FINAL REPORT

SEARCH FOR MAJOR GENES WITH PROGENY TEST DATA TO ACCELERATE THE DEVELOPMENT OF GENETICALLY SUPERIOR LOBLOLLY PINE

DOE-Agenda 2020

Award Number: DE-FC36-99GO10380

September 22, 2003

Department of Forestry
North Carolina State University
Raleigh, North Carolina 27695-8002

Project Title:

Search for major genes with progeny test data to accelerate the development of genetically superior loblolly pine

Funding Period:

October 1, 1998 to June 30, 2003

Date of Final Report:

September 22, 2003

Recipient:

North Carolina State University, Raleigh, NC 27695-8002

Award Number:

DE-FC36-99GO10380

Subcontractors:

None

Other Partners (cost sharing):

Bowater, Inc.,
The Plume Creek Timber Company,
International Paper Company,
MeadWestvaco Corporation,
Weyerhaeuser Company.

Contact: Bailian Li (919-515-6845) bailian_li@ncsu.edu

Project Team:

Dr. Bailian Li
Dr. David O'Malley
Dr. Ben Liu
Dr. Robert Weir
Dr. Ron Sederoff

BACKGROUND AND OBJECTIVES

This research project is to develop a novel approach that fully utilized the current breeding materials and genetic test information available from the NCSU-Industry Cooperative Tree Improvement Program to identify major genes that are segregating for growth and disease resistance in loblolly pine. If major genes can be identified in the existing breeding population, they can be utilized directly in the conventional loblolly pine breeding program. With the putative genotypes of parents identified, tree breeders can make effective decisions on management of breeding populations and operational deployment of genetically superior trees. Forest productivity will be significantly enhanced if genetically superior genotypes with major genes for economically important traits could be deployed in an operational plantation program.

Project Objectives

The overall objective of the project is to develop genetic model and analytical methods for major gene detection with progeny test data and accelerate the development of genetically superior loblolly pine. Specifically, there are three main tasks:

- 1) Develop genetic models for major gene detection and implement statistical methods and develop computer software for screening progeny test data
- 2) Confirm major gene segregation with molecular markers
- 3) Develop strategies for using major genes for tree breeding

PROJECT SUMMARY

We have successfully completed this research project as originally proposed, after a no-cost extension to complete additional laboratory work with molecular markers and complete a Ph. D. student's dissertation. Three Ph.D. students have been trained and their dissertations have been successfully completed within the time frame of this research grant. A total of 25 related papers have been published in some high regarded scientific journals, such as *Biometrics*, *Heredity*, *Forest Science*, and *Canadian Journal of Forest Research*, and *Silvae Genetica*. The project PIs and students have been invited to give presentations at several scientific conferences (both international and domestic), and articles have been published in conference proceedings.

A genetic model with parent block Gibbs sampling method was developed in this study to make statistical inferences about the major gene and polygenic effects on quantitative traits for progenies derived from a half-diallel mating design. Using simulated data sets with different major and polygenic effects, the proposed method accurately estimated the major and polygenic effects of quantitative traits, and possible genotypes of parents and progenies. A special computer program "Mgene" has been developed based on this Bayesian analytical method, which was designed to detect a major gene in a diallel mating progeny population using a mixed inheritance model (MIM), assumed that the genetic

structure of a quantitative trait is controlled by a major gene and polygenic effects. The MGene was written with C++ for detecting major gene in current diallel progeny populations. The program has been effective in analyzing progeny test data for detecting major genes affecting quantitative traits, such as height or volume of pine trees. The analysis is done in four steps that use SAS for data manipulation and analysis, MGene program or SPLUS.

Computer simulations were conducted to investigate the factors affecting the mixed inheritance model (MIM) and relate the results to a traditional pure polygenic model. A parameter-based scheme has been used to simulate diallel data sets in the MIM framework. Each realization of diallel data sets consists of phenotypic observations simulated by summing up random deviations of the model effects. Different combinations of genetic parameters were used for simulations, e.g., three levels of additive major gene effect, two levels of dominant major gene effect, and two levels of genotype by environment interaction. The MIM model could have sufficient power when the magnitude of additive effect a has more than half the phenotypic deviation unit. The simulation results indicated that the MIM model had high power (near or above 90%) to detect a major gene when $a=1.0 \sigma_p$, e.g., one standard deviation of phenotypic variance, and with $d>0$. The power of detection was low when $a=0.5\sigma_p$ or less. The negative d values reduced the power of detection to about 20%. The true major gene variance σ_m^2 had the similar effect as genetic effects a and d in that a higher σ_m^2 resulted in high power for detecting major gene. Major gene was falsely detected in 1% to 3% for all simulated diallels with no true major gene effect. Such low levels of error probability were below a commonly used 0.05 alpha level. High power and low type one error strongly supported the reliability and accuracy of using MGene to analyze diallel progeny tests for major gene detection. The genotype by environment interaction had very little effects on the power of detection and the type one error, although slightly increased due to interaction with other parameters in the model. The variance estimates from the polygenic model and the MIM were generally similar and were close to the true values, although the GCA and SCA variance were slightly larger in the MIM as they were affected by skewed posterior distribution. The parameters regarding major gene effects, i.e. a , d , σ_m^2 were underestimated in most parameter settings, mainly because of the low power of major gene detection when the true a value was half of the σ_p^2 . When a value increased to one unit of σ_p^2 , the parameters became much closer to the true values. The GxE effect added some noises into genetic parameter estimates such as GCA variance, additive major gene effect and major gene variance.

The program has been effective in searching for major genes affecting growth and rust resistance in loblolly pine progeny populations. This program has been used to run through many progeny test data sets from the NC State University Tree Improvement Program. For a total of 190 diallels that have been analyzed with this MIM model, if the major gene is defined as the one having an effect of at least one standard deviation between homozygotes of the metric traits ($2a > 1.0$), nineteen diallels have been found to have the major gene effect. Dominant major gene effects (d) were positive for all 19 diallels. Additive major gene effect (a) ranged from 0.51 to 1.11 with an average of 0.66. In the diallels with a large major gene effect, genotypes of major gene for six parents were estimated with high probabilities. Such high confidence in estimating parental genotypes is very useful to select parent trees with the favorable allele of major genes. The major genotypes of parents showed high percent heterozygotes in the population, indicating major gene segregation in the progeny population. In most diallels, estimated parental major genotypes showed evidences of major gene segregation. In these diallels, the families of heterozygotes or homozygotes with favorable major gene alleles provided promising materials for further molecular analysis to detect QTL. The MIM, as the extension of

polygenic model, should provide more accurate estimates for genetic components and have a broader application in plant breeding schemes.

The Bayesian analysis was further modified to make inference about rust infection in progeny tests by a mixed inheritance model (MIM) that included both polygenic effects and major gene effects. Results showed that the MIM was a better model to explain the inheritance of rust-resistance than the pure polygenic model in the diallel population. A large major gene variance component (around 40-50%) indicated the existence of major genes for rust resistance.

Molecular markers of RAPDS and micro satellite have been used to examine for association with rust phenotypes in a diallel progeny population. RAPD markers analysis identified three QTLs from one parent that explained 21% of the phenotypic variation in the mapping family. One QTL was validated in progeny from 3 other families that shared this common parent (i.e., a half-sib family comprised of 3 full-sib families). This QTL explained 6.9% of the phenotypic variation in the mapping family and 5.6% of the phenotypic variation in the validation set ($P < 0.01$). The ability to detect and validate QTLs with breeding value in operational breeding populations could facilitate marker-assisted selection for forest tree breeding. Bulk segregate analyses and marker / trait co-segregation analyses were also used to search for major genes in rust resistance. Additional RAPD markers were found to be linked with the Fr2 locus in one parent. An improved map (framework linkage map) of the Fr2 linkage group was developed. This information will be valuable for using molecular markers for selecting for fusiform rust resistance in loblolly pine.

Strategies of using major genes and markers for breeding have been evaluated in this study. The low heritability of economic traits in forest trees made it difficult for effective selection and breeding. Clonal propagation could increase the heritability of individual tree performance, but loblolly pine genotypes are not easily cloned and field test size would be greatly increased. Marker assisted selection (MAS) has not been considered feasible in forest trees for several reasons. A major problem is that QTL detection often uses large mapping populations where hundreds of individuals are planted in a large contiguous block to reduce environmental heterogeneity and increase heritability. Such large testing populations are not common for tree breeding programs. New approaches are needed for QTL detection in the existing breeding populations, as QTLs discovered in breeding populations would be easier to integrate with breeding programs. The probability of detecting QTLs in breeding populations could be improved by screening biometrical data for families that show evidence of major gene segregation, as shown by the new method in this study. The application of QTLs for breeding requires that polygenic breeding values be integrated with breeding values for QTLs. Inheritance can be described using the mixed model, which explains phenotypic variation using a single gene effect, a polygenic effect, and an error effect attributed to environmental variation. Breeding value depends on both gene frequency and effect. The computer program, MGene, based on the Bayesian analytical method has been developed for detecting major genes affecting quantitative traits in diallel progeny test populations. The program can detect the major gene segregation, major genotypes and associated probabilities. With the putatively identified populations, molecular confirmation can be used to identify the markers associated with the major genes. The availability of DNA markers has made QTL analysis and molecular confirmation feasible for many species. A well-defined linkage map will be valuable to estimate more accurate QTL effects and QTL position in the whole genome, which would be more powerful for detecting the QTL effects of major genes among and within families. This information will be valuable for selection in the loblolly pine breeding program.

DEVELOP GENETIC MODELS AND ANALYTICAL METHODS

SIMPLE TESTS FOR DETECTING SEGREGATION OF MAJOR GENES

Classic quantitative genetics is based on the assumption that quantitative traits are controlled by “polygenes,” i.e., many genes each having small but roughly equal effect. Because the effects of individual genes cannot be estimated directly, some summary statistics, such as mean, variance and heritability, are often used to describe the genetic effects of quantitative traits (Falconer and MacKay, 1996, p.104). Recent studies, however, have shown that some quantitative traits may be under the influence of a number of genes having major effects, i.e. so-called “oligogenic” inheritance, in *Drosophila* (Long et al., 1995), domestic animals (Piper and Shrimpton, 1989), rice (Jiang et al., 1994) and tree species (Wilcox et al., 1996; Wu et al., 2000). If major genes that control quantitative traits can be detected and then manipulated at the individual gene level within a breeding population, breeding for quantitative traits would be much more efficient.

The methods for detecting major gene segregation with phenotypic data can generally be classified into two groups. The first group contains simple non-parametric tests for departure from normality. Major gene segregation can be detected but gene effects are not estimated. This group can be further divided into the tests based on phenotypic distributions at the population level (Hoeschele, 1988), and the tests based on within family distribution of phenotypes (Lynch and Walsh, 1998, p. 359). The other group of methods is based on mixture models of segregation analysis, where the phenotypic distribution is explained by a weighted mixture of several underlying distributions of a few major genotypes. An example of segregation analysis is the maximum-likelihood-based approach (Elston and Stewart, 1971; Morton and MacLean, 1974; Janss, et al., 1995). In this analysis, complicated pedigrees are used to construct likelihood-ratio tests for different models of major genes, and then heavy computation is required to estimate the major gene effects (Lynch and Walsh, 1998 p. 359; Zeng 2000 p. 40-60). Simple tests are useful tools for screening many populations to identify potential populations or pedigrees with major gene segregation. The segregation analysis with heavy computation and linkage analysis with molecular markers can then be used to locate major genes and estimate major gene effects.

Le Roy and Elsen (1992) examined several simple statistic tests for major gene detection in a hierarchical phenotypic data set with a nested mating design. The most powerful tests were identified as the Bartlett test (Bartlett, 1937), the within family-variance heterogeneity test, and the Fain test (Fain, 1978), the within family mean-variance regression. Uimari et al. (1996) found that the most suitable robust simple test for major gene detection under a hierarchical data structure was the log-ANOVA test (Box, 1953). The Bartlett test was more powerful than the log-ANOVA test, but not robust to the skewness. The data structure in their studies precluded partitioning the polygenic effect into detailed components, such as general combining ability (GCA), specific combining ability (SCA), and their interactions with environments. Effects of these polygenic components and their interactions with a major gene on the power and robustness of these simple tests are unknown. Furthermore, these simple tests have not been applied to detect major gene segregation for progeny populations derived from the commonly used diallel mating designs in tree species.

A traditional way to detect segregation of a major gene is by crossing two different inbred lines and then generating a segregating progeny population. This cannot be easily done with forest tree species because of the long life cycle and the nature of outcrossing species, where inbreeding depression for fitness and the costs and time for developing and maintaining inbred lines are prohibitive. Recent research in farm animals has shown that these methods have potentials for detecting major gene segregation using phenotypic observations from a progeny population derived from nested mating designs (Mayo, 1989; Hill and Knott, 1990; Le Roy and Elsen, 1992; and Uimari et al., 1996). A diallel mating design, commonly used in tree and plant breeding programs, contains both half- and full-sib families so polygenic genetic effects can be further partitioned into general combining ability (GCA) due to additive and specific combining ability (SCA) due to dominance effects. It has been difficult to incorporate genetic properties of major genes with the two levels of polygenic effects. Furthermore, diallel progenies from tree breeding are usually planted at several locations or site types to determine their growth potential under different environments. The potential larger environmental variation and genotype by environmental interaction, relative to animal breeding, may affect the statistical power of these methods for major gene detection. Thus, it is important to evaluate the power and robustness of these simple statistical methods for detecting the segregation of major genes with a diallel progeny population across environments.

In the study, we applied three simple tests to eight diallel progeny populations of loblolly pine (*Pinus taeda* L.) for detecting major gene segregation on 6-year tree height. The statistical power and robustness were further evaluated by using simulated data with different degrees of major gene and polygenic effects, assuming a mixed inheritance model with a major gene plus polygenic effects (GCA and SCA).

Simple Statistical Tests

Three commonly used statistical tests for major gene detection in farm animal populations were evaluated for diallel progeny populations in this study, i.e., the Fain test, the Bartlett test, and the log-ANOVA test. The null hypothesis (H_0) is that there is no major gene segregation, i.e., absence of major gene segregation, and the alternative hypothesis (H_1) is the presence of major gene segregation. H_1 for the Fain test (Fain, 1978) is that a curvilinear relation exists between family mean ($\bar{y}_i$) and variance (s_i^2). The families with major gene segregation have intermediate means and large variances, while families without major gene segregation have large means (either positive or negative) and small variances. The corresponding model is:

$$\log s_i^2 = \beta_0 + \beta_1 \bar{y}_i + \beta_2 \bar{y}_i^2 + \beta_3 \bar{y}_i^3 + e_i$$

where β_j , $j = 0, 1, 2, 3$, are coefficients, $\bar{y}_i$, $i = 1, \dots, k$, is the mean of family i , s_i^2 is the i^{th} family variance, e_i is the random error, and k is the total number of full-sib families. H_0 is not rejected if the F statistics for the model is not significant. This method has been evaluated for human family structure (Fain and Ott, 1976; Fain, 1978; Mayo et al., 1980), and for farm animal family structure (Le Roy and Elsen, 1992), where it was powerful for major gene detection, but not robust to skewness.

The Bartlett test and the log-ANOVA test are widely used for testing homogeneity of variances. When a major gene is segregating in a population, the within-family (full-sib or half-sib) distribution of the

trait depends on the major genotypes of the parents. If no major gene is segregating, within-family variances is expected to be homogenous if there is no other non-major gene related variance heterogeneity.

The Bartlett test statistics is (Bartlett 1937):

$$B = \chi^2 / C \sim \chi^2_{k-1}$$

$$\chi^2 = \left\{ \sum_{i=1}^k (n_i - 1) \ln \bar{s}^2 + \sum_{i=1}^k (n_i - 1) \ln s_i^2 \right\}$$

$$\bar{s}^2 = \sum_{i=1}^k (n_i - 1) s_i^2 / \sum_{i=1}^k (n_i - 1)$$

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^k (n_i - 1)} \right]$$

where, k is the total number of full-sib families, n_i is the number of progenies for full-sib family i , $i=1, \dots, k$, s_i^2 is the variance of full-sib family i , B is the Bartlett test statistics, and χ^2 is chi-square distribution. H_0 is rejected if the test statistic B exceeds the χ^2 critical value with $(k-1)$ degrees of freedom at the 5% significant level. The Bartlett test was found to be sensitive to the skewness of data, suggesting that significance could result from non-normality rather than heterogeneity of variances (Le Roy and Elsen, 1992; Uimari et al., 1996).

An alternative test for heteroscedasticity, which is less sensitive to departures from normality, is the log-ANOVA or Scheffe-Box test (Sokal and Rohlf, 1981, p. 105). This approach is to divide observations within a family into $\sqrt{n_i}$ sub-samples. The natural logarithm of variances for each subsample is calculated and treated as observations for final analysis of variance (ANOVA) with family as a main effect. If the F test for the family effect is significant, differences in variances are significantly greater among families than that would be expected on the basis of the average variance within families, i.e. variances among families are heterogeneous, and H_0 will be rejected (Sokal and Rohlf, 1981, p. 105).

Experimental Data Analysis

Eight diallel progeny populations were selected randomly from a loblolly pine (*Pinus taeda* L.) breeding population for applying three simple tests in this study. Each progeny population was generated with a half-diallel mating of six parents with no reciprocal crosses, and thus there were 15 full-sib families in each diallel population. Progenies were established with a randomized complete block design with six trees (in a row-plot) per family in each block. The experiments were replicated over four different locations with six blocks at each location (Li et al., 1996). Tree heights (cm) were measured at age six for all surviving trees in the tests. The survival rates ranged from 79% to 93% and skewness measured from 0.367 to 0.926 among diallels.

The Bartlett, the Fain, and the log-ANOVA tests were performed on the eight diallels. The 15 full-sib families within a diallel were treated as units for applying these simple tests. The significance level used to test for heterogeneity of variance and for the Fain test (F test) was $P \leq 0.05$.

Evaluation of Simple tests by Simulation

Simulation data sets were generated with the same diallel mating and experimental layout as the experimental data. The full statistical model, including polygenic effects (GCA and SCA) and no major gene effects, for the diallel population tested over multiple locations was:

$$[1] \quad Y_{ijklr} = \mu + T_i + B(T)_{j(i)} + G_k + G_l + S_{kl} + TG_{ik} + TG_{il} + TS_{ikl} + S * B(T)_{ijkl} + \varepsilon_{ijklr}$$

μ overall mean
 T_i $i = 1-4$, test location effects (E, fixed effect))
 $B(T)_{j(i)}$ $j = 1-6$, replicate effects (fixed effect for simulation)
 $G_k, G_l, k, l = 1-6$, general combining ability (GCA) effects $\sim N(0, \sigma_g^2)$
 S_{kl} specific combining ability (SCA) effects $\sim N(0, \sigma_s^2)$
 TG_{ik}, TG_{il} GCA*E $\sim N(0, \sigma_{tg}^2)$
 TS_{ikl} SCA*E $\sim N(0, \sigma_{ts}^2)$
 $S * B(T)_{ijkl}$ plot effects $\sim N(0, \sigma_{tbs}^2)$
 ε_{ijklr} $r = 1-6$, within plot error (random effects) $\sim N(0, \sigma_e^2)$

where N represented the normal distribution with mean 0 and variance σ_e^2 . This statistical model was used to generate data for a quantitative trait affected by environmental variation and polygenic genetic effects.

Polygenic effects

The total phenotypic variance was: $\sigma_p^2 = 2\sigma_g^2 + \sigma_s^2 + 2\sigma_{tg}^2 + \sigma_{ts}^2 + \sigma_{tbs}^2 + \sigma_e^2$, which was set equal to 1.0. Three genetic parameters were used to specify the polygenic parameter space. The values for heritability, $h^2 = \frac{4\sigma_g^2}{\sigma_p^2}$, were 0.1, 0.2, and 0.3. The ratios of dominance to additive genetic variance, $\gamma = \frac{\sigma_s^2}{\sigma_g^2}$, were 0.25, 0.50, and 0.75. The type-B genetic correlations, $r_B = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{tg}^2}$, were 0.60, 0.75, and 0.90. In order to separate σ_{tbs}^2 from σ_e^2 , we assumed that $\frac{\sigma_{tbs}^2}{\sigma_{tbs}^2 + \sigma_e^2} = 12\%$, which was similar to that used by Huber et al. (1992). These parameter ranges were chosen according to general results of diallel data analysis of loblolly pine (Li et al., 1996). Seven parameter combinations were used to simulate the polygenic mode (Table 1).

The polygenic model was modified to study robustness of tests when 1) the phenotypes followed a skewed distribution, and 2) the data were unbalanced due to missing crosses and poor survival. A skewed distribution was created by transforming the original polygenic phenotypic values (y) into new ones (x) using the inverse of the generalized Box-Cox transformation. The inverted version of the formula was given by MacLean et al. (1976):

$$x = -r + r \exp \left[\frac{1}{w} \ln \left(\frac{wy}{r} + 1 \right) \right]$$

where w was a transformation parameter, and r was a scaling factor to ensure that the logarithm would be positive. The scaling factor r was set to 5 and w was varied from 0.1 to 0.9. With these values, skewness of the transformed distribution varied almost linearly from 0.02 ($w=0.9$) to 0.41 ($w=0.1$) with the square of the multiple correlation coefficient $R^2 = 0.957$.

Table 1: Genetic parameter combinations used to simulate all cases: no genetic model, polygenic model and mixed model. The total phenotypic variance of the trait was set to $\sigma_p^2 = 1.0$ for all cases. Polygenic effects were controlled by h^2 , γ_B and γ , and the major gene effects were simulated by $2a$, $p(A_1)$ and d .

<i>Model</i>	<i>Case</i>	$2a$	$p(A_1)$	d	h^2	γ_B	γ
No genetic effect	0	0.0	0.0	0.0	0.0	0.0	0.0
Polygeni c Model (p)	1	0.0	0.0	0.0	0.1	0.75	0.50
	2	0.0	0.0	0.0	0.2	0.75	0.50
	3	0.0	0.0	0.0	0.3	0.75	0.50
	4	0.0	0.0	0.0	0.2	0.60	0.50
	5	0.0	0.0	0.0	0.2	0.90	0.50
	6	0.0	0.0	0.0	0.2	0.75	0.25
	7	0.0	0.0	0.0	0.2	0.75	0.75
Mixed Model (m)	1 / 2 / 3	1.0	0.5	$d = a/2$	0.1 / 0.2 / 0.3	0.75	0.5
	4 / 5 / 6	1.5	0.5	$d = a/2$	0.1 / 0.2 / 0.3	0.75	0.5
	7 / 8 / 9	2.0	0.5	$d = a/2$	0.1 / 0.2 / 0.3	0.75	0.5
	10 / 11 / 12	1.5	0.1	$d = a/2$	0.1 / 0.2 / 0.3	0.75	0.5
	13 / 14 / 15	1.5	0.9	$d = a/2$	0.1 / 0.2 / 0.3	0.75	0.5
	16 / 17 / 18	1.5	0.5	$d = a$	0.1 / 0.2 / 0.3	0.75	0.5
	19 / 20 / 21	1.5	0.5	$d = 0$	0.1 / 0.2 / 0.3	0.75	0.5
	22	1.5	0.5	$d = a/2$	0.2	0.75	0.25
	23	1.5	0.5	$d = a/2$	0.2	0.75	0.75
	24	1.5	0.5	$d = a/2$	0.0	0.0	0.0

Major gene effects

A major gene effect was added into the polygenic model to create a mixed inheritance model. We assumed a major gene with two alleles, A_1 and A_2 , affecting the quantitative trait. There were three genotypes, A_1A_1 , A_1A_2 , and A_2A_2 , with genetic values a , d , and $-a$, respectively (a as the additive and d as the dominant genetic effect). The parameters for major gene effects (expressed as the difference between homozygotes, $2a$ in σ_p units) were 1.0, 1.5, or 2.0. The frequencies of the favorable allele $p(A_1)$ were set to 0.1, 0.5, or 0.9. The dominance effects (d) took the values of a (complete dominance), $a/2$ (partial dominance), or 0 (additive). Twenty-four combinations for the mixed model of major and polygenic effects were evaluated (Table 1).

Simulation Scheme

All parents in a half-diallel were generated from a base-population under Hardy-Weinberg and linkage equilibrium. Large random mating and non-inbred populations were assumed to avoid the confounding effects of inbreeding on the estimation of additive genetic variance (De Boer and Hoeschele, 1993). The parental genotypes for the major gene were generated from a uniform distribution using genotype frequencies, calculated from given allele frequency as thresholds.

Parent genotypes were simulated by two methods called Level I and Level II sampling. For Level I sampling, six parent genotypes were randomly drawn from a base population for 500 replicates. Then, we calculated the genotype frequencies of progenies produced by crossing parents following a half-diallel mating design, one frequency for each of 15 full-sib families. Using these 15 sets of genotype frequencies as threshold, the genotypes of all 15 full-sib progenies were generated. With a given diallel with the six parent genotypes fixed, 500 sets of progeny population were generated (500 replicates for level II sampling).

For the polygenic effect, the statistical model (Eq. 1) was used for data generation for a quantitative trait. The polygenic and major gene effects were assumed independent. The major gene effects were added on to the polygenic portion in the same way, regardless of major genotypes. Each combination of parameters had 500 replicates. The location and replicate (block within location) effects were considered as fixed effects and adjusted before conducting simple tests.

