Theor Appl Genet (2006) 113:921–929
DOI 10.1007/s00122-006-0351-x

ORIGINAL PAPER

Major gene detection for fusiform rust resistance using Bayesian complex segregation analysis in loblolly pine

Hua Li · Sujit Ghosh · Henry Amerson · Bailian Li

Received: 22 November 2005 / Accepted: 21 June 2006 / Published online: 29 July 2006

Abstract The presence of major genes affecting rust resistance of loblolly pine was investigated in a progeny population that was generated with a half-diallel mating of six parents. A Bayesian complex segregation analysis was used to make inference about a mixed inheritance model (MIM) that included polygenic effects and a single major gene effect. Marginalizations were achieved by using Gibbs sampler. A parent block sampling by which genotypes of a parent and its offspring were sampled jointly was implemented to improve mixing. The MIM was compared with a pure polygenic model (PM) using Bayes factor. Results showed that the MIM was a better model to explain the inheritance of rust resistance than the pure PM in the diallel population. A large major gene variance component estimate (> 50% of total variance), indicated the existence of major genes for rust resistance in the studied loblolly pine population. Based on estimations of parental genotypes, it appears that there may be two or more major genes affecting disease phenotypes in this diallel population.

Communicated by J. L. Jannink.

H. Li · H. Amerson · B. Li (✉)
Department of Forestry and Environmental Resources,
North Carolina State University, Campus Box 8002,
Raleigh, NC 27695, USA
e-mail: Bailian_Li@ncsu.edu

S. Ghosh
Department of Statistics, North Carolina State University,
Raleigh, NC 27695, USA

Introduction

Fusiform rust, a disease of southern pines caused by *Cronartium quercuum* (Berk.) Miyabe ex Shirai f.sp. *fusiforme*, continues to be the most economically important tree disease in commercial forests of the southern U.S. Deployment of genetically resistant trees is viewed as environmentally friendly and the only feasible means of control (Kinloch and Walkinshaw 1991). Studies with loblolly and slash pine suggest that fusiform rust resistance in pines, at least in part, is the result of major resistance genes in the host interacting with avirulence genes in the pathogen (Carson and Carson 1989: Kinloch and Walkinshaw 1991; Nelson et al. 1993; Wilcox et al. 1996; Kuhlman et al. 1997; Amerson et al. 1997; Stelzer et al. 1997, 1999) in gene-for-gene fashion (Flor 1956). Non-infected trees in high hazard rust areas may be candidate carriers for major resistance genes and these trees could potentially be used for rust resistance breeding. In a tree breeding program, phenotypic data including growth, disease resistance and wood properties from progeny tests could be easily accessed by researchers. By utilizing these available breeding materials and genetic testing information, and by identifying major segregating genes with economic values (like rust resistance), breeders may make more effective decisions regarding stock management, enhancement of productivity or detection for quantitative trait loci (QTL). The purpose of this study is to provide putative genotypes of parents in the diallel mating design for the trait of rust resistance by statistical inference using easily accessed phenotypic data. The methods presented here for recognizing major genes could have broad application.

To our knowledge, complex segregation analysis is considered to be the most powerful statistical test for major gene detection. It was proposed by Elston and Stewart (1971) and Morton and MacLean (1974), and was further developed by geneticists. Bayesian complex segregation analysis was first introduced for the animal model by Hoeschele (1988). Janss et al. (1995, 1997) applied Gibbs sampling within a Bayesian framework to search for a major gene affecting meat quality traits in a crossed F2 population based on a mixed inheritance model (MIM) in animal breeding. Another application was based on an investigation for major genes affecting carcass traits in Japanese black cattle populations (Miyake et al. 1999). In forest tree breeding, a major gene affecting height in loblolly pine was detected in some half-diallel progeny populations using Bayesian complex segregation analysis (Zeng et al. 2004).

Most complex segregation analyses have been used to study traits with continuous phenotypic measurement. However, many traits in animal and plant breeding, such as survival scores, or resistance to insects and diseases that are postulated to be continuously inherited are categorically scored. A widely used model for genetic analysis of categorical data is based on the threshold liability concept, first introduced by Wright (1934). In the threshold model, one assumes that there exists a latent or underlying variable (liability) that has a continuous distribution. The threshold concept was applied to complex segregation analysis for a binary trait by Thaller et al. (1996a, b). In their study, numerical integration methods were used to determine the mode of inheritance for two traits in swine. Albert and Chib (1993) developed Bayesian inference and used Gibbs sampling algorithm to obtain parameter estimates based on the posterior distribution. They used latent variables within a data augmentation framework that lead to a computationally simple strategy. Sorensen et al. (1995) applied the above methodology to estimate genetic parameters of the animal model.

