作物学报 ACTA AGRONOMICA SINICA 2023, 49(9): 2321–2330 http://zwxb.chinacrops.org/
ISSN 0496-3490; CN 11-1809/S; CODEN TSHPA9 E-mail: zwxb301@caas.cn

**DOI: 10.3724/SP.J.1006.2023.23072**

# 基于 Adaptive Lasso 的两阶段全基因组关联分析方法

杨文宇[1,2] 吴成秀[1] 肖英杰[1,3,*] 严建兵[1,3]

[1] 作物遗传改良全国重点实验室, 湖北武汉 430070; [2] 华中农业大学理学院, 湖北武汉 430070; [3] 湖北洪山实验室, 湖北武汉 430070

**摘 要:** 作为进行全基因组关联分析的主流方法, 混合线性模型类方法得到了广泛的应用。但是, 现有方法仍存在检测功效不高的问题。本文提出一种基于 Adaptive Lasso 的 2 阶段全基因组关联分析方法(two-stage Adaptive Lasso-based genome-wide association analysis, ALGWAS), 该方法在第 1 阶段通过变量选择方法 Adaptive Lasso 筛选出与目标性状相关联的单核苷酸多态性位点(single nucleotide polymorphism, SNP), 第 2 阶段将第 1 阶段筛选出的 SNP 作为协变量放入线性模型中进行全基因组扫描。在模拟实验中, ALGWAS 方法与 3 种常用的全基因组关联分析方法 fastGWA、GEMMA 和 EMMAX 相比具有最高的检测功效, 同时具有较低的错误发现率(false discovery rate, FDR)。将以上 4 种方法应用到包含 1341 份材料的玉米 CUBIC (Complete-diallel plus Unbalanced Breeding-like Inter-Cross)群体的全基因组关联分析中, ALGWAS 方法可检测到与开花期相关基因 *ZmMADS69*、*ZmMADS15/31*、*ZmZCN8* 和 *ZmRAP2.7*, 与株高相关基因 *ZmBRD1* 和 *ZmBR2*, 与产量相关基因 *ZmUB2*、*ZmKRN2* 和 *ZmCLE7* 等, 而其他 3 种常用的全基因组关联分析方法检测功效较低。本研究提出了一种非混合线性模型类的全基因组关联分析方法, 对解析微效多基因决定的复杂遗传性状具有更高的检测效率, 为基因挖掘提供了新的途径。

**关键词:** 玉米; 全基因组关联分析; 变量选择; Adaptive Lasso

# ALGWAS: two-stage Adaptive Lasso-based genome-wide association study

YANG Wen-Yu[1,2], WU Cheng-Xiu[1], XIAO Ying-Jie[1,3,*], and YAN Jian-Bing[1,3]

[1] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, Hubei, China; [2] College of Science, Huazhong Agricultural University, Wuhan 430070, Hubei, China; [3] Hubei Hongshan Laboratory, Wuhan 430070, Hubei, China

**Abstract:** As mainstream methods for genome-wide association analysis, mixed linear model methods have been widely used. However, the existing methods still have the problem of low detection power. In this study, a two-stage Adaptive Lasso-based genome-wide association analysis (ALGWAS) method was proposed. In the first stage, single nucleotide polymorphism (SNP) associated with target traits were screened by Adaptive Lasso, a variable selection method. In the second stage, SNPs selected from the first stage were put into the linear model as the covariates for genome-wide scanning. Compared with fastGWA, GEMMA and EMMAX, the ALGWAS method had the highest detection power and lower false discovery rate (FDR) in the simulation experiments. The above four methods were applied to genome-wide association analysis of Complete-diallel plus Unbalanced Breeding-like Inter-Cross (CUBIC) population of 1341 individuals in maize. ALGWAS method can detect the genes (*ZmMADS69*, *ZmMADS15/31*, *ZmZCN8*, and *ZmRAP2.7)* related to days to tasseling, the genes (*ZmBRD1* and *ZmBR2*) related to plant height, and the genes (*ZmUB2*, *ZmKRN2*, and *ZmCLE7*) related to yield, while the other three commonly used genome-wide association analysis methods had low detection efficiency. In this study, a non-mixed linear model class of genome-wide association analysis method was proposed, which had higher detection advantage for microeffect polygenes and provided a new way for genetic analysis of complex traits.

**Keywords:** maize; genome-wide association study; variable selection; Adaptive Lasso

本研究由国家自然科学基金项目(32201855, 32122066)资助。

This study was supported by the National Natural Science Foundation of China (32201855, 32122066).

* 通信作者(Corresponding author): 肖英杰, E-mail: yxiao25@mail.hzau.edu.cn

第一作者联系方式: E-mail: yangwenyurain@126.com

Received (收稿日期): 2022-10-28; Accepted (接受日期): 2023-02-21; Published online (网络出版日期): 2023-03-03.

URL: https://kns.cnki.net/kcms/detail/11.1809.S.20230302.1544.007.html

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

全基因组关联分析(Genome-Wide Association Study, GWAS)是在全基因组水平上分析高密度的SNP 与性状相关性的分析, 从而发现影响复杂性状的基因变异的一种统计方法。遗传学家最先使用的是简单易算的线性模型(Linear Model, LM), 但该模型没有考虑群体结构的影响, 会挖掘出很多基因位点与复杂性状的假阳性关联。在一般线性模型中控制群体结构效应后, 假阳性检测大大降低。此外, 遗传学家发现复杂的亲缘关系也会带来假阳性的关联结果, 因此 Zhang 等[1]和 Yu 等[2]提出了混合线性模型。混合线性模型能同时控制群体结构和亲缘关系的影响, 降低了假阳性率。此后, 很多研究者致力于改善混合线性模型。Kang 等[3]2008 年提出有效的混合线性模型(Efficient Mixed-Model Association, EMMA)。EMMA 是一种被广泛使用的精确方法, 它将求解混合线性模型时涉及的优化问题转化成了一维的优化问题, 提高了计算效率, 并通过谱分解方法避免了每次迭代计算似然函数时的大量矩阵相乘和求逆运算, 进一步提高了计算效率。但是 EMMA 难以处理由数千个个体组成的数据集, 为了解决这个问题, Kang 等[4]2010 年在 EMMA 的基础上提出了 EMMAX (EMMA eXpedited)。EMMAX 将 EMMA 扫描每个位点时均需估计的多基因方差与残差方差之比用无效应模型中得到的比值近似, 大幅减少了计算量。同年, Zhang 等[5]在混合线性模型的基础上提出了压缩的混合线性模型(Compressed MLM, CMLM)和 P3D (Population Parameters Previously Determined)方法。CMLM 采用聚类方法将群体进行分组, 减少了有效样本数量。P3D 通过固定多基因方差与残差方差的比值, 减少了全基因组扫描时需要估计的参数数目, 提升了计算效率。2012 年 Zhou 等[6]提出一种高效的精确方法, 全基因组高效混合线性模型(Genome-wide Efficient Mixed-Model Association, GEMMA)。GEMMA 大约比 EMMA 快 *n* 倍(*n* 为样本数目), 它的出现使得处理大样本数据集时采用精确全基因组关联分析方法变得可行。近年来, 混合线性模型类方法得到了广泛的应用[7-13], 与之相关的快速算法也陆续被提出, 如 Fast-LMM[14]、Fast-LMM-Select[15]和 BOLT-LMM[16]等。2019 年 Jiang 等[17]针对大规模数据分析, 开发了一种基于混合线性模型的新方法 fastGWA, 它通过将亲缘关系矩阵中较小系数替换成 0 值, 增加矩阵稀疏性, 提高了模型功效和运算速度, 并用模拟实验证明了 fastGWA 的可靠性和鲁棒性。

