Genet. Res., Camb. (2003), **81**, *pp.* 157–163. © 2003 Cambridge University Press
DOI: 10.1017/S0016672303006141 Printed in the United Kingdom

The EIM algorithm in the joint segregation analysis of quantitative traits

YUAN-MING ZHANG[1]*, JUN-YI GAI[1] AND YONG-HUA YANG[2]

[1] *National Key Laboratory of Crop Genetics and Germplasm Enhancement, Soybean Research Institute, Nanjing Agricultural University; and Chinese National Center for Soybean Improvement, Ministry of Agriculture, Nanjing 210095, P.R. China*

[2] *State Key Laboratory of Pharmaceutical Biotechnology, Plant Cell Physiology and Molecular Biology Laboratory, Department of Biological Science and Technology, College of Life Science, Nanjing University, Nanjing 210093, P.R. China*

(*Received 12 July 2002 and in revised form 25 November 2002*)

Summary

In this article, a new algorithm for obtaining the maximum likelihood estimators (MLEs) of parameters in the joint segregation analysis (JSA) of multiple generations of P_1, F_1, P_2, F_2 and $F_{2:3}$ (MG5) for quantitative traits was set up. Firstly, owing to the fact that the component variance of the heterogeneous genotype in $F_{2:3}$ included both the first-order genetic parameters (denoted by the means of distributions) and the second-order parameters, a simple closed form for the MLEs of the means of component distributions did not exist while the expectation and maximization (EM) algorithm was used. To simplify the estimation of parameters, the first partial derivative of the above variance on the mean in the sample log-likelihood function was omitted. However, this would be remedied by the iterated method. Then, variances of component distributions for segregating populations were partitioned into major-gene, polygenic and environmental variances so that the generally iterated formulae for estimating the means as well as polygenic and environmental variances of component distributions in the maximization step (M-step) of the EM algorithm were obtained. Therefore, the EM algorithm for estimating parameters in the JSA model for the MG5 was simplified. This is called the expectation and iterated maximization (EIM) algorithm. Finally, an example of the inheritance of the resistance of soybean to beanfly showed that the results of mixed inheritance analysis in this paper coincided with those in both Wang & Gai (2001) and Wei *et al.* (1989), so the EIM algorithm was appropriate.

1. Introduction

The results from both plant breeding and QTL (quantitative trait loci) mapping show that the inheritance system of quantitative traits consists of both a few major genes and a number of polygenes (Elston *et al.*, 1973; Paterson, 1997; Kearsey & Farquhar, 1998; Gai & Wang, 1998; Wang & Gai, 2001; Zhang, 2001). The mixed major-gene plus polygenes inheritance model was first studied in human genetics and animal breeding (Elston *et al.*, 1973). Recently, Wang (1996) and Gai *et al.* (2003) applied the mixed inheritance model to the genetic study of plant quantitative traits, such as maturity, tofu quality, cyst nematode resistance and foliar feeding insect resistance in soybean, bacterial blight and wide compatibility in rice, maturity in rapeseed and dwarf mosaic virus resistance in maize, where the expectation and maximization (EM) algorithm (Dempster *et al.*, 1977) was used for parameter estimation.

Using the EM algorithm, the joint segregation analysis (JSA) method for the multiple generations P_1, F_1, P_2, F_2 and $F_{2:3}$ (denoted by MG5) was set up, involving a mixed one-major-gene plus polygenes inheritance model (Wang & Gai, 1998). It was noticed that the distribution variance of heterogeneous genotype line in $F_{2:3}$ included both the first-order genetic parameters (denoted by the means of distributions) and the second-order parameters, so that a simple closed form for the maximum likelihood estimators (MLEs) of the means of the component distributions did not exist. To simplify the estimation of parameters, the first partial derivative of the above variance on the mean in the sample log-likelihood function was omitted. However, when the mixed

* Correponding author. e-mail: soyzhang@njau.edu.cn

Table 1. *Genetic models in the joint segregation analysis of the five generations of* P_1, F_1, P_2, F_2 *and* $F_{2:3}$

