作物学报 ACTA AGRONOMICA SINICA 2026, 52(6): 1743−1756 http://zwxb.chinacrops.org/
ISSN 0496-3490; CN 11-1809/S; CODEN TSHPA9 E-mail: zwxb301@caas.cn

DOI: 10.3724/SP.J.1006.2026.55064

大豆株高性状全基因组关联分析及预测模型构建

唐宽强[1] 李公允[1] 宋美毅[1] 赵 雪[1] 常春玲[1,2,3,4,*]

[1] 齐齐哈尔大学生命科学与农林学院, 黑龙江齐齐哈尔 161006; [2] 黑龙江省农业微生物制剂产业化工程技术研究中心, 黑龙江齐齐哈尔 161006; [3] 黑龙江省农用生物制剂产业化协同创新中心, 黑龙江齐齐哈尔 161006; [4] 黑龙江省高浓度有机废液转化及产业化概念验证中心, 黑龙江齐齐哈尔 161006

摘 要: 株高是影响大豆产量与抗倒伏能力的核心农艺性状, 对保障大豆高产稳产具有重要意义。对 17,148 份世界范围内的大豆种质资源进行全基因组关联分析, 共鉴定到 139 个与株高性状显著关联的区间, 其中候选基因 *Dt1* 对株高的贡献最大。在全基因组选择模型中, BayesA 和 rrBLUP 模型优于其他模型, 基于前 500~1000 个 SNP 的预测模型在精度与成本效益间达到最优平衡。本研究鉴定的调控株高的关键基因位点, 为大豆株高的基因组选择提供了高效的预测模型, 也为大豆株高性状的分子设计育种提供了重要理论依据。

关键词: 大豆; 株高; 全基因组关联分析; 单倍型; 基因组选择

Genome-wide association analysis and prediction model construction for soybean plant height

Tang Kuan-Qiang[1], Li Gong-Yun[1], Song Mei-Yi[1], Zhao Xue[1], Chang Chun-Ling[1,2,3,4,*]

[1] College of Life Sciences, Agriculture and Forestry, Qiqihar University, Qiqihar 161006, Heilongjiang, China; [2] Heilongjiang Provincial Agricultural Microbial Preparations Industrialization Engineering and Technology Research Center, Qiqihar 161006, Heilongjiang, China; [3] Heilongjiang Provincial Agricultural Bio-Preparation Industrialization Collaborative Innovation Center, Qiqihar 161006, Heilongjiang, China; [4] Heilongjiang Provincial Concept Verification Center of High Concentration Organic Waste Liquid Biotransformation and Industrialization, Qiqihar 161006, Heilongjiang, China

Abstract: Plant height is a key agronomic trait that is closely associated with soybean yield potential and lodging resistance, and it is essential for achieving high and stable yields in soybean production. Here, we conducted a genome-wide association study (GWAS) and genomic selection (GS) analyses of plant height using a diverse panel of 17,148 soybean germplasm accessions with worldwide geographic representation. We identified 139 significant genomic loci associated with plant height, with the candidate gene *Dt1* showing the largest genetic effect. In the GS analysis, the BayesA and rrBLUP models outperformed the other models in prediction accuracy. Further comparisons indicated that prediction models based on the top 500–1000 SNPs provided the best balance between predictive performance and cost effectiveness. Collectively, the loci and models identified here support efficient genomic selection for soybean plant height and provide a theoretical basis for molecular design breeding to improve this trait.

Keywords: soybean; plant height; GWAS; haplotype; GS

大豆(*Glycine max* [L.] Merr.)作为全球重要的经济作物, 在人类膳食和动物饲料中占据关键地位[1]。据统计, 全球约 60%的油料种子和 69%的植物蛋白来源于大豆[2]。随着全球人口持续增长和膳食结构不断改善, 人们对大豆的需求日益增加, 对产量潜力提出了更高的要求[3]。株高是大豆理想株型育种

本研究由黑龙江省自然科学基金项目(YQ2024D013), 黑龙江省省属本科高校“优秀青年教师基础研究支持计划”项目(YQJH2023099)和黑龙江省省属高等学校基本科研业务费项目(145309211)资助。

This study was supported by the Heilongjiang Provincial Natural Science Foundation of China (YQ2024D013), the Excellent Young Teachers Program of Basic Research in Heilongjiang Province in China (YQJH2023099), and the Basic Scientific Research Fund for Universities in Heilongjiang Province of China (145309211).

