Published in final edited form as:

Genet Epidemiol. 2009 November ; 33(7): 604–616. doi:10.1002/gepi.20412.

Bayesian Intervals for Linkage Locations

Ritwik Sinha^{1,2}, Robert P. Igo Jr.¹, Shiv K. Saini³, Robert C. Elston¹, and Yuqun Luo¹

¹ Case Western Reserve University, Cleveland, Ohio

² Bristol-Myers Squibb Company, Wallingford, Connecticut

³ University of Wisconsin, Madison, Wisconsin

Abstract

Intermediate fine mapping has received considerable attention recently, with the goal of providing statistically precise and valid chromosomal regions for fine mapping following initial identification of broad regions that are linked to a disease. The following classes of methods have been proposed and compared in the literature: (1) LOD-support intervals, (2) Generalized estimating equations, (3) Bootstrap, (4) Confidence set inference framework. These methods provide confidence intervals either with coverage levels deviating from the nominal confidence levels or that are not fully efficient. Here, we propose a novel Bayesian method for constructing such intervals using affected sibling pair data. The susceptibility gene location is treated as a parameter in this method, with a uniform prior. A Metropolis-Hastings algorithm is implemented to sample from the posterior distribution and Highest Posterior Density Intervals of the disease gene locations are constructed. Correct coverage levels are maintained by our method. Both simulation studies and an application to a Rheumatoid Arthritis dataset demonstrate the improved efficiency of the Bayesian intervals compared with existing methods.

Keywords

BILL; disease gene localization; intermediate fine mapping; model-free linkage

1 Introduction

Confidence sets for the location of a disease gene constructed after initially identifying a linkage region can help narrow the region in which fine mapping studies are conducted. Along with reducing the financial cost of follow-up studies, such confidence sets can also alleviate the multiple testing associated with fine mapping studies. Until recently, LOD-support intervals, where a 95% confidence interval (CI) is taken to be the chromosomal region where the LOD score has dropped less than 1 from the linkage peak, had been the most widely applied. Note that this CI construction approach can be applied to different types of LOD scores, including the Kong and Cox (KAC) LOD score [Kong and Cox, 1997] and the maximum LOD score (MLS) [Risch, 1990]. However, the true coverage level of 1-LOD-support intervals can vary wildly under different scenarios [Papachristou and Lin, 2006a]. Thus, several methods have been proposed recently for affected sib pairs (ASPs). These include bootstrap methods [Papachristou and Lin, 2006a], generalized estimating equations (GEE) [Liang et al., 2001], and the confidence set inference (CSI) framework [Papachristou and Lin, 2006c; Sinha and Luo, 2007b].

All these methods construct CIs for the unknown yet unique location of the trait locus on a broad chromosomal region that has been shown by linkage studies to potentially contain a trait locus. Before we provide a brief review of the above approaches, let us first consider the difficulty of constructing a CI based on the maximum likelihood estimate (MLE) of the disease gene location. Let l be the location of the disease gene and let (z_0, z_1, z_2) be the probabilities of sharing 0, 1, or 2 alleles identical-by-descent (IBD) at the disease locus by an ASP ($z_0 + z_1 + z_2 = 1$). Maximization of the likelihood, under the triangle constraints ($z_1 \leq 1/2$ and $2z_0 < z_1$) of Holmans [1993], provides a consistent estimator of the location of the putative gene [Risch, 1990]. In theory, this MLE, together with an estimate of its standard error, may be used to provide asymptotic CIs for the parameters. Several issues complicate this approach: (1) it is difficult to obtain an estimate of the standard error of the MLE; (2) asymptotic normality might not hold when the maximization of the likelihood is conducted under the triangle constraints [Papachristou and Lin, 2006a]; (3) asymptotic normality will be violated if the disease gene is located near or at the end of a chromosome.

As a common tool to approximate the complex sampling distributions of estimators, two bootstrap approaches have been proposed by Papachristou and Lin [2006a] to construct a CI for l : the non-parametric bootstrap (NPB) and the parametric bootstrap (PMB). A bootstrap sample either consists of ASP units sampled with replacement from the observed sample of ASPs (under NPB), or consists of ASPs simulated using the MLEs of (l, z_0, z_1, z_2) obtained from the original ASP sample. The totality of MLEs from each of the bootstrap samples constitutes an approximation to the sampling distribution of the MLE from the original sample and confidence intervals can be obtained from this sampling distribution.

On the other hand, Liang et al. [2001] explored formal modeling of the genetic effect and the gene location. Let $\mu(l)$ be the expected number of alleles shared IBD at the disease gene location l by a sib pair given that both sibs are affected. Suppose there are k markers. Let M_i be the map location of marker i and let $\theta_i(l)$ be the recombination fraction between this marker and l . The GEE approach views the imputed IBD sharing at the k markers for an ASP as repeated measures with a mean function

$$\mu(M_i; l) = 1 + (1 - 2\theta_i(l))^2(\mu(l) - 1) = 1 + C(1 - 2\theta_i(l))^2, \quad i = 1, 2, \dots, k,$$

that is parameterized as a function of the gene effect (C) and the gene location (l). Solving the generalized estimating equations provided an estimate of (C, l) and the associated standard errors. CIs can be constructed by invoking normality as an approximation of the sampling distribution of the estimates.

Contrary to the bootstrap and the GEE approaches' focus on approximating the sampling distributions of the location estimates, the CSI framework [Papachristou and Lin, 2006c] employs yet another general approach to CI construction: the duality of CI construction and hypothesis testing. Specifically, each genomic location within the candidate region is tested under the null hypothesis that it is the disease gene location, using a level α statistical test. A $(1 - \alpha)$ CI then consists of all the genomic locations where the null is accepted. In order to carry out a CSI approach, the IBD sharing probabilities at the trait locus, (z_0, z_1, z_2) , or a function of them, needs to be known. Thus a two-step procedure is often adopted: estimates of (z_0, z_1, z_2) from the first step are used as known in the second, CI construction, step. Counterparts to various traditional linkage test statistics (under the null of no linkage) can be used as the test statistics in the CSI framework: using the MLS test statistics (CSI-MLS) results in some improvement over the use of the Mean statistics in the original CSI (CSI-Mean) [Sinha and Luo, 2007b].

With all these recent developments, Papachristou and Lin [2006a] conducted a thorough comparison of their performance. This revealed the need for improvement: the GEE and the 1-LOD-support intervals can be very liberal, especially for relatively small sample sizes and moderate genetic effects attributable to the trait locus; CSI-Mean is too conservative and thus results in longer intervals than justified; and the bootstrap methods are not efficient even though the nominal coverage is well maintained. Though most methods, notably the 1-LOD-support, improve considerably when the sample size is sufficiently large, the linkage data for complex traits that are of interest nowadays will almost certainly fall within the scenarios where a more efficient method is desirable.

One attractive alternative to the above approaches is to utilize Risch's likelihood in a Bayesian framework for estimating the location of the disease gene. Under certain conditions, a Bayesian credible interval is asymptotically equivalent to a confidence interval of the same level [Gelman et al., 2004]. Noting the similarity of estimating a trait locus under certain idealized models (affected half-sibling design, completely informative markers, etc.) to the general class of problems of estimating a change-point, which has been studied extensively, Siegmund [1998] suggested that a uniform prior on the disease gene location provided satisfactory confidence regions even for small sample sizes. Application of the uniform prior to quantitative trait loci mapping in experimental crosses confirmed this expectation [Dupuis and Siegmund, 1999]. In what follows, we first define our Bayesian model with a uniform prior on l and describe an algorithm for sampling from the posterior distribution. Extensive simulation studies and an application of our Bayesian Intervals for Linkage Locations (BILL) program to a Rheumatoid Arthritis dataset follow.