在过去的几十年, GWAS 在人类、动物和植物中识别了成千上万的相关基因座, 为疾病诊断和动植物育种提供了帮助。但是, GWAS 识别出的基因座只能解释很小的一部分表型变异, 这种现象被称为“消失的遗传力”[18]。例如, GWAS 识别到了约 50 个与人类身高相关的基因座, 但是他们仅能解释 5%的身高变异[19]。Yang 等 2010 年指出遗传力并没有消失, 而是基因组中存在大量的微效位点 GWAS 检测不到[20]。这说明长期以来复杂性状 GWAS 一直都存在检测功效不足的问题。为了提高 GWAS 的检测功效, 主要有以下 3 个方面的探索: (1) 增加标记的类型, Song 等[21]采用 InDel (short insertion/deletion)作为标记进行 GWAS 分析, 发现使用 SNP 进行 GWAS 检测不到的 *TFL1* 基因; (2) 采用多变量模型, Zhang 等[22]通过模拟实验和真实数据验证了多位点模型 MrMLM 的优越性; (3) 采用非参数模型, Yang 等[23]提出 A-D test 方法, 对不服从正态分布的表型可提高 GWAS 的检测功效。本研究在参数模型的范畴下, 为了提高 GWAS 的检测功效提出一种基于 Adaptive Lasso 的 2 阶段全基因组关联分析方法(ALGWAS), 该方法先通过 Adaptive Lasso 筛选出与目标性状相关联的 SNP, 再将筛选出的 SNP 作为协变量放入一般线性模型中进行全基因组扫描。本研究选用包含 1341 份材料的玉米 CUBIC 群体的基因型和模拟的表型, 采用 2 种模拟方法进行数值实验, 并与 3 种常用的全基因组关联分析方法 fastGWA、GEMMA 和 EMMAX 进行对比。试验结果显示 ALGWAS 具有最高的检测功效且具有较低的错误发现率。

本文使用以上 4 种方法对玉米 CUBIC 群体的开花期、株高和产量数据进行全基因组关联分析, 发现 ALGWAS 方法可检测到与开花期相关的已知基因 *ZmMADS69*、*ZmMADS15/31*、*ZmZCN8* 和 *ZmRAP2.7*, 与株高相关的已知基因 *ZmBRD1* 和 *ZmBR2*, 与产量相关的已知基因 *ZmUB2*、*ZmKRN2* 和 *ZmCLE7* 等, 而其他 3 种常用的全基因组关联分析方法只能检测到少量已知基因。

## 1 材料与方法

### 1.1 试验材料的基因型和表型

本研究所用的 1341 份材料来源于玉米 CUBIC 群体[24]。该群体通过以“黄改系”为核心的 24 个优良玉米自交系作为亲本, 采用一代不完全的双列杂交

和 6 代的随机交配, 再进行 6 代的连续自交得到。利用第 2 代测序技术对 CUBIC 群体的 1341 个后代自交系进行低覆盖度的测序(~1X), 选择最小等位基因频率大于 0.02, 获得 11,800,000 高质量的 SNP, 本文从中随机挑选标记 60,000 个。在全国选取 5 个典型玉米种植生态区种植 CUBIC 群体, 进行大规模的田间表型试验。对每份材料调查抽雄期(days to tasseling)、株高(plant height)和穗重(ear weight)性状。本研究利用的基因型和表型性状数据来自 Liu 等[24]已发表文章。

### 1.2　模拟试验方案

为了测试 ALGWAS 方法, 本研究采用 2 种方案来模拟表型。在 2 种方案中均使用 CUBIC 群体的基因型。第 1 种方案是从头模拟表型, 在 60,000 个 SNP 标记中随机选择 20 或 50 指定为 QTN (Quantitative Trait Nucleotides), 并设置 2 个水平的狭义遗传力($h^2$=0.5, 0.8)。QTN 的加性遗传效应服从几何级数分布, 第 $i$ 个 QTN 的效应为 $a^i$ (当 QTN 数目为 20 时, $a$=0.9; 当 QTN 数目为 50 时, $a$=0.96)[25-26]。由于 QTN 的加性遗传效应按指数递减, 本研究按照 QTN (20 个/50 个)效应大小将其分为高效应(6 个/15 个)、中等效应(8 个/20 个)和低效应(6 个/15 个) 3 种。第 $j$ 个自交系的模拟表型值为: $\sum_i z_{ji}a^i+\varepsilon_j$, $z_{ji}$ 和 $\varepsilon_j$ 分别代表基因型和残差。第 2 种方案是基于真实性状遗传结构模拟表型, 在保留原有性状遗传结构的前提下测试 ALGWAS 方法, 在观察到的表型上添加额外的 QTN 效应, QTN 的效应用表型标准差的 $l$ (l=0.1, 0.15, …, 0.5)倍来表示, 即第 $j$ 个自交系的模拟表型为: $y_j+z_{ji}ly_{sd}$, $y_j$ 为原始表型, $z_{ji}$ 和 $y_{sd}$ 分别代表基因型和原始表型的标准差。2 种方案的模拟实验均随机重复 50 次。生成的模拟表型数据用 60,000 个 SNP 标记, 利用多种方法进行 GWAS 分析, 定义的 QTN 左右 500 kb 内有显著的 SNP 则认为该 QTN 被检测到, QTN 被 GWAS 检测显著的比例定义为统计功效(Power), 检测到非 QTN 占 GWAS 检测显著位点比例定义为错误发现率(False Discovery Rate, FDR)。

