Multipoint Oligogenic Analysis of Age-at-Onset Data with Applications to Alzheimer Disease Pedigrees

E. Warwick Daw,[1,2] Simon C. Heath,[2*] Ellen M. Wijsman[1,3]

[1]Department of Medicine, Division of Medical Genetics; [2]Department of Statistics; and [3]Department of Biostatistics, University of Washington, Seattle

Summary

It is usually difficult to localize genes that cause diseases with late ages at onset. These diseases frequently exhibit complex modes of inheritance, and only recent generations are available to be genotyped and phenotyped. In this situation, multipoint analysis using traditional exact linkage analysis methods, with many markers and full pedigree information, is a computationally intractable problem. Fortunately, Monte Carlo Markov chain sampling provides a tool to address this issue. By treating age at onset as a right-censored quantitative trait, we expand the methods used by Heath (1997) and illustrate them using an Alzheimer disease (AD) data set. This approach estimates the number, sizes, allele frequencies, and positions of quantitative trait loci (QTLs). In this simultaneous multipoint linkage and segregation analysis method, the QTLs are assumed to be diallelic and to interact additively. In the AD data set, we were able to localize correctly, quickly, and accurately two known genes, despite the existence of substantial genetic heterogeneity, thus demonstrating the great promise of these methods for the dissection of late-onset oligogenic diseases.

Introduction

Many diseases with a genetic component are not apparent at birth. For example, schizophrenia, glaucoma, Alzheimer disease (AD), and various cancers all typically occur later in life. Since their presence is not immediately apparent, there is hope that, after the genetic mechanisms that lead to these diseases are determined, treatments may be developed that prevent or delay their onset. Unfortunately, late-onset diseases tend to exhibit complex modes of inheritance, and pedigrees segregating for such diseases tend to contain many members who are deceased, thereby reducing the number of individuals for whom genetic material is available. These factors contribute to the extreme difficulty associated with the genetic dissection of late-onset traits.

The question of which analysis methods are best suited to localizing genes for complex traits has been the subject of some discussion. Proposals have ranged from association-based methods (Risch and Merikangas 1996) and other small-pedigree methods (e.g., Haseman and Elston 1972; Olson and Wijsman 1993; Kruglyak et al. 1995) to methods applied to large, extended pedigrees, (e.g., Ott 1979; Amos 1994; Greenberg et al. 1996). A recent workshop confirmed the potential value of multipoint analysis and also demonstrated that the use of extended, rather than nuclear, pedigrees could provide additional information (Wijsman and Amos 1997). Unfortunately, a multipoint linkage analysis with many markers in extended pedigrees, using traditional exact methods, can be computationally intractable and will likely remain so for the foreseeable future, even with steady exponential increases in computer speeds and the algorithmic improvements found in programs such as VITESSE (O'Connell and Weeks 1995) and FASTLINK (Cottingham et al. 1993).

Another difficulty associated with dissecting a complex trait is that multiple trait loci may contribute to the phenotype. Most analysis methods do not explicitly allow for multiple trait loci in the inheritance model, primarily because of computational limitations. Analysis of a complex trait under two-locus models has been shown to provide substantial gains in power to detect linkage, as compared with single-locus models (Schork et al. 1993; Knapp et al. 1994). Although single-locus models can be used to find traits influenced by two genes (Vieland et al. 1992*a*, 1992*b*), this difference in power to map trait loci between single-locus and oligogenic models may become more pronounced as the number of

Received April 10, 1998; accepted for publication January 6, 1999; electronically published February 22, 1999.

Address for correspondence and reprints: Dr. E. Warwick Daw, Department of Medicine, Division of Medical Genetics, Box 357720, University of Washington, Seattle, WA 98195-7720. E-mail: warwickd@u.washington.edu

[*] Present affiliation: Laboratory of Statistical Genetics, The Rockefeller University, New York.

0002-9297/99/6403-0022$02.00

genes influencing the trait increases (Rice et al. 1993). Furthermore, several results suggest that specifying the wrong model for the trait locus linked to the marker may cause biased location estimates or outright failure to detect linkage (Clerget-Darpoux et al. 1986; Greenberg and Hodge 1989). These issues suggest that methods not computationally limited to single-locus models may be necessary for efficient genomic localization of trait loci for complex diseases. Unfortunately, it appears that carrying out such an analysis, using exact computation and accounting for all marker and family data, is a computationally intractable problem. However, a compromise that does not require an exact solution can result in a feasible analysis that considers a large parameter space.

One way to implement this compromise is to use Monte Carlo Markov chain (MCMC) methods (Metropolis et al. 1953; Hastings 1970). These methods use statistical sampling of the parameter space to estimate a result that is difficult or impossible to obtain from exhaustive enumeration of all genotype probabilities. MCMC methods have been used to estimate linkage likelihoods in problems in which compromise in pedigree size or model complexity was not desired (Guo and Thompson 1992; Thompson 1994*a,* 1994*b*). These methods also provide a way to implement Bayesian genetic analysis, in which computationally tractable evaluation of high-dimensional integrals is required (Stephens and Smith 1993; Hoeschle 1994; Heath 1995; Satagopan et al. 1996). Additionally, MCMC samplers have been developed for sampling over a space of different models, providing an estimate of which models best fit the data (Carlin and Chib 1995; Green 1995; Phillips and Smith 1996).

Recently, an MCMC multiple-trait-locus joint linkage and segregation analysis method for quantitative trait loci (QTLs) has been developed (Heath 1997). This method, although extremely promising, is limited to the analysis of continuous trait data. There are a number of types of real data that do not fit the QTL model. In particular, age-at-onset data for a trait that is not fully penetrant or that typically has a late age at onset fits this model poorly. Although one could do an analysis, using only the ages at onset of affected individuals, with the original model, this would result in the loss of a large amount of information that would have come from the censored individuals.

Our primary goal was to expand the methods of Heath (1997) to age-at-onset data. We also demonstrate this expanded method, in an application to real age-at-onset data for a complex trait, and we provide some practical insight gained from this application. The real data set available to us was an AD data set, which had been used previously to localize two AD genes, PS1 and PS2 (Schellenberg et al. 1992; Levy-Lahad et al. 1995*a,* 1995*b*). This sample is heterogeneous, containing families in which either PS1 or PS2 mutations have been found, as well as families in which the cause of AD is as yet unknown. Our methods quickly and correctly localized the known AD genes, demonstrating the utility of this approach.

Analysis Methods

QTL Model Summary

The MCMC algorithm described by Heath (1997) and implemented in the program Loki (PANGAEA) is a combined segregation and linkage analysis for oligogenic quantitative traits, using full-chromosome multipoint data. We modified the algorithm to deal with censored data, so that age at onset could be analyzed as a right-censored quantitative trait.

