

DOI: 10.3724/SP.J.1006.2009.00239

数量性状基因的完备区间作图方法

王建康

中国农业科学院作物科学研究所 / 国家农作物基因资源与基因改良重大科学工程 / CIMMYT 中国办事处, 北京 100081

摘要: 结合分子标记和表型数据的 QTL 作图已成为数量性状遗传分析的常规方法。复合区间作图是近 10 多年来广泛应用的一种 QTL 定位方法, 但它在算法上有一些缺陷, 致使 QTL 效应可能会被侧连标记区间之外的标记变量吸收, 同时不同的背景标记选择方法对作图结果的影响较大, 并且难以推广到上位型互作 QTL 的定位。针对这些问题, 笔者提出完备区间作图方法。本文介绍了该方法的遗传和统计原理, 并通过一个大麦加倍单倍体群体说明其在定位加性 QTL 和加性×加性互作 QTL 中的应用。完备区间作图包含两个步骤: 首先利用所有标记的信息, 通过逐步回归选择重要的标记变量并估计其效应; 然后利用逐步回归得到的线性模型校正表型数据, 通过一维扫描定位加(显)性效应 QTL, 通过二维扫描定位上位型互作 QTL。这种作图策略简化了复合区间作图中控制背景遗传变异的过程, 提高了对 QTL 的检测功效。

关键词: 数量性状; QTL 作图; 完备区间作图; 加显性效应; 上位型互作

Inclusive Composite Interval Mapping of Quantitative Trait Genes

WANG Jian-Kang

Institute of Crop Sciences / National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences / CIMMYT China Office, Beijing 100081, China

Abstract: Rapid increase in the availability of fine-scale genetic marker maps has led to the intensive use of QTL mapping in the genetic study of quantitative traits. Composite interval mapping (CIM) is one of the most commonly used methods for QTL mapping with populations derived from biparental crosses. However, the algorithm used in CIM cannot completely ensure that the effect of QTL at current testing interval is not absorbed by the background marker variables, and may result in biased estimation of QTL effect. We proposed a statistical method for QTL mapping, which was called inclusive composite interval mapping (ICIM). Two steps were included in ICIM. In the first step, stepwise regression was applied to identify the most significant regression variables. In the second step, a one-dimensional scanning or interval mapping was conducted for detecting additive (and dominance) QTL and a two-dimensional scanning was conducted for detecting digenic epistasis. ICIM provides intuitive statistics for testing additive, dominance and epistasis, and can be used for most experimental populations derived from two inbred parental lines. The EM algorithm used in ICIM has a fast convergence speed and is therefore less computing intensive. ICIM retains all advantages of CIM over interval mapping, and avoids the possible increase of sampling variance and the complicated background marker selection process in CIM. A doubled haploid (DH) population in barley was used to demonstrate the application of ICIM in mapping additive QTL and additive by additive interacting QTL.

Keywords: Quantitative trait; QTL mapping; Inclusive composite interval mapping; Additive and dominance effects; Epistatic interaction

随着分子标记技术的发展和标记连锁图谱的建立, 人们已经可以像研究质量性状基因一样研究数量性状基因, 也可以把单个数量性状基因(quantitative trait gene or locus, QTL)定位在染色体上, 并估计出单个数量性状基因的遗传效应, 寻找 QTL 在染色体上的位置并估计其遗传效应的过程称为 QTL 作图^[1-2]。自 1989

年 QTL 区间作图方法提出以来^[3], QTL 作图逐渐成为数量遗传学的研究重点^[4-13], 根据 QTL 定位结果对数量性状基因进行图位克隆、利用标记对性状进行间接选择等, 都已有成功的例子^[14-19]。

QTL 作图方法研究大致经历以下几个过程^[2]。一是单标记或单点分析, 通过比较不同标记基因型

本研究由国家自然科学基金项目(30771351), 国家高技术研究发展计划(863 计划)项目(2006AA10Z1B1)资助。

作者联系方式: E-mail: wangjk@caas.net.cn; jkwang@cgiar.org; Tel: 010-62133890

Received(收稿日期): 2008-06-19; Accepted(接受日期): 2008-10-04.

