网络出版时间:2022-03-08 11:23 DOI:10.13207/j.cnki.jnwafu.2022.09.016
网络出版地址:https://kns.cnki.net/kcms/detail/61.1390.S.20220304.1635.001.html

基于协方差估计的多因变量回归模型的多性状 QTL 定位研究

张 慧[1a,1b],叶景山[2],申佳瑜[1a,1b],刘慧铭[1a,1b],尹 宁[1a,1b],李立婷[3],温永仙[1a,1b]

(1 福建农林大学 a 计算机与信息学院,b 统计及应用研究所,福建 福州,350002;
2 漳州农业发展集团有限公司,福建 漳州,363000;3 厦门华厦学院,福建 厦门 361021)

[摘 要] 【目的】基于协方差估计的多因变量回归(multivariate regression with covariance estimation,MRCE)模型进行多性状 QTL 定位分析,为动植物数量性状基因定位提供理论参考。【方法】构建适用 QTL 定位的 MRCE 模型,设计 3 个模拟试验对模型进行检验,通过计算机生成基因型和 2 个相关性状的表型值,并用 2 组数据对模型进行实际应用,其中一组为水稻 DH 群体数据,选自 qtlnetwork 软件;另一组为水稻永久 F_2 群体数据,由珍汕 97×明恢 63,含有 210 个株系的重组自交系(RIL)群体随机交配生成,分析 MRCE 模型在以上 2 组数据多性状 QTL 定位中的应用效果。【结果】用 MRCE 模型进行 QTL 定位的模拟试验结果表明,遗传变异所占方差比越大,相关系数绝对值越大,遗传率越大,则功效越好,估计值越接近效应值。MRCE 的 QTL 定位应用结果显示,从水稻 DH 群体中识别出 8 个 QTL 与 ph6 性状有关,6 个 QTL 与 ph8 性状有关;从 1998 年水稻永久 F_2 群体数据中识别出 3 个 QTL 与穗粒数相关,10 个 QTL 与粒质量相关;从 1999 年数据识别出 3 个 QTL 与穗粒数相关,6 个 QTL 与粒质量相关。【结论】利用 MRCE 模型进行多性状 QTL 定位是可行的。

[关键词] 数量性状;QTL 定位;多因变量;协方差估计;回归模型
[中图分类号] Q348 [文献标志码] A [文章编号] 1671-9387(2022)09-0135-09

QTL mapping analysis of multiple quantitative traits based on multivariate regression with covariance estimation

ZHANG Hui[1a,1b],YE Jingshan[2],SHEN Jiayu[1a,1b],LIU Huiming[1a,1b],
YIN Ning[1a,1b],LI Liting[3],WEN Yongxian[1a,1b]

(1a *College of Computer and Information Science*,b *Institute of Statistics and Applications*,
Fujian Agriculture and Forestry University,*Fuzhou*,*Fujian* 350002,*China*;2 *Zhangzhou Agricultural Development Group Co.*,*Ltd*,
Zhangzhou,*Fujian* 363000,*China*;3 *Xiamen Huaxia University*,*Xiamen*,*Fujian* 361021,*China*)

Abstract:【Objective】A QTL mapping method of multiple traits was proposed based on multivariate regression with covariance estimation (MRCE) to provide reference for mapping quantitative traits of animals and plants in practice.【Method】The genetic model by MRCE was constructed for QTL mapping and validated by three simulations. Genotype and 2 related phenotype values were generated by computer simulation. The data of DH population in rice were selected from qtlnetwork software for first application example. The immortalized F_2 population of rice was generated by random hybridization of a recombinant inbred

[收稿日期] 2021-10-07
[基金项目] 国家自然科学基金项目(32071892);福建省自然科学基金项目(2021J01126);福建农林大学科技创新专项基金项目(CXZX2019127G)
[作者简介] 张 慧(1997—),女,福建闽侯人,在读硕士,主要从事统计信息技术与数据挖掘研究。E-mail:392019334@qq.com
[通信作者] 温永仙(1966—),女,福建永泰人,教授,博士,博士生导师,主要从事统计信息技术与数据挖掘研究。E-mail:wen9681@sina.com

line population (210 lines) from Zhenshan 97×Minghui 63 for second application example. Then, the two examples were used to analyze the application of MRCE in QTL mapping analysis on multiple quantitative traits.【Result】 The QTL mapping by MRCE showed that the power of QTL detection increased and the estimation accuracies increased as the increase of genetic variant effect of variance, absolute value of correlation coefficient of phenotype and QTL heritability. For first application example, 8 QTLs were identified for ph6 traits and 6 QTLs were related to ph8 traits for DH population of rice. For second application example, 3 QTLs were identified for grains per panicle and 10 QTLs were identified for grain weight by the data of immortalized F_2 population of rice in 1998. 3 QTLs were identified for grains per panicle and 6 QTLs were related to grain weight by the data of immortalized F_2 population of rice in 1999.【Conclusion】 It was feasible to use MRCE for QTL mapping of multiple quantitative traits.

