(20) 358-366

生 物 数 学 学 报 2000,15(3) 358-366

Journal of Biomathematics

# 利用回交 $B_1$ 和 $B_2$ 及 $F_2$ 群体鉴定数量性状两对主基因 + 多基因混合遗传模型

Q348

Q-332

章元明[1] 盖钧镒[1] 王建康[2]

(1 南京农业大学大豆研究所, 农业部国家大豆改良中心, 江苏 南京 210095;
2 河南省农科院实验中心, 河南 郑州 450002)

**摘 要:** *QTL* 作图和主基因 + 多基因混合遗传分析表明: 拓展两对主基因 + 多基因混合遗传模型十分必要. 本文利用混合分布理论、AIC 准则在回交 $B_1$ 和 $B_2$ 群体或 $F_2$ 群体中鉴定两对主基因的存在, 当主基因存在时估计其遗传参数; 同时还改进了利用亲本、$F_1$ 和回交 $B_1$ 和 $B_2$ 群体, 或亲本、$F_1$ 和 $F_2$ 群体鉴定多基因存在的方法. 分布参数的估计采用 IECM 算法. 以水稻株高性状为例说明该方法的应用.

**关键词:** 数量性状; 主基因 + 多基因混合遗传; 混合模型; IECM 算法; 水稻株高

**中图分类号:** Q348 **文献标识码:** A

**文章编号:** 1001-9626(2000)03-0358-09

最早认为数量性状是由效应相等的微效多基因控制. *QTL* 作图结果表明其效应大小有别, 也发现了如水稻矮秆等数量性状还受主基因的控制. 由此盖钧镒等 (1999) 将主基因 + 多基因 (简称主 + 多基因) 混合遗传作为数量性状的通用性模型, 单纯主基因和单纯多基因作为其特例[1], 在 Elston 等 (1971,1973) 工作的基础上提出了适合植物的主 + 多基因混合遗传分析的试验设计和统计分析方法[2-5], 现已有初步应用成果, 例如提出水稻广亲和性、白叶枯病抗性和玉米矮花叶病抗性的遗传模式, 以及大豆抗食叶性害虫和孢囊线虫病的育种策略. 随着应用的深入, 该方法还将从以下三方面进行扩展: ①扩展不同类型的分离世代, 包括重组近交家系群体; ②扩展至有重复的家系世代, 包括多个家系世代的联合; ③将遗传模型扩展至 2 对主 + 多基因. 因此, 本文利用混合分布理论、AIC 准则和 IECM 算法在回交或 $F_2$ 世代中扩展数量性状 2 对主基因 + 多基因混合遗传模型, 还改进了鉴定多基因存在的方法[2,4]. 主 + 多基因混合遗传模型的数学模型和符号与文献 [2.4] 相同, 其基本假定参见文献 [8], 这里从略.

## 1 回交 $B_1$ 和 $B_2$ 及 $F_2$ 群体的遗传组成

当存在两对主基因时, 假定亲本主基因型分别为 $AABB(P_1)$ 和 $aabb(P_2)$, $B_1 = F_1 \times P_1$, $B_2 =$

**收稿日期:** 1999-04-12

**基金项目:** 国家自然科学基金项目; 国家 973 项目和重庆市科委应用基础研究项目资助

**作者简介:** 章元明 (1965-), 男, 重庆永川人, 南京农业大学大豆研究所农业部国家大豆改良中心副教授, 硕士.

$F_1\times P_2$.$B_1$ 群体为 $AABB, AABb, AaBB$ 和 $AaBb$ 4 种主基因型等比例混合，分别用 $N(\mu_{41},\sigma_4^2)\sim N(\mu_{44},\sigma_4^2)$ 表示这 4 种基因型个体观测值的分布；$B_2$ 群体为 $AaBb, Aabb, aaBb$ 和 $aabb$ 4 种主基因型等比例混合，分别用 $N(\mu_{51},\sigma_5^2)\sim N(\mu_{54},\sigma_5^2)$ 表示这 4 种基因型个体观测值的分布；$F_2$ 群体为 $AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb$ 和 $aabb$ 9 种主基因型按孟德尔分离比的混合，分别用 $N(\mu_{61},\sigma_6^2)\sim N(\mu_{69},\sigma_6^2)$ 表示这 9 种基因型个体观测值的分布．在两个主基因服从等加性、完全显性和等显性下，$B_1, B_2$ 和 $F_2$ 的成分分布数分别为：3(1 : 2 : 1), 3(1 · 2 : 1) 和 5(1　4 . 6 : 4 : 1); 1. 4(1 . 1 · 1 · 1) 和 4(9　3　3 . 1). 1、3(1 : 2 : 1) 和 3(9 . 6 : 1)．若主基因符合加性－显性－上位性模型，$B_1$、$B_2$ 和 $F_2$ 群体的主基因遗传方差分别为

