王连铮、戴景瑞(主编), 《全国作物育种学术讨论会论文集》: 19-20。

复杂数量性状基因定位的混合模型方法[①]

朱　军
(浙江农业大学 生物数学研究中心, 杭州 310029)

提　要

本文运用混合线性模型的分析原理，提出了复杂数量性状基因定位的方法，可以分析 QTL 复杂的遗传效应及 QTL×环境互作效应。采用混合线性模型随机效应的无偏预测方法，可以预测基因型值和基因型×环境互作效应值，再运用区间作图法或复合区间作图法间接分析 QTL 的加性、显性遗传主效应及其与环境的互作，还能定位在特定发育阶段表达的 QTL。基于混合模型的复合区间作图法(MCIM 法)可以分析多环境的遗传实验资料，直接分析包括上位性效应的遗传主效应及其与环境的互作效应。

关键词: 数量性状基因定位, QTL 定位方法, 上位性效应, 基因型×环境互作, 发育性状基因定位

作物产量、品质和抗逆等重要的农艺性状大多为数量性状基因所控制。近十年来，随着分子标记检测技术的发展，有关数量性状基因座位(Quantitative Trait Loci，简称 QTL) 的定位方法如区间作图法[1]、复合区间作图法[2]、 标记回归法[3]等已能将众多的数量基因定位在相应的连锁图上。一些学者利用这些定位方法已对水稻、小麦、玉米等农作物的一些重要性状进行了 QTL 定位研究。

许多重要的农艺性状都是复杂的数量性状。它们除了受到简单的加性、显性效应控制以外，可能还受上位性效应、母体遗传效应、及其各项遗传效应与环境的互作等控制。国内外现有的 QTL 分析方法大都运用简单回归或多元回归的统计方法，只能分析简单的数量遗传模型(如加性–显性模型)，尚不能分析复杂的遗传现象及在不同时空下表达的基因效应，如基因型与环境的互作、不同发育阶段的基因表达等。QTL 分析方法的滞后，使作物遗传育种工作者不能有效地对一些重要的农艺性状进行精细定位及遗传效应分析。

七十年代以来发展的混合线性模型分析方法，可以同时分析固定效应和多项随机效应，已成为研究数量性状遗传的重要统计分析方法[4]。近年来，我们在发展混合线性模型统计方法，创立分析复杂数量性状的遗传模型等领域开展了系统的基础理论研究[5]，为经典数量遗传分析提供了一些稳健的遗传模型和无偏的统计方法。

本文运用混合线性模型的分析原理，提出分析复杂数量性状的基因定位新方法。

[①]本研究受国家自然科学基金重大项目的资助。
作者E-mail地址: jzhu@zjau.edu.cn

1. 区间作图方法及复合区间作图方法

1.1. 加性-显性效应的QTL定位

近十年来，在植物中已先后发表了蕃茄、玉米、水稻等 30 多种分子标记连锁图谱[6]。这些连锁图谱可用于对控制质量性状或数量性状的目标基因进行定位，并检测基因与分子标记的连锁关系。

利用方差分析的方法，可以检测出数量性状与单一分子标记之间的相互关联。但是这种方法无法对数量性状基因的位置及其效应进行有效地定位和估计，因而不能满足 QTL 精细作图的研究需要。Lander 和 Botstein(1989)提出了基于最大似然分析原理的区间作图(Interval Mapping，简称 IM)QTL 分析方法[1]，可用于推断相邻分子标记(M_{i-} 和 M_{i+})之间的某个 QTL(Q_i)的位置及其遗传效应。如果假定所分析的数量性状只受一对基因控制，并且不存在基因型与环境的互作效应，个体 j 的表现型值(y_j)可以用以下简单回归模型表示：

$$y_j = b_0 + b^* X_j^* + \varepsilon_j$$

其中 b_0 是群体平均数，b^* 是 QTL 遗传效应，X_j^* 是遗传效应的系数，ε_j 是随机机误。

数量性状一般都受多基因控制。当搜索某个 Q_i 时，其它 QTL 的影响可能会干扰区间作图分析的结果。Zeng (1994)提出的复合区间作图(Composite Interval Mapping，简称 CIM)分析方法[2]，在回归模型中包括了与其它 QTL 紧密连锁的分子标记

$$y_j = b_0 + b^* X_j^* + \sum_f b_f X_{fj} + \varepsilon_j$$

其中 b_0 是群体平均数，b^* 是 QTL 遗传效应，X_j^* 是遗传效应的系数，b_f 是第 f 个分子标记基因型的效应，X_{fj} 是个体 j 的 M_f 分子标记效应的系数。CIM 方法可以在一定程度上消除背景遗传变异的干扰。

IM 或 CIM 作图方法是基于简单或多元回归分析方法，在分析复杂数量性状时会遇到一些难以克服的困难。比如，回归模型中除了机误为随机效应以外，其它的效应均为固定效应，因而不能分析属于随机效应的基因型与环境的互作；控制复杂数量性状的一些遗传效应，如果都设为回归模型的固定效应，它们之间因相互混杂而无法有效地被区分。