2 Methods

2.1 Model

The data consist of n independent ASPs. Let G_i be the genotypes of all markers in the candidate chromosomal region under consideration for the i th ASP (and possibly other family members), $i = 1, \dots, n$. The main parameter of interest is l , the chromosomal position of the susceptibility locus. To completely specify the probability of the observed data, the probabilities of IBD sharing are needed as additional parameters. Specifically, $z_j = \Pr(\text{IBD}_i = j \mid \text{ASP})$, $j = 0, 1, 2$, where IBD_i denotes the number of alleles shared IBD by the i th pair of siblings at position l . Let $\pi(l, z_0, z_1, z_2)$ be the prior distribution of the parameter vector. Our inference is based on the posterior distribution

$$\pi(l, z_0, z_1, z_2 \mid \text{Data}) \propto \pi(l, z_0, z_1, z_2) \Pr(\text{Data} \mid l, z_0, z_1, z_2) \quad (1)$$

The probability of the data given the parameter vector or, equivalently, the likelihood of the parameter vector given the observed data, needs to be evaluated for all possible values of the parameter vector, which is computationally impractical. For computational reasons, we allow l to take on values only from a fine grid of positions (say every 0.25 cM) within the candidate region. Further manipulation of the likelihood [Risch, 1990; Sinha and Luo, 2007b] greatly simplifies the computational demand and takes advantage of quantities that are readily available from several well tested software. Specifically,

$$\begin{aligned}
\Pr(\text{Data}|l, z_0, z_1, z_2) &= \prod_{i=1}^n \Pr(G_i|l, z_0, z_1, z_2) \\
&= \prod_{i=1}^n \sum_{j=0}^2 \Pr(G_i|\text{IBD}_i=j) z_j \\
&= \prod_{i=1}^n \Pr(G_i) \sum_{j=0}^2 \frac{\Pr(\text{IBD}_i=j|G_i)}{\Pr(\text{IBD}_i=j)} z_j \\
&\propto \prod_{i=1}^n \sum_{j=0}^2 \frac{\Pr(\text{IBD}_i=j|G_i)}{\Pr(\text{IBD}_i=j)} z_j.
\end{aligned} \tag{2}$$

We are able to omit $\Pr(G_i)$ from the last expression, since it is not needed in the Markov chain Monte Carlo (MCMC) sampling of the posterior in (1) (details will be given later).

Note that both the denominator and numerator of $\frac{\Pr(\text{IBD}_i=j|G_i)}{\Pr(\text{IBD}_i=j)}$ in (2) are not conditioned on the fact that both siblings are affected. That is, $\Pr(\text{IBD}_i=j)$ are 1/4, 1/2, 1/4, for $j = 0, 1, 2$, respectively. Hidden Markov Models and algorithms have been used by several genetic epidemiology software to compute $\Pr(\text{IBD}_i=j | G_i)$ (e.g., S.A.G.E. [2008], Merlin [Abecasis et al., 2002] and Allegro [Gudbjartsson et al., 2000]). These quantities need to be computed only once for each possible value of l , and combined with any value of z_j s in equation (2) to provide the corresponding likelihood.

The parameterization in (1) and (2) with z_j s is straightforward, but the z_j s should conform to the Holman's triangle constraint ($z_1 \leq 1/2$ and $2z_0 < z_1$). This poses difficulties for the specification of appropriate prior distributions and the MCMC sampling from the posterior distribution. After some experimentation, we found a practical and equivalent alternative is to use the genetic variance components. Consider a single-locus disease model. Let K be the prevalence of the disease, $\alpha = \frac{V_A}{K(1-K)}$ and $\delta = \frac{V_D}{K(1-K)}$, where V_A is the additive genetic variance and V_D is the dominance genetic variance. Then [James, 1971]

$$\begin{cases} z_0(\alpha, \delta) = \frac{\frac{K}{1-K}}{4\frac{K}{1-K} + 2\alpha + \delta} \\ z_1(\alpha, \delta) = \frac{\frac{2K}{1-K} + \alpha}{4\frac{K}{1-K} + 2\alpha + \delta} \\ z_2(\alpha, \delta) = \frac{\frac{K}{1-K} + \alpha + \delta}{4\frac{K}{1-K} + 2\alpha + \delta} \end{cases} \tag{3}$$

Substituting (K, α, δ) for (z_0, z_1, z_2) in (1) and (2), it is then easier to specify sound priors for (α, δ) and to sample from the posterior distribution of these parameters, together with l , using MCMC. However, the practical parameterization comes with a price; the data, consisting only of ASPs, provide no information on K . Thus, we assume K is known without uncertainty. Because accurate prior estimates of K are not always available, we investigate the effect of misspecification of K on the performance of our method later.

The alternative parameterization might seem unnecessarily contrived at first sight. It might further be argued that it is better to put a prior on K rather than fixing it at a more or less arbitrary value. First note that $K(1-K)$ is the phenotypic variance for a binary trait. Thus α is the narrow sense heritability under a single-locus disease model. Both α and δ lie between 0 and 1. It is evident that the z_j s as defined in (3) automatically satisfy the Holman's triangle for any α and δ in this range. In addition, it is the relative magnitudes of $\frac{K}{1-K}$, α and δ that determines the z_j s, not their absolute magnitudes. Thus, allowing K to vary would create a non-identifiability problem.

Because K , α and δ can be interpreted as anything that is proportional to a baseline configuration, anchoring K at roughly the prevalence enables selecting the appropriate priors on α and δ such that the configuration of the z_{js} is likely to be supported by the data. This is a clear advantage over the original parameterization. At times of doubt, one should intuitively fix K at the lowest reasonable guess, since the prior we put on α and δ (a beta distribution for $\alpha+\delta$) does not allow them to go over 1, but there is substantial probability for them to be quite close to 0. This suggestion has been corroborated by later results on the effect of “misspecifying” K . Alternatively, one might even obtain the MLE of the z_{js} from the data and make an educated selection of K so that the corresponding α and δ solved from the MLE of the z_{js} would lie in roughly the center of their prior distribution (see below). Lastly, the parameterization in terms of (3) is necessary when, in addition to ASPs, the data contain other types of affected relative pairs (e.g., uncle-nephew and half-sib). The allele sharing probabilities at the trait locus for different types of affected relative pairs will have some functional relationship that is better represented by using K , α and δ .

2.2 Prior distributions

To complete the model specification in (1), the prior distribution, $\pi(l, \alpha, \delta)$, is needed. Let l be independent of (α, δ) . Following Siegmund [1998], the prior of l is chosen to be a uniform distribution on a fine grid of positions within the candidate region, which is every 0.25 cM in our current implementation.

For the prior on α and δ , we first provide a prior for $\gamma = \alpha + \delta$, the locus-specific broad sense heritability. A natural and convenient prior is a Beta distribution with shape parameters β_1 and β_2 . Conditional on γ , we let the prior on α be a uniform distribution between 0 and γ . Transformation leads to the joint prior on (α, δ)

$$\pi(\alpha, \delta; \beta_1, \beta_2) = \frac{1}{B(\beta_1, \beta_2)} (\alpha + \delta)^{\beta_1 - 2} (1 - \alpha - \delta)^{\beta_2 - 1}, \quad 0 \leq \alpha, \delta, \alpha + \delta \leq 1, \quad (4)$$

where $B(\beta_1, \beta_2)$ is the Beta function.

