Genetic Epidemiology 31: 103–114 (2007)

MCMC-Based Linkage Analysis for Complex Traits on General Pedigrees: Multipoint Analysis With a Two-Locus Model and a Polygenic Component

Yun Ju Sung,[1] Elizabeth A. Thompson,[2] and Ellen M. Wijsman[1,3,4]*

[1]*Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington*
[2]*Department of Statistics, University of Washington, Seattle, Washington*
[3]*Department of Biostatistics, University of Washington, Seattle, Washington*
[4]*Department of Genome Science, University of Washington, Seattle, Washington*

We describe a new program `lm_twoqtl`, part of the MORGAN package, for parametric linkage analysis with a quantitative trait locus (QTL) model having one or two QTLs and a polygenic component, which models additional familial correlation from other unlinked QTLs. The program has no restriction on number of markers or complexity of pedigrees, facilitating use of more complex models with general pedigrees. This is the first available program that can handle a model with both two QTLs and a polygenic component. Competing programs use only simpler models: one QTL, one QTL plus a polygenic component, or variance components (VC). Use of simple models when they are incorrect, as for complex traits that are influenced by multiple genes, can bias estimates of QTL location or reduce power to detect linkage. We compute the likelihood with Markov Chain Monte Carlo (MCMC) realization of segregation indicators at the hypothesized QTL locations conditional on marker data, summation over phased multilocus genotypes of founders, and peeling of the polygenic component. Simulated examples, with various sized pedigrees, show that two-QTL analysis correctly identifies the location of both QTLs, even when they are closely linked, whereas other analyses, including the VC approach, fail to identify the location of QTLs with modest contribution. Our examples illustrate the advantage of parametric linkage analysis with two QTLs, which provides higher power for linkage detection and better localization than use of simpler models. *Genet. Epidemiol*. 31:103–114, 2007.

Key words: Markov chain Monte Carlo; quantitative trait locus; large pedigree; lod score; multilocus

The Supplemental materials described in this article can be found at http:// www. interscience.wiley.com/jpages/0741-0395/suppmat
Contract grant sponsor: NIH; Contract grant numbers: GM46255 and HD35465.
*Correspondence to: Dr. Ellen M. Wijsman, Division of Medical Genetics, University of Washington, Box 357720, Seattle, WA 98195-7720.
E-mail: wijsman@u.washington.edu
Received 23 June 2006; Accepted 4 October 2006
Published online 22 November 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20194

INTRODUCTION

Complex traits are influenced by multiple, possibly interacting, loci. Most linkage programs, originally developed for finding simple Mendelian traits, assume that only a single locus influences the trait. These programs are routinely used for complex traits as well, but with only modest success [Altmuller et al., 2001; Glazier et al., 2002]. This has led to increased interest in developing linkage methods that accommodate more than one trait locus. Modeling two trait loci as one can lead to incorrect inference: for discrete traits, it has been shown that both parametric and nonparametric analyses with two trait loci can give higher power for linkage detection and can provide more accurate localization than analyses with one trait locus [Schork et al., 1993; Knapp et al., 1994; Strauch et al., 2003]. For discrete traits, several methods have been developed for linkage analyses for two trait loci. The program TMLINK [Lathrop and Ott, 1990] can perform parametric lod score analysis with two unlinked trait loci for large pedigrees with a limited number of markers. The program GENEHUNTER-TWOLOCUS [Strauch et al., 2000] can perform both parametric and nonparametric analyses with two unlinked trait loci for many markers with limited pedigree sizes. For analysis with two linked trait loci, Biernacka et al. [2005] use identity-by-descent

(IBD) sharing in affected sib pairs and Biswas et al. [2003] use a Bayesian approach. All of these two-locus analysis approaches have practical limitations on the number of markers and pedigree sizes.

For continuous traits, few methods have been developed for linkage analyses that accommodate more than one trait locus. The variance components (VC) method is very popular and has been extended to several trait loci in SOLAR [Almasy and Blangero, 1998]. However, the VC method typically has lower power for linkage detection and provides less accurate localization than full model-based approaches, as do other IBD sharing methods [Greenberg et al., 1998; Abreu et al., 1999; Sham et al., 2000; Badzioch et al., 2005]. Trait localization with VC methods does not improve with dense markers [Atwood and Heard-Costa, 2003]. The VC method is also sensitive to pedigree ascertainment issues, yielding both false-positive and false-negative results [Allison et al., 1999; Forrest and Feingold, 2000]. PAP [Hasstedt, 1982], and can perform linkage analysis with one trait locus and a polygenic component, but uses an analytic approximation on larger pedigrees and can handle only a limited number of markers. Loki [Heath, 1997] can perform joint linkage and segregation analysis with multiple Quantitative Trait Loci (QTLs) on general pedigrees. However, the Bayesian approach of Loki prevents easy comparison of significance of results with those obtained by more traditional lod score methods.

Markov chain Monte Carlo (MCMC) methods are useful when the exact likelihood is intractable due to the number of markers and pedigree sizes. Programs for exact likelihood computation for one trait locus implement either the Elston-Stewart algorithm [Elston and Stewart, 1971] that handles large pedigrees but only a limited number of markers (LIPED [Ott, 1974], LINKAGE [Lathrop and Lalouel, 1984], FASTLINK [Cottingham et al., 1993; Schäffer et al., 1994], and VITESSE [O'Connell and Weeks, 1995; O'Connell, 2001]) or the Lander-Green algorithm [Lander and Green, 1987] that handles multiple markers but only limited pedigree sizes (GENEHUNTER [Kruglyak et al., 1996; Markianos et al., 2001], ALLEGRO [Gudbjartsson et al., 2000], and MERLIN [Abecasis et al., 2002]). SimWalk [Lange and Sobel, 1991; Sobel and Lange, 1996] and MORGAN [Thompson et al., 1993; Thompson, 2005] use MCMC methods to handle more markers and larger pedigrees for one trait locus; both have been used for real-data analyses [Cader et al., 2005; Gagnon et al., 2005; George et al., 2005; Hwu et al., 2005; Orlacchio et al., 2005; Igo et al., 2006]. An MCMC method also has been implemented for discrete traits with two unlinked trait loci to handle more markers and larger pedigrees [Lin, 2000].