In brief, MCMC methods are used to sample possible sets of all model values and full-genotype information consistent with the observed data (observed trait values, observed genotypes, and pedigree information). Such a set will be referred to as a “state.” The model allows for multiple QTLs such that the number of trait loci is one of the sampled values, and interactions between loci are additive. In the analysis, each state, S, was specified by

$$S = (k,G,M,\lambda,\delta,\eta,\alpha,\sigma_e^2,\mu,Y),$$

in which k is the number of QTLs and had a Poisson prior distribution; G is the matrix of complete genotypes (including phase) for all the QTLs; M is the matrix of complete genotypes (including phase) for all the markers; λ is the vector of linked QTL map positions (including the chromosome on which each QTL is located); δ is a vector indicating which QTLs are linked to chromosomes present in the analysis; η is the matrix of allele frequencies for the QTLs and markers; α is the matrix of additive and dominance effects for each QTL, that is,

$$\alpha = \begin{pmatrix} a_1 & d_1 \\ \vdots & \vdots \\ a_k & d_k \end{pmatrix},$$

in which a_i is the additive and d_i is the dominance effect for the i^{th} QTL; σ_e^2 is the variance of e, the residual environmental effect; μ is the overall mean; and Y is the observed matrix of genotype data and QTL values, on the available pedigrees. Note that the QTLs are assumed to be diallelic, and the values for the quantitative trait, y, are:

$$y = \mu + \sum_{i=1}^{k} Q_i\alpha_i + e,$$

in which Q_i is the incidence matrix for each QTL effect and can be derived from G, α_i is the i^{th} row of α, and e is the normally distributed residual environmental effect.

The sampling is done by taking the initial state S, making a proposal for a new state S' (a different set of model values), and computing a Metropolis-Hastings acceptance ratio (Metropolis et al. 1953; Hastings 1970), A, in which:

$$A = \frac{p(S')q(S;S')}{p(S)q(S';S)}.$$

In this formula, p(S′)/p(S) is the probability ratio of the two states, q(S;S′) is the probability of proposing the reverse move (from S' to S), and q(S′;S) is the probability of proposing the forward move. The new state is then randomly accepted, with probability $min(1,A)$, or it is rejected and the old state is kept. For details about the computations entering into the acceptance ratio, see Heath (1997). The complete updating of the state is, in fact, broken into a number of smaller steps, each updating a subset of the state. The steps in each sampling iteration were:

(1) Update ages at onset for individuals with censored data.
(2) Update the residual variance, σ_e^2.
(3) Propose either:
   (a) the birth or death of a QTL, or
   (b) splitting one QTL into two or combining two QTLs into one.
(4) Update QTLs and markers:
   (a) QTL effects,
   (b) QTL positions, and
   (c) QTL and marker genotypes locus by locus.
(5) Update QTL and marker allele frequency.
(6) Update the "overall mean," μ.
(7) Output state and repeat.

Age-at-Onset Model

The model used in the current study differs from that described by Heath (1997), in step 1 (above). All other steps remain the same. In the original model, individuals had either observed trait values or missing trait values. For individuals with missing trait values, no weighting for trait values was done in other steps. Note that, under the original model, no updating of trait values was done. The current model deals with right-censored quantitative traits: for each individual, we have either an observed trait value for affected individuals (uncensored data); an observed censored value, less than the true unobserved trait value for unaffected individuals (censored data); or a missing trait value. Observed or missing trait values are treated in the same manner as before, but a new treatment is required for censored values. In this case, trait values were sampled, conditional on the current QTL genotypes and censored value: a value was sampled from the truncated normal distribution for each individual's genotype, with the truncation point at the censored value (current age or age at death; see fig. 1). The actual trait values for individuals with observed values, or the sampled trait values for individuals with censored values, were then used in other updating steps, such as QTL genotype updating, that were conditional on trait values (see fig. 1), as described elsewhere (Heath 1997).

The censoring model also contains two independence assumptions, which should be noted: (1) censoring age and age at onset are independent, and (2) the observation of disease status is independent among individuals within each family. For unaffected living individuals, there is no reason to suspect that censoring age and the putative age at which they will get the disease have any dependence, other than that induced by survival to a particular age. However, this may not be true for deceased individuals. For example, for AD, there is some suggestion of correlation with risk factors also associated with coronary artery disease, which might influence censoring age (Jarvik et al. 1994). We believe that mild violation of this assumption will have only minor effects on the linkage analysis, because censoring age in a family-based linkage analysis is only one constraint on the latent age at onset. Latent onset age is also constrained by the distribution of observed onset ages, both in the

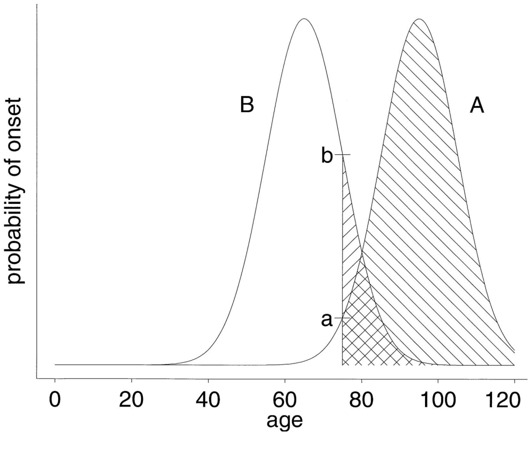

Figure 1 Hypothetical example with two genotype onset distributions, A and B. An unaffected 75-year-old will be assigned an age at onset from the hatched area under the B curve, if that person is currently assigned genotype B, and from the hatched area under the A curve, if currently assigned genotype A. Heights a and b indicate the relative likelihoods of genotypes A and B, respectively, for an individual with onset at age 75 years.

sample and in the family; the segregation patterns of the disease within the family; and if linked, the genetic marker data. The second assumption, that observations are made independently within each family, is probably also violated for many diseases. When a person is identified with a particular disease, family members tend to be more closely watched. This has the effect of reducing the within-family variance for affected individuals, whereas other sources of variance remain the same, possibly causing a false boost in a genetic signal. This may result in an upward bias in the estimate of the effect of a QTL, but the location estimate should remain unbiased.

An additional assumption is that age at onset is normally distributed for each genetic group (those sharing the same genotype at all QTLs), with all variances equal, so that the normal distributions differ only in their means. That the distributions themselves are normally distributed is not an unreasonable assumption, since many processes do result in distributions that are well approximated by the normal distribution. In some cases one may wish to consider a transformation such as a Box-Cox power transformation, but we did not do so here. The assumption that each group has the same variance may be more questionable, but we found it necessary for computational reasons. Note that these assumptions are essentially a genetic survival analysis in which the survival curve is restricted to the same cumulative normal curve, shifted for each group. Also, in using a right-censored quantitative trait model for age at onset, there is an implicit assumption that all individuals will eventually get the disease, if they live long enough.

Practical Considerations

It should be emphasized that, in addition to placing QTLs at a particular location on a chromosome, this procedure can place a QTL in a state that is not on any of the chromosomes included in the analysis. This is useful, because it allows modeling of QTLs that are linked to chromosomes about which marker data are not available. Interpretation of the results therefore requires examination of all QTLs, both those placed on a specific chromosome and those placed "elsewhere."