Choice of hyperparameters—Judicious choice of the hyperparameters (β_1, β_2) may be important for the performance of our method. To this end, we conducted a survey on the possible range of locus-specific heritabilities of complex traits that our method is designed for. The heritability of Rheumatoid Arthritis (RA) was estimated to be 65% in a Finnish population and 53% in a UK population [MacGregor et al., 2000]. About one-third of this heritability may be attributable to the HLA locus on chromosome 6 [Deighton et al., 1989]. This places the locus-specific heritability of the HLA locus for RA between 0.17 and 0.22. An estimate of the heritability of early age-related Maculopathy is 45% [Hammond et al., 2002], which places heritability for a major locus in the range of 0.1 to 0.2 if there is one locus that contributes a significant fraction to this heritability. These heritability estimates, together with findings for a wide variety of complex traits in diverse populations, led us to postulate that a reasonable prior for γ should take values between 0 and 0.45, with a peak around 0.15, when K is specified as the population prevalence. We choose $\beta_1 = 3$ and $\beta_2 = 13$, which results in a mode of 0.14, a mean of 0.19, a standard deviation of 0.095. The 99% quantile is 0.45. Sensitivity analysis (Section 3.2) showed that the estimation of the susceptibility locus l is not sensitive to choice of β_1 and β_2 ,

2.3 Sampling from the Posterior using Markov chain Monte Carlo

The inference of l is based on a sample drawn from the posterior distribution of (l, α, δ) . A Metropolis-Hastings algorithm [Gilks et al., 1996] was employed to construct the Markov chain from which the posterior was drawn. We briefly review a generic Metropolis-Hastings algorithm here. Let x be the observed data and θ be the parameter vector. In a Bayesian framework, the posterior, $\pi(\theta | x) \propto \pi(\theta)\text{Pr}(x | \theta)$, is the target distribution that we wish to draw a sample from. A proposal distributions $q(\theta' | \theta_t)$ is required, which proposes the next possible update of the parameter (θ') when the current value of the parameter is θ_t . The newly proposed θ' will be accepted as the state of the chain in the next iteration with probability

$$A(\theta_t, \theta') = \min \left[1, \frac{\pi(\theta' | x)q(\theta_t | \theta')}{\pi(\theta_t | x)q(\theta' | \theta_t)} \right] = \min \left[1, \frac{\pi(\theta')\text{Pr}(x | \theta')q(\theta_t | \theta')}{\pi(\theta_t)\text{Pr}(x | \theta_t)q(\theta' | \theta_t)} \right]. \quad (5)$$

Otherwise the chain will take the current realization of θ in the next step. Note that the acceptance probability is a ratio, thus the posterior needs only be known up to a constant, as does the probability of the observed data. This is the basis for the manipulation of the likelihood expression in (2) to improve computational efficiency. We start the chain with an arbitrary value θ_0 for θ . Then a candidate θ' is sampled from $q(\theta' | \theta_0)$ and

$$\theta_1 = \begin{cases} \theta' & \text{with probability } A(\theta_0, \theta') \\ \theta_0 & \text{otherwise} \end{cases}.$$

Further realizations of θ are then similarly sampled in subsequent iterations. When the chain is run long enough, $\theta_0, \theta_1, \dots, \theta_n$ will form a correlated sample from the target distribution $\pi(\theta | x)$ and inference on θ can be based on this sample.

Proposal distribution—After some experimenting, we adopted a proposal distribution for our problem at hand that seems to perform well. Each iteration consists of two steps, where l and (α, δ) are updated separately. The proposal for a new l is a uniform distribution on 7 points in the grid of possible chromosomal positions, consisting of the current location and 3 immediately adjacent locations on either side of the current location. The proposal distribution $q(\alpha', \delta' | \alpha_t, \delta_t)$ consists of independent proposals $\alpha' | \alpha_t \sim N(\alpha_t, \sigma_0^2)$ and $\delta' | \delta_t \sim N(\delta_t, \sigma_0^2)$. We update α and δ simultaneously out of two considerations. First, these two parameters represent the proportions of trait variance attributable to the additive and dominance components at the locus. Given a fixed heritability, these two parameters are negatively correlated. One way to overcoming possible poor mixing of the Markov chain when parameters in a Bayesian problem are correlated is to update all parameters simultaneously [Gilks et al., 1996]. Second, joint updating speeds up the chain because the requirement to compute likelihoods is only half that required for separate updating. Typical values of α and δ will be in the range (0, 0.45), thus we let $\sigma_0 = 0.01$, which seems to produce well-mixing chains. For each dataset, the Markov chain was run for 50,000 iterations, after allowing for a burn-in of 5,000 iterations (See Section 3.2 for convergence diagnostics that lead to these choices). The output of every 50th iteration was retained to provide a practically independent sample from the posterior distribution. Using this sample, Highest Posterior Density credible intervals were constructed for l [Chen and Shao, 1999] and the posterior mode was used as a point estimate of the gene location.

3 Simulation study

3.1 Simulation setting

Disease models—To compare the Bayesian intervals for linkage locations (BILL) approach to other available methods for intermediate fine mapping, we carried out simulation studies employing three single-locus and eight two-locus disease models. In the two-locus models, the trait loci were assumed to segregate independently and be in linkage equilibrium. The trait locus (loci) in each model is diallelic and follows Hardy-Weinberg proportions. The three one-locus disease models (Table I) have prevalences ranging from 4.5% to 8.5% and heritabilities ranging from 0.094 to 0.143. The modes of inheritance include Recessive, Intermediate, and Dominant. These models represent characteristics observed in complex traits of interest today. The Dominant model has been investigated in other studies [Papachristou and Lin, 2006b; Sinha and Luo, 2007b].