Here we describe a new MORGAN program, `lm_twoqtl`. This is the first available linkage program that can perform analysis with a parametric model that includes two QTLs and a polygenic component, which is a more complex trait model than can be analyzed with previously available programs. Because the program uses MCMC, it has no restriction on number of markers or complexity of pedigrees. We provide two examples to illustrate its practical utility. The first has data sets with various sized pedigrees, and the second has data sets with various QTL spacings. All data sets were analyzed with four different models: two by `lm_twoqtl`, one by `lm_markers` [Thompson, 2005; Sieh et al., 2005], and one by SOLAR [Almasy and Blangero, 1998] using its VC model. Our examples illustrate the advantage of parametric linkage analysis with two QTLs, providing higher power for linkage detection and better localization than analyses with simpler models.

METHODS

TWO-QTL AND POLYGENIC COMPONENT MODEL

The data consist of marker genotypes Y_M and a quantitative trait Y_T, and possibly covariates, measured on observed individuals in a pedigree. The model we use for trait data is

$$Y_T = Q_1 + Q_2 + Z + X\beta + E \qquad (1)$$

where Q_1 and Q_2 are two QTL effects; Z is the polygenic value; X is the matrix of covariates, such as sex or age; β is the covariate effects parameter; and E is an environmental effect. QTL effects Q_1 and Q_2 are discrete with three levels μ_{AA}, μ_{Aa}, μ_{aa} and μ_{BB}, μ_{Bb}, μ_{bb}, respectively, corresponding to the three genotypes at each locus; the polygenic value Z is normally distributed with mean 0 and variance $2\Phi\sigma_a^2$, where Φ is the kinship matrix [Lange, 2002, p. 81–84]; and E is normally distributed with mean 0 and variance $I\sigma_e^2$, where I is the identity matrix.

We follow a standard genetic model for unobserved phased genotypes, which specifies the joint distribution of marker data and unobserved

QTL genotypes. The model for the phased genotypes of founders is specified by Hardy-Weinberg equilibrium at each locus and linkage equilibrium over loci, which is generally adequate when markers are not extremely dense. The phased genotypes of nonfounders are specified by the phased genotypes of founders and the segregation indicators [Thompson, 2000, p. 3–4], which indicate which genes nonfounders inherit from their parents. We assume there is no cross-over interference, which implies the Haldane map function. While an approximation, this map function has served well for linkage analysis of human data. Marker data, when observed, are assumed to be observed without error. The allele frequencies and locations of markers are assumed known. Most linkage analysis programs impose these assumptions.

MCMC ESTIMATION OF LOD SCORE

The lod score at a hypothesized bivariate QTL location (λ_1, λ_2) is the log of the likelihood ratio

$$\mathrm{lod}(\lambda_1, \lambda_2) = \log_{10} \frac{L(\lambda_1, \lambda_2)}{L(-\infty, \infty)} \tag{2}$$

where $L(\lambda_1, \lambda_2)$ denotes the likelihood at (λ_1, λ_2). Other parameters, allele frequencies of QTLs and marker loci, genotypic means of QTLs, additive genetic and environmental variances, are fixed and assumed known for the purpose of analysis. PAP [Hasstedt, 1982] and Loki [Heath, 1997] are two possible programs that can provide estimates of these parameters. Our null hypothesis is that both QTLs are unlinked to the marker loci and to each other: we denote their bivariate location under this null hypothesis by $(-\infty, \infty)$ in (2). The likelihood of (λ_1, λ_2) is the joint probability of trait data Y_T and marker data Y_M and can be written as

$$L(\lambda_1, \lambda_2) = \mathrm{Prob}_{\lambda_1,\lambda_2}(Y_T|Y_M) \cdot \mathrm{Prob}|(\mathrm{Y_M}).$$

Since $\mathrm{Prob}(Y_M)$ does not depend on the QTL location (λ_1, λ_2), the lod score in (2) can be written as

$$\mathrm{lod}(\lambda_1, \lambda_2) = \log_{10} \frac{\mathrm{Prob}_{\lambda_1,\lambda_2}(Y_T|Y_M)}{\mathrm{Prob}_{-\infty,\infty}(Y_T|Y_M)}. \tag{3}$$

We use MCMC to estimate the lod score in (3) because exact computation is intractable for large or complex pedigrees genotyped with multiple markers. In particular, we use MCMC to estimate both conditional probabilities in (3). Note that

$$\begin{aligned}\mathrm{Prob}_{\lambda_1,\lambda_2}(Y_T|Y_M) &= \sum_{S_{\lambda_1},S_{\lambda_2}} \mathrm{Prob}_{\lambda_1,\lambda_2}(Y_T, S_{\lambda_1}, S_{\lambda_2}|Y_M) \\ &= \sum_{S_{\lambda_1},S_{\lambda_2}} \mathrm{Prob}(Y_T|S_{\lambda_1}, S_{\lambda_2}) \\ &\quad \cdot \mathrm{Prob}_{\lambda_1,\lambda_2}(S_{\lambda_1}, S_{\lambda_2}|Y_M) \end{aligned} \tag{4}$$

where S_{λ_1} and S_{λ_2} are segregation indicators at λ_1 and λ_2, respectively. The second equality follows from $\mathrm{Prob}(Y_T|S_{\lambda_1}, S_{\lambda_2}, Y_M) = \mathrm{Prob}(Y_T|S_{\lambda_1}, S_{\lambda_2})$, which results from the conditional independence of Y_T and Y_M, given S_{λ_1} and S_{λ_2}: marker data Y_M do not contain any more information about Y_T, once S_{λ_1} and S_{λ_2} are determined [Thompson, 2000, Section 6.1]. We first sample segregation indicators $S_M^{(i)}$ at the marker loci from $\mathrm{Prob}(S_M|Y_M)$ by MCMC, using a mixture of the L-sampler [Heath, 1997] and the M-sampler [Thompson and Heath, 1999] implemented in the MORGAN package. Then we sample $(S_{\lambda_1}^{(i)}, S_{\lambda_2}^{(i)})$ from $\mathrm{Prob}_{\lambda_1,\lambda_2}(S_{\lambda_1}, S_{\lambda_2}|S_M^{(i)})$ by ordinary (independent) Monte Carlo. By the conditional independence of S_λ's and Y_M given S_M [Thompson, 2000, Section 6.1], this produces $(S_{\lambda_1}^{(i)}, S_{\lambda_2}^{(i)})$ as MCMC realizations from $\mathrm{Prob}_{\lambda_1,\lambda_2}(S_{\lambda_1}, S_{\lambda_2}|Y_M)$. We estimate $\mathrm{Prob}_{\lambda_1,\lambda_2}(Y_T|Y_M)$ in (4) by

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{Prob}(Y_T|S_{\lambda_1}^{(i)}, S_{\lambda_2}^{(i)}). \tag{5}$$

This is a standard Monte Carlo technique [Hammersley and Handscomb, 1964; Gentle, 2002, Section 7.1]: replace a theoretical expectation by the sample average over realizations from the relevant probability distribution.