1.2. 遗传主效应及GE互作效应的QTL定位

如果在不同环境下实施定位 QTL 的遗传试验，需要分析 QTL 的遗传主效应及基因型×环境互作效应。第 j 种基因型在环境 h 中第 k 次重复中的表现型值可以用以下线性模型表示，

$$y_{hjk} = \mu + G_j + E_h + GE_{hj} + e_{hjk} \qquad (1)$$

其中 y_{hjk} 是第 j 种基因型在第 h 个环境内第 k 次重复中的观察值；μ 是群体平均数，固定效应；G_j 是基因型效应，$G_j \sim N(0, \sigma_G^2)$；$E_h$ 是环境效应，$E_h \sim N(0, \sigma_E^2)$；$GE_{hj}$ 是基因型×环境互作效应，$GE_{hj} \sim N(0, \sigma_{GE}^2)$；$e_{hjk}$ 是剩余效应，$e_{hjk} \sim N(0, \sigma_e^2)$。以上公式可以改写为矩阵形式的混合线性模型,

$$\begin{aligned}\mathbf{y} &= \mathbf{1}\mu + \mathbf{U}_G\mathbf{e}_G + \mathbf{U}_E\mathbf{e}_E + \mathbf{U}_{GE}\mathbf{e}_{GE} + \mathbf{e}_\varepsilon \\ &= \mathbf{1}\mu + \sum_{u=1}^{4}\mathbf{U}_u\mathbf{e}_u \\ &\sim N(\mathbf{1}\mu,\quad \mathbf{V} = \sum_{u=1}^{4}\sigma_u^2\mathbf{U}_u\mathbf{U}_u')\end{aligned} \tag{2}$$

其中$\mathbf{y}$是表现型值向量；μ是固定效应的群体平均数；$\mathbf{1}$是系数为1的向量；$\mathbf{e}_1 = \mathbf{e}_G \sim N(\mathbf{0}, \sigma_G^2\mathbf{I})$是基因型效应的随机向量；$\mathbf{e}_2 = \mathbf{e}_E \sim N(\mathbf{0}, \sigma_E^2\mathbf{I})$是环境效应的随机向量；$\mathbf{e}_3 = \mathbf{e}_{GE} \sim N(\mathbf{0}, \sigma_{GE}^2\mathbf{I})$是基因型×环境互作效应的随机向量；$\mathbf{e}_4 = \mathbf{e}_\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$是剩余效应随机向量。

采用随机效应的预测方法(如 BLUP 法、LUP 法、AUP 法)[7]可以获得基因型效应值G_j及基因型×环境互作效应值GE_{hj}的无偏预测，然后可进一步计算$y_{j(G)} = \mu + G_j$和$y_{hj(GE)} = \mu + E_h + GE_{hj}$的预测值。

如果采用区间作图法或复合区间作图法，分析基因型效应值的预测数据$\hat{y}_{j(G)}$，可以定位具有遗传主效应(加性主效应、显性主效应)的 QTL，

区间作图法：$\hat{y}_{j(G)} = b_{0(G)} + b_{(G)}^{*}X_j^{*} + \varepsilon_{j(G)}$

复合区间作图法：$\hat{y}_{j(G)} = b_{0(G)} + b_{(G)}^{*}X_j^{*} + \sum_f b_{f(G)}X_{fj} + \varepsilon_{j(G)}$

其中$b_{0(G)}$是群体平均数，$b_{(G)}^{*}$是 QTL 的遗传主效应，b_f是第f个分子标记基因型的主效应。

如果采用区间作图法或复合区间作图法，分析基因型与第 h 个环境的互作效应的预测数据$\hat{y}_{hj(GE)}$，则可定位具有基因型×环境互作效应 (加性×环境互作效应、显性×环境互作效应)的 QTL，

区间作图法：$\hat{y}_{hj(GE)} = b_{0(GE_h)} + b_{(GE_h)}^{*}X_{hj}^{*} + \varepsilon_{j(GE_h)}$

复合区间作图法：$\hat{y}_{hj(GE)} = b_{0(GE_h)} + b_{(GE_h)}^{*}X_{hj}^{*} + \sum_f b_{f(GE_h)}X_{fhj} + \varepsilon_{j(GE_h)}$

其中$b_{0(GE_h)}$是在第 h 个环境中的群体平均数，$b_{(GE_h)}^{*}$是 QTL×环境 h 的互作效应，$b_{f(GE_h)}$是第f个分子标记基因型×环境 h 的互作效应。

1.3. 发育数量遗传的QTL定位

数量性状的最终表现是生物体不同发育阶段基因表达的综合结果。研究基因在特定时刻的表达及其对数量性状的影响，是发育数量遗传学的一项重要研究内容。

环境 h 中第 k 次重复的第 j 种基因型在 t 时刻(t=1，2，…)的表现型值可以用以下线性模型表示，

$$y_{hjk(t)} = \mu_{(t)} + G_{j(t)} + E_{h(t)} + GE_{hj(t)} + e_{hjk(t)} \tag{3}$$

以上公式可以改写为矩阵形式的混合线性模型,

$$\begin{aligned}\mathbf{y}_{(t)} &= \mathbf{1}\mu_{(t)} + \mathbf{U}_G\mathbf{e}_{G(t)} + \mathbf{U}_E\mathbf{e}_{E(t)} + \mathbf{U}_{GE}\mathbf{e}_{GE(t)} + \mathbf{e}_{\varepsilon(t)} \\ &= \mathbf{1}\mu_{(t)} + \sum_{u=1}^{4}\mathbf{U}_u\mathbf{e}_{u(t)} \\ &\sim N(\mathbf{1}\mu_{(t)},\quad \mathbf{V}_{(t)} = \sum_{u=1}^{4}\sigma^2_{u(t)}\mathbf{U}_u\mathbf{U}'_u)\end{aligned} \tag{4}$$