The major goal of this study was to detect major rust resistance genes using a Bayesian complex segregation model for binary data in a half-diallel population. We describe the phenotypic data and present statistical models in Materials and methods. An efficient Gibbs sampling algorithm is proposed to obtain parameter estimates. In particular, a parent blocking strategy is implemented to improve the mixing of the Gibbs iterates. The proposed method is illustrated by major rust-resistance gene detection for a loblolly pine population.

Materials and methods

Phenotypic data

The progeny population of loblolly pine was generated from a half-diallel mating of six parents with no selfing or reciprocal crosses, and thus there were 15 full-sib families in this diallel population. All six parents were selected from the upper piedmont of Alabama in plantation forests, and were rust-free at the time of selection. The test was established using a randomized complete block design with six blocks and six trees/full-sib family in each block. The total number of trees that were evaluated in this experiment was 540, with 36 trees for each cross. Rust disease was the result of natural infection and was recorded for six consecutive years, starting from year 3. Disease status (disease phenotype) was assessed as no gall/no disease = 1 or gall/disease = 0 for each tree based on the 6 years of measurements. If a tree was recorded as dead due to the rust disease, that tree in subsequent years was counted as a diseased tree (gall) instead of a missing value. Table 1 shows non-diseased rates for each cross in this diallel. We propose two classes of statistical models to annotate the data.

Statistical model

Mixed inheritance model for binary response

In a MIM for diallel mating, it is assumed that the rust resistance is influenced by a single major gene and polygenes and further it is assumed that the major gene and polygene effects are additive. Block effects are very small, so we do not include block effects in our model. The single locus is assumed to be a diallellic locus with Mendelian transmission probabilities. Assuming that the base population, where that parents are selected, is in Hardy–Weinberg and linkage equilibrium, *R*/*r* denote alleles and *f* denotes the frequency of the favorable allele *R*. Three genotypes, *RR*, *Rr* (= *rR*), *rr* are denoted by 2, 1, and 0, respectively. The polygenic effects include general combining ability (GCA) caused by additive polygenic effects and specific combining ability (SCA) caused by dominant polygenic effects. The phenotypic measurement y_i is recorded as 1 (non-diseased) and 0 (diseased). The y_i has a Bernoulli distribution with probability of rust-free p_i. Assume that,

$$p_i = \Pr(y_i = 1) = \Phi(H_i^{\mathrm{T}}\theta),$$

where Φ is a standard normal cumulative distribution function (probit model), H_i denotes a vector of

Table 1 Non-diseased rates of progeny trees for each full sib-family in the six-parent (A–F) half-diallel mating design used in this study

Female/male	A	B	C	D	E	F
B	0.39					
C	0.31	0.15				
D	0.56	0.22	0.09			
E	0.63	0.19	0.14	0.22		
F	0.17	0.00	0.03	0.09	0.00	
	0.41	0.19	0.14	0.24	0.24	0.06

The bottom row shows the non-diseased rates for individual parents

explanatory variables and θ denotes a vector of parameters. Equivalently, this model can be represented by introducing a latent variable U_i. Specifically, if $Ui \sim N(H_i^{\mathrm{T}}\theta, 1)$, then we define $y_i = 1$ if $U_i > 0$ and $y_i = 0$ if $Ui \leq 0$. Notice that U_is are not observed; however, they are used to simplify the likelihood. Define $H = [H_1^{\mathrm{T}},..., H_n^{\mathrm{T}}]$ to be the design matrix and we write,

$$H\theta = \mu + X\beta + wLm,$$

where μ is the overall mean, X is the incidence matrix of the polygenic effects for all progeny; β is a vector of random polygenic effects, e.g., GCA (additive genetic effects, g_1, g_2, ...) and SCA (dominant genetic effects, s_1, s_2, ...). Specifically, $\beta = (g_1, g_2, ..., g_6, s_1, s_2,..., s_{15})^{\mathrm{T}}$;w would be an $n \times 3$ design matrix of major genes at a single locus and L would be a 3×2 indicator matrix of the major gene effects for major genotypes, where $L = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}$. Let m be a vector of major gene effects, $m = (a,d)$, where a is an additive major genotypic effect, d is a dominance major genotypic effect. The product wLm generates three possible genotypes of progeny ($a = RR$,$d = Rr/rR$, $- a = rr$). The likelihood of θ is given by,

$$p(Y|\theta) \propto \prod_{i=1}^{n} (\Phi(H_i^{\mathrm{T}}\theta))^{y_i}(1 - \Phi(H_i^{\mathrm{T}}\theta))^{1-y_i}.$$

Let $p(\theta)$ denote the prior density of θ. According to Bayes theorem, the joint posterior density of the parameters and latent variable $U = (U_1, ..., U_n)$ given the data $Y = (y_1, ..., y_n)$ is given by

$$p(\theta, U|Y) \propto p(Y|U,\theta)p(U|\theta)p(\theta),$$

where $p(Y|U, \theta)$ denotes the conditional density of Y given U and θ, and $p(U \mid \theta)$ denotes the conditional density of U given θ.