Since MCMC methods are stochastic, it is important not to base conclusions on a single analysis run. It is possible for the sampling procedure to become "stuck" in a particular region of the state space (a local maximum). Although the MCMC chain used here is theoretically irreducible (i.e., any point of the state space can be reached from any other), there are some very pathological cases in which the transition probabilities are so low that the MCMC chain is effectively reducible. Thus, it is important to run several analyses with different random number seeds and parameter values and then to examine these analyses for consistency. For a true strong signal, we expect all analysis runs to show QTL placements near that location in the majority of their iterations. For a weaker signal, we expect QTL placements in the signal region at a more frequent rate than expected from the prior distribution. Note that, in the absence of any linkage markers, the number of QTL placements in the region is proportional to the prior distribution. Thus, fewer QTL placements than expected under the prior distribution can be taken as evidence for exclusion. Another feature we look for, in both strong and weak signals, is a series of iterations in which the QTL is removed from the signal region, for a number of iterations, and then is placed back in the same region. This is indicative of a true signal and of good mixing of the MCMC sampler.

Sample

The AD data, used to determine the ability of these methods to localize disease genes, were obtained through a large study of the genetic basis of Alzheimer disease, which included obtaining informed consent for all subjects. We elected to use a real rather than a simulated data set because we feel that simulated data almost invariably fall short of the complexity found in real data. Furthermore, the basic method, without age-at-onset censoring, has already been tested on several simulated data sets (Heath 1997; Heath et al. 1997). There was a certain amount of missing marker data in this real data set, stemming from the history behind its collection, but additional marker typing for the purpose of testing the new methodology presented here could not be justified. Even in the absence of additional data collection, the data available to us were sufficient to test the method adequately.

The data set consisted of 1,150 individuals in 84 families, which fall into three subgroups: Volga German early-onset families (VG), non–Volga German early-onset families (EO), and a group of other families with generally later onset of AD (LO) (see table 1). Observed ages at onset were used as trait values for affected individuals, whereas age at last examination or age at death was used as the censored value for unaffected individuals. The VG families share a common ethnic heritage and were used to localize and clone the PS2 gene on chromosome 1 (Levy-Lahad et al. 1995*a*, 1995*b*). A single mutation in this gene accounts for AD in five of the seven VG families (Levy-Lahad et al. 1995*a*), representing 167 of 241 VG individuals. The EO families were used to localize the PS1 gene on chromosome 14 (Schellenberg et al. 1992), and the PS1 gene was subsequently cloned (Sherrington et al. 1995). Analysis of the EO families has identified several mutations in the

Table 1

AD Subgroups

	Stratification Group			
	VG	EO	LO	Total
Families	7 (5)	9 (8)	68	84
Individuals	241 (167)	295 (261)	624	1,150
Family sizes	6–58	12–53	4–26	4–58
Marker data:				
Chromosome 1	81–119[a]	...	...	...
Chromosome 5	99–114[a]	...	...	...
Chromosome 14	56–113/0–113[a,b]	46–159[a]	0–243[a]	61–489[a]

Note.—Numbers in parentheses for VG families indicate numbers for families found to have a PS2 mutation; for EO families, families found to have PS1 mutations.

[a] Minimum and maximum number of individuals typed for each marker used.

[b] (VG-only run)/(chromosome 14–only run).

PS1 gene accounting for AD in eight of the nine families, representing 261 of the 295 EO individuals (Ikeda et al. 1996; Poorkaj et al. 1998). The cause of AD in the LO families has not been conclusively determined. Those screened have not been found to have either PS1 (Schellenberg et al. 1993) or PS2 mutations (E. Wijsman and G. Schellenberg, unpublished data). Other candidate genes, such as apoE (Corder et al. 1993), as well as genes as yet unidentified, remain possibilities.

For historical reasons, the most extensive genetic marker data were available in the VG families, and the least extensive in the LO families. The emphasis in the early years of the genome screen was on finding genes for early-onset AD, and once the PS1 gene had been mapped, the EO families were excluded from further genome screening. Note, however, that some of the early screening markers were available for some of the LO families. LO families were also typed for markers in regions identified in the EO and VG families. On chromosome 14, the number of markers available for LO families ranged from 1 family with 6 markers to 17 families with no markers available. In the 51 LO families with marker data, the mean number of markers was three. The coverage in the LO families was somewhat better on the telomeric side of PS1 than on the centromeric side. Since the PS1 gene in the EO families on chromosome 14 was localized earliest, later efforts at localization on other chromosomes focused only on the VG families. As a result, ample marker data were available for VG and EO families on chromosome 14, with 11–16 markers available in each EO family (12 markers were used in this analysis) and 3–10 for each VG family. Only in the VG families were sufficient markers available on chromosome 1 (⩽19 markers) and chromosome 5 (⩽11 markers). Chromosome 5 was included as a negative control in the current analysis. Also, as is to be expected with a late-onset disease, genetic material was unavailable for approximately one-half of the individuals within each family because of the large number of deceased individuals.

Note that, although there are differences between groups in marker availability, there was no within-group bias among markers typed on the basis of PS1 or PS2 mutation status. Furthermore, since the group stratification was performed before the linkage analysis, each group is, in fact, heterogeneous. In the case of the VG sample, heterogeneity has been documented, both within and between families. Most individuals with genetic material available were screened for the PS2 mutation. No PS2 mutation was found in two of the seven VG families, and several individuals in the remaining five families showed AD but no PS2 mutation.

The genetic maps used in the current study were based on information obtained from web sites (Genome Database; Rockefeller database). Since none of these linkage maps contained all the markers used, it was necessary to combine information from several, to form a "consensus" map. Roughly linear interpolation was performed with the markers contained on multiple maps, to combine the results into one map. If there was a disparity between distances in different maps, the larger distance generally was used. If a marker could not be convincingly ordered, it was not used. This resulted in sex-averaged map lengths of ~310 cM for chromosome 1, ~220 cM for chromosome 5, and ~171 cM for chromosome 14. Because of a lack of marker data near the q-arm telomers of chromosomes 5 and 14, these maps were truncated in some analysis runs to 190 cM and 161 cM, respectively. The markers used were RH, PGM, AMY2B, D1S238, D1S422, D1S412, D1S306, D1S249, D1S245, D1S205, D1S425, D1S237, D1S229, D1S227, D1S479, D1S459, D1S446, D1S235, D1S180, and D1S102, on chromosome 1; D5S392, D5S395, D5S424, D5S428, D5S421, D5S414, D5S436, D5S434, D5S410, D5S412, and D5S422, on chromosome 5; and TCRD, D14S47, D14S52, D14S66, D14S77, D14S43, D14S53, D14S55, D14S48, PI, AACT, and D14S1, on chromosome 14. Note that D14S66 and D14S48 were not available for VG families and thus were not used in analyses that included only those families.