In addition to a linked major locus, given a set of ASPs, for many complex diseases of interest today we expect other factors, genetic and/or environmental, to contribute to disease susceptibility. To investigate the performance of the intermediate fine mapping methods under a more realistic setting, additional data were simulated with eight two-locus disease models. These are a subset of the 12 models considered by Knapp et al. [1994]. They chose the allele frequencies and penetrances for each model such that the prevalence is roughly 0.1, the total offspring risk is 0.3, and the two loci contribute equally to the trait prevalence. Only four of the models are not symmetric in the two loci and have been considered in the intermediate fine mapping context (e.g., Papachristou and Lin [2006c], Sinha and Luo [2007b]). These four models were termed EP-2 (epistatic recessive-dominant), HET-2 (heterogeneous recessive-dominant), S-2 (special case of HET-2 with the second locus fully penetrant), and EP-4 (epistatic). In addition to these four two-locus models, we considered another four two-locus models in Knapp et al. [1994] that are symmetric in the two trait loci: EP-1 (epistatic dominant-dominant), HET-1 (heterogeneous dominant-dominant), EP-3 (epistatic recessive-recessive), and HET-3 (heterogeneous recessive-recessive). Let A and a be the high-risk and low-risk alleles, respectively, at the 1st locus and similarly define B and b at the 2nd locus. Let p_1 and p_2 be the frequency of the high-risk alleles at the two loci, respectively. Among the eight two-locus models, four are heterogeneous: HET-1, HET-2, S-2, and HET-3. Each of these models is determined by two parameters: ϕ_1 as the penetrance of two-locus genotypes where the 1st locus genotype is high-risk and the 2nd locus genotype is low-risk, and ϕ_2 as the penetrance of two-locus genotypes where the 2nd locus genotype is high-risk and the 1st locus genotype is low-risk. Penetrance for genotypes that are high-risk at both loci is $\phi = \phi_1 + \phi_2 - \phi_1\phi_2$, while it is zero for those that are low-risk at both loci. For example, under the heterogeneous recessive-dominant (HET-2) model, penetrance is ϕ_1 for genotype $AAbb$, ϕ_2 for ($AaBB$, $aaBB$, $AaBb$, $aaBb$), ϕ for ($AABB$, $AABb$), and zero otherwise. These heterogeneous models are approximately additive when both ϕ_1 and ϕ_2 are small. The other models are epistatic, each determined by a single parameter ϕ , the common penetrance of high-risk genotypes. EP-1, EP-2, and EP-3 are all multiplicative. For example, under EP-2 (epistatic recessive-dominant), the penetrance is ϕ for genotypes made up of AA at the 1st locus and BB or Bb at the 2nd locus, and is zero for all the other genotypes. Lastly, the high-risk genotypes under EP-4 include $AABB$, $AABb$, $AaBB$, and $aaBB$. The other models can be similarly visualized from their suggestive names. The parameters (p_1, p_2, ϕ) are (.6, .199, .778) under EP-2, and (.372, .243, .911) under EP-4, respectively. Under HET-2, $p_1 = .279$, $p_2 = .04$, and $\phi_1 = \phi_2 = .66$. Under S-2, $p_1 = .228$, $p_2 = .045$, and $\phi_2 = .66$. The other models are symmetric and each is determined by a single allele frequency and a single penetrance parameter; these are (.210, .707) for EP-1, (.577, .9) for EP-3, (.053, .495) for HET-1, and (.194, 1) for HET-3.

Datasets of 100, 250, and 500 ASPs were simulated under the one-locus models and datasets of 500 ASPs were simulated under the two-locus models.

Marker maps—For one-locus models, a chromosome of 60 cM was simulated with one SNP (minor allele frequency 0.3) every 0.25 cM. Linkage equilibrium was assumed among all loci, including the trait locus. The trait locus was placed at 30.125 cM. Both disease chromosomes had the same characteristics as described above when two-locus models were considered. The genotype data of the parents of each ASP were available unless otherwise noted.

Confidence set construction—Confidence sets of the disease susceptibility locus were constructed, for each set of simulated ASPs, using five methods: (1) BILL, (2) parametric Bootstrap [Papachristou and Lin, 2006a] since it was shown that the parametric version performed better than the non-parametric version, (3) CSI-MLS [Sinha and Luo, 2007b], (4) LOD-support interval with 1-LOD drop using KAC LOD score, (5) GEE [Liang et al., 2001]. Since all these methods were proposed as intermediate fine mapping tools, we shall employ these methods only in regions that have shown evidence of linkage. Specifically, we divided the marker data into two maps, a step-one map and a step-two map, each consisting of half of the SNPs with 0.5 cM spacing. Confidence sets were constructed using the step-two map only when the KAC LOD score using data from the step-one map exceeded 2.33, a threshold commonly used to flag suggestive linkage. We note that it is commonly known now [e.g., Papachristou and Lin, 2006a] that using the totality of the two maps to construct CIs in the 2nd step would yield essentially the same result. Furthermore, Papachristou and Lin [2006a] suggested focusing the CI construction effort to within 25 cM on either side of the linkage peak. The size of the candidate region should be dictated by the premise that there is only one trait locus within the region. Most of the five methods run sufficiently fast that markers on the entire chromosome can be used to construct CI if there is no danger of having more than one trait locus. BILL, GEE, and CSI-MLS, in their current implementations, require input files from other software (e.g., Merlin, S.A.G.E. GENIBD) that contain IBD sharing probabilities at all potential disease gene locations. The size of these IBD files limits their applications to candidate regions of no more than 100 cM. The candidate region is the entire 60 cM in the simulated data. For BILL, K was given the value of the population prevalence of the disease under the simulation model. For CSI-MLS, the step-one map is used to estimate the MLEs of z_j s to be supplied to the 2nd, CI construction, step.

3.2 Results

Sensitivity analysis—To investigate the sensitivity of the posterior distribution to the choice of prior distribution on (α, δ) , we simulated two datasets from the Intermediate disease model (Table I), one with 100 ASPs and the other with 500 ASPs. Posterior distributions from five different Beta priors applied to these two datasets are summarized in Table II. The inference on the main parameter of interest, the position of the susceptibility locus l , is not sensitive to choice of prior distributions, even with only 100 ASPs. While the posterior distributions of α and δ show some sensitivity to prior choice when the sample size is small, the influence is minimal.

Convergence diagnostics—A chain should have migrated into the target distribution for the inference based on the chain sample to be valid. While convergence characteristics for each chain will be different, we performed some convergence diagnostics to decide, *a priori*, on the required running length. Our test dataset, for this purpose, consisted of 100 ASPs simulated under the Intermediate model (Table I). First, visual inspection of an extremely long chain suggested that the chain converged after around 5,000 iterations. A

formal diagnostics was also performed. The Potential Scale Reduction Factor (PSRF) [Gelman and Rubin, 1992] is the square root of the ratio of the “between” and “within” variances, computed from several independent Markov chain sequences started from an initial sample of a distribution overdispersed relative to the target distribution. When each chain has been run indefinitely, the distribution of each sequence will all be the same as the target distribution and thus the PSRF equals 1. Because of the overdispersion of the starting points, PSRF will be greater than 1 when the chains have not yet converged to the target distribution. Gelman et al. [2004] suggest that PSRF less than 1.1 indicates convergence. The PSRF of the three parameters were computed based on four chains, each of 200,000 iterations and started from an overdispersed distribution. Since α and δ are proportions, a “logit” transform was applied to them. The PSRF of (l, α, δ) is (1.00, 1.03, 1.03) at iteration 10,000, reduces to (1.00, 1.00, 1.02) at iteration 50,000 and stays stable thereafter. Therefore, a sample of 50,000 iterations following a burn-in of 5,000 iterations was retained for inference based on each dataset.

Precision and coverage of nominal confidence sets—We investigated the empirical coverage and precision of 95% confidence intervals constructed using each of the five intermediate fine mapping approaches, for data simulated under the one-locus or two-locus disease models. For each model, 250 replicates were simulated. Empirical coverage is defined as the ratio of the number of intervals that captured the location of the trait locus, over the number of intervals that signaled linkage (KAC LOD score > 2.33). Precision is measured by the length of the interval. The number of replicates showing suggestive linkage was 243, 217 and 155, for datasets of 100 ASPs simulated under each of the three one-locus models, respectively. This number is > 230 for datasets of 250 ASPs and = 250 for samples of 500 ASPs. This suggests that if the true coverage of a confidence interval estimator is 95%, the empirical coverage should fall between 93% and 97% for at least 80% of the simulations.