The overall mean μ, additive major gene effect (a), and dominant major gene effect (d) are given independent normal priors, i.e., $\mu \sim N(0, k_1^2)$, $a \sim N(0, k_2^2)$, and $d \sim N(0, k_3^2)$. k_is are chosen based on the simulation studies in Zeng et al. (2004) as $k_i = 4$, for $i = 1, 2, 3$. Polygenic effects are assumed to be random. The GCA effects of six parents are assumed to be identically independently normally distributed, i.e., g_1, ... $g_6 \sim N(0, \sigma_g^2)$, where σ_g^2 is assumed to have an inverted gamma distribution, $IG(\gamma_1, v_1)$, with hyper-parameters γ_1, v_1. Similarly, the SCA effects of 15 crosses are assumed to be identically independently normally distributed, i.e., s_1 ..., $s_{15} \sim N(0, \sigma_s^2)$, and $\sigma_s^2 \sim$ IG(γ_2, v_2). We used $\gamma_i = 2$, and $v_i = (\gamma_i - 1) \times \hat{\sigma}_i$, where $\hat{\sigma}_i$ are estimates from frequentist approach. A conjugate beta prior is assumed for the allele frequency f, i.e., $f \sim Beta(\alpha_f, \beta_f)$, where $\alpha_f = \beta_f = 1$ are chosen to express prior ignorance. Assuming that the progeny population is in Hardy–Weinberg equilibrium, the distribution of progeny genotype [$p(w)$] is obtained from the parental genotypes (Wp) following Mendelian transmission probabilities given the favorable gene frequency (f) in the base population. Assuming that the major gene effects and polygenic effects are independent, the joint posterior distribution could be easily derived from the likelihood function and above prior distributions.

The Gibbs sampler for MIM

A Gibbs sampling algorithm can be used to generate samples from the posterior distribution of target parameters *theta* and variance components. However, high dependence of parent and progeny genotypes causes slow convergence of a naive Gibbs sampler and hence a parent blocking strategy is needed. The parent blocking strategy of Zeng et al. (2004) for a complex segregation analysis in the diallel mating design of forest trees, can be adapted here for binary data. In the parent blocking method, the genotypes of a parent and its offspring are blocked and updated simultaneously. The full conditional distributions for each unknown parameter are derived in order to implement the Gibbs sampler.

Conditional on observed values, Ys, and all parameters, the conditional distribution of latent variables U_is has a standard distributional form. In particular, it follows that,

$$U_i|Y,\mu,g,s,a,d,w,Wp,f,\sigma_g^2,\sigma_s^2 \sim N(\mu+X\beta+wLm,1)$$
$$\text{truncated at the left by 0 if } y_i=1$$
$$U_i|Y,\mu,g,s,a,d,w,Wp,f,\sigma_g^2,\sigma_s^2 \sim N(\mu+X\beta+wLm,1)$$
$$\text{truncated at the right by 0 if } y_i=0. \quad (1)$$

In the parent blocking, the joint conditional distribution of a parent and all its offspring is proportional to the product of the distribution of parent i's genotype marginalized over all progeny and the joint distribution of its offspring given all parent's genotypes, i.e.,

$$p(Wp_i, w_{pi(1)} \ldots w_{pi(o_i)}|Wp_{-i}, w_{-i(k)}, U, Y, \mu, a, d, g, s, \sigma_g^2, \sigma_s^2, f) \propto p(Wp_i|Wp_{-i}, w_{-i(k)}, U, Y, \mu, a, d, g, s, \sigma_g^2, \sigma_s^2, f)\, p(w_{pi(1)}, \ldots w_{pi(o_i)}|Wp, w_{-i(k)}, U, Y, \mu, a, d, g, s, \sigma_g^2, \sigma_s^2, f) \quad (2)$$

where o_i denotes the number of offspring of parent i, and the offspring are indexed by $i(1)$, $i(2)$, ..., $i(o_i)$. $Wp_{-\ i}$ denotes all other parents except for parent Wp_i. The marginalized conditional distribution of a parent and the conditional distribution of its offspring could be easily derived and used for calculation. In the parent blocking strategy, each offspring is updated twice in each cycle.

