropolis-Hastings sampling scheme consists of a seven-step procedure. We denote the current parameter set by x, $\mathbf{\Gamma}$, $\mathbf{p}$, and w. In addition, the current set of recombination rates is denoted $\mathbf{y}$, modeled as a first-order autoregressive process with known mean recombination rate μ and current parameters ϱ and σ^2. If we do not wish to allow for heterogeneity in recombination rates across the candidate region, $\mathbf{y} = 1$ and we ignore step 7 of the sampling scheme. The likelihood of a sample of cases and controls for the current parameter set is denoted $L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}$. The likelihood of the set of recombination rates for the current parameter set is denoted $L(\mathbf{y}|\mu,\varrho,\sigma^2)_{AR1}$. Throughout, we assume each ε to be drawn at random from the proposal distribution U(−.5,.5) and each υ to be drawn from U(0,1).

1. For each marker j in turn, propose a new allele frequency, $p_j^* = p_j + \nu_p\varepsilon$, where ν_p determines the maximum possible change from the current allele frequency. Since $p_j \in [0,1]$, proposed allele frequencies outside this range are reflected back into the parameter space. The likelihood for the proposed parameter set is denoted $L(\text{data}|x,\mathbf{\Gamma},\mathbf{p}^{j*},w,\mathbf{y})_{\text{TOT}}$ where $\mathbf{p}^{j*}$ is the vector of current allele frequencies with p_j replaced by the proposed p_j^*. The proposed allele frequency is accepted to the current parameter set if the acceptance probability is

$$\alpha = \min\left[\frac{L(\text{data} \mid x,\mathbf{\Gamma},\mathbf{p}^{j*},w,\mathbf{y})_{\text{TOT}}}{L(\text{data} \mid x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}},1\right] > \upsilon \ .$$

2. For each marker j, in turn, propose a new ancestral indicator:

$$\omega_j^* = \begin{cases} 0 & \text{if } \varepsilon \leqslant 0 \\ 1 & \text{if } \varepsilon > 0 \end{cases} .$$

The likelihood for the proposed parameter set is denoted $L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w^{j*},\mathbf{y})_{\text{TOT}}$ where w^{j*} is the vector of current ancestral indicators with ω_j replaced by the proposed ω_j^*. We then accept the proposed ancestral indicator to the current parameter set if the acceptance probability is

$$\alpha = \min\left[\frac{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w^{j*},\mathbf{y})_{\text{TOT}}}{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}},1\right] > \upsilon \ .$$

3. Propose a new location for the disease gene, $x^* = x + \nu_x\varepsilon$, where the parameter ν_x determines the maximum change from the current disease gene location. We restrict the location of the disease gene to the candidate region, so that proposed locations distal to the first and last markers on the map are reflected back into the candidate region. The likelihood for the proposed parameter set is denoted $L(\text{data}|x^*,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}$ and the proposed location is accepted to the current parameter set if

$$\alpha = \min\left[\frac{L(\text{data}|x^*,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}}{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}},1\right] > \upsilon \ .$$

4. Propose a new penetrance parameter, $\beta^* = \beta + \nu_\beta\varepsilon$, where ν_β determines the maximum change from the current penetrance parameter. The penetrance parameter is restricted to $\beta \in [0,1]$ so that proposed penetrances outside this range are reflected back into the parameter space. The likelihood for the proposed parameter set is denoted $L(\text{data}|x,\mathbf{\Gamma}^{\beta*},\mathbf{p},w,\mathbf{y})_{\text{TOT}}$ where $\mathbf{\Gamma}^{\beta*}$ is the vector of current hidden Markov–model parameters with β re-

placed by the proposed β^*. The proposed penetrance parameter is then accepted to the current parameter set if the acceptance probability is

$$\alpha = \min\left[\frac{L(\text{data}|x,\mathbf{\Gamma}^{\beta*},\mathbf{p},w,\mathbf{y})_{TOT}}{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}},1\right] > v \ .$$

5. Propose a new age of the mutation, $\gamma^* = \gamma + \nu_\gamma\varepsilon$, where ν_γ determines the maximum change from the current age of the mutation. The age of the mutation is restricted to be positive so that a negative proposed age is reflected back into the valid parameter space. The likelihood for the proposed parameter set is denoted $L(\text{data}|x,\mathbf{\Gamma}^{\gamma*},\mathbf{p},w,\mathbf{y})_{\text{TOT}}$, where $\mathbf{\Gamma}^{\gamma*}$ is the vector of current hidden Markov–model parameters with γ replaced by the proposed γ^*. The proposed age of the mutation is then accepted to the current parameter set if the acceptance probability is

$$\alpha = \min\left[\frac{L(\text{data}|x,\mathbf{\Gamma}^{\gamma*},\mathbf{p},w,\mathbf{y})_{\text{TOT}}}{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}},1\right] > v \ .$$