			Model code	
Class	Major gene	Polygenes	Only major gene	Mixed major gene and polygenes
Polygenes	–	Additive-dominant-epistasis, [d], [h], [i], [j], [l]	–	C
	–	Additive-dominant, [d], [h]	–	C-1
A major gene	Additive-dominant, d, h	Additive-dominant-epistasis, [d], [h], [i], [j], [l]	A-1	D
	Additive-dominant d, h	Additive-dominant, [d], [h]	A-1	D-1
	Additive, $d(h=0)$	Additive-dominant, [d], [h]	A-2	D-2
	Completely dominant, $d(h=d)$	Additive-dominant, [d], [h]	A-3	D-3
	Completely negative dominant, $d(h=-d)$	Additive-dominant, [d], [h]	A-4	D-4
Two major genes	Additive-dominant-epistasis, $d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l$	Additive-dominant-epistasis, [d], [h], [i], [j], [l]	B-1	E
	Additive-dominant-epistasis, $d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l$	Additive-dominant, [d], [h]	B-1	E-1
	Additive-dominant, $d_a, d_b, h_a, h_b, i=j_{ab}=j_{ba}, l$	Additive-dominant, [d], [h]	B-2	E-2
	Additive, $d_a, d_b, h_a=h_b=0$	Additive-dominant, [d], [h]	B-3	E-3
	Equally additive, $d(=d_a=d_b, h_a=h_b=0)$	Additive-dominant, [d], [h]	B-4	E-4
	Completely dominant, $d_a=h_a, d_b=h_b$	Additive-dominant, [d], [h]	B-5	E-5
	Equally dominant, $d=d_a=h_a=d_b=h_b$	Additive-dominant, [d], [h]	B-6	E-6

d, h: additive and dominance effects of major gene for model A and D; d_a, h_a: additive and dominance effects of the first major gene for model B and E; d_b, h_b: additive and dominance effects of the second major gene for model B and E; i, j_{ab}, j_{ba} and l: additive × additive, additive × dominance, dominance × additive, dominance × dominance epistatic effects between the two major genes; [d], [h], [i], [j], [l]: additive effects, dominance effects, additive × additive, additive × dominance (or dominance × additive) and dominance × dominance epistatic effects for the polygene system.

two-major-gene plus polygenes inheritance model was extended (Zhang, 2001), the inexact results of the parameter estimation could be ascribed to the above simplification. In the present analysis, the iterated method is used to remedy that simplification. Under the condition that the component variances of segregating generations were partitioned into major-gene, polygenic and environmental variances, and that the iterated method was used in the maximization step of EM algorithm, the EM algorithm could be brought into effect. This is called the expectation and iterated maximization (EIM) algorithm. An illustrative example is given at the end of the paper.

2. The EIM algorithm of JSA of MG5 for parameter estimation

(i) *Basic assumption*

The underlying assumptions are as follows: diploid nuclear inheritance with no maternal or cytoplasmic effects, no interaction or linkage between major genes and polygenes, and no selection; the polygenic effect and the environmental effect in any segregating population follow a normal distribution; and the variances within the two homozygous parents (P_1 and P_2) and the F_1 populations are equal.

(ii) *Notion*

Let x_{1i}, x_{3i}, x_{2i}, x_{4i} and x_{5i} be a random sample of observations or means of lines from a finite mixture distribution with one, one, one, k_1 and k_2 normal components for parent P_1 and P_2, the F_1, F_2 and $F_{2:3}$, respectively and n_j and m_j ($j=1, \ldots, 5$) the corresponding sample size and mean of populations including the polygenic effects. The Mather & Jinks (1982) notation was used in this paper. The notations and their meanings of genetic parameters are shown in Table 1.

(iii) *Genetic model*

Five kinds of genetic models, A, B, C, D and E, are considered as listed in Table 1. If the two parents differ at only two major loci for a specific quantitative trait, then only nine major genotypes are possible. Let A-a and B-b represent the alleles of the loci, then the major genotypes for the two parents and F_1 will be AABB, aabb and AaBb, while for F_2 a 1:2:1:2:4:2:1:2:1 mixture of (1) AABB, (2) AABb, (3) AAbb, (4) AaBB, (5) AaBb, (6) Aabb, (7) aaBB, (8) aaBb and (9) aabb is expected. For $F_{2:3}$ the same mixture of major genotypes corresponding to the major genotypes in F_2 is expected. The general distribution forms of the five populations can be written as:

$$P_1\colon x_{1i} \sim N(\mu_1, \sigma^2); \quad F_1\colon x_{2i} \sim N(\mu_2, \sigma^2);$$

$$P_2\colon x_{3i} \sim N(\mu_3, \sigma^2); \quad F_2\colon x_{4i} \sim \sum_{t=1}^{k_1} p_{1t} N(\mu_{4t}, \sigma_{4t}^2);$$

$$F_{2:3}\colon x_{5i} \sim \sum_{t=1}^{k_2} p_{2t} N(\mu_{5t}, \sigma_{5t}^2),$$

where $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2; p_{1t} and p_{2t} are the segregation proportions of the nine major genotypes in the segregating populations F_2 and $F_{2:3}$, respectively; $\sigma_{41}^2 = \cdots = \sigma_{49}^2 = \sigma_4^2$; x_{5i} is the mean of the ith line in $F_{2:3}$. In the genetic experiment, some F_2 seeds were reserved and planted to form $F_{2:3}$, where each line in $F_{2:3}$ population was formed by some F_2 seeds on a plant. In the next year, P_1, F_1, P_2, F_2 and $F_{2:3}$ populations were simultaneously planted by the randomized design. Thus $F_{2:3}$ was independent of F_2. Therefore, the sample likelihood function for JSA of MG5 is

