DOI: 10.3724/SP.J.1006.2009.01569

基于贝叶斯统计的谷物胚乳性状 QTL 多区间作图方法

王亚民^{1,2} 汤在祥¹ 陆鑫¹ 徐辰武^{1,*}

¹扬州大学江苏省作物遗传生理重点实验室 / 教育部植物功能基因组学重点实验室, 江苏扬州 225009; ²连云港职业技术学院基础部, 江苏连云港 222006

摘 要: 贝叶斯统计学已被广泛地应用在现代科学的各个研究领域。本研究将贝叶斯统计方法和谷物三倍体胚乳性状数量遗传模型相结合, 以 F_2 群体中各植株的分子标记基因型以及植株上若干粒自交种子胚乳性状的单粒观测值为数据模式, 提出了胚乳数量性状基因座(QTL)多区间作图的贝叶斯方法。该方法首先构建胚乳性状的多区间多 QTL 遗传模型, 然后通过基于 Gibbs 抽样和 Metropolis-Hastings 算法实现的马尔可夫链蒙特卡罗(MCMC)方法同时获得多个 QTL 效应和位置的估计。方法的有效性通过一条长染色体的模拟实验进行了验证, 结果表明, 本文提出的贝叶斯多区间方法能够准确地估计胚乳性状 QTL 的位置和效应, 并可有效区分两种显性效应。

关键词: 贝叶斯统计; 胚乳性状; 马尔可夫链蒙特卡罗; 数量性状基因座位

Bayesian Statistics-Based Multiple Interval Mapping of QTL Controlling Endosperm Traits in Cereals

WANG Ya-Min^{1,2}, TANG Zai-Xiang¹, LU Xin¹, and XU Chen-Wu^{1,*}

¹ Jiangsu Provincial Key Laboratory of Crop Genetics and Physiology / Key Laboratory of Plant Functional Genomics of Ministry of Education, Yangzhou University, Yangzhou 225009, China; ² Basis Course of Lianyungang Technical College, Lianyungang 222006, China

Abstract: The endosperm of plants is a major source of food, feed and industrial raw materials. The genetic analysis of endosperm traits poses numerous challenges due to its complex genetic composition and unique physical and developmental properties. Modern molecular techniques and statistical methods have greatly improved the mapping of quantitative trait loci (QTL) underlying endosperm traits. In recent years, Bayesian statistics-based analyzing methods have been developed for mapping QTL underlying diploid quantitative traits, but these methods have not been effective to the mapping of triploid endosperm characters. On the basis of Bayesian statistics and quantitative genetic model of triploid endosperm traits, a Bayesian multiple interval method for mapping QTL underlying endosperm traits was proposed. This method used the DNA molecular marker genotypes of each plant in F_2 segregation population and the single endosperm observation of a few endosperms of each plant as data set to analyze endosperm QTL. After constructing the multiple-QTL model, the Bayesian estimates of multiple QTL position and effects were obtained through MCMC algorithm implementing via Gibbs and Metropolis-Hastings sampling. The validation of the statistical procedure was verified through chromosome level simulation studies. The results showed that the proposed Bayesian method can estimate the multiple QTL positions and effects as well as distinguish the two dominance effects.

Keywords: Bayesian statistics; Endosperm traits; Markov chain Monte Carlo; Quantitative trait loci

谷类作物品质性状的遗传改良自从 20 世纪 90 年代就已成为国际性的一大育种目标。这类性状主要是指受三倍体遗传控制的胚乳数量性状。因此, 了解胚乳品质性状的遗传基础是实现谷类作物品质改良的前提。莫惠栋^[1-2]提出了胚乳性状的数量遗传模型, 徐辰武等^[3]根据三倍体胚乳性状的数量遗传模型, 率先提出胚乳性状 QTL 图谱的构建方法。其

后, 国内外学者先后提出了一系列基于经典数理统计的分析方法以定位控制胚乳数量性状的 QTL, 如 Wu 等^[4-5]提出的基于 EM 算法的极大似然法和基于一阶设计和二阶设计的作图方法, Xu 等^[6]提出的迭代重新加权最小平方法。这些方法均是基于单 QTL 模型发展起来的, 当在同一连锁群上存在多个 QTL, 以及 QTL 间存在互作时, 这些方法对有关 QTL 的估

本研究由国家重点基础研究发展规划(973 计划)项目(2006CB101700)和教育部“新世纪优秀人才支持计划”项目(NCET-05-0502)资助。

* 通讯作者(Corresponding author): 徐辰武, E-mail: qtls@yzu.edu.cn; Tel: 0514-87979358

第一作者联系方式: E-mail: qtls@hotmail.com

Received(收稿日期): 2008-12-18; Accepted(接受日期): 2009-04-25.