Key words: quantitative trait; QTL mapping; multiple-dependent variables; covariance estimation; regression model

数量性状基因座(quantitative trait loci, QTL)与连续变化的数量性状表型有密切关系,常用 DNA 分子标记技术对数量性状基因遗传位置进行标记,QTL 定位研究是遗传学领域的一个重点。

早期的 QTL 定位方法是利用分子标记与 QTL 之间的连锁关系,定位出 QTL 在染色体上的位置,并估算出相应 QTL 效应值。但初期的单个性状 QTL 定位存在一些问题。随后,研究者提出了多性状联合定位分析方法。Jiang 和 Zeng[1] 提出了一种多性状的复合区间定位方法(composite interval mapping, CIM),利用所考虑性状的相关结构进行定位,可以提高 QTL 检测的准确性。还有研究结果表明,同时检测多个性状比单独检测 1 个性状更有效[2-4]。

多性状 QTL 定位的实质是在多因变量回归模型的基础上进行变量选择。近年来,很多学者尝试对多性状 QTL 定位进行研究。Jansen 和 Stam[5] 提出了参数多变量回归模型,研究多个性状与分子标记之间的关系,并通过极大似然比检验来找出与性状相关的 QTL 位点。但这种方法计算量较大,为此,Lange 和 Whittaker[6] 提出广义估计方程,该方程不需要对具体分布进行假设,大大缩短了计算时间。肖静等[7] 和 Xiao 等[8] 提出了多性状主基因联合分离分析方法(multivariate segregation analysis, MSA),通过对比单个性状和多个性状联合分析的模拟结果发现,多个性状联合分析效果较好,统计功效和效应估计值的准确度也较高。Banerjee 等[9] 在多性状分析中引入贝叶斯模型,并结合马尔科夫链蒙特卡洛(markov chain monte carlo, MCMC)算法进行模拟,建立相关表型和不相关表型两个模型。Xu 等[10] 利用贝叶斯模型分析多个性状与分子标记之间的关系,通过压缩系数方式来估计所有标记区间内的遗传效应。

关于多性状联合基因关联分析,O′Reilly 等[11] 用 MultiPhen 方法,以可解释的方式同时快速模拟了多种表型,提高了功效[11]。Bolormaa 等[12] 和 Zhu 等[13] 用 meta 分析方法对多个性状进行了基因关联分析。Cheng 等[14] 用混合先验贝叶斯回归方法进行多性状回归分析发现,其效果优于单性状基因关联分析。Tong 等[15] 结合期望最大化算法(expectation maximization, EM),提出多性状特征多区间下估计参数的方法(multiple trait multiple-interval mapping-new, MTMIM-NEW)。Montesinos-López 等[16] 用基于奇异值分解(singular value decomposition, SVD)的四阶段分析方法进行多性状基因关联分析发现,其在参数估计和预测精度方面与使用贝叶斯多性状多环境模型(bayesian multiple-trait and multiple-environment model, BMTME)获得的结果类似。Yang 等[17] 提出了一个具有多性状的全关联的整合函数线性模型,利用惩罚函数解决了单核苷酸多态性(single nucleotide polymorphism, SNP)的高维性和多性状相关性问题。Lin 等[18] 提出一种基于混合线性模型的多性状联合基因关联分析方法,模拟结果表明,多性状全基因组关联研究(genome-wide association studies, GWAS)在检测多效性位点的影响方面较单个性状效果更好。Tran 等[19] 在绘制多数量性状位点的统计方法中考虑到 X 染色体,扩展了一种多 QTL 模型选择的惩罚似然方法。

此外,还有一些降维方法被应用于解决多性状基因关联分析问题,包括主成分分析[20-21]、典型相关分析[22]、偏最小二乘法[23] 和贝叶斯 Lasso 方法[24]。

本研究采用 Rothman 等[25]提出的基于协方差估计的多因变量回归（multivariate regression with covariance estimation，MRCE）模型，通过计算机模拟产生基因型数据和性状表型数据，利用 MRCE 模型进行参数估计，探究基因位点解释的方差比、表型相关系数、遗传率对模拟效果的影响，并将此模型应用于水稻群体标记数据中，完成基因定位，估计其参数，以期为多性状 QTL 定位研究提供参考。

1　研究模型的构建

1.1　模型建立

假设一个遗传群体包含 n 个个体，若不考虑群体结构等因素，对第 i 个个体，在遗传关联分析中假设有 p 遗传标记为 $x_{i1},x_{i2},\cdots,x_{ip}(i=1,2,\cdots,n)$，若有 q 个数量性状，线性遗传模型可以表示为：

$$y_{ij}=\beta_{0j}+\sum_{k=1}^{p}\beta_{kj}x_{ik}+\varepsilon_{ij},(i=1,2,\cdots,n;j=1,2,\cdots,q;k=1,2,\cdots,p)。$$

式中：y_{ij} 表示第 i 个个体第 j 个性状表型值。x_{ik} 表示第 i 个个体在第 k 个基因标记位点的指示变量值，若 A 和 a 表示 1 对等位基因，当基因型是 AA 时，x_{ik} 取 1；当基因型是 Aa 时，x_{ik} 取 0；当基因型是 aa 时，x_{ik} 取 −1。β_{0j} 代表第 j 个数量性状的均值，β_{kj} 代表第 k 个基因标记位点对第 j 个数量性状所表现的遗传效应值。ε_{ij} 为随机误差，一般 ε_{ij} 之间不是相互独立的，假定它们服从均值均为 0，协方差矩阵为 Σ 的多元正态分布。当 $q=1$ 时，模型为经典的单因变量回归模型。将线性遗传模型写成矩阵形式，分别用 $\boldsymbol{X}$、$\boldsymbol{Y}$、$\boldsymbol{B}$、$\boldsymbol{\varepsilon}$ 表示：

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix},\quad \boldsymbol{Y}=\begin{bmatrix}y_{11} & y_{12} & \cdots & y_{1q}\\ y_{21} & y_{22} & \cdots & y_{2q}\\ \vdots & \vdots & & \vdots\\ y_{n1} & y_{n2} & \cdots & y_{nq}\end{bmatrix},$$

$$\boldsymbol{B}=\begin{bmatrix}\beta_{01} & \beta_{02} & \cdots & \beta_{0q}\\ \beta_{11} & \beta_{12} & \cdots & \beta_{1q}\\ \vdots & \vdots & & \vdots\\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq}\end{bmatrix},\quad \boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1q}\\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2q}\\ \vdots & \vdots & & \vdots\\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nq}\end{bmatrix}。$$