Our MCMC estimate in (5) requires evaluation of $\mathrm{Prob}(Y_T|S_{\lambda_1}, S_{\lambda_2})$. Note that

$$\begin{aligned}\mathrm{Prob}(Y_T|S_{\lambda_1}, S_{\lambda_2}) &= \sum_{G_{\lambda_1},G_{\lambda_2}} \mathrm{Prob}(Y_T|S_{\lambda_1}, S_{\lambda_2}, G_{\lambda_1}, G_{\lambda_2}) \\ &\quad \cdot \mathrm{Prob}(G_{\lambda_1}, G_{\lambda_2}), \end{aligned} \tag{6}$$

where G_{λ_1} and G_{λ_2} are the phased genotypes of founders at λ_1 and λ_2, respectively. The sum on the right in (6) is not ''peelable'' for our trait model having one or two QTLs and a polygenic component [Thompson, 2000, Section 6.6]. Hence summing over all possible G_{λ_1}, G_{λ_2} is not practical when the number of founders in any pedigree is large. In that case, we use Monte Carlo with

importance sampling

$$\frac{1}{n}\sum_{j=1}^{n}\text{Prob}(Y_T|S_{\lambda_1},S_{\lambda_2},G_{\lambda_1}^{(j)},G_{\lambda_2}^{(j)})\frac{\text{Prob}(G_{\lambda_1}^{(j)},G_{\lambda_2}^{(j)})}{\text{Prob}^*(G_{\lambda_1}^{(j)},G_{\lambda_2}^{(j)})} \quad (7)$$

where $(G_{\lambda_1}^{(j)},G_{\lambda_2}^{(j)})$ are a Monte Carlo sample from $\text{Prob}^*(G_{\lambda_1},G_{\lambda_2})$.

Monte Carlo with importance sampling for founder genotypes has not previously been used. This approach makes possible the computation of lod scores for these complex models on general pedigrees. Importance sampling is a standard Monte Carlo technique [Hammersley and Handscomb, 1964; Gentle, 2002, Section 7.5.2]: instead of sampling from the true probability distribution $\text{Prob}(G_{\lambda_1},G_{\lambda_2})$, we sample from another probability distribution $\text{Prob}^*(G_{\lambda_1},G_{\lambda_2})$ and use the weighted sample average in (7). Our $\text{Prob}^*(G_{\lambda_1},G_{\lambda_2})$ corresponds to using allele frequencies of QTLs that are different from the true allele frequencies. This is particularly useful when pedigrees are ascertained on one tail of the phenotypic distribution: such pedigrees often segregate relatively rare alleles, in which case sampling with more common allele frequencies increases the chance of including such rare alleles in the sample.

Finally, we evaluate $\text{Prob}(Y_T|S_{\lambda_1},S_{\lambda_2},G_{\lambda_1},G_{\lambda_2})$ in both (6) and (7) by peeling the polygenic component. We use an algorithm that does not require the inverse of the variance matrix [Henderson, 1976; Quaas, 1976] and is applicable to general pedigrees with loops. In order to improve efficiency, each pedigree is separately evaluated.

PROFILE AND SLICED LOD SCORES

For the trait model having two QTLs and a polygenic component, two-dimensional lod scores are evaluated, simultaneously changing both QTL locations. To compare these two-dimensional lod scores with one-dimensional lod scores from trait models having one QTL, with and without a polygenic component, we summarize two-dimensional lod scores with two one-dimensional lod scores: profile and sliced lod scores. We explain these lod scores for the case where one QTL is weaker than the other, measured by the variance of their contribution to phenotypes: we call the QTL with smaller variance contribution ''weak'' and the QTL with larger variance contribution ''strong''.

The profile lod score at a hypothesized strong QTL location λ_2 is

$$\text{plod}(\lambda_2)=\max_{\lambda_1}\text{lod}(\lambda_1,\lambda_2). \quad (8)$$

This is standard practice [Barndorff-Nielsen and Cox, 1994, Section 3.4] for making inference about one parameter (in this case, the strong QTL location) in the presence of other parameters (in this case, the weak QTL location). The profile lod score function is maximized at the same point where the full two-dimensional lod score function is maximized.

The sliced lod score at a hypothesized weak QTL location λ_1 is

$$\text{slod}(\lambda_1)=\log_{10}\frac{L(\lambda_1,\hat{\lambda}_2)}{L(-\infty,\hat{\lambda}_2)}, \quad (9)$$

where $\hat{\lambda}_2$ is the strong QTL location at which the two-dimensional lod score is maximized, or equivalently the one-dimensional profile lod score for λ_2 is maximized. We call this the ''sliced'' lod score because it corresponds to the slice of the original two-dimensional lod score at $\hat{\lambda}_2$. The null hypothesis in (9) is that the weak QTL is unlinked to the marker loci but the strong QTL is fixed at the most likely position, which differs from the null hypothesis in (2). Unlike the use of the profile lod score, the use of the sliced lod score is not standard practice in the statistical literature. However, this is similar to the sequential approach in SOLAR [Almasy and Blangero, 1998] that estimates the weak QTL location after estimating the strong QTL location. Note that we estimate the strong QTL location using the two-QTL model with a polygenic component, whereas SOLAR estimates the strong QTL location using a one-QTL model with a polygenic component.

MONTE CARLO STANDARD ERRORS FOR MCMC LOD SCORES

Results of Monte Carlo calculations have Monte Carlo standard errors, which give the approximate size of the difference between the Monte Carlo (here MCMC) approximation of an object (here a lod score) and its exact value. Monte Carlo standard errors can, in theory, be made as small as one pleases by running the Monte Carlo calculation longer. These standard errors are different from ordinary standard errors, which can, in theory, be made as small as one pleases by collecting very large amounts of data.

We estimate Monte Carlo standard errors by the method of batch means [Fishman, 1978]. The same

method is used to estimate Monte Carlo standard errors in `lm_markers` [Thompson, 2005; Sieh et al., 2005]. When MCMC runs are too short, estimates of lod scores can be highly inaccurate, in which case estimates of their Monte Carlo standard errors are also inaccurate. When MCMC runs are sufficiently long, Monte Carlo standard errors estimated by the method of batch means do accurately reflect Monte Carlo variability Fishman, 1978; Schmeiser, 1982].

