Theor Appl Genet (1998) 97:1162–1168

J.-Y. Gai · J.-K. Wang

Identification and estimation of a QTL model and its effects

Received: 11 February 1998 / Accepted: 28 May 1998

Abstract A joint segregation analysis of a genetic system and the effects of QTLs based on the six populations P_1, F_1, P_2, B_1, B_2 and F_2 is proposed in this paper. The major steps were as follows. Firstly, under the supposition that the segregating population was composed of component distributions controlled by a major gene(s) and modified by both polygenes and environments, four groups and 17 types of genetic models, including a one major-gene model, a two major-gene model, a polygene model, and a mixed one-major gene and polygene model, were set up. Secondly, the joint maximum-likelihood function was constructed from the six generations so as to estimate the parameters of component distributions through an EM algorithm. Thirdly, the best-fitting genetic model was chosen according to Akaike's information criterion, a likelihood-ratio test, and tests for goodness of fit. Fourthly, the related genetic parameters, including gene effects, as well as the genetic variances of major genes and polygenes, were obtained from the estimates of component distributions. Finally, the individuals in segregating populations were classified into their major-gene genotypes according to their posterior probabilities. An example of the genetic analysis of plant height of a rice cross between Nanjing No. 6 and Guangcong was used to illustrate the above procedure. The method was especially appropriate to those crops with easy to obtain hybrid seeds.

This project was supported by the National 863 Program of China

Communicated by G. Wenzel

J.-Y. Gai (✉) · J.-K. Wang[1]
Soybean Research Institute, Nanjing Agricultural University, Nanjing 210095, PR China
Fax: +86-25-4395110
E-mail: nausri@public1.ptt.js.cn

Present address:
[1]Laboratory Center, Henan Academy of Agricultural Sciences, Henan 450002, PR China

Key words Quantitative trait loci (QTLs) · Mixed major gene and polygene inheritance model · Maximum-likelihood estimate · EM algorithm · Joint segregation analysis

Introduction

The genetics of a quantitative trait can often be deduced from the statistical analysis of several segregating populations (Mather and Jinks 1982). The fundamental assumption of classical quantitative genetics is the polygene system. But a number of genetic phenomena in plant and animal breeding have indicated that the effects of individual QTLs in the system may differ from each other, and change from environment to environment. Thus, there may be one or a few genes in the QTL system with relatively large genetic effects, referred to as major genes. Those with relatively small effects are called minor genes (or polygenes). Therefore, the inheritance system of a quantitative trait might consist both of a few major genes and a number of polygenes. This genetic model has wide representability and is called mixed the major-gene and polygene inheritance model (or mixed-inheritance model, or mixed-genetic model, in brief). Quite a number of methods have been studied by various researchers to analyze the mixed-inheritance model in human and animal populations (Elston and Stewart 1973; Morton and MacLean 1974; Elston 1984; Famula 1986; Hoeschele 1988; Knott et al. 1991; Guo and Thompson 1992; Fernando et al. 1994; Shoukri and McLachlan 1994; Janss et al. 1995). But these methods are not immediately available for the genetic analysis of plant quantitative traits due to the different mating systems and different breeding objectives involved in plant and animal breeding.

The production of a saturated genetic map by molecular biology, coupled with the genetic analysis of

quantitative traits, has led to the method of QTL mapping, which provides the possibility for determining the inheritance of individual QTLs. But, because of the high cost of the molecular technique, population and sample-size restriction, and the interference of errors, such a QTL technique has not yet become practicable in breeding. Further efforts have, therefore, been undertaken for the improvement of precision in detecting and locating QTLs, as well as estimating their effects. Using a statistical approach, based on the literature cited, Wang (1996) and Wang and Gai (1997) developed the segregation-analysis method to identify the mixed-inheritance model of QTLs and to estimate related genetic parameters; this takes large advantage of the sample size available for plant quantitative traits. The method for individual segregating populations, such as F_2, backcrosses and $F_{2:3}$, has been developed and reported. Based on this, the joint analysis of multiple generations for the five populations P_1, F_1, P_2, F_2 and $F_{2:3}$ and for the six populations P_1, F_1, P_2, B_1, B_2 and F_2 were developed separately, with respect to the degree of difficulty in obtaining hybrid seeds. The latter approach is the subject of the present paper.