$$\sigma^2_{mg(B_1)}=\frac{1}{4}[(d_a-h_a)^2+(d_b-h_b)^2+(d_a-h_a)(i+j_{ab}-j_{ba}-l)+(d_b-h_b)(i-j_{ab}+j_{ba}-l)]$$
$$+\frac{1}{16}[(i-j_{ab})^2+(i-j_{ba})^2+(i-l)^2+(j_{ab}-j_{ba})^2+(j_{ab}-l)^2+(j_{ba}-l)^2], \tag{1a}$$

$$\sigma^2_{mg(B_2)}=\frac{1}{4}[(d_a+h_a)^2+(d_b+h_b)^2+(d_a+h_a)(-i+j_{ab}-j_{ba}+l)+(d_b+h_b)(-i-j_{ab}+j_{ba}+l)]$$
$$+\frac{1}{16}[(i+j_{ab})^2+(i+j_{ba})^2+(i-l)^2+(j_{ab}-j_{ba})^2+(j_{ab}+l)^2+(j_{ba}+l)^2], \tag{1b}$$

$$\sigma^2_{mg(F_2)}=\frac{1}{4}\left[d_a^2+d_b^2+i^2+(d_a+j_{ab})^2+(d_b+j_{ba})^2+\left(h_a+\frac{l}{2}\right)^2+\left(h_b+\frac{l}{2}\right)^2+\frac{l^2}{4}\right]. \tag{1c}$$

其中，$d_a$ 和 $d_b$ 为主基因基因型 $AA$ 和 $BB$ 的加性效应，$h_a$ 和 $h_b$ 为主基因基因型 $Aa$ 和 $Bb$ 的显性效应，$i, j_{ab}, j_{ba}$ 和 $l$ 分别为主基因加性 $(AA)\times$ 加性 $(BB)$、加性 $(AA)\times$ 显性 $(Bb)$、显性 $(Aa)\times$ 加性 $(BB)$ 和显性 $(Aa)\times$ 显性 $(Bb)$ 的互作效应．主基因遗传率 $h^2_{mg}$ 为主基因遗传方差 $\sigma^2_{mg}$ 与相应群体的表型方差 $\sigma^2_P$ 之比．　$F_2$ 群体成分分布平均数与遗传参数间的关系分别为 [5]

$$\mu_{61}=m_6+C_1,\quad \mu_{62}=m_6+C_2,\quad \mu_{63}=m_6+C_3,$$
$$\mu_{64}=m_6+C_4,\quad \mu_{65}=m_6+C_5,\quad \mu_{66}=m_6+C_6,$$
$$\mu_{67}=m_6+C_7,\quad \mu_{68}=m_6+C_8,\quad \mu_{69}=m_6+C_9,$$

其中 $C_1\sim C_9$ 分别是 $d_a+d_b+i, d_a+h_b+j_{ab}, d_a-d_b-i, h_a+d_b+j_{ba}, h_a+h_b+l, h_a-d_b-j_{ba}, -d_a+d_b-i, -d_a+h_b-j_{ab}$ 和 $-d_a-d_b+i$．显然 $\mu_{41}\sim\mu_{44}$ 分别为 $m_4+C_1, m_4+C_2, m_4+C_4, m_4+C_5$; $\mu_{51}\sim\mu_{54}$ 分别为 $m_5+C_5, m_5+C_6, m_5+C_8, m_5+C_9$，其中 $m_4, m_5$ 和 $m_6$ 分别为双亲平均数 $m$ 与各世代多基因效应之和．此外，$B_1, B_2$ 和 $F_2$ 群体成分分布方差 $(\sigma_4^2,\sigma_5^2$和$\sigma_6^2)$ 分别剖分为多基因方差 $(\sigma^2_{40},\sigma^2_{50}$和$\sigma^2_{60})$ 与环境方差 $\sigma_e^2$ 两组分之和，即 $\sigma_j^2=\sigma^2_{j0}+\sigma_e^2$ $(j=4,5,6)$.

## 2　利用回交 $B_1$ 和 $B_2$ 群体及 $F_2$ 群体鉴定两对主基因的存在

### 2.1　遗传模型的建立

根据 §1 建立利用回交 $B_1$ 和 $B_2$ 或 $F_2$ 群体鉴定两对主基因的遗传模型见表 1，无主基因的 $A-0$ 模型以及一对主基因的加性－显性、加性、完全显性和负向完全显性的 $A-1\sim A-4$ 模型参见文献 [1, 4]　若两对主基因符合加性－显性－上位性模型，利用 $B_1$ 和 $B_2$ 群体鉴定两

对主基因存在的样本似然函数中的一阶分布参数个数少于相应一阶遗传参数个数，无法估计一阶遗传参数，因此这种情况在这里未加以讨论，只能在多世代的联合分析中考虑.