则有：

$$\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{\varepsilon}。$$

Rothman[25]提出了 MRCE，$\boldsymbol{B}$ 的稀疏估计量，该方法在负对数似然函数上加入了两个惩罚项，求解 $\boldsymbol{B}$ 的稀疏估计值，具体形式为：

$$(\hat{\boldsymbol{B}},\hat{\boldsymbol{\Omega}})=\underset{\mathrm{B},\Omega}{\operatorname{argmin}}\{tr[\frac{1}{n}(\boldsymbol{Y}-\boldsymbol{XB})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{XB})\boldsymbol{\Omega}]-\lg|\boldsymbol{\Omega}|+\lambda_1\sum_{j'\neq j}|\omega_{j'j}|+\lambda_2\sum_{j=1}^{p}\sum_{k=1}^{q}|\beta_{kj}|\}。$$

式中：$\boldsymbol{\Omega}=\Sigma^{-1}=[\omega_{j'j}]$，$\Sigma^{-1}$ 是协方差矩阵 Σ 的逆矩阵，$\omega_{j'j}$ 是逆矩阵中的元素。$\lambda_1\geqslant 0$，$\lambda_2\geqslant 0$，二者均是调整参数，用 k 折交叉验证来选择参数 λ_1 和 λ_2。

1.2　模型的假设检验

首先，原假设效应系数都为 0，通过基于 Pillai-Bartlett 迹、Hotelling-Lawley 迹和 Wilks's Lambda 的近似 F 分布检验进行模型检验[26-30]。其次，用 $\boldsymbol{L\beta}=0$ 的方法对基因标记位点的遗传效应 β_{ij}（$i=1,2,\cdots,p$；$j=1,2,\cdots,q$）进行检验[31]，其中 $\boldsymbol{L}$ 是 $c\times p+1$ 阶的矩阵，用来识别检验假设的遗传效应。如对 $\boldsymbol{\beta}_{1j}$ 的假设检验可以写：

$$\boldsymbol{L\beta}=[0,1,0,0,\cdots,0]\begin{bmatrix}\beta_{0j}\\ \beta_{1j}\\ \vdots\\ \beta_{pj}\end{bmatrix}=\boldsymbol{\beta}_{1j}=0,(j=1,2,\cdots,q)。$$

对于假设可采用 F 检验进行，F 检验的形式为：

$$F_{(df_h,df_e)}=\frac{SS_h/\mathrm{d}f_h}{SS_e/\mathrm{d}f_e}\sim F(c,n-p-1)。$$

式中：假设平方和 $SS_h=(\boldsymbol{L}\hat{\beta})^{\mathrm{T}}(\boldsymbol{L}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{L}^{\mathrm{T}})^{-1}\times(\boldsymbol{L}\hat{\beta})$，误差平方和 $SS_e=\sum(\boldsymbol{Y}_j-\boldsymbol{X}\hat{\beta})^2$，其中 $\boldsymbol{Y}_j=\begin{bmatrix}y_{1j}\\ \vdots\\ Y_{nj}\end{bmatrix}$，$j=1,2,\cdots,q$；自由度 $df_h=c$，c 是矩阵 $\boldsymbol{L}$ 的行数；$df_e=n-p-1$。

2　研究模型的模拟与实际应用

2.1　模拟试验设计

2.1.1　SNPs 生成　参照黄杨岳等[32]的 SNPs 数据仿真方法，生成纯合 SNP 模拟数据 SNPs，包含 500 个个体和 200 个基因位点，其基因型为 AA、aa。

2.1.2　数量性状表型值生成的多元仿真框架　(1) 给定截距 b 的合适值。

(2) 按照 Porter 和 O'Reilly[33]的方法，给定 v 值，v 是遗传变异所解释的表型方差遗传效应向量。例如，当 $v=(0,1,0.5)$ 时，对应于 SNP，表示解释了

性状1的0.1%表型方差,性状2的0.5%表型方差。

(3)根据 v 值计算回归效应 $f(v)$。$f(v)=\sqrt{\frac{s}{2pq}}$,其中 q 是次等位基因频率,$p=1-q$;s 是当误差方差等于1时,SNP解释的变换表型方差,$s=\frac{v}{1-v}$。例如,当 $v=0.5$ 时,表示解释了0.5%的表型方差,此时 $s=0.005/(1-0.005)=0.005\ 025$。

(4)协方差矩阵计算。在经典数量遗传学中,在一个位点加性方差被当作通过基因值回归所解释的遗传方差,显性方差被当作残差遗传方差,不能用回归所解释[34-36],假设 A 是等位基因置换平均效应,根据Falconer等[34]和Lynch等[35]研究,性状基因型加性方差为 $\sigma_A^2=2pqA^2$。给定遗传率 h^2:

$$h^2=\frac{\sigma_A^2}{\sigma_A^2+\sigma_e^2}。$$

则随机误差方差 σ_e^2 为:

$$\sigma_e^2=\frac{(1-h^2)\sigma_A^2}{h^2}。$$

性状表型相关矩阵为 $\boldsymbol{R}$:

$$\boldsymbol{R}=\begin{bmatrix}1 & \cdots & \rho_{1j} & \cdots & \rho_{1q}\\ \vdots & & \vdots & & \vdots\\ \rho_{i1} & & \rho_{ij} & & \rho_{iq}\\ \vdots & & \vdots & & \vdots\\ \rho_{q1} & \cdots & \rho_{qj} & \cdots & 1\end{bmatrix}。$$

式中:$\rho_{ij}(i,j=1,2,\cdots,q)$ 表示第 i 个性状与第 j 个性状的表型相关系数。

则协方差$\sum$为:

$$\sum=\sqrt{\sigma_{e1}^2}\times\sqrt{\sigma_{e2}^2}\times\boldsymbol{R}。$$

式中:σ_{e1}^2、σ_{e2}^2分别代表性状1和性状2的随机误差方差,利用R语言中的rmvnorm函数生成随机误差矩阵 ε,$\varepsilon\sim N(0,\sum)$。