A reduced model contained only environment effects, was used to generate data without polygenic genetic and/or without major gene effects (Eq. 2).

$$[2] \quad Y_{ijklr} = \mu + T_i + B(T)_{j(i)} + \varepsilon_{ijklr}$$

The type I error (false positive) was measured as the proportion of reduced model replicates where H_0 was rejected ($P \leq 0.05$). Using the simulated progeny test data of half-diallel mating, the Bartlett test, the log-ANOVA test, and the Fain test were evaluated for detection of major gene segregation. With the statistical model (Eq. 1) plus the additional major gene effects, the power of each test was measured as the proportion of replicates where H_0 was rejected. With the reduced statistical model (Eq. 2), the type I error of each test was measured as the proportion of replicates where the hypothesis of a major gene was accepted. With the statistical model (Eq. 1), the robustness of simple tests to the half-sib relationship was measured as the proportion of

replicates where H_0 was accepted. With the modified datasets generated from polygenic statistical model (Eq. 1) by skewing the distribution or deleting some of observations, the robustness of simple tests to the skewed distribution and survival rate were measured as the proportion of replicates where H_0 was rejected.

Results of Experimental Data

All eight diallels showed significance (at the 0.05 probability level) for the Bartlett χ^2 test, while only diallels 1, 2, 3, 5, and 8 showed significance for the log-ANOVA test (Table 2). The significant heterogeneity among family indicated possible segregation of a major gene in these diallel progeny populations for 6-year height. The Fain test showed that diallels 2, 5, and 8 were significant for regressions of family means and variance (Table 2).

The type I errors, i.e., percentage of false rejection of the null hypothesis (H_0), were 7.6% for the Bartlett test, 5.2% for the Fain test, and 4.6% for the log-ANOVA test when the simulated data sets with no genetic effects were analyzed. All three tests had a low false positive rate over replicates, close to the expected 5%.

Table 2: The Probability (P) values for three tests, the Bartlett test (B), the Fain test (F) and the log-ANOVA test (L) for eight real half-diallel data sets from the NCSU - Industry Tree Improvement Program. The survival rate and skewness of each set are noted. The P-values from diallels likely to have major gene segregation are bold.

Diallel #		1	2	3	4	5	6	7	8
Method	B	0.0	0.0	0.0	0.017	0.0	0.011	0.0	0.0
	F	0.144	0.009	0.245	0.247	0.0	0.379	0.234	0.011
	L	0.015	0.0	0.0	0.57	0.037	0.811	0.333	0.0
Skewness (S)		0.432	0.367	0.630	0.720	0.780	0.926	0.725	0.659
Survival rate (%)		90.3	92.8	84.9	92.2	79.0	80.3	89.9	80.6

Power of Detection

The power of detection for the log-ANOVA test was generally the same as that for the Bartlett test, but was much higher than that for the Fain test for level I sampling (Table 3). The size of the major gene effect had the strongest effects on the power of tests. With the increment of $2a$ (the difference between two homozygotes, in σ_p units), the powers of three tests increased

dramatically. The three tests had the largest power when the frequency of the favorable allele was intermediate, $p(A_1)=0.5$. At the high frequency $p(A_1)=0.9$, when the favorable allele was very common in the population, three tests had the lowest power of detection. With the increase of dominance (d) from pure additive to partial dominance and to dominance, power to detect increased for all three methods. Numerically, when $2a \geq 1.5\sigma_p$ and $d \geq a/2$, the power of the Bartlett test and log-ANOVA test was no less than 58% for $p(A_1)=0.1$, and at least 81% for $p(A_1)=0.5$. The power was 94% or more when $p(A_1)=0.5$ and either $2a \geq 2.0\sigma_p$ or $d \geq a$. For the Fain test, however, the power was less than 44% when $2a \leq 1.5\sigma_p$, $a \leq a/2$, and with any allele frequency.

The heritability of the polygenic inheritance had a small effect on the power of the three tests (Table 3). The power for the Bartlett test and log-ANOVA test increased with the increment of heritability, especially for the lower power cases. However, the power of the Fain test generally decreased with increasing heritability. The ratio of dominance to additive genetic variance for polygenic inheritance (γ) had almost no effect on the power of the Bartlett test and log-ANOVA test. The power of the Fain test, however, decreased slightly as γ increased.

Table 3: The power (%) of three simple tests for mixed model cases where six parents were chosen randomly for each replicate (level I sampling). There were 500 replicates for every case. The power was measured by the percentage of replicates rejecting H_0 at the $P \leq 0.05$ level.

Parameters		$2a$			$p(A_1)$			d			γ		
		1.0	1.5	2.0	0.1	0.5	0.9	a	a/2	0	0.25	0.50	0.75
Method $h^2 =$ 0.0	B	79			79			79					
	F	70			70			70					
	L	76			76			76					
Method $h^2 =$ 0.1	B	25	83	94	59	83	32	94	83	53			
	F	9	44	86	40	44	16	87	44	11			
	L	21	81	94	58	81	24	95	81	46			
Method $h^2 =$ 0.2	B	27	85	95	66	85	38	94	85	63	88	89	90
	F	6	34	72	32	34	14	69	34	9	37	34	31
	L	21	85	95	65	85	30	94	85	58	89	88	91
Method $h^2 =$ 0.3	B	38	88	94	67	88	45	95	88	70			
	F	6	25	90	23	25	8	59	25	8			
	L	26	87	94	65	87	35	95	87	65			

Notes: B = Bartlett test, F = Fain test, L = log-ANOVA test, $2a$ = the difference of two homozygotes, $p(A_1)$ = frequency of favorable allele, d = dominance of the major gene, h^2 = heritability of polygenic inheritance, and γ = ratio of dominance vs. additive variance component of polygenic inheritance.

For level II sampling (Table 4), the powers of three tests were generally higher than those in the level I except when the frequency of favorable allele was 0.9, and $d = a$ and $\gamma = 0.25$ or 0.75 for the Fain test. The power of tests increased with the difference of two homozygotes ($2a$) and dominance (d) of the major gene, except that the power of the Fain test decreased when d increased from $a/2$ to a . When $2a \geq 1.5\sigma_p$, $d \geq a/2$, and small or intermediate frequency of favorable allele, the powers of the Bartlett test and the log-ANOVA test were at least 95%. For the Fain test in the same situation, however, the power was no less than 46% except when $d = a$, where the power was only 29%. The power of the Bartlett test and log-ANOVA test increased slightly with increasing heritability, while the power of the Fain test mostly decreased. The ratio of dominance to additive (γ) for polygenic inheritance had no effect on the power of the Bartlett test and log-ANOVA test (Table 4). The power of the Fain test changed as γ varied, but without any trend compared with level I cases (Table 3).

Table 4: The power (%) of three simple tests for mixed model cases where six parents were fixed for all 500 replicates (level II sampling). Power was the percentage of replicates rejecting H_0 at the $P \leq 0.05$ level. Bold values show cases where the power is less than that in level I sampling.

Parameters		$2a$			$p(A_1)$			d		γ			
		1.0	1.5	2.0	0.1	0.5	0.9	a	a/2	0	0.25	0.50	0.75
Method $h^2 =$ 0.0	B		97			97			97				
	F		82			82			82				
	L		95			95			95				
Method $h^2 =$ 0.1	B	29	98	100	97	98	11	100	98	52			
	F	14	68	91	59	68	6	93	68	31			
	L	25	98	100	95	98	6	100	98	45			
Method $h^2 =$ 0.2	B	39	99	100	99	99	10	100	99	69	100	99	100
	F	12	72	72	48	72	6	29	72	30	33	57	20
	L	32	99	100	98	99	5	100	99	64	99	99	99
Method $h^2 =$ 0.3	B	43	100	100	100	100	13		100	67			
								100					
	F	11	46	90	35	46	3		46	4			
								51					
	L	35	100	100	100	100	5		100	67			
								100					

Notes: B = Bartlett test, F = Fain test, L = log-ANOVA test, $2a$ = the difference of two homozygotes, $p(A_1)$ = frequency of favorable allele, d = dominance of the major gene, h^2 = heritability of polygenic inheritance, and γ = ratio of dominance vs. additive variance component of polygenic inheritance.

Major gene effect

The power of major gene detection increased with the increment of major gene effect measured by the proportion of major gene variance (σ_m^2) in the total phenotypic variance (σ_p^2) (Figure 1). The power for level II was higher than for level I (Table 5). The power of the Bartlett test and the log-ANOVA test were almost the same and higher than the Fain test. The degree of dominance was also important, i.e., when $d=0$, the powers of tests for both level I and level II decreased even though the percentage of σ_m^2 / σ_p^2 increased (Figure 1).

Table 5: The percentage of total major gene variance component and the percentage of the dominant variance of the major gene in the total phenotypic variance $\sigma_p^2 = 1.0$, and the average power of three tests for all mixed model cases.

Case		<i>m13-15</i>	<i>m1-3</i>	<i>m10-12</i>	<i>m19-21</i>	<i>m4-6</i>	<i>m16-18</i>	<i>m7-9</i>
Special parameter		$p(A_1)=0.9$	$2a=1.0$	$p(A_1)=0.1$	$d=0$	$2a=1.5$	$d=a$	$2a=2.0$
Percentage of dominance variance for the major gene in the total variance		0.46	1.56	0.46	0.0	3.52	14.06	6.25
Percent of a major gene		4.10	14.06	20.30	28.13	31.64	42.19	56.25
Average power for level I cases								
Method	B	38	30	64	62	85	94	94
	F	13	7	32	9	34	72	83
	L	30	23	62	56	84	95	94
Average power for level II cases								
Method	B	11	37	99	63	99	100	100
	F	5	12	47	22	57	58	84
	L	5	31	97	59	99	100	100

Notes: m = mixed inheritance cases, B = Bartlett test, F = Fain test, L = log-ANOVA test, $2a$ = difference of two homozygotes, $p(A_1)$ = frequency of favorable allele, d = dominance of the major gene.

Half-sib relationship effect

The log-ANOVA and Fain tests were not affected by the half-sib relationship in a diallel progeny population. However, the Bartlett test resulted in more false detection of major gene segregation, i.e. having a high type I error (Table 6). One assumption of these simple tests was that these full-sib families are independent. However, for a half-diallel mating design with 6 parents, there were six half-sib families among these 15 full-sib families. Only three full-sib families were totally independent. In the parameter space studied, the log-ANOVA test and the Fain test were very robust to this departure from independence, with a type I error of 5% as expected. The Bartlett test was sensitive to this half-sib relationship. Its type I error was about 10% for most cases, and was even 16% for the type-B genetic correlation $\gamma_B=0.60$ cases. γ_B affected the sensitivity of the Bartlett test, as well as the heritability (h^2) of polygenic inheritance.

Survival rate effect

Survival rate of 80-90% had no effects on the type I error of these simple tests. Under the half-sib relationship, survival rate not only did not increase the type I error of the log-ANOVA test and the Fain test, but it also cancelled out the effects of the half-sib relationship on the Bartlett test and forced the type I error to about 7% (as in the case of 100% survival rate and no half-sib relationship). Except for the case where $\gamma_B = 0.6$, the strong genotype by environmental interaction, the type I error was remained about 14%.

With the increase of skewness of distributions to 0.25 under the half-sib relationship, type I error for the log-ANOVA test decreased to zero for all cases. For the Fain and Bartlett test, however, the type I errors increased with increased skewness. The Fain test had less than 33% false detection (type I error), even if skewness of distribution was 0.41. The most sensitive test for skewness was the Bartlett test. When skewness was larger than 0.25, the type I error for the Bartlett test exceeded 50%, and reached 92% and greater with 0.41 skewness for all cases. This type I error rate was larger than the power to detect major gene segregation for most cases. With increased skewness, the type I error rate for the log-ANOVA test decreased under the half-sib relationship. Heritability (h^2) of polygenic inheritance had a small effect on the type I error (slightly increased with h^2) of the Bartlett test and the Fain test.

Figure 1. Average power of statistical tests for simulated data at level I (a) and level II (b) under different percentage of major gene variance in the total phenotypic variance (σ_m^2 / σ_p^2). B, F and L referred to the Bartlett test, Fain test, and Log-ANOVA test, respectively.

Discussion of *Power and Robustness of Simple Tests*

The analysis of computer simulated data sets showed that the three simple statistical tests had different statistical power and robustness for detecting major gene segregation in diallel progeny populations. In general, all three methods can detect the segregation of a major gene when the effect of the major gene was large ($2a \geq 1.5\sigma_p$) (Table 3 and 4). The frequency of the favorable major gene allele and degree of dominance of the major gene were also important factors for detecting major gene segregation. The intermediate frequency of 0.5 had the highest power of detection, while low dominant effect reduced the power of detection, which was consistent with the results by Le Roy and Elsen (1992) and Uimari et al. (1996). Heritability of polygenic inheritance had only small effects on the power of detection. The ratio of dominance to additive genetic variance of polygenic inheritance had no systematic effect on the power of detection.

The power of major gene detection for the three simple tests in level II sampling, where the six randomly chosen parents were fixed for all 500 replicates, was generally high at given gene frequency levels. When the six parents were chosen randomly for each replicate in level I sampling, the power to detect the segregation of major gene was always less than that in level II except at high gene frequency $p(A_1)=0.9$. Thus, for a given half-diallel population with low to moderate frequencies of a major gene, the three simple tests were always more powerful in detecting major gene segregation than in a large base population with random sampling.

The Bartlett and log-ANOVA tests were almost equal in the power of detecting major gene segregation, but the Bartlett test was sensitive to skewness and half-sib relationship within the diallel progeny population. The Fain test was less powerful but also the least sensitive to skewness of the three tests. This was different from results of animal progeny population from a nested mating design, where the Fain test had similar power as the Bartlett test (Le Roy and Elsen, 1992) and was more sensitive to skewness (Uimari et al., 1996).

The effect of full-sib relationships was not considered by previous studies with the nested mating design (Le Roy and Elsen, 1992; Uimari et al., 1996). Uimari et al. (1996) reported that the proportion of false detection for the Bartlett test with skewness 0.58 was no more than 50% for a nested mating population. However, in the diallel progeny populations in this study, the proportion of false detection for the Bartlett test with skewness 0.41 was above 90%. This may be due to the use of full-sib families for the analysis in the diallel mating population, while the half-sib family was used for the Bartlett test in the nested design by Uimari et al. (1996).

The log-ANOVA test had almost the same power as the Bartlett test and was robust to skewness in the diallel populations, unlike in the nested design where the log-ANOVA test had inflated type I error ($\leq 25\%$) under 0.58 skewness (Uimari et al., 1996). This difference could be also due to the full-sib family structure used in our analysis. The effect of sub-sample size within full-sib families on the power of the log-ANOVA test was not studied. Different sub-sample sizes other than the square root of the number of progenies in a family may have better power than what we observed here. We did not evaluate the effect of the progeny size per family on the power of tests. Previous studies on the nested design (Le Roy and Elsen, 1992; Uimari et al., 1996) indicated that the power of tests increased with half-sib family size.

For the diallel progeny populations simulated in this study, the percent major gene variance in the total phenotypic variance and the dominance of the major gene were the two most important factors affecting the detection of major gene segregation for a quantitative trait. Considering all aspects of the analysis, the log-ANOVA test is the best simple test for half-diallel data structure, because it is powerful and robust to half-sib relationship and skewness. When $2a \geq 1.5\sigma_p$ and $d \geq a/2$, the power to detect the segregation of a major gene in a large base population is at least 58% and 81% for $p(A_1)=0.1$ and $p(A_1)=0.5$ respectively (level I sampling). With these same parameters of $2a$, d and $p(A_1)$, the power to detect the segregation of a major gene in a given half-diallel progeny population is no less than 95% (level II sampling).

Detection of Major Gene Segregation in Diallel populations

Diallels 2, 5, and 8 were most likely to have major gene segregation for 6-year height, while diallels 4, 6, and 7 were unlikely to have major gene segregation although the Bartlett tests were significant (Table 2). Based on the simulation results, the type I error (false detection) was expected to be high for the Bartlett test because of the half-sib relationship and high skewness for the experimental data (no less than 0.36, Table 2). The results from the log-ANOVA test should be more reliable than the Bartlett test, while the Fain test may be conservative because of the low power to detect the segregation of a major gene. Thus, the combination of results from the log-ANOVA test and the Fain test would be more plausible.

Applications of Simple Tests

These simple tests have been used in animal breeding programs to detect major genes and have made significant contribution to improve the commercially important traits (Mayo, 1989; Hill and Knott, 1990; Uimari et al., 1996). Results from this study showed that these tests could be useful for detecting major genes for growth traits in tree breeding populations. The log-ANOVA test is the best simple test for half-diallel data structure and could be the first step to identify candidate populations. Based on screening diallel populations for the possible major gene segregation, other segregation analysis methods, such as maximum likelihood (Janss, et al., 1995) and Bayesian analysis (Zeng, 2000) could be used to estimate major gene effects and identify major genotypes in the diallel populations. For a chosen population, linkage analysis with molecular markers could be carried out with greater chances and less costs to tag meaningful genes that control quantitative traits.

Simple tests could also offer a way to identify specific populations that may be more interesting to tree breeders in a breeding program. Populations identified with major gene segregation may have higher potential for improvement than parents in diallels without major gene segregation. Generally, major genes have relatively low allele frequency in the population as our results have showed. The power of detection for major gene segregation at low allele frequencies was above 50% for a large base population (Table 3) and above 95% for a given diallel population (Table 4). By purposely making more crosses and testing intensively for this population, the frequency of the major gene can be increased and may eventually be fixed through a few cycles of selection and breeding.

Literature Cited

- Bartlett, M.S. 1937. Some examples of statistical methods of research in agriculture and applied biology. *J. Roy. Stat. Soc. [suppl]* 4: 137-183.
- Box, G.E.P. 1953. Non-normality and tests on variances. *Biometrika*. 40: 318-335.
- De Boer, I.J.M., and I. Hoeschele. 1993. Genetic evaluation methods for populations with dominance and inbreeding. *Theor. Appl. Genet.* 86: 245-258.
- Elston, R.C., and J. Stewart. 1971. A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21:523-542.
- Fain, P.R. 1978. Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance. *Ann. Hum. Genet.* 42:109-120.
- Fain, P.R., and J. Ott. 1976. Heterogeneity of within sibship variance as a test of the major gene hypothesis. *Proc. V th Int. Congr. Hum. Genet. Excerpta. Med. Int. Congr. Ser.* 397,1809.
- Falconer, D.S., and T.F.C. MacKay. 1996. *Introduction to Quantitative Genetics*. Ed.4, New York: Longman. 464p.
- Hill, W.G., and S. Knott. 1990. Identification of genes with large effects. P. 517-538, *in* *Advances in Statistical Methods for Genetic Improvement of Livestock*, Springer-Verlag, Berlin.
- Hoeschele, I. 1988. Statistical techniques for detection of major genes in animal breeding data. *Theor. Appl. Genet.* 76:311-319.
- Huber, D.A., T.L. White, and G.R. Hodge. 1992. The efficiency of half-sib, half-diallel and circular mating designs in the estimation of genetic parameters in forestry: a simulation. *For. Sci.* 38: 757-776.
- Janss, L.L.G., R. Thompson, and J.A.M. van Arendonk. 1995. Application of Gibbs sampling for inference in a mixed major gene-polgenic inheritance model in animal population. *Theor. Appl. Genet.* 91: 1137-1147.
- Jiang, C.J., Pan X.B., and Gu, M.H. 1994. The use of mixture models to detect effects of major genes on quantitative characters in a plant breeding experiment. *Genetics* 136: 383-394.
- Le Roy, P., and J.M. Elsen. 1992. Simple test statistics for major gene detection: a numerical comparison. *Theor. Appl. Genet.* 83:635-644.
- Li, B., S.E. McKeand and R.J. Weir. 1996. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the southeastern U.S. p. 20-30 *in* Dieters M.J. Matheson, A.C. Nikles, D.G. Hardwood, C.E., and Walker, S.M. (eds.) *Tree improvement for sustainable tropical forestry*. QFRI, Australia.
- Li, B., S.E. McKeand, and R.J. Weir. 1999. Tree improvement and sustainable forestry -impact of two cycles of loblolly pine breeding in the US. *Forest Genetics* 6: 229-234.
- Lynch, M., and B. Walsh. 1998. *Genetics Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, MA. 980p.
- Long, A.D., S.L. Mullaney, L.A. Reid, J.D. Fry, CH. Langley, and T.F.C. Mackay. 1995. High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. *Genetics* 139:1273-1291.
- MacLean, C.J., N.E. Morton., R.C. Elston, and S. Yee. 1976. Skewness in commingled distributions. *Biometrics* 32:695-699.
- Mayo, O. 1989. Identification of genes which influence quantitative traits. p. 141-146 *in* Hill, W.G. and MacKay, T.F.C. (eds) *Evolution and Animal Breeding. Reviews on Molecular and Quantitative Approaches in Honor of Alan Robertson*. Wallingford, CAB Int.
- Mayo, O., T.W. Hancock, and P.A. Baghurst. 1980. Influence of major genes on variance within sibships for a quantitative trait. *Ann. Hum. Genet.* 43: 419-421.
- Morton, N.E., and C.J. MacLean. 1974. Analysis of family resemblance. III. complex segregation of quantitative traits. *Am. J. Hum. Genet.* 26:489-503.
- Piper, L.R., and A.E. Shrimpton. 1989. The quantitative effects of genes which influence metrics traits. p. 147-151 *in* Hill, W.G., Mackay, T.F.C. (eds) *Evolution and animal breeding: Reviews on molecular and quantitative approaches in honor of Alan Robertson*. Wallingford, CBA Int.
- Sokal, R.R., and F.J. Rohlf. 1981. *Biometry*, W.H. Freeman and Company, San Francisco. 368p.

- Uimari, P., B.W. Kennedy, and C.M. Dekkers. 1996. Power and sensitivity of some simple tests for detection of major genes in outbred populations. *Anim. Breed. Genet.* 113:17-28.
- Wilcox, P.L., Amerson, H.V. Kulman, E.G., Liu, B.H., O'Malley, D.M., and Sederoff, R.R. 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.
- Wu, R., Li, B., and Zeng Z-B. 2000. Molecular dissection of quantitative traits: new perspectives from populus. S.M. Jain and S.C. Minocha (eds.) *Molecular Biology of Woody Plants*. Vol I p. 475-490.
- Zeng, W. 2000. Statistical methods for detecting major genes of quantitative traits using phenotypic data of a diallel mating. Ph.D. dissertation, North Carolina State University, Raleigh, USA. 145p.

GIBBS SAMPLING METHOD TO DETECT MAJOR GENES

Several statistical approaches have been developed for the detection of major genes for quantitative traits, based on phenotypic data. Methods based on analysis of variance have been used to infer the number of major loci contributing to growth variation of interspecific aspen hybrids (Li and Wu, 1996; Wu and Li, 1999; Wu and Li, 2000). Several statistical methods using simple non-parametric tests for departure from normality have been used for detecting major gene segregation, but not for estimating major gene effects (Lynch and Walsh 1998; Zeng and Li 2003). Other approaches based on maximum likelihood and Bayesian inference have been developed for the MIM to detect major genes affecting quantitative traits in animal (Hoeschele, 1988; Knott, Haley and Thompson, 1991; Janss et al., 1997; Lund and Jensen, 1999), crop (Wang, et al., 2001) and tree species (Wu, et al., 2001). Most of these methods are based on either a multiple-generation pedigree, or else a progeny population derived from either a nested mating design (in the case of animal breeding) or a factorial mating design (in the case of tree hybrids). To our knowledge, however, no statistical methods have been developed for a progeny population derived from a diallel mating design.

Diallel mating is one of the most commonly used designs in plant and tree breeding programs (Hallauer and Miranda, 1981; Zobel and Talbert, 1984). Unlike factorial mating, where male parents from one group are crossed with females from a second group, parents in a diallel design are crossed either as male or female with other parents in a single group (Griffing, 1956). A half-diallel mating is the diallel mating without self and reciprocal crosses, as shown for a half-diallel mating with 6 parents, where both half-sib and full-sib progenies are produced for each of six parents. Diallel mating yields two levels of polygenic effects, i.e., the general combining ability (GCA) of parents due to additive polygenic effects, and the specific combining ability (SCA) of crosses due to dominant polygenic effects. The unique feature of diallel mating, one observation having two main effects has made it difficult to analyze with standard statistical programs for even polygenetic effects (Xiang and Li 2001). Thus, it has been difficult to incorporate genetic properties of major genes with the two levels of polygenic effects in a MIM model for analyzing diallel data. Because of high-dimensional marginalization of the joint density over the unknown single genotype and polygenic effects, it is practically impossible to maximize the likelihood function associated with such a model using analytical and/or numerical techniques (e.g., see Le Roy Elsen and Knott, 1989; Knott et al., 1991). For animal breeding, the Gibbs sampling algorithm has been found to be reasonably effective in making inference for a mixed inheritance model in a nested mating design, in which parents can be either male or female, but not both (Janss et al., 1997). Such analyses were primarily based on the half-sib relationships of parents (male or female) and their progenies. In the case of tree-breeding programs, diallel progenies are usually planted at several locations or site types to determine their growth potential under different environments. The potentially large environmental variation, as well as genotype by environmental interaction, relative to animal breeding, may affect the statistical power for major gene detection (Zeng and Li 2003). Although the Bayesian approach may have potential for major gene detection, its usefulness for MIM analysis of diallel data is unknown, because of the two types levels of polygenic effects and the heterogeneous environmental variance. It is important to evaluate the ability of the Bayesian approach for detecting the segregation of major genes with a diallel progeny population across environments.

In the second study, we developed a Bayesian approach, using a parent blocking Gibbs sampling, to make inferences about major genes and polygenic effects (GCA and SCA) that control quantitative traits for a progeny population derived from a half-diallel mating design without self and reciprocal crossings. Computer simulations were done to examine the effects of different prior distributions and design matrix, either full-ranked or non-full-ranked, on the proposed statistical method. A case study with one half-diallel progeny population of Loblolly pine. was used to detect a major gene for height growth and to illustrate the application of the method.