Tables III and IV show the average length and the empirical coverage of the intervals constructed from data simulated under the single-locus and two-locus disease models, respectively. Increased sample size results in more precise interval estimates (Table III). Only the Bayesian and the Bootstrap methods maintained the nominal coverage level well. The intervals constructed using BILL are, on average, only 50% to 81% of the length of that using Bootstrap, except for the HET-3 model, where the two methods perform comparably. CSI-MLS is conservative, with empirical coverage being 100% under most settings, and consequently provides confidence intervals that are 1.64 to 4.76 times as long as the BILL intervals. The GEE intervals exhibit substantial under coverage, with most empirical coverage being around 85% and the most extreme one being only 72%. Even with this level of under coverage, 80% of the GEE intervals are longer than the BILL intervals, sometimes twice as long. LOD-support intervals show a moderate level of under coverage. When the empirical coverage of the LOD-support intervals are close to 95% (defined as between 93% and 97%, which occurred in 8 out of 21 simulation settings), 62.5% of the settings result in shorter BILL intervals, while 37.5% of the settings yield shorter LOD intervals (the bold entries in the two Tables). We will examine the relative performance of these two approaches in more details in what follows.

It is evident from Table III that the Recessive model results in more precise intervals than does the Intermediate model. The Dominant model yields least precision. The heritability of these three models, however, follows the reverse order.

Precision of confidence sets under controlled empirical coverage—The widely divergent empirical coverage levels of the different methods render direct comparison of efficiency difficult, although it is quite clear from the last paragraph that BILL is the most

favorable, with LOD intervals being a serious contender. Another experiment was conducted: empirical coverage was fixed at several levels and the precision of the intervals were compared. The results are displayed in Figures 1 (one-locus models), 2 (non-symmetric two-locus models), and 3 (symmetric two-locus models - note that only results from one locus are plotted, in view of the symmetry). The GEE intervals are the least precise. While the differences in precision of the other intervals tend to narrow with increased sample size (Figure 1), lengths of the GEE intervals remain widely separated from those of the others. The Bootstrap intervals are also long compared to others for small sample sizes (100 ASPs) and large empirical coverage levels. CSI-MLS intervals are often close to the most efficient intervals in precision.

The comparison of BILL and LOD intervals confirms the findings from the nominal intervals. For single-locus disease models, BILL intervals are the most precise except for the case of 250 ASPs under the Dominant model, where they are slightly less precise than the LOD and CSI-MLS intervals. LOD intervals outperform BILL intervals in five cases under two-locus models: EP-1, HET-1, the second locus of each of EP-2, HET-2, and S-2. The mode of inheritance of each of these disease loci is dominant. Close examination of these disease models shows that the probability that two affected siblings share 1 allele IBD at each of these loci is close to 0.5, which means that $\delta \approx 0$. Irrespective of this fact, the Bayesian method is estimating it in the model. To investigate whether this is the reason for BILL not outperforming LOD-support intervals, a variant of BILL, BILL-V2, where δ is set to be zero, was also developed. BILL-V2 was applied to datasets of 500 ASPs simulated under some of the models with marginal mode of inheritance being dominant. BILL-V2 intervals are improved over the BILL intervals, but are still inferior to the LOD intervals for the dominant loci under the non-symmetric two-locus models (Supplementary Table S1).

While the LOD intervals are more precise than the BILL intervals for data simulated under dominant models, it is not clear what drop size corresponds to a desired level of coverage. Both mode of inheritance and sample size influence the drop size, ranging from 1.04 to 1.62 when the three single-locus models and sample sizes of 100, 250, and 500 were considered (Supplementary Table S2). The drop size can be as much as 1.48 for 500 ASPs. Other uninvestigated factors might also play a role, further complicating the choice.

Estimates of susceptibility gene locations—Four of the intermediate fine mapping methods—BILL, Bootstrap, GEE, and LOD—also provide estimates of the susceptibility gene location. In particular, the location with the maximum K&C LOD score is used as a point estimate. Root Mean Squared Errors (RMSE) of these estimates are given in Table V for data simulated under the one-locus models. BILL estimates are the best under both the Recessive and the Intermediate models, while Bootstrap estimates are less precise for small sample size (150 ASPs) and are just as precise for sample sizes on or above 250 ASPs. The estimates from GEE and LOD are much less precise. On the other hand, under the Dominant model, LOD provides the most precise estimate with 250 ASPs and slightly better estimate than those from Bootstrap and BILL under 500 ASPs, while the estimate under 100 ASPs is much less precise than those from the other three methods.

Effect of the specification of K —As discussed following equation (3), the role of K is that of a scaling parameter. When it is fixed at the prevalence of the disease attributable to the trait locus under consideration, the other two parameters, α and δ , have the interpretation of narrow sense heritability and the proportion of dominance genetic variance, respectively, under a single-locus disease model. This interpretation facilitates selection of reasonable priors for α and δ , demonstrated through the good performance of BILL when applied to all the simulated data. The effect of specifying K as $\zeta \times (\text{prevalence})$ is shown in Table VI for five different values of ζ (0.1, 0.5, 1.0, 2.0 and 4.0), based on samples of 250 ASPs

simulated under the single-locus models. The effect of ζ on the length of the credible interval is minimal and hence is not shown. A value of $\zeta \leq 1$ has little or no effect on the empirical coverage, while $\zeta \geq 2$ leads to reduced coverage. This effect is more pronounced for the Dominant model (bold entries).

Effect of missing parental genotypes—We also investigated the effect of missing parental marker data on the performance of the five intermediate fine mapping methods, based on samples of 250 ASPs simulated under the single-locus Intermediate model (Supplementary Table S3). Although increasing percentage of missing parental marker data lead to slightly longer intervals, the empirical coverage and efficiency comparison of these methods confirm what was observed when all parental data were available. The change in the length of the CI with increasing percentage of missing parental information is more pronounced with the LOD-support intervals, probably due to its unstable coverage level, compared to the other approaches.

4 Application to data from the North American Rheumatoid Arthritis Consortium study

4.1 Data description, preprocessing, and analysis strategy

To further evaluate BILL and other competing methods, we analyzed the Rheumatoid Arthritis (RA) data contributed to Problem 2 of Genetic Analysis Workshop 15 by the North American Rheumatoid Arthritis Consortium (NARAC) [Amos et al., 2007]. Microsatellite marker data for 511 families and SNP data for 757 families were provided. The locus HLA-DRB1 has been consistently implicated by numerous studies as a highly significant risk factor for RA. In our analysis we used only the SNP data as it has been shown by numerous papers that the SNP panels available nowadays are satisfactory. Only the physical map of the SNPs was provided in the GAW15 Problem 2 data. Both physical distances and genetic map positions were available, however, for a denser map of SNPs on chromosome 6 in the GAW15 Problem 3 data. Thus, we obtained the genetic map positions for SNPs in the Problem 2 data by matching information on SNPs in the two problems, using interpolation when necessary. HLA-DRB1 is located at 49.46 cM in this genetic map. There were 404 available SNPs on chromosome 6. Because marker-marker linkage disequilibrium (LD) can lead to biased estimation of IBD sharing probabilities and the SNPs are sufficiently dense, we opted for filtering out SNPs in LD instead of incorporating the LD in the methods. Haploview [Barrett et al., 2005] was employed to filter the SNPs so that the remaining SNPs had pairwise r^2 less than 0.02. This resulted in 284 SNPs across 197.4 cM. The candidate region was set to be the first 100 cM (142 SNPs), around half of chromosome 6.