均值间的差异显著性来测验 QTL 的存在,这一方法只有在 QTL 的位置与标记完全重合、每条染色体上至多包含一个 QTL 的条件下,才能获得准确的定位结果。二是区间作图^[3],该方法的基本假定是每条染色体上至多包含一个 QTL、QTL 的遗传效应满足加一显性模型,因此当实际情况不符合此假定时,例如一条染色体上有 2 个 QTL 时,若两 QTL 的作用方向相反,往往检测不到;若作用方向相同,在两 QTL 间可能会出现一个“幻影”QTL,且 QTL 位置的置信区间较大,一般在 10~30 cM 之间,会造成待估 QTL 位置与效应估计值的偏差。三是复合区间作图^[20],通过引入其他标记作为协变量来消除区间以外 QTL 对作图区间的影响,从而消除“幻影”QTL 现象,适用于同一染色体上有多个 QTL 的情形。四是多 QTL 的同时定位^[9-13],多 QTL 定位的方法主要有极大似然法和贝叶斯方法两大类,极大似然法主要包括多区间作图法^[9]和惩罚最大似然法^[10]。贝叶斯方法主要包括可逆跳跃马尔可夫蒙特卡罗(Markov chain Monte Carlo, 简称 MCMC)方法^[11]和压缩估计方法^[12]。多 QTL 同时定位的缺点在于算法太复杂、收敛速度太慢、运算时间太长、而且需要较大的样本量。若标记较多时,难以进行参数估计,很多贝叶斯模型无法合理分析一个包含 127 标记、群体大小为 145 的大麦 DH 家系^[13],因此,如何将贝叶斯方法应用于实际作图群体仍有许多问题值得探讨。

复合区间作图(Composite Interval Mapping, 简称 CIM)是近 10 多年来广泛应用的 QTL/基因定位方法^[20]。最近我们发现该方法在算法上存在着一些缺陷,致使 QTL 的效应可能会被侧连标记区间之外的标记变量吸收,同时不同的背景标记选择方法对作图结果的影响很大,并且难以推广到上位型互作 QTL 的定位^[21-22]。针对这些问题,我们提出完备区间作图方法(Inclusive Composite Interval Mapping, 简称 ICIM)^[21-23]。本文介绍这一方法的基本遗传学和统计学原理,并通过一个真实群体说明完备区间作图方法的应用。

1 QTL 完备区间作图方法

1.1 单个 QTL 的加显性遗传模型

在只有 1 个 QTL(Q 和 q 表示该位点上的 2 个等位基因)的加显性遗传模型下,3 种 QTL 基因型 QQ 、 Qq 和 qq 的基因型值可表示为:

$$G = \mu + aw + dv \quad (1)$$

式中, μ 代表 2 种纯合基因型 QQ 和 qq 的平均, a 为

加性效应, d 为显性效应^[1-2], w 和 v 是 QTL 基因型的指示变量, $w = 1$ 和 $v = 0$ 代表基因型 QQ , $w = 0$ 和 $v = 1$ 代表基因型 Qq , $w = -1$ 和 $v = 0$ 代表基因型 qq 。如果不考虑显性效应,遗传模型可进一步简化为:

$$G = \mu + aw \quad (2)$$

QTL 作图时,个体的 QTL 基因型是未知的,即模型(1)和(2)中的 w 和 v 未知,参数 a 和 d 有待估计。QTL 作图前,已知个体的标记型,但由于标记和 QTL 之间存在连锁,标记型提供了 QTL 基因型的信息,因此需要寻求 QTL 基因型与标记基因型间的关系。

1.2 单个 QTL 的线性统计模型

假定 2 个共显性标记(分别用 $A-a$ 和 $B-b$ 表示)之间存在 1 个 QTL, F_2 群体中有 9 种不同的标记型,每种标记型下 3 种 QTL 基因型的频率可通过 QTL 与 2 个标记位点间的重组率来估计^[23]。可以证明, QTL 基因型指示变量 w 的期望值是标记型指示变量 x_1 和 x_2 的线性函数,指示变量 v 的期望值是标记型指示变量 y_1 、 y_2 以及乘积项 x_1x_2 、 y_1y_2 的线性函数,即

$$E(w | x_1, x_2, y_1, y_2) = \lambda_1 x_1 + \lambda_2 x_2 \quad (3)$$

$$E(v | x_1, x_2, y_1, y_2) = \delta + \rho_1 y_1 + \rho_2 y_2 + \lambda \lambda_{12} x_1 x_2 + \rho \rho_{12} y_1 y_2 \quad (4)$$

其中,

$$\begin{bmatrix} \delta \\ \lambda_1 \\ \lambda_2 \\ \rho_1 \\ \rho_2 \\ \lambda \lambda_{12} \\ \rho \rho_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(g_1 + g_3) \\ f_2 \\ \frac{1}{2}(f_1 - f_3) \\ (-\frac{1}{2}g_1 - \frac{1}{2}g_3 + g_4) \\ (-\frac{1}{2}g_1 + g_2 - \frac{1}{2}g_3) \\ \frac{1}{2}(g_1 - g_3) \\ (\frac{1}{2}g_1 - g_2 + \frac{1}{2}g_3 - g_4 + g_5) \end{bmatrix}$$