其中 $\mathbf{y}_{(t)}$ 是在t时刻的表现型值向量；$\mu_{(t)}$ 是t时刻的的群体平均数；$\mathbf{1}$ 是系数为1的向量；$\mathbf{e}_{1(t)} = \mathbf{e}_{G(t)} \sim N(\mathbf{0}, \sigma^2_{G(t)}\mathbf{I})$是$t$时刻的基因型主效应的随机向量；$\mathbf{e}_{2(t)} = \mathbf{e}_{E(t)} \sim N(\mathbf{0}, \sigma^2_{E(t)}\mathbf{I})$是$t$时刻的环境效应的随机向量；$\mathbf{e}_{3(t)} = \mathbf{e}_{GE(t)} \sim N(\mathbf{0}, \sigma^2_{GE(t)}\mathbf{I})$是$t$时刻的基因型×环境互作效应的随机向量；$\mathbf{e}_{4(t)} = \mathbf{e}_{\varepsilon(t)} \sim N(\mathbf{0}, \sigma^2_{\varepsilon(t)}\mathbf{I})$是$t$时刻的剩余效应随机向量。

采用随机效应的预测方法[7]可以获得 $E_{h(t)}$、$G_{j(t)}$ 及 $GE_{hj(t)}$ 的无偏预测，然后进一步计算 $y_{j(G)(t)} = \mu_{(t)} + G_{j(t)}$ 和 $y_{hj(GE)(t)} = \mu_{(t)} + E_{h(t)} + GE_{hj(t)}$ 的预测值。用区间作图法或复合区间作图法分析 t 时刻基因型主效应的预测值数据 $\hat{y}_{j(G)(t)}$，定位的 QTL 具有初始时刻至 t 时刻(0→t)的遗传主效应。采用区间作图法或复合区间作图法，分析基因型与第 h 个环境在 t 时刻的互作效应的预测值数据 $\hat{y}_{hj(GE)(t)}$，可分析在发育阶段(0→t)具有基因型×环境互作效应的QTL。

给定 t–1 时刻的表现型值，t 时刻的条件表现型值是条件随机变量，

$$y_{hjk(t|t-1)} = \mu_{(t|t-1)} + G_{j(t|t-1)} + E_{h(t|t-1)} + GE_{hj(t|t-1)} + e_{hjk(t|t-1)} \tag{5}$$

具有条件变量分布,

$$\begin{aligned}\mathbf{y}_{(t|t-1)} &= \mathbf{1}\mu_{(t|t-1)} + \mathbf{U}_G\mathbf{e}_{G(t|t-1)} + \mathbf{U}_E\mathbf{e}_{E(t|t-1)} + \mathbf{U}_{GE}\mathbf{e}_{GE(t|t-1)} + \mathbf{e}_{\varepsilon(t|t-1)} \\ &= \mathbf{1}\mu_{(t|t-1)} + \sum_{u=1}^{4}\mathbf{U}_u\mathbf{e}_{u(t|t-1)} \\ &\sim N(\mathbf{1}\mu_{(t|t-1)},\quad \mathbf{V}_{(t|t-1)} = \sum_{u=1}^{4}\sigma^2_{u(t|t-1)}\mathbf{U}_u\mathbf{U}'_u)\end{aligned} \tag{6}$$

采用条件随机效应的预测方法[8], 可以获得条件遗传效应 $G_{j(t|t-1)}$ 及条件互作效应 $GE_{hj(t|t-1)}$ 的无偏预测。

用区间作图法或复合区间作图法分析条件变量 $y_{j(G)(t|t-1)} = \mu_{(t|t-1)} + G_{j(t|t-1)}$ 的预测值，所定位QTL的可以推断特定发育阶段(t–1→t)的净遗传主效应。采用区间作图法或复合区间作图法分析条件变量 $y_{hj(GE)(t|t-1)} = \mu_{(t|t-1)} + E_{h(t|t-1)} + GE_{hj(t|t-1)}$ 的预测值，则可以分析(t–1→t)阶段的QTL及净QTL×环境互作效应。

2. 混合线性模型的QTL定位方法

2.1. 混合线性模型的定位方法

在运用分子标记定位数量性状基因的分析中，确定 QTL 的位置和基因的遗传效应是分析的主要目的。随着分子标记技术的飞速发展，某一作物基因组可能获得的分子标记数目远多于基因定位所采用的。对于任何一个基因定位遗传实验，所采用的分子标记可以视为该基因组所

具有的分子标记的一个随机样本。因此在复合区间作图模型中，用于控制背景干扰的分子标记可以作为随机效应，而 QTL 的效应则可设为固定效应。基于这个原理建立的定位基因的遗传模型，可用混合线性模型的通式表示为，

$$\mathbf{y} = \mathbf{Xb} + \sum_u \mathbf{U}_u \mathbf{e}_u \tag{7}$$

该模型是具有平均数 $\mathbf{Xb}$ 和方差 $\mathbf{V} = \sum_u \sigma_u^2 \mathbf{U}_u \mathbf{R}_u \mathbf{U}_u'$, 的混合多变量正态分布。平均数和方差的似然函数 (L) 为

$$L(\mathbf{b},\mathbf{V}) = (2\pi)^{-\frac{n}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})] \tag{8}$$

似然函数的对数(l) 为

$$l(\mathbf{b},\mathbf{V}) = -\tfrac{n}{2}\ln(2\pi) - \tfrac{1}{2}\ln|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}) \tag{9}$$