Of the 757 families, 722 independent ASPs, together with genotypes from other family members who might provide information on IBD sharing, comprised our full set of data. The unusually large sample size and the substantial genetic effect of the HLA-DRB1 on RA presented an opportunity to investigate in some detail the performance of the intermediate fine mapping methods, including their asymptotic behavior. To this end, we applied the methods to seven sets of nested samples of sizes 150, 200, 300, 400, 500, 600, and 722 ASPs, respectively. The corresponding KAC LOD scores are 2.51, 3.15, 4.22, 5.21, 8.91, 10.54, and 16.33. We chose the minimum size of 150 ASPs so that the LOD score is still > 2.33 , the threshold we used for when it is worth pursuing intermediate fine mapping. For CSI-MLS, the two-step procedure was applied to each sample with the z_{β} s estimated in the 1st step using all 142 SNPs in the 100 cM region. The same set of SNPs was used in the 2nd step for CI construction. Since the underlying IBD process is the same for the same set of ASPs and the SNPs are sufficiently dense for tracking the IBD process, this strategy yields essentially the same result as one that used exclusive sets of SNPs in the two steps. GEE

(implemented in GeneFinder version 1.0) requires IBD sharing probabilities that is output from GENEHUNTER. GENEHUNTER eliminated some pedigrees, claiming numerical underflow during likelihood calculation or insufficient information available on IBD sharing. The numbers of pedigrees dropped were 8, 8, 16, 18, 25, 31, and 38 for sample sizes increasing from 150 ASPs to 722 ASPs.

4.2 Results

Kong and Cox LOD scores and 95% confidence/credible intervals constructed using the five methods are displayed in Figure 4. Numerical details are provided in supplementary Table S4. As the sample size increases, point estimates of the disease gene location converge to the HLA-DRB1 locus for BILL, 1-LOD-support, and Bootstrap. CSI-MLS does not provide a point estimate. On the other hand, GEE consistently yielded point estimates on the opposite side of the true location, compared to the other three methods. The series of point estimates seem to converge towards the true location, for sample size going from 150 to 300, only to go further away thereafter, with the point estimate being more than 5 cM away from the HLA-DRB1 locus for the full data set. As a consequence, although the width of the CI decreases from 18.84 cM to 7.80 cM from the 150-ASP sample to the 722-ASP sample, the CI does not contain the true location when the sample size is at or above 500. Each of the other four methods provides series of CIs that captures the true location and consistently becomes more precise with increasing sample size. The most dramatic improvement occurs with Bootstrap: the CI from the full data set is 5.3 times as precise as that from the 150-ASP sample; this number is 1.9, 3.6, and 4.0, for CSI-MLS, BILL, and 1-LOD-support intervals, respectively. A few exceptions to the monotone trend do occur: at sample size 600 for CSI-MLS, 200 for Bootstrap, and 400 for BILL and LOD-support.

Comparing the precision of the intervals from all methods, except those from the GEE approach due to their bias, shows that BILL is quite efficient relative to the other methods. Using BILL intervals as baseline, the ratio of CSI-MLS interval lengths to BILL interval lengths are 1.9, 2.0, 2.1, 2.0, 3.1, 3.3, and 3.5 for sample size going from 150 to 722. These numbers are 1.6, 2.4, 1.3, 1.09, 1.7, 1.16, and 1.04 for Bootstrap and 1.16, 1.00, 0.90, 0.87, 1.18, 1.11, and 1.05 for 1-LOD-support intervals. Thus the conservative nature of CSI-MLS is evident. Bootstrap is apparently not efficient compared to BILL, despite the fact that they both maintain the nominal coverage of the intervals well. BILL and 1-LOD-support intervals are mostly similar. However, the uncertainty of the appropriate drop size to obtain a 95% confidence level and the LOD-support intervals' frequent tendency to be liberal makes it difficult to conduct an overall fair comparison of the LOD-support approach with BILL. Recall that our limited investigation showed that the appropriate drop size can go from 1.04 to 1.6. The relative precision of BILL to the 1.05-LOD-support intervals are, going from the 150-ASP sample to the full sample, 1.17, 1.02, 0.98, 0.92, 1.21, 1.18, and 1.14. These ratios become 1.27, 1.18, 1.07, 1.02, 1.29, 1.21, and 1.33 for a drop size of 1.20, and 1.92, 1.42, 1.41, 1.24, 1.61, 1.46 and 1.52 for a drop size of 1.60 (Supplementary Table S5).

Convergence of the Markov chain we used for BILL was examined for one of the samples. The Potential Scale Reduction Factors, computed after running four chains started from an overdispersed distribution, were 1.01, 1.03, and 1.04, for l , α , and δ , respectively. Thus, there was no evidence against convergence. We set $K = 0.0107$ for BILL.

5 Discussion

After identifying a genomic region that is potentially linked to a complex disease, construction of precise yet valid statistical confidence intervals of disease gene locations within this region is an important component in positional cloning. Several approaches have been proposed for this endeavor, but comparison of their performance indicates there is

much room for improvement. In this article, we have proposed a novel Bayesian approach, BILL, that jointly estimated the disease location and the IBD sharing probabilities at the trait locus. The posterior distribution of these parameters is approximated using Markov chain Monte Carlo. The posterior mode of the disease gene location serves as a point estimator and the Highest Posterior Density intervals are obtained for desired credible levels.

Comparison of BILL with existing methods (LOD-support, GEE, CSI-MLS, Bootstrap), under simulations of single- and two-locus disease models and an extensive analysis of the NARAC Rheumatoid Arthritis data, suggests that BILL is statistically both valid and efficient. In addition to corroborating findings found by others [Papachristou and Lin, 2006a], the comparison makes some surprising discoveries. In terms of faithfulness to the nominal coverage, intervals from CSI-MLS are conservative, while those from GEE and 1-LOD-support are liberal, sometimes substantially so. BILL and Bootstrap are the only two methods that remain truthful to the nominal coverage across all simulation settings, and BILL intervals are from 1.64 to 4.76 times as precise as Bootstrap intervals. When empirical coverage was held constant so that efficiency of the information utilization could be separated from variation in empirical coverage resulting from the methods's ability to approximate the sampling distributions, LOD-support emerges as a serious contender to BILL under simulation models with $z_1 \approx 0.5$. This might be due to the fact that the LOD-support intervals were based on KAC LOD scores, which are efficient LOD scores for disease models with only additive genetic components. A fairer comparison might be to base the LOD-support interval on the MLS. It was a surprise to discover via the analysis of the RA data that GEE is substantially biased.

Why do these methods behave so differently, given each one is based on sound statistical reasoning? Efficiency of CSI-MLS is close to the optimal among the methods compared, so it is not the test statistic (MLS) that is causing the coverage level to be conservative. We believe the reason lies in that the CSI framework, to be practical, use a two-step procedure but does not account for the inflation in the coverage level that results from using the first step estimates of the z_j s as known without error in the CI construction step. Improvement of this framework, thus, would need to focus on ways to correct for this inflation. On the other hand, as a general tool for maintaining nominal coverage level, the parametric Bootstrap, which uses MLEs as the point estimates of the disease gene location, is faithful but not fully efficient - despite its reliance on the efficient MLEs. This lack of efficiency might arise because the sample size required, for the parametric model estimated from the sample to closely approximate the sampling distribution, is larger than most current linkage studies can provide. Suggestive evidence for this comes from the improvement of the Bootstrap intervals in the RA analysis with increasing sample size, to the point of being very similar to the BILL intervals (the optimal) for the 722-ASP data. The remaining approaches, BILL, GEE, and LOD-support with the KAC LOD score, share one commonality: modeling the data with both a parameter for the gene location and a set of nuisance parameters that can be loosely interpreted as genetic effects of the locus, in the form of z_j s, or mean IBD sharing at the trait locus, or the KAC parameter in a cleverly constructed linear allele-sharing model, respectively. Allowing the nuisance parameters to vary with the postulated gene location, together with full likelihood specification, might account for the efficiency of BILL and LOD-support intervals. The alarming bias of the GEE intervals warrant further investigation: is this due to the unsatisfactory modeling of the discrete number of alleles shared IBD with the first two statistical moments, is it due to some implementation idiosyncrasy, or is it due to some other factors? The formal modeling aspect of the GEE framework is desirable, as covariates can then be incorporated relatively easily, compared to the other frameworks, to account for linkage heterogeneity [Glidden et al., 2003]. Lastly, the fidelity of BILL intervals to nominal coverage levels stems, we believe, from its employing Monte Carlo simulation to approximate the posterior distribution, which is only possible

with today's computing power, instead of invoking an asymptotic approximation. Furthermore, the posterior distributions of the gene location from the NARAC samples are skewed, so HPD intervals are more efficient than symmetric credible intervals around the mode, and the effort to obtain HPD intervals is insignificant compared to the other components of BILL.

