

植物数量性状遗传体系检测中回交 或自交家系重复试验数据的分析方法

章元明, 盖钧镒, 戚存扣

(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)

摘要:为提高主基因+多基因混合遗传分析的精度,降低试验误差,采用重复内分组随机区组设计,对低遗传力性状的 $B_{1:2}$ 和 $B_{2:2}$ 或 $F_{2:3}$ 家系平均数资料进行遗传分析。通过 AIC 准则和适合性检验比较无主基因(A-0)、1 对主基因(A)、2 对主基因(B)、多基因(C)、1 对主基因+多基因(D)和 2 对主基因+多基因(E)模型以鉴定其遗传模式。采用 IECM 算法估计混合模型参数。通过油菜 HSTC14 × 宁油 7 号初花期 $F_{2:3}$ 家系平均数资料阐明该方法。

关键词:数量性状; 主基因+多基因混合遗传; IECM 算法; 油菜

中图分类号: Q348

文献标识码: A

文章编号: 0253-9772(2001)04-0329-04

Detection of Genetic System of Quantitative Traits Using Backcross and Selfing Families

ZHANG Yuan-ming, GAI Jun-yi, QI Cun-kou

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

Abstract: To improve the precision in the genetic analysis of quantitative traits, the $B_{1:2}$ and $B_{2:2}$, or $F_{2:3}$ families in a randomized blocks design were used to identify the mixed major gene plus polygene inheritance model while error variance was estimated from the analysis of variance. Five kinds of genetic models were established, including: one-major-gene model, two-major-gene model, polygene model, mixed one-major-gene plus polygene model, and mixed two-major-gene plus polygene model. The AIC value and a set of tests of goodness-of-fit were used to identify the most fitted model among the possible ones. The iterated ECM (IECM) algorithm was used to obtain maximum likelihood estimates of the parameters in sample likelihood function. An example of the genetic analysis of days from planting to flowering of a rape cross was used to illuminate the above procedure.

Key words: quantitative trait; major genes plus polygene mixed inheritance; IECM algorithm; mixture model; *brassica napus* L.

QTL 作图启示数量性状可由效应大的主基因和效应小的多基因所控制,因而支持主基因+多基因混合遗传模式理论。由此,王健康和盖钧镒在 Elston & Steward 的基础上提出了 1 对主基因+多基因的统计分析方法^[1-4,7,8],章元明和盖钧镒拓展到 2 对主基因+多基因混合遗传模型^[5-6]。应用研究

发现,试验误差大小和性状遗传力高低是影响其分析效果的重要因素,特别是以单株观测值为基础的方法。因此,发展以家系为试验单位并结合有重复试验来降低试验误差提高低遗传力性状检测 QTL 体系的精度。其中最便于获得的试验材料是回交或自交家系世代。为此,本文提出利用 $B_{1:2}$ 和 $B_{2:2}$ 或 $F_{2:3}$

收稿日期:2000-05-29;修回日期:2000-08-10

基金项目:国家 973 项目、重庆市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,硕士研究生毕业,副教授,研究方向:数量遗传学和生物统计学。E-mail:soy Zhang@njau.edu.cn

家系的主基因 + 多基因混合遗传分析法。为提高参数估计的效率和准确性,采用 IECM 算法^[5]。

1 分析方法的导演

1.1 试验数据的统计模型

利用家系进行遗传试验近来一些研究者常采用重复内分组随机区组设计。由于这种设计也属不完全区组设计,特别适用于供试样本为随机样本的情形,这时假定分组后的每组家系均为总体代表性样本,组间应无显著差异,若存在显著差异,应为环境误差,可予以剔除。设将家系群体供试家系随机分为 a 组,每组有 b 个家系,每家系设置 c 个重复,按重复内分组随机区组设计,每一重复种 ab 小区。若数据单位为小区时,数量性状 x_{ijk} 的数学模型为:

$$x_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \beta_{j(i)} + \epsilon_{ijk}$$

其中, μ 为群体平均数; α_i 为第 i 组的效应,服从 $N(0, \sigma_\alpha^2)$; γ_k 为重复效应,组与重复间的互作效应 $(\alpha\gamma)_{ik} \sim N(0, \sigma_{\alpha\gamma}^2)$; $\beta_{j(i)}$ 为第 i 组内第 j 个家系的效应,服从 $N(0, \sigma_\beta^2)$; 试验误差随机变量 $\epsilon_{ijk} \sim N(0, \sigma^2)$ 。其期望均方见应用实例中的表 2。

若数据单位为小区内单株,其数学模型为:

$$x_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \beta_{j(i)} + (\beta\gamma)_{jk(i)} + \delta_{ijk}$$