* 通信作者(Corresponding author): 常春玲, E-mail: changchunling0709@163.com

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

Received (收稿日期): 2025-10-16; Accepted (接受日期): 2026-03-16; Published online (网络出版日期): 2026-03-20.

URL: https://link.cnki.net/urlid/11.1809.S.20260319.1807.002

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

的重要性状之一, 与产量、抗倒伏能力和环境适应能力密切相关[4]。培育理想矮化品系并配套密植栽培模式, 是当前提升我国大豆产量的有效途径[5-6]。因此, 鉴定控制大豆株高的相关基因, 对解析株高形成的遗传基础及大豆理想株型育种与产量提升具有重要意义。

大豆株高受到多基因协同调控, 其遗传机制解析一直是作物遗传育种领域的研究重点[5-6]。目前, 已通过连锁分析和关联分析等方法, 鉴定出大量调控株高性状的遗传位点[6-12]。例如, 大豆生长习性基因 *Dt1* 和 *Dt2* 通过调控节数形成直接影响株高。无限生长习性(基因型为 Dt1_/dt2dt2)和亚有限生长习性(基因型为 Dt1_/Dt2Dt2)的大豆材料节数较多, 株高较高; 而有限生长习性(基因型为 dt1dt1/__)的大豆材料的节数相对较少, 株高较矮[6,10,13]。此外, *RIN1* [14]、*GmILPA1* [15]、*GA2ox* [8]、*GmDW1* [16]、*miR166* [17]、*GmFULc* [18]和 *GmRAV* [19]通过调控内源赤霉素的含量进而影响大豆株高。*GmPH13* 编码 WD40 蛋白, 与 COP1 互作, 调控 STF1/2 的蛋白丰度, 最终通过调节大豆节间长度来调控株高[9]。大豆作为严格的光周期作物, 与花期相关的 *E1*、*E2*、*GmAP1a* 和 *GmMRF2* 等基因也被鉴定为与大豆株高相关[12,20-22]。目前, 通过 QTL 和 GWAS 等方法鉴定到的与株高相关的位点达 392 个(soybase 数据库) [23]。如何有效整合多基因信息以实现大豆株高的精准分子设计育种, 仍是当前研究面临的重要挑战。

随着生物信息的快速发展, 基因组选择育种技术通过构建基于基因组分子标记的预测模型, 实现了对个体早期精准预测, 显著缩短了育种周期、提高了选择效率, 成为研究作物多基因控制数量性状的重要技术手段[24]。基于基因组最佳线性无偏预测(genomic best linear unbiased prediction, GBLUP)、岭回归最佳线性无偏预测(ridge regression best linear unbiased prediction, rrBLUP)、贝叶斯方法(Bayes)、机器学习和深度学习等的全基因组选择模型已在作物中展现出良好的应用前景[25]。然而, 当前基因组选择分析仍需对大规模群体进行基因型测定, 造成较高的经济成本。基于全基因组关联分析, 筛选与目标性状高度关联的遗传位点, 构建精简后的分子标记集, 进行基因组选择分析[26-27], 显著降低了测序和计算资源需求, 提升了模型在特定性状上的预测效率。本研究以 17,148 份分布于世界各地的大豆种质资源的株高表型为训练群体, 利用多种全基因组关联分析方法, 鉴定控制株高的遗传位点, 解析主要关联位点不同单倍型对株高性状调控的差异。通过多种统计指标评估 rrBLUP、贝叶斯 A 方法(BayesA)、支持向量回归(support vector regression, SVR)及全连接神经网络(deep neural network, DNN) 4 种基因组选择模型的预测性能。基于表现最优的 BayesA 和 rrBLUP 模型, 利用与株高性状显著关联的前 98、500、1000 和 5000 个 SNP 标记, 对 914 份和 932 份独立的大豆品系进行了株高预测。本研究基于大规模种质资源的 GWAS-GS 联合研究, 有望为大豆株高遗传机制解析及预测模型优化提供新思路, 为培育理想株高大豆新品系提供关键基因资源和理论支撑。