In the simulation study, we used equally informative, equally spaced SNPs (one per 0.25 cM). The extensive coverage of the human genome by today's marker maps implies that sufficiently dense and informative markers will almost always be readily available to recover the underlying IBD process at the resolution suitable for linkage studies. The application of the methods to the NARAC data, with real SNP maps, yielded the same conclusion as those drawn from the simulation. Furthermore, CI construction by these methods will only be invoked when suggestive linkage has been indicated, so any data that call for the application of these methods would have had IBD sharing distortion at the trait locus that is sufficiently substantial to yield a linkage signal with a realistic sample size. The two-locus and one-locus disease models employed in the simulation studies have been commonly used to study properties of linkage methods in the face of complex traits/disease. Similarity between conclusions from the simulation study and those from the RA data analysis lend support to the generality of the relative performance of the investigated approaches. Further simulations with more realistic disease models could be of interest, but the trait locus under consideration will need to make a substantial contribution to the disease to be able to pass the "suggestive linkage" screen.

As a reviewer pointed out, the rapidly decreasing cost of large-scale sequencing may prompt resurgence of linkage studies in the near future. Rare but high-penetrance variants, which tend to be missed by genetic association studies, are easier to map using large pedigrees enriched with the disease. Thus, it is highly desirable to develop intermediate fine mapping approaches applicable to general pedigrees. The statistical problem would, undoubtedly, be much more difficult. The observations obtained from the current study, however, provide insights that will be helpful in guiding this effort. LOD-support intervals based on traditional parametric LOD scores are a start. But the mode of inheritance in such LOD score computation is fixed across genomic locations in the candidate region. Thus such intervals cannot be expected to perform similar to the KAC-based LOD-support intervals. KAC LOD scores were designed for general pedigrees, but there are several difficulties associated with extending this LOD-support approach to general pedigrees: (1) Simulations by others [Papachristou and Lin, 2006a] showed that even for ASPs, the KAC-LOD-support intervals can be extremely liberal, to the point that a 95% nominal level yields only a 26% empirical coverage. This problem would be aggravated when the data consist of a small number of pedigrees. (2) Optimal scoring functions and weighting factors will depend on the mode of inheritance and the ascertainment scheme, which are usually never known. (3) KAC scores make use of the inheritance vector distribution, the exact computation of which places limits on the size and complexity of pedigrees that can be analyzed. We are currently developing an extension to BILL to address this problem.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is partly supported by a U.S. Public Health Service Resource Grant RR03655, Research Grants R01 HG003054 and R01 GM28356, and Cancer Center Support Grant P30CAD43703. GAW15 was supported by Research Grant R01 GM031575. The NARAC data contributed to GAW15 was gathered with the support of grants from the National Institutes of Health (N01-AR-2-2263 and R01-AR-44422) and the National Arthritis Foundation.

We would like to express our gratitude to two anonymous reviewers whose insightful comments and questions have resulted in a much improved manuscript.

References

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.* 2002; 30:97–101. [PubMed: 11731797]
- Amos CI, Chen W, Remmers E, Siminovich KA, Seldin MF, Criswell LA, Lee AT, John S, Shephard ND, Worthington J, Cornelis F, Plenge RM, Begovich AB, Dyer TD, Kastner DL, Gregersen PK. Data for Genetic Analysis Workshop (GAW) 15 Problem 2, genetic causes of rheumatoid arthritis and associated traits. *BMC Proc.* 2007; 1(Suppl 1):S3. [PubMed: 18466527]
- Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005; 21:263–265. [PubMed: 15297300]
- Chen M, Shao Q. Monte Carlo estimation of Bayesian credible and HPD intervals. *J Comput Graph Stat.* 1999; 8:69–92.
- Deighton CM, Walker DJ, Griffiths ID, Roberts DF. The contribution of HLA to rheumatoid arthritis. *Clin Genet.* 1989; 36:178–182. [PubMed: 2676268]
- Dupuis J, Siegmund D. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics.* 1999; 151:373–386. [PubMed: 9872974]
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis.* Boca Raton, Florida: Chapman & Hall; 2004.
- Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci.* 1992; 7:457–511.
- Gilks, WR.; Richardson, R.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall; 1996.
- Glidden DV, Liang KY, Chiu YF, Pulver AE. Multipoint affected sibpair linkage methods for localizing susceptibility genes of complex diseases. *Genet Epidemiol.* 2003; 24:107–117. [PubMed: 12548672]
- Gudbjartsson DF, Jonasson K, Frigge M, Kong A. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet.* 2000; 24:12–13. [PubMed: 10802644]
- Hammond CJ, Webster AR, Snieder H, Bird AC, Gilbert CE, Spector TD. Genetic influence on early age-related maculopathy: a twin study. *Ophthalmology.* 2002; 109:730–736. [PubMed: 11927430]
- Holmans P. Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet.* 1993; 52:362–374. [PubMed: 8430697]
- James JW. Frequency in relatives for an all-or-none trait. *Ann Hum Genet.* 1971; 35:47–49. [PubMed: 5106369]
- Knapp M, Seuchter SA, Baur MP. Two-locus disease models with two marker loci: the power of affected sib-pair tests. *Am J Hum Genet.* 1994; 55:1030–1041. [PubMed: 7977340]
- Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet.* 1997; 61:1179–1188. [PubMed: 9345087]
- Liang K-Y, Chiu Y-F, Beaty TH. A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. *Hum Hered.* 2001; 51:64–78. [PubMed: 11096273]
- MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, Silman AJ. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* 2000; 43:30–37. [PubMed: 10643697]
- Olson JM. A general conditional-logistic model for affected-relative-pair linkage studies. *Am J Hum Genet.* 1999; 65:1760–1769. [PubMed: 10577930]
- Papachristou C, Lin S. A comparison of methods for intermediate fine mapping. *Genet Epidemiol.* 2006a; 30:677–689. [PubMed: 16917817]
- Papachristou C, Lin S. Microsatellites versus single-nucleotide polymorphisms in confidence interval estimation of disease loci. *Genet Epidemiol.* 2006b; 30:3–17. [PubMed: 16355404]
- Papachristou C, Lin S. A two-step procedure for constructing confidence intervals of trait loci with application to a rheumatoid arthritis dataset. *Genet Epidemiol.* 2006c; 30:18–29. [PubMed: 16355402]

- Risch N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet.* 1990; 46:242–253. [PubMed: 2301394]
- S.A.G.E. 2008. Statistical Analysis for Genetic Epidemiology, version 5.4.
- Siegmund D. Genetic linkage analysis: an irregular statistical problem. *Doc Math J DMV Extra Vol ICM.* 1998; III:291–300.
- Sinha R, Luo Y. Efficient intermediate fine mapping: confidence set inference with likelihood ratio test statistic. *Genet Epidemiol.* 2007a; 8:922–936.
- Sinha R, Luo Y. Two-step intermediate fine mapping with likelihood ratio test statistics: applications to Problems 2 and 3 data of GAW15. *BMC Proc.* 2007b; 1(Suppl 1):S146. [PubMed: 18466490]

Figure 1. Average lengths of confidence intervals over fixed empirical coverages. Data were simulated under the single-locus models.

