Published in final edited form as:

Genet Epidemiol. 2012 September ; 36(6): . doi:10.1002/gepi.21653.

Computationally efficient multipoint linkage analysis on extended pedigrees for trait models with two contributing major loci

Ming Su¹ and Elizabeth A. Thompson^{2,*}

¹Department of Electrical Engineering, University of Washington, Seattle, Washington

²Department of Statistics, University of Washington, Seattle, Washington

Abstract

We have developed a computationally efficient method for multipoint linkage analysis on extended pedigrees for trait models having a two-locus QTL effect. The method has been implemented in the program, *hg.lod*, which uses MCMC method to sample realizations of descent patterns conditional on marker data, then calculates the trait likelihood for each realization by efficient exact computation. Given its computational efficiency, *hg.lod* can handle data on large pedigrees with a lot of unobserved individuals, and can compute accurate estimates of lod scores at a much larger number of hypothesized locations than can any existing method. We have compared *hg.lod* to *lm.twoqtl*, the first publically available linkage program for trait models with two major loci, using simulated data. Results show that our method is orders of magnitude faster while the accuracy of QTL localization is retained. The efficiency of our method also facilitates analyses with multiple trait models, e.g. sensitivity analysis. Additionally, since the MCMC sampling conditions only on the marker data, there is no need to resample the descent patterns to compute likelihoods under alternative trait models. This achieves additional computational efficiency.

Keywords

lod score; quantitative trait locus; Markov Chain Monte Carlo; graphical models; exact inference

1 Introduction

In genetic linkage analysis, lod scores can be calculated for pedigrees with trait and marker data on observed members. The size of the pedigree becomes a choice in the study design of such analyses. In general, large and complex pedigrees are more informative than small and simple pedigrees in terms of power to detect linkage, accuracy of gene localization and accuracy of parameter estimates, especially for complex traits [Wijsman and Amos, 1997]. In addition, multipoint analysis has higher power than does single markers analysis [Wijsman and Amos, 1997]. For multipoint analysis on extended pedigrees with data on a large number of markers, lod scores are typically estimated by sampling-based methods, because exact methods, such as the Elston-Stewart algorithm [Elston and Stewart, 1971] and the Lander-Green algorithm [Lander and Green, 1987], become infeasible, due to their limitation on either the size of the pedigree or the number of markers that can be analyzed jointly. Lange and Sobel [1991] developed a Monte Carlo approach to the estimation of lod

*Corresponding author: Elizabeth Thompson, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, Phone: 206-685-0108, eathomp@u.washington.edu.

scores on extended pedigrees. In this approach, Markov Chain-Monte Carlo (MCMC) is used to sample realizations of multi-locus descent states conditional on marker data, then the trait likelihood contribution for each realization is computed by pedigree peeling. More recently, marker based realizations of descent patterns across a chromosome have been generated for subsequent use in trait analyses [Thompson, 2011]. In this paper, we show that how these realizations may be used not only for single-locus trait models, but also in a new computationally efficient method to estimate lod scores on extended pedigrees for trait models with two contributing major loci. Previously, only sampling-based approximate methods existed for trait likelihood computation in analyses with two loci [Sung and Wijsman, 2007; Sung et al., 2007b]. We propose an efficient way of computing exact trait likelihoods which increases both the computational efficiency and the accuracy of lod score estimation.

Multipoint linkage analysis for trait models with more than one major locus is important because it is generally hypothesized that complex traits are jointly affected by multiple genomic loci, potentially on the same chromosome. It has been shown that both parametric and nonparametric analyses with two-major-locus trait model can yield higher power for linkage detection and more accurate localization of trait loci than analyses with single-locus models, for both discrete [Knapp et al., 1994; Schork et al., 1993; Strauch et al., 2003] and quantitative traits [Sung et al., 2007a].

Several programs and methods have been developed to facilitate linkage analyses with two-locus model. However these approaches all have their own limitations. The program TMLINK [Lathrop and Ott, 1990] can perform parametric multipoint linkage analysis for only a small number of loci using a trait model with two unlinked loci. The program GENEHUNTER-TWOLOCUS [Strauch et al., 2000] can perform both parametric and non-parametric analyses using a trait model with two loci, either linked or unlinked. Biswas et al. [2003] used a Bayesian parametric approach to detecting linkage with possible locus heterogeneity. These approaches are limited either in the number of markers in the multipoint analysis or in the size of the pedigree. Biernacka et al. [2005] used IBD sharing in affected sib pairs to detect linkage of two linked trait loci. Lin and Schaid [2007] generalized this approach to affected sib pairs to affected relative pairs. These IBD sharing approaches often require a large number of affected relative pairs which are usually not available in genetic studies.

The MCMC-based MORGAN program *lm.twoqtl* was the first publically available program for lod scores with a parametric model including two Quantitative Trait Loci (QTL) and a polygenic component. This is a more complex trait model than can be analyzed with previously available programs [Sung et al., 2007b]. The program uses MCMC thus can analyze extended pedigrees with a large number of marker loci. [Sung et al., 2007b] showed that in an analysis using a model with known parameters, *lm.twoqtl* yielded higher power in linkage detection and more accurate localization than the Variance Components (VC) approach in SOLAR [Almasy and Blangero, 1998]. However, for large pedigrees with a lot of missing data, it becomes very computationally costly to achieve accurate estimates of lod scores using *lm.twoqtl*.

We propose a method that is more computationally efficient than *lm.twoqtl*. Rather than using the importance sampling approach in *lm.twoqtl*, we compute exact trait likelihoods using an efficient graphical model algorithm. In addition, the multiple-meiosis sampler [Tong and Thompson, 2008] is used in our method whereas a single-meiosis sampler was used in the early implementation of *lm.twoqtl* [Sung et al., 2007b]. For large pedigrees with a lot of missing data, the multiple-meiosis sampler can achieve MCMC standard errors

smaller than that of the single-meiosis sampler, using same amount of computation time [Tong and Thompson, 2008].