其中, $(\beta\gamma)_{jk(i)}$ 与以小区为数据单位的 ϵ_{ijk} 同义,可检测组内家系间的差异显著性; $\delta_{ijk} \sim N(0, \sigma_\delta^2)$ 为小区内株间差异。这时,可利用小区平均数进行以下的分析。显然,相同材料不同重复间差异为误差差异,该误差是小区内单株间误差的 $1/n$ (n 为小区内观测植株数)。

若组间差异显著,则剔除组效应后进行以下分析。若组内不同家系间差异不显著,则分析到此为止。通过方差分析得到的误差方差作为混合模型成分分布方差中的误差方差组分。

1.2 B_{1:2}、B_{2:2}和 F_{2:3}家系群体的遗传组成

当存在 1 对主基因时,假定亲本主基因型为 $AA(P_1)$ 和 $aa(P_2)$, $B_1 = F_1 \times P_1$, $B_2 = F_1 \times P_2$ 。按主基因型家系分, $B_{1:2}$ 为 AA 和 aa 家系等比例混合,用 $N(\mu_{1t}, \sigma_{1t}^2)$ ($t = 1, 2$) 表示其家系平均数分布; $B_{2:2}$ 为 Aa 和 aa 家系等比例混合,用 $N(\mu_{2t}, \sigma_{2t}^2)$ ($t = 1, 2$) 表示其家系平均数分布; $F_{2:3}$ 为 AA 、 Aa 和 aa 家系按 1:2:1 混合,用 $N(\mu_{3t}, \sigma_{3t}^2)$ ($t = 1, 2, 3$) 表示其家系平均数分布。记成分分布方差的主基

因、多基因和环境三种方差组分分别为 C_0 、 σ_{j0}^2 和 σ_e^2/n , 则

$$\sigma_{11}^2 = \sigma_{10}^2 + \sigma_e^2/n, \sigma_{12}^2 = \sigma_{11}^2 + C_0$$

$$\sigma_{21}^2 = \sigma_{22}^2 + C_0, \sigma_{22}^2 = \sigma_e^2/n$$

$$\sigma_{31}^2 = \sigma_{33}^2 = \sigma_{30}^2 + \sigma_e^2/n, \sigma_{32}^2 = \sigma_{31}^2 + C_0$$

其中, $C_0 = (1/2d^2 + 1/4h^2)/n_0$ 。

2 对主基因的情形参见文献^[5]。

1.3 遗传模型的建立、分布参数和遗传参数的估计

主 + 多基因遗传体系的基本假定参见文献^[4]。 $B_{1:2}$ 和 $B_{2:2}$ 群体鉴定主基因 + 多基因遗传模型的样本似然函数为:

$$L_1(Y | \theta) = \prod_{i=1}^{n_1} \sum_{i=1}^{k_1} \pi_{1i} f(x_{1i}; \mu_{1i}, \sigma_{1i}^2) \prod_{i=1}^{n_2} \sum_{i=1}^{k_2} \pi_{2i} f(x_{2i}; \mu_{2i}, \sigma_{2i}^2)$$

$F_{2:3}$ 的样本似然函数为:

$$L_2(Y | \theta) = \prod_{i=1}^{n_3} \sum_{i=1}^{k_3} \pi_{3i} f(x_{3i}; \mu_{3i}, \sigma_{3i}^2)$$

其中, $f(x_i; \mu, \sigma^2)$ 是正态分布 $N(\mu, \sigma^2)$ 的密度函数, n_1 与 k_1 、 n_2 与 k_2 以及 n_3 与 k_3 分别是 $B_{1:2}$ 、 $B_{2:2}$ 和 $F_{2:3}$ 群体的家系数和成分分布个数。由此可建立 A(1 对主基因)、B(2 对主基因)、C(多基因)、D(1 对主基因 + 多基因)和 E(2 对主基因 + 多基因)5 类遗传模型,表 1 列出了 E 类模型和利用 $B_{1:2}$ 和 $B_{2:2}$ 的 C、D 类模型。应当注意,多基因存在时,可不作多基因符合加性 - 显性模型的假定。

采用 IECM 算法^[5]获得似然函数 $L_1(Y | \theta)$ 和 $L_2(Y | \theta)$ 中分布参数的极大似然估计值。根据五类模型 AIC 值和适合性检验选择最适遗传模型。用最小二乘法估计一阶遗传参数,用文献^[1, 3, 4]的方法估计二阶遗传参数。以上模型分布参数的 IECM 算法均用 Turbo C++ 语言编写计算程序,欢迎读者来函联系。