The Mixed Inheritance Model

A mixed inheritance model (MIM) is adopted in this study for the diallel analysis, in which phenotypes are assumed to be influenced by a single major gene and the polygenic effects (Zeng and Li 2003). A half-diallel mating design (Figure 1), with n_p parents selected from a base population under Hardy-Weinberg and linkage equilibrium (Falconer and Mackey, 1996), and $n_p(n_p - 1)/2 = n_s$ full-sib families, is used to study the MIM. Each full-sib family is tested at several sites, following a randomized complete block design with several trees per full-sib family within each block, and several blocks within each site. The statistical model for a MIM can be written as a mixed linear model:

$$Y = X\mu + Zu + WLn + e \quad (1)$$

In model (1), Y is a $(n \times 1)$ vector of n progeny observations; μ is the overall mean (it can be extended to a $(c \times 1)$ vector of c fixed non-genetic effects (e.g. site effect and block within site effect); X is a $(n \times 1)$ vector with value 1 of overall mean for all progenies; u is a $(q \times 1)$ vector of q random polygenic effects including n_p GCAs (one for each parent, notated as g) and n_s SCAs (one for each cross, notated as s), i.e. $u' = (g', s')$; Z is a $(n \times q)$ incidence matrix of GCA and SCA for all progenies; m is a (2×1) vector of major gene effects, i.e. $m' = (a, d)$, where a is the additive major genotypic effect, and d is the dominance major genotypic effect of the major gene under the one gene with two alleles traditional genetic model (Falconer and Mackey, 1996). L^1 is a (3×2) indicator matrix of the major gene effects for major genotypes, and W is an unknown $(n \times 3)$ random design matrix of major genotypes at the single locus for n progenies. Unlike MIM for an animal population, there is no incidence matrix Z in the major gene effect term, because data Y are only the phenotypic observations of progenies in a tree population, rather than progenies plus parents (sires and dams) as in an animal population. Finally, e is a $(n \times 1)$ vector of iid errors.

The distribution of e is assumed to be $N(0, \sigma_e^2 I)$, where N denotes the Normal distribution, σ_e^2 is the residual variance, and I is the $(n \times n)$ identity matrix. Consequently, given location and scale parameters, the vector of data Y is also normally distributed as:

$$Y | \mu, u, W, m, \sigma_e^2 \sim N(X\mu + Zu + WLn, \sigma_e^2 I)$$

For GCA, $g' = \{g_i, i=1, \dots, n_p\}$ are assumed to be mutually independent normal distributions, i.e. $g | \sigma_g^2 \sim N(0, \sigma_g^2 I)$, where σ_g^2 is the GCA polygenic variance due to additive polygenic effects. For SCA, $s' = \{s_j, j=1, \dots, n_s\}$ are assumed to be mutually independent normal distributions, i.e. $s | \sigma_s^2 \sim N(0, \sigma_s^2 I)$, where σ_s^2 is the SCA polygenic variance due to dominance polygenic effects.

The single major gene is assumed to be a bi-allelic (A_1 and A_2), autosomal locus with Mendelian transmission probabilities, such that each progeny has one of the three possible genotypes: A_1A_1 , A_1A_2 , and A_2A_2 with genetic effects a, d , and $-a$ respectively. For progeny $k (k=1, \dots, n)$, the genotype is represented as a random vector w_k , with values (1,0,0), (0,1,0), or (0,0,1) corresponding to the three possible genotypes of A_1A_1 , A_1A_2 , and A_2A_2 respectively. We denote $w_1=(1,0,0)$, $w_2=(0,1,0)$, and $w_3=(0,0,1)$. Let w_T represent w_1 , w_2 , and w_3 with T taking values 1, 2, and 3 respectively. These vectors w_k form the rows of W . Given the two parent genotypes $w_{p1(k)}$ and $w_{p2(k)}$, the genotype distribution of progeny k is denoted as $p(w_k | w_{p1(k)}, w_{p2(k)})$. This distribution describes the probability of alleles constituting genotype w_k being transmitted from parents with genotypes $w_{p1(k)}$ and $w_{p2(k)}$ when segregation of allele follows Mendelian transmission probabilities. Because of the independent structure of the genotypes, the joint genotype distribution of progenies can be written as:

$$p(W | w_p) = \prod_{k=1}^n p(w_k | w_{p1(k)}, w_{p2(k)}) \quad (2)$$

where w_p are the genotypes of n_p parents. The parent genotypes are sampled from a base population with genotypes in Hardy-Weinberg equilibrium (Falconer and Mackey, 1996). This is a reasonable assumption for tree breeding populations because individual trees serving as parents are usually selected independently from natural populations. Given the favorable allele frequency in the base population $f = p(A_1)$, the probability distribution of the genotype of parent i is assumed to be $p(w_{pi} | f)$ which follows Hardy-Weinberg proportions. Because of the independence among parents, the joint genotype distribution of parents can be written as:

$$p(w_p | f) = \prod_{i=1}^{n_p} p(w_{pi} | f) \quad (3)$$

In order to fully specify the Bayesian model, normal priors are assigned to overall mean μ , and major gene effects $m' = (a, d)$, i.e. $\mu \sim N(0, k_1^2)$, $a \sim N(0, k_2^2)$, and $d \sim N(0, k_3^2)$, where k_i^2 , $i=1, 2$ and 3 are the hyper-parameters of prior distribution. In the simulation and real data analysis we used $k_i=4$, $i=1,2,3$. Variance components, σ_g^2 , σ_s^2 and σ_e^2 are assumed to arise independently from conjugate inverted gamma distributions, i.e. $\sigma_g^2 \sim IG(u_1, v_1)$, $\sigma_s^2 \sim IG(u_2, v_2)$ and $\sigma_e^2 \sim IG(u_3, v_3)$, where IG stands for the inverted gamma distribution, and u_i 's and v_i 's are

hyper-parameters. For our application we used $u_i=2$, and $v_i = (u_i - 1) * \hat{\sigma}_i$, $i = 1, 2$ and 3 for $\hat{\sigma}_1 = \hat{\sigma}_g$, $\hat{\sigma}_2 = \hat{\sigma}_s$, and $\hat{\sigma}_3 = \hat{\sigma}_e$, where $\hat{\sigma}_i$'s are obtained from a preliminary study using frequency distribution method. The conjugate Beta prior is used for allele frequency, i.e., $p(f) \sim \text{Beta}(\alpha_f, \beta_f)$, where α_f and β_f are prior distribution parameters. We have chosen $\alpha_f = \beta_f = 1$ to express prior ignorance.

The joint posterior density of all unknowns, given the data Y , is proportional to the product of the likelihood function and the prior densities:

$$\begin{aligned}
 & p(\beta, m, u, W, w_p, f, \sigma_g^2, \sigma_s^2, \sigma_e^2 | Y) \propto \\
 & p(Y | \beta, m, u, W, \sigma_e^2) p(\sigma_e^2) p(g | \sigma_g^2) p(s | \sigma_s^2) p(\sigma_g^2) p(\sigma_s^2) p(\mu) p(a) p(d) p(W | w_p) p(w_p | f) p(f) \\
 & \propto (\sigma_e^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_e^2} (y - X\beta - Zu - Wlm)' (y - X\beta - Zu - Wlm)\right\} \\
 & \times (\sigma_e^2)^{-(u_3+1)} \exp\left\{-\frac{v_3}{\sigma_e^2}\right\} (\sigma_g^2)^{-\frac{n_g}{2}} \exp\left\{-\frac{1}{2\sigma_g^2} \sum_{i=1}^{n_g} g_i^2\right\} (\sigma_g^2)^{-(u_1+1)} \exp\left\{-\frac{v_1}{\sigma_g^2}\right\} \\
 & \times (\sigma_s^2)^{-\frac{n_s}{2}} \exp\left\{-\frac{1}{2\sigma_s^2} \sum_{j=1}^{n_s} s_j^2\right\} (\sigma_s^2)^{-(u_2+1)} \exp\left\{-\frac{v_2}{\sigma_s^2}\right\} \\
 & \times \exp\left\{-\frac{1}{2\kappa_1^2} \mu^2\right\} \exp\left\{-\frac{1}{2\kappa_2^2} a^2\right\} \exp\left\{-\frac{1}{2\kappa_3^2} d^2\right\} \\
 & \times \prod_{k=1}^n p(w_k = w_g | w_{p1(k)}, w_{p2(k)}) \prod_{i=1}^{n_g} p(w_{pi} = w_g | f) f^{\alpha_f-1} (1-f)^{\beta_f-1}
 \end{aligned} \tag{4}$$

In order to study the prior effects on the method behavior with this data structure, improper flat priors are used for the overall mean μ , and major gene effect a and d , i.e., $p(\mu) \propto \text{constant}$, $p(a) \propto \text{constant}$, and $p(d) \propto \text{constant}$, besides normal priors. The prior for σ 's is always inverted Gamma.

Gibbs Sampling

Parent Blocking

In order to make statistical inferences about unknowns, the marginal posterior distributions for the model parameters are of interest. However, it appears to be almost impossible to obtain such margins for our model. In analytical approaches, the study of marginal densities would require integration and/or summation. Often such marginalizations are not feasible to compute or even express in closed form for a high-dimensional model like MIM, as presented in equation (4). But this difficulty can be circumvented by means of simulation-based methods. The Gibbs sampler is based on sampling random varieties from a Markov chain (MC) with its stationary distribution as the posterior distribution, and the sampling from the MC used to perform the high dimensional Monte Carlo integration (Gelfand and Smith, 1990; Brooks, 1998). Samples are obtained from the full conditional distributions (up to a multiplicative constant), which form the transition probabilities of the Markov chain. Each time a full conditional distribution is visited, it is used to

sample the corresponding parameter, while other parameters are considered to be fixed, and then the realized value is substituted into the full conditional distribution of all other parameters.

To improve the mixing and hence the speed of convergence, it is possible to sample several parameters simultaneously, called a ‘block,’ from their joint conditional distribution instead of updating all parameters univariately. As long as all parameters are updated, the new Markov chain will still have equation (4) as its stationary distribution. Unlike an animal population where data Y include parents and their offspring, and usually span several generations, in a tree population we consider only the progeny observations in the data Y with just two generations. The sire block strategy (Janss, et al., 1997) has worked well for animal populations. Since in a diallel mating design, one tree served as a male as well as a female, we modified ‘a sire blocking’ into ‘a parent blocking’. In a parent blocking, the genotypes of a parent and its offspring are treated as a block and updated simultaneously. Consequently, in each cycle, the genotype of every offspring is updated twice instead of once a cycle as usual because of the combination of full-sib and half-sib structure in the diallel mating. Given the work of Liu, Wong and Kong (1994) and Robert and Sahu (1997), it seemed to us that the block Gibbs sampler would mix faster than the ordinary one-at-a-time version that updates each component sequentially. Blocking is generally effective when the elements within the block are highly correlated compared to the correlation between blocks.

Full Conditional Distributions

Full conditional distributions are derived from the joint posterior distribution (4). For notational convenience, the MIM can be rewritten as: $Y = H\theta + e$, where $H = [X : WL : Z]$ is a $(n \times p)$ matrix, $\theta' = (\mu, m', u') = (\mu, a, d, g', s') = (\theta_1, \theta_2, \dots, \theta_p)$ is a $(p \times 1)$ parameter vector.

In order to implement the ‘parent blocking,’ an exact calculation of the joint conditional distribution of a parent and its all offspring is required. The joint conditional distribution for parent i is: $p(w_{pi}, w_{i(1)}, \dots, w_{i(n_i)} | W_{-i(l)}, w_{-pi}, \theta, f, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y)$, where n_i denotes the number of offspring of parent i , and the offspring are indexed by $i(1), i(2), \dots, i(n_i)$, or simply $i(l)$, where $l = 1, \dots, n_i$. By definition, this distribution is proportional to $p(w_{pi} | W_{-i(l)}, w_{-pi}, \theta, f, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) * p(w_{i(1)}, \dots, w_{i(n_i)} | W_{-i(l)}, w_p, \theta, f, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y)$.

The first term is the genotypic distribution of the parent i , marginalized with respect to the genotypes of its offspring. The second term is the joint distribution of offspring genotypes conditional on the parents’ genotypes. To calculate the genotype distribution of parent i , the three possible genotypes of all offspring must be summed after weighting each genotype by its relative probability. The final marginalized full conditional distribution for the major genotype of the parent i is:

$$p(w_{pi} = w_T | W_{-i(l)}, w_{-pi}, \theta, f, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) \propto p(w_{pi} = w_T | f) * \prod_{k \in i(l)} \sum_{b=1}^3 p(w_k = w_b | w_{p1} = w_T, w_{p2}) p(\tilde{y}_k | w_k = w_b) \quad (5-1)$$

where $\tilde{y}_k = y_k - \mu - Z_k u$ is called the adjusted record, and Z_k are the k^{th} rows of the matrices Z , and w_b has the same notation as w_T . The penetrance function (or weight) is:

$p(\tilde{y}_k | w_k = w_b) \propto \exp\left\{-\frac{1}{2\sigma_e^2}(\tilde{y}_k - w_k Lm)^2\right\}$. The probabilities here are given up to a constant

of proportionality and must be normalized to ensure that $\sum_{T=1}^3 p(w_{pi} = w_T) = 1$.

For the genotypes of offspring, the marginalized full conditional distributions are the same as the usual full conditional distributions found by extracting from equation (4) the term in which $w_k, k \in i(l)$, is present. i.e.

$$\begin{aligned} p(w_k = w_T | W_{-k}, w_{-p}, \theta, f, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) &\propto \\ p(w_k = w_T | w_{p1(k)}, w_{p2(k)}) * p(\tilde{y}_k | w_k = w_T) &\end{aligned} \quad (5-2)$$

The full conditionals for allele frequency, location parameters and variance components are obtained by just extracting the relevant terms from the joint posterior density in equation (4) (for details see Zeng, 2000).

Simulation

To evaluate the procedure with this data structure, simulated data with both major gene and polygenic components were generated for this study. A 6-parent half-diallel mating design with 4 test sites and 6 blocks per site is used to simulate phenotypic observations, although the site effects and block within site effects are both set to zero. Six parents are chosen randomly from a base population in which the major gene and polygenic parameters are defined. There are 15 full-sib families, 6 progenies per family per block per test site, and a total of 2160 progenies across 4 test sites.

For all progenies, phenotypic observations were simulated according to the model (1). The polygenic effect (Zu) included $G_1 + G_2 + S$, where G_1 and G_2 are GCA effects for 2 parents with prior distribution $N(0, \sigma_g^2)$, and S is SCA effects with prior distribution $N(0, \sigma_s^2)$. Two genetic parameters, the narrow sense heritability of polygenic inheritance $h^2 = 4\sigma_g^2 / \sigma_p^2$ (where σ_p^2 is the total phenotypic variance), and the ratio of dominance to additive genetic variance of polygenic inheritance, $r = \sigma_s^2 / \sigma_g^2$, are used to calibrate these polygenic quantities (Huber et al., 1992).

The major genotypes of parents and progenies are simulated according to equation (3) and (2). The major gene variance component is calculated as following: $\sigma_m^2 = \sigma_a^2 + \sigma_d^2 = 2f(1-f)[(1-2f)d+a]^2 + [2f(1-f)d]^2$. The total phenotypic variance is $\sigma_p^2 = \sigma_m^2 + 2\sigma_g^2 + \sigma_s^2 + \sigma_e^2 = 1.0$. The relationship between polygenic effects and major gene effects is assumed to be additive. For our simulation, the parameters are set to $h^2=0.2$, $r=0.5$, $a=1.0$, $d=0.0$, and $f=0.2$. The realized favorable allele frequency (f) is 0.167. The major genotypes of six parents are A_2A_2 , A_2A_2 , A_2A_2 , A_1A_2 , A_1A_2 , and A_2A_2 . The effect of the single gene in this case is to be detected and estimated.

Effects of Prior Distribution, Initial Value, and Design Matrix

For the overall mean (μ), the additive major gene effect (a) and the dominance major gene effect (d), the priors are chosen as a flat distribution, i.e., $p(\mu), p(a)$ and $p(d)$, and are proportional to a constant, or chosen as a normal distribution, i.e. $p(\mu), p(a)$ and $p(d)$, and have $N(0, K^2)$, where $K = 4$ is used in the analysis. Both uniform and normal priors are used in the model, respectively, to see the effect of priors on posterior inference.

The design matrix for the random polygenic effects (GCAs and SCAs) can be either full rank by putting constraints $\sum_{i=1}^{n_g} g_i = 0, \sum_{j=1}^{n_s} s_j = 0$, or singular. Both design matrices are used to test its effect on the MCMC method, especially on the convergence of MCMC.

Initial values of GCA, SCA and variance components, $\sigma_g^2, \sigma_s^2, \sigma_e^2$, are obtained from the traditional genetic model analysis (without major gene effect). These estimates are used as initial values for the Markov chain. For the major gene, the ranges for a, d , and f are $[0.0, 1.0]$, $[0.0, 0.5]$, and $[0.1, 0.5]$, respectively (Table 1). The genotypes of parents are generated by f , assuming that parents are all from a base population with Hardy-Weinberg and linkage equilibrium. Given the parents' genotypes, major genotypes of progenies are generated by following the Mendelian transmission probabilities of allele segregation. These multiple independent parallel runs of Gibbs sampler can be used as a diagnostic tool to examine the mixing property of MCMC. For each case, two independent chains, with 40,000 iterations, each are run.

Table 1: The six combinations of initial values (N0-N5) for each of three combinations of prior and design matrix: Uniform priors for a , d , and μ with singular design matrix, normal priors for a , d , and μ with singular design matrix, and normal priors for a , d , and μ with full-rank design matrix to test the effect of prior, initial value, and design matrix on the mixed inheritance model (MIM) using a blocking Gibbs sampling. In MIM model, the polygene background setup was the narrow sense heritability of polygenic inheritance $h^2=0.2$, and the ratio of dominance to additive genetic variance of polygenic inheritance $r^*=0.5$, and the major gene effect was simulated by $2a$, d , and f .

Initial value			
Case \ Parameter	a	d	f
N0 (true values)	1.0	0.0	0.2
N1	0.0	0.0	0.1
N2	0.25	0.25	0.2
N3	0.5	0.0	0.3
N4	0.75	0.5	0.4
N5	1.0	0.0	0.5

Convergence Diagnostics

Bayesian Output Analysis (BOA version 0.5.0) (Smith, 2000) is used to analyze these outputs. The Gelman and Rubin Shrink Factors (Gelman and Rubin, 1992) plot is used to determine the burn-in time as well as the convergence. The autocorrelation plot is then used to determine the length of thinning lag in order to get a relatively independent sample for the final analysis. Brooks, Gelman and Rubin's Corrected Scale Reduction Factors (for multiple chains) and Raftery and Lewis's Dependence Factors (for a single chain) are also used to diagnose the convergence of MCMC chains (Brooks and Roberts, 1998). As a rule of thumb, if the 0.975 quantile of Corrected Scale Reduction Factors is less than 1.2, the sample may be considered to have arisen from the stationary distribution. For a single chain, Dependence Factors greater than 5.0 often indicates convergence failure and a need to reparameterize the model. Trace plots are used as indicators of mixing and convergence of chains.

In the Gibbs chain, the additive major gene effect (a) may be positive as well as negative. The sign of a is relevant, i.e. the favorable allele is A_1 when a is positive, and A_2 when a is negative. From the Gibbs samples, we are interested in the absolute value of a . For consistency, we change the frequency of the favorable allele (f) to $1-f$ when a is changed from a negative value to a positive value.

Results of Gibbs Sampling Method

When the design matrix was singular, both uniform prior or normal prior provided good frequentist coverage estimates, except for the fact that σ_g^2 estimates were lower than expected from simulations (Table 2). The Gelman and Rubin plot for a set of initial values (as in N1) indicated that the burn-in iteration was about 25,000 iterations (see Figure 1). The corrected scale reduction factors were approximately 1.0, and Raftery dependence factors were found to be much less than 5.0. Posterior densities of six genetic parameters, $a, d, f, \sigma_e^2, \sigma_g^2$ and σ_s^2 , for a different set of initial values (as in N2) are listed in Figure 2. These numerical diagnostic summaries (Table 3 and Figure 1) indicated that the chains mix well and there was no severe problem with MCMC convergence. We also found that the prior and initial values did not have any effects on the Gibbs sampler under the singular design matrix.

When the design matrix was chosen to be of full rank, each individual chain converged with the Raftery dependence factors less than 5.0. The parameter estimates from the five different initial value sets N0, N1, N2, N4, and N5 were very good in term of precision. For the set of initial values N3, however, the estimates of major gene genotype for the 6th parent, as well as the corresponding genetic parameters, were not correct (see Table 2). As a result, the 0.975 quantile of corrected scale reduction factors for parameter d, f , and σ_g^2 were 1.28, 1.29, and 2.87 respectively (see Table 3). This may indicate a possible mixing problem for the N3 set due to the combination of full-rank design matrix with this data structure. To avoid this possibility, a normal prior with a singular design matrix was chosen as a model for the further analysis of our case study.

Table 3: Convergence diagnostics of the six parameters, a , d , f , σ_e^2 , σ_g^2 , σ_s^2 , for testing the effects of prior distribution, initial value, and design matrix on the mixed inheritance model. In the MIM model, the polygene background setup was $h^2=0.2$, $r=0.5$, and the major gene effect was simulated by $2a=2.0$, $d=0.0$ and $f=0.2$ (actual value is 0.167). There were six runs, N0 to N5, for each of three cases.

<i>Parameters</i>			<i>a</i>	<i>d</i>	<i>f</i>	σ_e^2	σ_g^2	σ_s^2
Uniform Prior	CS Reduction Factor ¹	Est	1.00	1.00	1.00	1.00	1.00	1.00
		.975	1.01	1.01	1.00	1.00	1.00	1.00
	Raftery Dependent Factor ²	N0	12.2(1.7) ³	3.0(1.5)	1.0	1.1	1.1	1.0
		N1	9.8(2.0)	5.4(1.7)	1.0	1.3	1.1	1.1
		N2	10.4(1.7)	4.9(1.3)	1.0	1.3	1.0	1.2
		N3	5.0(1.7)	5.5(1.5)	1.0	1.2	1.0	1.0
		N4	8.2(3.3)	6.3(2.6)	1.0	1.2	1.2	1.0
		N5	6.2(1.5)	5.8(1.1)	1.0	1.2	1.0	1.2
Normal Prior	CS Reduction Factor	Est	1.00	1.00	1.00	1.00	1.00	1.00
		.975	1.01	1.01	1.00	1.00	1.00	1.00
	Raftery Dependent Factor	N0	13.0(3.1)	8.7(2.0)	1.1	1.3	1.0	1.2
		N1	13.6(2.2)	7.6(1.5)	1.1	1.3	1.1	1.2
		N2	19.9(4.3)	7.1(1.5)	1.0	1.0	1.2	1.0
		N3	4.4(1.3)	4.6(1.5)	1.2	1.2	1.0	1.1
		N4	4.9(1.2)	8.2(1.8)	1.1	1.1	1.1	1.2
		N5	5.4(1.3)	4.6(1.5)	1.0	1.3	1.3	1.1
Normal Prior With Full-rank Design matrices	CS Reduction Factor	Est	1.04	1.11	1.12	1.04	1.98	1.03
		.975	1.09	1.28	1.29	1.11	2.87	1.08
	Raftery Dependent Factor	N0	5.2(1.5)	5.7(1.3)	1.0	1.1	1.0	1.0
		N1	6.4(1.7)	4.9(1.3)	1.0	2.0	1.2	1.2
		N2	5.9(1.5)	4.7(1.3)	1.1	1.2	1.2	1.2
		N3	9.1(2.2)	3.2(1.3)	1.0	1.3	2.6	1.0
		N4	31.1(5.9)	6.2(1.3)	1.0	1.3	1.2	1.1
		N5	8.0(1.7)	7.4(3.3)	1.1	1.2	1.2	1.2

- 1: CS Reduction Factor: corrected score reduction factor;
- 2: Raftery Dependent Factor is calculated under Quantile = 0.025, Accuracy = ± 0.05 , and Probability = 0.9.
- 3: The dependent factor in the parenthesis was calculated by using lag=30, instead of lag=5 in the regular base because of strong autocorrelation.

Figure 1: Gelman and Rubin plots of six genetic parameters (a , d , f , σ_e^2 , σ_g^2 , σ_s^2) of the initial value set N1 for normal prior for a , d , and μ with singular design matrix for testing the effects of prior distribution, initial values, and design matrix on mixed inheritance model.

Table 4: Estimated means, and standard deviations of posterior densities for seven genetic parameters (a , d , f , σ_e^2 , σ_g^2 , σ_s^2 , σ_m^2), and general combining ability (GCA) ($g_1 - g_6$) and major gene genotypes of six parents ($P_1 - P_6$) for the diallel from Bayesian based segregation analysis. There were two independent runs, chain 1 and chain 2.