数量性状表型值 y 的计算公式为:

$$y=b+f(v)x+\varepsilon。$$

式中:x 代表基因型指示变量,基因型为AA时取1,基因型为aa时取-1。

2.1.3 *QTL检验功效的计算* 对于染色体上的某个基因位点需要对其进行参数估计及统计检验,若对于给定的显著性水平,该位点的遗传效应值达到显著,说明在该位点检测到QTL。若假设情况的计算机模拟共重复 m 次,染色体基因位点能检测到 m_0 次,则该位点的QTL检测功效为 m/m_0。效应值是 $f(v)$,估计值是用MRCE方法估计的 $f(v)$。

2.1.4 *模拟案例* 假设有2个相关的数量性状 y_1、y_2,2个性状分别与基因位点50(假设为QTL1)和基因位点150(假设为QTL2)有关,给定生成表型值 y_1、y_2 方程的截距 $b_1=1$,$b_2=0.8$,此时性状表型相关矩阵 $\boldsymbol{R}$ 为2×2的矩阵,即 $\boldsymbol{R}=\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}$,协方差$\sum=\sqrt{\sigma_{e1}^2}\times\sqrt{\sigma_{e2}^2}\times\boldsymbol{R}$,显著性水平 $\alpha=0.05$,每种情况重复100次,即 $m=100$,计算每个QTL的检测功效。

(1)模拟1。给定3组 v 值,$v_1=(0.5,0.5)$,$v_2=(0.5,0.1)$,$v_3=(0.5,0.0)$,计算相应的 $f(v)$:$f(v_1)=(0.125\ 3,0.125\ 3)$,$f(v_2)=(0.125\ 3,0.055\ 9)$,$f(v_3)=(0.125\ 3,0.000\ 0)$,进而得到相应性状表型值,利用MRCE模型进行参数估计,根据功效,比较 v 值对QTL定位模拟效果的影响。

(2)模拟2。设 $v=(0.5,0.5)$,相关系数从-0.9每次增加0.1直到0.9,分别产生相应的性状表型值,利用MRCE方法进行参数估计,探究相关系数对模拟效果的影响;同时设 $v=(0.5,0.1)$、遗传率为(0.05,0.05)时,研究相关系数对QTL定位模拟效果的影响。

(3)模拟3。设 $v=(0.5,0.5)$,给定不同遗传率组合(0.05,0.05),(0.05,0.10),(0.05,0.15),(0.10,0.05),(0.10,0.10),(0.10,0.15),(0.15,0.05),(0.15,0.10),(0.15,0.15),对相应性状表型值进行功效模拟,分析遗传率对QTL定位模拟效果的影响。

2.2 实例数据收集

2.2.1 *实例1数据* 实例1数据选自qtlnetwork软件,是一个水稻DH群体,包含12条水稻染色体中的3条染色体,共54个标记,每条染色体上标记数量不等,99个个体,2个环境(1998年和1999年)。由于水稻DH群体数据中存在缺失数据,需通过相邻平均值方法进行填补,再将1998年与1999年数据进行整合,最终得到的数据集为54个标记和198个样本量,提取ph6、ph8作为性状表型值,两性状的相关系数为0.946 4。

2.2.2 *实例2数据* 实例2数据是包含12条染色体的水稻永久 F_2 群体试验数据[37-40],该群体由来自珍汕97×明恢63,含有210个株系的重组自交系(RIL)群体随机交配生成,共产生278个杂种株系,其遗传图谱共有1 619个标记序号(Bin1~Bin1619),包含单株产量、分蘖数、穗粒数、粒质量4个性状,本研究仅对穗粒数和粒质量进行联合分析,

剔除缺失数据后获得 2 组完整数据，其中 1998 年有 246 个，1999 年有 276 个，为了简化考虑，本研究仅考虑其加性效应。

3　结果与分析

3.1　MRCE 模型对 QTL 定位效果的模拟试验结果

3.1.1　模拟试验 1　通过 MRCE 模型对不同 v 值情况下的 QTL 进行定位发现，任意给定一个固定的相关系数和遗传率时，不同 v 值对 QTL 定位的影响规律大致相同，所以本研究选择其中 3 个相关系数(0.1，0.5，0.9)且遗传率为(0.05，0.05)时进行分析，结果见表 1。从表 1 可以看出，当相关系数分别为 0.1，0.5，0.9，遗传率为(0.05，0.05)时，v 值越大，功效越大；v 为 0 时，功效为 0 或接近 0。所以，如果遗传变异所解释的方差比大小合适，则利用 MRCE 模型进行 QTL 定位是可行的。

表 1　不同 v 值情况下 QTL 定位的模拟结果

Table 1　Simulation of QTL mapping with different v values

v	相关系数 Correlation coefficient	QTL1 效应值 Real value	QTL1 估计值 Estimated value	QTL1 功效 Power	QTL2 效应值 Real value	QTL2 估计值 Estimated value	QTL2 功效 Power
(0.5,0.5)	0.1	0.125 3	0.082 4±0.018 0	0.81	0.125 3	0.076 9±0.016 1	0.73
(0.5,0.1)	0.1	0.125 3	0.106 5±0.011 8	0.96	0.055 9	0.040 6±0.012 1	0.61
(0.5,0.0)	0.1	0.125 3	0.097 6±0.014 4	0.94	0.000 0	0.000 0±0.000 0	0.00
(0.5,0.5)	0.5	0.125 3	0.087 2±0.014 9	0.85	0.125 3	0.083 1±0.015 4	0.83
(0.5,0.1)	0.5	0.125 3	0.107 1±0.008 9	0.96	0.055 9	0.037 0±0.010 1	0.63
(0.5,0.0)	0.5	0.125 3	0.100 6±0.013 4	0.91	0.000 0	0.000 0±0.000 0	0.00
(0.5,0.5)	0.9	0.125 3	0.106 1±0.011 5	0.89	0.125 3	0.104 9±0.010 0	0.94
(0.5,0.1)	0.9	0.125 3	0.111 6±0.007 5	0.96	0.055 9	0.043 8±0.006 6	0.72
(0.5,0.0)	0.9	0.125 3	0.106 4±0.011 4	0.91	0.000 0	0.000 0±0.000 0	0.00