CODE TESTING

The program `lm_twoqtl` was written as a part of the MORGAN package to implement linkage analysis with a trait model having one or two QTLs plus a polygenic component. This involved adding new C code to pre-existing code. We tested the code for various trait models using a small data set—a 6-member nuclear pedigree with two markers. The program was run with MCMC sample size 10^6 so that MCMC variability is negligible. Resulting lod scores were then verified with exact lod scores, computed using FASTLINK [Cottingham et al., 1993; Schäffer et al., 1994] for the one-QTL model, and PAP [Hasstedt, 1982] for the one-QTL model with a polygenic component. For the two-QTL model with a polygenic component, because no program is available for the exact lod scores, a completely independent implementation in R [Ihaka and Gentleman, 1996], limited to this small pedigree size, was written for comparison.

SIMULATION AND ANALYSES OF DATA

We provide two examples using simulated data sets, both simulated with a model that includes two linked QTLs and a polygenic component, using the program `genedrop` in the MORGAN package. Example 1 shows the use of three different pedigree sizes on MCMC lod score estimation, and example 2 shows the use of four different QTL spacings. Three pedigree structures (Fig. 1) were used: a 6-member nuclear pedigree (ped6), a 16-member three-generation pedigree (ped16) and a 52-member five-generation pedigree (ped52). Only ped52 contains missing data. In data sets, we combine 300 replicates of ped6 (300ped6), or 100 replicates of ped16 (100ped16), or 40 replicates of ped52 (40ped52). Example 1 uses all three data sets, and example 2 uses 600ped6 and 100ped16. Table I shows the parameter values used for simulation in both examples. The parameter values for example 2 were simplified from the parameter values used for simulation of Q1 in GAW9 [MacCluer et al., 1995]. For both examples, the first QTL is weaker than the second QTL, measured by the variance of their respective contribution to phenotypes. The strong QTL is always at 55 cM. The weak QTL is always at 15 cM in example 1 but at 15, 25, 35, or 45 cM in

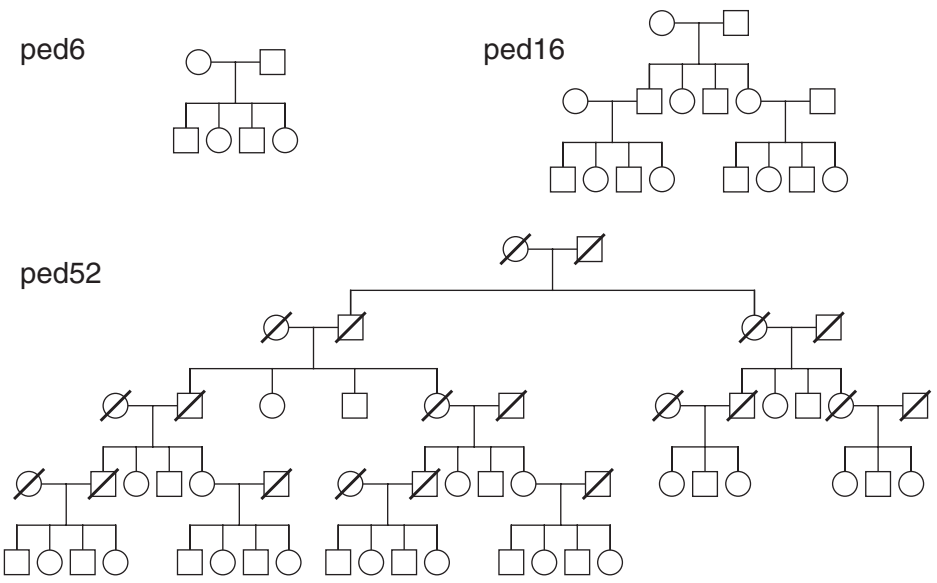

Fig. 1. Pedigrees for simulated data sets. No data are missing for ped6 and ped16, but data for 20 members are missing for ped52 (indicated by slashes).

TABLE I. Parameter values of trait models used for simulation

	Example 1					Example 2				
	p_A	μ_{AA}	$\mu_{A\alpha}$	μ_{aa}	σ^2 (σ^2/σ_y^2)	p_A	μ_{AA}	$\mu_{A\alpha}$	μ_{aa}	σ^2 (σ^2/σ_y^2)
QTL 1	0.10	−2	0	2	0.72 (0.11)	0.15	−3	0	3	2.29 (0.12)
QTL 2	0.30	−3	0	3	3.78 (0.58)	0.80	−3	0	6	4.38 (0.22)
Polygenic					1.00 (0.15)					1.00 (0.20)
Environment					1.00 (0.15)					9.00 (0.46)

TABLE II. Iterations and run times for analyses with *lm_twoqtl* for example 1

	1Q+P model			2Q+P model		
Data	MCMC	MC	Time	MCMC	MC	Time
ped6	2×10^4	Exact	0.01	1×10^5	Exact	0.07
ped16	1×10^5	5×10^2	1.84	1×10^5	1×10^3	1.71
ped52	2×10^5	5×10^5	73.72	2×10^5	1×10^4	196.56

MCMC: iterations for segregation indicators. MC: iteration for founder genotypes; exact summation is negligible for ped6. Time: run time (in minutes) per evaluation point on a 2.66 GHz Intel CPU. Sample sizes shown for computation of likelihoods under the alternative hypothesis. Sample sizes for computation of likelihoods under the null hypothesis were 10 times larger.

example 2. For both examples, eight markers are spaced 10 cM apart, from 0 to 70 cM, each with six alleles (frequencies 0.3, 0.2, 0.15, 0.16, 0.1, and 0.1).

The data sets in both examples were analyzed with four different models to mimic different models that might be used: two QTLs plus polygenic component (2Q+P) model, one-QTL plus polygenic component (1Q+P) model, one QTL without polygenic component (1Q) model, and VC model. For the 1Q+P model, the variance due to the weak QTL was added to the polygenic variance σ_a^2, and for the 1Q model, both the variance due to the weak QTL and the polygenic variance σ_a^2 were added to the environmental variance σ_e^2. We used `lm_twoqtl` for the 2Q+P and 1Q+P models, `lm_markers` [Thompson, 2005; Sieh et al., 2005] for the 1Q model, and SOLAR [Almasy and Blangero, 1998] for the VC model. For the VC model, multipoint IBD matrices were computed by MERLIN [Abecasis et al., 2002] for ped6 and ped16, and by Loki [Heath, 1997] for ped52. We used MCMC programs, even for the 1Q model, because exact computation is intractable for ped52 with eight multiallelic markers. Only the 2Q+P and VC models allow for the second QTL. For all parametric lod scores, MCMC estimates used every 10th iteration, Monte Carlo sample sizes are in Table II, and Monte Carlo standard errors were computed using 20 batches. Lod scores were evaluated every 5 cM (and every 2.5 cM) for the parametric models for example 1 (and for example 2) and every 1 cM for the VC model.