1 材料与方法

1.1 表型及基因型数据集分析

以 17,148 份栽培大豆种质资源[28-29]以及分别于 2014 年播种于中国山西的 914 份[30]和 2013 年播种于中国北京的 932 份[30]栽培大豆为研究材料。其中 17,148 份栽培大豆种质资源的株高表型数据来自大豆种质资源库数据库[23], 基因型数据来自于 50K 芯片[28], 共包含 42,195 个 SNP 位点, 按最小等位基因频率大于 5%和缺失率小于 10%过滤后, 共获得 35,215 个 SNP 位点, 用于全基因组关联分析和全基因组选择分析。模型验证环节采用 2 份独立数据集, 即 2014 年播种于中国山西的 914 份栽培大豆和 2013 年播种于中国北京的 932 份栽培大豆, 其株高表型数据均来源于 SoyOmics 数据库, 其基因型数据来自国家生物信息中心[30-31]。原始测序数据用 Fastp 软件(v. 0.20.1) [32]进行质量控制与过滤, 用 BWA (v. 0.7.17)软件进行序列比对。用 GATK (v. 4.6.1.0) [33]工具包进行后续处理, Picard 模块对比对数据进行排序并去除 PCR 重复, HaplotypeCaller 模块获得各样品的变异信息(gvcf 文件), GenomicsDBImport 模块整合所有样品的 gvcf 文件, GenotypeGVCFs 模块鉴定所有样品的变异信息, VariantFiltration 模块筛选高质量变异位点。基于上述处理结果, 将重测序获得的 SNP 与 50K 芯片 SNP 位点进行坐标匹配, 构建用于 BayesA 和 rrBLUP 模型的预测集。

1.2 群体结构分析

采用 Admixture (v. 1.3.0)软件进行群体结构分析[34], *K* 值设定范围为 1~18。利用 FastTree (v. 2.1.11)软件的 GTR 模型构建系统进化树[35]。利用

Plink (v. 1.90)软件计算不同亚群连锁不平衡(linkage disequilibrium, LD)的半衰减距离[36]。

1.3　全基因组关联分析

全基因组关联分析选用 MLM、FarmCPU 和 Fast3VmrMLM 三种模型。MLM 和 FarmCPU 模型由 GAPIT [37]实现, 显著性阈值采用 Bonferroni 校正法确定($-\log_{10}(\alpha/N)$, $\alpha = 0.05$, N 为分子标记总数)。Fast3VmrMLM 模型利用 Fast3VmrMLM 包实现[27]。PLINK (v. 1.90)软件计算标记间的连锁不平衡值。合并连锁紧密(LD $r^2 > 0.2$)的显著位点为同一个区块, 以距离区块两侧最近的 2 个非显著 SNP 位点作为区块的边界, 若超过 100 kb, 则以距离区块 100 kb 处为边界[4], 结果仅展示区块内峰值处的 SNP 编号(表 1)。连锁区块内所有显著关联的 SNP 用于单倍型分析, 携带相同单倍型的个体归为一组, 用 R 语言进行单因素方差分析(ANOVA), 并对候选基因在 314 个组织(12 个器官) [38]中的时空表达模式进行分析。

1.4　预测模型构建与评估

选用 4 种基因组选择模型和全部 SNP 构建株高预测模型。rrBLUP 模型通过 rrBLUP 包实现[39], BayesA 模型通过 BGLR 包构建[40], SVR 模型通过 Scikit-learn 框架[41]实现。DNN 模型通过 Keras 框架[42]搭建, 网络结构包括 3 个隐藏层, 每层设 256 个神经元。模型评估采用 5 倍交叉验证方法, 即将供试群体随机分成 5 份, 每次选择 4 份构成训练群体, 剩余 1 份为验证群体, 在训练群体中利用表型数据和基因型数据建立基因组选择模型, 重复 5 次, 将得到的预测值与真实值进行评估。评估指标包括平均绝对百分比误差(mean absolute percentage error, MAPE)、对称平均绝对百分比误差(symmetric mean absolute percentage error, SMAPE)、平均绝对误差(mean absolute error, MAE)、均方误差(mean squared error, MSE)和皮尔逊相关系数(Pearson correlation coefficient, r)。

在 GWAS 分析结果中提取与株高表型显著关联的前 98、500、1000 和 5000 个 SNP 标记, 利用 BayesA 和 rrBLUP 方法分别构建不同 SNP 数量的训练集, 以 914 份和 932 份栽培大豆[30-31]对应的基因组变异信息作为预测集, 评估模型在不同数据集上的预测能力。