3.1.2　模拟试验 2　图 1 表明，当 v 相同时，两端功效略高于中间部分，说明相关系数绝对值越大，其功效越高。

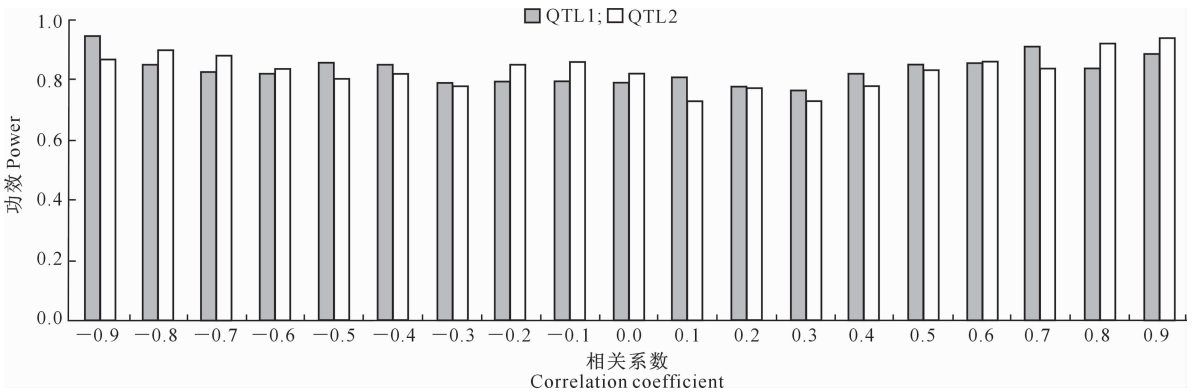

图 1　v=(0.5，0.5)时不同相关系数对 MRCE 模型用于 QTL 定位模拟效果的影响

Fig. 1　Simulation of QTL mapping based on MRCE for different correlation coefficients and v=(0.5，0.5)

表 2 是当 v=(0.5，0.1)、遗传率为(0.05，0.05)时相关系数对模拟效果的影响。从表 2 可以看出，相关系数绝对值越大，QTL1 和 QTL2 估计值越接近效应值，功效也越高。可见 MRCE 模型可用于 QTL 定位。

表 2　v=(0.5，0.1)时不同相关系数情况下 QTL 定位的模拟结果

Table 2　Simulation of QTL mapping with different correlation coefficients and v=(0.5，0.1)

相关系数 Correlation coefficient	QTL1 效应值 Real value	QTL1 估计值 Estimated value	QTL1 功效 Power	QTL2 效应值 Real value	QTL2 估计值 Estimated value	QTL2 功效 Power
−0.9	0.125 3	0.111 8±0.007 4	0.98	0.055 9	0.042 1±0.006 3	0.68
−0.8	0.125 3	0.110 9±0.007 4	0.98	0.055 9	0.039 8±0.007 8	0.70
−0.7	0.125 3	0.109 1±0.008 5	0.97	0.055 9	0.040 2±0.009 9	0.66
−0.6	0.125 3	0.108 2±0.010 0	0.96	0.055 9	0.039 6±0.011 3	0.60
−0.5	0.125 3	0.107 4±0.011 2	0.97	0.055 9	0.037 8±0.010 3	0.60
−0.4	0.125 3	0.107 2±0.011 3	0.97	0.055 9	0.036 8±0.012 6	0.66

表 2(续) Conutinued table 2

相关系数 Correlation coefficient	QTL1 效应值 Real value	QTL1 估计值 Estimated value	QTL1 功效 Power	QTL2 效应值 Real value	QTL2 估计值 Estimated value	QTL2 功效 Power
−0.3	0.125 3	0.106 0±0.011 8	0.95	0.055 9	0.039 0±0.009 9	0.55
−0.2	0.125 3	0.104 4±0.012 7	0.98	0.055 9	0.040 1±0.011 4	0.51
−0.1	0.125 3	0.105 0±0.013 6	0.97	0.055 9	0.036 7±0.010 7	0.58
0.0	0.125 3	0.103 6±0.013 7	0.95	0.055 9	0.039 6±0.009 8	0.63
0.1	0.125 3	0.106 5±0.011 8	0.96	0.055 9	0.040 6±0.012 1	0.61
0.2	0.125 3	0.103 6±0.013 1	0.97	0.055 9	0.037 6±0.011 1	0.59
0.3	0.125 3	0.106 9±0.010 0	0.93	0.055 9	0.037 1±0.011 6	0.68
0.4	0.125 3	0.107 6±0.010 6	0.97	0.055 9	0.040 3±0.009 2	0.60
0.5	0.125 3	0.107 1±0.008 9	0.96	0.055 9	0.037 0±0.010 1	0.63
0.6	0.125 3	0.108 3±0.009 1	0.98	0.055 9	0.041 0±0.010 0	0.64
0.7	0.125 3	0.108 5±0.009 9	0.97	0.055 9	0.040 7±0.009 1	0.64
0.8	0.125 3	0.112 3±0.008 7	0.95	0.055 9	0.040 4±0.009 4	0.67
0.9	0.125 3	0.111 6±0.007 5	0.96	0.055 9	0.043 8±0.006 6	0.72