### 2.2 利用回交或 $F_2$ 群体鉴定两对主基因的似然函数和分布参数的极大似然估计

利用回交 $B_1$ 和 $B_2$ 或 $F_2$ 群体可将主基因的效应分解出来. 记 $x_{4i}, x_{5i}$ 和 $x_{6i}$，以及 $n_4, n_5$ 和 $n_6$ 分别为 $B_1, B_2$ 和 $F_2$ 群体的观测值和样本容量. 由 §1 的假定，$B_1$ 群体为 $k_1$ 个 $N(\mu_{4j}, \sigma_4^2)$ 的混合；$B_2$ 群体为 $k_2$ 个 $N(\mu_{5j}, \sigma_5^2)$ 的混合；$F_2$ 群体为 $k_3$ 个 $N(\mu_{6j}, \sigma_6^2)$ 的混合. 利用 $B_1$ 和 $B_2$ 群体群体鉴定主基因存在的样本似然函数为

$$L_1(Y|\theta) = \prod_{i=1}^{n_4}\sum_{t=1}^{k_1}\pi_{4t}f(x_{4i};\mu_{4t},\sigma_4^2)\prod_{i=1}^{n_5}\sum_{t=1}^{k_2}\pi_{5t}f(x_{5i};\mu_{5t},\sigma_5^2), \tag{2}$$

其中，$f(x_{6i};\mu_{6t},\sigma_6^2)$ 是 $N(\mu_{6t},\sigma_6^2)$ 的密度函数. 利用 $F_2$ 群体的样本似然函数参见文献 [4]

表 1 两对主基因遗传模型下 $B_1$ 和 $B_2$ 或 $F_2$ 群体所包含的成分分布数及相应的遗传参数

Table 1 The Number of Component Distributions and Corresponding Genetic Parameters of $B_1$ and $B_2$ or $F_2$ under Two Major Genes Inheritance Models

| 群体 | 模型 | 成分分布个数 | 独立参数个数 | 一阶遗传参数 | 二阶分布参数 | 约束条件个数 |
|---|---|---|---|---|---|---|
| | $B-1$ | 8 | 12 | $m_4, m_5, d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l$ | $\sigma_4^2, \sigma_5^2$ | — |
| | $B-2$ | 8 | 8 | $m_4, m_5, d_a, d_b, h_a, h_b$ | $\sigma_4^2, \sigma_5^2$ | 2 |
| $B_1$ | $B-3$ | 8 | 6 | $m_4, m_5, d_a, d_b (h_a = h_b = 0)$ | $\sigma_4^2, \sigma_5^2$ | 4 |
| & | $B-4$ | 6 | 5 | $m_4, m_5, d (= d_a = d_b, h_a = h_b = 0)$ | $\sigma_4^2, \sigma_5^2$ | 3 |
| $B_2$ | $B-5$ | 5 | 6 | $m_4, m_5, d_a (= h_a), d_b (= h_b)$ | $\sigma_4^2, \sigma_5^2$ | 1 |
| | $B-6$ | 4 | 5 | $m_4, m_5, d (= d_a = d_b = h_b = h_b)$ | $\sigma_4^2, \sigma_5^2$ | 1 |
| | $B-1$ | 9 | 10 | $m_6, d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l$ | $\sigma_6^2$ | — |
| | $B-2$ | 9 | 6 | $m_6, d_a, d_b, h_a, h_b$ | $\sigma_6^2$ | 4 |
| $F_2$ | $B-3$ | 9 | 4 | $m_6, d_a, d_b (h_a = h_b = 0)$ | $\sigma_6^2$ | 6 |
| | $B-4$ | 5 | 3 | $m_6, d (= d_a = d_b, h_a = h_b = 0)$ | $\sigma_6^2$ | 3 |
| | $B-5$ | 4 | 4 | $m_6, d_a (= h_a), d_b (= h_b)$ | $\sigma_6^2$ | 1 |
| | $B-6$ | 3 | 3 | $m_6, d (= d_a = d_t = h_a = h_b)$ | $\sigma_6^2$ | 1 |

采用 IECM 算法 [16] 获得分布参数的极大似然估计值. IECM 算法按下列两步骤进行. 对 $F_2$ 群体，$E$ 步骤：求在给定 $Y_{obs}$ 下 $Y_{miss}$ 条件分布的完全资料对数似然函数的条件期望值 $Q(\theta|\theta^{(t)})$，

$$Q(\theta|\theta^{(t)}) = E(L_C(Y|\theta)|X,\theta^{(0)}) = \sum_{i=1}^{k_3}\sum_{j=1}^{n_6} w_{6ji}^{(0)}[\ln\pi_{6i} + f(x_{6j};\mu_{6i},\sigma_6^2)]. \tag{3}$$