所有图表的可视化均通过 R 语言的 ggplot2 包实现[43]。

2　结果与分析

2.1　群体结构及表型分析

17,148 份大豆种质资源的株高表型变异范围为 50~250 cm, 平均值为 87 cm。Shapiro-Wilk 正态性检验结果表明, 该群体的株高表型符合正态分布($P = 0.12$, 图 1-a)。基于 35,215 个 SNP 标记进行群体遗传结构分析, 17,148 份大豆种质资源可分为 11 个亚群, 记为 pop01~pop11 (图 1-b)。同一亚群内的种质在发育树上高度聚集, 不同亚群的种质在分支上呈分散分布特征(图 1-c)。不同亚群间的株高表型存在分化(图 1-d, $P < 0.05$), 其中 pop05 (112.0±33.9 cm)、pop07 (105.00±33.4 cm)和 pop10 (96.5±25.1 cm)的平均株高较高, pop02 (88.9±17.9 cm)、pop11 (85.7±17.0 cm)和 pop04 (84.8±18.9 cm)次之, pop09 (73.3±23.7 cm)、pop08 (71.4±17.4 cm)、pop03 (71.0±24.5 cm)、pop01 (69.6±18.7 cm)和 pop06 (69.4±10.8 cm)的平均株高较矮(图 1-d)。连锁不平衡分析显示, 不同亚群间 LD 半衰减距离存在差异, 试验群体的连锁不平衡平均半衰减距离为 51.23 kb, 其中 pop09 亚群衰减速率最快, 其半衰减距离约为 35.46 kb, 而 pop10 亚群衰减速率最慢, 其半衰减距离达 134.59 kb (图 1-e)。

2.2　全基因组关联分析及候选位点筛选

利用 Fast3VmrMLM、FarmCPU 和 MLM 三种模型进行全基因组关联分析(图 2-a~c)。其中 Fast3VmrMLM 检测到 98 个显著关联的 SNP 位点(图 2-a), FarmCPU 检测到 154 个显著关联的 SNP 位点(图 2-b), MLM 仅检测到 27 个显著关联的 SNP 位点(图 2-c)。3 种模型共检测到 244 个与株高显著关联的 SNP 位点, 进一步合并连锁紧密的 SNP 位点, 共鉴定到 139 个与株高相关的区间, 其中 53 个区间存在已报道的与株高相关的 QTL 或功能基因(表 1)。

19 号染色体 43.64~48.28 Mb 区间在 3 种模型中均检测到与株高性状的强关联信号(图 2), 该区间包含 3 个已报道的 QTL 位点(*PH3-1* [44], *PH4-4* [45]和 *PH4-g1* [46]) (表 1), 该候选区间内的基因 *Glyma.19G194300* (*Determinate stem1*, *Dt1*), 在多篇研究中均被证实与株高发育有关[6,12,47-48], 该基因编码磷脂酰乙醇胺结合蛋白, 在大豆株高和生长习性调控中起关键作用。Dt1/_基因型通常表现为无限、亚有限生长习性, 株高较高; 而 dt1/dt1 基因型则表现为有限生长习性, 株高较矮[6,10,12,47,49]。该基因在大豆植株顶端生长点和根部表达量最高, 暗示其在大豆顶端生长点中具有调控叶芽向花芽转化的作用(图 3-a)。

该区间共包含 29 个显著关联的 SNP 位点, H1、H2、H3、H4、H5 和 H6 单倍型株高较高, 其中 H1 单倍型株高最高(117.76±35.88) cm, 显著高于其他单倍型($P < 0.05$), 而 H7、H8、H9、H10、H11 和 H12 单倍型株高较矮, 其中 H12 单倍型株高仅为(55.90±15.78) cm, 显著低于其他单倍型($P < 0.05$) (图 3-g)。