We have implemented the proposed method in a new lod score program *hg.lod*. Using previously simulated data [Sung et al., 2007b], we compare the performance of *hg.lod* and *lm.twoqtl*, for analyses using a two-major-locus additive model. Results show that *hg.lod* achieves the same level of power and accuracy in localizing QTL using much less computation time. We also evaluate the linkage information, especially that for localizing the weak QTL, that is conveyed by different amount of data by examining results on subsets of the whole dataset. Results shows that there is a fair amount of uncertainty in the inferred positions of the weak QTL, among subsets of data.

To demonstrate the effectiveness of *hg.lod* for SNP data, we use previously simulated marker data [Wijsman et al., 2006] to generate trait data using the trait model used by Sung and Wijsman [2007]. This model has no polygenic component but has epistatic effect to allow the interaction between two loci. In addition, the environmental effect takes a larger portion in this model than it does in model used by Sung and Wijsman [2007]. Results show that with *hg.lod*, both trait loci can be accurately localized.

The efficiency of our method also facilitates sensitivity analyses to be performed across a wide range of parameter values. As for the VC method, the marker-data-based approach decouples the computation of trait likelihoods from the sampling. This allows analyses with different trait models be performed without re-sampling the realizations of multi-locus descent states. Results shows that the localization of two trait loci are relatively insensitive to the allele frequencies. The power of the linkage detection drops continuously when parameter values are gradually deviated from the simulation truth.

2 Methods

2.1 Lod Score Methods for QTL Data

Traditionally there are two approaches to the computation of an exact lod score. The Elston-Stewart algorithm efficiently computes exact lod score on large pedigrees for only a small number of markers [Elston and Stewart, 1971], while the method of choice becomes the Lander-Green algorithm for small pedigrees but with a large number of markers [Lander and Green, 1987]. Both the Elston-Stewart and the Lander-Green algorithms involve marginalizing a joint probability using either genotypes or inheritance vectors [Lander and Green, 1987] as variables. In that sense, they both belong to a general class known as peeling algorithms in which a large nested summation is achieved by sequentially summing over smaller subsets of variables. Given a peeling sequence, i.e., the sequence of subsets in the sequential summation, the computational complexity is exponential in the size of the largest subsets. For multipoint analysis on extended pedigrees with data on a large number of markers, peeling algorithms become infeasible, and lod scores are typically estimated by MCMC methods [Lange and Sobel, 1991; Thompson, 2011].

Our analysis model assumes that the quantitative trait is affected by two major genes. The vector of trait values Y_T can be decomposed as follows:

$$Y_T = Q + E,$$

where Q is a two-locus QTL effect, and E is the residual effect. Q is modeled using discrete genotypic mean for each of the 9 two-locus genotypes. When there is no epistasis, $Q = Q_1 + Q_2$, where Q_1 and Q_2 are single-locus QTL effects, each specified by 3 genotypic

means. E is modeled as normally distributed with mean 0 and variance $I\sigma_e^2$. We assume the trait model parameters, i.e., the genotypic means and σ_e , are known. We assume that the marker data Y_M are observed without error and that the genetic map positions and the population allele frequencies of the marker loci are known.

For two-QTL model, the lod score at a pair of hypothesized map positions (λ_1, λ_2) is

$$\text{lod}(\lambda_1, \lambda_2) = \log_{10} \frac{P_{\lambda_1, \lambda_2}(Y_T, Y_M)}{P_{-\infty, \infty}(Y_T, Y_M)}$$

where the null hypothesis $(-\infty, \infty)$ denotes that the trait loci are not linked either to the marker loci or to each other. Since the trait data are independent of the marker data under the null hypothesis, we have

$$\frac{P_{\lambda_1, \lambda_2}(Y_T, Y_M)}{P_{-\infty, \infty}(Y_T, Y_M)} = \frac{P_{\lambda_1, \lambda_2}(Y_T | Y_M)}{P(Y_T)}.$$

The inheritance vector [Lander and Green, 1987] at locus j is defined as $S_j = \{S_{i,j}, i = 1, \dots, m\}$, where $S_{i,j}$ is a binary variable indicating in meiosis i at locus j , whether the DNA transmitted to the offspring is paternal. Let S be the set of inheritance vectors at all marker loci, i.e., $S = \{S_j, j = 1, \dots, l\}$. Using inheritance vectors as latent variables, we have

$$\begin{aligned} P_{\lambda_1, \lambda_2}(Y_T | Y_M) &= \sum_S P_{\lambda_1, \lambda_2}(Y_T, S | Y_M) \\ &= \sum_S P_{\lambda_1, \lambda_2}(Y_T | S) P(S | Y_M) \\ &= \sum_S P(Y_T | S_{\lambda_1}, S_{\lambda_2}) P(S | Y_M). \end{aligned} \quad (1)$$

The above derivation uses the fact that conditional on the inheritance vectors, the probability of the trait data is independent of the marker data and the trait data depend on the inheritance vectors only through the ones at the QTL. At the hypothesized position (λ_1, λ_2) , the computation in Eq. 1 can be estimated using MCMC methods by first sampling S conditional on Y_M , then averaging over N realizations using the inheritance vectors, i.e.,:

$$\tilde{P}_{\lambda_1, \lambda_2}(Y_T | Y_M) = \frac{1}{N} \sum_{i=1}^N P(Y_T | S_{\lambda_1}^i, S_{\lambda_2}^i). \quad (2)$$

2.2 FGL graph and *ibd* Graph

We first show that how $\Pr(Y_T | S_\lambda)$ can be evaluated at a single locus λ , under a single-locus trait model. Genes are said to be *identical by descent (ibd)* if they are copies of the same founder gene. Note that *ibd* is defined relative within a pedigree, so founder genes are not *ibd* by definition. At each locus, we can assign a unique identifier to each of the two haploid genomes of every founder. These identifiers are called *founder genome labels (FGL)*.