计可能出现偏差。为此, Kao 等^[7]发展出多区间作图方法以解决这一问题。种子胚乳作为着生在母体植株上的子代, 其遗传表达可能还受母体植株基因型的影响, 为分析母体基因型的效应, Cui 等^[8]发展出用于同时分析母体和子代效应的遗传模型和分析方法, 该方法假设存在亲代和子代两套遗传体系共同控制胚乳性状的遗传表达。而 Hu 等^[9]认为母体 QTL 和子代 QTL 应属于同一套遗传体系, 进而发展出包含母体遗传效应的胚乳 QTL 定位方法。值得注意的是, 上述这些方法在株平均值数据模式下并不能分辨出胚乳性状的两个显性效应, 为解决这一问题, Wen 等^[10]提出基于随机交配设计的胚乳性状 QTL 区间定位方法, 王学枫等^[11]用经典数量遗传学的 NCIH 和 TTC 设计试图精确估计两种显性效应。此外, 为考虑胚乳性状不同位点间的上位性互作, He 和 Zhang^[12]提出基于随机交配设计的惩罚似然方法可对胚乳性状上位性 QTL 进行分析。

上述方法均是基于经典数理统计发展出来的。近年来, 随着高性能计算机的快速发展以及 MCMC 算法的提出, 有别于经典数理统计的贝叶斯统计方法在科学研究的各个领域获得了广泛的应用。与此同时, 在统计基因组学研究中已先后发展出一系列的 QTL 定位的贝叶斯方法^[13-16]。然而如何将贝叶斯方法应用于胚乳性状的 QTL 作图, 国内外均未见报道。在基于单 QTL 模型的胚乳性状 QTL 区间作图的贝叶斯方法的基础上^[17], 本文进一步将胚乳性状单 QTL 模型扩展为多 QTL 模型, 并发展出基于 MCMC 算法实现的贝叶斯方法和程序, 以供实际数据分析应用。

1 原理与方法

1.1 胚乳性状统计遗传模型

假定控制某一胚乳性状的 QTL 有 p 个, 依次位于基因组上 p 个标记区间内。由于双亲杂交衍生分离群体的任一 QTL 仅涉及两个等位基因, 故任一胚乳 QTL 有 4 种可能的基因型 $Q_mQ_mQ_m$ 、 $Q_mQ_mq_m$ 、 $Q_mq_mq_m$ 和 $q_mq_mq_m$ ($m = 1, 2, \dots, p$), 其基因型值分别为 $3a_m/2$ 、 $a_m/2+d_{1m}$ 、 $-a_m/2+d_{2m}$ 和 $-3a_m/2$ ^[2]。其中 a_m 为加性效应, d_{1m} 和 d_{2m} 分别为第一和第二显性效应。设 Y_{ij} 表示 F_2 分离群体中第 i 个植株 ($i = 1, 2, \dots, k$) 上的第 j 粒自交种子 ($j = 1, 2, \dots, n_i$) 胚乳表型值。则 Y_{ij} 具有如下统计模型。

$$Y_{ij} = \mu + \sum_{m=1}^p (a_m X_{1ijm} + d_{1m} X_{2ijm} + d_{2m} X_{3ijm}) + e_{ij} \quad (1)$$

式中, μ 为群体均值, X_{1ijm} 、 X_{2ijm} 和 X_{3ijm} 分别是第 i 植株上第 j 粒种子胚乳在基因组第 m 个标记区间内 QTL 基因型的指示变量, 具有如下定义: 对 $Q_mQ_mQ_m$ 基因型, $X_{1ijm}=3/2$, $X_{2ijm}=X_{3ijm}=0$; 对 $Q_mQ_mq_m$ 基因型, $X_{1ijm}=1/2$, $X_{2ijm}=1$, $X_{3ijm}=0$; 对 $Q_mq_mq_m$ 基因型, $X_{1ijm}=-1/2$, $X_{2ijm}=0$, $X_{3ijm}=1$; 对 $q_mq_mq_m$ 基因型, $X_{1ijm}=-3/2$, $X_{2ijm}=X_{3ijm}=0$ 。 $e_{ij} \sim N(0, \sigma_0^2)$, σ_0^2 为剩余方差。令:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \dots & \mathbf{X}_{1p} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & \dots & \mathbf{X}_{2p} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{X}_{k1} & \mathbf{X}_{k2} & \dots & \mathbf{X}_{kp} \end{bmatrix}$$

$$\text{其中 } \mathbf{X}_{im} = \begin{bmatrix} X_{i1m} & X_{i2m} & X_{i3m} \\ X_{i1m} & X_{i2m} & X_{i3m} \\ \vdots & \vdots & \vdots \\ X_{i1m} & X_{i2m} & X_{i3m} \end{bmatrix}_{n_i \times 3}$$

记

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1n_1} & Y_{21} & Y_{22} & \dots & Y_{2n_2} & \dots \\ Y_{k1} & Y_{k2} & \dots & Y_{kn_k} \end{bmatrix}_{1 \times n}^T$$

$$\mathbf{b} = \begin{bmatrix} a_1 & d_{11} & d_{21} & a_2 & d_{12} & d_{22} & \dots \\ a_p & d_{1p} & d_{2p} \end{bmatrix}_{1 \times 3p}^T$$

$$\mathbf{e} = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1n_1} & e_{21} & e_{22} & \dots & e_{2n_2} & \dots \\ e_{k1} & e_{k2} & \dots & e_{kn_k} \end{bmatrix}_{1 \times n}^T$$