The full conditional distribution for allele frequency f is a Beta distribution with parameters ($\alpha_f + n_1$, $\beta_f + n_2$), where n_1 and n_2 are the number of R and r alleles in the base population, respectively.

$$f|Y, U, \mu, a, d, g, s, w, Wp, \sigma_g^2, \sigma_s^2 \propto f^{\alpha_f+n_1-1}(1-f)^{\beta_f+n_2-1}. \quad (3)$$

Let $\eta = (\mu, a, d, g_1, \ldots g_6, s_1, \ldots s_{15})^{\mathrm{T}}_{p\times 1}$, where $p = 24$. The full conditional distribution of $\eta_j (j = 1\ldots\ p)$ is normally distributed, i.e.,

$$\eta_j|Y, U, \eta_{-j}, w, Wp, f, \sigma_g^2, \sigma_s^2 \sim N(\tilde{\eta}_j, \sigma^2_{\tilde{\eta}_j}), \quad (4)$$

where $\tilde{\eta}_j = \frac{\sum_{k=1}^n H_{kj}(U_k - \sum_{r=1,r\neq j}^p H_{kr}\eta_r)}{\sum_{k=1}^n H_{kj}^2 + \frac{1}{\sigma_j^2}}, \sigma^2_{\tilde{\eta}_j} = \frac{1}{\sum_{k=1}^n H_{kj}^2 + \frac{1}{\sigma_j^2}}$,

where H_{kj} is the kth row and jth column element of the matrix H, and σ_j^2 is the corresponding variance for η_j. The full condition distributions for GCA and SCA variances are given by

$$\sigma_g^2|Y,U,\mu,a,d,g,s,\sigma_s^2,w,Wp,f \sim \mathrm{IG}\left(\frac{n_g}{2}+\gamma_1, \frac{\sum_{i=1}^{n_g} g_i^2}{2}+v_1\right)$$
$$\sigma_s^2|Y,U,\mu,a,d,g,s,\sigma_g^2,w,Wp,f \sim \mathrm{IG}\left(\frac{n_s}{2}+\gamma_2, \frac{\sum_{j=1}^{n_s} s_j^2}{2}+v_2\right). \quad (5)$$

In order to initiate the Gibbs sampler, the starting values of parental genotypes (Wp) were generated from the initial values of the favorable allele frequency (f) assuming Hardy–Weinberg equilibrium in the population. Progeny genotypes were generated based on Mendelian transmission probability given the initial values of related parental genotypes. The other parameters, σ_g^2, σ_s^2 and η are also initiated using reasonable guesses from their support. Samples were drawn using the following scheme, starting at time t = 0:

1. Sample latent variables $U_i^{(t)}$ given $w^{(t)}$, $\eta^{(t)}$, $\sigma_g^{2(t)}$, $\sigma_s^{2(t)}$ using Eq. 1.
2. Sample genotypes $w^{(t+1)}$ from by parent blocking using Eq. 2.
3. Sample allele frequency $f^{(t+1)}$ using Eq. 3.
4. Sample location parameters $\eta^{(t+1)}$ using Eq. 4.
5. Sample variance components $\sigma_g^{2(t+1)}$ and $\sigma_s^{2(t+1)}$ from their full conditional distributions using Eq. 5.
6. Repeat steps 1–5 until the sufficient samples are generated to achieve convergence to the stationary distribution.

Three independent chains were generated with dispersed sets of initial values. Bayesian Output Analysis (BOA version 1.0.0, Smith 2001, http://www.public-health.uiowa.edu/boa/) was used for convergence diagnostics and posterior distribution summarization. Convergence of a single chain was checked by Raftery and Lewis dependence factors (Raftery et al. 1992); mixing of multiple chains was checked by Gelman and Rubin (Gelman et al. 1992) shrink factors.

The negative value of the additive effect (a) was artificially changed to be positive with the consideration of allele R being the favorable allele. The predicted parental genotypes were also changed along with the additive effect for consistency purpose. Major gene variance (σ_m^2) was calculated as the sum of additive major gene variance (σ^2_{ma}) dominant major gene variance (σ^2_{md}), i.e.,

$$\sigma^2_{\mathrm{m}} = \sigma^2_{\mathrm{ma}} + \sigma^2_{\mathrm{md}} = 2f(1-f)[(1-2f)d+a]^2 + [2f(1-f)d]^2.$$

The total variance was calculated as the sum of the major gene variance and polygenic variance, i.e.,

$$\sigma^2_{\mathrm{T}} = 2\sigma_g^2 + \sigma_s^2 + \sigma^2_{\mathrm{m}}.$$

Polygenic model

The polygenic model (PM) is the subset of the full model generated by suppressing the major gene effect part. A similar Gibbs sampler algorithm (but much simpler) was used to obtain parameter estimates. The prior distributions for parameters in the PM were the same as in the MIM. The joint posterior distribution and full conditional distributions were derived using Bayes theorem. The updating scheme for parameters was also similar except that it did not involve updating steps for genotypes, gene frequency and additive and dominant effects.