2 应用实例与讨论

油菜 HSTC14 × 宁油 7 号的 120 个 $F_{2:3}$ 家系分 10 组种植,每组 12 个家系,每家系重复 2 次,每重复观测 4 株初花期(以 3 月 1 日记为 1),其方差分析见表 2。从表 2 可知,家系间差异显著,分组间差异不显著。因此,可直接利用原始资料(表 3)进行以下分析。

表 1 B_{1:2} 和 B_{2:2} 或 F_{2:3} 家系群体所包含的成分分布数及可估的遗传参数
 Table 1 The number of component distributions (NCD) and estimable genetic parameters of B_{1:2} and B_{2:2}, or F_{2:3} under mixed major gene plus polygene inheritance models

模型 Model	成分 分布 个数 NCD	独立参数 个 数 No. independent parameter	一阶遗传参数 1st order genetic parameter			约束条件数 No. restrictive conditions
			B _{1:2}	&	B _{2:2}	
C	2	4		m_1, m_2		-
D-1	4	6		m_1, m_2, d, h		-
D-2	4	5		$m_1, m_2, d, (h=0)$		1
D-3	4	5		$m_1, m_2, d (h=h)$		1
D-4	4	5		$m_1, m_2, d (h=-h)$		1
E-1	8	12		$m_1, m_2, d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l$		-
E-2	8	8		$m_1, m_2, d_a, d_b, h_a, h_b, (i=j_{ab}=j_{ba}=l=0)$		2
E-3	8	6		$m_1, m_2, d_a, d_b, (h_a=h_b=0)$		4
E-4	6	5		$m_1, m_2, d (d_a=d_b, h_a=h_b=i=j_{ab}=j_{ba}=l=0)$		3
E-5	8	6		$m_1, m_2, d_a, d_b, h_a, h_b, (i=j_{ab}=j_{ba}=l=0)$		4
E-6	6	5		$m_1, m_2, d (d_a=h_a=d_b=h_b, i=j_{ab}=j_{ba}=l=0)$		3
F _{2:3}						
E-1	9	10		$m_3, d_a, d_b, h_a, h_b, i, j_{ab}, j_{ba}, l$		-
E-2	9	6		$m_3, d_a, d_b, h_a, h_b, (i=j_{ab}=j_{ba}=l=0)$		4
E-3	9	4		$m_3, d_a, d_b, (h_a=h_b=0)$		6
E-4	6	3		$m_3, d (d_a=d_b, h_a=h_b=i=j_{ab}=j_{ba}=l=0)$		3
E-5	9	4		$m_3, d_a, d_b, h_a, h_b, (i=j_{ab}=j_{ba}=l=0)$		6
E-6	6	3		$m_3, d (d_a=h_a=d_b=h_b, i=j_{ab}=j_{ba}=l=0)$		4

表 2 方差分析表
 Table 2 Analysis of variance

变异来源 Source of variation	DF	SS	MS	F	期望均方 Expectation mean squares
重复 Replication	1	6.35	6.35	1.50	$\sigma^2 + b\sigma_{\sigma^2}^2 + ab\sigma_f^2$
分组 Set	9	63.50	7.0556	0.84	$\sigma^2 + c\sigma_f^2 + b\sigma_{\sigma^2}^2 + bc\sigma_f^2$
重复 × 分组 Replication × Set	9	38.15	4.2389	2.45	$\sigma^2 + b\sigma_{\sigma^2}^2$
家系间 Family	110	684.21	6.2201	3.60**	$\sigma^2 + c\sigma_f^2$
误差 Error	110	190.04	1.7276		σ^2
总 Total	239	982.25			

表 3 油菜 HSTC14 × 宁油 7 号 F_{2:3} 家系初花期的次数分布
 Table 3 Distributions of days from planting to flowering for a cross between two rape cultivars

	28	30	32	34	36	38	40	n	\bar{x}
P ₁								40	42.3077
F ₁								20	35.0000
P ₂								40	31.8000
F _{2:3}	3	8	61	120	33	14	1	240	34.6303

主基因存在的鉴定结果见表 4。由表 4 可知, E-1 模型的 AIC 值最低, 且 E-1 模型的适合性检验结果也最好 (表 5), 因此, E-1 模型为最适模型。在遗传参数估计时, 为了使估计值更符合实际情况并估计出多基因的加性与显性效应, 可利用亲本和 F₁ 的群体平均数与 F_{2:3} 成分分布平均数共同估计一阶遗传参数, 其估计值为: $m = 36.17, d_a = d_b = 0.99, h_a = -3.17, h_b = -1.43, i = 0.95, j_{ab} = -$