3.1.3 模拟试验 3 分析 $v=(0.5,0.5)$时遗传率 对模拟结果的影响,结果见表 3。

表 3 $v=(0.5,0.5)$时不同遗传率下 QTL 定位的模拟结果

Table 3 Simulation of QTL mapping for different heritability and $v=(0.5,0.5)$

遗传率 Heritability	QTL1 效应值 Real value	QTL1 估计值 Estimated value	QTL1 功效 Power	QTL2 效应值 Real value	QTL2 估计值 Estimated value	QTL2 功效 Power
(0.05,0.05)	0.125 3	0.106 1±0.011 5	0.89	0.125 3	0.104 9±0.010 0	0.94
(0.05,0.10)	0.125 3	0.107 7±0.011 2	0.96	0.125 3	0.108 0±0.009 5	0.96
(0.05,0.15)	0.125 3	0.111 2±0.007 4	0.95	0.125 3	0.110 1±0.008 4	1.00
(0.10,0.05)	0.125 3	0.109 2±0.009 4	0.92	0.125 3	0.108 9±0.008 7	0.99
(0.10,0.10)	0.125 3	0.112 1±0.007 7	0.98	0.125 3	0.111 9±0.008 7	1.00
(0.10,0.15)	0.125 3	0.113 7±0.007 8	1.00	0.125 3	0.112 0±0.008 3	1.00
(0.15,0.05)	0.125 3	0.111 0±0.008 6	0.98	0.125 3	0.109 0±0.008 7	0.98
(0.15,0.10)	0.125 3	0.112 2±0.009 5	1.00	0.125 3	0.112 3±0.009 6	1.00
(0.15,0.15)	0.125 3	0.112 1±0.009 5	1.00	0.125 3	0.111 8±0.009 7	1.00

从表 3 可以看出,遗传率越高,其效应估计值越接近真值,功效也越好,在其他不同遗传率假设下也有上述相似结果。综上可知,利用 MRCE 模型进行 QTL 定位分析是可行的,同时遗传变异所占方差比越大,相关系数绝对值越大,遗传率越大,则模拟效果越好。

3.2 MRCE 模型对 QTL 定位效果的应用实例

3.2.1 应用实例 1 表 4 和表 5 分别为用 qtlnetwork 软件和 MRCE 模型得出的水稻 DH 群体数据 QTL 定位结果,定位到的 QTL 均通过了显著性检验。

表 4 基于 qtlnetwork 的水稻 DH 群体数据 QTL 定位结果

Table 4 QTL mapping for DH population of rice by qtlnetwork

ph6 性状 ph6 trait		ph8 性状 ph8 trait	
QTL	间隔 Interval	QTL	间隔 Interval
1～6	MK6－MK7	1～15	MK15－MK16
1～15	MK15－Mk16	2～12	MK30－MK31
2～12	MK30－MK31	3～20	MK53－MK54
3～20	MK53－MK54		

表 5 基于 MRCE 模型的水稻 DH 群体数据 QTL 定位结果

Table 5 QTL mapping for DH population of rice based on MRCE

定位相同的数量性状基因座 Consistent QTL			多定位的数量性状基因座 More mapped QTL		
编号 Number	ph6	ph8	编号 Number	ph6	ph8
MK6	−0.091 3	0.000 0	MK18	0.448 8	0.510 6
MK15	0.121 4	0.299 8	MK32	−0.115 6	0.000 0
MK16	0.100 9	0.178 3	MK52	0.313 5	0.389 8
MK31	−0.097 1	−0.153 5			
MK54	0.079 2	0.072 4			

从表 5 可以看出，通过 MRCE 模型发现，8 个标记 MK6、MK15、MK16、MK18、MK31、MK32、MK52、MK54 与 ph6 性状有关，6 个标记 MK15、MK16、MK18、MK31、MK52、MK54 与 ph8 性状有关。

由表 5 还可以看出，与 qtlnetwork 软件定位的 QTL 结果对比，基于 MRCE 模型选出的标记中，有 6 个标记与真实结果一致，尤其是 MK6 这个标记仅与 ph6 有关；此外，还多定位到了 3 个标记，分别为 MK18、MK32、MK52；且这些标记与 qtlnetwork 软件定位的 QTL 结果相邻。MK18 与 qtlnetwork 软件定位的 MK15－MK16 相邻，MK32 与 MK30－MK31 相邻，MK52 与 MK53－MK54 相邻，多定位到的标记可能与邻近 QTL 效应的影响以及在 qtlnetwork 软件定位过程中的阈值设定有关，由此可知，基于 MRCE 模型的 QTL 定位与用 qtlnetwork 软件的定位结果基本相符，进一步说明 MRCE 模型应用于 QTL 定位是可行的。

3.2.2　*应用实例 2*　由于 MRCE 模型不能用于样本量(n)小于遗传标记个数(p)的情况，所以本研究计算了遗传标记与性状表型值的边际相关系数，且边际相关系数越高，则该遗传标记与对应性状表型值的相关性越高，最终选取边际相关系数绝对值较大的 200 个标记数据进行初步降维，QTL 定位结果见表 6。表 6 表明，利用 MRCE 模型检测到 1998 年穗粒数在第 3、第 6 和第 7 条染色体上各有 1 个 QTL；粒质量在第 1 和第 5 条染色体上各有 3 个 QTL，第 3 和第 7 条染色体上各有 2 个 QTL。利用 MRCE 模型检测到 1999 年穗粒数在第 3 条染色体有 1 个 QTL，第 7 条染色体有 2 个 QTL，粒质量第 1 和第 3 条染色体各有 1 个 QTL，第 5 和第 7 条染色体各有 2 个 QTL。对比 1998 和 1999 年的定位结果可知，穗粒数都定位到 Bin436，粒质量都定位到 Bin65、Bin439、Bin699、Bin769 和 Bin1008，是因为 1998 年穗粒数与粒质量 2 个性状之间的相关系数(0.15)大于 1999 年(0.05)，故 1998 年定位出更多 QTL。

