Two-locus Modeling of Asthma in a Hutterite Pedigree via Markov Chain Monte Carlo

Yuqun Luo, Shili Lin*, and Mark E. Irwin

Department of Statistics, The Ohio State University, Columbus, Ohio 43210

Bayesian Markov chain Monte Carlo (MCMC) segregation analysis for asthma was performed on the whole 1,544-member Hutterite pedigree. Heterogeneous and epistatic two-locus models and complex one-locus models were investigated, with trait loci postulated to be linked to markers in regions previously found to be possibly linked to asthma or atopy. The epistatic two-locus dominant-dominant model provided the best estimates, among the models investigated, in terms of prediction of population prevalence and relative risk for sibs of the affecteds.

Key words: complex pedigree, complex traits, epistasis, heterogeneity, Markov chain Monte Carlo, segregation analysis, two-locus models.

INTRODUCTION

Being one of the most common chronic childhood diseases in developed countries, asthma is genetically complex. Previous studies have suggested that, together with environmental factors, multiple loci might be involved in the etiology of asthma. The disease susceptibility loci may affect the trait independently (heterogeneity), or interactively (epistasis). Founder population offers many advantages in mapping complex traits such as asthma [Ober et al., 1998]. Large pedigrees such as the Hutterites may provide more information on the mode of inheritance if analyzed as a whole instead of being broken into smaller pieces. Also, it is essential for the segregation analysis of complex traits to make use of information from linked markers, since the number of parameters to be estimated is large. To our knowledge, such an analysis has not been attempted to date for a pedigree of such size (1,544-member) and complexity (many inbreeding and marriage loops) as the Hutterite pedigree.

* Address reprint request to Dr. Shili Lin, Department of Statistics, The Ohio State Univeristy, 1958 Neil Avenue, Columbus, OH 43210.

We propose a Bayesian Markov chain Monte Carlo approach that tries to identify, among one-locus and two-locus disease models, a plausible model for (partially) explaining asthma among the Hutterites. Information from two regions (3p24.2-22 and 5q23-31) that had been shown to be possibly linked to asthma [Ober et al., 1998] is incorporated in our study. We hope this analysis will shed some light on the mode of transmission for this complex trait, thereby providing models for further studies, including parametric linkage analysis.

MODELS AND METHODS

Models. We analyzed both one-locus and two-locus models, with the belief that there is potentially more than one locus involved in the etiology of asthma. The trait locus or loci is assumed to be diallelic, with high-risk and low-risk alleles denoted as *A* and *a* respectively (with *B* and *b* for the second locus of a two-locus model). For a one-locus model, the parameters to be estimated include the trait allele frequency, p, the map position of the trait locus relative to a known marker map, reflected as the recombination fraction between the trait and an adjacent marker locus, r, and the penetrances, f_0, f_1, and f_2, for genotypes *aa*, *Aa* and *AA*, respectively. For a two-locus model, where the two trait loci are assumed to be on two different chromosomes, we have two trait allele frequencies, p_1 and p_2, and the positions of the two trait loci relative to a known marker map, parameterized as two recombination fractions between the trait loci and the adjacent markers, r_1, r_2, one for each trait locus. The trait allele frequencies are all constrained to be less than 0.5. We investigated both epistatic and heterogeneous two-locus models, with the mode of inheritance postulated to be either dominant or recessive for each of the two trait loci. The naming of the classes of models reflects the model setup. For example, HDR refers to heterogeneous two-locus models with the first locus being dominant and the second being recessive. ERD is a class of epistatic models with the first trait locus being recessive and the second being dominant. The penetrance parameters for these two classes of models are listed in table 1. Penetrance parameters for other classes are similarly defined. We investigated seven classes of two-locus models: HDD, HDR, HRD, HRR, EDD, EDR and ERD. ERR was not investigated since the population prevalence is quite high for asthma among the Hutterites (11.7%); an ERR model could not have given rise to such a high prevalence. For example, even with a penetrance of 1 for susceptible genotypes and trait allele frequencies of 0.5, there can be at most a prevalence of 6.3% for an ERR model.

TABLE 1. Penetrance parameters for the HDR and ERD classes of two-locus models.

HDR	*BB*	*Bb*	*Bb*	ERD	*BB*	*Bb*	*bb*
AA	f *	f_1	f_1	*AA*	f	f	0
Aa	f_2	f_1	f_1	*Aa*	0	0	0
aa	f_2	0	0	*aa*	0	0	0

* $f = f_1 + f_2 - f_1 f_2$.