Genotype	P_1	P_2	P_3	P_4	P_5	P_6			
Chain 1	$A_2 A_2$ (.95)	$A_1 A_2$ (.90)	$A_2 A_2$ (.86)	$A_2 A_2$ (.84)	$A_2 A_2$ (.97)	$A_2 A_2$ (.91)			
Chain 2	$A_2 A_2$ (.97)	$A_1 A_2$ (.86)	$A_2 A_2$ (.82)	$A_2 A_2$ (.80)	$A_2 A_2$ (1.0)	$A_2 A_2$ (.93)			
GCA	g_1	g_2	g_3	g_4	g_5	g_6			
Chain 1	1.30±0.81 ²	-0.88±0.64	0.03±0.69	-0.45±0.67	0.66±0.67	-0.27±0.75			
Chain 2	1.53±0.92	-0.91±0.62	-0.05±0.71	-0.48±0.68	0.84±0.76	-0.26±0.76			
Param.	a	d	f	σ_e^2	σ_g^2	σ_s^2	σ_m^2	σ_p^2	σ_m^2 / σ_p^2
Chain 1	2.29±0.65	-2.31±1.05	0.33±0.16	6.96±0.25	0.96±0.92	1.41±0.90	2.18±1.50	12.47	0.175
Chain 2	2.17±0.65	-2.25±0.78	0.37±0.17	6.91±0.27	1.12±1.09	1.29±0.92	2.16±1.41	12.60	0.171

1: Probability of this genotype;

2: the estimated mean ± standard deviation from the Gibbs samples with lag=5;

The Case Study

The blocking Gibbs sampling method was applied to a progeny data set derived from a 6 parents, half-diallel mating of loblolly pine by the N. C. State University Tree Improvement program pine (Li et al., 1999). Similar to the simulated data, 15 full-sib families from the diallel mating were planted at 4 different sites with 6 blocks each. Tree heights of progenies at age 6 were measured as the quantitative trait for this analysis. First, a mixed linear model $y = t + b(t)$ was fitted to adjust the site effect (t) and block within site effect ($b(t)$). The residuals were used as the phenotypic observation vector Y . Initial values for GCA, SCA and variance components were all taken from the estimates from a traditional polygenic model without the major gene. The prior distribution for μ, a and d were normal $\sim N(0, K^2)$ with $K = 4$. The hyper-parameter for allele frequency f were $\alpha_f = 1$ and $\beta_f = 1$. A singular design matrix was chosen for the analysis.

The results for two independent chains, with different sets of initial values for a, d and f , were very close to each other using 240,000 iterations (Table 4). The 0.975 quantile of corrected Scale Reduction Factors for parameters was less than 1.2. The percentage of major gene effects was estimated as 17% of the total phenotypic variance. The estimated major genotype of parent 2 was $A_1 A_2$, while the others were $A_2 A_2$. The additive effect of the major gene was $a \approx 2.3$ and the dominance effect was $d \approx -2.3$. That is, $A_1 A_1$ had $a \approx 2.3$, $A_1 A_2$ had $d \approx -2.3$ and $A_2 A_2$ had $a \approx -2.3$. The results indicated that there might be a detectable recessive major gene controlling

the height growth of loblolly pine in this diallel population, although the effect of the major gene is small compared with the polygenic effects (83%). High estimated GCA values also indicated that the polygenic component was more important for height growth of loblolly pine at age 6. Given the limitations of the experimental design and relatively small effect of the major gene, the major genotypes may not be accurate in this case study. Furthermore, the validity of the model assumptions and possible interaction of polygenic and major gene effects may make this genotype interpretation difficult.

Discussion of Gibbs Sampling Method

The Bayesian approach with parent-blocking Gibbs sampling has shown to be effective in this study for analyzing data from a half-diallel mating design using a mixed inheritance model. The method can be successfully used to detect major gene segregation, estimate major gene effects and putative genotypes of a major gene for parents and progenies, as well as polygenic parameters of a quantitative trait. To our knowledge, this is the first statistical approach that incorporates the polygenic effects of GCA and SCA with a major gene in the MIM for a diallel mating design. The results from this model have provided a better understanding of mixed inheritance of quantitative traits in diallel populations, particularly for tree-breeding.

Although major-gene genotypes detected are putative based on the statistical inference, this information of segregation could be valuable for identifying parents with major genes affecting quantitative traits. The proposed method is based on the existing half-diallel mating design, and hence it can be used to analyze actual progeny test data for breeding purposes. By systematically screening progeny test data with this method, putative major genes, genotypes of parents and progenies, and their probabilities can be estimated. This is in addition to the polygenic effects of GCA and SCA, and other variance component estimates from the traditional analysis. The detectable major gene and putative genotypes would be valuable for selecting materials in an active breeding program. By combining the GCA and SCA estimates and possible major genotypes, suitable combinations of parents or progenies can be chosen to provide maximum genetic gains for a breeding program.

The putative genotypes of major genes identified with this method could also be valuable for molecular mapping experiments by providing a mapping population with a high probability of segregation for the quantitative traits. This should improve the effectiveness in search for quantitative trait loci (QTL) in the laboratory and reduce the experimental costs of such a search. Often no QTLs can be detected due to inadequate segregation in the experimental population. Our analytical approach can thus be first used to identify parents or families that are most likely segregating for a quantitative trait before further molecular mapping and linkage analysis are pursued. The detection of major genes using statistical approaches and confirmation of the presence of a major gene using genetic markers are very important for designing more effective breeding strategies and would make breeding for quantitative traits much more efficient.

One problem with the traditional ML approach is that it is not feasible to obtain mle's either by maximizing the likelihood of incomplete data directly or by using an iterative algorithm such as EM. By adapting a Bayesian framework into the segregation analysis, we avoided the necessity of performing such an optimization. In addition, estimates of the parameters are based on finite

sample posterior distribution and thus avoid the use of asymptotic approximation using Fisher information. The marginal Bayesian estimators take into account the uncertainty of a single parameter that is due to the uncertainty in all other parameters in the model. Thus, it can take into account all model uncertainty based on finite samples. Model selection methods based on theoretic criteria such as AIC and BIC can be used to choose models with different number of major genes.

Efficiency of Gibbs sampling depends on the mixing property of the Markov chain, which in turn is determined by the parameterization used in the model and the sampling scheme applied. From the consistent results of multiple chains and convergence tests, we conclude that the chains have mixed well for parent block sampling of genotypes. However, if the size of progeny population is small and/or the major gene effect is small, mixing may become a problem even with the parent block sampling. If additional molecular marker information is included in the model or the overall mean μ in the model is extended to a vector by including other non-genetic parameters, such as site effects and block within site effects, the mixing problem may be worse. One possible way to avoid this is to use the hybrid Markov chain embedding a Hasting or Metropolis updating step in the basic Gibbs sampling scheme, as used in pedigree analysis (Tierney, 1994). Another way would be to use Metropolis jumping kernel to make transition between communicating classes (Lin, 1995). A Bayesian network is also an alternative solution (Lund and Jensen, 1999).

Although only one major gene with bi-allele was considered in this study, this method can be extended to more general situations by considering $2n_p$ alleles and/or two or more major genes. When multiple alleles and genes are involved in the model, many important issues such as Hardy-Weinberg disequilibrium (among alleles for one gene), linkage disequilibrium (association among genes), and epistasis (non-allelic interaction) should be examined. In these cases, model selection can be adapted by means of the Bayes Factor (Kass and Raftery, 1998), or by means of a predictive loss approach (Gelfand and Ghosh, 1998).

Literature Cited

- Brooks, S.P. (1998). Markov chain Monte Carlo method and its application. *The Statistician* **47**, 69-100.
- Brooks, S.P., and Roberts, G.O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* **8**, 319-335.
- Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics. Ed.4, New York: Longman.
- Gelfand, A.E., and Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**(1), 1-11.
- Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.
- Gelfand, A.E., and Sahu, S.K. (1998). Identifiability, improper priors and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* **94**, 247-253.
- Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-511.
- Gilks, W.R., and Roberts, G.O. (1996). Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), pp. 89-114. London: Chapman & Hall.

- Griffing, G. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Aust. J. Biol. Sci.* **9**:463-493.
- Hallauer, A.R., and Miranda, J.B. (1981). Quantitative genetics in maize breeding. Iowa State University press, Ames, Iowa. P468.
- Hoeschele, I. (1988). Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theoretical and Applied Genetics* **76**, 81-92.
- Huber, D.A., White, T.L., and Hodge, G.R. (1992). The efficiency of half-sib, half-diallel and circular mating designs in the estimation of genetic parameters in forestry: a simulation. *Forest Science* **38**, 757-776.
- Janss, L.L.G., Van Arendonk, J.A.M., and Brascamp, E.W. (1997). Bayesian statistical analysis for presence of single gene affecting meat quality traits in a crossed pig population. *Genetics* **145**, 395-408.
- Jiang, C.J., Pan, X.B., and Gu, M.H. (1994). The use of mixture models to detect effects of major genes on quantitative characters in a plant breeding experiment. *Genetics* **136**, 383-394.
- Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90**(430), 773-795.
- Kaya, Z., Sewell, M.M., and Neale, D.B. (1999). Identification of quantitative trait loci influencing annual height- and diameter increment growth in loblolly pine (*Pinus taeda* L.). *Theoretical and Applied Genetics* **98**, 586-592.
- Kinghorn, B.P., Van Arendonk, J. and Hetzel, J. (1994). Detection and use of major genes in animal breeding. *AgBiotech News and Information* **6**(12), 297-302.
- Knott, S.A., Haley, C.S., and Thompson, R. (1991). Methods of segregation analysis for animal breeding data: a comparison of power. *Heredity* **68**, 299-311.
- Le Roy, P., Elsen, J.M., Knott, S.A. (1989). Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genetics Selection and Evolution* **21**, 341-357.
- Li, B., and Wu, R. (1996). Genetic causes of heterosis in juvenile aspen: a quantitative comparison across intra- and interspecific hybrids. *Theo. Appl. Genetics*, **93**, 380-391
- Li, B., McKeand, S.E., and Weir, R.J. (1999). Tree improvement and sustainable forestry – impact of two cycles of loblolly pine breeding in the U.S.A.. *Forest Genetics*, **6**(4), 229-234.
- Lin, S., (1995). A scheme for constructing an irreducible Markov Chain for pedigree data. *Biometrics* **51**, 318-322.
- Liu, J.S., Wong, W.H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81** (1), 27-40.
- Long, A.D., Mullaney, S.L., Reid, L.A., Fry, J.D., Langley, CH., and Mackay, T.F.C. (1995). High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. *Genetics* **139**:1273-1291.
- Lund, M.S., and Jensen, C.S. (1999). Blocking Gibbs sampling in the mixed inheritance model using graph theory. *Genetics Selection and Evolution* **31**, 3-24.
- Piper, L.R., and Shrimpton, A.E.. (1989). The quantitative effects of genes which influence metrics traits. in Hill, W.G., Mackay, T.F.C. (eds) *Evolution and animal breeding. Reviews on molecular and quantitative approaches in honor of Alan Robertson.* Wallingford, CBA Int, 147-151.
- Remington, D.L. AND O'Malley, D.M. (2000) "Evaluation of major genetic loci contributing to inbreeding depression for survival and early growth in a selfed family of *Pinus taeda*", *Evolution* **54**:1580-1589
- Roberts, G.O., and Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of Royal Statistical Society Serial B* **59**, 291-317.
- Smith, B. (2000). Bayesian output analysis program (BOA) Version 0.5.0 Programmer Manual.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annual Statistics*. **22**, 1701-1762.
- Wang, C.S, Rutledge, J.J., and Gianola, D. (1994). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics Selection and Evolution* **26**, 91-115.

- Wang, J., Podlich, D.W., Cooper, M. and Delacy, I.H. (2001). Power of the joint segregation analysis method for testing mixed major-gene and polygene inheritance models of quantitative traits. *Theoretical and Applied Genetics* **103**, 804-816.
- Wilcox, P.L., Amerson, H.V., Kulman, E.G., Liu, B.H., O'Malley, D.M., and Sederoff, R.R. (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.
- Wu, R. and Li, B. (1999). A multiplicative epistatic model for analyzing interspecific differences in outcrossing species. *Biometrics* **55**, 355-365.
- Wu, R. and Li, B. (2000). A quantitative genetic model for analyzing species differences in outcrossing species. *Biometrics* **56**, 1098-1104.
- Wu, R., Li, B., Wu, S.S. and Casella, G. (2001). A maximum likelihood-based method for mining major genes affecting a quantitative character. *Biometrics* **57**, 764-768.
- Xiang, B. and B. Li. 2001. A new mixed analytical method for genetic analysis of diallel data. *Canadian J. For. Research* **31**, 1-8.
- Zeng, W. (2000). Statistical methods for detecting major genes of quantitative traits using phenotypic data of a diallel mating. Ph.D. dissertation, North Carolina State University, Raleigh, USA. 145p.
- Zeng, W. and BL Li. 2003. Simple tests for detecting segregation of major genes with phenotypic data from a diallel mating. *Forest Science* **49(2)**, 000-000.
- Zobel, B.J., and Talbert, J.T. (1984). Applied forest tree improvement. John Wiley and Sons. New York. NY.

MOLECULAR CONFIRMATION OF MAJOR GENE SEGREGATION

The probability of detecting QTLs in breeding populations could be improved by screening biometrical data for families or pedigrees that show evidence of major gene segregation (Leroy and Elsen 1992, Zeng and Li 2003). Under the assumptions of the infinitesimal model of polygenic inheritance, a normal distribution is expected for full-sib progeny phenotypic values (Fain 1978). The phenotypic frequency distribution for a family could depart from normality (e.g., appear bimodal, or show skewness or kurtosis), but statistical tests for departure from normality are not powerful (Lynch and Walsh 1998). Assuming a single major gene with two alleles and additive effects, three genotypes and six kinds of families are possible. A cross between heterozygous parents has an intermediate mean and a larger variance than crosses between heterozygotes and homozygotes, or between homozygotes. The statistical tests that reject variance homogeneity of full-sib families (e.g., Bartlett's test, Bartlett 1937, and the Log ANOVA test, Martin and Games 1977) have been used for detection of major gene segregation (Fain 1978; Leroy and Elsen 1992). Application of these simple statistical tests (the Bartlett test, the log-ANOVA test, and the Fain test) showed that there could be major gene segregation for 6-year height in three diallel mating populations of loblolly pine (Zeng and Li 2003). With the simulated data in this study, the log-ANOVA test was found to be the most effective simple test for detecting major gene segregation in diallel population with half-sib structures. These biometric tests may be useful to screen breeding populations for possible major gene segregation, and then the QTL detection could be carried out with these specific pedigrees. This approach should increase the chances of tagging meaningful genes that control quantitative traits (Zeng and Li 2003) and reduce the laboratory costs of QTL detection.

The application of QTLs for breeding requires that polygenic breeding values be integrated with breeding values for QTLs (Dekkers 1999). Inheritance can be described using a mixed model, which explains phenotypic variation with a single gene effect, a polygenic effect, and an error effect attributed to environmental variation (Zeng and Li 2003). Breeding value depends on both gene frequency and effect (Falconer 1989). The effect of gene substitution, α , is $a + d(q - p)$, where a and d are the additive and dominance effects, and p and q are gene frequencies, assuming a simple one locus, two allele model. A QTL allele with breeding value will have a substitution effect in many families that have a common parent that is a heterozygote for the QTL (O'Malley and McKeand 1996). Most QTL studies in plants have not assessed substitution effects in related families and those that have often find many QTLs specific to the "tester" family (Austin et al. 2000; Lubberstedt et al. 1998; Marsan et al. 2001).

DNA markers have made QTL analysis feasible for many species. Different marker methods have advantages for different purposes. Random Amplified Polymorphic DNA (RAPD) markers provide a powerful means to resolve many anonymous markers in a family with little prior knowledge (Williams et al. 1990). RAPD markers have been used extensively to "tag" QTLs, resistance genes, and morphological trait genes (Ranade et al. 2001). RAPDs are relatively inexpensive DNA markers, and genotyping cost is an important determinant of MAS efficiency (Xie and Xu 1998). RAPDs usually have a dominant mode of inheritance, but provide many parent specific, 1:1 segregating markers in a pseudo testcross configuration (i.e., one parent is a heterozygote and the other parent is a homozygote band absent) (Grattapaglia and Sederoff 1994). One disadvantage of RAPD markers is that the linkage groups from the two parents are

difficult to align because the proportion of RAPD markers that segregate in both of a pair of parents is small (O'Malley and Whetten 1997). Aligning maps and locating QTLs on a genomic map were not important for our objectives. RAPDs can be difficult to reproduce across laboratories, but often can be converted to sequence characterized amplified regions (SCARs, Paran and Michelmore 1993). SCARs are DNA fragments derived from RAPD polymorphisms that are amplified by conventional two-primer polymerase chain reaction.

This part of study is to assess the feasibility of QTL detection in a loblolly pine breeding population with both biometric and DNA marker approaches. Phenotypic data on 6-year height of two loblolly pine diallel populations were analyzed to identify a family likely to have major gene segregation, and then RAPD markers were scored and associated with tree height. If QTL with relatively large effects could be detected in progeny from the selected family and then validated with other related families, then such QTLs would be useful in the loblolly pine breeding program.

Materials and Methods for DNA isolation

DNA samples were prepared from parents and families in one selected diallel based on the statistical tests above. One full-sib family, 5 x 6 was chosen as the mapping population because its large variance and intermediate mean. Needles were collected from 115 progeny of family 5 x 6 and their parents. DNA was isolated using the method of Doyle and Doyle (1987). Additional needles were sampled from two tests for all 15 families (sample size ranged from 30 to 42), and DNA was extracted. DNA concentration was estimated by visual comparison of fluorescence under ultra violet light after electrophoresis in 0.8% agarose gels stained with ethidium bromide.

Molecular marker analysis

DNA samples from 96 progeny were diluted to 1 ng/μl. RAPD reactions were carried out following Williams et al. (1990) with some modifications. The volume for each of the reactions was 15 μl, including 5 μl DNA solution and 10 μl of a master mix. The 10 μl of master mix consisted of 1.5 μl BSA (10 mg/ml), 0.9 μl MgCl₂ (25mM), 1.5 μl 10x RAPD reaction buffer, 0.9 μl dNTP (3.3mM for each nucleotide), 5 μl primer (6 ng/μl), 0.05 μl dH₂O and 0.15 μl Taq polymerase (5 U/μl). The 10x RAPD reaction buffer contained 10 mM Tris and 1 mM EDTA. The thermal cycler program was: 92°C, 1 min; 35°C, 1 min; 72°C, 2 min; 41 cycles.

Electrophoresis was carried out using 1.5% agarose gels, 1x TBE buffer, and 150 volts for 4-5 hours. Gel images were recorded with a CRT and printed for scoring. Intense, polymorphic bands that segregated in approximately a 1:1 ratio were scored by visual inspection. Polymorphic bands that segregated in approximately a 3:1 ratio (by χ^2 test) with band present in both parents were treated as F₂ markers. Polymorphisms were excluded from the analysis if the segregation departed significantly from the expected Mendelian ratios ($P < 0.05$). DNA was extracted from RAPD bands cut from gels for 4 markers, M71, M135, M137, and M237. The DNA fragments were cloned and sequenced to convert them to a sequence characterized amplified regions (SCARs, Paran and Michelmore 1993). Primers were designed to specifically amplify these fragments. The forward and reverse primer sequences for the SCAR marker corresponding to

M135 were GGCCTCACGTAGCAGAAAAC and TAACACGACCAACCTGCTGA (both 5' to 3').

Linkage map construction

The mapping population in this study was the progeny from a full-sib cross of two parents (Figure 1), with no knowledge of grandparental genotypes. The linkage phase was unknown for the molecular marker data. Coupling linkage phase was assigned if $P[\chi^2_{d.f. = 1} \geq G] \leq \alpha$ and $\theta < 0.5$, where G is the likelihood ratio calculated assuming coupling linkage, and α is the significance level assuming that linkage phase is in coupling. Otherwise, repulsion linkage phase was assigned (Liu 1998). Criteria for grouping markers were recombination frequency $q < 0.28$ and significance level $P < 0.0001$. Markers were ordered using PGRI (Liu 1998). Confidence for gene order was estimated using a jackknife-bootstrap approach. The multi-locus recombination frequencies were estimated by maximum likelihood (ML). The map distances were computed from the multilocus estimates using Haldane's mapping function. Map coverage was estimated by dividing the sum of the map distances by 1800 cM, the best estimate of the pine genome map length (Remington et al. 1999).

QTL analysis

Single marker QTL analyses were performed using simple linear regression, and multiple linear regressions. The model for simple linear regression was:

$$y_i = \mu + B_0 + B_1x_i + \varepsilon$$

where y_i was the trait value for the i^{th} individual, B_0 was the intercept (mean), B_1 , the regression coefficient, x_i was an indicator variable for marker genotype for the i^{th} individual ($x_i = 1$ or 2), and ε was the error term. Both simple linear regression analysis on single markers and composite interval mapping on linkage groups were carried out using QTL Cartographer (Basten et al. 1994).

QTL validation

The other families within the diallel that shared common parent 5 were used as the validation population). Markers linked to QTLs in the mapping family 5 x 6 that segregated in either population were analyzed for QTL effects using a general linear model that included family and family x marker interaction terms (SAS GLM and PROC MEANS, SAS Institute Inc. Cary NC 1996). Type III sum of squares was used in the analysis to avoid possible confounding of effects between family and marker (Rawlings 1988).

Results of QTL Detection

With DNA samples from progenies and parents of this full-sib family 5 x 6, a linkage map was made using RAPD markers. RAPD gels for 87 primers yielded 270 polymorphic bands (Table 3, Figure 3a). There were 27 polymorphisms in apparent agreement with 3:1 segregation ($P \leq 0.05$). The 1:1 bands originated in almost equal numbers from the maternal and paternal parents (121 vs. 122). The 223 polymorphisms that were scored as genetic markers with 1:1 segregation were used for map construction and QTL analysis (Table 1). Four RAPD markers, M71, M135, M137,

and M237 were chosen for conversion to SCARs because they were associated with phenotypic effects in QTL analysis. Only marker M135 yielded a useful SCAR marker, S135 (Figure 3b).

Table 1. Primer names (Operon), primer sequences, and polymorphic markers amplified

Primer Name	Sequence	Markers
A5	AGGGGTCTTG	M17 M18 M19 M20 M21
A9	GGGTAACGCC	M122
A10	GTGATCGCAG	M14 M15 M16
A14	TCTGTGCTGG	M10 M11 M12 M13
A15	TTCCGAACCC	M22 M23
A19	CAAACGTCGG	M01 M02 M03 M04
A20	GTTGCGATCC	M07 M08 M09
B2	TGATCCCTGG	M121
B5	TGCGCCCTTC	M25
B8	GTCCACACGG	M248 M249 M250 M251 M252
B12	CCTTGTACGCT	M116 M117
B13	TTCCCCCGCT	M118 M119 M120
B16	TTTGCCCGGA	M253 M254
B20	GGACCCTTAC	M32 M33 M34
C6	CCAGAACGGA	M 28 M29 M30 M31
C12	TGTCATCCCC	M35 M36
C16	CACACTCCAG	M37 M38 M39 M40 M41
D3	GTCGCCGTCA	M51
D10	GGTCTACACC	M48 M49
D11	AGCGCCATTG	M50
D20	ACCCGGTCAC	M46 M47
E1	CCCAAGGTCC	M58
E12	TTATCGCCCC	M54 M55 M56 M57
E18	GGACTGCAGA	M52 M53
F1	ACGGATCCTG	M69 M70
F3	CCTGATCACC	M59 M60 M61 M62 M63
F5	CCGAATTCCC	M64 M65 M66 M67 M68
F6	GGAATTCGG	M269 M270
F13	GGCTGCAGAA	M71 M72
G10	AGGGCCGTCT	M73 M74
G11	TGCCCCGTCGT	M86 M87 M88 M89
G12	CAGCTCACGA	M111 M112 M113
G13	CTCTCCGCCA	M84 M85
G15	ACTGGGACTC	M114 M115
J5	CTCCATGGGG	M110
J6	TCGTTCCGCA	M97 M98
J7	CCTCTCGACA	M105 M106 M107 M108 M109
J10	AAGCCCGAGG	M93 M94 M95 M96
J11	ACTCCTGCGA	M90 M91
J13	CCACACTACC	M125 M126 M127 M128 M130
J18	TGGTCGCAGA	M92 M158 M159 M160
K3	CCAGCTTAGG	M99 M100 M101
K5	TCTGTCGAGG	M102 M103 M104
K9	CCCTACCGAC	M75 M76 M77 M78
K11	AATGCCCCAG	M79 M80
K15	CTCCTGCCAA	M81 M82 M83
P1	GTAGCACTCC	M131 M132 M133
W6	AGGCCCCGATG	M242 M243 M244
W11	CTGATGCGTG	M140 M141 M142

AA2	GAGACCAGAC	M225 M226 M227
AA8	TCCGCAGTAG	M235 M236 M237 M238
AB12	CCTGATCCGA	M123 M124
AB14	AAGTGCGACC	M239 M260
AB18	CTGGCGTGTA	M215 M216 M217 M218
AB19	ACACCGATGG	M134 M135 M136 M137 M138
AB20	CTTCTCGGAC	M219 M220 M221
AC9	AGAGCGTACC	M161 M162
AC19	AGTCCGCCTG	M187 M188 M189 M190
AH9	AGAACCGAGG	M240 M241
AH19	GGCAGTTCTC	M204 M205 M206
AJ10	GTTACCGCGA	M222 M223 M224
AJ15	GAATCCGGCA	M228 M229 M230
AK4	AGGGTCGGTC	M167 M168 M169
AK6	TCACGTCCCT	M143 M144 M145
AK10	CAAGCGTCAC	M194 M195 M196 M197
AM1	TCACGTACGG	M210 M211 M212 M213 M214
AM6	CTCGGATGT	M231 M232 M233 M234 M234
AM19	CCAGGTCTTC	M191 M192 M193
AP1	AACTGGCCCC	M207 M208 M209
AP6	AGGCAGCCTG	M146 M147 M148 M149
AP7	ACCACCCGCT	M151 M152 M258
AP9	CCAGATGGGG	M153 M154 M155
AP11	CTGGCTTCTG	M163 M164 M165 M166
AP20	GGCTTGCCTG	M156 M157 M257
AQ2	ACCCTCGGAC	M177 M178 M179
AQ18	GGGAGCGAGT	M180 M256
AQ19	AGTAGGGCCT	M181 M182 M183 M184
AR4	CCAGGAGAAG	M261
AR8	GTGAATGCGG	M170 MM171 M172
AS6	GGCGCGTTAG	M265 M266 M267
AS13	CACGGACCGA	M198 M199 M200 M201
AS18	GTTGCGCAGT	M202 M203
AT15	TGACGCACGG	M245 M246 M247
AV3	TGTAGCCGTG	M263 M264
AV6	CCCAGATCC	M262
AV13	CTGACTTCCC	M185 M186
AV15	GGCAGCAGGT	M268
AV20	TCATGCGCAC	M139 M255 MM259

Parent specific linkage maps were constructed for full-sib family 5 x 6 from the 121 RAPD markers segregating from 5 and 122 RAPD markers segregating from parent 6. Using the criteria $\theta \leq 0.28$ and $P < 0.0001$ for significant linkage, all but 28 markers were assigned to groups of at least 2 markers (Figure 4). The number of linkage groups (≥ 3 markers) from parent 5 was 19, and the number of linkage groups from parent 6 was 15 (Table 3). Pairwise recombination matrices were analyzed for each linkage group to determine marker order. The sum of map distance for all linkage groups was 2328 cM, or approximately 1164 cM per parent.