The joint segregation analysis method of the six populations

Basic assumptions and genetic models

Four kinds of genetic models, i.e. one major-gene inheritance, two major-gene inheritance, polygene inheritance, and mixed one major-gene and polygene inheritance, were considered. It was assumed that each sample observation was an individual from one of the six populations: the two homozygous parents (denoted by P_1 and P_2), the F_1, the two backcrosses (denoted by B_1 and B_2), and the F_2. Some characteristics of the related populations are listed in Table 1. The underlying assumptions were as follows: diploid nuclear inheritance with no maternal or cytoplasmic effects, no interaction or linkage between major genes and polygenes, and no selection; the genetic effect of polygenes and the effect of the environment in any segregating population followed a normal distribution, and variances within the P_1, P_2 and F_1 populations were equal. Based on these assumptions, four groups and 17 types of genetic models were established as listed in Table 2.

If the two parents differ at only one major locus for a specific quantitative trait, then only three major genotypes are possible. Let A-a represent the alleles of the locus, then the major genotypes for the two parents and the F_1 will be AA, aa and Aa, respectively. The genotypes for backcross B_1 is a 1:1 mixture of AA and Aa, for B_2 a 1:1 mixture of Aa and aa, and for the F_2 a 1:2:1 mixture of AA, Aa and aa. The general distribution forms of the six populations can be written as:

$$P_1 : X_{1i} \sim N(\mu_1, \sigma^2),\ F_1 : X_{2i} \sim N(\mu_2, \sigma^2),\ P_2 : X_{3i} \sim N(\mu_3, \sigma^2),$$

$$B_1 : X_{4i} \sim (1/2)N(\mu_{41}, \sigma_4^2) + (1/2)N(\mu_{42}, \sigma_4^2),$$

$$B_2 : X_{5i} \sim (1/2)N(\mu_{51}, \sigma_5^2) + (1/2)N(\mu_{52}, \sigma_5^2),$$

$$F_2 : X_{6i} \sim (1/4)N(\mu_{61}, \sigma_6^2) + (1/2)N(\mu_{62}, \sigma_6^2) + (1/4)N(\mu_{63}, \sigma_6^2).$$

The populations of P_1, P_2 and F_1 are all distributed as single normal curves; B_1 and B_2 populations are all 1:1 mixtures of two normal curves; and the F_2 population is distributed as a 1:2:1 mixture of three normal distributions. Altogether there are ten component distributions in the six populations. When the genetic model is the mixed one major-gene and polygenes, and all possible genetic effects exist, the ten components are different. Under some specific cases, some components may be the same. For example, when only the major gene exists without polygenes (A-group model), the components will have the following relationships:

$$N(\mu_1, \sigma^2) = N(\mu_{41}, \sigma_4^2) = N(\mu_{61}, \sigma_6^2),$$

$$N(\mu_2, \sigma^2) = N(\mu_{42}, \sigma_4^2) = N(\mu_{62}, \sigma_6^2),$$

$$N(\mu_3, \sigma^2) = N(\mu_{52}, \sigma_5^2) = N(\mu_{63}, \sigma_6^2).$$

In the situation with two major-genes without polygenes, there will be nine component distributions contained in the six populations. Under the polygenic-inheritance model, each of the six populations is considered as a single normal distribution, and there are six different components in the six populations. The genetic parameters contained in each model are given in Table 2. In the present paper, the mixed two major-genes and polygenes model and more complicated models will not be included and are to be left for future papers due to their complication.

Joint multiple-generation likelihood and an EM algorithm for parameter estimation

The EM algorithm (Dempster et al. 1977; Wang and Gai 1997; McLachlan 1988) was exploited to calculate the maximum-likelihood estimates, and will be expounded here for model D. The main principle for other models is almost the same. In the E-step, the logarithm likelihood function of the complete data which are classified by Bayesian rules can be written as:

$$\begin{aligned} L_c(\Phi) = C &+ \Sigma \log f(X_{1i};\, \mu_1, \sigma^2) + \Sigma \log f(X_{2i};\, \mu_2, \sigma^2) \\ &+ \Sigma \log f(X_{3i};\, \mu_3, \sigma^2) \\ &+ \Sigma[W_{4i1} \log f(X_{4i};\, \mu_{41}, \sigma_4^2) + W_{4i2} \log f(X_{4i};\, \mu_{42}, \sigma_4^2)] \\ &+ \Sigma[W_{5i1} \log f(X_{5i};\, \mu_{51}, \sigma_5^2) + W_{5i2} \log f(X_{5i};\, \mu_{52}, \sigma_5^2)] \\ &+ \Sigma[W_{6i1} \log f(X_{6i};\, \mu_{61}, \sigma_6^2) + W_{6i2} \log f(X_{6i};\, \mu_{62}, \sigma_6^2) \\ &+ W_{6i3} \log f(X_{6i};\, \mu_{63}, \sigma_6^2)], \end{aligned}$$