其中，$L_C(Y|\theta)$ 是对数似然函数，$w_{6ji}$ 是 $x_{6j}$ 归属第 $i$ 种主基因型的后验概率. 迭代 CM 步骤：分步骤极大化 $Q(\theta|\theta^{(t)})$，并用极大值点处的 $\theta$ 值代替 $\theta^{(t-1)}$ 作为下一轮循环的初值. $Q(\theta|\theta^{(t)})$

的极大值点由下列公式确定

$$\pi_{6i}^{(t)} = \sum_{j=1}^{n_6} \frac{w_{6ji}^{(t-1)}}{n_6}, \qquad i = 1, \cdots, k_3, \tag{4a}$$

$$\partial[\ln L_C(Y|\theta) - \sum_{m=1}^{k} \lambda_m g_m]/\partial\theta\theta = 0, \tag{4b}$$

其中, $g_m$ 是分布平均数间的第 $m$ 个约束条件, $\lambda_m$ 是 Lagrange 常数, $k$ 是约束条件个数.

首先, 固定多基因方差组分和环境方差组分, 求分布平均数的条件极大似然估计: ①由约束条件方程和平均数迭代公式得到的联立方程组求 $\lambda_m$ ; ②由分布平均数迭代公式得到其极大条件似然估计值; 然后, 固定已得到的分布平均数, 求多基因方差组分 $\sigma_{60}^2$ 与环境方差组分 $\sigma_e^2$ 之和的条件极大似然估计

$$\sigma_{60}^2 + \sigma_e^2 = \sum_{i=1}^{k_3} \sum_{j=1}^{n_6} w_{6ji}(x_{6j} - \mu_{6i})^2/n_6, \tag{5}$$

对似然函数 (2) 中分布参数的估计, 可仿此进行.

### 2.3　遗传模型的 AIC 判定和适合性检验

采用文献 [8] 的方法, 通过比较 0 、 1 和 2 对主基因的 $A-0, A-1 \sim A-4, B-1 \sim B-6$ 共 11 个遗传模型的 AIC 值以选择最优遗传模型, 并且进行遗传模型的均匀性检验 ($U_1^2, U_2^2$和$U_3^2$统计量) 、 Smirnov 检验 ($nW^2$ 统计量) 和 Kolmogorov 检验 ($D_n$ 统计量) 以检验其适合性, 其基本公式参见文献 [8]. 在选择遗传模型时, 要综合考虑极大对数似然值、 AIC 值和适合性检验结果.

### 2.4　遗传参数的估计

确定了数量性状最适遗传模型后便由分布参数的估计值估计相应遗传参数. §1 已经给出了 $F_2$ 群体各成分分布平均数与一阶遗传参数的相互关系, 由此获得分布平均数向量 $\theta(=(\mu_{61}, \cdots, \mu_{6k_3})^T)$ 与一阶遗传参数向量 $G(=(m_6, d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, 1)^T)$ 的关系: $\theta = AG$ , 其中, $A$ 是系数矩阵. 由最小二乘法原理可得到一阶遗传参数的最小二乘估计

$$\hat{G} = (A^T A)^{-1} A^T \hat{\theta}.$$

同理, 容易获得 $B_1$ 和 $B_2$ 群体各成分分布平均数与一阶遗传参数间的关系, 用最小二乘法估计一阶遗传参数. 若主基因符合加性 - 显性模型, 利用 $B_1$ 和 $B_2$ 得到的一阶遗传参数为

$$\hat{m}_4 = \frac{1}{4}(\mu_{41} + \mu_{42} + \mu_{43} + \mu_{44} - 2\mu_{51} + 2\mu_{54}),$$

$$m_5 = \frac{1}{4}(2\mu_{41} - 2\mu_{44} + \mu_{51} + \mu_{52} + \mu_{53} + \mu_{54}),$$

$$\hat{d}_a = \frac{1}{4}(\mu_{41} + \mu_{42} - \mu_{43} - \mu_{44} + \mu_{51} + \mu_{52} - \mu_{53} - \mu_{54}).$$

$$\hat{d}_b = \frac{1}{4}(\mu_{41} - \mu_{42} + \mu_{43} - \mu_{44} + \mu_{51} - \mu_{52} + \mu_{53} - \mu_{54}),$$

$$\hat{h}_a = \frac{1}{4}(-\mu_{41} - \mu_{42} + \mu_{43} + \mu_{44} + \mu_{51} + \mu_{52} - \mu_{53} - \mu_{54}).$$

$$\hat{h}_b = \frac{1}{4}(-\mu_{41} + \mu_{42} - \mu_{43} + \mu_{44} + \mu_{51} - \mu_{52} + \mu_{53} - \mu_{54})$$

同理可获得其它模型参数的极大似然估计. 关于二阶遗传参数的估计结合 §1 和 §2.1 有关内容获得. 关于分离世代个体的主基因型归类可参考文献 [2,4,8]