其中, T 表示矩阵的转置, $n = \sum_{i=1}^k n_i$ 。则模型(1)可以进一步表示为:

$$\mathbf{Y} = \mu \mathbf{I} + \mathbf{Xb} + \mathbf{e} \quad (2)$$

式中, 的 $\mathbf{I}$ 为 n 行元素皆为 1 的列向量。

1.2 模型中参数的条件后验分布

模型(2)中待估参数(μ , $\mathbf{b}$ 和 σ_0^2)较多, 需采用贝叶斯压缩估计方法进行估计。参照 Wang 等^[13]定义各参数的先验分布为 $p(\mu) \propto 1$ 、 $p(\sigma_0^2) \propto 1/\sigma_0^2$ 和 $p(b_l) = N(0, \sigma_l^2)$, 其中 b_l 为矩阵 $\mathbf{b}$ 的第 l 行元素 ($l=1, 2, \dots, 3p$)。在贝叶斯压缩估计中, 需进一步定义 σ_l^2 具先验分布 $p(\sigma_l^2) \propto 1/\sigma_l^2$, 并记 $\mathbf{v} = (\sigma_0^2 \ \sigma_1^2 \ \sigma_2^2 \ \dots \ \sigma_{3p}^2)$ 。然而, 值得注意的是, 模型(2)中的 $\mathbf{X}$ 并不已知, 需要根据各 QTL 的位置 $\lambda = [\lambda_1 \ \lambda_2 \ \dots \ \lambda_p]_{1 \times p}^T$ 以及分子标记基因型矩阵 $\mathbf{M}$ 联合推断。因此, 所有未知变量包括 $\theta = \{\mu, \mathbf{b}, \mathbf{v}, \lambda, \mathbf{X}\}$ 。

根据贝叶斯原理, 在 $\mathbf{Y}$ 和 $\mathbf{M}$ 已知情况下, θ 的后

验分布可表示为:

$$p(\theta|Y, M) = \frac{p(Y, M, \theta)}{p(Y, M)} \propto p(Y, M, \theta) = p(Y, M|\theta)p(\theta) \quad (3)$$

其中, $p(\theta)$ 为各未知变量的联合先验分布, $p(Y, M|\theta)$ 为观察值向量的似然函数。公式(3)是贝叶斯推断的核心, 据之可导出有关未知变量的条件后验分布。

$$1) \mu | \dots \sim N(\bar{\mu}, s_0^2)$$

其中

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} [Y_{ij} - \sum_{m=1}^p (a_m X_{1ijm} + d_{1m} X_{2ijm} + d_{2m} X_{3ijm})],$$

$$s_0^2 = \sigma_0^2 / n。$$

$$2) b_l | \dots \sim N(\bar{b}_l, s_l^2)$$

其中, $\bar{b}_l = (X_l^T X_l + \sigma_0^2 / \sigma_l^2)^{-1} [X_l^T (Y - \mu I - X_{-l} b_{-l})]$, $s_l^2 = (X_l^T X_l + \sigma_0^2 / \sigma_l^2)^{-1} \sigma_0^2$ 。 X_l 为矩阵 X 的第 l 列向量, X_{-l} 为矩阵 X 去掉第 l 列向量后所得到的矩阵, b_{-l} 为矩阵 b 去掉第 l 行元素后所得到的列向量。

$$3) \sigma_0^2 | \dots \sim \text{Inv} - \chi^2[n, (Y - \mu I - Xb)^T (Y - \mu I - Xb)]$$

$$4) \sigma_l^2 | \dots \sim \text{Inv} - \chi^2(1, b_l^2)$$

上述 1) 至 4) 式中的“...”表示除相应参数外的所有参数 θ 。根据各参数的相应条件后验分布可更新有关参数。

5) 矩阵 X 的更新较为复杂, 其方法如下:

记第 i 个植株在基因组第 m 个 QTL (相应位置为 λ_m) 的双侧标记基因型为 M_{im}^l 和 M_{im}^r 。根据 Haldane 作图函数, 计算出该 QTL 与 M_{im}^l 和 M_{im}^r 之间的重组率, 并进一步计算该 QTL 3 种可能植株基因型 $Q_m Q_m$ 、 $Q_m q_m$ 和 $q_m q_m$ 的先验概率 $p_{1im} = p(Q_m Q_m | \lambda_m, M_{im}^l, M_{im}^r)$ 、 $p_{2im} = p(Q_m q_m | \lambda_m, M_{im}^l, M_{im}^r)$ 和 $p_{3im} = p(q_m q_m | \lambda_m, M_{im}^l, M_{im}^r)$ 。需要注意的是, 在计算 QTL 的先验概率时, 如该 QTL 的双侧标记中存在不完全信息标记或标记信息缺失情况, 可采用 Jiang 和 Zeng^[18] 的方法进行处理。在此基础上, 可得该植株上第 j 粒种子胚乳 QTL 的 4 种基因型 $Q_m Q_m Q_m$ 、 $Q_m Q_m q_m$ 、 $Q_m q_m q_m$ 和 $q_m q_m q_m$ 的先验概率分别为 $p_{1ijm} = p_{1im} + p_{2im}/4$ 、 $p_{2ijm} = p_{3im} + p_{2im}/4$ 和 $p_{4ijm} = p_{3im} + p_{2im}/4$ 。根据 4 种胚乳基因型的先验概率以及胚乳表型值 Y_{ij} 可计算出该粒种子 4 种胚乳基因型的后验概率为:

$$p_{hijm}^* = p_{hijm} f_{hijm}(Y_{ij} | G_h) / \sum_{h=1}^4 p_{hijm} f_{hijm}(Y_{ij} | G_h)$$

式中, $G_h (h = 1, 2, 3, 4)$ 分别代表胚乳 4 种可能基因型,

$f_{hijm}(Y_{ij} | G_h)$ 为 Y_{ij} 的条件概率密度。由 p_{hijm}^* 可确定该粒种子胚乳基因型。该植株上 n_i 粒胚乳的 QTL 基因型均以此方法确定, 从而可得矩阵 X_{im} , 进而得到 X 。

6) QTL 位置 λ 的更新

记第 m 个 QTL 双侧标记的位置分别为 ζ_l 和 ζ_r , 并假定 λ_m 的先验分布为 $\lambda_m \sim U(\zeta_l, \zeta_r)$ 。则有 $P(\lambda_m) = 1/(\zeta_r - \zeta_l)$, 其条件后验分布 $p(\lambda_m | \dots)$ 无明确解析式, 因此, 参数 λ_m 无法像其它参数那样进行 Gibbs 抽样, 而需采用 Metropolis-Hastings 算法实现。首先, 从以 λ_m 为中心, c 为半径的区间内按均匀分布抽取一个候选位置 λ_m^* , 其中 c 是一个给定的常数, 取 1 cM 或 2 cM, 则在 λ_m 已知情况下获得 λ_m^* 的条件概率密度为:

$$q(\lambda_m^* | \lambda_m) = 1/[(\lambda_m + c) - (\lambda_m - c)] = 1/2c$$

当获得 λ_m^* 时, 必须确保有同样的概率密度抽取到原来的位置 λ_m , 因此,

$$q(\lambda_m | \lambda_m^*) = 1/[(\lambda_m^* + c) - (\lambda_m^* - c)] = 1/2c$$

故有 $q(\lambda_m^* | \lambda_m) = q(\lambda_m | \lambda_m^*)$ 。采用 Metropolis-Hastings 算法来决定是否接受新位置 λ_m^* , 该法以 $\min(1, \alpha)$ 的概率来接受新位置 λ_m^* , 其中:

$$\alpha = \frac{\prod_{i=1}^k \prod_{j=1}^{n_i} f(Y_{ij} | \lambda_m^*)}{\prod_{i=1}^k \prod_{j=1}^{n_i} f(Y_{ij} | \lambda_m)} \frac{p(\lambda_m^*) q(\lambda_m | \lambda_m^*)}{p(\lambda_m) q(\lambda_m^* | \lambda_m)} \quad (4)$$

式中, $f(Y_{ij} | \lambda_m^*)$ 和 $f(Y_{ij} | \lambda_m)$ 分别为 λ_m^* 和 λ_m 下 Y_{ij} 的概率密度函数。由于任一粒胚乳可能有 4 种基因型, 故 Y_{ij} 为一混合分布。显然, 当 $\alpha \geq 1$ 时, 将百分之百接受新位置 λ_m^* , 当 $\alpha < 1$ 时, 将以概率 α 接受新位置 λ_m^* 。一旦新位置被接受, λ_m 就被 λ_m^* 所替换。如果新位置 λ_m^* 被拒绝, 则原来位置 λ_m 保留不变。

但值得注意的是, 在一些特殊情况下, $q(\lambda_m^* | \lambda_m) \neq q(\lambda_m | \lambda_m^*)$ 。这是由于当 QTL 位置 λ_m 靠近区间边界以至于 $\lambda_m - \zeta_l = d_m < c$ 时, 新位置 λ_m^* 则不得不在区间 $[\lambda_m - d_m, \lambda_m + c]$ 内按均匀分布抽取一个新位置 λ_m^* , 即 $\lambda_m^* \sim U[\lambda_m - d_m, \lambda_m + c]$; 同理, 如果 λ_m 非常接近右侧标记 ζ_r 以至于 $\zeta_r - \lambda_m = d_m < c$ 时, 新位置 λ_m^* 不得不按 $\lambda_m^* \sim U[\lambda_m - c, \lambda_m + d_m]$ 进行抽取。无论上述哪种情况, 概率密度 $q(\lambda_m^* | \lambda_m)$ 与 $q(\lambda_m | \lambda_m^*)$ 都必须修改, 正确的形式应该为:

$$q(\lambda_m^*|\lambda_m) = \begin{cases} 1/[c+(\lambda_m-\zeta_l)] & \text{当 } \lambda_m-\zeta_l < c \\ 1/[c+\zeta_r-\lambda_m] & \text{当 } \zeta_r-\lambda_m < c \\ 1/2c & \text{其他情况} \end{cases}$$

$$q(\lambda_m|\lambda_m^*) = \begin{cases} 1/[c+(\lambda_m^*-\zeta_l)] & \text{当 } \lambda_m^*-\zeta_l < c \text{ 时} \\ 1/[c+(\zeta_r-\lambda_m^*)] & \text{当 } \zeta_r-\lambda_m^* < c \text{ 时} \\ 1/2c & \text{其他情况} \end{cases}$$

若记 δ_m 、 δ_m^* 分别为 λ_m 、 λ_m^* 与区间最近的边界的距离, 则上面形式可进一步写成 $q(\lambda_m^*|\lambda_m)=1/[c+\min(c, \delta_m)]$, $q(\lambda_m|\lambda_m^*)=1/[c+\min(c, \delta_m^*)]$ 。又因为 $p(\lambda_m)$ 和 $p(\lambda_m^*)$ 都是同一区间上均匀分布的先验概率密度, 所以 $p(\lambda_m^*)/p(\lambda_m)=1$ 。因此, 公式(4)可改写为:

$$\alpha = \frac{\prod_{i=1}^k \prod_{j=1}^{n_i} f(Y_{ij}|\lambda_m^*)}{\prod_{i=1}^k \prod_{j=1}^{n_i} f(Y_{ij}|\lambda_m)} \frac{c + \min(c, \delta_m^*)}{c + \min(c, \delta_m)}$$

1.3 MCMC 过程

除 QTL 位置 λ 需由上述 Metropolis-Hastings 算法实现更新外, 其余参数均有确定的条件后验分布, 可直接依据相应分布进行 Gibbs 抽样, 由此, MCMC 抽样过程简述如下:

- 1) 给未知变量赋初值 $\mu^{(0)}$, $\mathbf{b}^{(0)}$, $\sigma_0^{2(0)}$, $\mathbf{v}^{(0)}$, $\mathbf{X}^{(0)}$, $\lambda_1^{(0)}, \dots, \lambda_p^{(0)}$ 。
- 2) 从 $N(\bar{\mu}, s_0^2)$ 抽取随机数, 记作 $\mu^{(1)}$, 更新 $\mu^{(0)}$ 。
- 3) 从 $N(\bar{b}_l, s_l^2)$ 抽取随机数, 记作 $b_l^{(1)}$, 更新 $b_l^{(0)}$ 。
- 4) 从 $Inv-\chi^2[n, (\mathbf{Y}-\mu\mathbf{I}-\mathbf{Xb})^T(\mathbf{Y}-\mu\mathbf{I}-\mathbf{Xb})]$ 抽取随机数, 记作 $\sigma_0^{2(1)}$, 更新剩余方差 $\sigma_0^{2(0)}$ 。
- 5) 从 $Inv-\chi^2(1, b_l^2)$ 中抽取随机数, 记作 $\sigma_l^{2(1)}$, 更新 $\sigma_l^{2(0)}$ 。
- 6) 更新 QTL 位置 λ 。
- 7) 更新 QTL 的基因型指示变量 $\mathbf{X}^{(0)}$ 。
- 8) 重复 2)~7) t 轮。

对收集到的 t 轮后验样本, 剔除前面若干轮未收敛的数据, 然后按一定间隔收集后验样本, 依据该后验样本分布特征, 选取平均数或众数作为相应参数的贝叶斯估计值。

2 模拟研究

2.1 模拟设置

设 F_2 分离群体的样本容量为 200 株, 每株考察 30 粒。假设一条长 100 cM 的染色体, 其上均匀分布

11 个共显性分子标记, 控制胚乳性状的 3 个 QTL 分别位于 15、55 和 95 cM 处, 群体均值设为 20, 各个 QTL 的遗传力及效应设置见表 1。QTL 总遗传方差 σ_G^2 为各个 QTL 遗传方差 σ_g^2 的累加, 其中, 单个 QTL 遗传方差 $\sigma_g^2 = 5a^2/4 + (ad_1 - ad_2)/4 + (3d_1^2 + 3d_2^2 - 2d_1d_2)/16$ 。根据 QTL 总遗传力 H^2 以及 QTL 总遗传方差, 确定剩余方差 $\sigma_E^2 = \sigma_G^2(1-H^2)/H^2$ 。