6. Propose a new major mutation proportion, $\lambda^* = \lambda + \nu_\lambda\varepsilon$, where ν_λ determines the maximum change from the current proportion. Since λ is a proportion, it is restricted to [0,1]. Proposed proportions outside this range are reflected back into the valid parameter space. The likelihood for the proposed parameter set is denoted $L(\text{data}|x,\mathbf{\Gamma}^{\lambda*},\mathbf{p},w,\mathbf{y})_{\text{TOT}}$, where $\mathbf{\Gamma}^{\lambda*}$ is the vector of current hidden Markov–model parameters with λ replaced by the proposed λ^*. The proposed major mutation proportion is then accepted to the current parameter set if the acceptance probability is

$$\alpha = \min\left[\frac{L(\text{data}|x,\mathbf{\Gamma}^{\lambda*},\mathbf{p},w,\mathbf{y})_{\text{TOT}}}{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}},1\right] > v \ .$$

7. Propose a new set of K recombination rates so that, for each $i = 1,2,...,K$, $y_i^* = y_i + \nu_y\varepsilon$, where ν_y determines the maximum change from the current recombination rates. Each recombination rate is restricted to be nonnegative so that negative proposals are reflected back into the valid parameter space. In the same step, we also propose new autoregressive parameter values, $\varrho^* = \varrho + \nu_\varrho\varepsilon$ and $\sigma^* = \sigma^* + \nu_\sigma\varepsilon$, where ν_ϱ and ν_σ determine the maximum change in parameter values for ϱ and σ, respectively. The correlation parameter $\varrho \in [0,1]$ and the standard deviation σ is restricted to be positive. Proposals outside the permitted space are reflected back to valid parameter values. The likelihoods for the proposed parameters and recombination rates are denoted $L(\mathbf{y}^*|\mu,\varrho^*,\sigma^{*2})_{AR1}$ and $L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y}^*)_{\text{TOT}}$. The complete set of proposed recombination rates and autoregressive parameters are accepted to the current set if

$$\alpha = \min\left[\frac{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y}^*)_{\text{TOT}}L(\mathbf{y}^*|\mu,\varrho^*,\sigma^{*2})_{AR1}}{L(\text{data}|x,\mathbf{\Gamma},\mathbf{p},w,\mathbf{y})_{\text{TOT}}L(\mathbf{y}|\mu,\varrho,\sigma^2)_{AR1}},1\right] > v \ .$$

At any stage we can incorporate prior distributions for model parameters by multiplying the appropriate acceptance probability by the ratio of prior probabilities for the proposed and current parameter values.

References

Collins A, Frezal J, Teague J, Morton NE (1996) A metric map of humans: 23,500 loci in 850 bands. Proc Natl Acad Sci USA 93:14771–14775

Collins A, Morton NE (1998) Mapping a disease locus by allelic association. Proc Natl Acad Sci USA 95:1741–1745

Graham J, Thompson EA (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am J Hum Genet 63: 1517–1530

Gusella JF, Tanzi RE, Anderson MA, Hobbs W, Gibbons K, Raschtchian R, Gilham TC, et al (1984) DNA markers for nervous system disorders. Science 225:1320–1326

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–208

Hastings WK (1970) Monte-Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109

Huntington's Disease Collaborative Research Group (1993) A novel gene containing a tri-nucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72:971–983

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of complex disease genes. Nat Genet 22:139–145

Lam JC, Roeder K, Devlin B (2000) Haplotype fine mapping by evolutionary trees. Am J Hum Genet 66:659–673

Lander ES, Green P (1987) Construction of multi-locus genetic maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Lazzeroni LC (1998) Linkage disequilibrium and gene mapping: an empirical least squares approach. Am J Hum Genet 62:159–170

MacDonald ME, Lin C, Srinidhi L, Bates G, Altherr M, Whaley WL, Lehrach H, et al (1991) Complex patterns of linkage disequilibrium in the Huntington's disease region. Am J Hum Genet 49:723–734

McPeek MS, Strahs A (1999) Assessment of linkage disequi-

librium by the decay of haplotype sharing, with application to fine scale genetic mapping. Am J Hum Genet 65:858–875

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Moran P (1962) The statistical processes of evolutionary theory. Clarendon Press, Oxford

Morral N, Bertranpetit J, Estevill X, Nunes V, Casals T, Gimenez J, Reis A, et al (1994) The origin of the major cystic fibrosis mutation (delta-F508) in European populations. Nat Genet 7:169–175

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931–942

Ott J (1991) Analysis of human genetic linkage. John Hopkins University Press, Baltimore

Penisi E (1998) A closer look at SNPs suggests difficulties. Science 281:1787–1789

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77: 257–286

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Serre JL, Simon-Bouy B, Morret E, Jaume-Roig B, Balassopoulou A, Schwartz M, Taillander A (1990) Studies of RFLPs closely linked to the cystic-fibrosis locus throughout Europe lead to new considerations in population genetics. Hum Genet 84:449–454

Terwilliger J (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777–787

Wedderburn RWM (1974) Quasi-likelihood functions, generalized models, and the Gauss-Newton method. Biometrika 61:439–447

Xiong M, Guo S-W (1997) Fine scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531