RESULTS

EXAMPLE 1: VARIOUS PEDIGREE SIZES

For the 2Q+P model, two-dimensional lod scores were evaluated: row 1 in Figure 2 shows their contour plots. All three show the maximum lod scores near the true QTL locations. Not surprisingly, the strong QTL locations were estimated better than the weak QTL locations: the former estimates were the true strong QTL location for all three, whereas the latter estimates were at most two evaluation points (10 cM) away from the true weak QTL location.

Row 2 in Figure 2 shows the lod scores for the strong QTL location using four different models. For the 2Q+P model, the profile lod scores defined by (8) were computed from the two-dimensional lod scores (Fig. 2, row 1). First, linkage analyses with all four models correctly provided a single peak near the true strong QTL location for all three data sets. This is expected because, in this example, the strong QTL's contribution to the phenotypic variance (58%) was large compared to that of the weak QTL (11%). Second, analyses with more complex models provided much higher lod scores than analyses with simpler models. Among the para-

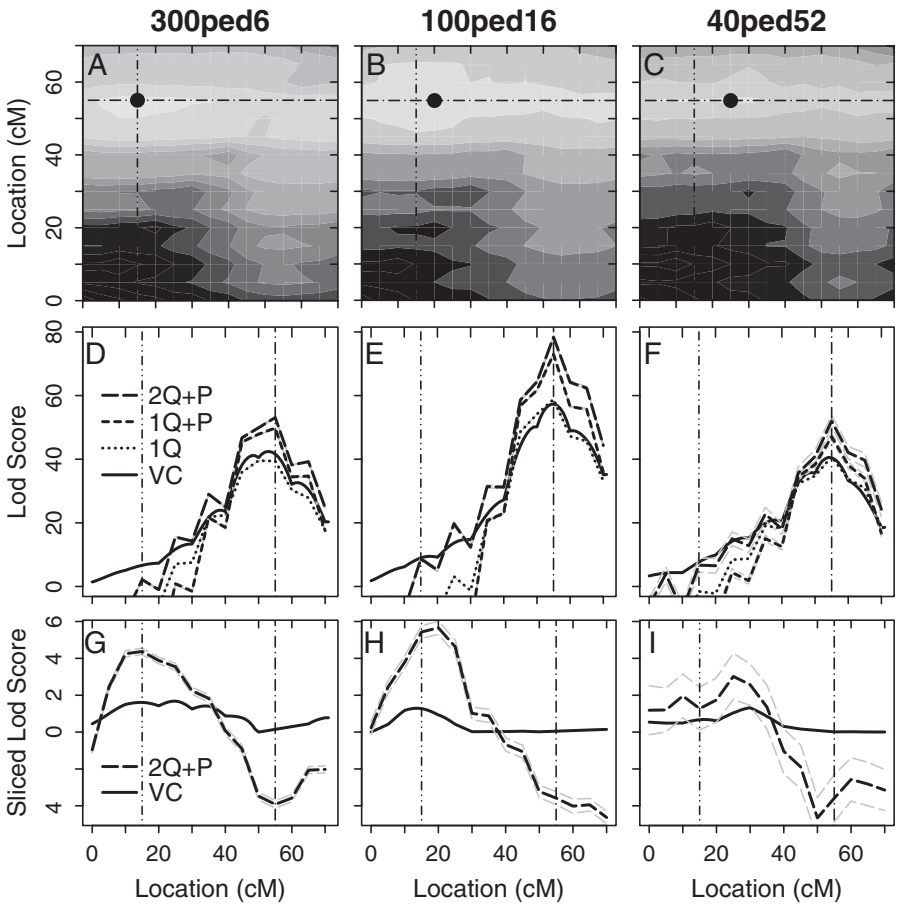

Fig. 2. Lod scores for example 1: various pedigree sizes. Row 1: two-dimensional lod scores from the 2Q+P model. Light shading indicates high and dark shading indicates low lod scores. Dots are at the maximum lod scores. Row 2: lod scores for the strong QTL location. Row 3: sliced lod scores for the weak QTL location. Vertical lines are true QTL locations (weak QTL at 15 cM, strong QTL at 55 cM). Gray lines show Monte Carlo variability (estimate ±2 MC standard error).

metric lod scores, the lod scores are the highest with the 2Q+P model, from which the data were simulated, next highest with the 1Q+P model and the lowest with the 1Q model. The lod scores from the VC model are a little higher than those from the 1Q model for 300ped6 and about the same for 100ped16 and 40ped52. The differences among these lod scores using different models were more noticeable from the 100ped16 data set than from the 300ped6 data set. Third, for all three data sets, the lod scores from the 1Q+P model were closer to those from the 2Q+P model, indicating that including a polygenic component appears to increase the lod score significantly over use of the 1Q model without a polygenic component. Fourth, lod scores from the 100ped16 data set were much higher than those from the 300ped6 data set, even though the latter data set contains more people. This is consistent with previous results suggesting that extended pedigrees contain more information than equivalent samples of smaller pedigrees [Wijsman and Amos, 1997].

Row 3 in Figure 2 shows the sliced lod scores for the weak QTL location using the 2Q+P model and the VC model. For all three data sets, the 2Q+P model gave higher sliced lod scores at the true weak QTL location, hence providing stronger evidence of linkage, than the VC model. The maximum sliced lod scores from the 2Q+P model were closer to the true weak QTL location (Table III) and gave narrower peaks, hence also providing better localization, than the VC model.

To investigate MCMC performance for the 2Q+P model, we ran 100 short runs and three long runs for the single ped16 and ped52 replicates with the largest Monte Carlo standard errors. MCMC sample sizes were 2×10^4 and 10^7 scans (short and long runs) for ped16, and 2×10^5 and 10^7 scans for ped52. Times for each run were 45 sec and 6 hr (short and long runs) for ped16, and 3.5

TABLE III. Highest lod score and associated location, cM, from 2Q+P and VC models in example 1

	2Q+P model		VC model	
Data	Strong QTL Lod (cM)	Weak QTL Lod (cM)	Strong QTL Lod (cM)	Weak QTL Lod (cM)
300ped6	53.10 (55)	4.37 (15)	42.44 (53)	1.68 (25)
100ped16	78.39 (55)	5.66 (20)	57.32 (55)	1.29 (14)
40ped52	52.05 (55)	3.02 (25)	40.58 (54)	1.33 (30)

The true strong QTL is at 55 cM and the true weak QTL is at 15 cM.