综合实例 1 和实例 2 的结果，说明 MRCE 模型不仅可以用于模拟 QTL 定位，而且在实际定位中也同样适用，结果较好。

表 6　基于 MRCE 模型的水稻永久 F_2 群体穗粒数和粒质量的 QTL 定位结果

Table 6　QTL mapping of grains per panicle and grain weight for immortalized F_2 population of rice by MRCE

性状 Trait	1998(0.15)			1999(0.05)		
	QTL	染色体 Chromosome	Bin 的位置 Bin location	QTL	染色体 Chromosome	Bin 的位置 Bin location
穗粒数 Grains per panicle	QTL3	3	436	QTL3	3	436
	QTL6	6	878	QTL7a	7	997
	QTL7	7	1 014	QTL7b	7	1 057
粒质量 Grain weight	QTL1a	1	31	QTL1	1	65
	QTL1b	1	33	QTL3	3	439
	QTL1c	1	65	QTL5a	5	699
	QTL3a	3	439	QTL5b	5	769
	QTL3b	3	453	QTL7a	7	1 008
	QTL5a	5	699	QTL7b	7	1 014
	QTL5b	5	761			
	QTL5c	5	769			
	QTL7a	7	1 008			
	QTL7b	7	1 020			

注：0.15，0.05 分别是 1998 年和 1999 年穗粒数与粒质量的相关系数。

Note：0.15 and 0.05 are correlation coefficients between grains per panicle and grain weight traits in 1998 and 1999，respectively.

4　讨　论

采用不同的 QTL 定位方法和不同数据检测到的 QTL 数目和位置可能有差异，若能定位到更多的 QTL，在一定程度上可以弥补用其他方法未找到的备选 QTL，但是否是真实的 QTL，需用生物检测方法进行验证。Yu 等[39]用超高密度 SNP 图谱，检测出穗粒数性状在第 1、第 3 和第 7 条染色体上各有 1 个 QTL，粒质量在第 1 和第 3 条染色体上各有 2 个 QTL，第 5 和第 9 条染色体上各有 1 个 QTL。对比本研究结果可以发现，利用 MRCE 模型检测出的穗粒数、粒质量 QTL 更多；其中 1998 年的数据中多检测出穗粒数第 6 条染色体上的 1 个 QTL，粒质量多检测出第 7 条染色体上的 2 个 QTL，且与 Yu 等[39]检测到的 QTL 位置大致相近；但本研究利用 MRCE 模型检测时丢失了穗粒数第 1 条染色体上

的 1 个 QTL 和粒质量性状第 9 条染色体上的 1 个 QTL、第 1 条染色体 Bin172 位置附近的 1 个 QTL。原因可能是只考虑了加性效应而没有考虑显性效应,或丢失 QTL 的 LOD 值都比较小,刚好超过给定的阈值[39]。

本研究仅验证了 MRCE 模型定位 QTL 的可行性和优势,即表型性状联合定位时相关系数越大,效果越好。且 MRCE 模型不适用样本量小于维度的情况,对此情况可利用降维手段,先将高维数据降为低维。对于水稻永久 F_2 群体数据分析中定位到的 QTL 少的问题,可以增加显性效应进一步研究。

[参考文献]

[1] Jiang C J, Zeng Z B. Multiple trait analysis of genetic mapping for quantitative trait loci [J]. Genetics, 1995, 140(3): 1111-1127.

[2] Zhu W S, Zhang H P. Why do we test multiple traits in genetic association studies? [J]. Journal of the Korean Statistical Society, 2009, 38(1): 1-10.

[3] Joehanes R. Multiple-trait multiple-interval mapping of quantitative-trait loci [D]. Manhattan: Kansas State University, 2009.

[4] Silva L, Wang S C, Zeng Z B. Multiple trait multiple interval mapping of quantitative trait loci from inbred line crosses [J]. BMC Genetics, 2012, 13(1): 1-24.

[5] Jansen R C, Stam P. High resolution of quantitative traits into multiple loci via interval mapping [J]. Genetics, 1994, 136(4): 1447-1455.

[6] Lange C, Whittaker J C. Mapping quantitative trait loci using generalized estimating equations [J]. Genetics, 2001, 159(3): 1325-1337.

[7] 肖 静,胡治球,汤在祥,等. 多个相关数量性状主基因的联合分析方法 [J]. 中国农业科学, 2005, 38(9): 1717-1724.
Xiao J, Hu Z Q, Tang Z X, et al. Joint analysis method for major genes controlling multiple correlated to quantitative traits [J]. Scientia Agricultura Sinica, 2005, 38(9): 1717-1724.

[8] Xiao J, Wang X, Hu Z, et al. Multivariate segregation analysis for quantitative traits in line crosses [J]. Heredity, 2007, 98(6): 427-435.

[9] Banerjee S, Yandell B S, Yi N J. Bayesian quantitative trait loci mapping for multiple traits [J]. Genetics, 2008, 179(4): 2275-2289.

[10] Xu C W, Wang X F, Li Z K, et al. Mapping QTL for multiple traits using Bayesian statistics [J]. Genetics Research, 2009, 91(1): 23-37.

[11] O'Reilly P F, Hoggart C J, Pomyen Y, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS [J]. PLoS One, 2012, 7(5): e34861.

[12] Bolormaa S, Pryce J E, Reverter A, et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle [J]. PLoS Genetics, 2014, 10(3): e1004198.

[13] Zhu X F, Feng T, Tayo B O, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension [J]. The American Journal of Human Genetics, 2015, 96(1): 21-36.

[14] Cheng H, Kizilkaya K, Zeng J, et al. Genomic prediction from multiple-trait Bayesian regression methods using mixture priors [J]. Genetics, 2018, 209(1): 89-103.