Figure 3. RAPD polymorphisms. A. Gel image for primer AA8 showing variation in RAPD bands for family 6 x 5. The first and the last lane contained molecular weight markers (band sizes in bp are shown to the left). B. SCAR marker S135 (arrow) segregating in family 1 x 5. The first lane (left) contained molecular weight markers (band sizes in bp are shown to the left).

QTL analysis

Marker and trait association was analyzed for family 5 x 6 using 6-year adjusted height data ($n = 96$). Simple linear regression showed that 11 markers were significantly associated with height ($P < 0.05$, Table 4). All markers except M23 were from parent 5, with 8 markers in two linkage groups, 16 and 27. The other 3 were in linkage groups 1, 16 and 11. Using all 223 markers, stepwise multiple regression found three markers, M71, M137 and M237, that explained a significant amount of phenotypic variation at age six (Table 5). The coefficient of determination was 21.4% for the whole model, and R^2 for the individual markers was 6.0, 6.9 and 5.1% respectively. The effects for the 3 markers were 0.53, 0.64, and 0.75 meters, respectively. M137 and M135 were closely linked and had nearly the same genetic effect.

Table 4. Simple regression results from QTL Cartographer. The 11 markers with significant genetic effects ($P < 0.05$) were from 5 linkage groups. All the markers were from parent 5 except for M8.

Marker	Linkage Group	Effect (B_1)	LR	Prob.
M8	1	0.481	4.812	0.030
M23	11	0.472	4.69	0.033
M129	16	-0.623	8.795	0.003*
M162	16	-0.560	6.657	0.011
M231	16	0.686	10.133	0.002*
M237	16	0.746	12.103	0.001*
M85	16	0.605	7.673	0.006
M 71	26	0.526	5.877	0.017
M135	27	0.584	7.275	0.008
M137	27	0.640	8.817	0.003*
M175	27	0.534	6.047	0.015

Table 5. Multiple regressions of height on RAPD markers for ages six and eight.

Age	Marker	SS	b	R square	P-value
6	M71	6.629	0.266	0.060	0.010
	M137	7.651	0.291	0.069	0.006
	M237	5.663	0.250	0.051	0.016
	Coefficient of Determination = 0.2139 $P < 0.0001$				
8	M71	13.689	0.388	0.079	0.005
	M137	13.026	0.381	0.075	0.006
	Coefficient of Determination = 0.1580 $P < 0.0004$				

Figure 4. Linkage groups determined for RAPD markers segregating in family 5 x 6. The linkage groups are labeled with the group name and parent ((p1 as parent 5), and the length in cM. The distance between markers (cM) is shown to the left of each group and marker names to the right.

The association of markers and phenotypes were also analyzed using adjusted height data from age eight ($n = 93$). The markers M71, M135 and M137 showed significant departures from the null hypothesis of no effect at age 8 as well as age 6 (Table 4). Multiple regressions at age eight included M71 and M137 but excluded M237 and had a coefficient of determination of 15.8% (7.9% and 7.5% for M71 and M137 respectively). Compared with age six, the magnitudes of effects for M71, M135 and M137 were larger at age eight.

Interval and composite interval mapping yielded a QTL peak over M135 in linkage group 27 that was significant at the $P < 0.05$ level, genome-wise (Figure 5a). The significance was determined from 500 permutations (QTL Cartographer) and corresponds with LOD 2. The QTL peak in linkage group 16 was near M237 and also was significant at the $P < 0.05$ level genome-wise (Figure 5b).

Figure 5. Composite interval maps for linkage groups 16 (A.) and 27 (B.) from the mapping family showing map locations and likelihood ratios. The $P < 0.05$ level genome-wise significance level was determined by permutation using QTL Cartographer.

QTL validation

Each of the six parents in diallel 2 was genotyped for markers S135 and S237. The S135 marker was present in parents 5 and 3. Family 3 x 5 segregated 3:1 present/absent for S135 and was not used for validation. The analysis of variance for progeny from the other three full-sib families sharing the parent 5 (i.e., the validation population, Vpop) showed that the S135 marker explained a significant amount of variation in height with no significant marker by family interactions ($P < 0.01$, Table 6). S135 explained approximately 5.6% of phenotypic variation in height. Relative to the phenotypic standard error, the effect of marker S135 on height for the Vpop was approximately $0.5 \sigma_p$. For age 8, the effect of marker S135 was almost significant at the $P < 0.05$ level. S135 explained approximately 3% of the total variation in height at age 8 for the Vpop.

Table 6. The family and marker (S135) effects in the analysis of variance on 6-year height in the validating population.

Population	Source	DF	Sum of Squares	Mean square	F Value	Pr > F
Vpop sharing 5	Model	5	18.8841	3.7768	4.64	0.0007
	Error	101	82.2489	0.8143		
	Corrected Total	106	101.1330			
	R-Square = 0.1867 Root MSE = 0.9024 Mean = -0.1443					
	FAMILY	2	12.8297	6.4149	7.88	0.0007
	S135	1	5.6670	5.6670	6.96	0.0097
	FAMILY*S135	2	0.0630	0.0315	0.04	0.9621

Two of the six parents in diallel 2, parents 5 and 2 were band present for M237, while the other four were band absent. Similar to the M135/SCAR marker, nine families were segregating for M237, including 5 x 6. The ANOVA for M237 showed neither a significant main effect, nor an interaction effect of M237 with family for the Vpop. There were no significant effects of M235 at age 8 for the Vpop.

Discussion of Molecular Confirmation

The molecular marker mapping study on family 5 x 6 provided supporting evidence from the biometric analysis for major gene segregation (QTL) in diallel 2. The segregation of 3 RAPD markers from parent 5 explained 21% of the within family phenotypic variation in adjusted 6 year height for cross 5 x 6 in diallel population 2. Composite interval mapping yielded QTL peaks significant at the $P < 0.05$ level genome-wide for two linkage groups and these two QTLs explained a significant amount of variation in 8 year height. The mean height of the mapping family was 7.3 m and the QTL effects were approximately 0.5 – 0.7 m. The QTLs associated

with M137 and M235 occurred in linkage groups that permitted composite interval mapping (Basten et al. 1994). By this method, the QTLs were significant at the $P < 0.05$ level genome-wide (i.e., a false QTL peak will meet this rejection criterion only once in 20 genomes studied). The significant QTL effects were transmitted by only parent 5.

The validation of the QTLs from parent 5 was important in this experiment because of the small progeny size of the full-sib ($n = 96$) and the large number of markers analyzed (223). The M137 QTL in linkage group 27 was validated in a sample of progeny from 3 families that shared parent 5 as a common parent. The QTL effect was approximately $0.5 \sigma_p$, compared with $0.7 \sigma_p$ in the mapping family, suggesting that the effect was overestimated in the mapping family (Beavis 1994), or that QTL and genetic background interaction existed (Leips and Mackay 2000, Yamamoto et al. 2000). The effect could also appear smaller because the phenotypic variance is larger for a set of 3 families than for a single mapping family.

The 223 DNA markers scored from 87 RAPD primers provided over 90% genomic map coverage for genetic dissection of phenotypic variation in the 5 x 6 mapping population. The 223 markers yielded 34 linkage groups that contained 198 markers which spanned a total distance of 2328 cM, accounting for 65% of the map length predicted for the two parents based on an estimated genomic map length of ~1800 cM (Remington et al. 1999). In addition, the markers at the ends of linkage ($34 \times 2 = 68$ markers) provide additional coverage (about 10 cM), and there were 25 unlinked markers that could cover at least 10 cM in either direction (assuming that the polymorphisms scored as unlinked markers were not artifacts). The sequence of RAPD marker M135 was determined to ensure the reproducibility of this marker using molecular methods.

The M135/M137 QTL effect transmitted by parent 5 supports our biometrical approach that identified family 5 x 6 as a good candidate for mapping. However, the magnitude of the effect seemed small, and the relationship between QTL effect and heterogeneity of variance among families in a half-diallel population needs further study. Variance heterogeneity is expected if genes with a major quantitative effect are segregating within some families and not in others (Fain 1978, Leroy and Elsen 1992, Zeng and Li 2003). The frequency distribution for tree height within family had the skewed distribution typical of within family height data and did not appear bimodal. The map coverage for family 5 x 6 suggests that a major QTL is unlikely to have been missed. One important conclusion from our study is that QTL effects can be detected using phenotypic data from a mapping population comprised of trees planted at large distances apart in a well-designed and well-executed field trial.

The utility of such QTLs for breeding, however, should not be inferred from a mapping study in a single, small sized family. Many QTL effects discovered in a mapping population are difficult to reproduce in replicate populations (Melchinger et al. 1998). QTL discovery has low statistical power and QTL effects are often biased upwards (Beavis 1994). The large number of independent variables (i.e., markers) and the small number of progeny ($n = 96$ for our mapping population) can cause problems for QTL analysis using linear models. Parameter estimates for over-fitted models have larger variances, while models that omit genetic effects (i.e., latent predictor variables) can yield biased estimates of parameters, predictions, and error variance. Inconsistent expression of QTLs in related families is expected for false QTLs, gene interactions, or QTLs that are recessive. QTL effects that result from epistatic interactions would depend on

genetic background. The expression of QTLs could also appear to be inconsistent in replicates due to genotype x environment interaction or inconsistent at different ages due to developmental specificity. Setting a high threshold for rejection of the null hypothesis (H_0 : no QTL effects) increases the probability that the QTLs discovered will be real, but that some real QTLs will be missed.

Parent 5 had the lowest breeding value in diallel 2, assuming polygenic inheritance, but the estimates of polygenic breeding values could be affected by undetected major gene segregation (Dekkers 1999). Breeding value depends on both the magnitude of the genetic effect and gene frequency. A QTL allele with a large breeding value would have a large additive or dominant effect and low to intermediate gene frequency. For parent 5, the M135/M137 QTL effect was segregating in the 4 families studied from diallel 2, as would be expected for a low frequency allele. (Both parents of family 4 x 5 were heterozygotes for the S135 marker so it was not used for analysis.) QTL effects measured this way are averaged over the possible contribution of segregation at the corresponding locus in the other parent or parents (Plomion and Durel 1996). The difference between QTL genotypic means was approximately 7% (mean difference ~ 0.5 m for a 6 year height of approximately 7 m). Assuming additive effects and that the frequency of the favorable QTL allele is 1/12 in diallel 2, then fixation of the favorable QTL allele could increase 6-year height by approximately 0.9 m (13%). The sample size of parents mated to 5 is small, but in principle, the M135/ M137 QTL effect is only $\alpha/2$, where α is the average effect of gene substitution (i.e., 1/2 of the breeding value of the QTL). If the M135/M137 QTL is dominant, its breeding value will be approximately half as large as an additive QTL.

Within family MAS could be economically feasible when h^2 is low, and the proportion (p) of the additive genetic variance (σ_A^2) explained by markers is large (Lande and Thompson 1990, Kumar and Garrick 2001). However, p is difficult to estimate within pedigrees of outbred forest trees, and could vary among families and populations if QTLs make a large contribution to σ_A^2 . Heritability of 6 year height in loblolly pine is approximately 0.2. Relative to the population level additive genetic variance, the M135/M137 QTL effect is approximately $2.5 \sigma_A$, assuming an additive model of inheritance. A QTL effect of this magnitude could have value for breeding. Two critical pieces of information are missing from our analysis and will be pursued in future studies. We do not know the degree of dominance for the inheritance of this QTL and whether the favorable QTL allele is rare (i.e., has breeding value). Families from a third generation of this pedigree would be required to separate the two QTL alleles, Q and q , near S135 and to determine whether S135+ is associated with a substitution effect.

This study showed that QTLs with breeding value can be discovered in tree breeding populations, despite the small numbers of progeny per family and the widely dispersed planting sites. The use of biometrical analysis in a breeding population may be useful to identify families likely to have a major gene segregating, and thus increase the odds for successful detection of QTLs in molecular analysis. For dairy cattle, the cost of establishing and maintaining separate QTL mapping populations is prohibitively expensive, so QTL discovery is carried out in the populations used for breeding (review by Georges 1999). This study showed the potential for a similar approach in forest trees. The availability of related families and multi generation pedigrees in tree breeding programs should facilitate the validation of QTL effects. The magnitude of the M135/M137 QTL effect is small but in the context of a low heritability trait, its

potential value could be large. Ultimately, the application of QTLs in tree breeding will depend on economic considerations and the system of propagation and deployment (Johnson et al. 2000, Kumar and Garrick 2001, Wilcox et al. 2001). Screening for evidence of major gene segregation, coupled with DNA marker analysis could provide new opportunities for tree breeding.

Literature Cited

- Austin DF, Lee M, Veldboom LR, Hallauer AR (2000) Genetic mapping in maize with hybrid progeny across testers and generations: Grain yield and grain moisture. *Crop Sci* 40: 30-39.
- Balocchi C.E., Bridgwater F.E., Zobel B.J. & Jahromi S. 1993: Age trends in genetic parameters for tree height in nonselected population of loblolly pine. *Forest Science* 39: 231-251.
- Bartlett MS (1937) Some examples of statistical methods of research in agriculture and applied biology. *J Royal Society* 4: 137-170 (Supplement)
- Basten CJ, Weir BS, Zeng Z-B (1994) Zmap-a QTL Cartographer, pp 65-66 in 5th World Congress on Genetics Applied to Livestock Production; Computing Strategies and Software, Edited by Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W. Organizing Committee, Guelph, Ontario, Canada
- Beavis WD (1994) The power and deceit of QTL experiments: Lessons from comparative QTL studies, pp 250-266 in 49th annual Corn and Sorghum industry Research Conference. American Seed Trade Association
- Bradshaw HD, Stettler RF (1995) Molecular genetics of growth and development in populus.4. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* 139: 963-973
- Byrne M, Murrell JC, Owen JV, Kriedemann P, Williams ER, Moran G (1997) Identification and mode of action of quantitative trait loci affecting seedling height and leaf area in *Eucalyptus nitens*. *Theor Appl Genet* 94: 674-681
- Dekkers JCM (1999) Breeding values for identified quantitative trait loci Under Selection. *Genet Sel Evol* 31: 421-436
- Doyle JJ, Doyle JL (1987) Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15
- Fain PR (1978) Characteristics of simple sibship variance tests for detection of major loci and application to height, weight and spatial performance. *An J Hum Genet* 42: 109-120
- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. 4th edn. Longman Scientific & Technical, Harlow, pp 108-121
- Georges M (1999) Towards marker assisted selection in livestock. *Reprod Nutr Dev* 39: 555-561
- Grattapaglia DF, Bertolucci LG, Penchel R, Sederoff RR (1996) Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. *Genetics* 144: 1205-1214
- Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. *Genetics* 137: 1121-1137
- Johnson GR, Wheeler NC, Strauss SS (2000) Financial feasibility of marker-aided selection in Douglas-fir. *Can J For Res* 30:1942-1952
- Kaya Z, Sewell MM, Neale DB (1999) Identification of quantitative trait loci influencing annual height- and diameter-increment growth in loblolly pine (*Pinus Taeda* L.). *Theor Appl Genet* 98: 586-592
- Kumar S, Garrick DJ (2001) Genetic response to within-family selection using molecular markers in some radiata pine breeding schemes. *Can J For Res* 31:779-785
- Lande R, Thompson R (1990) Efficiency of marker assisted selection in the improvement of quantitative traits. *Genetics* 124: 743-756
- Leips J, Mackay TFC (2000) Quantitative trait loci for lifespan in *Drosophila melanogaster*: Interactions with genetic background and larval density. *Genet* 155: 1773-1788
- Lerceteau E, Plomion C, Andersson B (2000) AFLP mapping and detection of quantitative trait loci (QTLs) for economically important traits in *Pinus sylvestris*: a preliminary study. *Mol Breed* 6: 451-458
- Leroy P, Elsen JM (1992) Simple test statistics for major gene detection - a numerical comparison. *Theor Appl Genet* 83: 635-644
- Li B, McKean SE, Weir RJ (1996) Genetic parameter estimates and selection efficiency for loblolly pine breeding in the southeastern U.S, pp 164-168 in QFRI-IUFRO Conference: Tree Improvement for Sustainable Tropical

- Forestry, Edited by Dieters MJ, Matheson. AC, Nikles DG, Harwood CE, Walker. SM. Queensland Forestry Research Institute; Gympie; Australia, Caloundra, Queensland, Australia.
- Li B, McKeand SE, Weir RJ (1999) Tree improvement and sustainable forestry -impact of two cycles of loblolly pine breeding in the US. *Forest Genetics* 6: 229-234.
- Liu B-H (1998) Statistical genomics: Linkage mapping and QTL analysis. CRC Press, Boca Raton, Fl pp 216 -218, 529-544.
- Lubberstedt T, Melchinger AE, Fahr S, Klein D, Dally A, Westhoff P (1998) QTL mapping in testcrosses of flint lines of maize: III. Comparison across populations for forage traits. *Crop Sci* 38: 1278-1289.
- Lynch M, Walsh B (1998). *Genetics Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, MA, pp 980.
- Marsan PA, Gorni C, Chitto A, Redaelli R, Van Vijk R, Stam P, Motto M (2001) Identification of QTLs for grain yield and grain-related traits of maize (*Zea Mays* L.) using an AFLP map, different testers, and cofactor analysis. *Theor Appl Genet* 102: 230-243
- Martin CG, Games PA (1977) Anova tests for homogeneity of variance: Monumentality and unequal samples. *J Educ Stat* 2: 187-206
- McKeand, S, Bridgwater F (1998) A strategy for the third breeding cycle of loblolly pine in the southeastern U.S. *Silvae Genetica* 47: 223-234
- Melchinger AE, Utz HF, Schon CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149: 383-403
- O'Malley D, McKeand S (1996) Marker assisted selection for breeding value in forest trees. *For Genet* 1: 207-218
- O'Malley D, and R. Whetten (1997) Molecular markers and forest trees, in *DNA markers: Protocols, applications, and overviews*, edited by G. Caetano-anolles and P. Gresshoff. Wiley-Liss, inc, New York, pp 237-257.
- Paran I, Michelmore RW (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. *Theor Appl Genet* 85: 985-993
- Plomion C, Durel CE (1996) Estimation of the average effects of specific alleles detected by the pseudo-testcross QTL mapping strategy. *Genet Sel Evol* 28: 223-235
- Ranade SA, Farooqui N, Bhattacharya E, Verma A (2001) Gene tagging with random amplified polymorphic DNA (RAPD) markers for molecular breeding in plants. *Crit Rev Plant Sci* 20: 251-275
- Rawlings JO (1988) *Applied regression analysis: A research tool*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, Ca. 285p.
- Remington DL, Whetten RW, Liu BH, O'Malley DM (1999) Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. *Theor Appl Genet* 98: 1279-1292
- Sewell MM, Bassoni DL, Megraw RA, Wheeler NC, Neale DB (2000) Identification of QTLs influencing wood property traits in loblolly pine (*Pinus taeda* L.). I. Physical wood properties. *Theor Appl Genet* 101: 1273-1281
- Sun S (2001) Detecting major genes for growth rate with RAPD markers using a diallel cross of loblolly pine, in *Departments of Forestry and Crop Science*. North Carolina State University, Raleigh, pp 125.
- Verhaegen D, Plomion C, Gion JM, Poitel M, Costa P, Kremer A (1997) Quantitative trait dissection analysis in eucalyptus using RAPD markers .1. Detection of QTL in interspecific hybrid progeny, stability of QTL expression across different ages. *Theor Appl Genet* 95: 597-608
- Wilcox PL, Carson SD, Richardson TE, Ball RD, Horgan GP, Carter P (2001) Benefit-cost analysis of DNA marker based-selection in progenies of *Pinus radiata* seed orchard parents. *Can J Forest Res* 31: 2213-2224.
- Williams J, Kubelik GAR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl Acids Res* 18: 6531-6535
- Xiang, B and Li B (2001) A new mixed analytical method for genetic analysis of diallel data. *Canadian J. For. Research* 31:1-8.
- Xie CQ, Xu SZ (1998) Efficiency of multistage marker-assisted selection in the improvement of multiple quantitative traits. *Heredity* 80: 489-498
- Yoshimaru H, Ohba K, Tsurumi K, Tomaru N, Murai M, Mukai Y, Suyuma Y, Tsumura Y, Kawahara T, Sakamaki Y (1998) Detection of quantitative trait loci for juvenile growth, flower bearing and rooting ability based on a linkage map of *Sugi* (*Cryptomeria japonica* D Don). *Theor Appl Genet* 97: 45-50
- Yamamoto T, Lin HX, Sasaki T, Yano M (2000) Identification of heading date quantitative trait locus Hd6 and characterization of its epistatic interactions with Hd2 in rice using advanced backcross progeny. *Genet* 154: 885-891

MAPPING MAJOR GENES WITH MICROSATELLITE MARKERS

In mapping major genes affecting height that are segregating in breeding program populations, the probability that a randomly chosen marker located within 20 cM of a major gene is small, perhaps 0.02. Thus, only 1 out of 50 markers are expected to be linked with the trait, and hundreds would be needed to ensure that we will get at least one that is linked. AFLP markers are ideal for tagging genes within a single family. The markers are highly multiplexed and efficiently produced. However, family sizes in the breeding program are small and the heritability is small. The key to mapping major genes will be combining information from several full sib families that share a common parent. Only a portion of the AFLP markers from one family will segregate in another. Mapping would be extremely complex. Microsatellite markers are ideal for this mapping problem because they could mark the same map location in all of the families. The markers are multiallelic, codominant, and highly variable.

Microsatellite markers are the best marker system to analyze QTL allele substitution effects for the common parent in a set of half-sib families from a diallel population. Not all microsatellite markers will be useful for our purposes. Most important, the common parent must be heterozygous for the marker. If the two alleles that comprise the common parent's genotype are unique, the common parent's contribution to each progeny can be determined unambiguously. The same is true if the common parent has only one unique allele, unless the other parent has a nonamplifying allele. In this case, the number of progeny with a "homozygous" band type will be inflated and we won't know which genomic region was contributed by the common parent. Substitution effects from the "other" parent can be described if it is a heterozygote for the marker and has at least 1 allele different from the common parent. These substitution effects are not our main objective, but accounting for variance due to these effects could help to resolve the common parent substitution effect. The set of progeny microsatellite genotypes from a half-sib family are complex and present challenging problems for data organization and management.

Microsatellite marker analysis is implemented for loblolly pine using Applied Biosystems Inc. (ABI) automated fluorescent fragment analysis. A set of 48 microsatellite primer pairs was synthesized using the sequence information provided on Claire William's microsatellite website at TAMU. The microsatellites were PCR-amplified using the [F]-dUTP labeling method, which incorporates a few dye molecules in each DNA fragment. The advantages of this approach are flexibility and lower cost especially at the early stages of a project. The conventional microsatellite approach is to PCR-amplify microsatellite DNA fragments using primers that are labeled with fluorescent dye. However, dye-labeled primers are 5 to 10 times more expensive to purchase, and many of the microsatellites amplified using these labeled primers could be uninformative for genetic analysis. The disadvantage of the [F]-dUTP labeling method is that the protocol is more technically demanding than for dye-labeled primers and PCR-amplification of several fragments in the same reaction tube (multiplexing) is much more difficult. Dye-labeled primers are the best approach for high-throughput screening. In the future, the ABI methods will allow us to use capillary sequencing instruments available in the NCSU Genomic Research Laboratory. These capillary instruments are ideal for high-throughput genotyping using automated methods, but require well-developed methods. The ABI microsatellite methods complement and extend the DNA marker methods that we have developed for Li-Cor sequencing instruments.