### 1.3　ALGWAS 方法

1.3.1　ALGWAS 方法第 1 阶段　　ALGWAS 方法第 1 阶段利用 Adaptive Lasso 模型快速对目标性状关联的 SNP 标记进行初步筛选。考虑线性模型: $y=X\beta+\varepsilon$, 这里 $X=[x_1, \ldots, x_p]$是 $n\times p$ 设计矩阵, $\varepsilon\sim N(0, I\sigma^2)$, $\beta=(\beta_1, \ldots, \beta_p)^T$ 是 $p\times 1$ 的未知系数向量。Lasso 估计定义为: $\hat{\beta}(\text{lasso})=\arg\min_\beta\left\|y-\sum_{j=1}^{p}x_j\beta_j\right\|^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$, 这里 $\lambda$ 是惩罚系数[27]。由定义可知 Lasso 对每个系数施加了相同的惩罚, 忽略了系数间重要性的差别, 所以 Zou 提出了加权的 Lasso, 即 Adaptive Lasso[28], 定义: $\hat{\beta}=\arg\min_\beta\left\|y-\sum_{j=1}^{p}x_j\beta_j\right\|^2+\lambda\sum_{j=1}^{p}w_j\left|\beta_j\right|$, 这里 $w=(w_1, \ldots, w_p)^T$ 是 $p\times 1$ 的已知权重向量。与 Lasso 相比, Adaptive Lasso 对零系数施加较大的惩罚同时对非零系数施加较小的惩罚, 减小了估计的偏差、提高了变量选择的准确性。

1.3.2　ALGWAS 方法的第 2 阶段　　假设第 1 阶段由 Adaptive Lasso 筛选出与性状相关的 SNP 集合为 B, 第 2 阶段全基因组扫描到第 $k$ 个 SNP, 定义 ALGWAS 方法第 2 阶段的模型为:

$$y=Xb+z_k\gamma_k+\varepsilon, \quad (1)$$

这里 $y$ 是 $n\times 1$ 表型向量, $z_k$ 是 $n\times 1$ 基因型向量, $\gamma_k$ 是第 $k$ 个 SNP 效应, $X=[x_1, \ldots, x_{q+1}]$是 $n\times(q+1)$设计矩阵, $x_{q+1}=(1, \ldots, 1)^T$, $b=(b_1, \ldots, b_{q+1})^T$是$(q+1)\times 1$系数向量, $b_{q+1}$ 为模型(1)的截距, $q$ 为集合 B 中选出进入模型(1) SNP 的个数($q<|B|$), $\varepsilon\sim N(0, I\sigma^2)$。规定扫描窗口大小为 10 Mb, 即第 $k$ 个 SNP 左右 5 Mb 以外的集合 B 中的 SNP, 作为检测第 $k$ 个 SNP 的协变量进入模型(1)。ALGWAS 方法的 R 语言程序可从 github (https://github.com/yangwenyurain/ALGWAS.git)下载。

### 1.4　常用 GWAS 方法

1.4.1　线性模型(LM)方法　　线性模型为: $y=\beta+Z_k\gamma_k+\varepsilon$, 这里 $y$ 是 $n\times 1$ 表型向量, $\beta$ 是截距, $Z_k$ 是 $n\times 1$ 基因型向量, $\gamma_k$ 是第 $k$ 个 SNP 的效应, $\varepsilon\sim N(0, I\sigma^2)$。

1.4.2　混合线性模型方法　　混合线性模型为: $y=X\beta+Zu+\varepsilon$, $y$ 是 $n\times 1$ 表型向量, $X$ 是固定效应对应的 $n\times p$ 设计矩阵, $\beta$ 是 $p\times 1$ 代表固定效应的系数向量, $Z$ 是随机效应对应的 $n\times n$ 设计矩阵, 多基因效应 $u\sim N(0, K\sigma_g^2)$, $K$ 为亲缘关系矩阵, 残差效应向量 $\varepsilon\sim N(0, I\sigma_e^2)$, $I$ 为单位矩阵, $\sigma_g^2$ 和 $\sigma_e^2$ 分别为估计的遗传方差和残差方差。本研究利用 EMMAX[4]、GEMMA[6]和 fastGWA[17] 3 种常用的混合线性模型进行模拟数据和真实数据的 GWAS 分析。

## 2 结果与分析

### 2.1 模拟试验方案 1: 从头模拟表型

利用 CUBIC 群体基因型数据, 定义 20 个和 50 个 QTN, 狭义遗传力为 0.5 和 0.8, 共 4 个模拟组合, 随机重复 50 次后, 共得到 200 个模拟表型。使用 LM、EMMAX、GEMMA、fastGWA 和 ALGWAS 分别对其进行全基因组关联分析, 得到的平均结果见表 1。从表 1 可以看出, ALGWAS 与 EMMAX、GEMMA 和 fastGWA 相比具有最高的平均检测功效和较低的错误发现率, 进一步可以看出 ALGWAS 检测功效高的原因在于 ALGWAS 对于低效应的 QTN 平均检测功效比较高。当 QTN 数目为 20, 遗传力为 0.8 时, ALGWAS 的平均检测功效为 0.802, fastGWA 的检测功效为 0.457, ALGWAS 对于低效应 QTN 的平均检测功效为 0.48, 比 fastGWA 的平均检测功效 0.04 高 12 倍。

### 2.2 模拟实验方案 2: 基于真实性状遗传结构的表型模拟

在 CUBIC 群体观察到的表型抽雄期、株高和穗重上分别随机选择标记, 添加 1 个 QTN 效应, QTN 的效应设置为表型标准差的 0.1 倍至 0.5 倍, 重复 50 次后, 共得到 1350 个模拟表型。使用 EMMAX、GEMMA、fastGWA 和 ALGWAS 分别对其进行全基因组关联分析, 得到的平均结果如图 1。从图 1 可以看出在不同表型上添加 QTN 效应, ALGWAS 均具有最高的平均检测功效, 尤其是添加小效应 QTN 时, ALGWAS 的优势更明显, 例如在穗重表型上添加表型标准差 0.1 倍的 QTN 效应时, EMMAX、GEMMA 和 fastGWA 的平均检测功效均为 0, 而 ALGWAS 的检测功效为 0.12。

### 2.3 真实数据结果

考虑 CUBIC 群体观察到的表型抽雄期、株高和穗重, 使用 EMMAX、GEMMA、fastGWA 和 ALGWAS 分别对其进行全基因组关联分析(图 2~图 4)。可以看出 EMMAX、GEMMA 和 fastGWA 方法检测到的 QTL, ALGWAS 均可检测到, 并且 ALGWAS 还可检测到更多的 QTL, 这说明了 ALGWAS 有更高检测功效。对于抽雄期, ALGWAS 方法可检测到与开花期相关的基因 *ZmMADS69*、*ZmMADS15/31*、*ZmZCN8*