Figure 2. Average lengths of confidence intervals over fixed empirical coverages. Data were simulated under the non-symmetric two-locus models.

Figure 3. Average lengths of confidence intervals over fixed empirical coverages. Data were simulated under the symmetric two-locus models.

Figure 4.

Confidence intervals (panel B) constructed using the five intermediate fine mapping methods and K&C LOD scores (panel A) across the first 100 cM of chromosome 6. The vertical line marks the HLA-DRB1 locus. Numbers in panel A are the sample sizes. Points in panel B are point estimates of the disease gene location and numbers are the width for BILL or 1-LOD-support intervals. CSI-MLS does not provide point estimates.

Table 1

Single-locus disease models for the simulation study

Model	K	$H = \frac{V_G}{V_T}a$	$\frac{V_A}{V_T}a$	$\frac{V_D}{V_T}a$	p_D^b	f_{dd}^b	f_{Dd}^b	f_{DD}^b
Recessive	0.045	0.094	0.014	0.08	0.08	0.04	0.040	0.840
Intermediate	0.058	0.110	0.050	0.06	0.10	0.04	0.100	0.800
Dominant	0.085	0.143	0.143	0.00	0.01	0.07	0.827	0.865

^a V_T , total variance of the trait; V_G , genetic variance of the trait; V_A , additive genetic variance; V_D , dominance genetic variance; H , broad-sense heritability.

^b D , high-risk allele; d , low-risk allele; P_D , allele frequency of D ; (f_{DD}, f_{Dd}, f_{dd}) , penetrances for genotypes (DD, Dd, dd).

Table II

Sensitivity of posterior inference to choice of hyperparameters

# ASP	Posterior Summary									
	Prior distribution				I (cM)				α	
	β_1	β_2	Mean	SD	Mean	SD	Mean	SD	Mean	SD
100	1	9	0.10	0.09	26.44	3.01	0.091	0.060	0.064	0.042
	2	8	0.20	0.12	25.90	3.51	0.121	0.068	0.062	0.046
	2	13	0.13	0.08	26.67	3.11	0.097	0.058	0.092	0.059
	3	12	0.20	0.10	26.13	2.26	0.118	0.061	0.095	0.052
	1	1	0.50	0.29	26.74	2.95	0.127	0.076	0.049	0.037
	1	9	0.10	0.09	29.95	0.62	0.089	0.026	0.036	0.019
500	2	8	0.20	0.12	30.01	0.81	0.080	0.025	0.033	0.018
	2	13	0.13	0.08	29.94	0.69	0.076	0.024	0.035	0.018
	3	12	0.20	0.10	29.99	0.80	0.081	0.025	0.035	0.017
	1	1	0.50	0.29	29.99	0.69	0.083	0.025	0.028	0.017
	1	9	0.10	0.09	29.95	0.62	0.089	0.026	0.036	0.019
	2	8	0.20	0.12	30.01	0.81	0.080	0.025	0.033	0.018

Mean lengths in cM (L) and empirical coverages (CP) of nominal 95% confidence sets from data simulated under the single-locus models

Table III

Model	#ASPs	BILL		CSE-MLS		LOD-Support		GEE		Bootstrap	
		L	CP	L	CP	L	CP	L	CP	L	CP
Recessive	100	8.91	0.95	23.03	0.97	11.97	0.88	10.95	0.79	15.87	0.93
	250	3.30	0.96	13.25	1.00	5.17	0.94	6.68	0.84	5.07	0.97
	500	1.81	0.97	8.69	1.00	2.90	0.94	4.63	0.80	2.53	0.95
Intermediate	100	10.44	0.94	25.57	0.98	11.69	0.93	11.80	0.83	21.08	0.94
	250	6.48	0.94	16.81	1.00	5.90	0.90	7.80	0.80	8.82	0.94
	500	2.92	0.95	11.81	1.00	3.32	0.92	5.58	0.84	4.41	0.92
Dominant	100	16.10	0.93	26.53	0.99	10.87	0.82	12.97	0.72	26.89	0.95
	250	8.89	0.94	19.55	1.00	6.58	0.93	10.25	0.85	16.21	0.95
	500	4.60	0.93	13.82	1.00	4.05	0.91	7.06	0.84	6.97	0.95

Table IV

Mean lengths in cM (L) and empirical coverages (CP) of nominal 95% confidence intervals from 500 ASPs simulated under two-locus models

Model	Locus	BILL		CSI-MLS		LOD-Support		GEE		Bootstrap	
		L	CP	L	CP	L	CP	L	CP	L	CP
EP-2	1	3.36	0.94	10.71	1.00	3.61	0.91	5.54	0.85	4.39	0.95
	2	3.20	0.94	11.03	1.00	2.36	0.96	5.77	0.85	4.69	0.94
EP-4	1	10.29	0.95	19.10	0.98	8.83	0.88	9.78	0.84	14.48	0.93
	2	1.66	0.96	7.43	1.00	1.88	0.93	3.74	0.87	2.16	0.96
HET-2	1	3.63	0.94	12.88	1.00	5.09	0.91	6.92	0.88	6.12	0.93
	2	4.87	0.93	12.60	1.00	2.85	0.90	6.77	0.86	6.41	0.92
S-2	1	2.01	0.96	8.71	1.00	3.09	0.95	4.63	0.88	2.79	0.96
	2	7.25	0.94	16.79	0.99	4.42	0.92	8.47	0.84	11.36	0.94
EP-1	1	3.25	0.94	12.50	1.00	2.62	0.93	3.54	0.92	5.02	0.96
EP-3	1	2.63	0.93	11.22	1.00	3.42	0.92	3.10	0.90	3.87	0.93
HET-1	1	8.27	0.95	17.66	1.00	4.22	0.92	4.86	0.87	10.21	0.96
HET-3	1	2.88	0.96	9.51	1.00	3.12	0.92	2.67	0.93	2.70	0.95

Table V
Root Mean Squared Error (RMSE, in cM) of the estimate of the susceptibility gene location

Model	# ASP	RMSE of $\hat{t}$			
		GEE	Bootstrap	BILL	LOD
Recessive	100	3.94	4.47	3.08	6.28
	250	2.51	1.06	1.02	1.84
	500	1.73	0.62	0.66	1.01
Intermediate	100	4.68	4.79	4.33	7.31
	250	3.00	2.00	1.90	2.57
	500	1.98	1.20	1.16	1.81
Dominant	100	5.90	6.20	6.47	8.20
	250	3.77	3.10	3.79	2.89
	500	2.61	2.03	2.08	2.00

Table VI

Empirical coverage of BILL when K is fixed at $\zeta \times (\text{prevalence})$, using 250 ASPs simulated under the single-locus models

Model	Prevalence	ζ				
		0.1	0.5	1.0	2.0	4.0
Recessive	0.045	0.970	0.975	0.960	0.955	0.940
Intermediate	0.058	0.970	0.970	0.950	0.940	0.915
Dominant	0.085	0.935	0.945	0.945	0.895	0.775