Bayes factor for model comparison

Bayes factor (BF)(Robert 2001) was approximated from the MCMC output using the following formula:

$$B = \frac{p(Y|M_{\mathrm{MIM}})}{p(Y|M_{\mathrm{PM}})} \approx \frac{m_1/\sum_{i=1}^{m_1} \frac{1}{p(\theta_i|Y,M_{\mathrm{MIM}})}}{m_2/\sum_{i=1}^{m_2} \frac{1}{p(\theta_i|Y,M_{\mathrm{PM}})}},$$

where m_1 and m_2 are lengths of Markov chains under each model, and $p(\theta \mid Y,M)$ denote(s) the likelihood function under the corresponding model. It is well known that the above formula might not be a stable one as it involves harmonic means. However, for our data, we find the formula to be rather stable in producing similar values for different choices of m_1 and m_2. We have used $m_1 = m_2 = 75{,}000$.

Results

Convergence diagnostics of Markov chains

For both models, trial chains were run to determine suitable starting values for a burn-in period and a thinning factor. From the trial chains, we decided that the random samples for all parameters could be obtained from 750,000 cycles of the chain, with discarding of the first 100,000 samples and using a lag of 1,000 cycles. Three Gibbs chains with independent starting values produced 19,500 final samples. For example, the initial values of gene frequency for three chains were taken as 0.25, 0.5 and 0.75, which covered the support of this parameter very well. Raftery and Lewis dependence factor and Gelman and Rubin shrink factor have been presented in Table 2. Most dependence factors of single chains were less than 5 and the 0.975 quantiles of corrected scale shrink factors were less than 1.2, which indicated that our samples were adequate for convergence and mixing.

Parameter estimates from the mixed inheritance model

Estimated posterior means and standard deviations of variance components including GCA variance (σ_g^2), SCA variance (σ_s^2) and major gene variance (σ_{m}^2) are shown in Table 3a, and posterior densities are plotted in Fig. 1. The additive polygenic variance (GCA variance) is similar to the dominant polygenic variance (SCA variance), while the major gene additive variance (σ_{ma}^2) is much larger than the major gene dominant variance (σ_{md}^2). The ratio of the total major gene variance to the total genetic variance reached 0.55. The ratio of the major gene additive variance to the total additive genetic variance was 0.61. These high percentages for major gene variance components suggest the presence of at least one major resistance gene segregating in this half-diallel loblolly pine population.

In terms of parental genotype estimates as in Table 3b, parent A was estimated as dominant homozygous (*RR*) with a probability of 0.63, suggesting that parent A is probably carrying two *R* alleles. Parent F was estimated as recessive homozygous (*rr*) with a probability of 0.85, suggesting parent F is likely to have no resistance allele. Parent A is almost certain to have at least one resistance allele given the probability of genotypes *RR* and *Rr* being 0.97. Other parents were estimated as heterozygous with probabilities ranging from 0.77 to 0.89 (Table 3b).

In the MIM, the parental GCA effect serves as polygenic additive effect, accounting for an important variance component. Although parent trees were rust free when selected, progeny of six parents showed different resistance levels in our experiment. The rust disease rates ranged from 59% (for progeny of parent A) to 94% (for progeny of parent F) (Table 1). Six parent's GCA predictions under the MIM showed that

Table 2 Convergence diagnostics (dependence factor and shrink factor) of the Gibbs sampler for additive effect (*a*), dominant effect (*d*), gene frequency (*f*), GCA variance (σ_g^2) and SCA variance (σ_s^2) in the mixed inheritance model (MIM)

		a	*d*	*f*	σ_g^2	σ_s^2
Shrink factor	Estimated	1.03	1.08	1.03	1.00	1.00
	0.975	1.07	1.23	1.06	1.00	1.00
Dependence factor	Chain 1	7.55	3.36	7.80	0.97	1.01
	Chain 2	3.25	2.45	3.85	1.04	0.97
	Chain 3	4.23	3.69	3.92	0.97	1.03

There are three independent chains

Table 3 Estimated posterior statistics from the MIM

(a) Marginal posterior means and standard deviations of variance

σ_g^2	σ_s^2	σ_{ma}^2	σ_{md}^2	σ_{m}^2	σ_t^2	$\sigma_{\mathrm{m}}^2/\sigma_t^2$
1.6 ± 1.4	1.5 ± 1.0	5.0 ± 4.1	0.7 ± 0.9	5.7 ± 4.2	10.4 ± 3.1	55%

(b) Estimated parental major gene genotypes

	Parent A	Parent B	Parent C	Parent D	Parent E	Parent F
Genotypes	*RR*	*Rr*	*Rr*	*Rr*	*Rr*	*rr*
Probability	0.63	0.79	0.89	0.89	0.77	0.85

parent A had the largest GCA effect, while parent F had the smallest GCA effect (Fig. 2). Progeny of parent A were highly resistant in the field, not only because parent A carries major genes but also it has high levels of polygenic resistance. Apparently, GCA effects are associated with the favorable resistance allele. The predicted high GCA effect and the high probability of parent A carrying at least one resistance allele suggest that the parent A is the most likely candidate among the six parents for further investigations to examine major resistance gene(s).