The descent of the founder genes to non-founders is determined by the inheritance vectors. Given inheritance vectors at one locus, the descent pattern can be illustrated by a descent graph [Sobel and Lange, 1996] which we will call a FGL graph [Thompson, 2011]. The nodes of an FGL graph are the set of unique founder genome labels. Each edge corresponds to a set of observed individuals, whose two genes at this locus are copies of the founder

genes represented by the endpoints of this edge. Figure 1(a) shows a nuclear family with all three offspring observed. For each individual, the two FGL at locus i , paternal then maternal, are also shown. The FGL graph at locus i is shown in Figure 1(b), where the FGL are numbers and edges are labeled with the observed individuals. Since the inheritance pattern changes along the chromosome, the FGL graph will change accordingly. Figure 1(c) shows the FGL graph at another locus j , potentially linked to locus i . At locus j , individual C receives FGL 2 and 3 instead of 1 and 4.

Suppose at a particular locus, the set of all FGL is K and the allelic type of FGL k is $A(k)$ with population allele frequency $q(A(k))$. Let E denote the set of edges and $M(\varepsilon)$ be the set of observed individuals on edge ε . Let $k_1(\varepsilon)$ and $k_2(\varepsilon)$ be the endpoints of edge ε . With the assumption that the founder genotypes are in Hardy-Weinberg equilibrium at each locus and linkage equilibrium among loci, the allelic types of distinct founder genomes are independent. So the probability of observed trait data conditional on the inheritance vector at locus λ , can be computed

$$\begin{aligned} \Pr(Y_T | S_\lambda) &= \Pr(Y_T | \text{FGL-graph}) \\ &= \sum_{\mathcal{A}} \left(\prod_{k \in K} q(\mathcal{A}(k)) \right) \left(\prod_{\varepsilon \in E} \left(\prod_{m \in M(\varepsilon)} \Pr(Y_m | \mathcal{A}(k_1(\varepsilon)), \mathcal{A}(k_2(\varepsilon))) \right) \right). \quad (3) \end{aligned}$$

The computation requires summation over $\mathcal{A}$ which is all possible assignments of allelic types to the FGL in the graph. It was first proposed by Thompson and Heath [1999] that this computation can be achieved by ‘peeling’ the allelic assignment of types to founder genes.

Notice that the labeling of founder genomes is arbitrary, any two FGL graphs having same topology with respect to the edges are equivalent in likelihood computation. We will refer to the node-unlabeled version as the *ibd* graph. Each distinct *ibd* graph corresponds to an equivalence class in the set of all FGL graphs. Examples given by Thompson [2011] show that across the set of MCMC realizations, the number of equivalence classes can be much smaller than the number of FGL graphs. This results in a significant reduction in computing $\Pr(Y_T | Y_M)$ because we only need to evaluate $\Pr(Y_T | S_\lambda)$ for each equivalence class.

Now we consider the computation of $\Pr(Y_T | S_{\lambda_1}, S_{\lambda_2})$ under the two-locus model. We will need to specify the inheritance patterns at two loci jointly using two-locus FGL graphs, where we have two sets of distinct founder genome labels, one for each locus. Since each observed individual can receive up to 4 FGL at two loci, individuals sharing same FGL are represented as hyperedges, which are sets of FGL, instead of regular edges. Thus two-locus FGL graphs are in fact hypergraphs. Figure 1(d) shows the two-locus FGL graph at loci (λ_1, λ_2) . FGL are suffixed to indicate their chromosomal positions and hyperedges are represented as shaded areas.

Similar to the single-locus case, there are equivalent two-locus FGL graphs, and the two-locus *ibd* graph can be similarly defined. In a two-locus *ibd* graph, nodes associated with different loci are distinguishable, while nodes associated with the same locus are not. Since there is no straightforward way to directly testing the equivalence of two-locus FGL graphs, we propose a way of constructing equivalence classes in the set of all two-locus FGL graphs by combining information from both loci. First we find the set of equivalence classes for each locus. We then take pairwise intersections between equivalence classes from different loci. We now use an example to demonstrate how this works. Suppose that we have 6 realizations of inheritance vectors, from which we construct two sets of single-locus FGL graphs, one for each locus, and one set of two-locus FGL graphs. Within each set, graphs are numbered from 1 to 6. Across all three sets, graphs with the same number are from the same

realization. Assume that the set of equivalence classes for the first locus is $\{\{1, 2, 3\}, \{4, 5, 6\}\}$, and for the second locus is $\{\{2, 4\}, \{1, 3\}, \{5, 6\}\}$. Taking pairwise intersections produces the set of equivalence classes for the two-locus graphs, i.e., $\{\{2\}, \{4\}, \{1, 3\}, \{5, 6\}\}$.

There is no existing peeling program for two-locus FGL graphs. Sung et al. [2007b] and Sung and Wijsman [2007] used importance sampling to approximate the likelihoods on two-locus FGL graphs. However, in this work, we aim to achieve a more accurate MCMC estimation by exact computation of the likelihood components, i.e., $P(Y_T|S_{\lambda 1}, S_{\lambda 2})$, using the Junction Tree algorithm that was developed for general graphical models.

2.3 Exact Computation on Factor Graphs

The hypergraph structure of two-locus FGL graphs fits well into a standard graphical model, the factor graph [Kschischang et al., 2001], for which many efficient inference algorithms have been developed over the years. Other popular graphical models, such as the Bayesian network [Jensen, 1997] and the Markov random field [Kindermann and Snell, 1980], can be equivalently transformed to the factor graph [Kschischang et al., 2001]. A factor graph $G = (X, F, E)$ is a bipartite graph consisting of random variable nodes $X = \{X_1, X_2, \dots, X_n\}$, factor nodes $F = \{f_1, f_2, \dots, f_m\}$, and edges E . Each factor node corresponds to a function $f_k(S_k)$, where $S_k \subseteq \{X_1, X_2, \dots, X_n\}$. A factor node f_k is connected to all variables in S_k . The factor nodes give a factorization of a joint function $g(X_1, X_2, \dots, X_n) = \prod_{k=1}^m f_k(S_k)$.