$l = 6.2, j_{ba} = -3.36, l = 6.90, [d] = 3.28, [h] = -3.42$ 。这一遗传参数估计结果既能有效地解释亲本和 F_1 的表现,又能解释 $F_{2:3}$ 群体的表现。因此,在利用单一分离群体进行分离分析时加进亲本和 F_1 群体平均数估计一阶遗传参数是十分有效的方法。

通过上述一阶遗传参数估计值计算 $F_{2:3}$ 群体各主基因型的理论值,由所获得的理论值计算 $F_{2:3}$ 群体主基因遗传方差: $\sigma_{mg}^2 = 1.2437$;进而计算多基因遗传方差: $\sigma_{pg}^2 = 3.2526 - 1.2437 - 1.7276 = 0.2813$ 。由此,主基因遗传率 $h_{mg}^2 = 1.2437/3.2526 = 38.24\%$,多基因遗传率 $\sigma_{pg}^2 = 0.2813/3.2526 = 8.70\%$ 。

表 5 $F_{2:3}$ 群体配合 A-0, A-1, B-1, C, D-1 和 E-1 模型的适合性检验

Table 5 Tests of goodness-of-fit for Models A-0, A-1, B-1, C, D-1 and E-1 in $F_{2:3}$ population

模型 Model	U_1^2	U_2^2	U_3^2	nW^2	D_n
A-0	0.677(0.41)	1.095(0.30)	54.344***	1.373***	0.160*
A-1	2.309(0.13)	4.650*	7.509***	0.766**	0.144*
B-1	0.148(0.70)	0.125(0.72)	8.443***	0.459(>0.05)	0.102*
C	0.118(0.73)	0.780(0.38)	4.84*	0.495*	0.114*
D-1	0.114(0.74)	0.769(0.38)	4.851*	0.494*	0.114*
E-1	0.001(0.98)	0.040(0.84)	0.491(0.48)	0.266(>0.10)	0.088(>0.05)

利用家系重复试验资料的分离分析与文献^[2]利用单株资料的分离分析比较,前者一方面可降低误差以提高参数估计精度,另一方面只需一次就可鉴定出数量性状的遗传模式。王健康等(1997)和盖钧镒等(1998)是首先确定成分分布数然后确定遗传模型的^[1,2],本文是利用文献[3,4]的数量性状联合多个分离世代的分离分析的思路完成的,这就避免了具有相同成分分布数时不能区分是 1 对主基因控制的还是 2 对主基因控制的缺陷,并且在分布参数的估计中考虑了相应的遗传背景。

在估计一阶遗传参数时利用亲本和 F_1 的平均数资料,这对确定主基因效应相对大小是有作用的并能有效地解释亲本、 F_1 和 $F_{2:3}$ 群体的表现。这是因为 P_1 与 P_2 之间差异可度量基因的加性效应,中亲值与 F_1 间差异可度量基因的显性效应的缘故。

表 4 遗传模型的 AIC 值

Table 4 AIC values under various genetic models

模型 Model	AIC	模型 Model	AIC	模型 Model	AIC
A-0	1181.1718	B-4	983.5978	E-1	960.2113
A-1	977.8405	B-5	992.8699	E-2	976.0257
A-2	1006.0497	B-6	990.8681	E-3	972.0565
A-3	1029.6340	C	968.0446	E-4	970.0588
A-4	999.5370	D-1	972.0274	E-5	972.0541
B-1	968.4349	D-2	970.0557	E-6	970.0546
B-2	977.7195	D-3	972.0528		
B-3	982.2790	D-4	972.0356		

参考文献(References):

- [1] 盖钧镒,王健康.利用回交世代或 $F_{2:3}$ 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.
- [2] 王健康,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.
- [3] 王健康,盖钧镒.数量性状主基因-多基因混合遗传的 P_1, F_1, P_2, F_2 和 $F_{2:3}$ 联合分析方法[J].作物学报,1998,24(6):651-659.
- [4] 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.
- [5] 章元明,盖钧镒.数量性状分离分析中分布参数估计的 IECM 算法[J].作物学报,2000,26(6):699-706.
- [6] 章元明,盖钧镒.利用 DH 或 RIL 群体鉴定 QTL 体系并估计其遗传效应[J].遗传学报,2000,27(7):634-640.
- [7] 盖钧镒,王健康.利用回交世代鉴定数量性状主基因-多基因混合遗传模型[J].遗传,1998,20(增刊):110-111.
- [8] 王健康,盖钧镒.利用因子分析鉴定数量性状遗传体系中的主基因[J].遗传,1998,20(增刊):139-140.