Genotypes of microsatellite loci were determined for a set of 240 DNA samples from a half-diallel cross of loblolly pine. The four families chosen for study all shared parent 9217. A previous study using RAPD markers and a biometrical analysis using MCMC methods both provided evidence that offspring from this parent could segregate for a major gene affecting height growth. RAPD markers are difficult to reproduce and more DNA markers are needed to characterize the variation from parent 9217. Microsatellite markers could be more useful for segregation analysis because they are codominant and segregate in many different genetic backgrounds. We screened half of the 48 TAMU microsatellite primer pairs and identified 10–12 microsatellites for genotyping in our families. The DNA samples are organized in 6 sample sets of 48 samples (some duplications among the 6 sets). The primer pairs were organized into panels (sets) of 3 microsatellites, each locus to be amplified using one of the 3 dyes. PCR-amplification of microsatellite loci and fluorescent fragment analysis were carried out on the ABI 377 sequencing instrument. We used 48 lane gels, Long Ranger Singel packs, and membrane combs. Amplification of one sample set has been difficult and could require new DNA preps. Panel 1 and 2 have been analyzed on gels for 5 of the 6 sample sets. Genotyping has been carried out on the gel images for panel 1. The gel images were “tracked” and peaks were analyzed using GeneScan software to estimate the molecular weight (bp) of the DNA fragments. The genotypes were “called” using GenoTyper software.

We analyzed the segregation of microsatellites from the 4 families. First, we reconciled parent genotypes and family segregations to identify “Mendelian errors” that could be due to loss of sample identity or artifacts of techniques (e.g., bands that bleed from one lane to the next or from one dye to another). Artifacts are greatly reduced using capillary sequencers. We determined that approximately 5% of the samples have illegitimate genotypes originating from errors in field collection of the samples or loss of identity in the original field trial. Correcting this level of error could substantially increase the power of QTL analysis to detect major gene segregation. Unrelated progeny have now been removed from our sample sets. Another problem is that some of the microsatellite loci had non-amplifying alleles in some parents. The parent genotype in these cases appeared to be homozygous, but the progeny segregation data contained homozygotes for both alleles of the other parent in ratios compatible with this genetic model. We also encountered some microsatellite patterns that were unambiguous but did not conform with Mendelian laws (e.g., 3 alleles in a diploid individual). The mutation rate for microsatellite loci is high and could explain some of these discrepancies. One microsatellite yielded clear Mendelian segregations in some families, but ambiguous patterns in another family. We underestimated the extent of repeat runs needed in establishing panels of markers sample sets for genotyping.

The SSR markers were analyzed for QTL detection in this loblolly pine population. To analyze marker data as a half-diallel structure, these populations contain 15 full sib families that can be combined to make 6 half-sib families. Large numbers of progeny per family (hundreds) are needed for marker analysis and QTL detection. In animal breeding, the contribution of only the common parent in the half-sib family is tracked in marker analysis because the small sizes of the full-sib families preclude meaningful analysis of substitution effects from the other parents. In pine breeding, the family sizes are too large to ignore (about 100 in our study). One important

question that our research is now addressing is whether analysis that exploits marker information from full sib families within half sib families is useful.

The marker analysis approach is based on simple linear regression using indicator variables. In the simplest case where a marker in parent 1 is heterozygous and parent 2 is homozygous for the marker, the progeny will be either heterozygotes or homozygotes (2 classes of progeny). The marker substitution effect is the difference in means for the 2 classes of progeny (i.e., if the marker was changed in one class of progeny, the phenotype would change by this amount). The model is $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$. β_1 is the substitution effect (the difference between the means of the 2 progeny classes). X_1 is an indicator variable (numerical) that takes value 1 if allele A1 is present and 0 if allele A2 is present.

The model for a half-sib family from a half-diallel breeding population is considerably more complicated than the simple regression model. Both the common parent and the specific parent could be segregating for a QTL in a full-sib family. Our model takes account of the potential for interactions between effects in different families, as well as substitution effects from each parent (Table 1). The interpretation of the markers and the effects is critical. We have no basis to assume that marker alleles with the same name track the same effects in different families. We assume that there are only 2 QTL alleles in the population. This perspective is justified by the major gene detection work carried out using biometrical data in Task 1. We do not know the parental genotypes for the QTL, nor their association with the marker alleles in different families. Each substitution effect has a corresponding dependent variable in the model. This variable takes value +1/2 if allele A_i was transmitted to the progeny and -1/2 if the other allele was transmitted by that parent. If the progeny comes from a family with a different parent, then the indicator takes value 0. The assumption that there are only 2 QTL alleles enables us to estimate an interaction between marker substitution effects from the 2 parents (Table 2). The main value of estimating the effects in the full sib families is to reduce the error variance to enable better detection and estimation of the substitution effect in the common parent.

Table 1. Example showing values of independent indicator variables for linear regression of height on marker substitution effects for a common parent and 2 specific parents, and for a family effect unrelated to marker

Ht	Common Parent	Parent 1	CP x P1 IA	Parent 2	CP x P2 IA	Family
##	A ₁ : +1/2	A ₃ : +1/2	+1/2	0	0	1
##	A ₁ : +1/2	A ₄ : -1/2	-1/2	0	0	1
##	A ₂ : -1/2	A ₃ : +1/2	-1/2	0	0	1
##	A ₂ : -1/2	A ₄ : -1/2	+1/2	0	0	1
##	A ₁ : +1/2	0	0	A ₅ : +1/2	+1/2	0
##	A ₁ : +1/2	0	0	A ₆ : -1/2	-1/2	0
##	A ₂ : -1/2	0	0	A ₅ : +1/2	-1/2	0
##	A ₂ : -1/2	0	0	A ₆ : -1/2	+1/2	0

Table 2. Model for interactions analysis for a full sib family, with parental genotypes $A_1A_2 \times A_3A_4$. Both parents are assumed heterozygotes for a major gene, Q, affecting phenotype. The linkage phase of the marker and Q are unknown (only one possible arrangement shown). Note that the substitution effect for A_1A_2 is parent is $X1 \times [(\text{mean of progeny that got } A_1) - (\text{mean of progeny that got } A_2)]$, which equals $2a$. Similarly, the other substitution effect is $2b$ and the interaction effect is $2d$.

	Q	q	-2d
Q	$A_1 A_3$ $a + b + d$	$A_1 A_4$ $a - b - d$	+2a
q	$A_2 A_3$ $-a + b - d$	$A_2 A_4$ $-a - b + d$	-2a
	+2b	-2b	+2d

SSR marker analysis

Cambiums were collected from 281 progeny and from 4 full sib families with one common parent planted in 3 tests. DNA was extracted using the DNEasy 96 plant kit (QIAGEN Inc., Valencia CA). The Kodak digital ID 3.0.2 USB software after electrophoresis in 1% agarose gel stained with ethidium bromide determined DNA concentration. And then DNA samples were all diluted to 10 ng/ul. Hydro 96 microdispenser was used during DNA concentration determination and dilution.

PCR based SSR reaction: The volume for each reaction was 12ul, including 2ul DNA and 10ul of master mix. The master mix consisted of 1.2ul 10x PCR buffer, 0.096ul 100uM dNTP (four nucleotides A,T,G and C), 0.24ul primers (0.12ul of each of the 50uM forward and reverse primer). , 8.364ul dH₂O and 0.1ul Hotstar Taq polymerase (0.5U/ reaction). The PCR amplification of DNA was performed on a MJ Research thermal cycle, programmed 15 min at 95 °C followed by 36 cycles of 30 s each at 94 C(denaturation), 57°C(annealing), 72°C(extension); and followed by incubation at 72°C for 9 min as a final extension step.

Electrophoresis and SSR scoring: The PCR products were run on ABI sequencer 377(ABI applied Biosystem Inc.,) as shown in Fig.1. 48 lane gels, long Ranger Singel packs, and membrane comb was used. The labeled primer pairs were organized into sets of 3 micorsatellites for each gel. Filter set C (FAM, TET and HEX) was used for primer label. The PCR product was diluted with dH₂O (1:30). 1.5ul diluted sample was mixed with 1.5ul loading dye. After denaturing at 90°C for 2min, 1.5ul of mixed sample was loaded onto gel. The gel images were tracked and peaks were analyzed using ABI PRISM GeneScan software 3.1 to estimate the size of the DNA fragment.

Table1. 13 primers were screened as shown in the table below.

Locus Name	Expected size (bp)	Sequence
SSR7		5'-CCAGACAACCCAAATGAAGG-3'
		3'-GCCTGCTATCGAATCCAGAA-5'
SST8		5'-TTGGCTTTAGATTGGCTTGG-3'
		3'-TGGATCTGTCACCTCCTCATC -5'
NZPR0006	219, 235	5'-ggAAg(A)5TTgggCCTTA-3'
		3'-CTCTCTATCTCTgCCCCA-5'
NZPR 0102	108 , 110	5'-ATCACTCACATCGGCAACTC-3'
		3'-GCTCCAAGGTGATTGAAATCTC-5'
NZPR 0104	154,156	5'-ACCTAATGCCAATAGAAGCAATATC-3'
		3'-AACTAGCACATCCACTTTGCTTC-5'
NZPR0116	151,163	5'-ATACCTAGTTATCATTTAAATAAATGC-3'
		3'-CTCTCAGTAAGTCGAAGAGAGTATC-5'
NZPR0119	134,161	5'-CATCTGAAACAACATCTAAAATGGA-3'
		3'-CCACTAGGTGAACATTAGCCCTTA-5'
NZPR0120	138	5'-GGCAAGGAACAATTGATACAACA-3'
		3'-AATCACTCTGACATAGCTACTGAGCA-5'
NZPR0127	254	5'-TTATTTTAAGCCACACACAAAACAA-3'
		3'-GATAACTTGAGATAATCATTGGGGATA-5'
NZPR0129	251,257	5'-CTTCAACCTCTCTCAGTCCA-3'
		3'-GAGCAGGGTTGTTAATAAAGAAGG-5'
NZPR0130	-	5'-GTCCACTTTATACTTTCTTGTCTTCCv
		3'-GTGTTGTGACCAGAAAAGAGCA-5'
NZPR0138B	159,183	5'-AGAGTGTGTGCATGTGCACATGA-3'
		3'-TTATCCTCCCCATTTTCTCTA-5'
NZPR0143	211,230	5'-GAAAGCATTAGCCATCTACATTCA-3'
		3'-TCATTGTGCATGCATTTATAATCTC-5'

With the established system for amplification of SSR loci in our lab., PCR based SSRs were successfully amplified, and the gel image were nicely generated by ABI sequencer 377 as shown in Figure.

Figure. DNA polymorphism with SSR7 microsatellite primer (blue) and NZPR0102 microsatellite primer (yellow). The red band is molecular weight marker.

Thirteen primers were screened for SSR, and two selected primers (SSR 7, NZPR 0102) were run and analysis on 281 progeny. These 13 primers have shown polymorphism in other *Pinus* species, but not all in *Pinus taeda*. As shown in Figure2, Primer SSR7 and NZPR 0102 showed good polymorphism.

Both markers did not show significant effects on adjusted height growth as shown in table 2 and 3, although P value for SSR 7 effect is 0.0587. As usually, 100~ 200 markers are required to construct a linkage map, it is expected that we could not find the significant association with these two markers. More useful SSR markers, together with AFLP markers may provide more information.

USING MGENE PROGRAM FOR PROGENY DATA ANALYSIS

The computer program MGene was written to detect a major gene in a diallel progeny population using a mixed inheritance model (MIM). The MIM model assumes that the quantitative trait of interest is controlled by a major gene and polygenic random effects. The program is to analyze progeny test data for detecting major genes affecting quantitative traits, such as height or volume of pine trees. The entire analysis can be summarized in the flow chart below. Four diamonds are four steps that use SAS for data manipulation and analysis, MGene program or SPLUS.

MGene program is a windows console mode program and it can be run in two modes: simple mode or argument mode. Regardless of the running mode, all three input files should be in the same physical folder as the program MGene.

In the simple mode no argument is supplied. The input files must be named as the following: the data file as **org.dat**, the design-matrix file as **incidcr.dat** and the initial values file **initial.dat**. The program can be run as any usual executable file, i.e. by double clicking or type the program name **MGxx** at command prompt mode, where **xx** is the version number. The user will be prompted to input the total number of iterations at running time. Thinning parameter and the number of parents is fixed at default value (**lag**=5, **NP**=6).

In the argument mode, the program can either be run by typing the program name followed by arguments in command prompt mode or be called by command lines in a batch file. The syntax of calling the program is

MGxx [file1 file2 file3 chainID iteration-number] [lag] [NP]

Examples:

MGxx 158one_org 158one_incidcr 158one_initial 158one 100000 10 5

****supply all arguments.**

MGxx 158one_org 158one_incidcr 158one_initial 158one 100000

****using default lag=5, NP=6.**

MGxx 20

****supply only lag, use default file1, file2, file3 names and NP. Iteration number needs to be inputted at running time.**

where **file1** is data file name, **file2** is design matrix file name and **file3** is initial values file name; **chainID** is the ID for identifying chains i.e. in output file names; **lag** is the lag distance for thinning the chains, i.e. output samples every **lag** iterations; **NP** is the number of parent.

The program prints out the iteration number every 100 iterations to give the running status. After the program finishes all iterations it output the chains into three files:

- variance and major gene effects as **[chainID_]bsigma.dat**,
- parental GCAs as **[chainID_]btheta.dat** and
- parental genotypes as **[chainID_]btgeno.dat**.

Tip: Since external programs can be executed within SAS using global command **X** '**command**', e.g. **X** '**MGxx** . . . ', a concise way of running a batch of datasets or simulated data sets is writing a micro program.

Workflow

The entire analysis can be summarized in the above workflow chart. Four diamonds are four steps which involve data manipulation and analysis using SAS, MGene program or Splus. These steps are explained in details in the following (a 6-parent diallel design is used in examples).

1. Prepare input files for MGene program

There are three files needed as input datasets:

- 1). The original data file: it should include **female**, **male** and **trait value** columns. **Female**, **male** are coded as 1, 2, ..., 6. The name should be called **org.dat** in the simple mode;
- 2). The incidence matrix file: it is the aggregated design matrix for the linear mixed model: $H = (X:WL:Z)$ is an $N \times 24$ matrix, where N is the number of observations and 24 is the number of parameters in the model. Parameter vector is $\beta^* = (\beta:m:u) = (u, a, d, g1, \dots, g6, s1, \dots, s15)$. The name should be called **incidic.dat** in the simple mode;

3). The initial value file: it includes all initial values for MCMC procedure. The name should be called **initial.dat** in the simple mode. The format of the initial value file is the following:

Line1: u (overall mean) a (additive effect of the major gene)
 d (dominant effect of the major gene)
Line2: $g1 \dots g6$ (GCA effects of six parents)
Line3: $s1 \dots s15$ (SCA effects of 15 crosses)
Line4: $\sigma_G^2 \sigma_S^2 \sigma_E^2 f$
Line5: $K_1^2 \nu_1 \alpha \beta$

where K_1^2 is the variance hyperparameter of normal prior for $u a d$; ν_1 is the first hyperparameter of inverse gamma prior for the variance component; the second hyperparameter ν_2 of inverse gamma prior is $(\nu_1 - 1) * \sigma_i^2$, $i = G, S, E$; α and β are the hyperparameters of beta prior for favorable allele frequency;

Here is an example:

```
0.0 1.1 0.0
-.1 0 .1 -.3 -.1 .2
-.1 -.2 -.2 .6 -.3 .1 -.2 .1 .2 -.1 -.4 .1 .7 0 .2
0.034 0.017 0.595 0.5
4 2 1 1
```

Note: in the latest version, K_1^2 is fixed as $0.5 * \sigma_E^2$.

All three input files can be prepared in SAS macro program MG_input.sas. This macro program fits a polygenic model using Proc Mixed and use the residuals $Y - X\beta$ as the adjusted trait values to remove fixed effects test, block. Estimates of variance and GCA, SCA parameters from polygenic model are used as initial values for MGene program.

2. Run GM program:

The details on running the program in different ways was already discussed in section ‘Running mode of MGene program’. Here, some brief inside knowledge of MGene program will be provided instead.

Understanding the coding in the GM program.

MAXOBS - is the total number of observations.

MAXPARA: - is the number of parameters used in the MIM model, default 24 for six parent diallel design. They are indexed as the following:

Index:	0	1	2	3	...	8	9	...	23
Location parameter θ									
Theta [MAXPARA]:	u	a	d	$g1$	...	$g6$	$s1$	...	$s15$
Variance of θ									
Sigma [MAXPARA]:	K_1^2	K_2^2	K_3^2	σ_G^2	...	σ_G^2	σ_S^2	...	σ_S^2

HyperPar[7]: Hyperparameters K_1^2 U_i U_1 V_2 V_3 α β
SigmaE: Error variance σ_E^2
y[MAXOBS]: Trait value Y
hMat[MAXOBS][MAXPARA]: Aggregated design matrix ($X:WL:Z$)
freq: frequency of the major gene allele $A_1 f$

After the program read in three input files, the above variables or matrices (array) were assigned values. **Theta**, **sigma**, **hMat**, **freq** are sequentially updated during MCMC loop. The updating order in each MCMC iteration is as follows:

- 1) Update major gene genotypes **W** from full conditional distributions by parent blocking, i.e. update parent 1 and its offspring, ..., parent 6 and its offspring;
- 2) Update allele frequency **freq** from its full conditional distribution;
- 3) Update location parameters **Theta** univariately from their full conditional distributions;
- 4) Update variance parameters **Sigma**, and **SigmaE** from their full conditional distributions.

Step 1 through 4 completes one iteration cycle and repeating the above steps forms a Markov chain. For every **lag** iterations, parameter values are written to buffer arrays and are then flashed to output files when buffers are full. The program finishes when all iterations are complete and output files are closed after the remaining values in the buffer are written to output files.

3. Modify output files

The program will generate three files:

- 1) **[chainID_]bsigma.dat** include variance parameters and major gene parameters σ_G^2 , σ_S^2 , σ_E^2 , f , a , d ;
- 2) **[chainID_]btheta.dat** includes parental GCA effects $g_1 - g_6$;
- 3) **[chainID_]btgeno.dat** includes parental genotypes.

In the Gibbs chain, additive major gene effect a may be positive or negative. A positive a indicate A is the favorable allele, and a negative a indicate A is the favorable allele. In our study, we are only interested in the absolute value of a and the frequency of the favorable allele. Hence we takes the value of $-a$ and $1-f$ when a is negative. The corresponding changes in genotype samples are also needed. Such changes can be made through the SAS macro program **change.sas**. This macro program also calculated the major gene variance, additive part and non-additive part for each Gibbs sample. The output files from **change.sas** are

- 1) **[chainID_]bsigmac.dat** based on **[chainID_]bsigma.dat**, modified a and f when a is negative;

- 2) **[chainID_]btgenoc.dat** based on **[chainID_]btgeno.dat**, modified genotype values when a is negative;
- 3) **[chainID_]varm.dat** includes major gene variances σ_m^2 , additive major gene variance σ_{ma}^2 and dominance major gene variance σ_{md}^2 .

Together with **[chainID_]btheta.dat**, there are four output files which need to be analyzed in BOA.

4. BOA output analysis

Bayesian Output Analysis Program (BOA) is developed by Brian J. Smith and can be used to analyze the chains from MGene program. The program files and help documentation may be downloaded from <http://www.public-health.uiowa.edu/BOA>

BOA.s is essentially a Splus script program thus requires installation of Splus or R software. It has two interfaces, menu-driven interface and command line. To enter the menu-driven interface, type **boa.menu()** at the command line. To use command line mode, type **boa.start()**. More details can be found in the BOA help documentation files.

To further facilitate using BOA, a short Splus script file MG_batch.ssc was written for analyzing output files produced by MGene program. The script includes three similar Splus functions **MGboabs**, **MGboabt** and **MGboavm** but customized for three output files **[chainId_]bsigmac.dat**, **[chainId_]btheta.dat** and **[chainId_]varm.dat**. In each function, the boa functions are called to obtain essential chain information such as summary statistics, highest probability density intervals, diagnostic statistics, diagnostic plots, trace plots and density plots etc.

Running from Graphical User's Interface

In order to make this program package available for most researchers, we construct a GUI version that successfully packs up all required programs for fulfilling this project. It is a friendly graphic user's interface and requires minimum amount of input work from researchers to finish the entire process.

1. Software and path set up

As explained in the workflow, several software packages are needed for the entire process, including SAS, S-plus and JAVA runtime environment. It is highly recommended that up-to-date version of each package is installed on the computer, and paths are required to be properly set up as the following examples:

For JAVA :C:\j2sdk1.4.0_02\bin;
 For SAS :C:\Program Files\SAS Institute\SAS\V8;
 For S-plus :C:\Program Files\sp2000\cmd;

For detail path setup information, please refer to instruction from the software distributors.

2. GUI set up

The software package is required to be installed under the directory, C:\MGGUI. Under this directory, all required “java.class” files and one C++ executable file are installed. These files are executable files, which are compiled from the corresponding source files, and are needed to fulfill the analysis. Any damage (such as deletion by mistake) of these files may result in malfunctioning in executing the program. So precaution is needed to protect these executable files from being impaired. (Due to the previous set up of C++ program, these files have to be stored and run from this directory. Newer version may come out to avoid the possible problem resulted from mis-conducting)

Three sub-directories are constructed under C:\MGGUI. ‘backup’ directory is used to store all source codes, including JAVA, C++, SAS, and S-plus codes; ‘executingCode’ directory holds all required executable files as a back for repairing purpose; ‘startCode’ directory holds required intermediate scripts for running this GUI, such as invoking SAS etc.

As the current set up, the actual running job will be carried out in C:\MGGUI directory as well. The user needs to load the original raw data file, such as “AIND035”, into this directory before starting to run the programs. All the output data files will be temporarily stored in this directory, it is highly recommended that the user transfer the output data files as well as original data file to other directory after finishing the whole procedure.

3. Running the GUI version

A shortcut icon will be put on the desktop, the whole procedure is invoked by double-clicking on the icon. One thing needs to be aware that the GUI is actually invoked by the following command: “C:\j2sdk1.4.0_02\bin\javaw.exe -cp C:\MGGUI\MGLinkerGUI” from the DOS command window. So it is obvious that the JAVA path set up and JAVA class files location are important for successfully running the GUI. Please follow the set up direction carefully for this GUI package.

Once the shortcut icon is activated, a GUI shows on the screen as following:

Eight buttons are lined up on the top for executing the programs, the text window shows information about which button is clicked and what is the function of that button. The first five buttons are arranged according to the order of the general workflow of the overall procedure (refer to the workflow section).

SAS_input button is clicked, SAS program will be invoked and the “MG Input.sas” will be loaded. This SAS code is used to process raw data, please refer to previous detail explanation. Raw data will be preprocessed and ready for C++ Gibbs Sampler program.

Major Gene Program button is clicked, another interface will show on the screen as following. The user needs to input individual name, which is numerical part of raw data file name. E.G. for raw data file AIND035, individual is “035”. The default inputs are shown for chained ID, iteration number, lag number, and number of parent, which often needed to be modified by user as input, for specific parameter set-up. Once all parameters are set up properly, the user can invoke C++ by clicking the “start” button. Two text windows are lined up on the interface, the top one is for interactive information, such as what command has been issued, etc. The bottom one is the re-direct window for C++ output stream. The user could visually monitoring the running of C++ program. If C++ runs successfully, a piece of message “Process exit with rc=0” will be shown on the bottom line.

Goes back to the first GUI window.

SAS_change button is clicked, SAS program will be invoked and the “change.sas” will be loaded. This SAS code is used to modify C++ program output files for BOA diagnosis.

S-plus_BOA button is clicked, S-plus program will be invoked and two s-plus scripts “MG-batch1.ssc” and “MG-batch4.ssc” will be loaded. The s-plus scripts combining with BOA package are used to diagnose the quality of Gibbs Sampler performance. Whether the joint posterior distribution is reached, and when? Once the proper converge information is confirmed, the user can proceed to finish the statistical analysis procedure.

SAS_result button is clicked, SAS program will be invoked and the “MG_result.sas” will be loaded. This SAS code is used to perform statistical analysis for all parameter estimation as well as reporting of major gene effect.

Output Data button is clicked, an open file window will show up. By default, it only shows text-related files under C:\MGGUI directory. The user can open any output data file to monitor the behavior of the programs.

Help button is clicked, this word –format documentation file will be opened for user's reference.

Quit button is clicked for exit the program.

SIMULATIONS FOR TESTING MGENE PROGRAM AND REAL PROGENY DATA ANALYSIS FOR MAJOR GENE DETECTION

With the computer program Mgene, additional computer simulations were performed to investigate the factors affecting the mixed inheritance model (MIM) and relate the results to a traditional pure polygenic model. We evaluated the power of Bayesian inference with Gibbs sampling technique and accuracy for detecting major genes with simulated phenotypic diallel progeny test data, and; We applied Gibbs sampling technique in analysis of actual progeny data of loblolly pine and discuss its implication for the breeding program; We evaluated genetic parameter estimation for both the MIM model (excluding the genetic by environmental effect) against the pure polygenic model (accuracy and bias, etc) and compare the genetic gain prediction based on these two models.

Simulated Study

A parameter-based scheme was used to simulate diallel data sets in the MIM framework. Each realization of diallel data sets consists of phenotypic observations simulated by summing up random deviations for model effects, which are generated using SAS random number generation functions. Different combinations of genetic parameters consist a total of 10 genetic parameter structures. Three levels of additive major gene effect a (none, low and high, labeled as N, H, L respectively), two levels of dominant major gene effect d (labeled as 1, 2), and two levels of type B genetic correlation r_B (labeled with or without e) were considered. The true genetic parameters are listed in Table 1. The mating design of one diallel test series consisted of two disconnected six-parent diallels. The field design was a randomized complete block design with 4 tests, 6 replicates. For each parameter combination, 50 simulated test series or 100 diallels were realized. Single-tree plot filed layout was used in order to reduce the computing time.