The two-locus FGL graph in Figure 1(d) can be converted to the factor graph [producttext] shown in Figure 2. The variable nodes correspond to the FGL, i.e., each variable node is a random variable whose sample space is all possible alleles at this locus. The factor nodes that are connected to only one variable node corresponds to the allele frequencies, i.e., $f_k = q(X_k)$, $k \in \{1.i, \dots, 4.j\}$. Each multi-connection factor node corresponds to a hyperedge, and its factor function is the probability of the trait data of the observed individuals on that hyperedge, conditional on the genotype, e.g., $f_A = \Pr(Y_A|X_{1.i}, X_{3.i}, X_{1.j}, X_{3.j})$. The joint function

$$g(X) = q(X_{1.i}) \cdot q(X_{2.i}) \cdot q(X_{3.i}) \cdot q(X_{4.i}) \cdot q(X_{1.j}) \cdot q(X_{2.j}) \cdot q(X_{3.j}) \cdot q(X_{4.j}) \cdot P(Y_A|X)$$

which is exactly in the form of the summand of Eq.3 and the summation over X is a standard graphical model computation, which can be carried out using the Junction Tree algorithm [Lauritzen and Spiegelhalter, 1988]. The Junction Tree algorithm is a generalization of the peeling algorithms and there exist advanced implementations which can be easily integrated into other software programs. However when there are major-gene QTL, the polygenic effect cannot be handled by peeling [Thompson, 2000, Section 6.6], so in comparing to previous work, in our analysis model we combine the variance of the polygenic component into that of the environmental component.

2.4 Implementation

The lod score method described above was implemented as a Java program called hg.lod. hg.lod makes calls to several standalone programs, as shown in the flow diagram in Figure 3. The input and output data for the programs are labeled on the transitional arrows. A brief description of the used programs is as follows:

- **gl.auto** is a MORGAN-3 program that implements the locus-meiosis sampler with multiple meiosis update [Tong and Thompson, 2008], to realize inheritance vectors conditional on marker data. **gl.auto** produces output in easy-to-parse FGL format in

a compact way such that the size of the output is proportional to the number of recombination events.

- **IBDgraph** [Koepke et al., 2011] is a C library that takes *gl.auto* output and produces the set of equivalence classes, i.e., distinct *ibd* graphs at a single locus. The library is available for download at <http://www.stat.washington.edu/thompson/Genepi/genepi.shtml>.
- **libDAI** is an open source C++ program with a collection of algorithms for making probabilistic inference in discrete graphical models [Mooij, 2010]. libDAI supports factor graph format and it is easy to integrate into one's own program. It is also organized in a way that one can easily experiment with novel algorithms by making modifications at various code levels.

We tested hg.lod by running several linkage problems from the MORGAN gold standard test suite. For one-locus models, hg.lod produces exactly same lod scores as those from the MORGAN-2.9 program *lm markers*, using the same MCMC realizations. For two-locus models, we test on examples with small pedigrees for which $\Pr(Y_T/S_i, S_j)$ can be computed exactly by brute force summation. For these small examples, the lod scores from hg.lod are exactly same as the lod scores produced by MORGAN program *lm twoqtl*.

2.5 Simulation and Analysis of Data

We have two simulated data sets, both using a 52-member 5-generation pedigree (ped52) that contains 20 unobserved individuals. In both examples, there are 40 replicates of data simulated on this pedigree. The pedigree is shown in Figure 4 with unobserved individuals marked by a slash. In both examples, the marker data Y_M and the QTL genotypes are simulated using the program *genedrop* in MORGAN. Then Q is generated from the simulated QTL genotypes. Example 1 were simulated using a model including two linked QTL with additive effect, a polygenic component and an environmental component. Example 2 includes two linked QTL in epistasis and an environmental component. There is no polygenic component in example 2. Simulation parameter values are shown Table 1. **Example 1** is same as the example 1 used by Sung et al. [2007b] for comparison purpose. Example 1 contains eight simulated microsatellite markers spaced 10 centi-Morgan (cM) apart, from 0 cM to 70 cM, each with six alleles (frequencies 0.3, 0.2, 0.15, 0.16, 0.1, and 0.1). The strong QTL (QTL 2) is at 55 cM and the weak QTL (QTL 1) is at 15 cM. Their parameters of the analysis model are shown in Table 1.

Sung et al. [2007b] computed lod scores both the markers and the midpoints between each pair of consecutive markers. To compute lod scores at the non-marker positions, we simulate inheritance vectors at these positions using the realized ones at markers. Let S_j be the inheritance vector at a non-marker locus j , i.e., the midpoint of two consecutive marker loci j and $j+1$. Note that inheritance vectors are independent among meioses and in the absence of genetic interference in the transmission of DNA from parents to offspring, the inheritance vectors form a Markov Chain along the chromosome, so that $S_{i,j}$ only depends on the inheritance vectors for the same i and at two flanking markers, i.e.,

$$\Pr(S_{i,j}|S) = \Pr(S_{i,j}|S_{i,j}, S_{i,j+1}) = \frac{\Pr(S_{i,j}, S_{i,j}, S_{i,j+1})}{\Pr(S_{i,j}, S_{i,j+1})}.$$

where S is the set of inheritance vectors at marker loci. Then we can use Eq. 2 where each of 1 and 2 can be either a marker position or a mid-point between two markers.

Example 2 is constructed to explore the effect of using SNP markers as opposed to the sparser microsatellite markers. It has 200 simulated SNP markers spaced 0.5 cM apart, from 0 to 99.5 cM. These simulated marker data were originally used by Wijsman et al. [2006]. The trait value was simulated using the trait model used by Sung and Wijsman [2007], which is epistatic. The strong QTL (QTL 2) is at 150th marker (74.5 cM) and the weak QTL (QTL 1) is at the 50th marker (24.5 cM). The parameters for the analysis model are shown in Table 1. Lod scores are computed at markers every 5 cM apart, starting from the 10th marker (4.5 cM) to the 200th marker (99.5 cM).