where the range of summations is over individuals and where $f(X_{1i};\, \mu_1, \sigma^2)$ represents the density function of the normal distribution $N(\mu_1, \sigma^2)$, and so on for the others. W_{4i1}, W_{4i2}, W_{5i1}, W_{5i2}, W_{6i1}, W_{6i2} and W_{6i3} are posterior probabilities of samples from B_1, B_2 and F_2 populations under the initial parameter values. In the M-step, the maximum point of $L_c(\Phi)$ can be obtained for model D by computing partial derivatives of $L_c(\Phi)$ for all parameters and letting derivatives be zero. But for models D-1 through D-4, there are still some constraints on the parameters. However, the Lagrange-multiplicator (or λ-multiplicator method) can be used in the maximisation step for those models with constraints. According to the above representation, the procedure to obtain the maximum-likelihood estimates of parameters can be summarized as follows:

(1) choose initial values of component parameters according to the observations;
(2) compute posterior probabilities W_{4i1}, W_{4i2}, W_{5i1}, W_{5i2}, W_{6i1}, W_{6i2} and W_{6i3}, and therefore obtain the logarithm likelihood $L_c(\Phi)$ (E-step) of the complete data;
(3) compute the maximum, or conditional maximum, of $L_c(\Phi)$ and obtain the estimates of means and variances of the component distributions (M-step);
(4) replace initial values with estimates from step (3) and then iterate steps (2) and (3) until a previously selected precision is achieved.

Table 1 The codes and parameters of P_1, F_1, P_2, B_1, B_2 and F_2

Generation	Code	Sample size	Observation	Mean	Variance	Distribution
P_1	1	n_1	X_{1i}	μ_1	σ^2	$N(\mu_1, \sigma^2)$
F_1	2	n_2	X_{2i}	μ_2	σ^2	$N(\mu_2, \sigma^2)$
P_2	3	n_3	X_{3i}	μ_3	σ^2	$N(\mu_3, \sigma^2)$
B_1	4	n_4	X_{4i}	μ_4	σ^2_{B1}	Mixture of two or more normal curves
B_2	5	n_5	X_{5i}	μ_5	σ^2_{B2}	Mixture of two or more normal curves
F_2	6	n_6	X_{6i}	μ_6	σ^2_{F2}	Mixture of two or more normal curves

Table 2 The number of component distributions and estimatable genetic parameters in various genetic models

Model group	Code and implication of model type	Number of component distributions	Number of independent parameters	First-order genetic parameter	Second-order parameter
One major gene	A-1: additive and dominance	3	4	m, d, h	σ^2
	A-2: additive	3	3	m, d (h = 0)	σ^2
	A-3: dominance	3	3	m, d (h = d)	σ^2
	A-4: negative dominance	3	3	m, d (h = −d)	σ^2
Two major genes	B-1: additive, dominance and epistasis	9	10	m, d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l	σ^2
	B-2: additive and dominance	9	6	m, d_a, d_b, h_a, h_b	σ^2
	B-3: additive	9	4	m, d_a, d_b ($h_a = h_b = 0$)	σ^2
	B-4: equal additive	9	3	m, d ($=d_a = d_b$, $h_a = h_b = 0$)	σ^2
	B-5: dominance	9	4	m, d_a ($=h_a$), d_b ($=h_b$)	σ^2
	B-6: equal dominance	9	3	m, d ($=d_a = d_b = h_a = h_b$)	σ^2
Polygene	C: additive, dominance and epistasis	6	10	m, [d], [h], [i], [j], [l]	σ^2_4, σ^2_5, σ^2_6, σ^2
	C-1: additive and dominance	6	7	m, [d], [h]	σ^2_4, σ^2_5, σ^2_6, σ^2
One major gene plus polygenes	D: mixed one major-gene and additive-dominance-epistasis polygenes	10	14	m, d, h, [d], [h], [i], [j], [l]	σ^2_4, σ^2_5, σ^2_6, σ^2
	D-1: mixed one major-gene and additive-dominance polygenes	10	9	m, d, h, [d], [h]	σ^2_4, σ^2_5, σ^2_6, σ^2
	D-2: mixed one additive major-gene and additive-dominance polygenes	10	8	m, d (h = 0), [d], [h]	σ^2_4, σ^2_5, σ^2_6, σ^2
	D-3: mixed one dominance major-gene and additive-dominance polygenes	10	8	m, d (=h), [d], [h]	σ^2_4, σ^2_5, σ^2_6, σ^2
	D-4: mixed one negative dominance major-gene and additive-dominance polygenes	10	8	m, d (= −h), [d], [h]	σ^2_4, σ^2_5, σ^2_6, σ^2

Model selection by AIC and test of fitness

Any constraints on the parameters will automatically lower the maximum likelihood. To cope with this effect, and in general to allow for the fact that different hypotheses depend on different numbers of unknown parameters, Akaike (1977) suggested that the hypothesis maximizing the expected entropy should be selected as the most suitable model. For this purpose, based on goodness-of-fit and parsimony, the hypothesis that leads to the smallest Akaike's Information Criterion (AIC) will be chosen. The AIC was defined as follows:

$$\mathrm{AIC} = (-2)\log(\text{maximum likelihood}) + 2\,(\text{number of independent parameters}).$$

Elston (1984) proposed to select non-nested genetic models by using AIC.