上式中,变量 f_1 、 f_2 、 f_3 、 g_1 、 g_2 、 g_3 、 g_4 和 g_5 是侧连标记间以及 QTL 与标记间重组率的函数^[23]。这样得到各种标记型的遗传型值:

$$\begin{aligned} E(G | x_1, x_2, y_1, y_2) &= \mu + aE(w | x_1, x_2, y_1, y_2) + dE(v | x_1, x_2, y_1, y_2) \\ &\triangleq [\mu + (d)\mu_d] + (a)A_1x_1 + (d)D_1y_1 + (a)A_2x_2 + (d)D_2y_2 + \\ &\quad (d)AA_{12}x_1x_2 + (d)DD_{12}y_1y_2 \end{aligned} \quad (5)$$

模型(5)中参数的取值为:

$$\begin{bmatrix} \mu + (d)\mu_d \\ (a)A_1 \\ (a)A_2 \\ (d)D_1 \\ (d)D_2 \\ (d)AA_{12} \\ (d)DD_{12} \end{bmatrix} = \begin{bmatrix} \mu + \delta d \\ \lambda_1 a \\ \lambda_2 a \\ \rho_1 d \\ \rho_2 d \\ \lambda \lambda_{12} d \\ \rho \rho_{12} d \end{bmatrix} = \begin{bmatrix} \mu + \frac{1}{2}(g_1 + g_3)d \\ f_2 a \\ \frac{1}{2}(f_1 - f_3)a \\ (-\frac{1}{2}g_1 - \frac{1}{2}g_3 + g_4)d \\ (-\frac{1}{2}g_1 + g_2 - \frac{1}{2}g_3)d \\ \frac{1}{2}(g_1 - g_3)d \\ (\frac{1}{2}g_1 - g_2 + \frac{1}{2}g_3 - g_4 + g_5)d \end{bmatrix}$$

如果不考虑显性效应, 即对于模型(2), 线性统计模型(5)可简化为:

$$E(G | x_1, x_2) = \mu + aE(w | x_1, x_2) \triangleq \mu + (a)A_1x_1 + (a)A_2x_2 \quad (6)$$

线性模型(5)和(6)中, 变量 x_1 、 x_2 、 y_1 、 y_2 以及乘积项 x_1x_2 、 y_1y_2 是已知的, 通过对(5)和(6)的求解, 就能获得包含 QTL 位置和效应的参数估计, 实现 QTL 作图的目标。

1.3 多个 QTL 的遗传和统计模型

将上述结果推广于 m 个 QTL 的情形, 加显性模型下个体的基因型值可表示为:

$$G = \mu + \sum_{j=1}^m [a_j w_j + d_j v_j] \quad (7)$$

式中, w_j 和 v_j 是第 j 个 QTL 基因型的指示变量。这时期望遗传型值为:

$$\begin{aligned} E(G) = \mu + \sum_{j=1}^m [(d_j)\mu_{d_j} + (a_j)A_jx_j + (d_j)D_jy_j \\ + (a_j)A_{j+1}x_{j+1} + (d_j)D_{j+1}y_{j+1} + (d_j)AA_{j,j+1}x_jx_{j+1} \\ + (d_j)DD_{j,j+1}y_jy_{j+1}] \triangleq \beta + \sum_{j=1}^{m+1} \lambda_jx_j + \sum_{j=1}^{m+1} \rho_jy_j \\ + \sum_{j=1}^m \lambda\lambda_{j,j+1}x_jx_{j+1} + \sum_{j=1}^m \rho\rho_{j,j+1}y_jy_{j+1} \end{aligned}$$

其中, $\beta = \mu + \sum_{j=1}^m (d_j)\mu_{d_j}$; $\lambda_1 = (a_1)A_1$; $\rho_1 = (d_1)D_1$;

$\lambda_j = (a_{j-1})A_j + (a_j)A_j$ 和 $\rho_j = (d_{j-1})D_j + (d_j)D_j$, 式中, $j = 2, \dots, m$; $\lambda_{m+1} = (a_m)A_{m+1}$; $\rho_{m+1} = (d_m)D_{m+1}$; $\lambda\lambda_{j,j+1} = (d_j)AA_{j,j+1}$, 和 $\rho\rho_{j,j+1} = (d_j)DD_{j,j+1}$, 式中, $j = 1, \dots, m$ 。从而得到表型值与标记变量间的完备线性模型为:

$$\begin{aligned} y = \beta + \sum_{j=1}^{m+1} \lambda_jx_j + \sum_{j=1}^{m+1} \rho_jy_j + \sum_{j=1}^m \lambda\lambda_{j,j+1}x_jx_{j+1} \\ + \sum_{j=1}^m \rho\rho_{j,j+1}y_jy_{j+1} + \varepsilon \end{aligned} \quad (8)$$