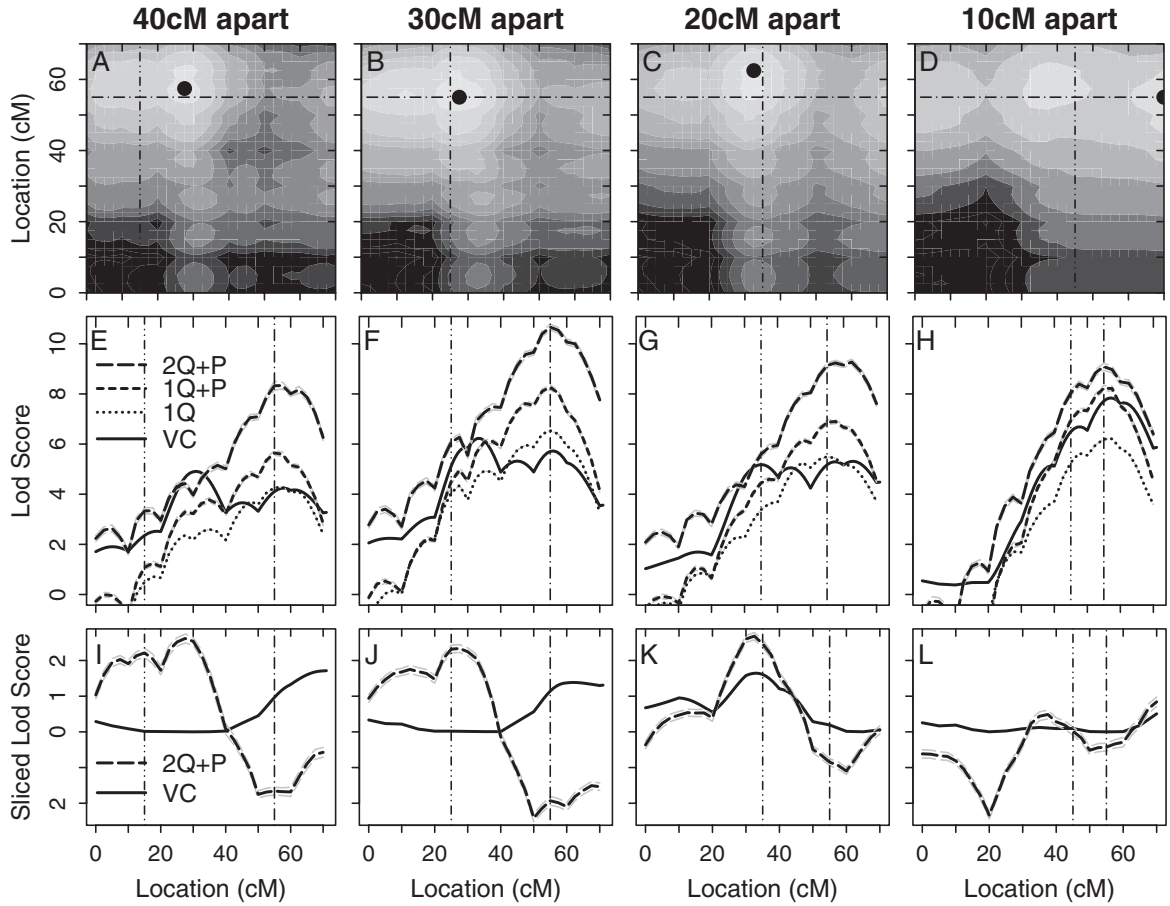

Fig. 3. Lod scores for example 2: various QTL spacings. Row 1: two-dimensional lod scores from the 2Q+P model. Row 2: lod scores for the strong QTL location. Row 3: sliced lod scores for the weak QTL location. Shadings and line types are as in Fig. 2.

and 40 hr for ped52 on a 1.8 GHz AMD Opteron. The estimates of the log likelihood at the true QTL location from these shorter runs tend to be lower than those from the longer runs: the mean of the estimates from the shorter runs is 0.007 lower for ped16 and 0.123 lower for ped52. For ped16, Monte Carlo standard error estimates (mean = 0.046) were accurate and reflected well the actual standard deviation (0.047) of 100 estimates. In contrast, for ped52, Monte Carlo standard error estimates (mean = 0.139) underestimated the actual standard deviation (0.253) of 100 estimates. Hence, to obtain accurate estimates of both lod scores and Monte Carlo standard errors, MCMC sample size 2×10^4 appears good enough for ped16, whereas MCMC sample size 2×10^5 is not long enough for ped52.

EXAMPLE 2: VARIOUS QTL SPACINGS

Three figures show lod scores for this example. For Figure 3 and Supplementary Figure 1, every-

thing is the same except for pedigree structure, 100ped16 and 600ped6, respectively. For Figure 3 and Supplementary Figure 2, everything is the same except for marker spacing, 10 and 5 cM, respectively. All three figures show that all parametric linkage analyses correctly provided a single peak near the true strong QTL location, even when the weak QTL was only 10 cM away. It is not surprising that the 2Q+P model did so, but it is surprising that the parametric models that take no account of the weak QTL (1Q+P and 1Q) also did so, because the strong QTL's contribution to the phenotypic variance (22%) was only moderately larger than that of the weak QTL (12%). All three figures also show that the 2Q+P model provided higher lod scores, which were sometimes considerably higher than those obtained from other models. The lod score difference between the 1Q+P model and the 1Q model was more noticeable for 100ped16 than for 600ped6. All three figures also show that the 2Q+P model provided higher sliced lod scores than the VC model, and the 2Q+P model correctly provided the highest sliced lod score near the true weak QTL.