## 3 多基因存在的鉴定

从 §1 可知, $\sigma_4^2, \sigma_5^2$ 和 $\sigma_6^2$ 可剖分为多基因方差组分和环境方差组分两部分. 在只有 $B_1$ 和 $B_2$ 群体或只有 $F_2$ 群体的情况下, 不能将多基因方差组分和环境方差组分分开. 因此, 在多基因鉴定时, 尚需利用亲本和 $F_1$ 群体以获得环境方差的估计, 从而鉴定多基因的存在. 记 $x_{1i}, x_{2i}$ 和 $x_{3i}$, 以及 $n_1, n_2$和$n_3$ 分别为 $P_1, F_1, P_2$ 群体观测值和样本容量. 由假定可知: $x_{ji} \sim N(\mu_j, \sigma_e^2)$ $(j = 1,2,3)$. 因此, 由 $B_1$ 和 $B_2$(或 $F_2$) 群体鉴定多基因存在的样本似然函数分别为 [2, 4]

$$L_2(Y|\theta) = \prod_{i=1}^{n_1} f(x_{1i}, \mu_1, \sigma_e^2) \prod_{i=1}^{n_2} f(x_{2i}; \mu_2, \sigma_e^2) \prod_{i=1}^{n_3} f(x_{3i}, \mu_3, \sigma_e^2)$$
$$\prod_{i=1}^{n_4} \sum_{t=1}^{k_1} \pi_{4t} f(x_{4i}; \mu_{4t}, \sigma_4^2) \prod_{i=1}^{n_5} \sum_{t=1}^{k_2} \pi_{5t} f(x_{5i}; \mu_{5t}, \sigma_5^2), \tag{6a}$$

$$L_3(Y|\theta) = \prod_{i=1}^{n_1} f(x_{1i}; \mu_1, \sigma_e^2) \prod_{i=1}^{n_2} f(x_{2i}, \mu_2, \sigma_e^2) \prod_{i=1}^{n_3} f(x_{3i}; \mu_3, \sigma_e^2)$$
$$\prod_{i=1}^{n_6} \sum_{t=1}^{k_3} \pi_{6t} f(x_{6i}; \mu_{6t}, \sigma_6^2). \tag{6b}$$

通过构造 $H_0$ : 不存在多基因 ($\sigma_4^2 = \sigma_e^2$ 和 $\sigma_5^2 = \sigma_e^2$ 或 $\sigma_6^2 = \sigma_e^2$, 多基因效应的平均数为 0) ; $H_a$ : 存在多基因 ($\sigma_{40}^2 > 0$ 和 $\sigma_{50}^2 > 0$ 或 $\sigma_{60}^2 > 0$, 多基因效应的平均数不为 0). 计算两种假设下的最大似然函数值 $L_0$ 和 $L_a$, 构造出似然比统计量 $\lambda = 2(\ln L_a - \ln L_0) \sim x_{df}^2$ 以鉴定多基因是否存在, 其中, $df$ 为两种假设下相差的遗传参数个数. 采用 IECM 算法获得分布参数的极大似然估计值, $E$ 步骤和迭代 $CM$ 步骤的 $CM_1$ 与 §2.2 相似, 下面给出利用 $P_1, F_1, P_2$ 和 $F_2$ 鉴定多基因存在的迭代 $CM_2$ 和迭代 $CM_3$ 步骤. 迭代 $CM_2$ 是固定分布平均数与误差方差条件下用迭代公式求多基因方差组分, 迭代 $CM_3$ 是固定分布平均数和多基因方差组分条件下用迭代公式求误差方差. $\sigma_{60}^2$ 和 $\sigma_e^2$ 的迭代公式如下

$$\sigma_{60}^2 = \sum_{i=1}^{k_3} \sum_{j=1}^{n_6} w_{6ji}(x_{6j} - \mu_{6i})^2 / n_6 - \sigma_e^2, \tag{7}$$

$$\sigma_e^2 = [\sum_{i=1}^{3} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 + \sum_{i=1}^{k_3} v_i^2 \sum_{j=1}^{n_6} w_{6ji}(x_{6j} - \mu_{6i})^2] / [\sum_{i=1}^{3} n_i + \sum_{i=1}^{k_3} v_i \sum_{j=1}^{n_6} w_{6ji}], \tag{8}$$

其中，$v_i = \sigma_e^2/(\sigma_e^2 + \sigma_{60}^2)$. 利用 $B_1$ 和 $B_2$ 群体鉴定多基因存在的算法可仿此进行.