Table 1. The true values of genetic parameters for each parameter structure level

Level	A	d	f	σ_m^2	h^2	γ	r_B	σ_g^2	σ_s^2	σ_{gt}^2	σ_{st}^2	σ_p^2	σ_e^2
N	0	0	0.2	0	0.2	0.3	1	5.000	1.500	0.000	0.000	8.850	79.650
Ne	0	0	0.2	0	2.0	0.3	0.75	5.000	1.500	1.766	0.500	8.467	76.200
L1	5	-2.5	0.2	4.56	0.2	0.3	1	4.772	1.432	0.000	0.000	8.446	76.018
L1e	5	-2.5	0.2	4.56	0.2	0.3	0.6	4.772	1.432	1.591	0.477	8.081	72.725
L2	5	2.5	0.2	14.16	0.2	0.3	1	4.292	1.288	0.000	0.000	7.597	68.372
L2e	5	2.5	0.2	14.16	0.2	0.3	0.75	4.292	1.288	1.431	0.429	7.268	65.410
H1	10	-5	0.2	18.24	0.2	0.3	1	4.088	1.226	0.000	0.000	7.236	65.122
H1e	10	-5	0.2	18.24	0.2	0.3	0.75	4.088	1.226	1.363	0.409	6.922	62.301
H2	10	5	0.2	56.64	0.2	0.3	1	2.168	0.650	0.000	0.000	3.837	34.536

H2e	10	5	0.2	56.64	0.2	0.3	0.75	2.168	0.650	0.723	0.217	3.671	33.040
-----	----	---	-----	-------	-----	-----	------	-------	-------	-------	-------	-------	--------

Simulation results

The power for each genetic parameter setting was calculated by taking the percentage of the number of diallels from which a major gene effect was found. The criterion for determining a significant major gene is based on the density plot of additive major gene effect, major gene frequency and major gene variance: 1) a significant positive maximum for additive major gene effect a , or a non-additive effect d that is clearly deviated from zero, i.e. larger than its standard deviation; 2) a clear mode in the gene frequency of f density plot; and 3) the mode or maximum of major gene variance σ_m^2 is significantly deviated from zero, i.e. larger than its standard deviation. All three criteria must be satisfied for major gene detection. Among all three criteria, the mode of σ_m^2 is the most important and has been used commonly for other purposes (Janss et al 1995, Pan et al 2001). In our simulated data sets, when this criterion is satisfied, other two criteria are also met.

It is showed that the MIM model does not have sufficient power (less than .1) to detect a major gene with an additive effect a less than half the phenotypic deviation unit (in level L1, L1e, L2, and L2e). When a true a value is of one unit of phenotypic deviation and true d value is also positive (in level H2 and H2e), the power increased above or near 0.9. The negative d value dramatically reduced the power to about 0.2. The true major gene variance σ_m^2 has the similar effect as genetic effects a and d , although at lower magnitude, in that a higher σ_m^2 resulted in higher power for detecting major gene. With 30% increase in true σ_m^2 value from parameter level L2 to H1, the power jumped from nearly zero to 0.2, mainly due to the increase of true a value.

The type one error was measured by the false major gene detection in diallels with no true major gene effect, i.e. level N and Ne of the parameter settings. Using the same criterion as in the power analysis, major gene was falsely signaled by a significant deviation of the major gene variance mode from zero in 1% to 3% of these simulated diallels (Table 2). Such level of error probability is below a commonly used 0.05, and hence strongly establishes the accuracy of the Bayesian method in diallel progeny tests for detecting a major gene.

The genetic by environment (GxE) effect seemed to have some effects on major gene detection, pushing the power slightly higher than that for data sets with no GxE effect. Such extra effect i.e. random GCAxTest and SCAxTest effects add the complicity to the model and also interact with other parameter estimates. Part of these effects may contribute to a slightly higher estimate of major gene effect, thus resulting in a false increase in power or type one error. Fortunately, this increase is small enough not to cause a serious concern.

Table 2. Power and type I error for major gene detection for the mixed inheritance model (over 100 simulated diallel data sets)

Parameter level ^a	# of converged diallels	# of diallels detected with major gene	Power	Type I Error
N	100	1	-	.010
Ne	100	3	-	.030
L1	100	0	0.000	-
L1e	100	1	0.010	-
L2	100	4	0.040	-
L2e	100	1	0.000	-
H1	100	19	0.190	-
H1e	100	22	0.220	-
H2	93	84	0.903	-
H2e	97	91	0.936	-

The variance estimates from both the polygenic model and the MIM were generally similar and they are close to the true values although the GCA and SCA variance were slightly larger in the MIM as they are affected by skewed posterior distribution. The parameters regarding major gene effects, i.e. a , d , σ_m^2 were underestimated in most parameter settings (Table 3), mainly due to the low power of major gene detection when the true a value was half the unit of σ_p . When a value increased to one unit of σ_p , the parameters became much closer to the true values although still underestimated. Such deviation from true population values also reflected the relatively small sample size (100). With a type B genetic correlation of 0.75, the GxE effect added some noise into genetic parameter estimates such as GCA variance, additive major gene effect and major gene variance.

Table 3. The average of mean estimates for additive and dominance gene effect (a and d), major gene variance σ_m^2 and frequency f of major gene from the mixed inheritance model (over 100 simulated diallel data sets)

Level ^a	<i>a</i>		<i>d</i>		σ^2_{m}		<i>f</i>		
	T. V. ^b	Est. ^c	T. V.	Est.	T. V.	Est.	T. V.	Est.	T. V.
N	0	3.96(0.74)	0	-0.08(1.1)	0	11.77(3.81)	-	0.50(0.04)	5
Ne	0	3.94(0.72)	0	-0.01(1.31)	0	11.94(3.86)	-	0.50(0.05)	5
L1	5	3.79(0.64)	-2.5	-0.02(0.85)	4.56	10.81(3.08)	0.2	0.50(0.03)	4.772
L1e	5	3.46(0.69)	-2.5	-0.09(1.28)	4.56	12.35(4.24)	0.2	0.50(0.04)	4.772
L2	5	3.99(0.84)	2.5	0.28(1.43)	14.16	12.34(4.98)	0.2	0.49(0.06)	4.292
L2e	5	3.89(0.48)	2.5	0.13(1.16)	14.16	11.54(3.11)	0.2	0.49(0.04)	4.292

H1	10	4.39(2.11)	-5	-1.12(2.24)	18.24	14.10(8.96)	0.2	0.46(0.07)	4.088
H1e	10	5.02(2.16)	-5	-1.03(2.36)	18.24	16.64(10.29)	0.2	0.45(0.07)	4.088
H2	10	6.21(1.83)	5	6.26(2.66)	56.64	35.36(12.52)	0.2	0.27(0.10)	2.168
H2e	10	7.02(2.12)	5	5.64(2.39)	56.64	39.74(13.01)	0.2	0.30(0.10)	2.168

Note: a-level: parameter settings (see Table 1);
b-T.V.: parameter true values;
c-Est.: average of mean estimate;
d-STD: standard deviation.

The probability of correct identification of parental major genotype, measured by the average percentage of parents whose major genotypes were correctly identified, increased gradually as the true major gene effect went up (Table 4). When true additive major gene effect is small i.e. $a=5$, only 41-52% of parents were correctly identified for their major genotypes. This figure increased above 60% with $a=10$, $d=-5$. When both the additive and dominance effects were large and positive as in setting H2 and H2e, the highest percentage of correct identification was achieved (95%). The breakdown of percentage into three genotypes indicated that heterozygote Aa has the highest probability of correct identification.

Table 4. Percentage of correct identification of parental genotypes by true genotypes

Level ^a	True parental genotype			Overall average
	aa	Aa	AA	
L1	35.7%	54.0%	40.0%	42.0%
L1e	34.3%	52.1%	35.3%	40.7%
L2	50.8%	54.3%	58.3%	52.3%
L2e	47.6%	58.2%	68.2%	51.7%
H1	61.1%	68.4%	52.2%	63.5%
H1e	56.4%	75.5%	37.5%	61.7%
H2	93.6%	100.0%	92.3%	95.4%
H2e	91.9%	100.0%	90.0%	94.6%

Note: a-level: parameter settings (see Table 1).

Real Progeny diallel data analysis

More progeny data sets from the N. C. State University Tree Improvement Program were analyzed with Mgene program for major gene detection. A total of 190 diallels progeny populations have been run through the MGene for testing major gene effect. The criteria for determining the present of major gene was based on diagnostic statistic such as Brooks, Gelman, and Rubin's Corrected Scale Reduction Factors and trace plots from BOA outputs.

Based on the above-mentioned three criteria in BOA diagnosis, 124 diallels were determined to have detectable major gene effect. However, the definition for any major gene is when the additive genetic effect is equal to or more than half unit of the phenotypic standard deviation. According to this definition, 19 diallels were further picked from the initial candidate pool, as having a major gene that affects the phenotypic value. Among these 19 dialleles, additive major gene effects (a) are significantly greater than zero, ranging from 0.98 to 4.110; and the additive major gene effect (a) over overall phenotypic standard deviation (σ_p) ranged from 0.51 to 1.11 (Table 5). The proportion of marginal mean estimates of the major gene variance component (σ_m^2) in the total phenotypic variance (σ_p^2) ranged from 20% to 44%. Here we want to bring some discussion on the criterion to claim a major gene effect. Traditionally, a major gene effect is reported when additive major gene effect estimated exceeds half unit of phenotypic standard deviation (σ_p). However, according to the simulation results shown above, test power will greatly increase when a is big and toward one unit of phenotypic standard deviation. So even with stringent criterion, we can say 035one has larger major gene effect with high confidence.

Table 5. The marginal posterior means and standard deviations for major gene effects a , d , and major gene frequency f , and the difference between two homozygotes in half σ_p unit (a/σ_p) and the proportion of major gene variance in the phenotypic variance σ_m^2/σ_p^2 .

diallel	a	d	f	σ_m^2/σ_p^2	a/σ_p
334one	4.1113	4.2978	0.78545	0.4442	0.50806
227one	3.50565	3.322	0.7718	0.41096	0.52213
434two	1.04675	1.40105	0.6108	0.19536	0.53366
482one	0.98095	0.7778	0.49785	0.24115	0.55787
445one	2.8659	1.8325	0.78555	0.33418	0.57305
434one	1.16785	1.14015	0.50145	0.26581	0.60987
356two	1.246	1.20785	0.4994	0.26755	0.61038
375one	1.21535	1.377	0.64345	0.2042	0.63431
357one	1.38635	1.3352	0.53675	0.26769	0.63941
404one	1.22555	1.1903	0.50055	0.30227	0.65204
357two	1.3686	1.3441	0.509	0.29946	0.65357
408two	1.37375	1.32255	0.4998	0.30613	0.65727
430two	1.47165	1.4211	0.5008	0.34176	0.69577
408one	1.2776	1.22195	0.49835	0.34696	0.69871
356one	1.27495	1.24145	0.5009	0.3557	0.70848
404two	1.49605	1.4728	0.5001	0.3964	0.74603
343two	1.47185	1.43615	0.49915	0.40551	0.75546
343one	1.4291	1.40185	0.50005	0.41204	0.76196
035one	3.15685	2.7424	0.78535	0.25634	1.11289

Mean	1.7406	1.6572	0.5751	0.3186	0.6648
Std Dev	0.9332	0.8640	0.1168	0.0732	0.1331
Minimum	0.9810	0.7778	0.4979	0.1954	0.5081
Maximum	4.1113	4.2978	0.7856	0.4442	1.1129

Dominant major gene effects (d) were positive in all diallels and consistently approached the magnitude of a (only slightly smaller than a on average), indicating a major gene is near complete or complete dominant. Such a consistent gene action in a single locus seems to suggest a major gene existing in the second generation breeding population of loblolly pine.

In these 19 diallels with evidences of major gene effect, genotypes of major gene for six parents were found to be heterozygotes with very high probabilities (data not shown). In another study on analysis of progeny data, 84% of the parents, the probability of estimated genotype was larger than 0.95, and for 77% of the parents the probability was higher than 0.99 (unpublished data from Xing, Bin). Such high confidence with parental genotypes is very useful in selecting parent trees with the favorable allele of major genes. The dominance of heterozygous genotypes of parents in the population indicates clear major gene segregation in the progeny population. The favorite major gene allele frequencies in the population are very close and greater than 0.5. In most diallels, estimated parental major genotypes provided evidences of major gene segregation. In these diallels, the families of heterozygotes or homozygotes with favorable major gene alleles provided promising materials for further molecular analysis to detect QTL.

Comparison of MIM model and polygenic model

Historically, the polygenic model that explains overall quantitative traits with many small gene effects has been successfully applied in animal and plant breeding for decades. Theoretically, the major gene information from MIM model could help develop better strategies in breeding programs. Thus, it would be of interest to compare the results from the MIM model and from the pure polygenic model on the same data.

In the MIM model, the variance components and GCA prediction obtained from a polygenic mixed model were served as the initial values for Gibbs sampling algorithm. The random GCAxTest, SCAXTest and plot effects are included in the polygenic model but not in the MIM model; on the other hand, the major gene variance σ_m^2 calculated from a , d and f only exists in the MIM model.

Variance components

Several important variance components obtained in the MIM models were compared to those estimated with a pure polygenic model (Table 6). When the true major gene effect is small ($a=5$ or $0.5\sigma_p$) or none ($a=0$), variance components σ_g^2 and σ_s^2 estimated from both models were very close, with the corresponding correlation between two models was at high range (.86~.96). However, when the true major gene effect was very large (a increased to 10 or one unit of σ_p) the correlation reduced to around .76~.86. The estimation of additional major gene parameters also had impact on other variance components. One of the side effects was the low correlation between error variance estimates from two models (Table 6).

Table 6. The average of estimates (standard deviate in parenthesis) for GCA variance σ_g^2 , SCA variance σ_s^2 and error variance σ_e^2 from the mixed inheritance model (MIM) and polygenic model (POLY) and their correlation (over 100 simulated diallel data sets).

Level ^a	σ_g^2				σ_s^2				σ_e^2			
	T. V.	MIM	POLY	r^b	T. V.	MIM	POLY	r	T.V.	MIM	POLY	r
N	5	6.86(3.18)	4.09(3.06)	0.86	1.5	1.83(2.23)	1.53(2.00)	0.92	88.500	76.34(6.37)	86.90(6.17)	0.83
Ne	5	7.64(4.04)	5.27(3.93)	0.91	1.5	1.64(2.06)	1.33(1.66)	0.89	84.667	74.98(6.91)	83.77(6.88)	0.85
L1	4.772	7.74(4.39)	4.97(4.18)	0.94	1.432	2.15(2.91)	1.68(1.94)	0.87	84.464	75.79(7.96)	85.59(7.17)	0.90
L1e	4.772	8.85(5.00)	6.35(5.45)	0.96	1.432	2.08(2.49)	1.81(2.07)	0.91	80.806	73.55(6.86)	82.63(6.87)	0.76
L2	4.292	9.68(4.18)	7.56(4.47)	0.94	1.288	1.51(1.84)	1.41(1.70)	0.97	75.968	70.96(8.60)	81.78(8.22)	0.88
L2e	4.292	9.14(5.64)	6.91(6.29)	0.97	1.288	1.88(2.49)	1.56(1.88)	0.92	72.678	70.88(7.07)	79.55(7.44)	0.89
H1	4.088	8.95(5.67)	8.54(6.35)	0.90	1.226	1.98(2.32)	2.06(2.54)	0.93	72.358	71.63(13.43)	83.24(12.5)	0.92
H1e	4.088	9.45(6.50)	9.14(8.76)	0.92	1.226	1.98(2.49)	2.05(2.55)	0.93	69.223	67.55(9.43)	79.14(10.5)	0.82
H2	2.168	8.87(8.59)	14.32(9.19)	0.86	0.650	0.95(1.22)	1.34(1.86)	0.94	38.374	39.24(6.49)	64.96(14.6)	0.60
H2e	2.168	7.76(6.33)	14.02(8.87)	0.76	0.650	1.22(1.43)	1.87(2.27)	0.93	36.711	37.88(3.97)	67.15(13.6)	0.42

Note: a-parameter settings (see Table 1);

b-correlation coefficient between variance estimates of the MIM model and the polygenic mode

In the polygenic model, the additive major gene variance was roughly equally transformed into GCA variance (including male and female parents) and error variance. Theoretically, dominance part of major gene variance should also be transformed into SCA variance and error variance according to the 1:3 ratio as in the case of poly-genetic variances. Unfortunately, such phenomenon can not be verified because of its relatively small magnitude and its confounding with the additive part in error variance.

For parameter settings with $r_B=.75$ (Ne, L1e, L2e, H1e and H2e), the GxE effects are confounded with genetic effects in the MIM model, hence slightly inflate GCA and SCA variance estimates. However, larger than expected averages of σ_g^2 and σ_s^2 were also found for parameter settings with $r_B=1$ (N, L1, L2, H1 and H2, in Table 3), which suggests that a upward bias is inherent in Bayesian estimates of small variance components with skewed posterior densities. As a result, σ_g^2 , σ_s^2 , σ_e^2 and the total phenotypic variance in the MIM model tends to be slightly larger than those in the polygenic model in all parameter settings.

GCA prediction from both the MIM and the polygenic model agreed well when there is no true major gene effect or its effect is small to medium (when $a=5$ or when $a=10$, $d=-5$). In parameter settings N through H1e, the correlation between GCA predictions from two models was as high as .98 (Table 3). Both the MIM model and polygenic model had similar correlation between their GCA predictions and the true values, which ranged from .61 to .78 and gradually declined as major gene variance increased. When the true major gene effect increased to a high level ($a=10$ or $1\sigma_p$ and $d=5$ or $.5a$), i.e. in parameter settings H2 and H2e, the disagreement between GCA

predictions from two models rose significantly as the their correlation dropped to .66~.70. In the meanwhile, GCAs from the polygenic model dropped its correlation with true values to .28~.32, while the MIM model maintained a relatively high correlation of its GCAs with true values (.55).

In the MIM model, after major gene effects were transformed into GCA (using Equ. 4a, 4b, 4c) and added to GCA predictions, both models had a much higher correlation (above .90) between their GCA predictions even when major gene effect is large (parameter setting H2, H2e). Similar adjustment of major gene effects is also possible with true GCA for parameter settings L1 through H2e. The adjusted true GCA values represents the total additive genetic values contributed by both major gene and polygenes. As a result, both models increased their correlations with true GCAs significantly across these parameter settings. Since the correlations between predicted and true values is higher in the polygenic model than that in MIM model, better GCAs will be predicted from polygenic model. This has long been the solid basis and a successful approach in animal breeding scheme. And it also explains some of the hurdles in the major gene components predictions with MIM model. However, the potential modification of MIM model will definitely provide precise evidence in the breeding plan.

In summary, the variance estimates from both the polygenic model and the MIM model were similar and close to the true values. The parameters regarding major gene effects were in most cases underestimated, mainly because of the low power of major gene detection when the true a value was less than half the unit of σ_p . When a value increased to one unit of σ_p , the power increased significantly up to .9, the probability of correct parental genotype identification went up to .95, and the major gene parameters also became closer to the true values although still underestimated. For the practical purpose, the main application of the MIM model is to detect major gene effects and to identify individuals with potential large major gene effects. In the real data analysis, instead of using single tree plot as in the simulated data set, a field design with 6 tree per plot resulted much more data, the power could even be higher than .9 when $a = 1\sigma_p$. The study also indicated that the type one error i.e. false major gene detection rate was as low as .01-.03.

When the major gene effect is small or none, GCA predictions were similar for both MIM model and polygenic model and they are both close to the true GCA values. When the true major gene effect is large, GCAs from the MIM model are apparently more precise than those from the polygenic model. However, when major gene effects are considered and counted toward true GCA values, GCAs from the polygenic model had respectable correlations with true values, and in fact slightly higher than adjusted GCAs from MIM models. From the perspective of selection and breeding strategies, such results are encouraging to tree improvement programs, which largely depends on a pure polygenic model.

The genetic by environmental effect does not play a significant role in causing the biasness of variance component estimation. It does, however, slightly affect the power of major gene detection.

SEARCH FOR MOLECULAR MARKERS ASSOCIATED WITH RUST RESISTANT GENES

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 of a threshold model was used to make inference about a mixed inheritance model that included polygenic effects and a single major gene effect. Bayesian complex segregation analysis was developed with Gibbs sampler and threshold model for binary data in a half-diallel population. A mixed inheritance model and a polygenic model for binary data are presented in the Bayesian concept. The threshold model and Gibbs sampler are applied for inference. All conditional posterior distributions needed for running the Gibbs sampler are derived and given. A parent blocking strategy is implemented to improve the move of the Gibbs chains. The method is illustrated in major rust-resistance gene detection in loblolly pine. Bayes factor is used for model comparison to determine the inheritance mode of rust-resistance in this half-diallel population of loblolly pine.

Loblolly pine seedlings from 12 full-sib families were obtained from a 6-parent half diallel mating design. These seedlings were challenged using two inocula of the fusiform rust fungus (*Cronartium quercuum* f.sp. *fusiforme*) at extremely high spore density (250,000sp/ml ~ 300,000sp/ml) with replications in the greenhouse (concentrated basidiospore spray method). Each basidiospore inoculum originated from a mixed gall collection of aeciospores obtained from field-infected trees. Presence or absence of rust galls was recorded at four and half months & nine months after inoculation and infection percentages were calculated. The rust infection rates were high for every full-sib family and were above 90% for most of full-sib families for both inocula. However, the infection in family of 28-321 by 28-301 progeny was relatively lower. The analysis showed that the rust infection of 28-321 by 28-301 progeny was not significant different from 75% for both inocula. Two pairs of complementary genes could explain the observed percent infection levels in this diallel based on a gene-for-gene hypothesis. The 75% infection in the full-sib family of 28-321 by 28-301 may be explained by epistasis between two resistance genes. The putative genotypes of host parents and virulence compositions of mixed inocula were postulated.

A bulk-segregant analysis based on phenotype (gall vs no gall) was used to search for dominant molecular markers associated with the potential resistance genes in the host parents. Few candidate marker polymorphisms were observed between the gall vs no gall bulks and none of the candidates co-segregated with phenotype when tested across the progeny set. The rust infection rates were high for every cross and were above 90% for most full-sib families for both mixed inocula. No significant differences were observed between two replications and between two inocula. Comparing the half-sib families, half-sibs of A showed approximately 85% galled for both inocula, while infection in all others was 90% or greater with some close to 100%. Notice that the cross E*A had consistent relatively low infection rates. On average, the half-sib infection rates of A and E were 84.5% and 91%, while the full sibs of this cross showed galled percentages of 71%, 74%, 68% and 76% in different replications. When two replications were pooled for each inoculum, the galled percentage for L-7-2-85 became 72.5%, while it was 72% for NCSU01. Nine-month readings for cross E×A continued to show the same pattern with slightly increase galled percentages. Results from PROC FREQ suggested that there were no

significant differences of these proportions from 75%. Based on the observed 75% rust infection rates for the cross E by A and high infection rates for other full-sib families, it was postulated that this diallel might have two independent resistance genes segregating in the progeny population.

In this diallel based on infection data, parents A and E are hypothesized to each have a resistance gene, respectively, at different loci. The pathogen population is a mixture of spores, hypothesized to include spores virulent to both individual R genes but not virulent for the two R-genes combined in a single individual. Two host parental genotypes could be derived as $R_1r_1R_2r_2$ and $r_1r_1R_2r_2$, where R is a dominant resistance allele and r is the recessive allele for lack of rust resistance. Pathogen diploid genotypes are speculated to contain $Avr_1avr_1Avr_2Avr_2$, $Avr_1Avr_1Avr_2avr_2$, or $Avr_1Avr_1Avr_2Avr_2$. Haploid basidiospore genotypes expectedly would be Avr_1avr_2 , Avr_2avr_1 , and Avr_1Avr_2 . When the inoculation density is high, individuals with a single R allele are overcome by the corresponding virulence allele and show no resistance to rust. Individuals without an R allele would similarly be galled. Only individuals with both R alleles would expectedly show no gall (rust-free). Assuming each parent has a heterozygous resistance gene and these two genes are unlinked or very weakly linked, a quarter of progeny will have both R alleles according to Mendelian segregation. So, 25% no gall in the progeny set is the expected result from the interaction of $R_1r_1R_2r_2$ progeny and Avr_1Avr_2 , Avr_1avr_2 and Avr_2avr_1 basidiospores. If avr_1avr_2 basidiospores were present at a significant frequency, all progeny expectedly would have been galled.

This hypothesis was investigated using marker / trait association analysis. By searching for markers polymorphic between no gall and gall groups based on BSA, it is possible to identify candidate polymorphic markers that may be linked to resistance loci. Because the RAPD markers are typically dominant markers, specific marker configurations have to be observed in order to find R-genes for the proposed two-gene model using phenotype based BSA. The “-” phase of the marker is expected to be associated with the no gall bulk and the “+” phase of the marker is expected to be associated with the gall bulk. Then when the progeny population is screened, 2/3 of galled individuals are expected to have the “+” phase of the marker, the other 1/3 of galled individuals and all no gall individuals should have the “-” phase of the marker. Non-galled individuals, those with both postulated R genes should be “-” for both markers 1 and 2 as depicted in figure 3-2b.

Figure 3. An illustration for the approach to identify resistance genes interacting in the proposed two-gene model using dominant molecular markers, (a) bulk segregant analysis, and (b) assessment of candidate markers in a sample progeny set.

Based on the approach proposed above and in figure 3-2, replicated bulks using diploid DNA (gall vs no gall) were constructed and 420 RAPD primers were screened. Five candidate markers were observed to be polymorphic with the “-“ phase of the marker associated with the no gall phenotype in replicated bulks. However, none of these candidates cosegregated with phenotype (gall vs no gall) when tested across a sample progeny population.