The likelihood-ratio test (LRT), utilizing the statistic $\lambda = 2\log(L_a)\text{-}2\log(L_0)$, is used to compare whether a restricted model (H_0) is compatible to the general model(H_a), where L_a and L_0 are the maximum likelihoods under H_a and H_0, respectively. This difference in λ asymptotically approaches a χ^2 distribution with the degrees of freedom equal to the difference in the number of independent parameters of each model.

After a genetic model is selected through AIC and/or LRT, it is still of importance to determine the goodness-of-fit between the expected values from the selected model and the observed values. Given H_0: $F(x) = F_0(x)$, when the n observations X_i ($i = 1 \ldots n$) are transformed by the accumulated probability transformation

$[Y_i = F_0(X_i)]$, n independent observations Y_i ($i = 1 \ldots n$) uniformly distributed on the interval (0, 1) can be obtained when H_0 holds. Consequently, the following three statistics can be used to test the hypothesis H_0.

$U_1^2 = 12[\Sigma F(X_i) - n/2]^2/n \sim \chi^2(1)$,
to test whether the mean of Y_i is 1/2;

$U_2^2 = (45/4)[\Sigma F(X_i)^2 - n/3]^2/n \sim \chi^2(1)$,
to test whether the second moment of Y_i is 1/3;

$U_3^2 = 180[\Sigma(F(X_i) - 0.5)^2 - n/12]^2/n \sim \chi^2(1)$,
to test whether the variance of Y_i is 1/12.

Let $F_n(x)$ be the empirical distribution function, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order-statistics, $F_0(x)$ be the expected distribution function (population distribution derived from the selected genetic model), the Smirnov statistic ${}_nW^2$ can be used to test H_0: $F_n(x) = F_0(x)$. The distribution of ${}_nW^2$ does not depend on $F_0(X)$, so the test is completely distribution-free. The asymptotic distribution of ${}_nW^2$ is reached remarkably rapidly, the exact distribution for $n = 3$ being close to it (Kendall and Stuart 1979).

Another important general test of goodness-of-fit is the Kolmogorov test. Like ${}_nW^2$, it is based on deviations of the sample density function $F_n(x)$ from $F_0(x)$. The measure of deviation used is very much simpler, being the maximum absolute difference between $F_n(x)$ and $F_0(x)$, i.e.,

$D_n = \mathrm{Sup}\,|F_n(x) - F_0(x)|$.

The distribution of D_n is completely distribution-free when H_0 holds.

In the joint segregation analysis of the mixed genetic model, AIC will be employed to determine which model group is most fitting, LRT will be used to choose the simplest type within the model group, and tests for goodness-of-fit will be used to determine whether the selected model sufficiently explains the data. If, for a particular genetic model, none of these five statistics are significant, one can be reasonably sure that the data adequately fit the model.

Estimation of genetic parameters

Genetic parameters can be computed from the estimates of component parameters in the corresponding model. Taking model D as an example, the first-order genetic parameters can be calculated by least squares from the following equations (Mather and Jinks 1982):

$\mu_1 = m + d + [d] + [i]$;

$\mu_2 = m + h + [h] + [l]$;

$\mu_3 = m - d - [d] + [i]$;

$\mu_{41} = m + d + (1/2)[d] + (1/2)[h] + (1/4)[i] + (1/4)[j] + (1/4)[l]$;

$\mu_{42} = m + h + (1/2)[d] + (1/2)[h] + (1/4)[i] + (1/4)[j] + (1/4)[l]$;

$\mu_{51} = m + h - (1/2)[d] + (1/2)[h] + (1/4)[i] - (1/4)[j] + (1/4)[l]$;

$\mu_{52} = m - d - (1/2)[d] + (1/2)[h] + (1/4)[i] - (1/4)[j] + (1/4)[l]$;

$\mu_{61} = m + d + (1/2)[h] + (1/4)[l]$;

$\mu_{62} = m + h + (1/2)[h] + (1/4)[l]$;