重复模拟 100 次, QTL 的统计功效以检测到 QTL 的次数表示, 位置和效应估计的准确度以 100 个重复样本相应 QTL 位置和效应估计值的平均值度量, 精确度则以 100 个重复样本相应 QTL 估计值的标准差度量。对每一样本 QTL 位置和效应的估计, 第一步, 收集后验样本, 设定循环次数 20 000, 去除最初的 2 000 轮, 在剩下的 18 000 轮中每间隔 20 个取样, 共收集 900 个后验数据。第二步, 将收集的 900 个 QTL 位置数据做成方柱图, 并计算该分布的平均数, 作为 QTL 位置的估计值 $\hat{\lambda}$ 。进一步计算方柱图中 $\hat{\lambda}$ 所在组区间内对应样本的加性效应和显性效应估计值的平均数, 即作为相应 QTL 的加性效应和显性效应估计值。

2.2 模拟结果

本模拟共 10 个标记区间, 故可能 QTL 有 10 个。每一 QTL 均收集 900 个位置数据。用 1 cM 为组距绘制 QTL 位置数据的次数分布。图 1-A 为某一随机样本下 10 个可能 QTL 位置数据的频次分布图, 由之可见在设定的 3 个 QTL 区间均出现明显的峰, 表明该样本的 3 个 QTL 均被发现。但由于每一标记区间长 10 cM, 按 1 cM 分组可分 10 个组区间, 900 个后验数据落在每一组区间的平均次数仅为 90 次, 为了清晰反映目标 QTL 所在的染色体位置, 参照 Wang 等^[13]以及 Xu 和 Yi^[19]的处理方法, 本文采用一种 QTL 位置的加权频次分布图。即先计算每 1 组区间后验样本相应的加性和两个显性效应平均值, 然后依据单个 QTL 的遗传方差估计公式计算出 σ_g^2 , 以之绘制出加权的 QTL 位置的频次分布图。图 1-B 为该样本 QTL 位置的加权频次分布图。显然加权的频次分布图可更清晰地展示 QTL 的存在与否及其相应位置。按上述分析, 可得 3 个 QTL 位置估计的平均值和标准差, 分别为 15.98±0.07、55.01±0.09 和 95.01±0.09。相应的加性效应估计值分别为 1.00±0.03、0.87±0.02 和 0.79±0.01; 第一显性效应估计值分别为 -0.53±0.47、0.48±0.45 和 -0.01±0.08; 第

二显性效应估计值分别为 -0.95 ± 0.57 、 2.25 ± 0.43 和 0.02 ± 0.09 。以此类推,可以得到其他 99 个样本的 QTL 位置和效应估计值,并进而得到各 QTL 的统计功效以及位置与效应估计值的平均值与标准差(表 1)。可

见本文提出的贝叶斯多区间方法能够较好地定位控制胚乳性状的 QTL 的位置,对群体均值和加性效应的估计也十分准确,相比而言,两个显性效应的估计略有偏差。

图 1 QTL 位置的后验分布图
Fig. 1 Posterior frequency distribution of QTL position

表 1 3 个 QTL 的统计功效及其位置与效应估计的平均值和标准差
Table 1 Powers, means and standard deviations of three simulated QTLs

QTL	遗传力 h^2 (%)	统计功效 Power (%)	QTL 位置 Position (cM)		群体均值 Mean		a		d_1		d_2	
			True	Estimated	True	Estimated	True	Estimated	True	Estimated	True	Estimated
<i>qtl1</i>	14.04	100	15	14.81±0.27			1	0.90±0.02	-1	-1.09±0.37	-2	-0.85±0.29
<i>qtl2</i>	10.82	100	55	54.91±0.32	20	20.32±0.16	1	0.85±0.03	1	0.48±0.08	2	1.39±0.33
<i>qtl3</i>	5.14	100	95	95.09±0.52			0.8	0.79±0.02	0	0.01±0.08	0	-0.06±0.08

3 讨论

三倍体胚乳性状的遗传研究是近 10 多年来谷物品质研究中较引人注意的一个重要课题。与常规的二倍体性状相比,胚乳性状有着更为复杂的遗传结构,采用解析二倍体性状的统计方法,如区间作图,复合区间作图法显然并不合适。徐辰武等^[3]首次报道了利用分子标记对胚乳性状进行 QTL 图谱构建的双侧标记基因型均值回归方法。其后,国内外学者先后提出了一系列的模型和方法,如极大似然法^[4,20-21],迭代重新加权最小平方方法^[6]等,以解析胚乳数量性状的遗传结构。但这些方法均是基于经典的数理统计学原理而发展起来的。近来,由于贝叶斯统计方法与经典数理统计方法相比,能更直接地解决问题,且可有效整合部分先验信息,已愈来愈引起研究者的重视,并已被成功地应用于二倍体性状的 QTL 作图^[13-16]。作者在前文^[17]中,率先将贝叶斯统计原理和胚乳性状的数量遗传模型相结合,以分离群体中各植株的分子标记基因型以及植株上若干粒种子胚乳性状的单粒观测值为数据模式,提出胚