**表 1 基于从头模拟表型的不同全基因组关联分析方法的平均检测功效和错误发现率**

**Table 1 Average detection power and false discovery rate of genome-wide association study methods**

| 遗传力 Heritability | QTN 数目 Number of QTN | 方法 Method | 检测功效 Detection power | 高效应 QTN 检测功效 Detection power of QTN with high effect | 中等效应 QTN 检测功效 Detection power of QTN with moderate effect | 低效应 QTN 检测功效 Detection power of QTN with low effect | 错误发现率 False discovery rate |
|---|---|---|---|---|---|---|---|
| 0.8 | 20 | LM | 0.811 | 0.997 | 0.838 | 0.590 | 0.853 |
| 0.8 | 20 | **ALGWAS** | **0.802** | **1.000** | **0.895** | **0.480** | **0.040** |
| 0.8 | 20 | GEMMA | 0.482 | 0.987 | 0.440 | 0.033 | 0.066 |
| 0.8 | 20 | fastGWA | 0.457 | 0.983 | 0.375 | 0.040 | 0.074 |
| 0.8 | 20 | EMMAX | 0.446 | 0.983 | 0.358 | 0.027 | 0.050 |
| 0.5 | 20 | LM | 0.808 | 0.947 | 0.825 | 0.647 | 0.811 |
| 0.5 | 20 | **ALGWAS** | **0.562** | **0.793** | **0.565** | **0.327** | **0.105** |
| 0.5 | 20 | GEMMA | 0.442 | 0.777 | 0.423 | 0.133 | 0.070 |
| 0.5 | 20 | fastGWA | 0.382 | 0.697 | 0.353 | 0.107 | 0.074 |
| 0.5 | 20 | EMMAX | 0.370 | 0.683 | 0.338 | 0.100 | 0.069 |
| 0.8 | 50 | LM | 0.763 | 0.937 | 0.769 | 0.581 | 0.780 |
| 0.8 | 50 | **ALGWAS** | **0.513** | **0.911** | **0.520** | **0.107** | **0.065** |
| 0.8 | 50 | GEMMA | 0.270 | 0.729 | 0.128 | 0.001 | 0.025 |
| 0.8 | 50 | fastGWA | 0.250 | 0.699 | 0.101 | 0.001 | 0.030 |
| 0.8 | 50 | EMMAX | 0.244 | 0.684 | 0.095 | 0.003 | 0.020 |
| 0.5 | 50 | LM | 0.600 | 0.843 | 0.571 | 0.397 | 0.708 |
| 0.5 | 50 | **ALGWAS** | **0.208** | **0.561** | **0.094** | **0.008** | **0.118** |
| 0.5 | 50 | GEMMA | 0.158 | 0.479 | 0.036 | 0.000 | 0.045 |
| 0.5 | 50 | fastGWA | 0.134 | 0.416 | 0.022 | 0.000 | 0.060 |
| 0.5 | 50 | EMMAX | 0.129 | 0.404 | 0.020 | 0.000 | 0.063 |

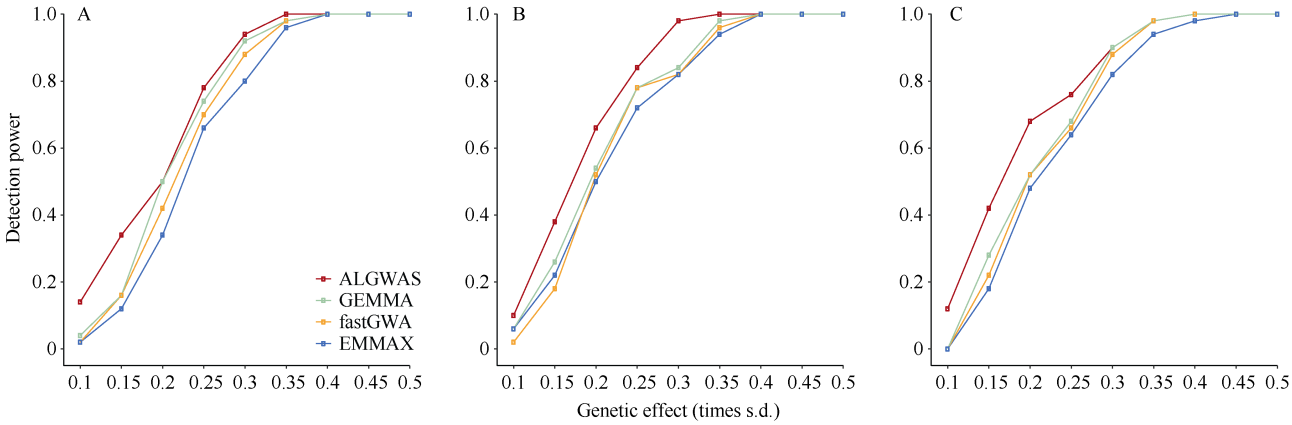



**图 1　基于真实性状遗传结构模拟表型的不同全基因组关联分析方法的检测功效**

**Fig. 1　Average detection power of genome-wide association study methods based on simulated phenotype on realistic genetic structure**

A: 抽雄期; B: 株高; C: 穗重。A: days to tasseling; B: plant height; C: ear weight.