Data Analysis

Two analyses of the AD data are presented here. We performed one analysis of all available families, using chromosome 14 data, and the other of the VG families alone, using marker data from chromosomes 1, 5, and 14. In both analyses, the mean of the Poisson prior distribution on the number of QTLs was 1. Limited experimentation has indicated that the identification of gene locations is not very sensitive to the value of this mean. The first analysis, of chromosome 14 with all available families, used 12 markers with heterozygosities of 0.32–0.93. This analysis explored the ability of the proposed methods to describe and localize disease genes in the presence of significant genetic heterogeneity, with only 23% of the individuals known to be in families segregating for chromosome 14 PS1 mutations. The analysis of the VG families on 3 chromosomes used 40 markers (19, 11, and 10 markers on chromosomes 1, 5, and 14, respectively). This analysis examined the ability of these methods to locate a disease gene, in the presence of modest genetic heterogeneity, when given a large section of the genome to explore. Although these analyses are sufficient to demonstrate the methods, ideally we also would like to have performed an all-family analysis of multiple chromosomes. Unfortunately, full screens for chromosomes 1 and 5 were not available for the LO and EO families, and thus a joint analysis, using all families and all three chromosomes, was not possible. The lack of these additional data is more than compensated by the use of real data.

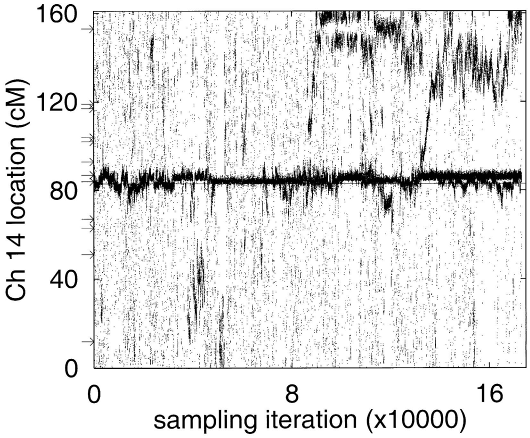

Figure 3 Location of each QTL placed on chromosome 14, in each iteration of the all-families analysis. Arrows indicate marker location; line indicates PS1 location.

Results

The age-at-onset model applied to analyses of the AD data produced correct localizations of both known genes

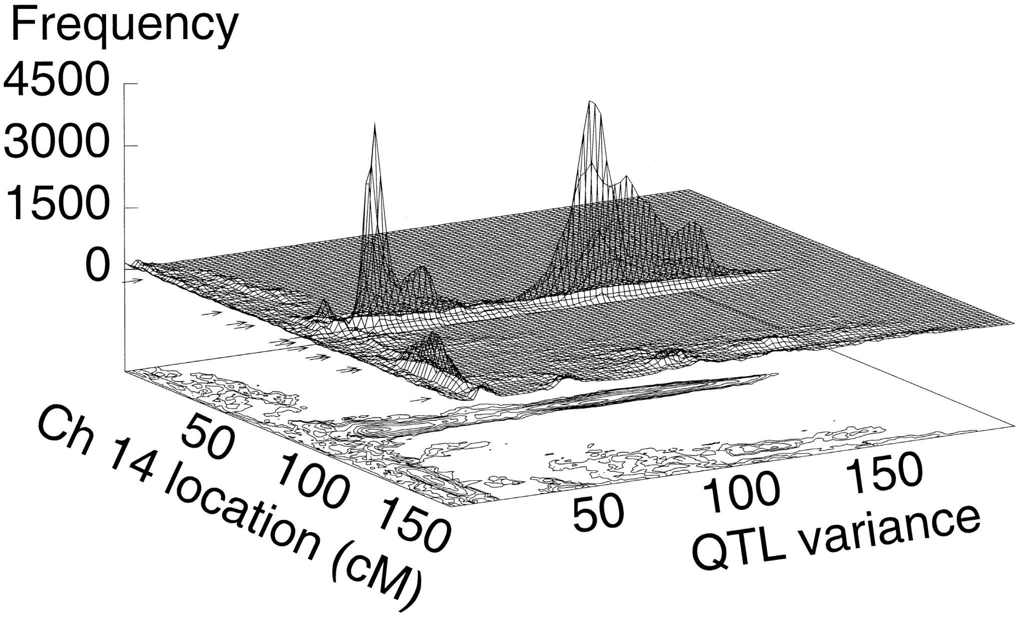

Figure 2 Histogram of chromosome location vs. variance, for every QTL placed on chromosome 14 in all iterations of one analysis run, using all available families and 12 markers on chromosome 14. Arrows indicate marker locations; line indicates PS1 location. Contour lines below histogram are at 25, 50, 100, 200, 400, 800, 1,600, and 3,200 counts.

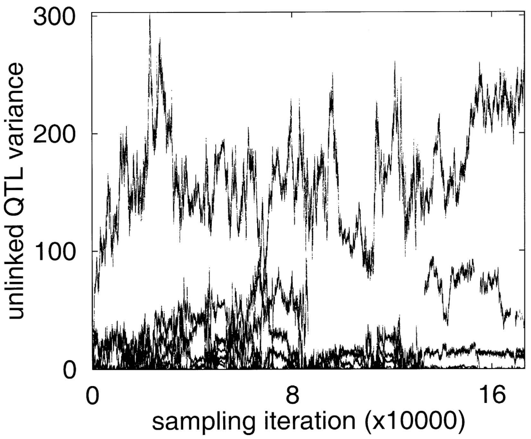

Figure 4 Variance of each QTL not placed on chromosome 14, in each iteration in the all-families analysis.

and evidence for additional genetic effects, demonstrating the potential usefulness of this method. The all-families chromosome 14 analysis correctly localized the PS1 gene, found evidence for a second gene on chromosome 14, and found segregation evidence for several other genes elsewhere, including one suggesting the PS2 gene in the VG families. The VG-families analysis with marker data on chromosomes 1, 5, and 14 correctly localized the PS2 gene and found segregation evidence for other AD genes, probably indicating a genetic factor in the non-PS2 AD cases that is not on chromosome 1, 5, or 14. The localizations appear to be at least as good as those in the original analysis and were obtained in significantly less time.

All-Families Chromosome 14 Analysis

The analysis of chromosome 14 with data from all 84 families provides validation for the age-at-onset model. This analysis not only produced a good localization of the PS1 gene; it also produced several additional signals, suggesting the presence of other genes. Figures 2, 3, and 4 are plots from one analysis run, focusing on three variables: map position (in centimorgans) of the QTLs placed on chromosome 14, the variance resulting from each QTL, and the MCMC sampling iteration number.

Figure 2 shows a strong signal, placing a large-effect QTL for age at onset on chromosome 14 in the all-families analysis. This figure is a histogram of all QTLs placed on chromosome 14, showing map position versus QTL variance. The variance that is attributed to each QTL is a measure of the effect size of each locus and depends on the allele frequencies and the effect size of each genotype. The pairwise plot of QTL positions versus variance is more effective for comparing the essential features of the results than is a simple histogram showing the number of QTL map placements at each map position, because very-small-variance QTLs can be more easily placed at random locations along the chromosome than larger-variance QTLs. Indeed, one can observe a smear of easily placed small-variance QTLs across chromosome 14. At larger variances, the histogram is relatively flat, away from the ridge produced by the many QTLs placed at the known PS1 location, which provide a localization for this gene. Thus, this plot helps separate signals from background noise. Note that there were two distinct peaks with different variances at PS1. The smaller of the two peaks at PS1 (variance ~25) had its maximum ~1 cM from the known gene location, whereas the larger (variance ~100) was ~3 cM away. We used the absolute measure of QTL variance in this plot rather than a relative measure, primarily because we had a fixed prior distribution on the QTL variance. Thus, the ease with which a QTL can be placed at a location is related to its absolute variance, and so absolute variance should be used as a means to help separate signals from noise.