Sung et al. [2007b] used the realizations from every 10th MCMC scans to calculate $P(\tilde{Y}_T | Y_M)$ due to the high cost of computing $P(Y_T | S_{\lambda 1}, S_{\lambda 2})$. To compare with the work of Sung et al. [2007b], we use the realizations from every 10th MCMC scan for example 1. For example 1, we also exploited a different number of Monte Carlo sample sizes listed in Table 2. For example 2, we choose to use the realizations from every 30th MCMC scan out of a total of 3×10^4 scans. This, in practice, produces accurate results in a reasonable amount of time.

3 Results

3.1 Example 1

Lod scores for 40 replicates of data on the 52-member pedigree at an array of hypothesized locations are shown in the contour plot in Figure 5 (a). The highest lod score is 45.420 and it is achieved at (30 cM, 55 cM), shown as a dot in the figure. It is close to the true QTL locations, which is at the intersection of the dotted lines, (15 cM, 55 cM). The strong QTL location is estimated accurately, while the weak QTL position is 10 cM away from the true weak QTL location. However, the three hypothesized positions on the left of the highest lod score position exhibit high lod scores very close to the highest one. These are 45.418, 45.302 and 44.987 at 25 cM, 20 cM and 15 cM, respectively.

To evaluate the linkage information when only a smaller amount of data is available, we examine the lod scores from combining subsets of the 40 replicates. For each size from $\{10, 15, 20, 25\}$, we generate 1000 random subsets with that size. We denote the data having a subset of size n as $n\text{ped}52$, e.g., data containing 10 replicates will be $10\text{ped}52$. The highest lod score position for each subset is shown in Figure 6, for sizes 10, 15, 20 and 25, respectively. Highest lod score positions are shown as dots and many of them are coincident. To depict the multiplicities, we plot the histogram along each axis. For example, out of 1000 $10\text{ped}52$, the number of subsets whose weak QTL positions are at 0 cM, 15 cM and 30 cM, is 173, 103 and 282, respectively. Out of 1000 $10\text{ped}52$, there are 890 subsets having strong QTL position at the true position. As the size of the subsets increases, the distributions of both strong and weak QTL positions become more centralized at (30 cM, 55 cM), the highest lod score position obtained using 40 replicates.

We also examine the impact of different MCMC sample sizes. We compare the computation time, MCMC standard errors, and highest lod score and its position with the work of Sung et al. [2007b], which used lm.twoqtl with 2×10^5 MCMC scans. The results are shown in Table 2. For 2×10^5 scans, $hg.lod$ is about 28 times faster than lm.twoqtl with an MCMC standard error about two times smaller. It also can be seen that for both programs, the estimated MCMC standard error is roughly two times smaller than the actual MCMC standard error. With much smaller number of MCMC sample size, i.e., 2×10^4 , we obtain the MCMC standard error that roughly equals the one obtained with lm.twoqtl using 2×10^5 scans. For $hg.lod$, reducing the MCMC sample size has monotonic but negligible drops in the magnitude of the highest lod score, while the QTL are still accurately localized.

3.2 Example 2

The contour plot of two-dimensional lod scores is shown in Figure 5 (b) and other statistics are listed in Table 3. The highest lod score is 49.22 and it is attained at the true QTL position, which is (50, 150). Comparing to example 1, the highest lod position is significantly distinguished from its neighboring positions in the magnitude of the lod scores. Lod scores for positions to the left and right are about 43 and 40, respectively. Interestingly, there is another high peak at position (150, 50), which is diagonally-symmetric to the true QTL position. The lod score at this position is about 39. Since the inheritance vectors for all 200 SNP markers are sampled, the sampling time is much longer than that of the run in example 1 with similar MCMC sample size. In Table 3, we report the estimated MCMC standard error for the smallest and the largest among all replicates. The average MCMC standard error is about two times smaller than that of the run in example 1 with similar MCMC sample size. For the replicate with largest estimated MCMC standard error, the actual standard error is about three times larger.

3.3 Sensitivity Analysis

The trait model in the previous analyses has the same parameter values used to generate the trait data, except that the polygenic variance is combined with the environmental variance. We are also interested in the variability of lod scores and highest lod score positions when the parameter values of the trait model differ from the simulation truth. Given the high computational efficiency of our method, we can re-evaluate lod score with many different parameter values. Further computational reduction is achieved with the elimination of the need to re-sample the inheritance vectors, given that our MCMC method samples the inheritance vectors conditional only on the marker data. We can reuse the MCMC realizations from previous computations and only recompute the likelihoods on the FGL graphs using different trait model parameter values. In this paper, we only perform single-parameter sensitivity analyses by varying the parameter of interest one at a time while keeping other parameters at their true values.

The results of sensitivity analyses are shown in Figure 7. The parameters under analysis are the allele frequencies of QTL 1 (row 1) and QTL 2 (row 2), and the environmental variance (row 3). The allele frequencies of QTL 1 and of QTL 2 ranges from 0.1 to 1 with stepsize 0.01. The environmental variance ranges from 0 to 20 with stepsize 0.5. In the first column of Figure 7 are shown the highest lod scores (crosses) and the QTL 1 positions (solid points) where the highest lod score is attained, versus the parameter values. The second column shows the highest lod scores (crosses) and the QTL 2 positions (solid points) where the highest lod score is attained. The true parameter values and the true QTL positions are indicated by dotted lines.