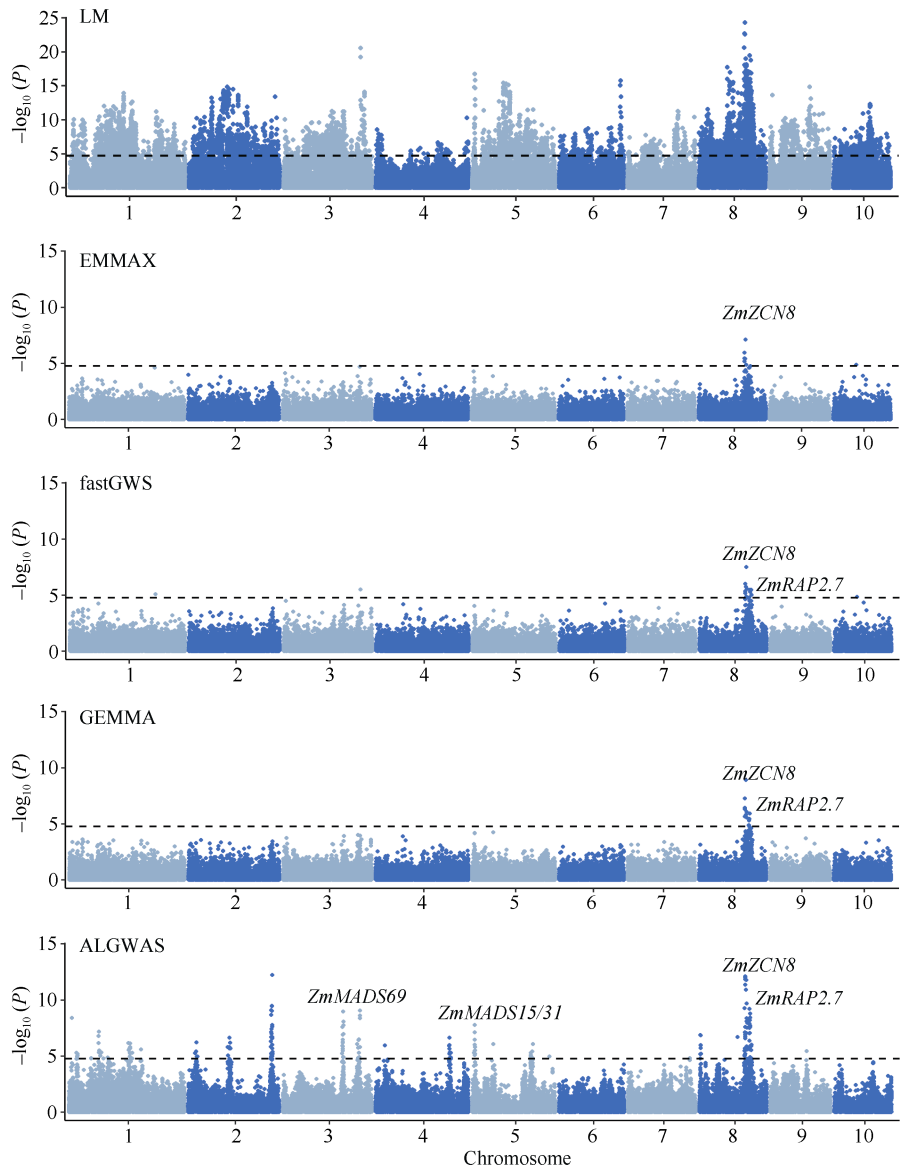



**图 2　CUBIC 群体抽雄期的曼哈顿图**

**Fig. 2　Manhattan plots for days to tasseling of CUBIC population**

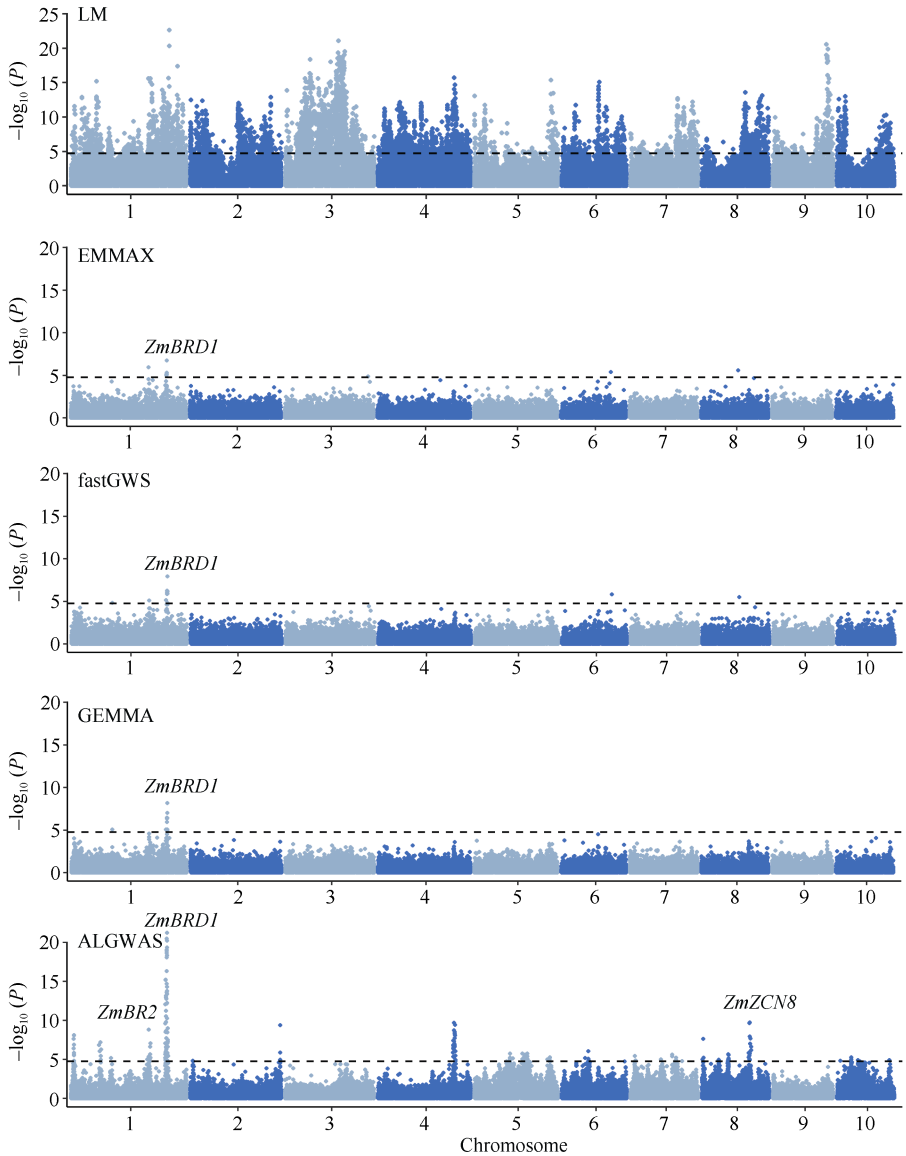
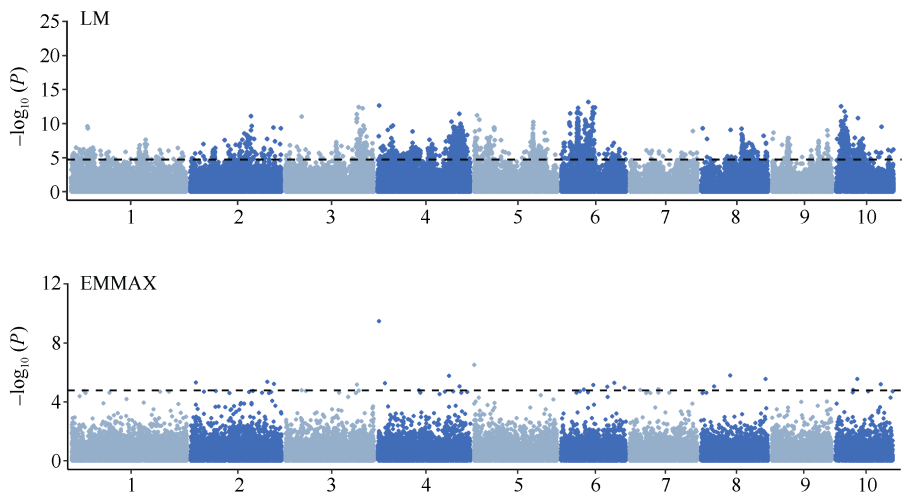