$$\begin{aligned} L = & \prod_{i=1}^{n_1} f(x_{1i}; \mu_1, \sigma^2) \prod_{i=1}^{n_2} f(x_{2i}; \mu_2, \sigma^2) \\ & \prod_{i=1}^{n_3} f(x_{3i}; \mu_3, \sigma^2) \prod_{i=t=1}^{n_4} \sum^{k_1} p_{1t}\, f(x_{4i}; \mu_{4t}, \sigma_4^2) \\ & \prod_{i=t=1}^{n_5} \sum^{k_2} p_{2t}\, f(x_{5i}; \mu_{5t}, \sigma_{5t}^2), \end{aligned} \qquad (1)$$

where $f(x_i; \mu, \sigma^2)$ represents the density function of a normal distribution $N(\mu, \sigma^2)$.

(iv) *Partitioning of variances of components*

If model E is considered for a quantitative trait, let σ_{40}^2 and σ_{50}^2 be the polygenic variances of F_2 and $F_{2:3}$ populations, respectively. The variances σ_4^2 and $\sigma_{5t}^2 (t=1, \ldots, k_2)$ in (1) will have the following relationships:

$$\begin{aligned} & \sigma_4^2 = \sigma_{40}^2 + \sigma^2 \\ & \sigma_{5t}^2 = \sigma_{50}^2 + \sigma^2/n + V_{MGt} \quad (t=1, \ldots, k_2), \end{aligned} \qquad (2)$$

where $V_{MGt} (t=1, \ldots, k_2)$ is the variance component involved in the genetic effects of major genes, called major-gene variance; n the number of plants observed in a line; σ^2 environmental variance. When $k_2 = 9$,

$$\begin{aligned} & V_{MG1} = V_{MG3} = V_{MG7} = V_{MG9} = 0 \\ & V_{MG2} = [1/2(d_b + i)^2 + 1/4(h_b + j_{ab})^2]/n \\ & V_{MG4} = [1/2(d_a + i)^2 + 1/4(h_a + j_{ba})^2]/n \\ & V_{MG5} = [d_a{}^2 + d_b{}^2 + i^2 + (d_a + j_{ab})^2 + (d_b + j_{ba})^2 \\ & \qquad + (h_a + 1/2l)^2 + (h_b + 1/2l)^2 + 1/4l^2]/4n \\ & V_{MG6} = [1/2(d_a - i)^2 + 1/4(h_a - j_{ba})^2]/n \\ & V_{MG8} = [1/2(d_b - i)^2 + 1/4(h_b - j_{ab})^2]/n. \end{aligned}$$

Thus, the variance is partitioned into major-gene, polygenic and environmental variances.

(v) *The EIM algorithm in the estimation of component parameters*

The EIM algorithm includes both the E-step and the iterated M-step (IM-step). In the E-step, the expected complete data log-likelihood for the JSA method of MG5 can be written as

$$\begin{aligned} L(\theta / Y) = & \sum_{j=1}^{3} \sum_{i=1}^{n_j} \log f(x_{ji}; \mu_j, \sigma^2) \\ & + \sum_{j=4}^{5} \sum_{i=1}^{n_j} \sum_{t=1}^{k_{j-3}} w_{jit} \log f(x_{ji}; \mu_{jt}, \sigma_{jt}^2), \end{aligned} \qquad (3)$$

where w_{jit} is the posterior probability of the tth major genotype from the ith individual of $F_2 (j=4)$ or the ith line of $F_{2:3} (j=5)$, $\sigma_{4t}^2 = \sigma_4^2 (t=1, \ldots, k_1)$, $\theta = (\mu_1, \mu_2, \mu_3 , \mu_{41}, \ldots, \mu_{4k_1}, \sigma_4^2, \mu_{51}, \ldots, \mu_{5k_2}, \sigma_{51}^2, \ldots, \sigma_{5k_2}^2)$. In this step, w_{jit} will be calculated as

$$w_{jit} = p_{j-3,t} f(x_{ji}; \mu_{jt}, \sigma_{jt}^2) \Big/ \sum_{m=1}^{k_{j-3}} p_{j-3,m} f(x_{ji}; \mu_{jm}, \sigma_{jm}^2)$$
$$(j=4, 5;\ i=1, \ldots, n_j;\ t=1, \ldots, k_{j-3}).$$