式中, y 是数量性状的观测值, ε 是随机误差。

对于只包含 2 种基因型的作图群体, 如回交、重组近交家系、加倍单倍体等, 只包含加性效应(加性效应的含义因群体而异^[20-21,23]), 这时多 QTL 遗传模型和完备线性模型分别为:

$$G = \mu + \sum_{j=1}^m a_jw_j \quad (9)$$

和

$$y = \beta + \sum_{j=1}^{m+1} \lambda_jx_j + \varepsilon \quad (10)$$

1.4 上位型互作 QTL 的遗传和统计模型

上位性是不同位点间基因的互作。上位 QTL 作图目前还没有一个公认的比较完善的方法。杂种 F_2 群体中包含 4 种互作类型的完备线性模型比较复杂, 这里以加性×加性的互作来说明上位性的遗传和统计模型^[22]。加性效应(a_i 表示第 i 个 QTL 加性效应)和加性×加性互作效应(aa_{ij} 表示第 i 个 QTL 和第 j 个 QTL 间的加性×加性互作效应)的遗传模型可表示为:

$$G = \mu + \sum_{j=1}^m a_jw_j + \sum_{j < k} aa_{jk}w_jw_k \quad (11)$$

与模型(11)对应的完备线性模型为:

$$y = b_0 + \sum_{j=1}^{m+1} b_jx_j + \sum_{j < k} b_{jk}x_jx_k + e_i \quad (12)$$

1.5 完备区间作图(ICIM)的加性 QTL 作图假设检验

ICIM 通过一个“两步策略”实现对 QTL 的定位^[21-23]。第一步利用逐步回归估计模型(8)或(10)中的参数; 第二步通过对全基因组的一维扫描寻找加性 QTL[模型(10)]和加显性 QTL[模型(8)]。现以 F_2 群体为例说明一维扫描时的检验方法。假定在标记区间 $(k, k+1)$ 上作图, 首先利用第一步模型选择的结果校正表型值, 即令

$$\begin{aligned} \Delta y_i = y_i - \sum_{j \neq k, k+1} [\hat{\lambda}_jx_{ij} + \hat{\rho}_jy_{ij}] - \sum_{j \neq k} [\lambda\hat{\lambda}_{j,j+1}x_{ij}x_{i,j+1} \\ + \rho\hat{\rho}_{j,j+1}y_{ij}y_{i,j+1}] \end{aligned} \quad (13)$$

如果标记区间上有一个 QTL, 3 种 QTL 基因型 QQ 、 Qq 和 qq 的分布为 $N(\mu_l, \sigma^2)$, 其中 $l = 1, 2, 3$ 代表这 3 种 QTL 基因型。检验 QTL 的备择假设和零假设分别为: $H_A: \mu_1, \mu_2$ 和 μ_3 至少有 2 个互不相等, $H_0: \mu_1 = \mu_2 = \mu_3$ 。备择假设 H_A 下的极大似然函数为:

$$L_A = \sum_{j=1}^9 \sum_{i \in S_j} \log \left[\sum_{l=1}^3 \pi_{jl} f(\Delta y_i; \mu_l, \sigma^2) \right],$$

式中, S_j 代表第 j 个标记基因型的集合($j = 1, 2, \dots, 9$), π_{jl} 是第 j 个标记型下第 l 个 QTL 基因型的概率, $f(x; \mu_l, \sigma^2)$ 表示正态分布 $N(\mu_l, \sigma^2)$ 的密度函数。在 H_0 下, 所有 QTL 基因型服从同样的分布 $N(\mu_l, \sigma^2)$, 这时的样本似然函数为:

$$L_0 = \sum_{i=1}^n \log \left[f(\Delta y_i; \mu_0, \sigma_0^2) \right].$$

因此可通过似然比检验统计量检验二种假设间

的差异显著性。 H_0 下参数的似然估计和最大似然函数容易通过似然方程求得,但 H_A 下参数的似然估计和最大似然函数需要借助迭代的EM算法^[21-24]。

1.6 完备区间作图(ICIM)的上位性 QTL 作图假设检验

第一步与加(显)性 QTL 作图类似,利用逐步回归估计模型(12)中的参数,第二步通过对全基因组的二维扫描寻找两个作图区间[分别位于标记区间($j, j+1$)和($k, k+1$)]之间的 QTL 互作。我们同样以 F_2 群体为例说明 ICIM 上位性作图的检验方法。首先利用第一步模型选择的结果校正表型值,即令

$$\Delta y_i = y_i - \sum_{r \neq j, j+1, k, k+1} \hat{\lambda}_r x_{ir} - \sum_{r \neq j, j+1, k, k+1} \hat{\rho}_r y_{ir} - \sum_{r \neq j, k} \lambda \hat{\lambda}_{r, r+1} x_{ir} x_{i, r+1} - \sum_{r \neq j, k} \rho \hat{\rho}_{r, r+1} y_{ir} y_{i, r+1} \quad (13)$$