如果固定效应 **b 是**可估计的，其最大似然估计值可以由下式算得

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

并具有抽样方差

$$\mathrm{Var}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

如果固定效应的系数矩阵奇异，且 **b 不**可无偏估计时，仍能获得固定效应的可估算函数 $\mathbf{c'b}$ 的最大似然估计值

$$\mathbf{c'}\hat{\mathbf{b}} = \mathbf{c'}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{+}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

其抽样方差矩阵为

$$\mathrm{Var}(\mathbf{c'}\hat{\mathbf{b}}) = \mathbf{c'}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{+}\mathbf{c}$$

当我们在相邻的两个分子标记 M_{i-} 和 M_{i+} 之间搜索 QTL，可以设置分子标记 M_{i-} 与 QTL 之间交换值 $r_{M_{i-}Q}$ 的先验值 $\hat{r}_{M_{i-}Q}$，然后计算似然比(LR) 统计量

$$LR = 2l_1(\hat{\mathbf{b}}, \hat{\mathbf{V}}, \hat{r}_{M_{i-}Q}) - 2l_0(\hat{\mathbf{b}}, \hat{\mathbf{V}}, r_{M_{i-}Q} = 0.5) \tag{10}$$

V 中的方差分量可用其无偏估计值替代,

$$\hat{\mathbf{V}} = \sum_u \hat{\sigma}_u^2 \mathbf{U}_u \mathbf{R}_u \mathbf{U}_u' .$$

对于无效假说 H_0: $r_{M_{i-}Q} = 0.5$ (QTL 与分子标记 M_{i-} 相互独立)及其相应的备择假说 H_1: $r_{M_{i-}Q} < 0.5$ (QTL 与分子标记 M_{i-} 相连锁)，可以采用似然比(LR)统计量检验 QTL 是否与分子标记 M_{i-} 相连锁。 LR 近似地具有 χ^2 分布。

当 LR > $\chi^2_{(\mathrm{df})}$ 时拒绝无效假设，可推断 QTL 与相邻的分子标记 M_{i-} 和 M_{i+} 连锁。QTL 的遗传距离可由 $\hat{r}_{M_{i-}Q}$ 计算，遗传效应则由 $\hat{\mathbf{b}}$ 算得。可以采用 t 测验对遗传效应进行检验。如果设置的无效假设与备择假设分别是

$$\mathrm{H}_0 : \mathbf{c}'\mathbf{b} = m \quad \text{vs.} \quad \mathrm{H}_1 : \mathbf{c}'\mathbf{b} \neq m$$

当统计量 $|\mathbf{c}'(\hat{\mathbf{b}}-\mathbf{b})/\sqrt{\mathbf{c}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})\mathbf{c}}\ | > t_\alpha$ 时，拒绝无效假设，接受备择假设。

下文介绍的各种QTL定位混合线性遗传模型都可以采用本节提出的方法分析。

2.2. 加性-显性效应的QTL定位

如果用二个纯系亲本的杂交 F_2 代 n 个个体，搜索分子标记 M_{i-} 和 M_{i+} 之间的数量性状基因位点 Q_i，估算其加性和显性效应，第 j 个个体的数量性状表现型值可由以下混合线性模型表示：

$$\begin{aligned} y_j &= \mu + ax_{A_j} + dx_{D_j} + \sum_{f\neq i-,i+} u_{M_{fj}} e_{M_f} + \varepsilon_j \\ &= \mathbf{x}'_j\mathbf{b} + \mathbf{u}'_{M_j}\mathbf{e}_M + \varepsilon_j \end{aligned} \tag{11}$$

其中 μ 是群体平均数；a 和 d 分别是被搜索的 QTL 的加性和显性效应，固定效应； x_{A_j} 和 x_{D_j} 是遗传效应的系数；e_{M_f} 是标记基因型 f 的随机效应，其三种基因型 M_fM_f、M_fm_f 和 m_fm_f 的系数 $u_{M_{fj}}$ 分别是 1、0 和–1； ε_j 是随机的剩余效应。$\mathbf{b}$ 是固定效应(μ、a 和 d)的参数向量；$\mathbf{e}_M$ 是标记基因型效应的随机向量，$\mathbf{u}'_{M_j}$ 是个体 j 的分子标记效应的系数行向量；$\mathbf{x}'_j$ 是个体 j 的固定效应 $\mathbf{b}$ 的系数行向量。由于个体 j 的 QTL 的基因型是未知的，遗传效应系数值(x_{A_j}，x_{D_j})需由相邻分子标记基因型推断的 QTL 基因型概率算得。表 1 列出了 F_2 遗传群体的 QTL 基因型概率期望值。

表 1. 对应于 F_2 个体相邻分子标记基因型的 QTL 各种基因型概率期望值。

Table 1 Probability of QTL genotype given observed flanking marker genotype in an F_2 population.

$M_{i-}M_{i-}$	$M_{i+}M_{i+}$	$Pr(Q_1Q_1)$	$Pr(Q_1Q_2)$	$Pr(Q_2Q_2)$
++	++	1	0	0
++	+–	$1-p$	p	0
++	––	$(1-p)^2$	$2p(1-p)$	p^2
+–	++	p	$1-p$	0
+–	+–	$\delta p(1-p)$	$1\text{-}2\delta p(1-p)$	$\delta p(1-p)$
+–	––	0	$(1-p)$	p
––	++	p^2	$2p(1-p)$	$(1-p)^2$
––	+–	0	p	$1-p$
––	––	0	0	1

$p = r_{M_{i-}Q} / r_{M_{i-}M_{i+}}$，$\delta = r^2_{M_{i-}M_{i+}} / [(1-r_{M_{i-}M_{i+}})^2 + r^2_{M_{i-}M_{i+}}]$.