In the IM-step, the iterated formulas are used to obtain the maximized point of $L(\theta | Y)$ for a specific genetic model by computing partial derivatives of $L(\theta | Y)$ for all parameters and letting the derivatives be zero. However, since there are still some constraints on the means, Lagrange multiplication (or λ-multiplicator method) can be used in the IM-step for those models. For example, for model D for JSA of MG5, the underlying relationships between the means of component distributions and genetic parameters (m_j, d, h) are as follows:

$$\begin{array}{lll} \mu_1 = m_1 + d & \mu_2 = m_2 + h & \mu_3 = m_3 - d \\ \mu_{41} = m_4 + d & \mu_{42} = m_4 + h & \mu_{43} = m_4 - d \\ \mu_{51} = m_5 + d & \mu_{52} = m_5 + \frac{1}{2}h & \mu_{53} = m_5 - d. \end{array}$$

https://doi.org/10.1017/S0016672303006141 Published online by Cambridge University Press

There are therefore two constraint conditions among the means of component distributions:

$$g_1 = \mu_{41} - \mu_{43} - \mu_{51} + \mu_{53} = 0$$
$$g_2 = 2\mu_{42} - 2\mu_{43} + \mu_{51} - 4\mu_{52} + 3\mu_{53} = 0.$$

Thus, $L_1 = L(\theta \mid Y) - \lambda_1 g_1 - \lambda_2 g_2$ is constructed. In the IM-step, if $\partial L_1/\partial\mu = 0$, the iterated formulae of means of component distributions is obtained,

$$\mu_j = \sum_{i=1}^{n_j} x_{ji}/n_j \quad (j = 1, 2, 3), \tag{4a}$$

$$\mu_{jt} = \left[\sum_{i=1}^{n_j} w_{jit} x_{ji} + c_{jt}\sigma_{jt}^2\right] \Big/ \sum_{i=1}^{n_j} w_{jit} \quad (j = 4, 5;\ t = 1, 2, 3), \tag{4b}$$

where $c_{41} = -\lambda_1$, $c_{42} = -2\lambda_2$, $c_{43} = \lambda_1 + 2\lambda_2$, $c_{51} = \lambda_1 - \lambda_2$, $c_{52} = 4\lambda_2$ and $c_{53} = -\lambda_1 - 3\lambda_2$. We have noticed that $\sigma_{52}^2 = \sigma_{51}^2 + (0{\cdot}5d^2 + 0{\cdot}25h^2)/n$ in model D, and both d and h were a function of means of component distributions. While the first partial derivative of σ_{52}^2 on the mean was omitted, the iterated formulae of means in (4a) and (4b) were obtained. This would be remedied by the iterated method as follows: (i) obtain λ_1 and λ_2 by the two constraint conditions while the values of the component parameters were given; (ii) calculate the means by (4a) and (4b); (iii) get the new estimates of σ_{5t}^2 by (2); (iv) replicate steps (i)–(iii) until estimates for λ_1 and λ_2 converge (the convergence of component parameters was confirmed by Monte Carlo simulation, data not shown). The items σ_{5t}^2 in (3) were included in the estimation of parameters by the iterated method to avoid the inexact results caused by omitting the items as in Wang & Gai (1998).

The polygenic variance is estimated by

$$\frac{\partial L_1}{\partial \sigma_{j0}^2} = \sum_{i=1}^{n_j} \sum_{t=1}^{k_{j-3}} w_{jit}\left[-\frac{(\sigma_{jt}^2)^{-1}}{2} + \frac{(x_{ji} - \mu_{jt})^2(\sigma_{jt}^2)^{-2}}{2}\right] = 0 \quad (j = 4, 5). \tag{5}$$

Taking $\sigma_4^2 = \sigma_{40}^2 + \sigma^2$, $\sigma_{5t}^2 = \sigma_{50}^2 + \sigma^2/n + V_{MGt}$ and $v_{2t} = \sigma_{51}^2/\sigma_{5t}^2\,(t = 1, \ldots, k_2)$, it is found that

$$\sigma_{40}^2 = \sum_{i=1}^{n_4} \sum_{t=1}^{k_1} w_{4it}(x_{4i} - \mu_{4t})^2/n_4 - \sigma^2, \tag{6a}$$

$$\sigma_{50}^2 = \sum_{t=1}^{k_2} v_{2t}^2 \sum_{i=1}^{n_5} w_{5it}(x_{5i} - \mu_{5t})^2 \Big/ \sum_{t=1}^{k_2} v_{2t} \sum_{i=1}^{n_5} w_{5it} - \sigma^2/n. \tag{6b}$$

Using the EM algorithm, the latter $k_2 - 1$ items in the numerator and denominator in the first item of the right-hand side of (6b) are omitted; the unequal σ_{5t}^2 included σ_{50}^2 in (5) for model E may result in a complicated equation in σ_{50}^2.