Markov chain Monte Carlo for Bayesian segregation analysis. Let θ be the vector of parameters to be estimated, which includes the trait allele frequencies, the map position of each trait locus on a known marker map, and the penetrances. Let d and M be the observed disease status and marker phenotypes, part of which may be missing. The terms in the distribution of interest, $\pi(\theta \mid d, M) \propto P(d, M \mid \theta)\, \pi(\theta)$, are hard to compute. However, we can simplify the computation greatly by augmenting the parameter space to

include some latent parameters, collectively denoted as G, which is composed of the marker descent graphs and the trait loci descent states [Sobel and Lange, 1996, Thompson, 1994]. Our target distribution is thus

$$\pi(\theta, G \mid d,M) \propto P(d, M \mid G)\pi(G \mid \theta)\, \pi(\theta). \qquad (1)$$

A combination of Metropolis-Hastings algorithms is used to generate the Markov chain, $\{(G^t, \theta^t): t = 1, 2, \ldots\}$, with the posterior distribution $\pi(\theta, G \mid d,M)$ as its stationary distribution. Components of θ are updated sequentially. Specifically, trait allele frequencies are updated with a uniform proposal on an interval of length $2l$ centered at the current values, where l is a pre-specified value. Some care should be taken when part of the interval is outside the range of all possible values. Penetrances and recombination fractions are updated in a similar fashion. The latent parameters are updated using a Metropolis algorithm, with symmetric proposal distributions. Random numbers of individuals are selected (from a Geometric distribution) to be updated for the trait descent state. If a founder is selected, a paternal or maternal allele will be switched to its opposite state. If the selected individual is a nonfounder, its inheritance vector will be changed. The descent graph approach proposed by Sobel and Lange [1996] is adopted to update marker descent graphs. A single update of all parameters in θ is referred to as a scan. After discarding an initial segment of the realizations to allow for convergence, the remaining realizations are then used for parameter estimation. Linkage equilibrium and Hardy-Weinberg equilibrium are assumed in our analysis.

Starting point. For any given marker locus, obtaining a starting genotype configuration (hereafter called starting point) that is compatible with the observed phenotypes demands nontrivial effort with the complexity of the Hutterite pedigree. A Gibbs sampling approach by Lin et al. [1993] is employed to propose a legal descent state, where each individual's genotype is updated in a random order, conditional on the neighboring individuals. The space of legal genotypic states is embedded into a larger space that allows penetrances for any heterozygous genotypes to be nonzero. All genotypic states in the larger space are ensured to be communicating under such a penetrance model, and the chain is stopped when a legal state is reached. For instance, suppose there are three alleles, $\{A, B, C\}$ for a marker, we set the probability of observing any phenotype in $\{AA, AC, BB, BC, CC\}$ given genotype AB to be $s = 0.01$, and the probability of observing AB to be 0.95. The tuning parameter s can be experimented with to speed up the search for a legal state. When a legal descent state is reached, the descent graph is then extracted to be used as a starting point for the MCMC segregation analysis.

RESULTS

Settings. Two regions, 3p24.2-22 and 5q23-31, are found to be possibly linked to asthma or atopy [Ober et al., 1998], and were chosen to be included in our analysis. Six markers, with three of them in or near 3p24.2-22, and the other three in or near 5q23-31, were investigated. Listed in table 2 is the information on these markers. We were able to find starting points for markers *M1*, *M2* and *M3* within a reasonable amount of computing time. So the analysis is based on these three markers.

Three possible settings of trait loci positions for each two-locus model were investigated. All three settings place the first trait locus (*T1*) within 25cM of *M1*. Setting 1 places the second trait locus (*T2*) between *M2* and *M3*. Setting 2 assumes the map on chromosome 5 to be *T2*-*M2*-*M3*, while setting 3 assumes the map to be *M2*-*M3*-*T2*. For complex one-locus models (OL), we investigated two settings. Setting 1 locates the trait

within 25cM of *M1* on chromosome 3, and setting 2 postulates that the trait is in the 25cM vicinity of *M3* on chromosome 5.

The starting values are 0.9 for allele frequencies and penetrances. A trait locus is started at the center of a marker interval if there is one, or at half of the maximum distance allowed from the marker if there is only one marker. Simulation studies indicate that the results generally are not sensitive to the choice of starting values.

Table 2. Information about markers in two regions (3p24.2-22, 5q23-31).