如果两区间上分别有一个 QTL,这 2 个 QTL 的 9 种 QTL 基因型,用 $Q_1 Q_1 Q_2 Q_2$, $Q_1 Q_1 Q_2 q_2$, $Q_1 Q_1 q_2 Q_2$, $Q_1 q_1 Q_2 Q_2$, $Q_1 q_1 Q_2 q_2$, $Q_1 q_1 q_2 Q_2$, $Q_1 q_1 q_2 q_2$, $q_1 Q_1 Q_2 Q_2$, $q_1 Q_1 Q_2 q_2$ 和 $q_1 Q_1 q_2 Q_2$ 表示,其分布 $N(\mu_l, \sigma^2)$,其中, $l = 1, 2, \dots, 9$,代表这 9 种 QTL 基因型。检验 QTL 存在的备择假设和零假设分别为: $H_A: \mu_1, \mu_2, \mu_3, \dots, \mu_9$ 至少有 2 个互不相等, $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_9$ 。 H_A 下的极大似然函数为:

$$L_A = \sum_{j=1}^{81} \sum_{i \in S_j} \log \left[\sum_{l=1}^9 \pi_{jl} f(\Delta y_i; \mu_l, \sigma^2) \right]$$

式中, S_j 代表第 j 个标记基因型的集合 ($j = 1, 2, \dots, 81$), π_{jl} 和 $f(x; \mu_l, \sigma^2)$ 符号意义同加—显性 QTL 作图的一维扫描。在 H_0 下,所有 QTL 基因型服从同样的分布 $N(\mu_l, \sigma^2)$,这时的样本似然函数为:

$$L_0 = \sum_{i=1}^n \log \left[f(\Delta y_i; \mu_0, \sigma_0^2) \right]$$

类似加性作图检验,通过似然比检验统计量检验二种假设间的差异显著性。 H_0 下参数的似然估计和最大似然函数依然通过求解似然方程求得, H_A 下参数的似然估计和最大似然函数则需要借助迭代的 EM 算法^[21-24]。根据 L_A 和 L_0 可以得到衡量互作显著性的 LOD 统计量(记做 LOD_A)。值得注意的是,从 LOD_A 中只能知道 9 种基因型均值之间是否存在显著性差异,但差异究竟是由两 QTL 本身的加性或显性效应导致,还是由互作效应导致,却无从获知。

为此,需要进一步考虑如下的备择假设: $H_{AA}: aa = ad = da = dd = 0$ 。 H_{AA} 下的极大似然函数为: $L_{AA} = L_A - (\lambda_1 aa + \lambda_2 ad + (\lambda_3 aa + \lambda_4 dd))$, 式中 λ_1 、 λ_2 、 λ_3

和 λ_4 为拉格朗日乘子。由 L_A 和 L_{AA} 计算得到的 LOD 统计量(记做 LOD_{AA})中去除了 2 个 QTL 自身加—显性效应的影响,只衡量二者互作的显著性。但是 LOD_{AA} 并没有将加加、加显、显加和显显这 4 种互作类型分开检验,如果想进一步检验每种互作的显著性,需要提出类似 LOD_{AA} 的统计量。

2 ICIM 在一个大麦 DH 作图群体中的应用

由亲本 Harrington 和 TR306 衍生的 145 个大麦 (*Hordeum vulgare* L.) DH 群体是国际上一个知名的 QTL 作图群体^[25],利用该群体和 127 个标记建立了均匀覆盖大麦 7 条染色体(用 1H 至 7H 表示)的连锁群。1992—1993 年间,在 17 个地点 25 个环境条件下评价各种数量性状的表现,这里我们用平均粒重(KWT)为例说明 ICIM 的应用。亲本 Harrington 的平均粒重为 38.7 mg, TR306 的粒重为 45.0 mg, DH 群体中的最低、平均和最高粒重分别为 35.8、42.0 和 48.1 mg。

ICIM 第一步变量选择时,采用的变量进、出回归方程的显著水平分别为 0.01 和 0.02; LOD 临界值取 2.5; 第二步的一维扫描时共检测到 9 个影响粒重的 QTL(表 1 和图 1),染色体 2H 上有 3 个, 3H 和 7H 上各有 2 个, 4H 和 5H 上各有 1 个,其中有 8 个也被其他作图方法所发现^[25]。位于 5H 染色体 5.0 cM 上的 $qKWT5H$ 和位于 7H 染色体 95.0 cM 上的 $qKWT7H$ 是加性遗传效应最大的 2 个 QTL,分别解释表型变异的 38.37% 和 17.20%(表 1)。大部分提高粒重的等位基因来源于粒重较高的亲本 TR306,但亲本 Harrington 仍携带有提高粒重的等位基因 $qKWT2H-1$ 、 $qKWT2H-3$ 和 $qKWT3H-2$,这些基因解释了群体中的超亲分离现象。