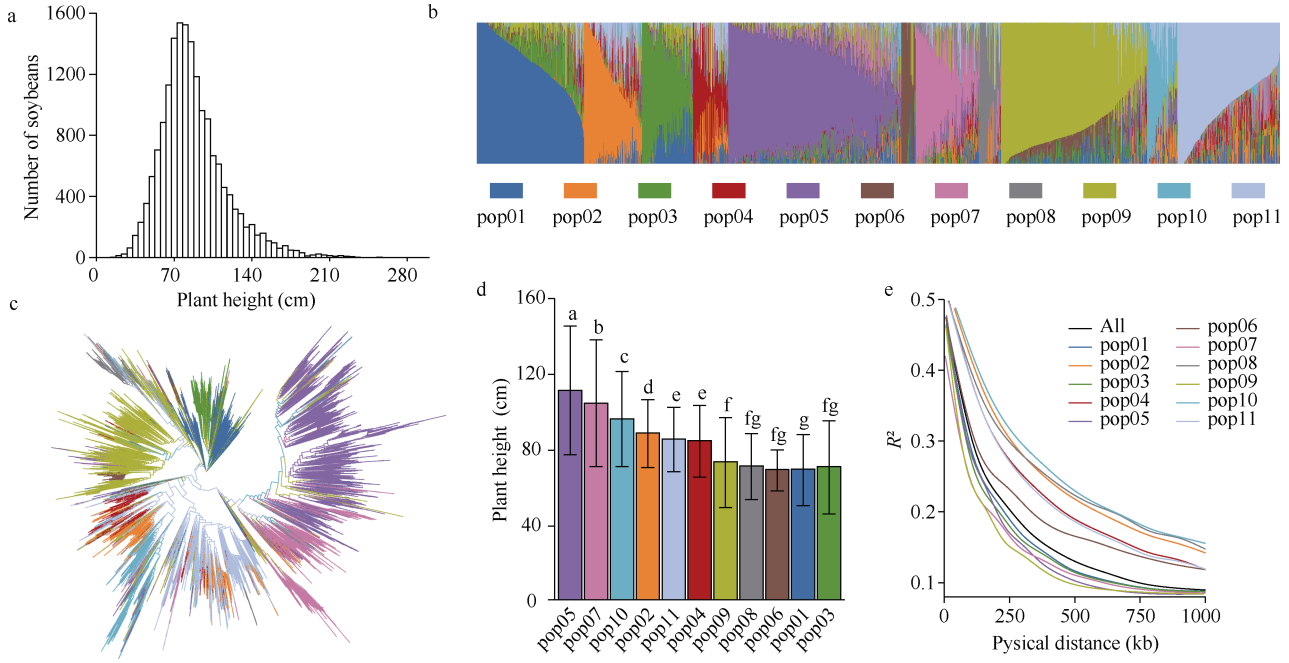


图 1 群体结构分析

Fig. 1 Population structure analysis

a: 株高分布; b: 群体结构 Admixture 分析($K = 11$); c: 系统进化树分析; d: 11 个亚群株高差异分析, 按亚群株高平均值从大到小排序, 不同小写字母表示差异显著性($P < 0.05$); e: 各亚群连锁不平衡分析。图 c–e 中颜色标注与图 b 保持一致, 相同颜色代表同一亚群。

a: distribution of plant height; b: population structure inferred by Admixture ($K = 11$); c: phylogenetic tree; d: differences in plant height among the eleven subpopulations, ranked by mean plant height; different lowercase letters indicate significant differences ($P < 0.05$); e: linkage disequilibrium decay in each subpopulation. Color coding in c–e matches that in b, with the same color indicating the same subpopulation.

10 号染色体 44.63~46.74 Mb 区间在 3 种模型中也均被检测到(图 2), 该区间包含 3 个已报道的与株高显著相关的 QTL 位点(*PH19-2* [50], *PH31-2* [51]和 *PH5-g1.2* [12]) (表 1), 研究表明, 该候选区间内的 *Glyma.10G221500* (*E2*)基因在大豆株高发育和生育期性状上起重要的调控作用[12,52-53], 编码 Gigantea 蛋白, 在茎的节间和节、生长点、花、叶和根等多个器官的多个时期中均表达, 暗示了该基因可能参与大豆生长发育中多种性状的调控(图 3-b)。该区间内包含 12 个与株高显著关联的 SNP 标记(表 1), 其中 H1 单倍型的株高(141.09±30.02 cm)最高, 显著高于其他单倍型($P < 0.05$), H2、H3、H4、H5 和 H6 单倍型次之(H2: (111.94±31.28) cm; H6: (95.06±18.27) cm), H7、H8、H9、H10、H11 和 H12 单倍型较矮(H7: (89.08±26.00) cm; H12: (70.84±21.42) cm), H12 单倍型最矮, 显著低于其他单倍型($P < 0.05$) (图 3-h)。