模型(11)是第 j 个个体表现型值的线性模型。所有个体的表现型值可以用以下混合线性模型的矩阵形式表示,

$$\begin{aligned} \mathbf{y} &= \mathbf{Xb} + \mathbf{U}_M\mathbf{e}_M + \mathbf{e}_\varepsilon \\ &= \mathbf{Xb} + \sum_{u=1}^{2}\mathbf{U}_u\mathbf{e}_u \\ &\sim N(\mathbf{Xb}, \mathbf{V} = \sum_{u=1}^{2}\sigma_u^2\mathbf{U}_u\mathbf{R}_u\mathbf{U}'_u). \end{aligned} \tag{12}$$

其中$\mathbf{y}$是表现型值向量； $\mathbf{b}$ 是固定效应的参数向量；$\mathbf{X}$是固定效应的系数矩阵，其行向量为

$\mathbf{x}'_j$； $\mathbf{e}_1 = \mathbf{e}_M \sim N(\mathbf{0}, \sigma_M^2 \mathbf{U}_M \mathbf{R}_M \mathbf{U}'_M)$是标记效应的随机向量，$\mathbf{R}_M$是描述标记$M_f$和$M_{f'}$之间相关性的常量矩阵；$\mathbf{U}_M$是标记效应的系数矩阵，并有转置向量矩阵$\mathbf{U}'_M$；$\mathbf{e}_2 = \mathbf{e}_\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$是剩余效应随机向量，其$\mathbf{U}_2 = \mathbf{R}_2 = \mathbf{I}$为单位矩阵。

QTL 简单加性和显性遗传模型(11)是基于不存在基因型×环境互作(GE)的假设。这一假设对大多数数量性状遗传并不适用，因此有必要分析 QTL 与环境的互作效应。如果对多环境下实施的遗传试验资料进行 QTL 定位分析，基因型 *j* 在环境 *h* 中的表现型值可用以下混合线性模型表示，

$$\begin{aligned} y_{hj} &= \mu + a x_{A_j} + d x_{D_j} + u_{E_{hj}} e_{E_h} + u_{AE_{hj}} e_{AE_h} + u_{DE_{hj}} e_{DE_h} \\ &\quad + \sum_{f \neq i-, i+} u_{M_{fj}} e_{M_f} + \sum_{l \neq i-, i+} u_{ME_{hlj}} e_{ME_{hl}} + \varepsilon_{hj} \end{aligned} \tag{13}$$

其中 μ 是群体平均数；*a* 和 *d* 分别是 QTL 的加性主效应和显性主效应，固定效应；x_{A_j} 和 x_{D_j} 是遗传主效应的系数；e_{E_h} 是环境 *h* 的随机效应，其系数为 $u_{E_{hj}}$；e_{AE_h} 是加性×环境互作随机效应，其系数为 $u_{AE_{hj}}$；e_{DE_h} 是显性×环境互作随机效应，其系数为 $u_{DE_{hj}}$；e_{M_f} 是标记基因型 *f* 的主效应，$u_{M_{fj}}$ 是标记主效应的系数；$e_{ME_{lh}}$ 是标记 *l*×环境 *h* 的互作随机效应，其系数为 $u_{ME_{lhj}}$；ε_{hj} 是随机的剩余效应。

模型(13)可用以下混合线性模型的矩阵形式表示,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{U}_E \mathbf{e}_E + \mathbf{U}_{AE} \mathbf{e}_{AE} + \mathbf{U}_{DE} \mathbf{e}_{DE} + \mathbf{U}_M \mathbf{e}_M + \mathbf{U}_{ME} \mathbf{e}_{ME} + \mathbf{e}_\varepsilon \\ &= \mathbf{X}\mathbf{b} + \sum_{u=1}^{6} \mathbf{U}_u \mathbf{e}_u \\ &\sim N(\mathbf{X}\mathbf{b}, \mathbf{V} = \sum_{u=1}^{6} \sigma_u^2 \mathbf{U}_u \mathbf{R}_u \mathbf{U}'_u). \end{aligned} \tag{14}$$

其中$\mathbf{y}$是表现型值向量；$\mathbf{b}$是固定效应的参数向量；$\mathbf{X}$是固定效应的系数矩阵；$\mathbf{e}_1 = \mathbf{e}_E \sim N(\mathbf{0}, \sigma_E^2 \mathbf{I})$是环境效应的随机向量；$\mathbf{e}_2 = \mathbf{e}_{AE} \sim N(\mathbf{0}, \sigma_{AE}^2 \mathbf{I})$是加性×环境互作效应的随机向量；$\mathbf{e}_3 = \mathbf{e}_{DE} \sim N(\mathbf{0}, \sigma_{DE}^2 \mathbf{I})$是显性×环境互作效应的随机向量；$\mathbf{e}_4 = \mathbf{e}_M \sim N(\mathbf{0}, \sigma_M^2 \mathbf{R}_M)$是分子标记基因型主效应的随机向量；$\mathbf{e}_5 = \mathbf{e}_{ME} \sim N(\mathbf{0}, \sigma_{ME}^2 \mathbf{R}_{ME})$是分子标记×环境互作效应的随机向量；$\mathbf{e}_6 = \mathbf{e}_\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$是剩余效应随机向量。