乳性状 QTL 区间作图的贝叶斯方法。该方法在单 QTL 模型的假定下,可估计胚乳性状 QTL 的位置和效应,并同时区分两种显性效应。然而,控制胚乳性状的 QTL 通常并非 1 个,因此这种基于单 QTL 模型的区间作图方法有待进一步拓展。本文即是在这一背景下进一步将胚乳性状单 QTL 模型扩展为多 QTL 模型,并发展出基于 MCMC 算法的贝叶斯方法和程序。

模拟结果显示,贝叶斯方法可以高效发现 QTL 并准确估计出 QTL 的遗传位置。例如,即使是遗传力只有 5%左右的 QTL,其统计功效也达 100%,这说明在本研究的模拟设置(200 个 F_2 单株,每株 30 粒胚乳)下利用贝叶斯统计对发现 QTL 是非常有效的。此外,从 QTL 效应估计的准确度以及精确度来看,只有两个显性效应的准确度略差。为了分析其原因,进一步对有关参数的条件后验样本分布进行了分析。图 2~图 4 分别展示了 QTL 位置、群体均值以及各 QTL 效应的后验样本收敛情况,除两个显性效应外,所有参数的后验样本分布平稳,且接近真值。两个显性效应的后验分布虽然平稳,但较真值略有偏

差和混杂。这一结论与 Wu 等^[5]、Kao^[7]以及 Wen 和 Wu^[22]研究结果一致。这一方面可能与模拟试验采用的群体有关, 由于本文采用一阶设计的 F_2 群体, F_3 胚乳 QTL 基因型以其所着生的母株 QTL 基因型推断, 由于世代的不对应, 必然会造成一定的信息丢失。另一方面, 即使加性和显性效应在量值上相等, 显性效应引起的变异在胚乳性状遗传方差中所占分量仍然相对很小。因此, 必然会带来对显性效应估计在准确性和精确性上的偏差。为解决这一问题, Wu 等^[5]和 Kao^[7]提出利用二阶设计, 即同时利用 F_2 植株和子代胚两代分子标记基因型联合推断子代胚乳 QTL 基因型, 该设计比一阶设计提供了更多的遗传信息, 明显优于一阶设计, 但却带来标记分析成本的提高。Wen 等^[10]提出随机交配群体来实现胚乳第一和第二显性效应的无偏估计。王学枫等^[11]新近发展了基于 NCIH 和 TTC 设计的胚乳性状 QTL 区间作图方法, 结果表明两种设计均有较高的 QTL 统计功效, 在供试的大多数处理下, 即使 QTL 的遗传力只有 5%, 其被发现能力也可高达 100%。两种设计也均可有效区分胚乳 QTL 的各种遗传效应, 特别是区分两种显性效应。此外, Wen 和 Wu^[23]提出利用 F_2 或 BC_1 种子的胚乳表现型以及各粒种子胚的分子标记基因型进行胚乳性状 QTL 定位, 不仅可消除母体效应和环境效应的影响, 还能准确估计各个遗传效应。因此, 将贝叶斯统计方法与这些遗传设计相结合必然能解决贝叶斯框架下两种显性效应估计的精

图 2 3 个 QTL 位置后验样本链
Fig. 2 Chain of posterior sample of three QTL positions

图 3 群体均值后验样本链
Fig. 3 Chain of posterior sample of population mean

图 4 3 个 QTL 效应后验样本链
Fig. 4 Chain of posterior sample of three QTL effects
黑色表示加性效应, 红色表示第一显性效应, 蓝色表示第二显性效应。
Black, red and blue lines denote a , d_1 , and d_2 , respectively

度问题。事实上, 两种显性效应的准确估计, 是建立在胚乳 QTL 的两种杂合基因型 QQq 和 Qqq 可以有效识别的基础上, 如能通过特种遗传设计结合分子标记分析, 区分出两种杂合基因型, 则不论显性效应量值的大小, 均可实现准确估计, 这一研究将是胚乳 QTL 分析的发展方向之一。

需要指出的是, 超负荷的运算量是贝叶斯方法应用的制约瓶颈。本研究在设定 200 个植株, 每株 30 个观测值, 10 个标记区间, 共 20 000 轮循环情况下, 在主频 3.0 GH、内存 1.0 GB 的 DELL 计算机上运行约需 3.2 h。可以预见, 在全基因组水平上, 如标记区间 150 个, 即便其他条件不变, 运算量也会激增。因此, 优化算法、提高运算速度、在效率和耗时上取得平衡也是下一步值得研究的问题。

4 结论

胚乳性状 QTL 的准确定位是谷物品质性状遗传改良的重要基础。在先前胚乳性状 QTL 区间作图方法的基础上, 发展出胚乳性状 QTL 多区间作图的贝叶斯方法, 在适当的样本容量下, 不仅能够准确估计各个 QTL 所在的染色体位置, 而且能够区分出胚乳 QTL 的两种显性效应。