The Fr2 locus was initially detected by Amerson (unpublished data) in parent A by inoculating progeny from its OP family with isolate 0-5-32. In that work, a RAPD marker (AK6_850) was found by Amerson to be significantly associated with phenotypic expression using diploid DNA. Subsequently, 5 AFLP markers and a RAPD marker AK6_850 were placed around the Fr2 locus in a 64 sample progeny set challenged with isolate 0-5-32. Haploid DNA was used for preliminary mapping of Fr2 (Amerson, unpublished data). In order to better map the Fr2 linkage group, seedlings from a controlled pollinated family A * F were inoculated with single isolate 0-5-32 for the current study. Nine month readings showed 60.5% rust infection. Replicated haploid DNA Bulks (16 samples/bulk) based on phenotypes (gall vs no gall) were assembled and 420 RAPD primers were screened. Candidate polymorphic markers observed in both bulks were tested across 96 individual samples. Six of these candidate markers co-segregated with phenotypes (based on LOD>3).

In order to integrate these new RAPD markers with previously found RAPD and AFLP markers (Amerson, unpublished data), these new RAPD markers were also tested on Amerson’s 64 sample progeny set. RAPD marker genotypes of all 160 (96+64) sample progeny of A were used to construct a framework linkage map of the Fr2 resistance linkage group in parent A (Figure 3-3). AFLP markers (Amerson, unpublished data), RAPD markers not in the framework map, and phenotype (Fr2) were placed on the framework map using Mappop software. Mappop placed Fr2 in the interval between RAPD markers AY12_1300 and C13_580 with a probability of 1.0. All framework markers were significantly associated with Fr2 at a LOD level of 6.9 or greater.

Figure 3-3. A linkage map of the Fr2 linkage group in parent A. Fr2 and accessory markers are placed around the markers in the framework map.

Given the results from studies by Wilcox (1995) and Amerson (unpublished), the Fr2 locus apparently is not the only resistance locus in parent A. Wilcox (1995) found markers A19_560 and A11_400 in parent A to be significantly associated with phenotype in the field and Amerson's unpublished data showed that the Fr2 locus would not account for observed field resistance in family A in Wilcox's diallel samples. Continued efforts to find markers associated with field resistance loci in parent A were undertaken using the A19_560 and A11_400 markers of Wilcox (1995) as bulking targets. 240 RAPD primers were screened using BSA based on the 2 markers of Wilcox (1995). Six candidate markers were tested across 96 samples. Two of them (BG12_570 and BE16_1300) were significantly associated with marker A19_560 with LOD values of (26.19 and 18.06) (Figure 3-4). The other four (AZ18_530, BE_1200, BH19_550, AY20_1300) were significantly associated with marker A11_400 with LOD values of at least 5. Markers BH19_550 and AY20_1300 showed no recombination with A11_400. An additional marker U1_800 that was known to be associated with A11_400 (Wilcox, 1995) was also again found to be associated with A11_400. A preliminary map of the A11_400 linkage group in parent A is given in figure 3-5. The two linkage groups around A19_560 and A11_400 were independent and could not be mapped together based on the analysis using MAPMAKER.

Seedlings of A by F were inoculated with the two mixed inocula used in the diallel study, but at inoculation densities much lower than that used in the diallel. Results from the lowest inoculation density (20,000sp/ml) showed a lower infection rate than that from the moderate inoculation density (50,000sp/ml) for both inocula (Table 3-5). At the same inoculation density, the infection rate for one inoculum (NCSU01) was consistently higher than the other one (L-7-2-85). The difference in rust infection between the two inocula suggests that virulence levels were likely to be different in two inocula. With the moderate infection level using inoculum L-7-2-85 at 50,000sp/ml, it is possible that a heterozygous resistant locus is segregating in the progeny population. Previous analysis and results suggested that the resistance was very likely inherited from the maternal parent A.

Haploid DNA (Megagametophyte DNA) was utilized to seek markers that may be associated with a presumptive maternally inherited resistance locus. Replicated Bulks based on phenotypes (gall vs no gall) were established and a total of 214 RAPD primers were screened. 10 candidate

markers were found to be polymorphic in the replicated bulks (gall vs no gall). However, cosegregation analysis showed no association between the trait locus (phenotype) and the candidate markers when tested across a 96 progeny population.

Discussion of Marker Association

This study applied bulk segregant analysis and RAPD markers to search for rust resistance genes. With a intermediate infection rate of 60.5% in the progeny population of parent A $\times$ F challenged with single spore isolate 0-5-32, 6 RAPD markers were found associated with the Fr2 locus in parent A using replicated haploid DNA (megagametophytes) bulks based on phenotype (gall vs. no gall). An improved map of Fr2 linkage group was constructed with RAPD markers found in this study and previously discovered markers (Amerson, unpublished data). Since most of the works by Amerson and co-workers mapping rust resistance loci have involved, and will continue to involve, RAPD markers and since full RAPD maps of some loblolly pine clones are available (Wilcox, 1995; Jordan, 1997), the additional RAPD markers included here may be useful for establishing regions of synteny between parent A and other clones. Such comparisons await investigation.

The BSA, co-segregation analysis approach successfully used to find markers for the Fr2 locus succeeded in large part because the inoculum was avirulent at the complementary locus and apparently virulent at all other loci for which parent A has resistance alleles. Using single spore inoculum offers the benefit of a simplified virulence composition but restriction of observations to a single spore inoculum may cause one to overlook other resistance loci which are not evident due to the virulence in the single inoculum. Such is the case with parent A.

The Fr2 locus found in this study and previously by Amerson using inoculum 0-5-32 is not the only resistance locus in parent A, as the marker AK6_850 linked with Fr2 locus was not significantly associated with phenotype (field resistance expression) observed in parent A progeny using materials from this diallel (Amerson, unpublished data). Also, statistical analyses in chapter II would support the presence of more than one major R-gene in parent A apart from Fr2. Efforts to marker identify another resistance locus in parent A using progeny of cross A $\times$ F inoculated with 50,000sp/ml of mixed gall isolate L-7-2-85 were not successful despite an intermediate (49%) infection rate at the 50,000 spore level. This 49% rate is suggestive of single gene regulation, but the high infection of cross F $\times$ A (95%) and half sibs of family A (85%) by inoculum L-7-2-85 at 250,000-300,000sp/ml indicated that there probably was some level of virulence in L-7-2-85 for any single gene present in parent A. The possibility of finding a single R gene in parent A at a reduced inoculum level depended on lowering the virulence level with the reduced spore numbers to the point that the inoculum would essentially perform as if avirulent at some complimentary locus where parent A had a resistance allele. This apparently did not occur and the 51% no gall response likely involves a number of as yet undefined genes.

A portion of this study investigated mixed spore inocula of the fusiform rust fungus interacting with progeny of 12 full-sib families from a six-parent half-diallel mating design. With the extremely high inoculation density (250,000 – 300,000 sp/ml), progeny of most crosses were above 90% galled, with the exception of cross E $\times$ A. Seventy-five percent of the progeny of

cross E × A were galled. It was hypothesized that this level of infection was explainable by two independent, epistatic resistance genes in the two parents and a spore population where individual spores could be virulent for one or neither of the proposed R-genes, but not both. Such a spore population should allow only the individuals with both R-alleles (in theory 25% of progeny) to show the resistance (no gall) phenotype. Bulk segregant analysis and co-segregation analysis were used to seek RAPD markers associated with the proposed resistance genes in the two parents, with a bulking strategy based on phenotype (gall vs. no gall). In the approach used to search for resistance genes in cross E × A, only polymorphisms with the “-” phase of the marker linked with “resistance” (NG) bulks were considered to be useful due to the dominance of RAPD markers (see Figure 3-2 for details). This specific marker configuration requirement greatly reduced the potential number of useful polymorphisms that could be observed in the bulks. Also, a very small proportion of unexpected dual virulence (both hypothesized virulence alleles in the same spore) in the pooled inocula could potentially confound cosegregation of bulk candidate markers and phenotype in the progeny set. A total of 420 RAPD primers were screened for this two-gene BSA work. Five candidate markers with the appropriate marker configuration were observed based on BSA; however, no markers were found to co-segregate with phenotype. Theoretically, two complementary gene pairs and the proposed spore population could explain the observed rust infection in the half-diallel study. However, with the complexity of interactions between the host and pathogen complementary loci and the constraints of the BSA approach, no molecular marker evidence was found to support the two-gene pair interaction hypothesis.

Wilcox (1995) used BSA and co-segregation analysis with parent A progeny to find two RAPD markers that were significantly associated with field resistance (gall vs. no gall phenotype) in parent A. Since the two markers were not associated with each other, Wilcox was uncertain if the markers denoted two different R loci or if they broadly flanked a single locus. This current work used the markers found by Wilcox (A19_560, A11_400) as bulking targets and attempted to place additional markers around them. Several markers were added to the regions around A11_400 and A19_560; however, these markers did not pull the mapped regions together. The uncertainty faced by Wilcox still exists and this study could not determine if the resistance that he observed resided at one or two loci. Had the markers overlapped to form a single linkage group, one locus would have been a very likely answer.

One can not be sure why efforts with mixed inocula failed to yield marker definable R-genes, but the complexity of mixed inocula and the almost certain presence of two or more resistance genes in the host population is not ideal for marker investigations. Ideally, one would like to have inoculum with a single avirulence locus interacting with a single corresponding host resistance locus. The best way to achieve this is through pathogen management and the use of single spore inoculum. Multiple, different inocula each with different avirulence factors could be used to search a particular host for the presence of more than one R-gene. The proposed multiple resistance genes in parent A could be defined one by one by studying the interactions of multiple single spore inocula and progeny of parent A. Amerson and coworkers have used this approach in loblolly pine families to demonstrate two distinct R-genes in a given mother tree (Amerson et al., 1997; Amerson unpublished data). Any further molecular studies undertaken with parent A should utilize this approach with multiple single spore isolates.

Literature Cited

- Amerson, H.V., A.P. Jordan, E.G. Kuhlman, D.M. O'Malley, and R.R. Sederoff. 1997. Genetic basis of fusiform rust disease resistance in loblolly pine. Proc. 24th South For. Tree. Imp. Conf. P. 403.
- Carson, S.D., and M.J. Carson. 1989. Breeding for resistance in forest trees – a quantitative genetics approach. Annu. Rev. Phytopathol. 27: 373-395.
- Crute, I. R., E. B. Holub, J. J. Burdon, Eds. 1997. The gene-for-gene relationship in plant-parasite interactions. CAB International, Wallingford, UK.
- Flor, H.H. 1956. The complementary genetic systems in flax and flax rust. Adv. Genet. 8: 29-54.
- Giovannoni, J.J., R.A. Wing, M.W. Ganai, and S.D. Tanksley. 1991. Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. Nucleic. Acids. Res. 19: 6553–6558.
- Hammond-Kosack, K.E., and J.D.G. Jones. 1996. Resistance gene dependent plant defense response. Plant Cell. 8: 1773-1791.
- Jordan, A.P. 1997. Fusiform rust disease resistance and genomic mapping in loblolly pine. M.S. thesis. North Carolina State University. Raleigh, NC.
- Kinloch, B.B. Jr. and C.H. Walkinshaw. 1991. Resistance to fusiform rust in Southern pines: how is it inherited? In: Proceedings of the IUFRO Rusts of Pine Working Party Conference, Banff, Alberta Canada. Northwest Region., North For. Cent Edmonton, Alberta. Inf. Rep. NOR-X-317, pp 219-228.
- Knighten, J.L., C.H. Young, T.C. McCartney, and R.L. Anderson. 1988. Resistance screening center procedures manual: a step-by-step guide used in the operational screening of southern pines for resistance to fusiform rust. USDA forest service, Forest pest management, Asheville, NC. P.1-62.
- Kuhlman, E.G. 1990. Frequency of single gall isolates of *Cronartium quercuum* f. sp. *fusiforme* with virulence toward three resistant loblolly pine families. Phytopathology. 80:614-617.
- Kuhlman, E.G. 1992. Interaction of virulent single-gall isolates of *Cronartium quercuum* f. sp. *fusiforme* and resistant families of loblolly pine. For. Sci. 38:641-651.
- Kuhlman, E.G. and E.R. Sluder. 1995. Interaction of single-aeciospore isolates of *Cronartium quercuum* f. sp. *fusiforme* with control pollinated loblolly pine families. In: proceedings of the 4th IUFRI Rusts of pines Working Party Conference. Tsukuba. Pp. 139-142.
- Kuhlman, E.G., and F.R. Matthews. 1993. Variation in virulence among single-aeciospore isolates from single-gall isolates of *Cronartium quercuum* f. sp. *fusiforme*. Can. J. For. Res. 23:67-71.
- Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln, and L. Newburg. 1987. MAPMARKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181.
- Martin, G.B., A.J. Bogdanove, and G. Sessa. 2003. Understanding the functions of plant disease resistance proteins. Annual Review of Plant Biology. Jun. Vol. 54: 23-61.
- Martin, G.B., S.H. Brommonschenkel, J. Chunwongse, A. Frary, and M.W. Ganai. 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science. 262:1432-36.
- Matthews, F. W., and S.J. Rowan. 1972. Plant Dis Rep. 56: 931-934.
- Nelson, C.D., R.L. Doudrick, W.L. Nance, J.M. Hamaker, and B. Capo. 1993. Specificity of host:pathogen genetic interaction for fusiform rust disease on slash pine. In: Proceedings of the 22nd Southern Forest Tree Improvement Conference. P. 403-410.
- Newton, A.C., Andrivon, D., 1995. Assumptions and implications of current gene-for-gene hypotheses in existing mapping population. Plant Pathol. 44: 4 607-618.
- Powers, H.R. 1980. Pathogenic variation among single-aeciospore isolates of *Cronartium quercuum* f. sp. *fusiforme*. For. Sci. 26: 280-282.
- Powers, H.R., R.A. Schmidt, and G.A. Snow. 1981. Current status and management of fusiform rust on southern pines. Annual Review. Phytopathology 19: 353-371.
- Robinson, R.A. 1987. Host management in crop pathosystems (1st ed.) New York: MacMillan Publishing Co.
- Snow, G.A., R.J. Dinus, and A.G. Kais. 1975. Variation in pathogenicity of diverse sources of *Cronartium fusiforme* on selected slash pine families. Phytopathology. 65: 170-175.
- Staskawicz BJ, D. Dahlbeck, and N.T. Keen. 1984. Cloned avirulence gene of *Pseudomonas syringae* pv. *glycinea* determines racespecific incompatibility of *Glycine max* (L) Merr. Proc Natl. Acad. Sci. USA 81: 6024-6028.

- Stelzer, H.E., and R.L. Doudrick. 1997. Derivation for host and pathogen genotypes in the fusiform rust pathosystem on slash pine using a complementary genetics model and diallel data. The 24th southern forest tree improvement conference. 320-330.
- Stelzer, H.E., and R.L. Doudrick. 1999. Prescreening slash pine and Cronatium pedigrees for evolution of complementary gene action in fusiform rust disease. Plant Dis. April. 385-388.
- Vision, T.J., D.G. Brown, D.B. Shmoys, R.T. Durrett, and S.D. Tanksley. 2000. Selective mapping: a strategy for optimizing the construction of high-density linkage maps. Genetics. 155: 407-420.
- Von Weissenberg, K. 1990. Host-parasite relationships in forest ecosystems: a review. Silva Fennica. 24 (91): 129-139.
- W.E. van de Weg, 1997. A gene-for-gene model to explain interactions between cultivars of strawberry and races of *phytophthora fragariae var fragariae*. Theor. Appl. Genet. 94: 445-451.
- Wilcox, P. 1995. Genetic Dissection of fusiform rust resistance in loblolly pine. Ph.D thesis. North Carolina State University. Raleigh, NC.
- Wilcox, P.L., H.V. Amerson, E.G. Kuhlman, B.H. Liu, D.M. O'Malley, and R.R. Sederoff, R.R. 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.
- Zobel, B., and J. Talbert. 1984. Applied Forest Tree Improvement. New York: John Wiley and Sons.

DEVELOP STRATEGIES FOR MARKER-ASSISTED BREEDING

The MIM method showed great potentials to provide additional information for the breeding program. Our simulation results showed that the traditional polygenic model would be sufficient for the purpose of variance component estimation and genetic effect prediction if major genes had small effects on the traits. However, when major gene effects are large ($a > 0.5$), the MIM can provide additional information by combining major and polygene effects. Breeding of parents with both high GCA value and favorable alleles of a major gene should accelerate genetic improvement of quantitative traits. Breeding values estimated from the MIM, which appropriately incorporate major genotype information, should be considered for future selection and breeding.

The heritability of economic traits is low in forest trees, approximately 0.20 and 0.15 respectively for height and diameter in loblolly pine. Clonal propagation could increase the heritability of individual tree performance, but loblolly pine genotypes are not easily cloned and field test size would be greatly increased. Marker assisted selection (MAS) could provide another means to enhance within family selection. MAS has not been considered feasible in forest trees for several reasons. A major problem is that QTL detection often uses large mapping populations where hundreds of individuals are planted in a large contiguous block to reduce environmental heterogeneity and increase heritability. Such large testing populations are not common for tree breeding programs. Instead, the progeny trees within families are planted in none-contiguous plots and the family sizes are small. New approaches are needed for QTL detection in the existing breeding populations, as QTLs discovered in breeding populations would be easier to integrate with breeding programs.

The probability of detecting QTLs in breeding populations could be improved by screening biometrical data for families that show evidence of major gene segregation. The application of QTLs for breeding requires that polygenic breeding values be integrated with breeding values for QTLs. Inheritance can be described using the mixed model, which explains phenotypic variation using a single gene effect, a polygenic effect, and an error effect attributed to environmental variation. Breeding value depends on both gene frequency and effect. A computer program, MGene, based on the Bayesian analytical method has been developed for detecting major genes affecting quantitative traits in diallel progeny test populations. The program can detect the major gene segregation, major genotypes and associated probabilities.

With the putatively identified populations, molecular confirmation can be done to identify the markers associated with the major genes. The availability of DNA markers has made QTL analysis and molecular confirmation feasible for many species. Different marker methods have advantages for different purposes. Random Amplified Polymorphic DNA (RAPD) markers provide a powerful means to resolve many anonymous markers in a family with little prior knowledge. RAPD markers have been used extensively to “tag” QTLs, resistance genes, and morphological trait genes. RAPDs are relatively inexpensive DNA markers, and genotyping cost is an important determinant of MAS efficiency. RAPDs usually have a dominant mode of inheritance, but provide many parent specific, 1:1 segregating markers in a pseudotestcross configuration (i.e., one parent is a heterozygote and the other parent is a homozygote band absent). One disadvantage of RAPD markers is that the linkage groups from each parent that

originate from the same chromosome in different parents are difficult to align because the proportion of RAPD markers that segregate in both of a pair of parents is small. Aligning maps and locating QTLs on a genomic map were not important for our objectives. RAPDs can be difficult to reproduce across laboratories, but often can be converted to sequence characterized amplified regions. SCARs are DNA fragments derived from RAPD polymorphisms that are amplified by conventional two-primer polymerase chain reaction. Multi-allelic hypervariable microsatellite markers (SSR) are superior for genotyping diploid individuals and for estimating the genetic effects among families. Microsatellite markers have advantages in QTL studies in polymorphism and reproducibility, especially compared to RAPD markers. A better well defined linkage map will be valuable to estimate more accurate QTL effects and QTL position in the whole genome, which would be more powerful for detecting the QTL effects of major genes among and within families.

The combined information of quantitative genetic analysis for genetic parameter estimates and molecular mapping should improve our breeding strategies for loblolly pine breeding.

MILESTONE OF THE RESEARCH PROJECT

The following is a list of milestones, anticipated completion dates and actual completion dates. Note: the project has been extended to June 30, 2003.

ID Number	Task / Milestone Description	Planned Completion	Actual Completion	Comments
1	Develop genetic models and statistical methods			
1.1	Literature Review	7/1/99	7/1/99	
1.2	Genetic model development	9/30/00	9/30/00	
1.3	Optimize genetic model	9/30/00	9/30/00	
1.4	Computer simulation	12/31/99	12/31/99	
1.5	Analyze progeny test data	9/30/01	12/31/02	Done more than planned
1.6	Develop computer software	9/30/01	12/31/01	
1.7	Identify parents for DNA sampling	5/15/99	5/15/99	
1.8	Verify models with more progeny data	5/15/01	12/31/01	
2	Confirm major genes with molecular markers			
2.1	Field sampling from selected trees	5/15/99	5/1/99	
2.2	DNA extraction	10/30/99	9/10/99	
2.3	Develop molecular markers	9/30/00	12/31/02	Added new markers
2.4	Marker and phenotype association	9/31/01	12/31/01	Added rust association
2.5	Verify markers association with more parent samples	9/31/01	12/31/01	Done more work on rust markers
3	Develop strategies for marker assisted breeding	9/31/01	6/30/03	

RELATED PUBLICATIONS OF THE RESEARCH PROJECT

- Zeng, W., S. Ghosh and B. Li. 2004. Blocking Gibbs Sampling with a Mixed Inheritance for Major Gene Detection. *Genetical Research* (In Press).
- Zeng, W. and B. Li. 2003. Power and robustness of statistical tests for major gene detection in diallel progeny test data. *Forest Science* 49:268-278.
- Sun, S., D.M. O'Malley, B. Li, and B-H. Liu. 2003. Detection of a QTL for 6 year height in a loblolly pine breeding population. (Submitted to *Theoretical & Applied Genetics*).
- Isik, F., B. Li, J. Frampton, and B. Goldfarb. 2003. Efficiency of seedlings and rooted cuttings for testing and selection in *Pinus taeda*. *For. Sci.* 50:00-00 (In Press).
- Sykes, R., F. Isik, B. Li, J. Kadla, and H-m. Chang. 2003. Genetic variation of juvenile wood properties in a loblolly pine progeny test. *TAPPI* 86(12):00-00 (In Press).
- Xiang, B., B. Li, and F. Isik. 2003. Time trend of genetic parameter estimates in growth traits of *Pinus taeda* L. *Silv. Genet.* 51:00-00 (In Press).
- Jansson, G., B. Li, and K. Hannrup. 2003. Time trends in genetic parameters for height and optimal age for parental selection in Scots pine. *For. Sci.* 49:696-705.
- Xiang, B. and B. Li. 2003. Best Linear Unbiased Prediction of Clonal Breeding Values and Genetic Values from Full-sib Mating Designs. *Can. J. For. Res.* 33:2036-2043.
- Isik, F. and B. Li. 2003. Rapid Assessment of Wood Density of Live Trees Using the IML Resi® for Selection in Tree Improvement Programs. *Can. J. For. Res.* 33:1-10.
- Hu, X.S., W. Zeng and B. Li. 2003 Impacts of one-way gene flow on genetic variance components in a natural population. *Silvae Genetica* 51:18-24.
- Xiang, B., B. Li, and S.E. McKeand. 2003. Genetic gain and selection efficiency of loblolly pine in three geographic regions. *For. Sci.* 49: 49:192-208.
- Isik, F., B. Li, and J. Frampton. 2003. Additive, dominance and epistatic genetic variance estimates from a replicated clonal test of loblolly pine. *For. Sci.* 49:77-88.
- McKeand, S.E., H.V. Amerson, **B. Li**, and T.J. Mullin. 2003. Families of loblolly pine that are most stable for resistance to fusiform rust are the least predictable. *Can. J. For. Res.* 33:1335-1339.
- Hu, X.S. and **B. Li**. 2003. On migration load of seeds and pollen grains in a local population. *Heredity* 90:162-168.

- Hu, X.S. and **B Li**. 2002 Linking the evolutionary qualitative genetics to conservation of genetic resources in natural forest populations. *Silvae Genetica* 51:177-183.
- Hu, X. and **B. Li**. 2002. Seed and pollen flow and cline discordance among genes with different models of inheritance. *Heredity* 88:212-217
- Hu, X.S. and B Li. 2002 Linking the evolutionary qualitative genetics to conservation of genetic resources in natural forest populations. *Silvae Genetica* 51: 177-183.
- Wu, R., B. Li., S. S. Wu, and G. Casella. 2001. A maximum likelihood-based method for mining major genes affecting a quantitative character. 57(3) 764-768.
- Xiang, B. and B. Li. 2001. A new mixed analytical method for genetic analysis of diallel data. *Canadian J. For. Research* 31:1-8.
- Hu, X. and B. Li. 2001. Assessment of the ratio of pollen to seed flow in a cline for genetic variation in a quantitative trait. *Heredity* 87(4): 400-409 .
- Wu, R. and B. Li. 2000. A quantitative genetic model for analyzing species differences in outcrossing species. *Biometrics* 56(4): 1098-1104
- Wu, R., B. Li, and Z.B. Zeng. 1999. Book Chapter: Molecular dissection of quantitative traits: new perspectives from *Populus*. In: S.M. Jain and S.C. Minocha (ed.). ***Molecular Biology of Woody Plants***. Vol. 1:475-490. Kluwer Academic Publishers, The Netherlands.
- Zeng, W. 2000. Ph.D. dissertation: Statistical methods for detecting major genes of quantitative traits using phenotypic data of a diallel mating. N. C. State University, Raleigh, NC. P.145.
- Sun, Shuku. 2001. Ph.D. dissertation: Detecting major genes for growth rate with RAPD markers using a diallel cross of loblolly pine. N. C. State University, Raleigh, NC. P.125.
- Li, Hua. 2003. Ph.D. dissertation: Major gene detection for fusiform rust resistance using segregation analysis and linkage analysis in Loblolly pine. N. C. State University, Raleigh, NC. P.150.