采用(2.1)介绍的混合线性模型分析方法，可以直接定位QTL，并估算其遗传主效应(加性*a*、显性*d*)，还可无偏预测QTL与环境的互作效应(加性×环境互作e_{AE}、显性×环境互作e_{DE})。

2.3. 上位性效应的QTL定位

数量遗传研究结果表明，非等位基因之间的互作效应(上位性效应)是不可忽略的遗传效应组成部分[9, 10]。目前，通常采用方差分析(ANOVA)方法分析分子标记之间的互作效应，然后推断 QTL 的上位性效应[10, 11]。这种分析方法不能精细定位具有上位性效应的 QTL，也无法估算上位性效应值。

采用混合线性模型的 QTL 定位方法，可以有效地推断具有上位性效应的 QTL 位置和遗传效应。利用 DH 群体或 RIL 群体，可以分析加性和加×加上位性效应。基因型 j 在的表现型值可由以下混合线性模型表示，

$$y_j = \mu + a_1 x_{A_1 j} + a_2 x_{A_2 j} + aax_{AA_j} + \sum_f u_{M_{fj}} e_{M_f} + \sum_l u_{MM_{lj}} e_{MM_l} + \varepsilon_j \qquad \textbf{(15)}$$

其中 μ 是群体平均数；a_1 和 a_2 分别是二个基因位点 Q_1 和 Q_2 的加性效应，aa 是 Q_1 和 Q_2 的加×加上位性效应；$x_{A_1 j}$、$x_{A_2 j}$ 和 x_{AA_j} 分别是加性和上位性效应的系数；e_{M_f} 是标记基因型 f 的随机效应，其系数是 $u_{M_{fj}}$；e_{MM_l} 是互作标记基因型 l 的随机效应，其系数是 $u_{MM_{lj}}$；ε_j 是随机的剩余效应。

所有个体的表现型值可用以下混合线性模型的矩阵形式表示,

$$\begin{aligned} \mathbf{y} &= \mathbf{Xb} + \mathbf{U}_M \mathbf{e}_M + \mathbf{U}_{MM} \mathbf{e}_{MM} + \mathbf{e}_\varepsilon \\ &= \mathbf{Xb} + \sum_{u=1}^{3} \mathbf{U}_u \mathbf{e}_u \\ &\sim N(\mathbf{Xb}, \mathbf{V} = \sum_{u=1}^{3} \sigma_u^2 \mathbf{U}_u \mathbf{R}_u \mathbf{U}'_u). \end{aligned} \qquad \textbf{(16)}$$

其中**y**是表现型值向量；**b** 是固定效应向量，**X**是其系数矩阵；$\mathbf{e}_1 = \mathbf{e}_M$ 是标记效应的随机向量，$\mathbf{R}_1 = \mathbf{R}_M$是描述相邻分子标记之间相关性的常量矩阵；$\mathbf{U}_1 = \mathbf{U}_M$是标记效应的系数矩阵；$\mathbf{e}_2 = \mathbf{e}_{MM}$ 是分子标记互作效应的随机向量，$\mathbf{R}_2 = \mathbf{R}_{MM}$是描述互作分子标记之间相关性的常量矩阵；$\mathbf{U}_2 = \mathbf{U}_{MM}$是互作分子标记效应的系数矩阵；$\mathbf{e}_3 = \mathbf{e}_\varepsilon$ 是剩余效应随机向量，其 $\mathbf{U}_3 = \mathbf{R}_3 = \mathbf{I}$是单位矩阵。

如果要分析上位性 QTL 的遗传主效应及与环境的互作效应，应该在多环境下实施遗传实验。DH 群体或 RIL 群体的个体 j 在环境 h 中的表现型值可用以下混合线性模型表示，

$$\begin{aligned} y_{hj} &= \mu + a_1 x_{A_1 j} + a_2 x_{A_2 j} + aax_{AA_j} \\ &+ u_{E_{hj}} e_{E_h} + u_{A_1 E_{hj}} e_{A_1 E_h} + u_{A_2 E_{hj}} e_{A_2 E_h} + u_{AAE_{hj}} e_{AAE_h} \\ &+ \sum_f u_{M_{fj}} e_{M_f} + \sum_l u_{MM_{lj}} e_{MM_l} + \sum_p u_{ME_{hpj}} e_{ME_{hp}} + \sum_q u_{MME_{hqj}} e_{MME_{hq}} + \varepsilon_{hj} \end{aligned} \qquad \textbf{(17)}$$

其中 μ 是群体平均数；a_1 和 a_2 分别是二个基因位点 Q_1 和 Q_2 的加性主效应，aa 是 Q_1 和 Q_2 的加×加上位性主效应；$x_{A_1 j}$、$x_{A_2 j}$ 和 x_{AA_j} 分别是加性和上位性主效应的系数；e_{E_h} 是环境 h 的随机效应，其系数为 $u_{E_{hj}}$；$e_{A_1 E_h}$ 是位点 Q_1 的加性×环境互作随机效应，其系数为 $u_{A_1 E_{hj}}$；$e_{A_2 E_h}$ 是位点 Q_2 的加性×环境互作随机效应，其系数为 $u_{A_2 E_{hj}}$；e_{AAE_h} 是双位点 Q_1 和 Q_2 的上位性×环境互作随机效应，其系数为 $u_{AAE_{hj}}$；e_{M_f} 是标记基因型的随机效应，其系数是 $u_{M_{fj}}$；e_{MM_l} 是互作分子标记的随机效应，其系数是 $u_{MM_{lj}}$；$e_{ME_{hp}}$ 是分子标记×环境 h 的互作效应，其系数为 $u_{ME_{hpj}}$；$e_{MME_{hq}}$ 是互作分子标记×环境 h 的互作效应，其系数为 $u_{MME_{hqj}}$；ε_{hj} 是随机的剩余效应。