$\mu_{63} = m - d + (1/2)[h] + (1/4)[l]$,

where m is the population mean, d and h are the additive and dominance effects of major genes respectively, and [d], [h], [i], [j] and [l] are additive, dominance, additive-additive, additive-dominance and dominance-dominance epistasis effects, respectively. The phenotypic variance (σ_p^2) of B_1, B_2 and F_2 can be directly calculated from the observation data. σ^2 in the phenotypic variance of P_1, P_2 and F_1 can be regarded as the environmental variance (σ_e^2) since there is no genetic variation in each of the three populations; σ_4^2 is the variance of component distribution in B_1 which consists of polygenic variance (σ_{pg}^2) and environmental variance (σ_e^2). Thus $\sigma_p^2 = \sigma_{mg}^2 + \sigma_{pg}^2 + \sigma_e^2$ and $\sigma_4^2 = \sigma_{pg}^2 + \sigma_e^2$ for the B_1 population. Therefore, the major-gene variance σ_{mg}^2 and the polygenic variance σ_{pg}^2 in B_1 can both be estimated, and the major-gene heritability (h_{mg}^2) and polygenic heritability (h_{pg}^2) can also be estimated from $h_{mg}^2 = \sigma_{mg}^2/\sigma_p^2$ and $h_{pg}^2 = \sigma_{pg}^2/\sigma_p^2$. The principle is the same for calculating σ_{mg}^2, σ_{pg}^2, h_{mg}^2 and h_{pg}^2 in B_2 and F_2.

Posterior genotype probabilities

For a general mixture having the density function form $p(x; \phi) = \sum_{j=1}^{k} \pi_j f(x; \theta_j)$, the posterior probabilities $W_t (t = 1, \ldots, g)$ of a sample having x can be computed as:

$$W_t = \pi_1 f(x; \theta_t)/p(x; \phi), \quad \sum_{t=1}^{g} W_t = 1.$$

For model D, the posterior probabilities of individuals in B_1, B_2, and F_2 will be:

B_1: $W_{4i1} = f(X_{4i}; \mu_{41}, \sigma_4^2)/[f(X_{4i}; \mu_{41}, \sigma_4^2) + f(X_{4i}; \mu_{42}, \sigma_4^2)]$

$W_{4i2} = f(X_{4i}; \mu_{42}, \sigma_4^2)/[f(X_{4i}; \mu_{41}, \sigma_4^2) + f(X_{4i}; \mu_{42}, \sigma_4^2)]$

B_2: $W_{5i1} = f(X_{5i}; \mu_{51}, \sigma_5^2)/[f(X_{5i}; \mu_{51}, \sigma_5^2) + f(X_{5i}; \mu_{52}, \sigma_5^2)]$

$W_{5i2} = f(X_{5i}; \mu_{52}, \sigma_5^2)/[f(X_{5i}; \mu_{51}, \sigma_5^2) + f(X_{5i}; \mu_{52}, \sigma_5^2)]$

F_2: $W_{6i1} = f(X_{6i}; \mu_{61}, \sigma_6^2)/[f(X_{6i}; \mu_{61}, \sigma_6^2)] + 2f(X_{6i}; \mu_{62}, \sigma_6^2) + f(X_{6i}; \mu_{63}, \sigma_6^2)]$

$W_{6i2} = 2f(X_{6i}; \mu_{62}, \sigma_6^2)/[f(X_{6i}; \mu_{61}, \sigma_6^2) + 2f(X_{6i}; \mu_{62}, \sigma_6^2) + f(X_{6i}; \mu_{63}, \sigma_6^2)]$

$W_{6i3} = f(X_{6i}; \mu_{63}, \sigma_6^2)/[f(X_{6i}; \mu_{61}, \sigma_6^2) + 2f(X_{6i}; \mu_{62}, \sigma_6^2) + f(X_{6i}; \mu_{63}, \sigma_6^2)]$.

An example

The frequency distributions of plant height in the six populations of a rice cross between Nanjing No. 6 and Guangcong are given in Table 3. It is obvious that the F_1 population has a tendency toward the high parent; B_1 shows a single mode in the high-plant height direction; B_2 shows two modes in both dwarf and high directions; and so does the F_2 population. The former conclusion from the Mendelian method was one recessive gene controlling the dwarf trait. But if the continuity in the populations and components is taken into consideration, it is important to distinguish the polygenic variation and the environmental variation from the continuous variation.