The environmental variance is estimated by setting $\partial L_1/\partial\sigma^2$ to zero:

$$\begin{aligned}\frac{\partial L_1}{\partial \sigma^2} = {} & \sum_{j=1}^{3}\sum_{i=1}^{n_j}\left[-\frac{(\sigma^2)^{-1}}{2} + \frac{(x_{ji} - \mu_j)^2(\sigma^2)^{-2}}{2}\right] \\ & + \sum_{i=1}^{n_4}\sum_{t=1}^{k_1} w_{4it}\left[-\frac{(\sigma_4^2)^{-1}}{2} + \frac{(x_{4i} - \mu_{4t})^2(\sigma_4^2)^{-2}}{2}\right] \\ & + \sum_{i=1}^{n_5}\sum_{t=1}^{k_2} w_{5it}\left[-\frac{(\sigma_{5t}^2)^{-1}}{2} + \frac{(x_{5i} - \mu_{5t})^2(\sigma_{5t}^2)^{-2}}{2}\right] \\ & \times \frac{1}{n} = 0.\end{aligned} \tag{7}$$

Let $v_4 = \sigma^2/\sigma_4^2$ and $v_{5t} = (\sigma^2/n)/\sigma_{5t}^2$ $(t = 1, \ldots, k_2)$, so that

$$\sigma^2 = \frac{\sum_{j=1}^{3}\sum_{i=1}^{n_j}(x_{ji} - \mu_j)^2 + \sum_{t=1}^{k_1} v_4^2 \sum_{i=1}^{n_4} w_{4it}(x_{4i} - \mu_{4t})^2 + n\sum_{t=1}^{k_2} v_{5t}^2 \sum_{i=1}^{n_5} w_{5it}(x_{5i} - \mu_{5t})^2}{\sum_{j=1}^{3} n_j + v_4 n_4 + \sum_{t=1}^{k_2} v_{5t}\sum_{i=1}^{n_5} w_{5it}}. \tag{8}$$

According to the above derivation, the procedure to obtain the maximum likelihood estimates of the parameters can be summarized as follows: (i) choose initial values for parameters according to the observations; for example, the means of P_1, F_1 and P_2 may be set equal to $\mu_1^{(0)} \sim \mu_3^{(0)}$, respectively, the pooled variance of P_1, F_1 and P_2 may be set equal to $\sigma^{2(0)}$; the mean $\bar{x}$ and its standard error $s_{\bar{x}}$ of F_2 are used to determine $\mu_{41}^{(0)} \sim \mu_{49}^{(0)}$ and $\sigma_4^{(0)}$ by $\mu_{4t}^{(0)} = \bar{x} + 0{\cdot}7(5 - t)s_{\bar{x}}\ \mathrm{sgn}(\mu_1^{(0)} - \mu_3^{(0)})$ $(t = 1 \sim 9)$ and $\sigma_4^{(0)} = n_4 s_{\bar{x}}^2/C$, $C \in (1{\cdot}5, 3)$; (ii) compute posterior probabilities w_{jit}, and obtain the expected complete data log-likelihood $L(\theta \mid Y)$ (E-step); (iii) compute the conditional maximum of $L(\theta \mid Y)$ and obtain $\mu^{(1)}$, $\sigma_{j0}^{2(1)}$ and $\sigma^{2(1)}$ (IM-step); (iv) replace initial values with estimates from step (iii) and then iterate steps (ii) and (iii) until a previously selected precision is achieved.

(vi) *Test of goodness of fit*

Let $F_0(x)$ be the expected distribution derived from the selected model. 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 $[y_i = F_0(x_i) = P(x < x_i)]$, n independent observations $y_i (i = 1, \ldots, n)$

Table 2. *The frequency distribution of number of insects in whole plant for P_1, F_1, P_2, F_2 and $F_{2:3}$ of the cross I*[a]

Generation	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9	Mean	Variance
P_1	3	4	5	8						1·90	1·2526
F_1	5	7	5	3						1·30	1·0632
P_2					6	6	5	2	1	5·30	1·3789
F_2	24	36	39	37	20	20	12	9	3	2·82	4·1182
$F_{2:3}$	16	68	15	0	5	18	5			2·26	3·1025

[a] Cross I is the soybean cross JNCWD × HJQDHY. The same is true for the later tables.