为了了解不同参数对 ICIM 作图结果的影响,图 1 还给出变量进入模型的显著水平(PIN)分别是 0.05 和 0.001 下的 LOD 曲线,变量离开模型的概率水平(POUT)均设为 PIN 的 2 倍。当 PIN=0.01 和 POUT=0.02,回归模型解释了 80.76% 的表型变异(表 1),已超过粒重的广义遗传力 0.71^[25]。因此基本断定该群体中加性 QTL 是主要的遗传变异。在进行二维扫描时,ICIM 给出 2 个 LOD 值,即 LOD_A (图 2-A) 和 LOD_{AA} (图 2-B), LOD_A 度量所有的遗传变异,而 LOD_{AA} 只度量上位性引起的变异。因此在图 2-A 中,具有显著加性效应的 QTL 在二维 LOD 图中呈现一

表 1 大麦 DH 作图群体中粒重性状的加性 QTL 作图结果(PIN=0.01, POUT=0.02)

Table 1 Nine additive QTL identified by ICIM (PIN=0.01, POUT=0.02) to control kernel weight (KWT) in the barley DH population

QTL 名称 QTL name	染色体上的位置 Position (cM)	LOD 得分 LOD score	加性效应 Additive effect (mg)	表型变异率 ¹⁾ PVE ¹⁾ (%)
<i>qKWT2H-1</i>	83	4.60	0.39	3.13
<i>qKWT2H-2</i>	140	7.23	-0.51	5.34
<i>qKWT2H-3</i>	201	5.59	0.43	3.77
<i>qKWT3H-1</i>	1	4.39	-0.39	3.04
<i>qKWT3H-2</i>	22	7.41	0.51	5.33
<i>qKWT4H</i>	125	4.12	-0.37	2.73
<i>qKWT5H</i>	5	34.28	-1.37	38.37
<i>qKWT7H-1</i>	4	8.27	-0.55	6.07
<i>qKWT7H-2</i>	95	19.81	-0.92	17.20
Total variation explained (%) ²⁾				80.76

¹⁾ PVE: percentage of variance explained.

²⁾ Total phenotypic variation explained by additive effects in the regression of phenotype on markers.

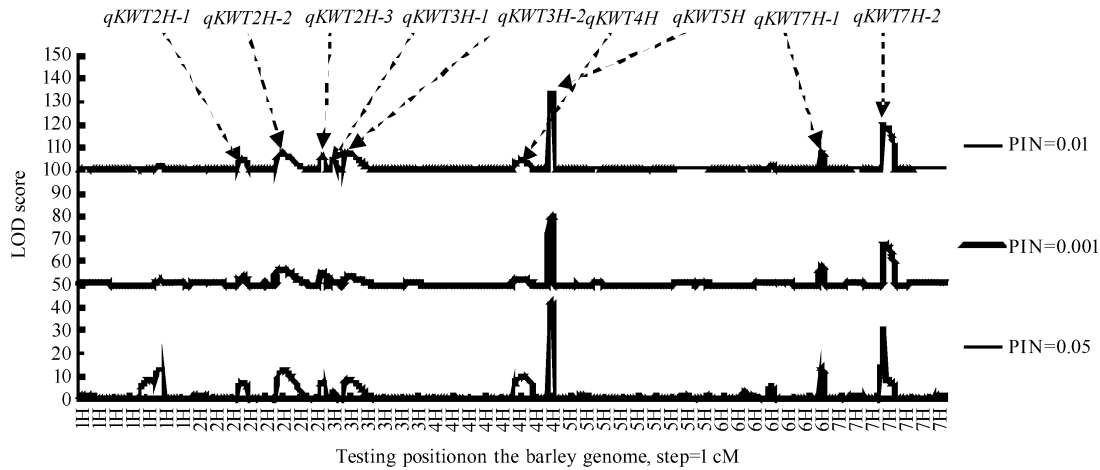


图 1 大麦 145 个 DH 系中对粒重性状 3 种作图参数的 LOD 曲线

Fig. 1 Mapping results from three mapping parameters for additive QTL affecting kernel weight in barley population consisting of 145 DH lines

从上到下, 逐步回归模型选择过程中变量进入的概率水平(PIN)分别为 0.01、0.001 和 0.05, 变量被剔除的概率水平(POUT)为 PIN 的 2 倍。扫描步长为 1 cM。1H 到 7H 表示大麦的 7 条染色体。为清楚起见, PIN=0.001 时的 LOD 值加以 50, PIN=0.01 时的 LOD 值加以 100。

Three probabilities for entering variables and removing variables were considered (i.e., PIN = 0.01, 0.001, 0.05 and POUT = 0.02, 0.002, 0.10, respectively). The scanning step is 1 cM. 1H to 7H represent the seven barley chromosomes. The LOD scores at PIN = 0.001, and 0.01 were added by 50 and 100, respectively.