From what has been discussed above, the maximum logarithm likelihood, the AIC value and the maximum-likelihood estimates in each genetic model were calculated, and the results are listed in Table 4. From the result of the test of fitness for model-C listed in Table 5, the following conclusions can be drawn: the homogenous populations P_1, F_1 and P_2 are distributed as a

Table 3 The frequency distribution of plant height in the P_1, F_1, B_1, B_2 and F_2 of the cross between Nanjing No. 6 (P_1) and Guangcong (P_2) cm

	80–	85–	90–	95–	100–	105–	110–	115–	120–	125–	130–	135–	140–	145–	150–	155–	160–	165–	170–	175–180–185
P_1														4	5	12	22	2		
F_1												3	17	8	4					
P_2			1	11	13	18	5													
B_1												20	14	24	13	12	3			
B_2	1	10	18	27	18	11	2				1	3	14	19	22	11	4	1		
F_2	5	9	15	13	13	11	14	14	8	2	9	9	41	30	69	85	72	50	23	6	2

Table 4 The AIC values and maximum-likelihood estimates under various genetic models. Note: "–" represents no such parameter in this model: the EM algorithm for model B-6 not converged

Model	AIC	μ_1	μ_2	μ_3	μ_{41}	μ_{42}	μ_{51}	μ_{52}	μ_{61}	μ_{62}	μ_{63}	σ^2	σ_4^2	σ_5^2	σ_6^2
A-1	7230.75	161.39	151.89	100.86	–	–	–	–	–	–	–	70.74	–	–	–
A-2	7976.79	168.54	139.18	109.81	–	–	–	–	–	–	–	180.29	–	–	–
A-3	7243.94	155.38	155.38	100.97	–	–	–	–	–	–	–	85.93	–	–	–
A-4	8251.19	157.05	134.51	134.51	–	–	–	–	–	–	–	560.78	–	–	–
B-1	7063.63	162.01	150.76	103.18	166.20	147.07	159.45	92.15	136.56	115.85	–	27.20	–	–	–
B-2	7089.38	161.94	149.77	103.51	143.64	168.07	159.60	93.68	153.47	111.98	–	30.36	–	–	–
B-3	7975.95	167.79	139.16	110.53	169.03	137.90	140.40	109.30	170.26	108.06	–	178.81	–	–	–
B-4	8192.44	157.45	133.47	109.49	169.46	169.46	121.48	121.48	133.47	133.47	–	388.88	–	–	–
B-5	7245.71	155.26	155.26	101.26	155.26	155.26	100.66	155.86	155.86	100.66	–	85.88	–	–	–
C	7583.21	162.54	148.31	103.58	150.95	–	123.54	–	145.11	–	–	19.35	150.95	123.54	145.11
C-1	7620.77	162.09	147.84	103.88	154.97	–	126.10	–	140.53	–	–	19.39	69.21	126.10	140.53
D	6993.71	162.54	148.31	103.58	156.98	146.01	154.17	97.13	156.68	155.70	103.25	19.35	24.96	44.14	106.93
D-1	7043.81	162.31	148.37	103.93	157.78	152.90	154.30	98.00	156.02	151.16	99.74	19.28	66.34	44.55	109.30
D-2	7618.73	162.12	147.91	103.89	161.77	148.25	132.66	119.14	153.97	140.45	126.94	19.37	25.38	809.40	489.79
D-3	7044.26	162.31	148.35	103.92	155.33	155.33	153.95	98.32	154.64	154.64	99.01	19.28	72.23	44.93	113.24
D-4	8336.24	161.24	140.84	100.36	145.81	156.26	73.30	73.30	165.72	176.16	176.17	35.76	25.63	3376.7	1351.9

Table 5 Tests of goodness-of-fit for models A-1, B-1, C and D in various populations. Note: In parentheses is the probability value