12 号染色体 4.93~5.90 Mb 区间在 3 种模型中也均被检测到(图 2), 该区间内的候选基因可能为 *Glyma.12G073900* (*Tof12*), 编码 pseudo-response regulator 7 蛋白, 在大豆株高及区域适应性调控中起重要作用[54-55]。*Tof12* 的时空表达趋势与 *E2* 基因类似, 在多个器官的多个时期均有表达, 可能参与大豆生长发育中多种性状的调控(图 3-c)。该区间内包含 6 个与株高显著关联的 SNP 标记(表 1), H1 (132.86±32.98 cm)、H2 (124.09±35.13 cm)、H3 (110.49±40.34 cm)和 H4 (113.33±32.03 cm)单倍型的株高较高, 其次为 H5 (101.46±35.88 cm)、H6 (97.66±35.03 cm)、H7 (88.57±21.28 cm)、H8 (88.63±32.81 cm)、H9 (83.92±26.08 cm)、H10 (79.91±16.13 cm)、H11 (83.65±31.41 cm)和 H12 (71.48±31.26 cm), H12 单倍型的株高显著低于其他单倍型($P < 0.05$) (图 3-i)。

13 号染色体 35.33~38.60 Mb 区间在 3 种模型中也均被检测到(图 2), 该区间包含 *PH6-g12* [56]位点, 区间内 *Glyma.13G287600* 和 *Glyma.13G288000* 编码赤霉素 β 氧化酶, 参与调控大豆株高的生长发育[8]。