The sensitivity of the estimation of the QTL location can be expressed by a pair of percentages by normalizing the parameter range over which the estimation is close to the simulation truth by the allowable range of the parameter. Suppose the parameter range over which the estimated location is close is $[A, B]$, the simulation truth is S , and the allowable range of the parameter is $[L, H]$, then the pair of percentages will be $[(A-S)/(S-L), (B-S)/(H-S)]$. Large ranges indicate insensitivity. When the varying parameter is QTL 1 allele frequency, the range for QTL 1 and QTL 2 are $[-56\%, 67\%]$ and $[-61\%, 100\%]$, respectively. When the varying parameter is QTL 2 allele frequency, the ranges for QTL 1 and QTL 2 are $[-100\%, 100\%]$ and $[-100\%, 100\%]$, respectively. When the varying parameter is the environmental variance, the ranges for QTL 1 and QTL 2 are $[-100\%, 11\%]$ and $[-100\%, 100\%]$, respectively. It can be seen that the estimated positions of both QTL are quite insensitive to the allele frequencies.

There are two interesting observations in the sensitivity analysis. The first one is that outside the insensitive region, the estimated position of one QTL coincides the estimated position of the other QTL in the original analysis. This can be seen for QTL 1 when the varying parameter is QTL 1 allele frequency and the value is less than 0.35, for QTL 2 when varying parameter is QTL 1 allele frequency and its value is less than 0.35, and for QTL 1 when the varying parameter is the environmental variance and its value is greater than 4. The second observation is that for certain parameter values, the estimated position of QTL 1 is closer to the simulation truth than the one in the original analysis. This happens when the allele frequency of QTL 1 is between 0.89 and 0.4, when the allele frequency of QTL 2 is outside of [0.7, 0.74], and when the environmental variance is less than 2.

Even though both QTL 1 and QTL 2 positions are insensitive in a reasonably large region around the true parameter value, the lod scores continuously change with the parameter values. For all three parameters, the lod score remains significantly high within the entire parameter range and peaks at the value that is very close to the simulation truth.

We have only considered single-parameter sensitivity, although in reality there is uncertainty about the true values of all parameters. We also did not show sensitivity results for the genotypic means, since these are jointly constrained. While our approach permits the computation of lod scores under multiple models, multi-parameter sensitivity analysis becomes computationally intensive, and the results hard to display. For specific joint parameter changes of interest, defining a line in the parameter space, the same analysis as we present for single parameters could be undertaken. Rather than to demonstrate any particular robustness, our goal in this paper is to show how our approach enables the computation of lod scores for multiple complex trait models without reconsideration of the marker data. This in turn enables any sensitivity analysis that might be needed in a real-data analysis.

In an analysis of real trait data, it is usual to start with a simple model for linkage detection, and then refine the joint two-locus model based on the location information obtained from the linkage detection. In some cases, the second QTL will only become apparent when the first is already taken into account. Linkage detection is far more robust to trait model misspecification than is the localization of the genes [Clerget-Darpoux et al., 1986].

4 Discussion

We have developed a new program, *hg.lod*, for linkage analysis of a two-locus QTL model. The program can handle general trait models, including two QTL effects and an environmental effect. *hg.lod* uses similar ideas in the MORGAN program, *lm.twoqtl*, which is the first available software for two-locus linkage analysis for large pedigree with a lot of missing data. *hg.lod* dramatically reduces the computation time which is a main issue in *lm.twoqtl*. As pointed out by Sung et al. [2007b], the capability of simultaneous detecting of two QTL is critical in accurate localization, especially when two QTL are linked and their contributions to the overall phenotypic variance are comparable. Also discussed by Sung et al. [2007b], a full model-based approach is expected to perform better than allele sharing methods and VC methods. These facts give rise to our program and whenever there are a large amount of data on extended pedigrees, our program provides a efficient tool for linkage analysis.

The lod score plots in Figure 5 both exhibit some diagonal symmetry which can be understood as due to the symmetry of the roles of two QTLs in the trait model. In both examples, the allele frequencies of both QTLs do not differ much (0.9 vs. 0.7 in example 1 and 0.87 vs. 0.79 in example 2). In example 1, the genotypic means of both QTLs have very

similar values. In example 2, the genotypic means for two genotypes of which the roles of two QTLs are swapped have similar values (7 for Aabb and 6 for aaBb). In Figure 5(b), there is a second peak at the mirror point to the highest lod score location, where the locations of the two QTLs are swapped. Note that there is a second peak in Figure 5(b), which is missing in Figure 5(a). This is explained by the SNP data providing more information on the latent inheritance, so that the symmetry in the trait model is more clearly reflected in the lod scores.

In our analysis model we combine the variance of the polygenic component into that of the environmental component, since the polygenic effect cannot be peeled in the presence of two major-gene QTL. In the example of this paper, the penalty associated with this simplification causes a 12% decrease in the highest lod score, compared with *lm.twoqtl*. There is also uncertainty induced in the neighborhood where the highest lod scores are attained. Along the axis for QTL 1, the weak QTL, we obtain a flat region rather than a single peak. However the true QTL 1 position is almost always included in this flat region. When there is no polygenic component as in example 2, we achieve a sharp peak at the true QTL position even though the distribution of the variance between two QTL still follows a strong-weak pattern. In case of dealing with a model with a large polygenic component, one might first obtain a quick identification of QTL positions using *hg.lod*, then perform a fine localization in the region of high lod score using *lm.twoqtl* and taking into account the polygenic effect.

Results in Figure 6 show the linkage information carried by different subsets of data. Let us regard the size of data, i.e., number of pedigree replicates, as a variable. The uncertainty of linkage information associated with each value of this variable is visualized as two histograms in Figure 6, one for each QTL. Not surprisingly, there is more uncertainty in the localization of QTL 1 than that of QTL 2. As the data size increases, the histogram for QTL 2 quickly converges to a single peak. Only about half of the data is sufficient to localize the strong QTL. The estimated position for the weak QTL converges slower and it ends up with two peaks, neither being the true QTL 2 position.