Model	Generation	U_1	U_2	U_3	${}_nW^2$	D_n
A-1	P_1	2.39(0.2)	0.06(0.81)	25.26***	1.39***	0.30***
	F_1	8.93**	12.63***	6.97**	1.22***	0.41***
	P_2	7.51**	3.51(0.06)	9.72**	1.45***	0.26**
	B_1	33.61***	30.42***	0.15(0.69)	3.10***	0.33***
	B_2	2.41(0.12)	0.20(0.65)	17.78***	0.76**	0.14**
	F_2	13.26***	13.67***	0.47(0.49)	1.58***	0.10***
B-1	P_1	1.99(0.16)	0.75(0.38)	3.96*	0.61*	0.21*
	F_1	9.19**	9.48**	0.33(0.57)	0.91**	0.37**
	P_2	0.59(0.44)	0.40(0.53)	0.20(0.66)	0.24(>0.10)	0.15(>0.10)
	B_1	30.31***	29.47***	0.15(0.70)	2.71***	0.29***
	B_2	2.17(0.14)	2.57(0.11)	0.51(0.47)	0.20(>0.10)	0.07(>0.10)
	F_2	23.35***	24.62***	1.29(0.26)	2.67***	0.14***
C	P_1	0.48(0.49)	0.19(0.66)	0.87(0.35)	0.26(>0.10)	0.14(>0.10)
	F_1	0.04(0.84)	0.31(0.58)	2.10(0.15)	0.19(>0.10)	0.16(>0.10)
	P_2	0.10(0.75)	0.32(0.57)	1.09(0.30)	0.18(>0.10)	0.13(>0.10)
	B_1	0.08(0.78)	0.00(0.95)	0.67(0.41)	0.10(>0.10)	0.09(>0.10)
	B_2	0.08(0.78)	0.82(0.36)	22.03***	2.33***	0.22***
	F_2	9.49**	7.04**	1.73(0.19)	5.57***	0.20***
D	P_1	0.48(0.49)	0.19(0.66)	0.87(0.35)	0.26(>0.10)	0.14(>0.05)
	F_1	0.04(0.84)	0.31(0.58)	2.10(0.15)	0.19(>0.10)	0.16(>0.05)
	P_2	0.10(0.75)	0.32(0.57)	1.09(0.30)	0.18(>0.10)	0.13(>0.10)
	B_1	0.43(0.51)	0.30(0.58)	0.11(0.73)	0.12(>0.10)	0.12(>0.10)
	B_2	0.69(0.40)	0.63(0.43)	0.00(0.96)	0.06(>0.10)	0.06(>0.10)
	F_2	3.97*	2.64(0.10)	1.50(0.22)	0.67*	0.07*

*, **, ***Represent the 0.05, 0.01, 0.001 significance levels, respectively

normal distribution, there is no requirement for data transformation and, if the segregation population is a mixture, it should be a mixture of normal distributions. From Table 4, model-D has the least AIC value, D-1 the second and D-3 the third. So the D-group model is the most fitting model to explain the inheritance of the quantitative trait in this example according to Akaike' s Information Criterion. The results from LRT between models D and D-1, D and D-3, all indicate that D is more suitable than D-1 and D-3. The results in Table 5 also show the fitness of model D. So one can reasonably deduce that the plant-height trait in the cross is dominated by a mixture of a partially dominant major gene plus additive-dominance-epistasis polygenes.

The first-order and second-order genetic parameters in model D, calculated from the results in Table 4, and the components in each segregating population, are given in Table 6. The plant height difference between Nanjing No. 6 and Guangcong is controlled by a mixed one major-gene and polygenes. The additive effect of the major gene is estimated as 29.09 cm. The high-plant trait is one of partial dominance, and the dominance ratio of the major-gene is about 0.83. The major-gene variations in B_1, B_2 and F_2 are 53.0%, 94.8% and 81.0% of their total phenotypic variations respectively, and are the main components. The polygenic variations are 10.6%, 2.9% and 15.5% of their phenotypic variations, and are less important components. Thus, to control the major gene means to control a large proportion of the phenotypic variation.

The most probable major-gene genotype of an individual in segregating populations is given in Table 7. For the B_1 population, individuals having a plant height 140–151 cm can be classified into the Aa genotype; but some of them with the plant height 145–151 cm have a 0.06–0.47 probability of being AA. Those with a plant height of 152–167 can be classified into the AA genotype; but some of them with a plant height of 152–158 cm have a 0.43–0.05 probability of being Aa. The genotypes of individuals in B_2 can be clearly determined, i.e. those with a plant height of 80–112 cm have the genotype aa and those with a plant height of 139–170 cm have the genotype Aa. For the F_2 population, individuals with a plant height of 80–127 cm can be classified as genotype aa; but some of them with a plant height of 123 cm have a 0.07 probability of being Aa, while some of them having a plant height of 124–127 cm have a 0.05–0.15 probability of being Aa and with 0.10–0.31 being AA. Those with a plant height of 130–180 cm can be classified as Aa genotypes; but some of them with a plant height of 130–132 cm have a 0.26–0.30 probability of being Aa and a 0.22–0.09 probability of being aa. Those with a plant height of 134–180 cm have a 0.65–0.66 probability of being Aa and a 0.32–0.34 probability of being AA. The Aa component distribution overlaps with the AA component, consequently individuals falling in this area can not be classified definitely. However, this is possible if further progeny tests are carried out.