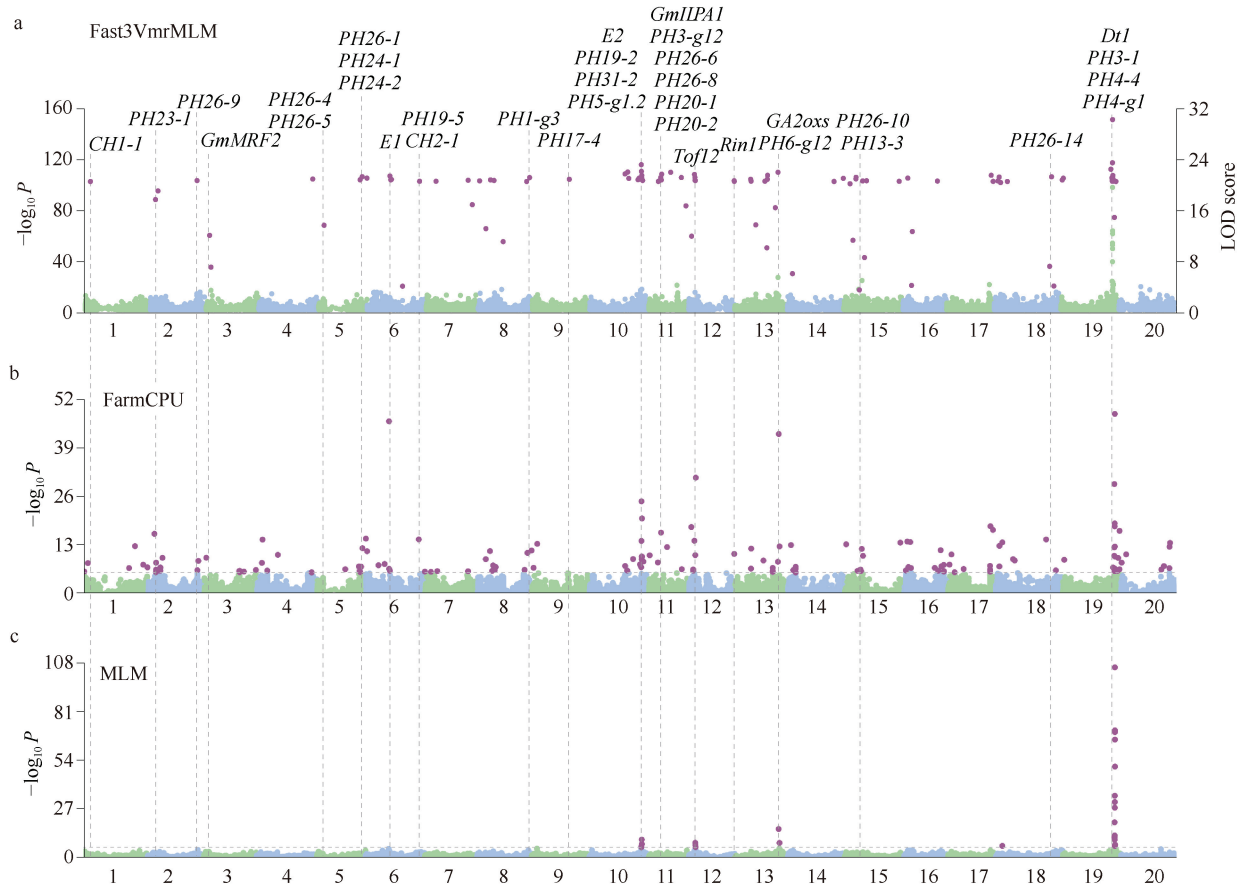


图 2　基于 Fast3VmrMLM (a)、FarmCPU (b)和 MLM (c)模型的大豆株高全基因组关联分析

Fig. 2　Genome-wide association analysis of soybean plant height using the Fast3VmrMLM (a), FarmCPU (b), and MLM (c) models

PH: 株高; CH: 冠层高度; Fast3VmrMLM: 快速三维多变量随机混合线性模型; FarmCPU: 固定和随机模型循环概率联合法; MLM: 混合线性模型。

PH: plant height; CH: canopy height; Fast3VmrMLM: fast three-vector multilocus random-SNP mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model.

表 1　株高性状相关的候选位点

Table 1　Loci associated with plant height

物理位置 Physical position (bp)	SNP 数目 SNP number	SNP 名称 Name of SNP	GWAS 模型 GWAS model	候选位点/基因 Candidate locus or gene
1: 4537369–4737369	1	ss715579566	Fast3VmrMLM	*CH1-1* [57]
2: 4462767–6836648	4	ss715582840	Fast3VmrMLM, FarmCPU	*PH23-1* [58]
2: 6712100–6912100	2	ss715583766	Fast3VmrMLM, FarmCPU	*PH23-1* [58]
2: 10255896–10553962	2	ss715580919	FarmCPU	*PH17-11* [59]
2: 12142977–12342977	1	ss715581055	FarmCPU	*PH33-3* [60]
2: 40658231–40858231	1	ss715582649	Fast3VmrMLM	*PH26-9* [61]
2: 42434287–43540684	2	ss715583006	FarmCPU	*PH26-9* [61]
3: 3011469–3211469	1	ss715585099	Fast3VmrMLM	*Glyma.03G106900* (*GmMRF2*) [22]
4: 4094550–4294550	1	ss715588091	FarmCPU	*PH33-4* [60]
4: 4685042–4885042	1	ss715588851	FarmCPU	*PH33-4* [60]
5: 4655898–4855898	1	ss715590521	Fast3VmrMLM	*PH26-4* [61], *PH26-5* [61]
5: 24093817–24293817	1	ss715590209	FarmCPU	*PH26-4* [61], *PH17-2* [59]
5: 35986874–38089147	5	ss715591542	Fast3VmrMLM, FarmCPU	*PH26-1* [61], *PH24-1* [62], *PH24-2* [62]
5: 39041972–39241972	1	ss715592108	FarmCPU	*PH26-1* [61], *PH24-1* [62], *PH24-2* [62]
5: 42019049–42219049	1	ss715591621	Fast3VmrMLM, FarmCPU	*PH26-1* [61], *PH24-1* [62], *PH24-2* [62]

(续表 1)