References

- [1] Mo H-D(莫惠栋). Genetic research of endosperm-quality traits in cereals. *Sci Agric Sin* (中国农业科学), 1995, 28(2): 1-7 (in Chinese with English abstract)
- [2] Mo H. Genetic expression for endosperm traits. In: Weir B S, Eisen E J, Goodman M M, Namkoong G, eds. *Proceedings of the Second International Conference on Quantitative Genetics*. Massachusetts: Sinauer Associates, Inc. 1988. pp 478-487
- [3] Xu C-W(徐辰武), He X-H(何小红), Kuai J-M(蒯建敏), Gu S-L(顾世梁). Mapping quantitative trait loci underlying endosperm traits in cereals. *Sci Agric Sin* (中国农业科学), 2001, 34(2): 117-122 (in Chinese with English abstract)
- [4] Wu R, Lou X Y, Ma C X, Wang X, Larkins B A, Casella G. An improved genetic model generates high-resolution mapping of QTL for protein quality in maize endosperm. *Proc Natl Acad Sci USA*, 2002, 99: 11281-11286
- [5] Wu R, Ma C X, Gallo-Meagher M, Littell R C, Casella G. Statistical methods for dissecting triploid endosperm traits using molecular markers: An autogamous model. *Genetics*, 2002, 162: 875-892
- [6] Xu C, He X, Xu S. Mapping quantitative trait loci underlying triploid endosperm traits. *Heredity*, 2003, 90: 228-235
- [7] Kao C H. Multiple-interval mapping for quantitative trait loci controlling endosperm traits. *Genetics*, 2004, 167: 1987-2002
- [8] Cui Y, Casella G, Wu R. Mapping quantitative trait loci interactions from the maternal and offspring genomes. *Genetics*, 2004, 167: 1017-1026
- [9] Hu Z, Xu C. A new statistical method for mapping QTLs underlying endosperm traits. *Chin Sci Bull*, 2005, 50: 1470-1476
- [10] Wen Y, Wu W. Methods for mapping QTLs underlying endosperm traits based on random hybridization design. *Chin Sci Bull*, 2006, 51: 1976-1981
- [11] Wang X-F(王学枫), Tang Z-X(汤在祥), Wang Y-M(王亚民), Song W(宋雯), Xu C-W(徐辰武). Interval mapping of quantitative trait loci underlying endosperm traits with NCIII and TTC designs. *Acta Agron Sin* (作物学报), 2008, 34(10): 1734-1743 (in Chinese with English abstract)
- [12] He X H, Zhang Y M. Mapping epistatic quantitative trait loci underlying endosperm traits using all markers on the entire genome in a random hybridization design. *Heredity*, 2008, 101: 39-47
- [13] Wang H, Zhang Y M, Li X, Masinde G L, Mohan S, Baylink D J, Xu S. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics*, 2005, 170: 465-480
- [14] Yi N. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics*, 2004, 167: 967-975
- [15] Xu S. Estimating polygenic effects using markers of the entire genome. *Genetics*, 2003, 163: 789-801
- [16] Yi N, Yandell B S, Churchill G A, Allison D B, Eisen E J, Pomp D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 2005, 170: 1333-1344
- [17] Wang Y-M(王亚民), Sun C-S(孙长森), Tang Z-X(汤在祥), Hu Z-Q(胡治球), Xu C-W(徐辰武). Bayesian method for mapping QTL controlling endosperm traits in cereals. *J Yangzhou Univ* (Agric Life Sci Edn) (扬州大学学报·农业与生命科学版), 2008, 29(3): 12-17(in Chinese with English abstract)
- [18] Jiang C, Zeng Z B. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica*, 1997, 101: 47-58
- [19] Xu S, Yi N. Mixed model analysis of quantitative trait loci. *Proc Natl Acad Sci USA*, 2000, 97: 14542-14547
- [20] Wang W(王伟), Hu Z-Q(胡治球), Sun C-S(孙长森), Xu C-W(徐辰武). Single grain observation-based mapping of quantitative traits loci underlying endosperm traits. *Acta Agron Sin* (作物学报), 2005, 31(8): 989-994 (in Chinese with English abstract)
- [21] Xu C-W(徐辰武), Wang W(王伟), Hu Z-Q(胡治球), Sun C-S(孙长森). Plant average-based maximum likelihood mapping of quantitative traits loci controlling endosperm traits. *Acta Agron Sin* (作物学报), 2005, 31(10): 1271-1276 (in Chinese with English abstract)
- [22] Wen Y, Wu W. Interval mapping of quantitative trait loci underlying triploid endosperm traits using F_3 seeds. *J Genet Genomics*, 2007, 34: 429-436
- [23] Wen Y, Wu W. Experimental designs and statistical methods for mapping quantitative trait loci underlying triploid endosperm traits without maternal genetic variation. *J Hered*, 2008, 99: 546-551