Table 3. *The AIC values under various genetic models for three soybean crosses*[a]

	AIC				AIC		
Model	I	II	III	Model	I	II	III
A-1	1281·33	1018·04	1387·99	D	1229·93	1029·63	1382·91
A-2	1354·69	1155·03	1494·18	D-1	1210·58	1005·40	1337·95
A-3	1290·20	1015·93	1390·72	D-2	1216·52	1042·45	1348·28
A-4	1512·00	1327·70	1673·06	D-3	1231·53	1048·04	1329·92
B-1	1249·77	1016·17	1371·20	D-4	1195·50	987·29	1357·30
B-2	1280·56	1022·22	1396·78	E	1225·94	1029·17	1367·36
B-3	1320·65	1103·41	1454·48	E-1	1255·81	1024·84	1382·97
B-4	1352·26	1127·94	1509·51	E-2	1276·81	1030·34	1398·99
B-5	1288·19	1017·40	1390·68	E-3	–	1183·22	1439·95
B-6	1266·65	1035·15	1416·96	E-4	1395·58	1072·44	1427·44
C	1376·73	1183·81	1552·55	E-5	1275·07	1021·86	1385·17
C-1	1396·94	1180·83	1571·92	E-6	1285·48	1067·24	1437·31

Minimum values are underlined.

[a] II is the soybean cross WXCQGJ × PXTED, III PXTED × 1138-2. The same is true for the later tables.

uniformly distributed on the interval (0, 1) can be obtained when H_0 holds, where the sample is derived from its population distribution $F(x)$. Consequently, three χ^2 statistics with 1 degree of freedom, namely $U_1^2=12[\sum y_i-n/2]^2/n$, $U_2^2=\frac{45}{4}(\sum y_i^2-n/3)^2/n$ and $U_3^2=180[\sum(y_i-1/2)^2-12/n]^2/n$, can be used to test whether the mean, second moment and variance of y_i are 1/2, 1/3 and 1/12, respectively under H_0. Moreover, Smirnov's statistics and Kolmogorov's statistics can also be used (Kendall & Stuart, 1979).

Table 4. *Maximum likelihood estimates of component parameters in model D-3 or D-4*

Parameter	Cross I (D-4)	Cross II (D-4)	Cross III (D-3)
μ_1	1·65	1·34	1·79
μ_2	1·46	1·32	1·58
μ_3	6·85	5·84	5·19
μ_{41}	1·97	1·47	5·21
μ_{42}	5·50	5·42	1·64
μ_{51}	1·26	1·33	5·47
μ_{52}	1·35	1·41	1·45
μ_{53}	5·11	5·50	1·29
σ^2	1·46	1·24	1·17
σ_{41}^2	1·90	1·24	1·60
σ_{51}^2	0·29	0·25	0·23

3. An example

To illustrate the application of the EIM algorithm, JSA of MG5 is used to re-analyse the inheritance of resistance to beanfly in soybean (*Melanagromyza sojae* Zehntner). Three soybean crosses – I: JNCWD (resistant: R) × HJQDHY (susceptible: S), II: WXCQGJ (R) × PXTED (S) and III: PXTED (S) × 1138-2 (R) – were made among five varieties (Wei *et al.*, 1989). A split plot design was used in the experiment, with crosses in main plots and parent hybrid generations in sub-plots. The number of insects (larvae plus pupae) in the stem (NIS) was used as an indicator of resistance. The frequency distribution of NIS for MG5 of cross I is shown in Table 2. It is obvious that the F_1 population tends toward the resistant parent. The distribution of NIS of the F_2 demonstrates a biased single mode toward the resistance parent while that of the $F_{2:3}$ lines demonstrates bi-modality, as do those of crosses II and III. This suggests that there is a major gene for resistance of soybean to the beanfly.

(i) *JSA of resistance of soybean to beanfly*

According to the procedures in Wang & Gai (1998), the maximized likelihood, Akaike's information

https://doi.org/10.1017/S0016672303006141 Published online by Cambridge University Press

Table 5. *Estimates of genetic parameters of resistance to beanfly of the three crosses*

First-order parameter	Estimates I	II	III	Second-order parameter	Estimates in F_2 I	II	III	Estimates in $F_{2:3}$ I	II	III
m	3·97	3·70	3·58	σ^2_P	3·47	3·56	4·12	3·02	3·94	3·10
d	−1·35	−1·52	1·45	σ^2_{mg}	1·37	1·73	1·58	1·30	1·65	1·50
h	1·35	1·52	1·45	σ^2_{pg}	0·65	0·60	1·37	1·43	2·05	1·37
$[d]$	−1·25	−0·73	−3·15	σ^2_e	1·46	1·24	1·17	0·29	0·25	0·23
$[h]$	−3·16	−3·00	−2·62	h^2_{mg} (%)	39	49	38	43	42	48
				h^2_{pg} (%)	19	17	33	47	52	44