[15] Tong L, Sun X X, Zhou Y. Simultaneous estimation of QTL parameters for mapping multiple traits [J]. Journal of Genetics, 2018, 97(1): 267-274.

[16] Montesinos-López O A, Montesinos-López A, Crossa J, et al. A singular value decomposition Bayesian multiple-trait and multiple-environment genomic model [J]. Heredity, 2019, 122(4): 381-401.

[17] Yang L, Fan W. Integrative functional linear model for genome-wide association studies with multiple traits. [J]. Biostatistics, 2020, 21(4): 1-17.

[18] Lin F, Qi G A, Xu T, et al. Joint association analysis method to dissect complex genetic architecture of multiple genetically related traits [J]. The Crop Journal, 2020, 8(5): 733-744.

[19] Tran Q, Broman K W. Treatment of the X chromosome in mapping multiple quantitative trait loci [J]. G3, 2021, 11(2): 1-6.

[20] Gao H, Zhang T, Wu Y, et al. Multiple-trait genome-wide association study based on principal component analysis for residual covariance matrix [J]. Heredity, 2014, 113(6): 526-532.

[21] Zhang W G, Gao X, Shi X P, et al. PCA-based multiple-trait GWAS analysis: a powerful model for exploring pleiotropy [J]. Animals, 2018, 8(12): 239.

[22] Cichonska A, Rousu J, Marttinen P, et al. Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis [J]. Bioinformatics, 2016, 32(13): 1981-1989.

[23] Gianola D, Fernando R L. A multiple-trait Bayesian lasso for genome-enabled analysis and prediction of complex traits [J]. Genetics, 2020, 214(2): 305-331.

[24] Chun H, Ballard D H, Cho J, et al. Identification of association between disease and multiple markers via sparse partial least-squares regression [J]. Genetic Epidemiology, 2011, 35(6): 479-486.

[25] Rothman A J, Levina E, Zhu J. Sparse multivariate regression with covariance estimation [J]. Journal of Computational and Graphical Statistics, 2010, 19(4): 947-962.

[26] Anderson T W. An introduction to multivariate statistical analysis [M]. 2nd ed. New York: John Wiley & Sons, 1984.

[27] Fox J. Applied regression analysis and generalized linear models [M]. 2nd ed. Los Angeles: Sage, 2008.

[28] Weisberg S, Fox J. An R companion to applied regression [M]. 2nd ed. Los Angeles: Sage, 2011.

[29] Morrison D F. Multivariate statistical methods [M]. 4th ed.

California:Thomson,2005.

[30] Rao C R. Linear statistical inference and its applications [M]. 2nd ed. New York:John Wiley & Sons,1973.

[31] 理查德·F·哈斯,臧晓露. 多元广义线性模型 [M]. 上海:上海人民出版社,2017:138-173.
Hass R F,Zang X L. Multivariate generalized linear models [M]. Shanghai:Shanghai People's Publishing House,2017:138-173.

[32] 黄杨岳,孔祥祯,甄宗雷,等. 全基因组关联研究中的多重校正方法比较 [J]. 心理科学进展,2013,21(10):1874-1882.
Huang Y Y,Kong X Z,Zhen Z L,et al. The comparison of multiple testing corrections methods in genome-wide association studies [J]. Advances in Psychological Science,2013,21(10):1874-1882.

[33] Porter H F,O'Reilly P F. Multivariate simulation framework reveals performance of multi-trait GWAS methods [J]. Scientific Reports,2017,7(8):132-148.

[34] Falconer D S,Mackay T F C. Introduction to quantitative genetics [M]. 4th ed. Glasgow:Benjamin Cummings,1996.

[35] Lynch M,Walsh B. Genetics and analysis of quantitative traits [M]. Sunderland:Sinauer Associates,1998.

[36] Fisher R A. The correlation between relatives on the supposition of Mendelian inheritance [J]. Transactions of the Royal Society of Edin-burgh,1918;52(2):399-433.

[37] Hua J P,Xing Y Z,Xu C G,et al. Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance [J]. Genetics,2002,162(4):1885-1895.

[38] Xing Y Z,Tan Y F,Hua J P,et al. Characterization of the main effects,epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice [J]. Theoretical and Applied Genetics,2002,105(43499):248-257.

[39] Yu H H,Xie W B,Wang J,et al. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers [J]. PLoS One,2011,6(3):e17595.

[40] Zhou G,Chen Y,Yao W,et al. Genetic composition of yield heterosis in an elite rice hybrid [J]. Proceedings of the National Academy of Sciences of the United States of America,2012,109(39):15847-15852.

(上接第 134 页)

[29] Ansari A,Andalibi B,Zarei M,et al. Combined effect of putrescine and mycorrhizal fungi in phytoremediation of *Lallemantia iberica* in Pb-contaminated soils [J]. 2021,28,58640-58659.

[30] 陈　雪,郑志鑫,石　青,等. AMF 和植物富集土壤中铅和镉的效应 [J]. 菌物研究,2017,15(1):33-38,52.
Chen X,Zheng Z X,Shi Q,et al. Effect of AMF and plants on accumulation of Pb and Cd in soil [J]. Journal of Fungal Research,2017,15(1):33-38, 52.

[31] Yang Y R,Liang Y,Han X Z,et al. The roles of arbuscular mycorrhizal fungi (AMF) in phytoremediation and tree-herb interactions in Pb contaminated soil [J]. Scientific Reports,2016,6:20469.

[32] Gu H H,Zhou Z,Gao Y Q,et al. The influences of arbuscular mycorrhizal fungus on phytostabilization of lead/zinc tailings using four plant species [J]. International Journal of Phytoremediation,2017,19(8):739-745.

[33] González-Chávez M D C A,Ortega-Larrocea M D P,Carrillo-González R,et al. Arsenate induces the expression of fungal genes involved in As transport in arbuscular mycorrhiza [J]. Fungal Biology,2011,115(12):1197-1209.