Model comparison with polygenic model

The PM that has a lesser number of parameters with polygenic additive effect (GCA) and polygenic dominant effect (SCA) was more easily fitted using the Gibbs sampler. The Gibbs chains of the PM were run 500,000 cycles with much faster speed than that of the MIM. The first 100,000 samples were discarded and the thinning factor was taken as 500 cycles. Again, three independent Gibbs chains were run and thus produced 2,400 samples. Single chains and multiple chains mixed well for all genetic parameters according to the dependence factors and corrected shrink factors. Enormous BFs (3,000) gave strong evidence for the MIM.

Discussion

The Bayesian complex segregation analysis with a block Gibbs sampler was developed for binary data in this study. The method was illustrated for the MIM and a PM in a half-diallel mating design for assessment of fusiform rust disease resistance in loblolly pine. Results of complex segregation analysis showed that the MIM with a major gene effect and polygenic effects fits the data better than the PM, as asserted by Bayes factor. In the MIM, > 50% of the total variance was estimated to be major gene variance, indicating the existence of major genes. Under the MIM, parent A was predicted

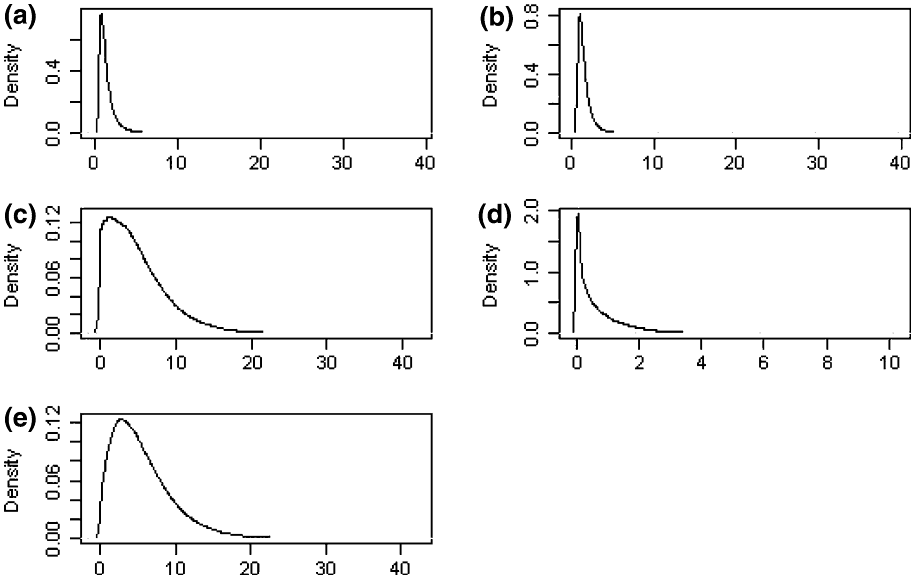

Fig. 1 Posterior density plots for estimated parameters. **a** σ_g^2, **b** σ_s^2, **c** σ_{ma}^2, **d** σ_{md}^2 and**e** σ_{m}^2

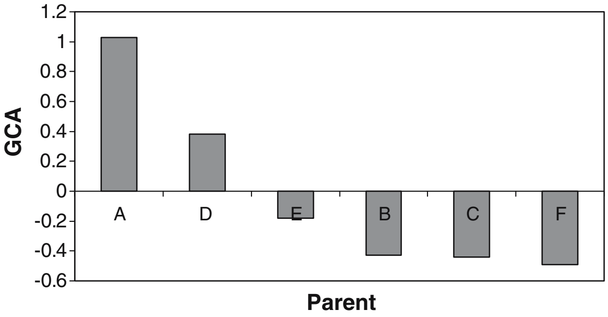

Fig. 2 Parental GCA predictions for six parents (A–F) by the mixed inheritance model. The GCA was estimated as the deviation from the population mean

to have the highest GCA, with parent F having the lowest GCA. The estimated genotypes of parents showed that five out of six parents have at least one dominant allele with high probabilities (> 70%). Among these five parents, parent A was estimated to carry at least one dominant allele with probability larger than 0.9. Further, the estimated genotype of this parent was dominant homozygous (*RR*) with largest probability (0.63) among the three genotypes. The estimated genotype of parent F was recessive homozygous (rr) with the largest probability among three genotypes. This result suggests that GCA prediction could be associated with estimated genotypes with the favorable allele. That is, high GCA and high probability of carrying dominate allele in parent A meant that this parent would be most likely to carry major genes for rust resistance.