**图 3 CUBIC 群体株高的曼哈顿图**

**Fig. 3 Manhattan plots for plant height of CUBIC population**




**(图 4)**

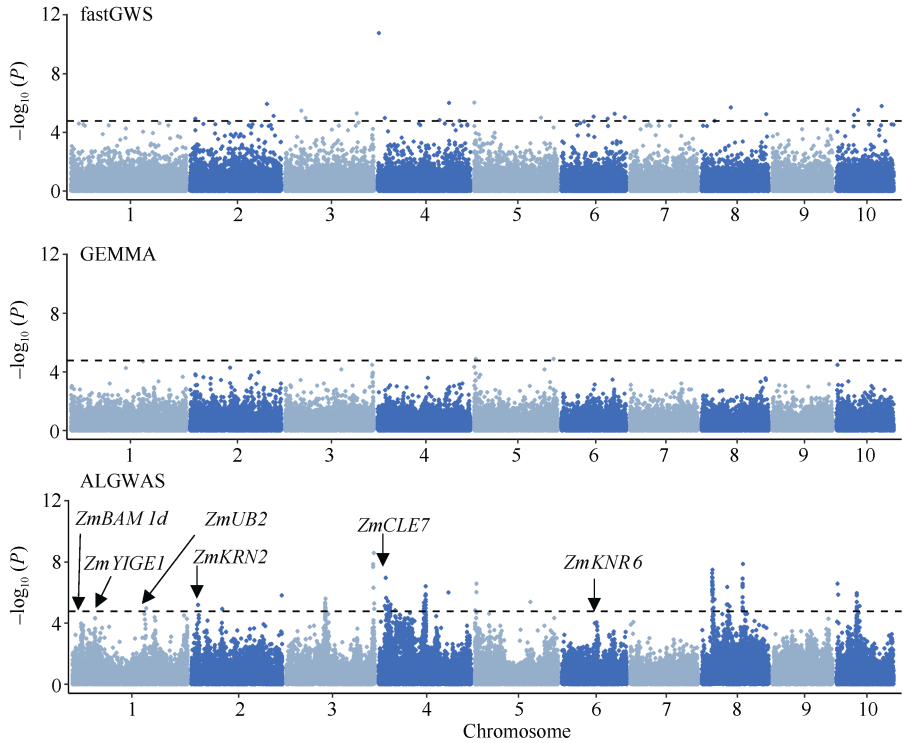



**图 4　CUBIC 群体穗重的曼哈顿图**

**Fig. 4　Manhattan plots for ear weight of CUBIC population**

和 *ZmRAP2.7*[24,29]，而 GEMMA 和 fastGWA 只能检测到基因 *ZmZCN8* 和 *ZmRAP2.7*, EMMAX 仅能检测到基因 *ZmZCN8*。对于株高, ALGWAS 可检测到与株高相关的基因 *ZmBRD1*[30]和 *ZmBR2*[31]，并检测到基因 *ZmZCN8*，该基因通过延迟开花进而影响株高，而其他 3 种方法只能检测到基因 *ZmBRD1*。对于穗重, ALGWAS 方法可检测到与产量相关的基因 *ZmBAM1d*[32]、*ZmY1GE1*[33]、*ZmUB2*[34]、*ZmKRN2*[35]、*ZmCLE7*[36]和 *ZmKNR6*[37]，而其他 3 种方法几乎检测不到相关基因。通过 QQ 图，可以发现 ALGWAS 相比于其他 3 种常用的混合线性模型方法均具有更高的统计功效，同时对背景噪音导致的假阳性有较好的控制(图 5)。ALGWAS 方法检测到的已知基因位置及其对应的 peakSNP 位置见表 2。

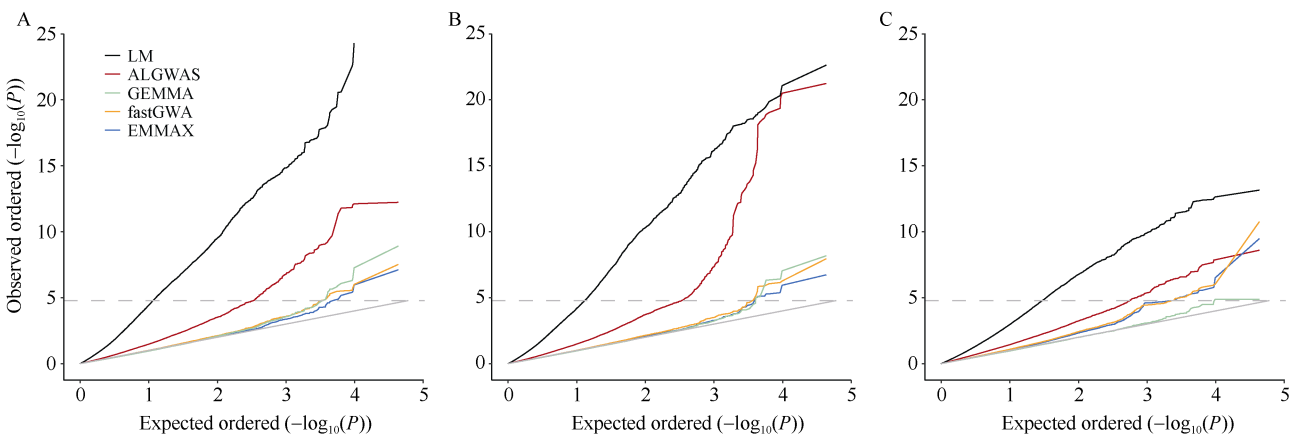



**图 5　不同全基因组关联分析方法的 QQ 图**

**Fig. 5　Quantitle-quantitle plot of genome-wide association study methods**

A：抽雄期; B：株高; C：穗重。A: days to tasseling; B: plant height; C: ear weight.