条带, 从 LOD_A 难以看出上位性的存在。 LOD_{AA} 中已扣除加性 QTL 的影响, 因此在图 2-B 中, 如某一点有显著的 LOD 值, 则说明对应的 2 个坐标轴的位置上存在显著的互作, 如图 2-B 中的互作 1 和互作 2。ICIM 通过二维扫描这种方法不仅能够检测到具有明显加性效应的 QTL 间的互作(如图 2-B 互作 1), 也能检测到没有加性效应的 QTL 间的互作(如图 2-B 互作 2)。

3 完备区间作图的统计性质和优点

从等式(3)和(4)可以看出, QTL 只影响它所在区

间上的标记变量的系数。在完备线性模型(8)和(10)中, 标记的系数是由标记左、右 2 个区间上的 QTL 唯一决定的, 而与其他标记区间上的 QTL 无关。这样在“分隔 QTL”的假定下(即 2 个 QTL 间包含至少一个不存在 QTL 的空白区间), 区间标记的系数包含了该区间上 QTL 位置和效应的所有信息; 通过等式(13)对表型的校正, 排除了当前区间之外的所有 QTL 对当前区间上 QTL 作图的影响, 因而有效地控制了背景遗传变异对当前位置上 QTL 检测的影响。ICIM 包含两个步骤: 首先利用所有标记的信息, 通

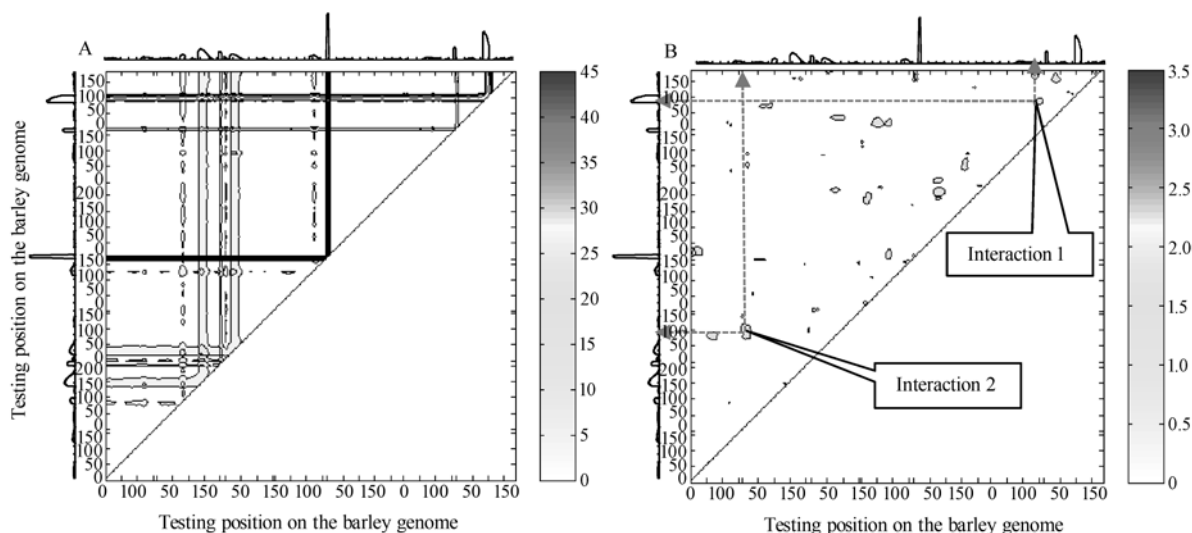


图 2 大麦 145 个 DH 系中同时检验粒重的加性和上位型效应(A)和只检验上位型效应(B)的二维扫描 LOD 曲面图
Fig. 2 Two-dimensional scanning from ICIM testing the significance of additive and epistasis (A), and epistasis only (B) affecting kernel weight (KWT) in barley population consisting of 145 DH lines

A: LOD_A 度量所有的遗传变异; B: LOD_{AA} 只度量上位性引起的变异。图形上方和左侧为一维扫描的 LOD 曲线。

A: LOD_A determines the significance of all genetic variations; B: LOD_{AA} determines the significance of epistatic variation.

LOD profile from one-dimensional scanning is shown at the top and left.