Marker	D3S1766	D5S2501	D5S1480	D3S2432	D3S1768	D5S1505
Notation	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
Map position	79.00	117.00	147.00	58.00	62.00	130.00
Heterozygosity	.709	.720	.793	.819	.701	.822

Estimates. A Markov chain was run for each combination of model and setting. It took approximately 207 hours (8.5 days) to run 2 x 10^7 scans using a Sun Ultra Enterprise 450 under Solaris 7**.** We found that 2 x 10^7 scans are sufficient for most chains to converge and provide valid estimates. We discarded the initial portion (1 x 10^7 scans for most cases) of each chain to allow the chain to navigate into the desired stationary distribution. The remaining scans were used for inferences. The batch mean method with 50 batches was used to estimate the standard error for each run.

We base our model selection on the predictions of two characteristics: the population prevalence (K), and the relative risk for sibs of an asthmatic subject (λ_s). An estimate of K = 11.7% for strict asthma was obtained from the data in Ober et al. [1998]. While relative risk cannot be estimated from the Hutterite data (Ober, personal communication), λ_s = 2.58 has been reported in Wjst et al. [1999] for a German population.

The estimated population prevalence and relative risks for sibs are reported in tables 3 and 4 for various one-locus and two-locus models. The class of EDD models did very well in predicting the population prevalence, but the estimates for λ_s are much lower than the one reported for the German population for settings 2 and 3. The two one-locus models over predicted the population prevalence, while EDR and ERD both under estimated the population prevalence and over estimated λ_s. For heterogeneous models, HRR under estimated both characteristics, while the other models tended to over estimate K. Among the EDD models, setting 1 provided the highest estimate of the relative risks for the sibs. Recall that this setting postulated the trait locus to be between *M2* and *M3* on chromosome 5, which turns out to be a region in which Greenwood et al. found some evidence for linkage [in this volume].

Table 5 gives the point estimates and standard errors of the parameters for the three settings of the EDD class of models. The estimates for settings 2 and 3, EDD2 and EDD3, where the trait locus on chromosome 5 is assumed to be on either side of the map *M2-M3*, are alike. They all have very high trait allele frequencies and low penetrance estimates compared to those of setting 1, where the trait locus is assumed to be lying between *M2* and *M3*.

Table 3. Estimates of population prevalence and relative risk for sibs under epistasis two-locus models and complex one-locus models.

Model	EDD		EDR		ERD		OL	
Setting	K	λ_s	K	λ_s	K	λ_s	K	λ_s
1	10.6%	2.23	2.9%	5.52	3.00%	5.37	18.4%	1.10
2	12.0%	1.50	6.1%	3.18	5.00%	3.28	18.2%	1.20
3	11.8%	1.54	5.6%	2.65	5.10%	3.38	-	-

Table 4. Estimates of population prevalence and relative risk for sibs under heterogeneous two-locus models.

Model	HDD		HDR		HRD		HRR	
Setting	K	λ_s	K	λ_s	K	λ_s	K	λ_s
1	7.4%	3.22	12.4%	1.45	21.9%	1.45	7.70%	2.04
2	21.4%	1.31	15.5%	1.62	19.5%	1.29	7.80%	1.92
3	12.2%	1.13	11.9%	1.39	17.1%	1.28	7.50%	1.99

Table 5. Estimates and standard errors of the parameters under the epistasis dominant-dominant (EDD) models.

Parameter	p_1	p_2	r_1	r_2	f
EDD1	.294± .011	.267± .012	.069± .008	.033± .005	.457± .004
EDD2	.428± .004	.407± .005	.169± .008	.190± .006	.276± .005
EDD3	.402± .006	.403± .005	.210± .004	.160± .007	.285± .005

Convergence Diagnostics. Figure 1 plots the realizations and estimates for some of the parameters under model and setting EDD1. Estimates for p_1 and f seem to stabilize after 1.5 x 10^7 scans, but there seems to be some correlation for the realizations of r_1. Overall, the chain appears to have moved adequately around the parameter space. Diagnostic plots on the other models and settings show similar patterns.