Table 6 The estimates of genetic parameters of the cross between Nanjing No. 6 (P_1) and Guangcong (P_2)

First-order parameter	Estimate	Second-order parameter	Estimate and component distribution		
			B_1	B_2	F_2
d	29.15	σ^2_p	53.11	853.26	563.86
h	24.12	σ^2_{mg}	28.15	809.12	456.93
h/d	0.83	σ^2_{pg}	5.61	24.79	87.58
[d]	0.33	σ^2_e	19.35	19.35	19.35
[h]	−49.24	h^2_{mg} (%)	53.00	94.82	81.03
[i]	−15.96	h^2_{pg} (%)	10.56	2.90	15.53
[j]	−7.27	Components	N(156.64, 24.96)	N(154.17, 44.14)	N(56.68, 106.93)
[l]	24.41		N(146.01, 24.96)	N(97.13, 44.14)	N(155.70, 106.93)
					N(103.25, 106.93)

Table 7 The Bayesian classification of individuals of the segregating populations

B_1			B_2			F_2		
X (cm)	f	Genotype	X (cm)	f	Genotype	X (cm)	f	Genotype
140–144	20	Aa	80–112	89	aa	80–122	98	aa
145–151	28	Aa + AA	139–170	75	Aa	123	2	aa + Aa
152–158	22	AA + Aa				124–127	4	aa + Aa + AA
159–167	16	AA				130–132	6	Aa + AA + aa
						134–180	390	Aa + AA

Discussion

The present procedure is basically established on the mixed one major-gene plus polygene inheritance theory according to Elston (1984), combining it with the joint maximum-likelihood fuction and the EM algorithm for model fitting, the AIC criterion, the likelihood-ratio test and the goodness-of-fit tests for model selection, and the posterior probability for the major-gene genotypic grouping of individuals, to form a system for handling the joint analysis of multiple generations, including P_1, F_1, P_2, B_1, B_2 and F_2. This procedure is especially appropriate for those crops with easy to obtain hybrid seeds since backcrosses need to be made. For those crops not easy to obtain hybrid seeds, the joint segregation analysis based on the five populations P_1, P_2, F_1, F_2 and $F_{2:3}$ (F_2-derived line) may be adopted.

For an effective use of this prodecure, sample size is of importance. Generally speaking, it should be greater than 30 for a homozygous population and greater than 100 for a segregating population. The greater the population size, the more precise are the results that will be achieved.

The present procedure takes into consideration the following four kinds of genetic models: one major-gene inheritance, two major-genes inheritance, polygenic inheritance, and mixed one-major gene and polygene inheritance. By using this procedure, the most suitable model can be selected for a set of data. However, the study still needs to be completed for the mixed model containing two, or even more, major, genes. Furthermore, significant errors might exist for the genetic data based on single-plant measurements in the six generations. To overcome this disadvantage, the data based on a plot measurement in F_2-, B_1- and B_2-derived lines (i.e. $F_{2:3}$, $B_{1:2}$ and $B_{2:2}$) should be used to reduce the experimental errors. The joint segregation analysis based on these family populations also remains to be developed.

Acknowledgements The authors wish to thank Dr. Eingegangen and an anonymous reviewer for their relevant and constructive comments and suggestions.

References

Akaike H (1977) On the entropy maximum principle. In: Krishnaiah PR (ed) Applications of statistics. North-Holland Publishing Company, Amsterdam, pp 27–41

Dempster AP, Laird NM, Robin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39 : 1–38

Elston RC (1984) The genetic analysis of quantitative trait differences between two homozygous lines. Genetics 108 : 733–744

Elston RC, Stewart J (1973) The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. Genetics 73 : 695–711

Famula TR (1986) Identifying single genes of large effect in quantitative traits using best linear unbiased prediction. J Anim Sci 63 : 68–76

Fernando RL, Stricker C, Elston RC (1994) The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance. Theor Appl Genet 88 : 573–580

Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51 : 1111–1126

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

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

Kendall MG, Stuart A (1979) The advanced theory of statistics, vol 2. Inference and relationship. Charles Griffin and Company Limited, London

Knott SA, Haley CS, Thompson R (1991) Methods of segregation analysis for animal breeding data: a comparison of power. Heredity 68 : 299–311

Mather K, Jinks JL (1982) Biometrical Genetics, 2nd edn. Chapman and Hall Ltd, London

McLachlan GJ (1988) Mixture models: inference and applications to clustering. Marcel Dekker, Inc., New York

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

Shoukri MM, McLachlan GJ (1994) Parametric estimation in a genetic mixture model with application to nuclear family data. Biometrics 50 : 128–139

Wang JK (1996) Studies on identification of major-polygene mixed inheritance of quantitative traits and estimation of genetic parameters. Doctorate dissertation, Department of Plant Breeding and Biometrics, Nanjing Agricultural University

Wang JK, Gai JY (1997) Identification of major gene and polygene mixed inheritance and estimation of genetic parameters in F_2 progeny. Chinese J Genet 24 : 181–190