Table 6. *The test of goodness of fit for model D-4 or D-3 of the three crosses*

Cross	Model	Generation	U^2_1	U^2_2	U^2_3	nW^2	D_n
I	D-4	P_1	0·00 (1·00)[a]	0·00 (0·99)	0·00 (0·98)	0·15 (>0·05)	0·20 (>0·05)
		F_1	1·03 (0·31)	1·21 (0·27)	0·22 (0·64)	0·24 (>0·05)	0·27 (>0·05)
		P_2	0·06 (0·81)	0·08 (0·77)	4·31 (0·04)*	0·21 (>0·05)	0·26 (>0·05)
		F_2	7·04 (0·01)**	8·22 (0·00)**	1·42 (0·23)	1·19 (<0·01)**	0·18 (>0·05)
		$F_{2:3}$	0·25 (0·62)	0·00 (0·95)	2·80 (0·09)	0·24 (>0·05)	0·14 (<0·05)*
II	D-4	P_1	0·00 (0·98)	0·03 (0·87)	0·54 (0·46)	0·16 (>0·05)	0·18 (>0·05)
		F_1	0·42 (0·52)	0·05 (0·83)	2·71 (0·10)	0·36 (>0·05)	0·34 (<0·05)*
		P_2	0·07 (0·79)	0·03 (0·86)	3·11 (0·08)	0·17 (>0·05)	0·20 (>0·05)
		F_2	2·42 (0·12)	2·83 (0·09)	0·50 (0·48)	0·95 (<0·05)*	0·23 (<0·05)*
		$F_{2:3}$	1·78 (0·18)	1·29 (0·26)	0·40 (0·53)	0·23 (>0·05)	0·12 (>0·05)
III	D-3	P_1	0·51 (0·48)	0·90 (0·34)	1·08 (0·30)	0·25 (>0·05)	0·27 (>0·05)
		F_1	1·51 (0·22)	1·17 (0·28)	0·19 (0·66)	0·29 (>0·05)	0·30 (>0·05)
		P_2	0·02 (0·89)	0·07 (0·79)	0·31 (0·58)	0·14 (>0·05)	0·17 (>0·05)
		F_2	3·55 (0·06)	3·61 (0·06)	0·10 (0·76)	0·78 (<0·05)*	0·16 (<0·05)*
		$F_{2:3}$	1·69 (0·19)	1·75 (0·19)	0·06 (0·80)	0·27 (>0·05)	0·11 (>0·05)

U^2_1, U^2_2, U^2_3: χ^2 statistics with 1 degree of freedom; nW^2: Smirnov's statistics; D_n: Kolmogorov's statistics.

[a] Values are the sample statistic and the corresponding p-value; *, **: the 0·05, 0·01 significance levels respectively.

criterion (AIC) and the maximum likelihood estimates in every model were calculated using the EIM algorithm. Here, $AIC = -2\,L(\theta\,|\,Y) + 2N$, where N is the number of independent parameters. The AIC values are listed in Table 3. From Table 3, model D-4 in crosses I and II, and model D-3 in cross III have the smallest AIC values and thus show the best fit. The difference is due to the fact that crosses I and II are R × S, and cross III is S × R. Therefore we can reasonably conclude that the resistance to beanfly is dominated by a mixture of dominant major gene plus additive-dominant polygenes.

The first-order and second-order genetic parameters in model D-4 or D-3, calculated from the results in Table 4, and the components in each segregating population, are given in Table 5. The additive effects of major genes in crosses I, II and III are estimated as –1·35, –1·52 and 1·45 heads/plant, respectively; the resistance trait is completely negative-dominant or dominant. The major-gene variations of the three crosses in F_2 are 38–49% of their total phenotypic variances, those in $F_{2:3}$ are 42–48% of the total variance. The polygenic variations of the three crosses in F_2 are 17–33% of their total phenotypic variances, those in $F_{2:3}$ are 44–52% of the total variance. Finally, the most probable major-gene genotype of an individual or a line in segregating populations can be determined by using the posterior probability w_{jit} (Wang & Gai, 2001).

4. Discussion

Thirteen genetic models, involving two groups – the two-major-gene models and the mixed two-major-gene plus polygenes models – were studied in this paper, which extended the results of Wang & Gai (1998, 2001). Therefore, more information was provided to infer whether there is one or two major gene(s) in the inheritance of a quantitative trait. Moreover, the efficiency of identification of our 'two major genes plus polygenes' inheritance model was also studied by the Monte Carlo simulation. The simulated genetic model was E-3 while $d_b = 0{\cdot}5d_a$, $h_a = b_b = 0$, and the heritabilities of the major gene and polygenes in the F_2 population were 0·4 and 0·2, respectively. The error variance may be set equal to one. Therefore, P_1, F_1, P_2,

F_2 and $F_{2:3}$ populations with sample sizes 20, 20, 20, 200 and 128 respectively are simulated by one, one, one, two and three normal distributions respectively. The replication is 100. According to the results of AIC and test of goodness of fit, the best model across 24 models for every replication was selected. The results showed that the power for identifying the two major genes above was 94%. However, sometimes the polygenes can not be identified, which is a problem that needs to be addressed. In addition, although JSA of quantitative traits does not provide the positions of QTLs, it can give useful information on quantitative traits at very low cost. Furthermore, the results of the JSA reciprocally confirm the results of QTL mapping.