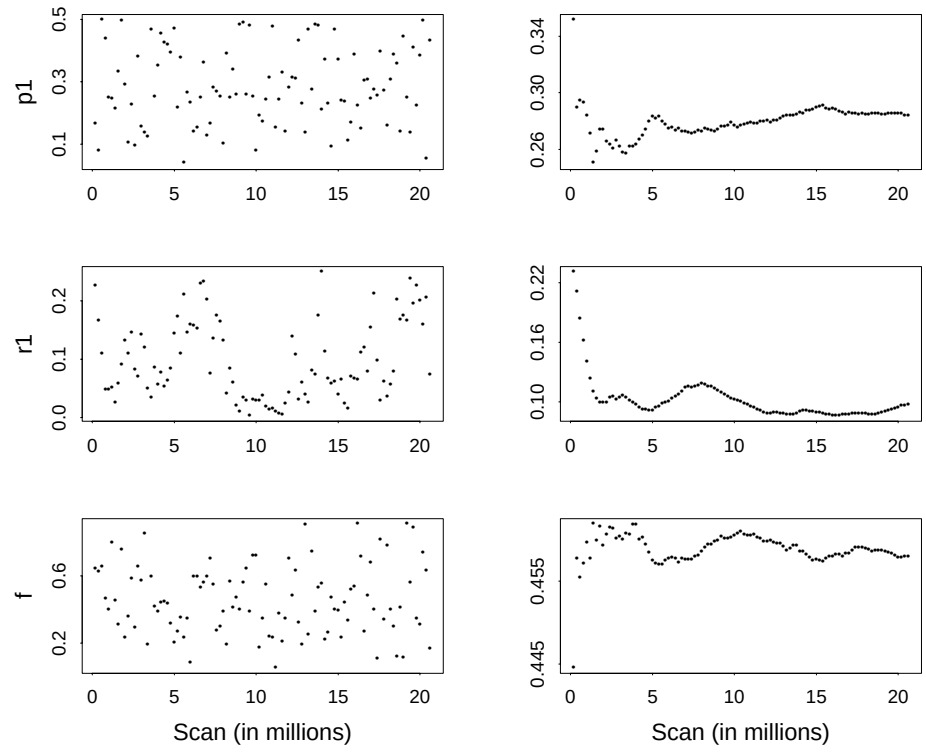

Figure 1. Diagnostic plots for model EDD1. Left column gives the realizations for every 200,000 scans. Right column gives the parameter estimates (cumulative averages). Plots for p_2 and r_2 are not shown, but reveal similar behaviors.

DISCUSSION

A Bayesian Markov chain Monte Carlo segregation analysis was performed on the whole 1,544-member Hutterite pedigree in the hope of gaining some understanding of the underlying genetic model for asthma. Information from markers previously identified to be possibly linked to the trait was incorporated in such an endeavor. Two-locus heterogeneous and epistatic models, as well as one-locus complex models were considered. It appears that the epstatic dominant-dominant model does the best, among the models considered, in terms of providing estimates closest to the population prevalence among the Hutterites. Furthermore, the setting that hypothesized the trait locus on chromosome 5 to be between markers *M2* and *M3* seemed to provide the most reasonable estimates of the parameters. More sophisticated model selection criteria and algorithms are under investigation.

Other markers (*M4, M5, M6* in table 1) were also considered but not included in our analysis. While it took about one hour to obtain a legal starting point for the three markers used in the analysis, a legal state could not be found for any of the above markers after more than 50 hours and various efforts in fine-tuning the program. For marker *M6* (D5S1505), for instance, the chain was unable to assign a legal genotype to only one individual. Thus, if we had treated this individual's phenotype at this marker locus as missing, a legal starting point could have been obtained. This suggests that the Mendelian errors present in the data (email correspondence from Vanessa Olmo) might have been the cause of our failing to find legal genotypic configurations for these markers. It is also possible that the chain might have gotten stuck in a local mode and had not escaped during the time period the chain was allowed to run.

Ascertainment correction was not incorporated in our analysis, as the effect of ascertainment biases is not as severe in the analysis of a single large pedigree as in smaller pedigrees with various ascertainment schemes.

ACKNOWLEDGEMENTS

This research was supported in part by NSF grant DMS-9971770 (to S.L.). Yuqun Luo was supported on a research assistantship by Dr. Fred A. Wright from NIH grant GM58934.

REFERENCES

Lin S, Thompson EA, Wijsman E (1993): Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. IMA J Math Appl in Med & Biol 10: 1-17.

Ober C, Cox NJ, Abney M, Rienzo AD, Lander ES, Changyaleket B, Gidley H, Kurtz B, Lee J, Nance M, Pettersson A, Prescott J, Richardson A, Schlenker E, Summerhill E, Willadsen S, Parry R, CSGA (1998): Genome-wide search for asthma susceptibility loci in a founder population. Hum Mol Genet 7: 1393-1398.

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

Thompson EA (1994): Monte Carlo likelihood in genetic mapping. Statistical Science 9: 355-366.

Wjst M, Fischer G, Immervoll T, Jung M, Saar K, Rueschendorf F, Reis A, Ulbrecht M, Gomolka M, Weiss EH, Jaeger L, Nickel R, Richter K, Kjellman NIM, Griese M, von Berg A, Gappa M, Riedel F, Boehle M, von Koningsbruggen S, Schoberth P, Szczepanski R, Dorsch W, Silbermann M, Loesgen S, Scholz M, Bickeboller H, Wichmann HE (1999): A genome-wide search for linkage to asthma. Genomics 58: 1-8.