本文作者应用 Turbo C++ 语言编制了以上全套IECM 算法的软件 F2.EXE 和 F2P.EXE(利用 $F_2$ 群体鉴定主基因和利用亲本、$F_1$ 与 $F_2$ 鉴定多基因的存在)、B.EXE(利用 $B_1$ 和 $B_2$ 群体鉴定主基因的存在) 和 BP EXE(利用亲本、$F_1$、$B_1$ 和 $B_2$ 群体鉴定多基因的存在)，由此计算复杂的工作变得十分简单，读者如有需要可与作者联系.

## 4　应用举例

根据水稻南京 6 号 × 广丛杂交组合 $B_1, B_2$ 群体[1,8] 鉴定主基因存在的结果见表 2. 由表 2 可知，$A-1$ 和 $B-2$ 模型的 AIC 值相对较小，$A-1$ 模型为最优遗传模型，$B-2$ 模型为次优模型. $A-1$ 模型鉴定出符合加性 - 显性的具有最大遗传效应的一对主基因的存在，$B-2$ 模型鉴定出符合加性 - 显性的具有最大和次大遗传效应的两对主基因的存在、$A-1$ 模型分布参数的遗传参数的极大似然估计值分别为：$m_4 = 124.61, m_5 = 132.31, d = 33.99, h = 22.40$, $B_1$ 群体的 2 成分分布比例不呈 1 . 1 的分离比 ($\chi^2 = 8.18, P = 0.0042$), $B_2$ 群体的 2 成分分布比例呈 1 : 1 的分离比 ($\chi^2 = 0.18, P = 0.67$). 若株高符合 $B-2$ 模型，其遗传参数估计值分别为：$m_4 = 124.59, m_5 = 132.80, d_a = 33.92. d_b = 0\ 72. h_a = 22.12. h_b = -0.19$.

在 $A-1$ 模型基础上进行多基因存在的鉴定结果为：$\lambda = 29.70$ ($P = 5\ 6 \times 10^{-6}$)，说明多基因存在，其遗传参数估计为：$m = 130.0670, d = 34\ 7233, h = 21.7356, [d] = -5.440, [h] = -6.5717, \sigma_e^2 = 19.2493, \sigma_{40}^2 = 10.5045, \sigma_{50}^2 \approx 0.0$，从而 $B_1, B_2$ 群体多基因遗传力分别为 19 78% 和 0 0. 在 $B-2$ 模型基础上进行多基因存在也有类似的结果.

利用 $F_2$ 群体进行主基因存在鉴定的结果亦列于表 2. 成分分布数为 4 的 2 对完全显性主基因遗传模型的 AIC 值最小，说明株高由表现为完全显性的 2 对主基因所控制. 4 个成分分布平均数的极大似然估计值分别为 158.48, 141 83, 111 40 和 94.76，其方差均为 65.36. 由此可得，$m_6 = 126\ 62$，两主基因的加性效应 $d_a = 23.54$ 和 $d_b = 8.32$，主基因遗传方差 $\sigma_{mg}^2 = 467.4078$，主基因遗传率 $h_{mg}^2 = 82.89\%$ 4 个成分分布所占比例并不呈 9 . 3 · 3 : 1 ($\lambda = 56.19, P$接近0)，这可能是由于第 1 与 2 成分分布平均数之间和第 3 与 4 成分分布平均数之间较接近，使相应的观测值归组模糊，造成不呈分离比例. 在 $B-5$ 模型的基础上进行多基因存在的鉴定结果为：$\lambda = 9\ 29$ ($P = 0.0023$)，说明多基因存在. 若对 $F_2$ 群体的表型方差进行剖分，则多基因的遗传方差 $\sigma_{pg}^2 = 44.9244$，多基因遗传率 $h_{pg}^2 = 7.97\%$. 以上述 $F_2$ 群体分离分析结果的 4 个成分分布按 9 . 3 : 3 : 1 的分离比例混合构成样本容量为 512 的 $F_2$ 模拟群体，按 $E-5$ 模型进行参数估计. 重复 500 次，得到分布参数和遗传参数的平均数与标准误 (见表 3). 从表 3 可知，①除成分分布比例外，分布参数和遗传参数的模拟平均值与真值均较接近；②遗传参数的变异大于分布参数的变异；③二阶参数的变异大于一阶参数的变异；④成分分布混合比例在 500 次中只有 53.2% 是符合 9 · 3 : 3 : 1 的分离比例的，这可能是由于第 1、2 成分分布平均数之间和第 3、4 成分分布的平均数之间较接近所造成的.

## 5　讨　论

比较本文中利用 $F_2$ 群体与利用 $B_1$ 和 $B_2$ 群体进行主基因鉴定的结果，发现两者的结果是

相对一致的，并且与利用亲本、 $F_1$、$B_1$、$B_2$ 和 $F_2$ 6 个世代的联合分析结果 (另文报道) 也是一致的. 通过 $F_2$ 群体及其模拟研究发现，若存在两对主基因时，各成分分布比例往往不符合孟德尔分离比例，这可能是由于成分分布数较多，其平均数间差异相对于标准差而言是较小的，致使观测值归组模糊所致.