Why did we not fit the model directly in terms of the Mather and Jinks parameterization? Of course, this would make it unnecessary to use Lagrange multipliers. However, the items involving genetic parameters *m*, *d*, etc., in the derivative of $L(\theta \mid Y)$ might be too complicated; as a result, it is difficult to obtain the expression of genetic parameters *m*, *d*, etc., as shown in formula (4).

The iterated formulas for estimating the means, polygenic variance and environmental variance of component distributions, such as (4), (6a), (6b) and (8) by the EIM algorithm, are obtained in this paper, and the iterated method is used in order to simplify the EM algorithm for JSA of MG5. When mixed two-major-gene plus polygenes inheritance models and more complicated models are extended, it is simple to estimate the component parameters. For the estimation of polygenic variance, the information of all components including $\sigma^2_{j0}(j=4,5)$ is used to improve the precision of parameter estimation.

Although the NIS did not theoretically obey a normal distribution, the test of goodness-of-fit between the expected values from the selected model and the observed values showed that hypothesis about the normal distribution almost held (Table 6). The results of the inheritance of resistance to beanfly in this paper are relatively consistent with those in Wang & Gai (2001) because of the similarly mixed 'one major gene plus polygenes' genetic model. But the former is more objective than the latter. Firstly, the former confirms that there is only one major gene. Secondly, the former reflects the difference that crosses I and II are R × S, and cross III is S × R by means of the best fitting genetic model. Then, the omitted two restrictions in model D in Wang & Gai (2001) are considered so that the relatively large maximum log-likelihood for model D in Wang & Gai (2001) was corrected. This can explain why the AIC values of model D for the three crosses in Wang & Gai (2001) are smaller than those in this paper and why the best-fitting genetic model in Wang & Gai (2001) is model D. Finally, the major gene heritability as well as polygene heritability estimates among the three crosses in this paper are more consistent than those in Wang & Gai (2001). Therefore, it is feasible to estimate the parameters using the EIM algorithm.

The authors wish to thank the editor and two anonymous referees for their comments and suggestions for the revision of our manuscript. The research was supported in part by the National 973 programme of China (no. G1998010206) and the National 863 programme of China (nos. 2002AA411052, 2001AA218061).

References

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

Elston, R. C. & Steward, J. (1973). The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. *Genetics* **73**, 695–711.

Gai, J. Y. & Wang, J. K. (1998). Identification and estimation of a QTL model and its effects. *Theoretical and Applied Genetics* **97**, 1162–1168.

Gai, J. Y., Zhang, Y. M. & Wang, J. K. (2003). *The Genetic System of Quantitative Traits in Plants*. Beijing: Science Press.

Kearsey, M. J. & Farquhar, A. G. L. (1998). QTL analysis in plants: where are we now? *Heredity* **80**, 137–142.

Kendall, M. G. & Stuart, A. (1979). *The Advanced Theory of Statistics, vol 2. Inference and Relationship*. Charles Griffin and Company Limited, London.

Mather, K. & Jinks, J. L. (1982). *Biometrical Genetics*, 2nd edn. London: Chapman and Hall.

Paterson, A. H. (1997). *Molecular Dissection of Complex Traits*. Boca Raton: CRC Press.

Tanner, M. A. (1993). *Tools for Statistical Inference*, 2nd edn. Berlin: Springer.

Wang, J. K. (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, J. K. & Gai, J. Y. (1998). Identification of major gene and polygene mixed inheritance model of quantitative traits by using joint analysis of P_1, F_1, P_2, F_2 and $F_{2:3}$. *Acta Agronomica Sinica* **24**, 651–659.

Wang, J. K. & Gai, J. Y. (2001). Mixed inheritance model for resistance to agromyzid beanfly (*Melanagromyza sojae* Zehntner) in soybean. *Euphytica* **122**, 9–18.

Wei, T., Gai, J. Y., Xia, J. K., *et al.* (1989). Inheritance of resistance to beanfly (*Melanagromyza sojae* Zehntner) in soybean. *Acta Genetica Sinica* **16**, 436–441.

Zhang, Y. M. & Gai, J. Y. (2000). Identification of mixed major genes and polygenes inheritance model of quantitative traits by using DH or RIL population. *Acta Genetica Sinica* **27**, 634–640.

Zhang, Y. M. (2001). A study on the improvement and expansion of segregation analysis in the inheritance of quantitative traits in plants. Doctoral dissertation, Department of Plant Breeding and Biometrics, Nanjing Agricultural University.

https://doi.org/10.1017/S0016672303006141 Published online by Cambridge University Press