For all three figures, the VC model provided several peaks for the strong QTL location, when the QTLs were at least 20 cM apart. For several data sets in 100ped16, the highest lod score was near the true weak QTL location and the highest sliced lod score was near the strong QTL location [Table IV]. This swapping would be expected especially if, as here, both QTLs' contributions were relatively similar. For 100ped16, when two QTLs are only 10 cM apart, both the 2Q+P and VC model had trouble estimating the weak QTL location correctly. For 600ped6, highest lod scores were correctly near the true strong QTL location. However, sliced lod scores from the VC model were lower than those from the 2Q+P model, indicating that the VC model had more trouble detecting the weak QTL for all data sets in 600ped6. Results from 100ped16 with markers 5 cM apart (Supplementary Figure 2) provided findings similar to those from 100ped16 with markers 10 cM apart, suggesting that marker density, per se, had little effect on the relative performance of analysis with the different models.

DISCUSSION

We have described a new MORGAN program, `lm_twoqtl`, for parametric linkage analysis with a quantitative trait model having one or two QTLs and a polygenic component. This program incorporates not only many existing ideas for MCMC sampling but also a novel approach for sampling founder genotypes. No previous program has been able to perform analysis with the complex models that `lm_twoqtl` can handle. The program has no restriction on number of markers or complexity of pedigrees. Our examples show possible advantages of analysis with these models, which can provide higher power for linkage detection and better localization than simpler models. Example 1 shown here illustrates the program's capability of handling large pedigrees with eight markers. Example 2 shows that for the data sets used, parametric analyses with two QTLs and polygenic component can accurately estimate both QTL locations even when they are 20 cM apart. In all data sets, the lod scores at the true strong QTL location are the highest with two QTLs and polygenic component, next highest with one QTL and polygenic component, and the lowest with one QTL only. This is similar to the findings for discrete traits that show analyses with two-trait-locus models can provide higher power for linkage detection than analyses with single-trait-locus models [Schork et al., 1993; Knapp et al., 1994; Strauch et al., 2003].

The program may detect multiple QTLs for complex traits when other methods fail. Of course, if the smaller model (e.g., 1Q) is correct, analysis with the smaller model provides higher power

TABLE IV. Highest lod score and associated location, cM, from 2Q+P and VC models in example 2

	2Q+P model		VC model	
Weak QTL location	Strong QTL Lod (cM)	Weak QTL Lod (cM)	Strong QTL Lod (cM)	Weak QTL Lod (cM)
15	8.33 (57.5)	3.82 (27.5)	4.91 (31)	1.71 (71)
25	10.66 (55.0)	2.34 (27.5)	6.23 (33)	1.38 (62)
35	9.27 (62.5)	2.41 (32.5)	5.31 (63)	1.64 (33)
45	9.09 (55.0)	0.84 (70.0)	7.84 (57)	0.51 (70)

The true strong QTL is at 55 cM.

than analysis with the larger model (e.g., 2Q). However, we envision that investigators would try linkage analysis with a two-QTL model after obtaining evidence for multiple loci, either from a segregation analysis or from results of a genome scan using a single-QTL model. Although a less accurate model may provide evidence of linkage, the resulting lod scores are generally smaller than with a more accurate model. Also, a single-QTL model is not guaranteed to detect multiple QTLs. Both our examples show that if one had only used a single-QTL model, then the weak QTL would have been missed.

In both examples, the lod scores from the VC analysis were lower than those from the parametric analysis with either one or two QTLs plus a polygenic component (Tables III and IV). Differences among analyses were more noticeable with extended pedigrees than with nuclear pedigrees. We observed this also in a recent extension of the model to include epistasis [Sung and Wijsman, 2007]; this may reflect the superior ability of more complex models to make use of the larger number of relationships in extended pedigrees. Also, as has been noted before [Malhotra et al., 2005], the VC analysis gave broader peaks (less accurate localization) than the parametric analyses. This confirms the general understanding that allele sharing methods, including VC, typically have lower power for linkage detection and provide less accurate localization than full model-based approaches with well-specified models [Greenberg et al., 1998; Abreu et al., 1999; Sham et al., 2000; Badzioch et al., 2005]. However, only two examples were explored here. More extensive simulations will be needed to compare the full model-based approach and the VC approach. With availability of a computationally practical program, future investigation of this problem is now possible.

The program `lm_twoqtl` simultaneously estimates the location of two linked QTLs. This differs from SOLAR's sequential approach: the first QTL location is estimated with the one-QTL model, then the second QTL location is estimated while fixing the first QTL location. This sequential approach can provide incorrect inference, especially when two trait loci are linked. Inaccurate estimates of the first QTL location can lead to poor estimates of the second QTL location. In our first example where the strong QTL's contribution to the phenotypic variance (58%) was large compared to that of the weak QTL (11%), localization of the strong QTL was accurate with all models even without taking account of the weak QTL. In contrast, in our second example where the strong QTL's contribution to the phenotypic variance (22%) was more moderate compared to that of the weak QTL (12%), localization of the strong QTL was poor without taking account of the weak QTL.

The MCMC sampler for `lm_twoqtl` is based on the MCMC sampler for `lm_markers` in the MORGAN package. Both simulate segregation indicators at marker loci conditional on marker data: the target distribution is the same as that of SimWalk [Lange and Sobel, 1991; Sobel and Lange, 1996]. Due to the complexity of trait models that include two QTLs and a polygenic component, `lm_twoqtl` uses additional sampling of segregation indicators and phased genotypes of founders, both at the hypothesized QTL locations. This framework provides potential for future extensions of linkage analysis with even more general trait models. It already has been extended to account for gene-gene interactions [Sung and Wijsman, 2007]. It could be extended to allow gene-environment interactions, discrete traits, or more than two QTLs.

ACKNOWLEDGMENTS

Y.J.S. thanks Charles J. Geyer for useful discussions. We thank two anonymous reviewers for advice on improving the manuscript.

WEB RESOURCES

The URLs for the programs used in this article are as follows: MORGAN, http://www.stat.washington.edu/thompson/Genepi/genepi.shtml (for *genedrop*, *lm_markers* and *lm_twoqtl*); MERLIN, http://www.sph.umich.edu/csg/abecasis/Merlin/index.html; SOLAR, http://www.sfbr.org/solar. Supplemental materials: http://www.interscience.wiley.com/jpages/0741-0395/suppmat.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Abreu PC, Greenberg DA, Hodge SE. 1999. Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. Am J Hum Genet 65: 847–857.

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–544.

Almasy L, Blangero J. 1998. Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211.

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. 2001. Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936–950.

Atwood LD, Heard-Costa NL. 2003. Limits of fine-mapping a quantitative trait. Genet Epidemiol 24:99–106.

Badzioch MD, Goode EL, Jarvik GP. 2005. The role of parametric linkage methods in complex trait analyses using microsatellites. BMC Genet 30(Suppl. 1):S48.