表 2 利用 $B_1$ 和 $B_2$ 群体或 $F_2$ 群体鉴定主基因存在的极大对数似然函数值 (MLV) 和 AIC 值

Table 2 Maximum Log-Likelihood Value (MLV) and AIC Value (AICV) of the Genetic Models for Identifying the Existence of Major Gene from $B_1$ and $B_2$ or $F_2$

| 模型 | $B_1$ and $B_2$ | | | $F_2$ | | |
|---|---|---|---|---|---|---|
| | MLV | AIC | AIC 排序 | MLV | AIC | AIC 排序 |
| $A-0$ | −709.020 | 1422 041 | 5 | −2293 170 | 4590.340 | 7 |
| $A-1$ | −638.246 | 1288 491 | 1 | −2124.873 | 4261 747 | 4 |
| $A-2$ | −706.261 | 1422 522 | 6 | −2293.171 | 4596.341 | 9 |
| $A-3$ | −901.720 | 1815.439 | 10 | −2124 873 | 4259.746 | 3 |
| $A-4$ | −706 285 | 1424 570 | 7 | −2293.172 | 4594.345 | 8 |
| $B-1$ | – | — | — | −2117 561 | 4271.121 | 5 |
| $B-2$ | −638.246 | 1292.492 | 2 | −2124.872 | 4277 743 | 6 |
| $B-3$ | −701 119 | 1414 237 | 3 | −2293.170 | 4610.339 | 11 |
| $B-4$ | −702.955 | 1415 911 | 4 | −2293.169 | 4600.338 | 10 |
| $B-5$ | −901.636 | 1815.273 | 9 | −2115 440 | 4244.880 | 1 |
| $B-6$ | −901.720 | 1813.439 | 8 | −2121.627 | 4253 254 | 2 |

表 3 蒙特卡罗模拟结果 (重复 500 次)

Table 3 The Results of Monte Carlo Simulation (500 Replications)

| 参数 | 平均数 | 标准误 | 变异系数 (%) | 变异幅度 |
|---|---|---|---|---|
| $\mu_1$ | 158 5553 | 0.0441 | 0.62 | 155.4102~161.3672 |
| $\mu_2$ | 141.3835 | 0.0891 | 1 41 | 125.2902~146.0431 |
| $\mu_3$ | 111 3964 | 0.0815 | 1 64 | 105.6792~116.7391 |
| $\mu_4$ | 94 2246 | 0.1227 | 2.91 | 75.9693~ 102.0431 |
| $\sigma_e^2$ | 64.2063 | 0.3800 | 13.24 | 46.8203~100.0028 |
| $m$ | 126 3900 | 0.0700 | 1 24 | 115.6953~130.5042 |
| $d_a$ | 23 5794 | 0.0426 | 4.04 | 20.5721~26.1612 |
| $d_b$ | 8 5859 | 0 0344 | 8 95 | 6.8476~15.0656 |
| $\sigma_{mg}^2$ | 473.4012 | 1.6684 | 7.88 | 365.5180~626.3334 |
| $h_{mg}^2$ | 0.8863 | 0.0029 | 7.31 | 0.6813~1 0000 |

在鉴定主基因存在时，王建康等 [4](1997) 和盖钧镒等 [2](1998) 是首先确定成分分布数然后确定遗传模型的，本文是通过 AIC 准则比较具有一定遗传背景遗传模型的 AIC 值来确定遗传模型以及相应分离群体的成分分布数. 这就避免了具有相同成分分布数时不能区分是一对主基因控制的还是两对主基因控制的缺陷. 例如，用文献 [2,4] 的方法， $F_2$ 群体由 2 个成分分布

构成，株高由 1 对主基因控制，从本文的分离分析结果看，这 2 成分分布中的每一成分分布都是 2 个成分分布的叠加. 但是，文献 [2,4] 的先确定成分分布数再确定遗传模型的方法能方便地确定分布谷值. 在鉴定多基因存在时，王建康等 (1997)[4] 和盖钧镒等 (1998)[1] 在设置 $H_0$ 和 $H_a$ 时仅考虑了多基因是否存在的二阶参数而忽略了多基因存在与否的一阶参数，本文在这一问题上进行了补充，以全面考虑了多基因是否存在. 在多基因存在的似然比检验中，其自由度并不是都为 1，而是 $H_0$ 和 $H_a$ 相差的遗传参数个数.

本文是通过利用个别分离世代先确定主基因的对数和作用方式，其模型符号暂用 $A$ 和 $B$ 表示；然后通过利用亲本、$F_1$ 和个别分离世代的联合鉴定多基因存在；以最终确定模型的类型. 例如若第 1 步为 $A-1$ 模型，第 2 步又鉴定出有多基因作用，则最终模型应该是 $D-1$.