A third peak, possibly indicating a previously unknown gene, can be seen, broadly centered between the PI/AACT markers and D14S1, at a modest QTL variance. A number of QTLs at larger variances were also placed in this region. The less accurate localization of this peak is not surprising, given the wider marker spacing and relatively low informativeness of the PI and AACT markers.

The plot of iterations versus location for all QTLs placed on chromosome 14 (fig. 3) indicates both the strength of the signal and the mixing of the sampler. The consistency of placements over iterations indicates the strength of a signal, with strong signals producing clear lines across this plot, as can be seen in figure 3 near the PS1 location. If the sampler is mixing properly, we also expect to see breaks in such a line, indicating iterations where the QTL is removed from the signal region and then replaced. Furthermore, in most iterations, two QTLs were placed in the region of PS1, corresponding to the two peaks seen in figure 2. QTLs contributing to the third peak appeared after ~90,000 iterations in the top of the plot. A QTL is not always placed between PI/AACT and D14S1 but comes and goes for hundreds or thousands of iterations. Similarly, two QTLs are not placed in the region of PS1 in all iterations, but whenever one is removed, it is eventually replaced, indicating proper mixing.

Evidence for genes on other chromosomes is illustrated by figure 4, which shows iteration versus variance for all QTLs, placed in each state, that were not linked to chromosome 14. A large-variance QTL that could not be placed on chromosome 14 can be seen in this plot. This QTL is probably indicative of the PS2 mu-

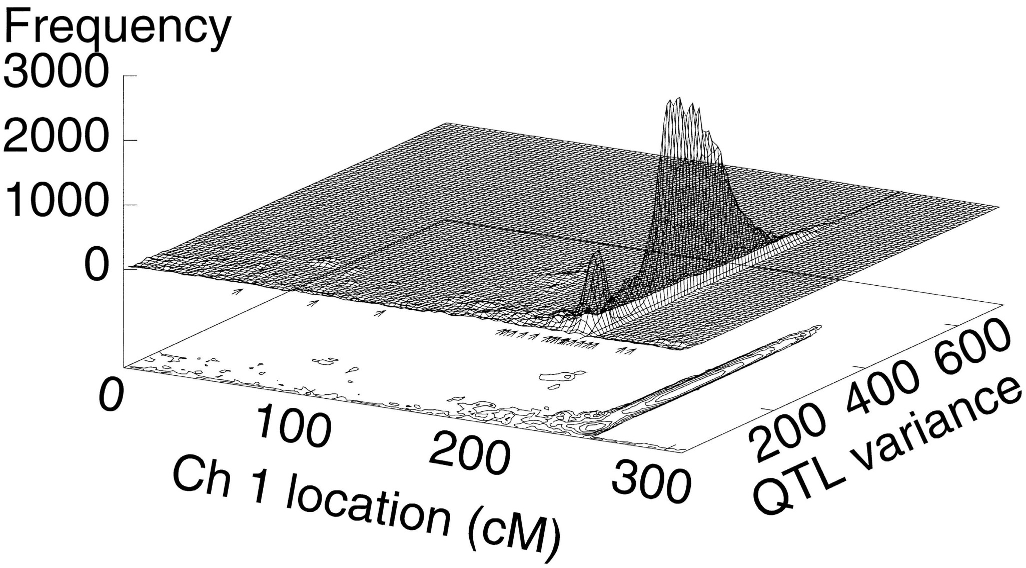

Figure 5 Histogram of chromosome location vs. variance, for every QTL placed on chromosome 1, in all iterations of one analysis run, using the VG families and 40 markers on chromosomes 1, 5, and 14. Arrows indicate the location of the 19 markers on chromosome 1.

tation in the VG families. This conclusion is supported by the variance contribution of this unlinked QTL, which is consistent with that found on chromosome 1 in the VG-family analysis (see figure 5). There was also evidence of several smaller-variance QTLs not linked to chromosome 14 (fig. 4), some of which might correspond to genes causing AD in the LO families.

With MCMC methods, the repeatability of an analysis run is an important question in determining whether the sampler is mixing well and running for enough iterations. Whereas figures 2–4 are plots from a single representative analysis run, ~24 runs were performed with different random number seeds and some with different parameter values. Each analysis run consisted of 20,000–200,000 iterations, and all had features in common with the run presented here. All had ⩾1 peak centered near PS1, and nearly all runs had 2 peaks, with variances tending to be centered near 40 and 100. Those that did not have two peaks at PS1 were all shorter (~20,000 iteration) analysis runs, that likely had not run long enough. Most runs found a large QTL, with variance of ~200, that was not placed on chromosome 14, which we believe is attributable to the PS2 gene in the VG families. Finally, like the analysis run shown here, most runs (including all the longer analysis runs) provided some evidence of an additional QTL telomeric to PS1. However, only a few analysis runs provided a distinct third peak like the one shown here. In those that did develop a third peak, the peak tended to appear after a large number of iterations, perhaps indicating that the program will need to be run for more iterations to localize weaker signals. Additionally, an analysis run using a sparser subset of the chromosome 14 markers also localized a QTL to the PS1 region, although the localization was not as accurate. This sparser subset of markers contained more-uniform information along the length of the chromosome.

Three-Chromosome Analysis of VG Families

Our analysis of the VG families, using 40 markers on chromosome 1, 5, and 14, further validated the age-at-onset model by providing a good localization of the known mutation (PS2), as well as some additional exclusion information. The histograms showing location versus QTL variance for the three chromosomes (figures 5, 6, and 7) showed a sharp peak at the PS2 location, with very few QTLs placed elsewhere on these chromosomes. This peak was centered at the location of the PS2 gene. Additionally, many more QTLs were placed on chromosome 1 (214,028) than on chromosome 5 (14,554) and chromosome 14 (44,776), and the vast majority of the points on chromosome 1 were placed in the region of PS2. On the basis of the prior probability of linkage and the number of QTLs found in each iteration, in the absence of any linkage information, each chromosome should have had ~47,500 placements, giving a threshold from which to evaluate evidence of linkage or exclusion. Although we plan to address calibration of results from this method in future work, these