表 2 ALGWAS 方法检测到的已知基因位置及其对应的 peakSNP 位置

**Table 2 ALGWAS detected the known gene location and its corresponding peakSNP location**

| 基因名称 Gene name | 基因 ID Gene ID | 基因位置 Gene location | peakSNP 位置 peakSNP location | *P* 值 *P*-value |
|---|---|---|---|---|
| *ZmMADS69* | *GRMZM2G171650* | Chr. 3: 159,022,119..159,050,063 | Chr. 3: 158,013,626 | 8.434874E-09 |
| *ZmMADS15/31* | *GRMZM2G553379* | Chr. 5: 6,993,294..7,011,505 | Chr. 5: 6,684,616 | 1.667779E-08 |
| *ZmZCN8* | *GRMZM2G179264* | Chr. 8: 123,028,887..123,033,675 | Chr. 8: 121,955,506 | 1.638071E-12 |
| *ZmRAP2.7* | *GRMZM2G700665* | Chr. 8: 131,575,389..131,581,816 | Chr. 8: 131,013,374 | 6.050159E-10 |
| *ZmBR2* | *GRMZM2G315375* | Chr. 1: 20,2334,824..202,342,008 | Chr. 1: 202,814,745 | 1.595160E-09 |
| *ZmBRD1* | *GRMZM2G103773* | Chr. 1: 249,371,977..249,376,239 | Chr. 1: 249,462,103 | 3.165227E-21 |
| *ZmZCN8* | *GRMZM2G107829* | Chr. 8: 123,028,887..123,033,675 | Chr. 8: 124,009,899 | 2.306359E-10 |
| *ZmBAM1d* | *GRMZM2G043584* | Chr. 1: 30,516,319..30,522,796 | Chr. 1: 27,500,593 | 1.373829E-04 |
| *ZmYIGE1* | *GRMZM2G008490* | Chr. 1: 51,075,171..51,161,917 | Chr. 1: 58,798,749 | 4.593378E-05 |
| *ZmUB2* | *GRMZM2G160917* | Chr. 1: 188,213,876..188,220,983 | Chr. 1: 192,566,425 | 1.078588E-05 |
| *ZmKRN2* | *GRMZM2G125656* | Chr. 2: 17,742,986..17,750,216 | Chr. 2: 17,280,224 | 6.421449E-06 |
| *ZmCLE7* | *GRMZM2G372364* | Chr. 4: 7,568,824..7,572,604 | Chr. 4: 16,087,155 | 7.774088E-06 |
| *ZmKRN6* | *GRMZM2G119714* | Chr. 6: 94,188,754..94,201,186 | Chr. 6: 90,084,654 | 8.809443E-05 |

## 3 讨论

ALGWAS 的第 1 阶段需要筛选与性状相关的 SNP, 这一步可通过变量选择方法实现, 本研究选用的是 Adaptive Lasso 方法, 因为 Zou 给出了该方法具有一致性的理论证明[28]。本研究提供的是一个 2 阶段方法的框架, 其他的变量选择方法也可用于 ALGWAS, 比如机器学习方法。在实际 GWAS 研究中, 如全基因组 SNP 数目达到百万级别时, ALGWAS 的变量筛选阶段建议从中随机抽取一部分 SNP 作分析。

ALGWAS 方法虽然在检测功效上具有优势, 但是它本身也有一定的局限性。ALGWAS 的第 2 阶段进行单点扫描时, 每一次都需要对进入模型的协变量进行判断, 这一步导致了 ALGWAS 的速度还有待提高, 在后期的研究中, 我们将参考 EMMAX[4]的做法, 通过固定进入模型的协变量来对其进行提速。

为了进一步提升 ALGWAS 方法的检测功效, 可以参考 Li 等[38]在 CIM (Composite Interval Mapping) 的基础上提出 ICIM (Inclusive CIM)[39]的作法, 将 ALGWAS 第 1 阶段通过 Adaptive Lasso 方法得到的 SNP 优化权重直接用于第 2 阶段模型的学习。采用此方法也可进一步对 ALGWAS 方法进行提速。

## 4 结论

本研究提出了一种基于 Adaptive Lasso 的 2 阶段全基因组关联分析方法 ALGWAS, 相比于目前常用的混合线性模型 GWAS 方法, ALGWAS 在较好控制假阳性情况下, 统计功效更高, 特别对于产量等微效多基因遗传的性状, ALGWAS 具有明显的检测优势, 这为复杂性状解析提供了新的解决途径。

## References


[1] Zhang Y M, Mao Y C, Xie C Q, Smith H, Luo L, Xu S Z. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics*, 2005, 169: 2267–2275.

[2] Yu J M, Pressoir G, Briggs H W, Vroh B I, Yamasakiet M, Doebley J F, McMullen M D, Gaut B S, Nielsen D M, Holland J B, Kresovich S, Buckler E S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 2006, 38: 203–208.

[3] Kang H M, Zaitlen N A, Wade C M, Kirby A, Heckerman D, Daly M J, Eskin E. Efficient control of population structure in model organism association mapping. *Genetics*, 2008, 178: 1709–1723.

[4] Kang H M, Sul J H, Service S K, Zaitlen N A, Kong S Y, Freimer N B, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 2010, 42: 348–354.

[5] Zhang Z W, Ersoz E, Lai C Q, Todhunter R J, Tiwari H K, Gore M A, Bradbury P J, Yu J, Arnett D K, Ordovas J M, Buckler E S. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet*, 2010, 42: 355–360.

[6] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 2012, 44: 821–824.

[7] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature*, 2007, 447: 661–678.

[8] Li H, Peng Z Y, Yang X H, Wang W D, Fu J J, Wang J H, Han Y J, Chai Y C, Guo T T, Yang N, Liu J, Warburton M L, Cheng Y B, Hao X M, Zhang P, Zhao J Y, Liu Y J, Wang G Y, Li J S, Yan J B. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet*, 2013, 45: 43–50.

[9] Huang X H, Wei X H, Sang T, Zhao Q, Feng Q, Zhao Y, Li C Y, Zhu C R, Lu T T, Zhang Z W, Li M, Fan D L, Guo Y L, Wang A, Wang L, Deng L W, Li W J, Lu Y Q, Weng Q J, Liu K Y, Huang T, Zhou T Y, Jing Y F, Li W, Lin Z, Buckler E S, Qian Q, Zhang Q F, Li J Y, Han B. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet*, 2010, 42: 961–969.

[10] Xiao Y J, Liu H J, Wu L J, Warburton M L, Yan J B. Genome-wide association studies in maize: praise and stargaze. *Mol Plant*, 2017, 10: 359–374.

[11] 彭勃, 赵晓雷, 王奕, 袁文娅, 李春辉, 李永祥, 张登峰, 石云素, 宋燕春, 王天宇, 黎裕. 玉米叶向值的全基因组关联分析. 作物学报, 2020, 46: 819–831.
Peng B, Zhao X L, Wang Y, Yuan W Y, Li C H, Li Y X, Zhang D F, Shi Y S, Song Y C, Wang T Y, Li Y. Genome-wide association studies of leaf orientation value in maize. *Acta Agron Sin*, 2020, 46: 819–831 (in Chinese with English abstract).

[12] 谢磊, 任毅, 张新忠, 王继庆, 张志辉, 石书兵, 耿洪伟. 小麦穗发芽性状的全基因组关联分析. 作物学报, 2021, 47: 1891–1902.
Xie L, Ren Y, Zhang X Z, Wang J Q, Zhang Z H, Shi S B, Geng H W. Genome-wide association study of pre-harvest sprouting traits in wheat. *Acta Agron Sin*, 2021, 47: 1891–1902 (in Chinese with English abstract).