过逐步回归选择重要的标记变量并估计其效应, 然后利用逐步回归得到的线性模型校正表型数据, 并利用校正后的数据进行全基因组的一维和二维扫描。这种作图策略简化了 CIM 中控制背景遗传变异的过程。模拟研究和实际数据的分析表明 ICIM 是一个行之有效的 QTL 定位方法; ICIM 有较低的抽样误差, 较高的作图效率; 有 QTL 的区域 ICIM 有显著高的 LOD 值, 没有 QTL 的区域 ICIM 的 LOD 值接近于 0; ICIM 对作图参数有着很好的稳健性(图 1), 同时也很容易推广到上位性作图^[21-23], 在上位性作图时, 不仅可以检测到有加性效应 QTL 间的互作(图 2-B 中的互作 1), 而且还可以检测到没有明显加性效应的 QTL 之间的互作(图 2-B 中的互作 2)。笔者已编制了实现 ICIM 的交互式用户友好软件 QTL IciMapping, 可从 <http://www.isbreeding.net> 网站免费下载。

References

- [1] Lynch M, Walsh B. Genetic and Analysis of Quantitative Traits. Sunderland, MA: Sinauer Associates, 1998
- [2] Zhai H-Q(翟虎渠), Wang J-K(王健康). Applied Quantitative Genetics (应用数量遗传). Beijing: China Agricultural Sciencetech Press, 2007 (in Chinese)
- [3] Lander E S, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 1989, 121: 185-199
- [4] Broman K W, Speed T P. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J Roy Statist Soc B*, 2002, 64: 641-656
- [5] Carlborg Ö, Kerje S, Schütz K, Jacobsson L, Jensen P, Andersson L. A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res*, 2003, 13: 413-421
- [6] Doerge R W. Mapping and analysis of quantitative trait loci in experiment populations. *Nat Rev Genet*, 2002, 3: 43-52
- [7] Feenstra B, Skovgaard I M, Broman K W. Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. *Genetics*, 2006, 173: 2269-2282
- [8] Sen S, Churchill G A. A statistical framework for quantitative trait mapping. *Genetics*, 2001, 159: 371-387
- [9] Kao C H, Zeng Z B, Teasdale R D. Multiple interval mapping for quantitative trait loci. *Genetics*, 1999, 152: 1203-1206
- [10] Zhang Y, Xu S. A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity*, 2005, 95: 96-104
- [11] Satagopan J M, Yandell B S, Newton M A, Osborn T C. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 1996, 144: 805-816
- [12] Wang H, Zhang Y, Li X, Masinde G, Mohan S, Baylink D, Xu S. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics*, 2005, 170: 465-480
- [13] Xu S, Jia Z. Genome-wide analysis of epistatic effects for quantitative traits in barley. *Genetics*, 2007, 175: 1955-1963
- [14] Frary A N, Nesbitt T C, Frary A M, Grandillo S, Knaap E V D, Cong B, Liu J P, Meller J, Elber R, Alpert K B, Tanksley S D. *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. *Science*, 2000, 289: 85-88
- [15] Xue W, Xing Y, Weng X, Zhao Y, Tang W, Wang L, Zhou H, Yu S, Xu C, Li X, Zhang Q. Natural variation in *Ghd7* is an important

- regulator of heading date and yield potential in rice. *Nat Genet*, 2008, 40: 761–767
- [16] Wan X, Weng J, Zhai H, Wang J, Liu X, Guo T, Su N, Wan J. QTL analysis for rice grain width and fine mapping of an identified QTL allele *gw-5* in a recombination hotspot region on chromosome 5. *Genetics*, 2008, 179: 2239–2252
- [17] Wang J, Wan X, Li H, Pfeiffer W, Crouch J, Wan J. Application of identified QTL-marker associations in rice quality improvement through a design breeding approach. *Theor Appl Genet*, 2007, 115: 87–100
- [18] Wang J, Wan X, Crossa J, Crouch J, Weng J, Zhai H, Wan J. QTL mapping of grain length in rice (*Oryza sativa* L.) using chromosome segment substitution lines. *Genet Res*, 2006, 88: 93–104
- [19] Haley C S, Knott S A. A simple regression method for mapping quantitative loci in line crosses using flanking markers. *Heredity*, 1992, 69: 315–324
- [20] Zeng Z B. Precision mapping of quantitative trait loci. *Genetics*, 1994, 136: 1457–1468
- [21] Li H, Ye G, Wang J. A modified algorithm for the improvement of composite interval mapping. *Genetics*, 2007, 175: 361–374
- [22] Li H, Ribaut J M, Li Z, Wang J. Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theor Appl Genet*, 2008, 116: 243–260
- [23] Zhang L, Li H, Li Z, Wang J. Interactions between markers can be caused by the dominance effect of QTL. *Genetics*, 2008, 180: 1177–1190
- [24] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc B*, 1977, 39: 1–38
- [25] Tinker N A, Mather D E, Rossmagel B G, Kasha K J, Kleinhofs A, Hayes P M, Falk D E, Ferguson T, Shugar L P, Legge W G, Irvine R B, Choo T M, Briggs K G, Ullrich S E, Franckowiak J D, Blake T K, Graf R J, Dofing S M, Saghai-Marooof M A, Scoles G J, Hoffman D, Dahleen L S, Kilian A, Chen F, Biyashev R M, Kudrna D A, Steffenson B J. Regions of the genome that affect agronomic performance in two-row barley. *Crop Sci*, 1996, 36: 1053–1062