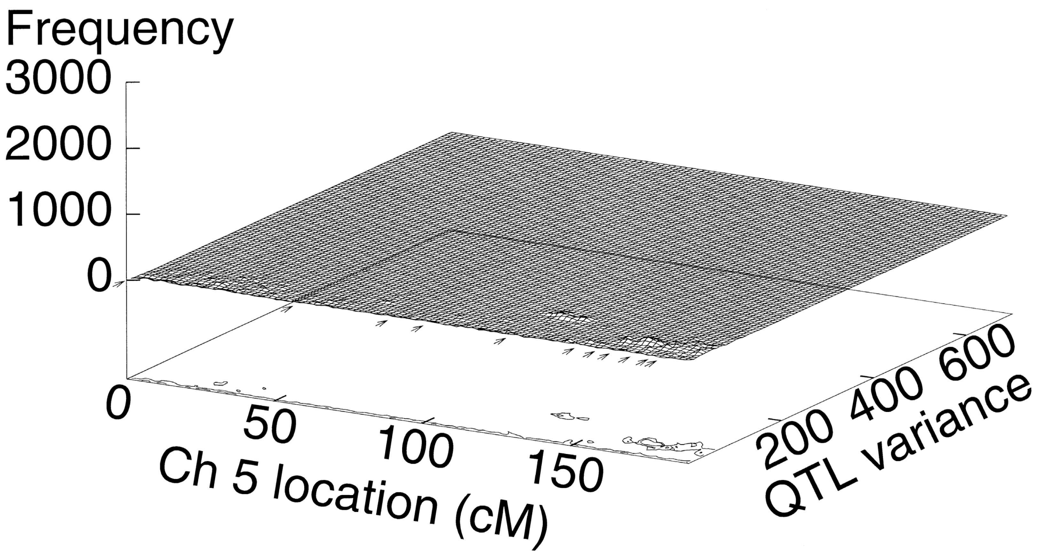

Figure 6 Histogram of location vs. variance, for every QTL placed on chromosome 5, in the VG-families analysis run. Arrows indicate the 11 marker locations.

results gave strong evidence for a QTL on chromosome 1 at or near PS2 but gave little evidence for any QTLs on chromosome 5, 14, or elsewhere on chromosome 1, in the VG families.

The exclusion of chromosomal regions for linkage appeared, not surprisingly, to be somewhat dependent on the quality of the markers, with the better overall marker quality on chromosome 5 leading to its "exclusion" from linkage. In some regions of chromosome 14, such as that near PS1, the marker information was better than for chromosome 5. Fewer QTLs were placed in these regions, suggesting that PS1 mutations did not occur in the VG families. However, when the heterozygosities of all the markers and the distribution of the markers over each chromosome were considered, the marker information spanning chromosome 5 was actually somewhat better than that spanning chromosome 14, resulting in a smaller number of QTLs placed on chromosome 5 than on chromosome 14. Consequently, chromosome 5 is excluded from linkage to AD QTLs in the VG families, whereas, although some regions of chromosome 14 can be excluded, the chromosome, as a whole, cannot.

Evidence for two genes segregating can be seen in the plot of iterations versus variance for all QTLs not placed on any of the three chromosomes (fig. 8). First, comparison of the unlinked QTLs and the iteration versus QTL location on the chromosome 1 plot (fig. 9) showed that, in those iterations where a QTL was not placed near PS2, there tended to be an unlinked large-variance QTL, and when a large-variance QTL was placed near PS2 there tended not to be a large-variance QTL that was unlinked. Second, there was evidence for a moderate-variance QTL not linked to chromosomes 1, 5, or 14, possibly representing a gene in the two families not found to have PS2 mutations.

The consistency was even greater among the multiple analysis runs of the VG families than among the chromosome 14–only analysis runs. In contrast to the chromosome 14–only analysis, the results from each of the 3-chromosome analysis runs for the VG families were nearly identical to those presented here, as were results from analysis runs, using only the VG families, of chromosomes 1 and 5, chromosomes 1 and 14, and chromosome 1 alone. In addition, analysis runs using only 9 of the 20 markers on chromosome 1 also localized a QTL in the PS2 region, although the localization was more accurate with the full set of markers.

Computer Requirements

Both of the analyses presented here were accomplished relatively quickly. Each took ~4 d on a DEC Alpha Station 500 running at 333 MHz. Although the all-families chromosome 14 analysis included more subjects, the three-chromosome analysis of the VG families had more markers, resulting in very similar needs for computing time. The memory requirements of this algorithm were a very modest 10–12 MB total memory, with only 3–4 MB of real memory. Both of these analyses were accomplished in much less time than is required for multipoint analysis using exact methods. For comparison, a previous three-marker, single disease–locus analysis of just the EO and VG families required comparable computer time (~3 mo on a computer ~30 × slower [Schellenberg et al. 1992]). The version of the analysis program (Loki 2.1) now available is ~4 × faster than the version available at the time of the analyses presented here. It is not possible to estimate accurately the time required for an exact 12-marker or 40-marker multiple disease–locus analysis with traditional methods, because not only would the computation time be exceptionally long, but the memory requirements would be far beyond the capacity of any computer currently available to us.

Discussion

The results presented here demonstrate the flexibility, robustness, and manageability of these MCMC methods in performing a multipoint analysis of large pedigrees with age-at-onset data. We used these methods, as implemented in the program, Loki, to analyze data from large pedigrees with many polymorphic markers for Alzheimer disease, a complex late-onset oligogenic trait. The utility of coupling the age-at-onset model with MCMC methods is indicated by the fact that, if we had no prior knowledge of the disease, we would still clearly conclude that there are large-effect QTLs for AD at approximately the PS1 and PS2 locations. Furthermore, evidence was found indicating that several additional genes may contribute to AD, and one of these may be located on chromosome 14. Although these results are interesting in themselves, the methods should be applicable to the genetic dissection of other "diseases of aging" and possibly to other diseases with age-at-onset data.

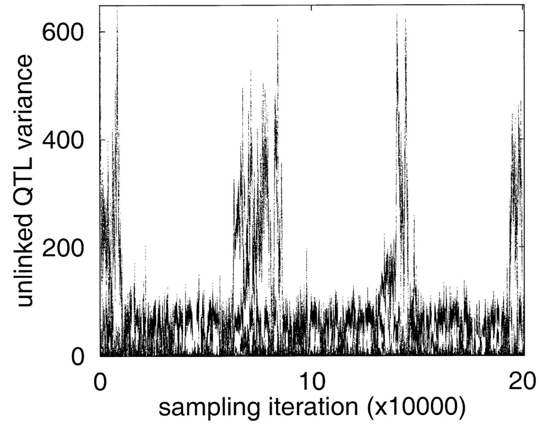

Figure 8 Variance of each QTL not placed on chromosomes 1, 5, or 14, in each iteration of the VG-families analysis.