[13] 杨飞, 张征锋, 南波, 肖本泽. 水稻产量相关性状的全基因组关联分析及候选基因筛选. 作物学报, 2022, 48: 1813–1821.
Yang F, Zhang Z F, Nan B, Xiao B Z. Genome-wide association analysis and candidate gene selection of yield related traits in rice. *Acta Agron Sin*, 2022, 48: 1813–1821 (in Chinese with English abstract).

[14] Lippert C, Listgarten J, Liu Y, Kadiel C M, Davidson R I, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods*, 2011, 8: 833–835.

[15] Listgarten J, Lippert C, Kadie C M, Davidson R I, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. *Nat Methods*, 2012, 9: 525–526.

[16] Loh P R, Bhatia G, Gusev A, Finucane H K, Bulik-Sullivan B K, Pollack S J. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet*, 2015, 47: 1385–1392.

[17] Jiang L D, Zheng Z L, Qi T, Kemper K E, Wray N R, Visscher P M, Yang J. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet*, 2019, 51: 1749–1755.

[18] Maher B. Personal genomes: the case of the missing heritability. *Nature*, 2008, 456: 18–21.

[19] Visscher P. Sizing up human height variation. *Nat Genet*, 2008, 40: 489–490.

[20] Yang J, Benyamin B, McEvoy B P, Gordon S, Henders A K, Nyholt D R, Madden P A, Heath A C, Martin N G, Montgomery G W, Goddard M E, Visscher P M. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 2010, 42: 565–569.

[21] Song B, Mott R, Gan X. Recovery of novel association loci in *Arabidopsis thaliana* and *Drosophila melanogaster* through leveraging INDELs association and integrated burden test. *PLoS Genet*, 2018, 14: e1007699.

[22] Zhang Y W, Tamba C L, Wen Y J, Li P, Ren W L, Ni Y L, Gao J, Zhang Y M. mrMLM v4.0.2: an R platform for multi-locus genome-wide association studies. *Genom Prot Bioinfor*, 2020, 18: 481–487.

[23] Yang N, Lu Y L, Yang X H, Huang J, Zhou Y, Ali F H, Wen W W, Liu J, Li J S, Yan J B. Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet*, 2014, 10: e1004573.

[24] Liu H J, Wang X Q, Xiao Y J, Luo J Y, Qiao F, Yang W Y, Zhang R Y, Meng Y J, Sun J M, Yan S J, Peng Y, Niu L Y, Jian L M, Song W, Yan J L, Li C H, Zhao Y X, Liu Y, Warburton M L, Zhao J R, Yan J B. CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biol*, 2020, 21: 20.

[25] Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 1990, 124: 743–756.

[26] Yu J M, Holland J B, McMullen M D, Buckler E S. Genetic design and statistical power of nested association mapping in maize. *Genetics*, 2008, 178: 539–551.

[27] Tibshirani R. Regression shrinkage and selection *via* the lasso. *J Royal Statist Society*, 1996, 58: 267–288.

[28] Zou H. The adaptive lasso and its oracle properties. *J Am Statist Assoc*, 2006, 101: 1418–1429.

[29] Liang Y M, Liu Q, Wang X F, Huang C, Xu G H, Hey S, Lin H Y, Li C, Xu D Y, Wu L S, Wang C L, Wu W H, Xia J L, Han X, Lu S J, Lai J S, Song W B, Schnable P S, Tian F. ZmMADS69 functions as a flowering activator through the regulatory module and contributes to maize flowering time adaptation. *New Phytol*, 2019, 221: 2335–2347.

[30] Makarevitch I, Thompson A, Muehlbauer G J, Springer N M. *Brd1* gene in maize encodes a brassinosteroid C-6 oxidase. *PLoS One*, 2012, 7: e30798.

[31] Xing A Q, Gao Y F, Ye L F, Zhang W P, Cai L C, Ching A, Llaca V, Johnson B, Liu L, Yang X H, Kang D M, Yan J B, Li J S. A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. *J Exp Bot*, 2015, 66: 3791–3802.

[32] Yang N, Liu J, Gao Q, Gui S T, Chen L, Yang L F, Huang J, Deng T Q, Luo J Y, He L J, Wang Y B, Xu P W, Peng Y, Shi Z, Lan L, Ma Z Y, Yang X, Zhang Q Q, Bai M Z, Li W, Liu L, Jackson D, Yan J B. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat Genet*, 2019, 51: 1052–1059.

[33] Luo Y, Zhang M L, Liu Y, Liu J, Li W Q, Chen G S, Peng Y, Jin M, Wei W J, Jian L M, Yan J, Fernie A R, Yan J B. Genetic variation in YIGE1 contributes to ear length and grain yield in maize. *New Phytol*, 2022, 234: 513–526.

[34] Du Y F, Liu L, Peng Y, Li M F, Li Y F, Liu D, Li X W, Zhang Z X. *UNBRANCHED3* expression and inflorescence development is mediated by *UNBRANCHED2* and the distal enhancer, *KRN4*, in maize. *PLoS Genet*, 2020, 16: e1008764.

[35] Chen W K, Chen L, Zhang X, Yang N, Guo J H, Wang M, Ji S G, Zhao X Y, Yin P F, Cai L C, Xu J, Zhang L L, Han Y J, Xiao Y N, Xu G, Wang Y B, Wang S H, Wu S, Yang F, Jackson D, Cheng J K, Chen S H, Sun C Q, Qin F, Tian F, Fernie A R, Li J S, Yan J B, Yang X H. Convergent selection of a WD40 protein that enhances grain yield in maize and rice. *Science*, 2022, 375: e7985.

[36] Liu L, Gallagher J, Arevalo E D, Chen R, Skopelitis T, Wu Q, Bartlett M, Jackson D. Enhancing grain-yield-related traits by CRISPR-Cas9 promoter editing of maize CLE genes. *Nat Plants*, 2021, 7: 287–294.

[37] Jia H T, Li M F, Li W Y, Liu L, Jian Y N, Yang Z X, Shen X M, Ning Q, Du Y F, Zhao R, Jackson D, Yang X H, Zhang Z X. A serine/threonine protein kinase encoding gene KERNEL NUMBER PER ROW6 regulates maize grain yield. *Nat Commun*, 2020, 11: 988.

[38] Zeng Z B. Precision mapping of quantitative trait loci. *Genetics,* 1994, 136: 1457–1468.

[39] Li H H, Ye G Y, Wang J K. A modified algorithm for the improvement of composite interval mapping. *Genetics*, 2007, 175: 361–374.