The mixed model used here assumes only one major gene and while these analyses strongly support the existence of major gene effects for rust resistance in this diallel loblolly pine population, this does not exclude the possibility of two or more major genes in the diallel. Simulation studies showed that major gene effects could be recognized using the MIM with a single gene assumption when actually there was more than one major gene affecting the phenotypes (Zeng 2000). The first one or two major genes determined the parental genotype estimations and high GCA estimates, plus the genotype estimates with the favorable allele were always associated with a good parent. With only a single major gene included in the MIM, it is impossible to distinguish between the effect of a single locus and the effects of two or more independently acting loci with Mendelian transmission patterns. If there are multiple genes, Zeng (2000) postulated that the major gene effects would be detected as if they were a single locus with an allele frequency and effect equaling to the sum of several alleles. Allowing for the impact of pathogen virulence variation in field populations of the fungus for the various resistance genes postulated, a somewhat similar situation could be expected for multiple fusiform rust resistance genes, regulating gall presence versus absence.

Wilcox (1995) investigating the same field diallel used in this current study recognized RAPD (Random Amplified Polymorphic DNA) markers A11_400 and A19_560, inherited from Parent A, as being strongly associated with fusiform rust resistance (i.e., disease phenotype, gall presence vs. absence). The two markers were not linked with each other suggesting at least one heterozygous resistance locus (flanked by the markers) or perhaps two heterozygous resistance loci in Parent A (Wilcox 1995). Amerson and coworkers (unpublished data) have molecular marker mapped a heterozygous resistance locus termed (*Fr2*) in parent A, that conferred resistance (assessed as gall presence vs. absence) against a single spore isolate of the fusiform rust fungus. However, analysis of Wilcox's family A diallel samples (part of the same diallel used in the current study) by Amerson (unpublished data) for an *Fr2* linked RAPD marker (AK6_850, inherited from Parent A), revealed that the marker was not significantly associated with field disease phenotype, nor was marker AK6_850 linked with either A11_400 or A19_560. Given the evidence of Wilcox (1995) for one or more resistance loci associated with disease phenotype in the diallel, the lack of association for Fr2 marker AK6_850 with either disease phenotype or the markers of Wilcox in the same diallel samples and the simulation study noted by Zeng (2000), it is very likely that Parent A has multiple resistance genes. In this study, the estimated high frequency of the favorable allele might be the virulence impacted sum of allele frequencies for several loci. With the current model, the above hypothesis was not verified. However, the method presented here could be easily extended to a mixed model with two major genes that essentially involves more complex genotype configurations.

By implementing Bayesian based segregation analysis for major disease resistance gene detection, the probability distributions of parental genotypes were directly obtained from the output of the Gibbs sampler and the putative parental genotypes for major genes were derived. The use of latent variables with the Gibbs sampler simplified complicated computations. Further, this study applied a parent blocking strategy (Zeng et al. 2004) into the Gibbs sampler updating scheme for the half-diallel mating design to improve mixing and eliminate the slow convergence due to the dependency of parents and progeny in the MIM. The easy access and low cost of phenotypic measurements, compared with genotypic data collection, make the method described in this study particularly useful as a first step in disease resistance gene detection with binary phenotypic measurements, and the method can be easily adapted to other genetic analyses for categorical data.

In theory, the irreducibility of the Markov Chain is satisfied under mild conditions (e.g., see Lemma 6.2.7 in Robert and Casella 1999). As stated by Cannings and Sheehan (2002), the single-site Gibbs Sampler is generally irreducible for binary traits determined by a diallelic locus. The only exception is when both homozygotes are affected with positive probability but the heterozygotes are affected with probability zero, which is not the case in our study. In general, the poor mixing caused by the flatness of the likelihood can not be improved with small sample size and non-informative prior as the resulting posterior surface turns out to be flat as well. However, when the sample size is relatively large and/or an informative prior distribution is used, the likelihood and resulting posterior surface become peaked near its mode, which in turn improves mixing. In our study, slow mixing that was likely caused by the latent variables was still noticed. It might be related to the data augmentation approach in this study. Similar problems were found by other authors (Liu et al. 1994; Sorensen et al. 1995). One possible scheme to accelerate mixing could be implementation of another blocking strategy, e.g., sampling jointly from the liability and selected parameters. In that case, another computational strategy and parameterization of models may be needed.

Complex segregation analysis can be applied to any pedigree structure and works with both qualitative and quantitative traits. The results from complex segregation analysis could be a useful starting point for defining major genes or detecting QTL in human genetics, animal and plant breeding (Jarvik 1998). In this study, the estimated large major gene variance component strongly supports the existence of major gene effects. The best performing parent with regard to rust disease was parent A, which was predicted to have the highest GCA effect and was estimated as *RR* (dominant homozygous) genotype with the highest probability among three genotypes. The estimated dominant homozygous genotype of parent A may be the confounded effects of multiple rust resistance genes. Although parent A appears to be the best source of resistance in this study, parents B, C, D and E all had high probabilities of being *Rr*, and these parents may also be useful sources of rust resistance for future rust research efforts.