Gene Localization

The accuracy of the localization of known genes in the AD data analyses was quite good, even in the presence of genetic heterogeneity (which was substantial in the all-family analyses). In the vast majority of iterations, a QTL was placed within 10 cM of the known gene, and the distribution of these QTLs, both in the analysis runs shown here and in those not shown, always had its peak within a few centimorgans of the known gene. An adequate localization usually was obtained after 50,000 iterations, although additional iterations improved the localization. The accuracy of the localization was improved by the relatively dense marker sets typed

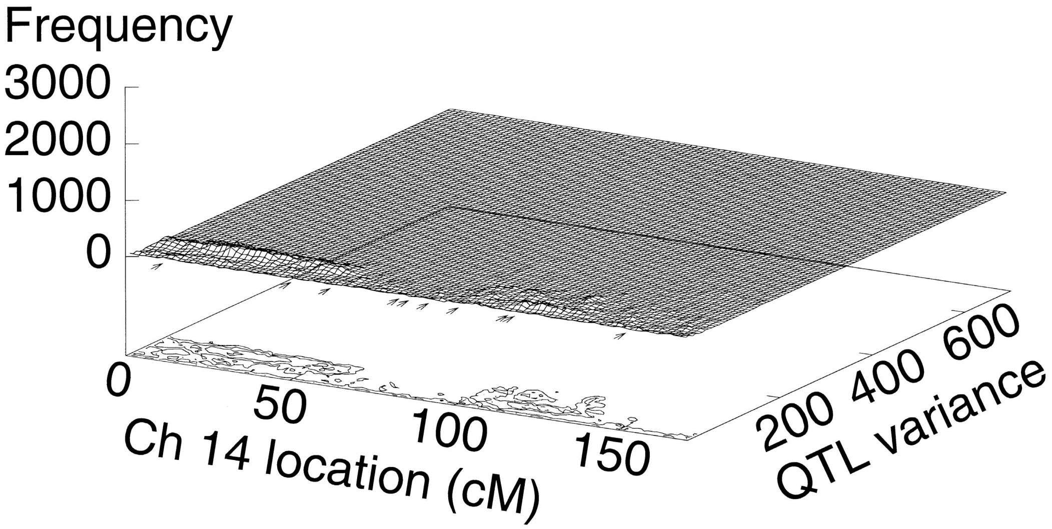

Figure 7 Histogram of locations vs. variance, for every QTL placed on chromosome 14, in the VG-families analysis run. Arrows indicate the 10 marker locations.

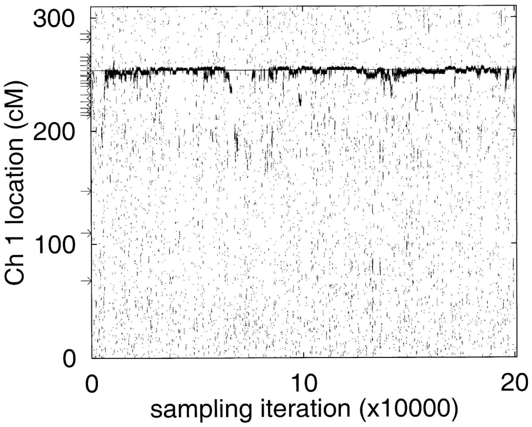

Figure 9 Location of each QTL placed on chromosome 1, in each iteration of the VG families analysis. Arrows indicate marker locations; line indicates PS2 location.

around the PS1 and PS2 genes, but the runs with sparser maps indicate that a dense map is not required for detection. This suggests that these analysis methods would be amenable to a two-stage marker screening process. A 20-cM initial screen, for example, could be followed up by a 5-cM screen in areas that produced positive evidence of linkage. These localizations are also comparable to, if not better than, those obtained with traditional linkage analyses. Furthermore, in the presence of the significant heterogeneity that defined this data set, traditional exact methods used on the full all-family data set would not have detected evidence of linkage to chromosome 14. In the original analysis of these data (Schellenberg et al. 1992), the EO group was analyzed separately, and this allowed evidence of linkage to be detected. Unfortunately, stratification schemes for many diseases are not readily apparent, or they require a great deal of time and effort to devise and carry out. The methods implemented here appear to eliminate some of the need to predetermine subgroups for analysis, although there is still a need to determine which QTLs are segregating in which families, after localizing a gene. This will be addressed in future work.

The double peak at PS1 has two likely explanations. First, the EO families are known to have several different PS1 mutations. However, the model used here assumes a diallelic QTL. If these PS1 mutations in fact result in different mean onset ages, the additional alleles could be fitted by placing a second QTL in the same position. Second, the assumption is made that the QTL distributions are normal, so a second QTL might represent a compensation for nonnormality. The first hypothesis is supported by the facts that all the VG families with PS2 mutations were known to have the same mutation and that only a single peak was found at PS2. Whether either of these explanations (or perhaps some other explanations) is true, the accurate gene localization of the two known AD genes, in the presence of substantial genetic heterogeneity, indicates that this analysis model is sufficiently robust to deal with the situation as it stands, since, in either case, we would conclude that there was a gene at PS1.

The most curious aspect of these analyses was the possible second QTL on chromosome 14. It could not be characterized as a clear signal because the signal seen in the analysis run presented here was not present in all analysis runs. The only consistent factor was the placement of somewhat more QTLs in the region than either were seen in regions with clearly no linkage to the disease or were expected, on the basis of the prior distribution of localizations. These placements led us to conclude that there was, in fact, a signal in the data at that region, but it was quite weak and could be a result of sampling error, of low heterozygosity of the markers in the region, or of the considerable missing data in the LO families. These results will need follow-up of typing additional markers on chromosome 14 in future work.

Unlinked QTLs

Our analyses found evidence for additional unlinked genetic effects, in addition to those localized on chromosomes 1 and 14. The largest of these unlinked QTLs in the chromosome 14 all-families analysis probably represents PS2. There were also more unlinked QTLs with larger variances in the all-families analysis than in the VG-families analysis, which suggests the presence of additional genetic factors in the LO families. It should be noted that a single QTL located elsewhere may, in fact, represent several trait loci. There is no way to distinguish between two loci with similar effects with segregation analysis. The number of unlinked QTLs may therefore indicate the minimum number of additional QTLs located elsewhere on the genome. However, a QTL could also arise in the model as a result of mitochondrial inheritance, a transmissible environmental factor, or simple noise due to, for example, deviation from the normality assumption. Also, an unlinked QTL may be the result of mistyping or genetic map misspecification that prevents placement into the correct location. In addition, QTLs may need some iterations for their parameter values to “settle down” before they can be placed on a chromosome, and, as the parameter values wander, a QTL may go through periods of not being linked, as can be seen with the large-effect QTL in the VG-families analysis. Thus, the number of unlinked QTLs typically present in an analysis can be taken as a rough estimate of the number of genes to be found on other chromosomes.

Exclusion

Since these methods allow exploration of a wide range of additive models, exclusion under these methods is more meaningful than exclusion with traditional mapping techniques for a complex trait. In a region with no linkage information from the markers, one would expect to see a distribution of QTLs congruent to those placed "elsewhere," with the frequency proportional to the prior possibility of linkage to that region. It is more difficult to exclude smaller-variance QTLs, because these effects can be placed almost anywhere. In the VG-families analysis, for example, we excluded chromosome 5 for all but the smallest-variance QTLs. On chromosome 14, with the VG families, in regions where the markers were good, all but the smallest-variance QTLs were also excluded. In regions where the marker coverage was poor, or where the markers were not very polymorphic, however, only large-variance QTLs could be excluded, although there was certainly no strong support for moderate-variance QTLs localized on chromosomes 14 in the VG families. On chromosome 1, there was a strong signal at the PS2 location, but other regions appeared to be excluded. The second signal on chromosome 14 in the all-families analysis was best characterized as something between a failure to exclude the region and positive evidence for localization of a gene.