Barndorff-Nielsen OE, Cox DR. 1994. Inference and asymptotics. London: Chapman & Hall.

Biernacka JM, Sun L, Bull SB. 2005. Simultaneous localization of two linked disease susceptibility genes. Genet Epidemiol 28:33–47.

Biswas S, Papachristou C, Irwin ME, Lin S. 2003. Linkage analysis of the simulated data - evaluations and comparisons of methods. BMC Genet 4(Suppl. 1):S70.

Cader MZ, Steckley JL, Dyment DA, McLachlan RS, Ebers GC. 2005. A genome-wide screen and linkage mapping for a large pedigree with episodic ataxia. Neurology 65:156–158.

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

Elston RC, Stewart J. 1971. A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542.

Forrest WF, Feingold E. 2000. Composite statistics for QTL mapping with moderately discordant sibling pairs. Am J Hum Genet 66:1642–1660.

Fishman GS. 1978. Principles of Discrete Event Simulation. New York: Wiley.

Gagnon F, Jarvik GP, Badzioch MD, Motulsky AG, Brunzell JD, Wijsman EM. 2005. Genome scan for quantitative trait loci influencing HDL levels: evidence for multilocus inheritance in familial combined hyperlipidemia. Hum Genet 117:494–505.

Gentle JE. 2002. Random number generation and Monte Carlo methods, 2nd edition. Berlin: Springer.

George AW, Wijsman EM, Thompson EA. 2005. MCMC multilocus lod scores: application of a new approach. Hum Hered 59:98–108.

Glazier AM, Nadeau JH, Aitman TJ. 2002. Finding genes that underlie complex traits. Science 298:2345–2349.

Greenberg DA, Abreu P, Hodge SE. 1998. The power to detect linkage in complex disease by means of simple LOD-score analyses. Am J Hum Genet 63:870–879.

Gudbjartsson DF, Jonasson K, Frigge M, Kong A. 2000. Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12–13.

Hammersley JM, Handscomb DC. 1964. Monte Carlo Methods. Methuen and Co.

Hasstedt SJ. 1982. A mixed model likelihood approximation for large pedigrees. Comput Biomed Res 15:295–307.

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

Henderson CR. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.

Hwu WL, Yang CF, Fann CS, Chen CL, Tsai TF, Chien YH, Chiang SC, Chen CH, Hung SI, Wu JY, Chen YT. 2005. Mapping of psoriasis to 17q terminus. J Med Genet 42:152–158.

Igo RP Jr, Chapman NH, Berninger VW, Matsushita M, Brkanac Z, Rothstein JH, Holzman T, Nielsen K, Raskind WH, Wijsman EM. 2006. Genomewide scan for real-word reading subphenotypes of dyslexia: novel chromosome 13 locus and genetic complexity. Am J Med Genet B Neuropsychiatr Genet 141: 15–27.

Ihaka R, Gentleman R. 1996. R: language for data analysis and graphics. J Comput Graph Stat 5:299–314.

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, Reeve-Daly MP, Lander ES. 1996. Parametric and non-parametric linkage analysis: unified multipoint approach. Am J Hum Genet 58:1347–1363.

Lander ES, Green P. 1987. Construction of multilocus linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367.

Lange K, Sobel E. 1991. A random walk method for computing genetic location scores. Am J Hum Genet 49:1320–1334.

Lathrop GM, Lalouel JM. 1984. Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 36:460–465.

Lathrop GM, Ott J. 1990. Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet 47(Suppl.):A188.

Lin S. 2000. Monte Carlo methods for linkage analysis of two-locus disease models. Ann Hum Genet 64:519–532.

MacCluer JW, Blangero J, Dyer TD, Kammerer CM. 1995. Simulation of a common oligogenic disease with quantitative risk factors. GAW9 problem 2: the answers. Genet Epidemiol 12:707–712.

Malhotra A, Cromer K, Leppert MF, Hasstedt SJ. 2005. The power to detect genetic linkage for quantitative traits in the Utah CEPH pedigrees. J Hum Genet 50:69–75.

Markianos K, Daly MJ, Kruglyak L. 2001. Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet 68:963–977.

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

O'Connell JR. 2001. Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 51:226–240.

Orlacchio A, Kawarai T, Gaudiello F, St George-Hyslop PH, Floris R, Bernardi G. 2005. New locus for hereditary spastic paraplegia maps to chromosome 1p31.1-1p21.1. Ann Neurol 58:423–429.

Ott J. 1974. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597.

Quaas RL. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32: 949–953.

Schäffer AA, Gupta SK, Shiram K, Cottingham RW. 1994. Avoiding recompilation in linkage analysis. Hum Hered 44: 225–237.

Schmeiser B. 1982. Batch size effects in the analysis of simulation output. Oper Res 30:556–568.

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.

Sham PC, Lin MW, Zhao JH, Curtis D. 2000. Power comparison of parametric and nonparametric linkage tests in small pedigrees. Am J Hum Genet 66:1661–1668.

Sieh W, Basu S, Fu AQ, Rothstein JH, Scheet PA, Stewart WC, Sung YJ, Thompson EA, Wijsman EM. 2005. Comparison of marker types and map assumptions using Markov chain Monte Carlo-based linkage analysis of COGA data. BMC Genet 6(Suppl. 1):S11.

Sobel E, Lange K. 1996. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323–1337.

Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP. 2000. Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. Am J Hum Genet 66: 1945–1957.

Strauch K, Fimmers R, Baur MP, Wienker TF. 2003. How to model a complex trait. 2. Analysis with two disease loci. Hum Hered 56:200–211.

Sung YJ, Wijsman EM. 2007. Accounting for Epistasis in Linkage Analysis of General Pedigrees. Hum Hered, In press.

Thompson EA. 2000. Statistical inference from genetic data on pedigrees. NSF-CBMS regional conference series in probability and statistics volume 6.

Thompson EA. 2005. Chapter 4: MCMC in the analysis of genetic data on pedigrees. In: Liang F, Wang J-S, Kendall W, editors. Lecture Note Series of the IMS, National University of Singapore. Singapore: World Scientific Co Pvt. Ltd. p 183–216.

Thompson EA, Lin S, Olshen AB, Wijsman EM. 1993. Monte Carlo analysis on a large pedigree. Genet Epidemiol 10:677–682.

Thompson EA, Heath SC. 1999. Estimation of conditional multilocus gene identity among relatives. Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology, IMS Lecture Note—Monograph Series 33, p 95–113.

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