References

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679

Amerson HV, Jordan AP, Kuhlman EG, O'Malley DM, Sederoff RR (1997) Genetic basis of fusiform rust disease resistance in loblolly pine. In: Proceedings of the 24th southern forest tree improvement conference, p 403

Cannings C, Sheehan NA (2002) On a misconception about irreducibility of the single-site Gibbs sampler in a pedigree application. Genetics 162:993–996

Carson SD, and Carson MJ (1989) Breeding for resistance in forest trees—a quantative genetics approach. Annu Rev Phytopathol 27:373–395

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

Flor HH (1956) The complementary genetic systems in flax and flax rust. Adv Genet 8:29–54

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–511

Hoeschele I (1988) Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. Theor Appl Genet 76:81–92

Janss LLG, Thompson R, Arendonk JV (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor Appl Genet 91:1137–1147

Janss LLG, Arendonk JV, Brascamp EW (1997) Bayesian statistical analyses for presence of single genes affecting meat quality traits in a crossed pig population. Genetics 145:395–408

Jarvik GP (1998) Complex segregation analysis: use and limitations. J Hum Genet 63:942–946

Kinloch BB, Walkinshaw CH (1991) Resistance to fusiform rust in southern pines: how is it inherited? In: Proceedings IUFRO rusts of pine working party conference. Banff, Alberta. Inf. Rep. NOR-X-317, pp 219–228

Kuhlman EG, Amerson HV, Jordan AP, Pepper, WD (1997) Inoculum density and expression of major gene resistance to fusiform rust disease in loblolly pine. Plant Dis 81(6):597–600

Liu JS, Wang WH, Kong A (1994) Covariance structure of the Gibbs sampler with application to the comparisons of estimators and augmentation schemes. Biometrika 81:27–40

Miyake T, Dogo T, Moriya K, Sasaki Y (1999) Bayesian analysis for existence of segregation of major genes affecting carcass traits in Japanese black cattle population. J Anim Breed Genet 116:207–215

Morton NE, MacLean CJ (1974) Analysis of family resemblance III. Complex segregation of quantitative traits. Am J Hum Genet 26:489–503

Nelson CD, Doudrick RL, Nance WL, Hamaker JM, Capo B (1993) Specificity of host:pathogen genetic interaction for fusiform rust disease on slash pine. In: Proceedings of the 22nd southern forest tree improvement conference. pp 403–410

Raftery AL, Lewis S (1992) How many iterations in the Gibbs sampler? Bayesian statistics, vol 4. Oxford University Press, Oxford, pp 763–774

Robert CP (2001) The Bayesian choice: from decision-theoretic foundations to computational implementation, 2nd edn. Springer, Berlin Heidelberg New York, pp 350–359

Robert CP, Casella G (1999) Monte Carlo statistical methods. Springer, Berlin Heidelberg New York

Sorensen DA, Anderson S, Gianola D, Korsgaard I (1995) Bayesian inference in threshold models using Gibbs sampling. Genet Sel Evol 27:229–249

Stelzer HE, Doudrick RL, Kubisiak TL, Nelson CD (1997) Derivation for host and pathogen genotypes in the fusiform rust pathosystem on slash pine using a complementary genetics model and diallel data. In: Proceedings of the 24th southern forest tree improvement conference, pp 320–330

Stelzer HE, Doudrick RL, Kubisiak TL, Nelson CD (1999) Prescreening slash pine and *Cronartium* pedigrees for evolution of complementary gene action in fusiform rust disease. Plant Dis 83:385–388

Thaller G, Dempfle L, Hoeschele I (1996a) Investigation of the inheritance of birth defects in swine by complex segregation analysis. J Anim Breed Genet 113:77–92

Thaller G, Dempfle L, Hoeschele I (1996b) Maximum likelihood of rare binary traits under different modes of inheritance. Genetics 143:1819–1829

Wilcox PL (1995) Genetic dissection of fusiform rust resistance in loblolly pine. Ph.D. thesis, North Carolina State University, Raleigh, NC

Wilcox PL, Amerson HV, Kuhlman EG, Liu BH, O'Malley DM, Sederoff RR (1996) Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proc Natl Acad Sci USA 93:3859–3864

Wright S (1934) Analysis of variability in number of digits in inbred strains of guinea pigs. Genetics 19:506–536

Zeng W (2000) Statistical methods for detecting major genes of quantitative traits using phenotypic data of a diallel mating. Ph.D. thesis, North Carolina State University, Raleigh, NC

Zeng W, Ghosh S, Li B (2004) Blocking gibbs sampling with a mixed inheritance for major gene detection. Genet Res 83:143–154