Age-at-Onset Model

The limitations of the age-at-onset model used here are not yet fully apparent. There is an implicit assumption that everyone will get the disease if they do not die of something else first. This means that, if a reasonable number of people do not get the disease, they would have to be given a QTL genotype with mean onset significantly after death. This could cause problems, if one were to use this method for, say, a childhood-onset disease. The distributions for the "well" and "sick" individuals would be so far apart that there might be a very low probability of midlife onset, causing computational problems or preventing proper mixing of the MCMC sampler. Thus, the model is best-suited to diseases in which onset can and does occur at a wide range of ages and will likely work particularly well for other "diseases of aging," such as heart disease, depression, and some cancers that become increasingly prevalent as people age. That this model did localize the PS2 gene in the VG-only sample, in which most of the onsets were in fact before an age at which death might occur more typically, suggests that this method might work for diseases with a midlife onset, such as schizophrenia, but the usefulness of this method for such mid-onset diseases remains to be determined.

Another factor that may limit these methods is the assumption that the ages at onset for each genotype are normally distributed, with each distribution having the same mean. Such an assumption would be consistent with, for example, a disease in which the genes determined a basic "fitness," and then random environmental "hits" occurred throughout life. This normality assumption is not unreasonable for many diseases, including AD, in which the onset ages of the chromosome 14 PS1 mutation–carrying individuals has been observed to be normally distributed. Even if the assumption of normality is violated, there may be a transformation of the age data that does not violate the assumption.

The model described in the present paper represents an important advance in the development of methods applicable for whole-genome linkage analysis of late-onset disease data. By taking age at onset as our quantitative trait with censoring information of unaffected individuals, we have a more informative phenotype than simply considering affectation status or age at onset in affected subjects, and thus more power to dissect complex traits. It is clear that this method worked well with our existing AD data, finding both known AD genes. We believe that this model will work well with most common diseases of a degenerative nature that have a variable onset age and that it may also help with other phenotypes with which a censoring model is appropriate.

Acknowledgments

The authors wish to thank Elizabeth Thompson for helpful discussions and comments. This work was supported by National Institutes of Health grants GM46255, AG11762, and AG05136.

Electronic-Database Information

URLs for data in this article are as follows:

Genome Database, http://www.gdb.org (for marker information)
PANGAEA, http://www.stat.washington.edu/thompson/pangaea.html (for Loki)
Rockefeller database, http://linkage.rockefeller.edu/index.html (for marker information)

References

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543

Carlin B, Chib S (1995) Bayesian model choice via Markov chain Monte Carlo methods. J R Stat Soc B 57:473–483

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, et al (1993) Gene dose of apolipoprotein-e type-4 allele and the risk of Alzheimer's Disease in late onset families. Science 261:921–923

Cottingham RW, Idury RM, Schaffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53: 252–263

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732

Greenberg DA, Hodge SE (1989) Linkage analysis under "random" and "genetic" reduced penetrance. Genet Epidemiol 6:259–264

Greenberg DA, Hodge SE, Vieland VJ, Spence MA (1996) Affecteds-only linkage methods are not a panacea. Am J Hum Genet 58:892–895

Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51:1111–1126

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3–19

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

Heath SC (1995) Inferences on the genetic control of quantitative traits from selection experiments. PhD thesis, Edinburgh University Press, Edinburgh

———(1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760

Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM (1997) MCMC segregation and linkage analysis. Genet Epidemiol 14:1011–1016

Hoeschle I (1994) Bayesian QTL mapping via the Gibbs sampler. In: Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW et al (eds) Fifth world congress on genetics applied to livestock production. Vol 21. University of Guelph Press, Ontario, pp 241–244

Ikeda M, Sharma V, Sumi SM, Rogaeva EA, Poorkaj P, Sherrington R, Nee L, et al (1996) The clinical phenotype of two missense mutations in the presenilin I gene in Japanese patients. Ann Neurol 40:912–917

Jarvik GP, Austin MA, Fabsitz RR, Auwerx J, Reed T, Christian JC, Deeb S (1994) Genetic influences on age-related change in total cholesterol, low density lipoprotein-cholesterol, and triglyceride levels: longitudinal apolipoprotein E genotype effects. Genet Epidemiol 11:375–384

Knapp M, Seuchter SA, Baur MP (1994) Two-locus disease models with two marker loci: the power of affected-sib-pair tests. Am J Hum Genet 55:1030–1041

Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56: 519–527

Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, Yu CE, et al (1995*a*) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science 269:973–977

Levy-Lahad E, Wijsman EM, Nemens E, Anderson L, Goddard KA, Weber JL, Bird TD, et al (1995*b*) A familial Alzheimer's disease locus on chromosome 1. Science 269:970–973

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

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402–408

Olson JM, Wijsman EM (1993) Linkage between quantitative trait and marker loci: methods using all relative pairs. Genet Epidemiol 10:87–102

Ott J (1979) Genetic linkage studies in man. Transplant Proc 11:1689–1691

Phillips D, Smith A (1996) Bayesian model comparison via jump diffusions. In: Gilks W, Richardson S, Spiegehalter D (eds) Markov chain Monte Carlo in practice. Chapman & Hal, London, pp 215–219

Poorkaj P, Sharma V, Anderson L, Nemens E, Alonso ME, Orr H, White J, et al (1998) Missense mutations in the chromosome 14 familial Alzheimer's disease presenilin 1 gene. Hum Mut 11:216–221

Rice JP, Neuman RJ, Burroughs TE, Hampe CL, Daw EW, Suarez BK (1993) Linkage analysis for oligogenic traits. Am J Hum Genet 53 Suppl:A66

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

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805–816

Schellenberg GD, Bird TD, Wijsman EM, Orr HT, Anderson LL, Nemens E, White JA, et al (1992) Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. Science 258:668–671

Schellenberg GD, Payami H, Wijsman EM, Orr HT, Goddard KAB, Anderson L, Nemens E, et al (1993) Chromosome-14 and late-onset familial Alzheimer disease (FAD). Am J Hum Genet 53:619–628

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127–1136

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, et al (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 375:754–760

Stephens DA, Smith AFM (1993) Bayesian inference in multipoint gene mapping. Ann Hum Genet 57:65–82

Thompson EA (1994*a*) Monte Carlo likelihood in genetic mapping. Stat Sci 9:355–366

———(1994*b*) Monte Carlo likelihood in the genetic mapping of complex traits. Phil Trans R Soc Lond B 29:345–351

Vieland V, Greenberg DA, Hodge SE, Ott J (1992*a*) Linkage analysis of two-locus diseases under single-locus and two-locus analysis models. Cytogenet Cell Genet 59:145–146

Vieland VJ, Hodge SE, Greenberg DA (1992*b*) Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet Epidemiol 9:45–59

Wijsman EM, Amos C (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 14: 719–735