基于模型(17)的所有个体，其表现型值可以用以下混合线性模型的矩阵形式表示,

$$\begin{aligned}\mathbf{y} &= \mathbf{Xb} + \mathbf{U}_E\mathbf{e}_E + \mathbf{U}_{A_1E}\mathbf{e}_{A_1E} + \mathbf{U}_{A_2E}\mathbf{e}_{A_2E} + \mathbf{U}_{AAE}\mathbf{e}_{AAE} \\ &\quad + \mathbf{U}_M\mathbf{e}_M + \mathbf{U}_{MM}\mathbf{e}_{MM} + \mathbf{U}_{ME}\mathbf{e}_{ME} + \mathbf{U}_{MME}\mathbf{e}_{MME} + \mathbf{e}_\varepsilon \\ &= \mathbf{Xb} + \sum_{u=1}^{9}\mathbf{U}_u\mathbf{e}_u \qquad (18) \\ &\sim N(\mathbf{Xb}, \mathbf{V} = \sum_{u=1}^{9}\sigma_u^2\mathbf{U}_u\mathbf{R}_u\mathbf{U}_u').\end{aligned}$$

其中$\mathbf{y}$是表现型值向量；$\mathbf{b}$是固定效应的参数向量；$\mathbf{X}$是固定效应的系数矩阵；$\mathbf{e}_1 = \mathbf{e}_E \sim N(\mathbf{0}, \sigma_E^2\mathbf{I})$是环境效应的随机向量；$\mathbf{e}_2 = \mathbf{e}_{A_1E} \sim N(\mathbf{0}, \sigma_{A_1E}^2\mathbf{I})$是位点$Q_1$的加性×环境互作效应的随机向量；$\mathbf{e}_3 = \mathbf{e}_{A_2E} \sim N(\mathbf{0}, \sigma_{A_2E}^2\mathbf{I})$是位点$Q_2$的加性×环境互作效应的随机向量；$\mathbf{e}_4 = \mathbf{e}_{AAE} \sim N(\mathbf{0}, \sigma_{AAE}^2\mathbf{R}_{AAE})$是双位点$Q_1$和$Q_2$的加加上位性×环境互作效应的随机向量；$\mathbf{e}_5 = \mathbf{e}_M \sim N(\mathbf{0}, \sigma_M^2\mathbf{R}_M)$是分子标记主效应的随机向量；$\mathbf{e}_6 = \mathbf{e}_{MM} \sim N(\mathbf{0}, \sigma_{MM}^2\mathbf{R}_{MM})$是互作分子标记效应的随机向量；$\mathbf{e}_7 = \mathbf{e}_{ME} \sim N(\mathbf{0}, \sigma_{ME}^2\mathbf{R}_{ME})$是分子标记×环境互作效应的随机向量；$\mathbf{e}_8 = \mathbf{e}_{MME} \sim N(\mathbf{0}, \sigma_{MME}^2\mathbf{R}_{MME})$是互作分子标记×环境互作效应的随机向量；$\mathbf{e}_9 = \mathbf{e}_\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$是剩余效应随机向量。

采用混合线性模型的分析方法，可以定位具有上位性的 QTL, 并估算其遗传主效应(加性a_1、加性a_2及其加性与加性的上位性aa)，还能预测 QTL 与环境的互作效应(加性×环境互作e_{AE}、 加加上位性×环境互作e_{AAE})。

3. 讨论

区间作图法和复合区间作图法都基于回归模型分析原理，并建立在一个简单的遗传假设上：数量性状的表现型变异受遗传效应(固定效应)和剩余机误(随机效应)控制，不存在基因型×环境的互作效应。可以用简单的数量遗传模型表示为，

$$y = \mu + G + \varepsilon \qquad (19)$$

其中y是个体的表现型值，μ是群体平均数，G是遗传效应值，ε是随机机误。

区间作图法假定遗传变异只受一对基因控制，因此遗传效应就是QTL的效应($G_Q = b^*X$)。考虑到数量性状实际上受多基因控制，复合区间作图法把总的遗传效应分解为被搜索的 QTL 效应($G_Q = b^*X$)及与其它 QTL 连锁的分子标记效应($G_M = \sum b_f X_{fj}$)两个分量。

但是在多环境下实施的遗传实验，其遗传群体的表现型变异除了受遗传效应(G)和剩余机误(ε)控制以外，还会受到环境效应(E)和基因型×环境互作效应(GE)的控制。包括环境及基因型×环境互作效应的遗传模型可简单表示为

$$y = \mu + G + E + GE + \varepsilon \qquad (20)$$

区间作图法和复合区间作图法是应用回归模型，分析 QTL 与分子标记的连锁关系及遗传效应。除了剩余机误以外，所有回归效应只能设为固定效应。因此，这二种方法不能直接分析环境及环境互作等随机效应。