A major advantage of *hg.lod* is that it enables lod scores to be computed efficiently for multiple trait models with no additional MCMC sampling. We have performed single-parameter sensitivity analysis for example 1. The analyzed parameters are the allele frequencies and the variance of the environmental effect. The localizations of both QTL are insensitive to the misspecification of the allele frequency for the strong QTL. The localizations of both QTL are still insensitive, yet to a lesser degree, to the allele frequency of the weak QTL. The specification of the value of environmental variance is important in accurate localization of the weak QTL. The sensitivity analysis also demonstrates the effectiveness of the MOD-score method [Risch, 1984] for the particular trait model of example 1. The lod score is always maximized at the true parameter value for the allele frequency of QTL 1 and the variance of the environmental effect. The lod score is maximized at a value that is very close to the truth for the allele frequency of QTL 2.

In any approach there is a balance between power and robustness. In contrast to trait-model-based lod score analyses, variance component methods do not lose power due to misspecified trait-model parameters. On the other hand, variance component approaches are more sensitive to lack of normality of trait values [Blangero, 1995], and hence are particularly sensitive to the use of selected samples [Cuenco et al., 2003]. By contrast, lod scores on a given data set under a specified trait model do not depend on the ascertainment. Although in our simulated data, there was no trait-based ascertainment, in real-data studies extended pedigrees may often be ascertained through individuals with extreme trait values. Kleinsang et al. [2010] have provided extensive comparisons of power and robustness of a

wide range of quantitative trait mapping approaches, including lod score and variance component methods, under a variety of scenarios of ascertainment and deviations from normality.

Table 2 shows that when the MCMC sample size increases by a factor of 100, the MCMC standard error only decreases by a factor of about 3, suggesting that there is much uncertainty in the estimation of *ibd* graphs using microsatellite markers. By contrast, Table 3 shows very small MC error using the SNP data. Once the marker data are dense enough to determine the relevant inheritance information lod score will not change with the addition of denser SNPs [Wijsman et al., 2006]. In fact, using dense SNP data will not only fail to increase the signal strength, but may introduce noise due to linkage disequilibrium (LD) and MCMC mixing problems [Sieh et al., 2005; Wilcox et al., 2005]. Although the optimal SNP density will depend on the specific dataset, on extended multi-generation pedigrees which provide substantial inheritance information typically 1 SNP per cM is sufficient to capture this information [Wijsman et al., 2006]. In our example, the SNPs are 0.5 cM apart, so that LD is not an issue, and the multiple-meiosis sampler in *lm.twoqt1* still works well to achieve good MCMC mixing and very small MC standard error.

Acknowledgments

The authors would like to thank Ellen M. Wijsman for useful discussions and Yun Ju Sung for providing comparison data. This work is supported in part by NIH grants HG004175 and GM046255.

References

- Almasy L, Blangero J. Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics*. 1998; 62:1198–1211. [PubMed: 9545414]
- Biernacka JM, Sun L, Bull SB. Simultaneous localization of two linked disease susceptibility genes. *Genetic Epidemiology*. 2005; 28:33–47. [PubMed: 15481103]
- Biswas S, Papachristou C, Irwin ME, Lin S. Linkage analysis of the simulated data - evaluations and comparisons of methods. *BMC Genetics*. 2003; 4(Suppl. 1):S70. [PubMed: 14975138]
- Blangero J. Genetic analysis of a common oligogenic trait with quantitative correlates: Summary of GAW9 results. *Genetic Epidemiology*. 1995; 12:689–706. [PubMed: 8787995]
- Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J. Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*. 1986; 42:393–399. [PubMed: 3741977]
- Cuenca KT, Szatkiewicz JP, Feingold E. Recent advances in human quantitative-trait-locus mapping: Comparison of methods for selected sibling pairs. *American Journal of Human Genetics*. 2003; 73:863–873. [PubMed: 12970847]
- Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Human Heredity*. 1971; 21(8):523–542. [PubMed: 5149961]
- Jensen, FV. Introduction to Bayesian networks. 1st Edition. Springer; 1997.
- Kindermann, R.; Snell, JL. Markov Random Fields and Their Applications. American Mathematical Society; 1980.
- Kleensang A, Franke D, Alcäis A, Abel L, Müller-Myhsok B, Ziegler A. An extensive comparison of quantitative trait loci mapping methods. *Human Heredity*. 2010; 69:202–211. [PubMed: 20203525]
- Knapp M, Seuchter SA, Baur MP. Two-locus disease models with two marker loci: the power of affected-sib-pair tests. *American Journal of Human Genetics*. 1994; 55:1030–1041. [PubMed: 7977340]
- Koepke H, Koepke L, Thompson EA. Efficient testing operations on dynamic structures using strong hash functions. *Journal of Computational and Graphical Statistics*. 2011 Under review.
- Kschischang FR, Frey BJ, Loeliger H. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*. 2001; 47(2):498–519.

- Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*. 1987; 84(8):2363–7. [PubMed: 3470801]
- Lange K, Sobel E. A random walk method for computing genetic location scores. *American Journal of Human Genetics*. 1991; 49:1320–1334. [PubMed: 1746559]
- Lathrop GM, Ott J. Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the linkage programs. *American Journal of Human Genetics*. 1990; 47(Suppl.):A188.
- Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1988; 50(2):157–224.
- Lin WY, Schaid DJ. Robust multipoint simultaneous identical-by-descent mapping for two linked loci. *Human Heredity*. 2007; 63:35–46. [PubMed: 17215580]
- Mooij JM. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*. 2010; 11:2169–2173. <http://cs.ru.nl/jorism/libDAI/>.
- Risch N. Segregation analysis incorporating linkage markers. I. single-locus models with an application to type I diabetes. *American Journal of Human Genetics*. 1984; 36(2):363–386. [PubMed: 6608876]
- Schork NJ, Boehnke M, Terwilliger JD, Ott J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *American Journal of Human Genetics*. 1993; 53:1127–1136. [PubMed: 8213836]
- Sieh W, Basu S, Fu AQ, Rothstein JH, Scheet PA, Stewart WCL, Sung YJ, Thompson EA, Wijsman EM. Comparison of marker types and map assumptions using Markov chain Monte Carlo-based linkage analysis of COGA data. *BMC Genetics*. 2005; 6(Suppl 1):S11. [PubMed: 16451566]
- Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*. 1996; 58(6):1323–37. [PubMed: 8651310]
- Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP. Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *American Journal of Human Genetics*. 2000; 66:1945–1957. [PubMed: 10796874]
- Strauch K, Fimmers R, Baur MP, Wienker TF. How to model a complex trait. 2. analysis with two disease loci. *Human Heredity*. 2003; 56:200–211. [PubMed: 15031621]
- Sung YJ, Wijsman EM. Accounting for epistasis in linkage analysis of general pedigrees. *Human Heredity*. 2007; 63(2):144–52. [PubMed: 17283443]
- Sung YJ, Di Y, Fu AQ, Rothstein JH, Sieh W, Tong L, Thompson EA, Wijsman EM. Comparison of multipoint linkage analyses for quantitative traits in the ceph data: parametric lod scores, variance components lod scores, and bayes factors. *BMC Proceedings*. 2007a; 1(Suppl 1):S93. [PubMed: 18466597]
- Sung YJ, Thompson EA, Wijsman EM. MCMC-based linkage analysis for complex traits on general pedigrees: multipoint analysis with a two-locus model and a polygenic component. *Genetic Epidemiology*. 2007b; 31:103–114. [PubMed: 17123301]
- Thompson EA. Statistical inference from genetic data on pedigrees. *NSF-CBMS regional conference series in probability and statistics*. 2000:6.
- Thompson EA. The structure of genetic linkage data: from LIPED to 1M SNPs. *Human Heredity*. 2011; 71:88–98.
- Thompson EA, Heath SC. 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*. 1999; 33:95–113.
- Tong L, Thompson EA. Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. *Human Heredity*. 2008; 65:142–153. [PubMed: 17934317]
- Wijsman EM, Amos CI. Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of gaw10 contributions. *Genetic Epidemiology*. 1997; 14(6):719–35. [PubMed: 9433569]

- Wijdsman EM, Rothstein JH, Thompson EA. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics*. 2006; 79:846–858. [PubMed: 17033961]
- Wilcox MA, Pugh EW, Zhang H, Zhong X, Levinson DF, Kennedy GC, Wijdsman EM. Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation groups 1, 2, and 3. *Genetic Epidemiology*. 2005; 29(Suppl 1):S7–S28. [PubMed: 16342186]

(a)

locus i

(b)

locus j

(c)

(d)

Figure 1.
Factor graph representation of the two-locus FGL graph

Figure 2.
Factor graph representation of the two-locus FGL graph

Figure 3.
Flowchart of *hg.lod*

Figure 4.
: Pedigree used in Example 1 and 2

Figure 5.

Contour plot of the lod scores for (a) example 1, and (b) example 2. The horizontal axis is the QTL 1 position, and the vertical axis is for QTL 2. Light color indicates high lod score, while positions of highest lod scores are marked by bold points

Figure 6. Distribution of highest lod score positions for 1000 (a) 10ped52, (b) 15ped52, (c) 20ped52 and (d) 25ped52

Figure 7. Sensitivity results for the allele frequency of QTL 1 (first row), QTL 2 (second row) and the environmental variance (third row) as a function of the parameter values. The true parameter values and the true QTL positions are indicated by dotted lines

Table 1

Trait model parameter values.

Example	Geno Means ²				Variance Components (%)					
	AA	Aa	aa	Allele Freq. ³	QTL1	QTL2	epi.	poly.	env.	
1	QTL1	-2	0	2	0.9	0.7	0.72 (11)	3.78 (58)	0(0)	1 (15)
	QTL2	-3	0	3						
2	BB	3	0	-3	0.87	0.79	2.42 (7)	10.6 (31)	6.1 (18)	0 (0)
	Bb	1	3.5	6						15 (44)
	bb	-1	7	15						

¹ Values for example 1 and 2 are same as those used by Sung et al. [2007b] and Sung and Wijsman [2007], respectively.

² For example 1, the alleles for both QTL are coded A and a. For example 2, the alleles for QTL1 are coded A and a, while the alleles for QTL2 are coded B and b.

³ Allele frequency are for the small alleles, i.e., a or b.

Table 2Comparison of *hg.lod* and *lm.twoqtl* with different MCMC sample sizes.

MCMC	Time (min.) ¹				Highest		MCMC		Positions
Sample	Samp.		Trait Lik.		Lod Score		Standard Error ²		
Size	lm	hg	lm	hg	lm	hg	lm	hg	
2 × 10 ³	-	0.12	-	0.026	-	44.94	-	0.120/0.402	{{20,25,30},55)}
2 × 10 ⁴	-	1.3	-	0.26	-	45.42	-	0.109/0.228	{{15,20,25,30},55)}
1 × 10 ⁵	-	6.3	-	1.28	-	45.79	-	0.076/0.160	{{15,20,25,30},55)}
2 × 10 ⁵	4.7	11.0	249	2.54	52.03	45.95	0.139/0.253	0.058/0.134	{{15,20,25,30},55)}

¹ Overall sampling time (*gl.auto*) and the average trait likelihood computation time for a single hypothesized position are listed.

² Batch mean estimate and actual MC standard error are listed, separately by back slashes. The MC standard errors shown are the largest among all replicates. The actual MC standard error is obtained by running 100 independent analyses for the replicate with the largest MC standard error.

Table 3

Results for example 2.

Sample Size	Time (min.)		Highest Lod Score	MC Standard Error				Position
	Samp.	Trait Lik.		Min.	Mean	Max.	Actual ^I	
3×10^4	25	0.19	49.22	0.004	0.018	0.054	0.18	(50,150)

^IThe actual MCMC standard error is obtained by running 100 independent analyses for the replicate with the largest MCMC standard error.