在估计样本似然函数中的分布参数时，一般是采用 EM 算法 [2−4,10−12,15]. 在多基因鉴定时，若采用 EM 算法就会出现估计环境方差时不能利用 $F_2$ 或 $B_1$ 和 $B_2$ 群体成分分布的信息，即在参数估计时没有把二阶分布参数的数量遗传学意义的关系体现出来；本文采用的 IECM 算法从参数估计角度体现了分布参数之间的数量遗传学意义，也提高了估计精度.

## [参 考 文 献]


[1] 盖钧镒、管荣展，王建康. 植物数量性状 $QTL$ 体系检测的遗传试验方法 [J]. 世界科技研究与发展，1999，21(1)：34–40.

[2] 盖钧镒，王建康. 利用回交世代或 $F_{2\cdot3}$ 家系世代鉴定数量性状主基因 – 多基因混合遗传模型 [J]. 作物学报. 1998，24(4)：402–409.

[3] 王建康、盖钧镒. 混合模型的理论及其应用 [J]. 生物数学学报，1995，10(4)：87–92.

[4] 王建康，盖钧镒. 利用杂种 $F_2$ 世代鉴定数量性状主基因 – 多基因混合遗传模型并估计其遗传效应 [J]. 遗传学报，1997，24(5)：432–440.

[5] 马育华. 植物育种的数量遗传学基础 [M]. 南京：江苏科技出版社，1982　120–157

[6] *Elston R C, Stewart J. The analysis of quantitative traits for simple genetic models from parental, $F_1$ and backcross data* [J]. *Genetics*, 1973, **73**(4)：695–711

[7] Fernando R L. The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance [J]. *Theor Appl Genet*, 1994, **88**(5)：573–580.

[8] Gai J Y, Wang J K. Identification and estimation of a *QTL* model and its effects [J]. *Theor Appl Genet*, 1998, **97**(7)：1162–1168

[9] Hoeschele I. Statistical techniques for detection of major genes in animal breeding data[J]. *Theor Appl Genet*, 1988, **76**(3)：311–319.

[10] Knott S A, Haley C S, Thanpson R. Methods of segregation analysis for animal breeding data. a comparison of power [J]. *Heredity*, 1992, **68A**(4)：299–311.

[11] Knott S A, Haley C S, Thanpson R. Methods of segregation analysis for animal breeding data: parameter estimates [J]. *Heredity*, 1992, **68B**(4)：313–320.

[12] Loisel P, Goffinet B, Monod H, *et al*. Dectmg a major gene in an $F_2$ population [J]. *Biometrics*, 1994, 50(2)：512–516

[13] Meng X L, Rubin D B. Maximum likelihood estimation via the ECM algorithm: a general framework [J]. *Biometrika*, 1993, **80**(2)：267–278.

[14] Paterson A H, Lander E S, Hewitt J D, *et al*. Resolution of quantitative traits into mendelian factors by using a complete RFLP linkage map [J] *Nature*, 1988, **335**(20)：721–726.

[15] Shoukri M M, McLachlan G J. Parameter estimation in a genetic mixture model with application to nuclear family data[J]. *Biometrics*, 1994, 50(1):128–139.

[16] 章元明, 盖钧镒. 数量性状主基因 + 多基因混合遗传分析中鉴定多基因存在的 IECM 算法 [J] 生物数学学报, 1999, 14(4):429–434.

# Identification of Two Major Genes Plus Polygenes Mixed Inheritance Model of Quantitative Traits in $B_1$ and $B_2$, and $F_2$

ZHANG Yuan-ming CAI Jun-yi

*(Soybean Research Institute, Nanjing Agricultural University, National Center of Soybean Improvement, Ministry of Agriculture, Jiangsu Nanjing 210095 China)*

WANG Jian-kang

*(Laboratary Center, Henan Academy of Agricultural Sciences, Henan Zhengzhou 450002 China)*

**Abstract:** Both *QTL* mapping and the segregation analysis indicated that major gene plus polygene mixed model needs to be extended from one major gene to two or more major genes. Fortunately, the iterated ECM (IECM) algorithm makes it possible. In the present paper, Akaike's Information Criterion are employed to identify the existence of major genes affecting quantitative traits. Meanwhile, IECM algorithm is used to obtain the maximum likelihood estimates of parameters of component distributions, as well as the first order or second order genetic parameters From $P_1, P_2, F_1, B_1$ and $B_2$ or $P_1, P_2, F_1$ and $F_2$, the likelihood ratio test is used to test the existence of polygenes. An example of the inheritance of rice plant height is provided.

**Key words:** Quantitative trait ; Major genes plus polygene mixed inheritance ; Mixture model ; Iterated ECM (IECM) algorithm ; Rice plant height