采用(1.2)中提出的混合线性模型分析方法，可以先分别预测基因型主效应值函数($y_{(G)} = \mu + G$)和基因型×环境互作效应函数($y_{(GE)} = \mu + E + GE$)。然后再借助于目前流行

的 QTL 分析方法(区间作图法或复合区间作图法)及相应的统计分析软件(MAPMAKER/QTL 或 QTL Cartographer)，便可间接地定位具有遗传主效应和环境互作效应的 QTL[12]。采用(1.3)中提出的发育数量遗传分析的方法，还可以定位在特定发育阶段($t \rightarrow t$–1)表达的、具有净遗传主效应以及净环境互作效应的发育特异性 QTL[13]。

复合区间作图方法是基于回归模型分析原理的(称为 Regression-model-based CIM 法,或 RCIM 法)，它把总的遗传效应(G)分解为两个固定效应的分量 ($G = G_Q + G_M = b^* X + \sum b_f X_{fj}$)。因而，在这个多元回归模型中，QTL 效应是由偏回归系数(b^*)算得，其效应值完全取决于模型中所包括的其它分子标记的效应。合理地选取分子标记，可以有效地控制背景遗传变异的干扰；但是如果选取不当，也会影响 QTL 效应的估算。

本文第二节中所提出的 QTL 分析方法是基于混合模型的复合区间作图(称为 Mixed-model-based CIM 法,或 MCIM 法)方法。该方法把控制背景遗传变异的分子标记效应($G_M = \sum u_{fj} e_f$)归为随机变量，使它们不会影响对 QTL 位置及遗传效应的无偏估算。这种分析方法还可以在模型中包括环境效应及环境互作效应，

$$y = \mu + G_Q + E + G_Q E + G_M + G_M E + \varepsilon \qquad (21)$$

其中 y 是个体的表现型值，群体平均数(μ)和 QTL 遗传主效应(G_Q)是固定效应；环境效应(E)、QTL×环境互作效应($G_Q E$)、分子标记效应(G_M)及其与环境的互作效应($G_M E$)、剩余机误效应(ε)都是随机效应。采用(2.1)中介绍的最大似然法，可以估算 QTL 的效应值或效应值的无偏线性函数。随机效应则可以采用混合线性模型的预测法无偏预测[8]。由此可有效地直接分析 QTL 的遗传主效以及 QTL×环境互作效应。

参考文献

[1] Lander, E.S. and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. ***Genetics,*** 121: 185-199.

[2] Zeng, Z.B., 1994 Precision mapping of quantitative trait loci. Genetics 136: 1457-1468.

[3] Moreno-Gonzalez, J. 1992. Estimates of marker-associated QTL effects in Monte Carlo backcross generations using multiple regression. ***Theor Appl Genet,*** 85:423-434.

[4] 朱军: 1994. 广义遗传模型与数量遗传分析新方法. **浙农大学报**, 1994, 20(6)551-559

[5] 朱军: 1997.《遗传模型分析方法》。**中国农业出版社**

[6] 陆朝福和朱立煌，1995. 植物育种中的分子标记辅助选择。 **生物工程进展** ,15(4): 11-17。

[7] Zhu, J. and B.S. Weir, 1996 Diallel analysis for sex-linked and maternal effects. ***Theor Appl Genet,*** 92: 1-9.

[8] Zhu, J. 1995 Analysis of conditional genetic effects and variance components in developmental genetics. ***Genetics,*** 141: 1633-1639.

[9] Li, Z. K., S. R. M. Pinson, W. D. Park, A. H. Paterson and J. W. Stansel. 1997. Epistasis for three grain components in rice (Oryza sativa L.). ***Genetics,*** 145: 453-465.

[10] Yu, S.-B., J.-X. Li, C.-G. Xu, Y.-F. Tan, Y.-J. Gao, X.-H. Li *et al*., 1997 Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. ***Proc Natl Acad Sci,*** USA 94: 9226-9231.

[11] Li, Z.-K., 1997 Molecular analysis of epistasis, pp. 119-130 in Molecular Dissection Of Complex Traits, edited by A. H. Paterson , ***CRC Press LLC***, Boca Raton, Florida.

[12] Yan J.-Q., J. Zhu, C. -X He, M. Benmoussa, and P. Wu, 1998a. Quantitative trait loci analysis for developmental behavior of tiller number in rice (*Oryza sativa* L.) ***Theor Appl Genet***, (in press)

[13] Yan J.-Q., J. Zhu, C.-X He, M. Benmoussa, and P. Wu, 1998b Molecular dissection of developmental behavior of plant height in rice (*Oryza sativa* L.). ***Genetics***, (in press)

Mixed Model Approaches of Mapping Genes for Complex Quantitative Traits

Jun Zhu
Research Center of Biomathematics, Zhejiang Agricultural University
Hangzhou 310029, China

Abstract

New QTL mapping methods based on mixed linear model approaches were proposed for analyzing complex genetic effects and their interaction with environments. Unbiased prediction can be applied for predicting genotype effects and genotype × environment interaction effects, which can then be further used for mapping QTL or developmental QTL with genetic main effects and GE interaction effects by interval mapping or composite interval mapping approaches. Mixed-model-based composite interval mapping approaches are capable of handling genetic data derived from multiple environments and directly analyzing genetic main effects (including epistasis) and GE interaction effects.

Key Words: Quantitative trait loci, QTL mapping methods, Epistasis, Genotype × environment interaction, Developmental QTL mapping