物理位置 Physical position (bp)	SNP 数目 SNP number	SNP 名称 Name of SNP	GWAS 模型 GWAS model	候选位点/基因 Candidate locus or gene
6: 19813355–20733420	3	ss715593833	Fast3VmrMLM, FarmCPU	*Glyma.06G207800* (*E1*) [20]
6: 50946599–51146599	1	ss715595278	FarmCPU	*PH27-1* [63]
7: 4818294–5018294	1	ss715598270	FarmCPU	*PH18-6* [50], *PH13-9* [64], *PH19-5* [50], *PH8-g2* [65]
7: 8576256–8776256	1	ss715598865	Fast3VmrMLM	*PH19-5* [50], *CH2-1* [66]
7: 10325142–10525142	1	ss715595759	FarmCPU	*CH2-1* [66]
8: 13345038–13545038	1	ss715599459	FarmCPU	*Glyma.08G163900* (*GmDW1*) [16]
8: 42962862–43162862	1	ss715602049	Fast3VmrMLM	*PH1-g3* [56]
9: 32325584–32525584	1	ss715603525	Fast3VmrMLM	*PH17-4* [59]
10: 44634052–46739345	12	ss715607448	Fast3VmrMLM, FarmCPU, MLM	*Glyma.10G221500* (*E2*) [12,53], *PH19-2* [50], *PH31-2* [51], *PH5-g1.2* [12]
11: 7571209–7771209	1	ss715611131	FarmCPU	*PH24-5* [62], *PH24-6* [62], *PH26-6* [61], *PH26-7* [61]
11: 8164638–8364638	1	ss715611223	Fast3VmrMLM	*PH26-6* [61], *PH26-8* [61], *PH20-1* [67]
11: 10250509–10491700	2	ss715608747	Fast3VmrMLM, FarmCPU	*PH26-8* [61], *PH20-1* [67]
11: 10829924–11029924	1	ss715608798	Fast3VmrMLM	*PH26-8* [61], *PH20-1* [67], *PH20-2* [67]
11: 11092462–11292462	1	ss715609445	Fast3VmrMLM	*PH26-8* [61], *PH20-1* [67], *PH20-2* [67]
11: 15578644–15778644	1	ss715609882	FarmCPU	*PH26-8* [61]
11: 18962168–19162168	1	ss715609800	Fast3VmrMLM	*PH3-g12* [68], *Glyma.11G026400* (*GmILPA1*) [15]
12: 4929086–5897709	6	ss715613141	Fast3VmrMLM, FarmCPU, MLM	*Glyma.12G073900* (*Tof12*) [54]
12: 39054817–39848127	3	ss715612990	Fast3VmrMLM, FarmCPU	*Glyma.12G224600* (*Rin1*) [14]
13: 24323315–24523315	1	ss715614116	FarmCPU	*CH1-2* [57]
13: 32362674–32574603	4	ss715615299	FarmCPU	*PH5-6* [69]
13: 35325581–38601835	5	ss715615977	Fast3VmrMLM, FarmCPU, MLM	*Glyma.13G287600/Glyma.13G288000* (*GA2oxs*) [8], *PH6-g12* [56]
14: 2423374–2623374	1	ss715618109	FarmCPU	*PH33-6* [60]
15: 10654448–11332649	3	ss715620250	Fast3VmrMLM, FarmCPU	*PH26-10* [52]
15: 13716717–14915803	4	ss715620652	Fast3VmrMLM, FarmCPU	*PH26-10* [61], *PH13-3* [64]
16: 1250020–1450020	1	ss715623432	FarmCPU	*PH13-5* [64]
16: 2502942–2702942	1	ss715623842	FarmCPU	*PH2-2* [70], *PH6-6* [71], *PH3-4* [44]
16: 3487595–3687595	1	ss715624811	FarmCPU	*PH5-9* [69]
16: 5699540–5899540	1	ss715625240	FarmCPU	*PH1-g17* [56]
16: 31450146–31650146	1	ss715624412	FarmCPU	*PH26-16* [61], *PH25-2* [72], *PH5-13* [69], PH7-g3 [73]
17: 36483136–37013821	5	ss715627165	FarmCPU	*PH33-7* [60]
18: 48304610–48504610	1	ss715631373	Fast3VmrMLM	*PH26-14* [61]
18: 49914203–50114203	1	ss715631551	Fast3VmrMLM	*PH26-14* [61]
18: 51949642–52149642	1	ss715631757	Fast3VmrMLM, FarmCPU	*PH26-14* [61]
19: 1189196–1389196	1	ss715633103	FarmCPU	*PH3-5* [44], *PH6-7* [71]
19: 43640270–48280446	29	ss715635425	Fast3VmrMLM, FarmCPU, MLM	*Glyma.19G194300* (*Dt1*) [12,47-48], *PH3-1* [44], *PH4-4* [45], *PH4-g1* [46]
19: 49180763–49434685	2	ss715635917	FarmCPU	*PH4-4* [45], *PH9-1* [74]
20: 276574–476574	1	ss715637977	FarmCPU	*PH1-g26* [56]
20: 36946555–37146555	1	ss715637802	FarmCPU	*PH1-g27* [56]

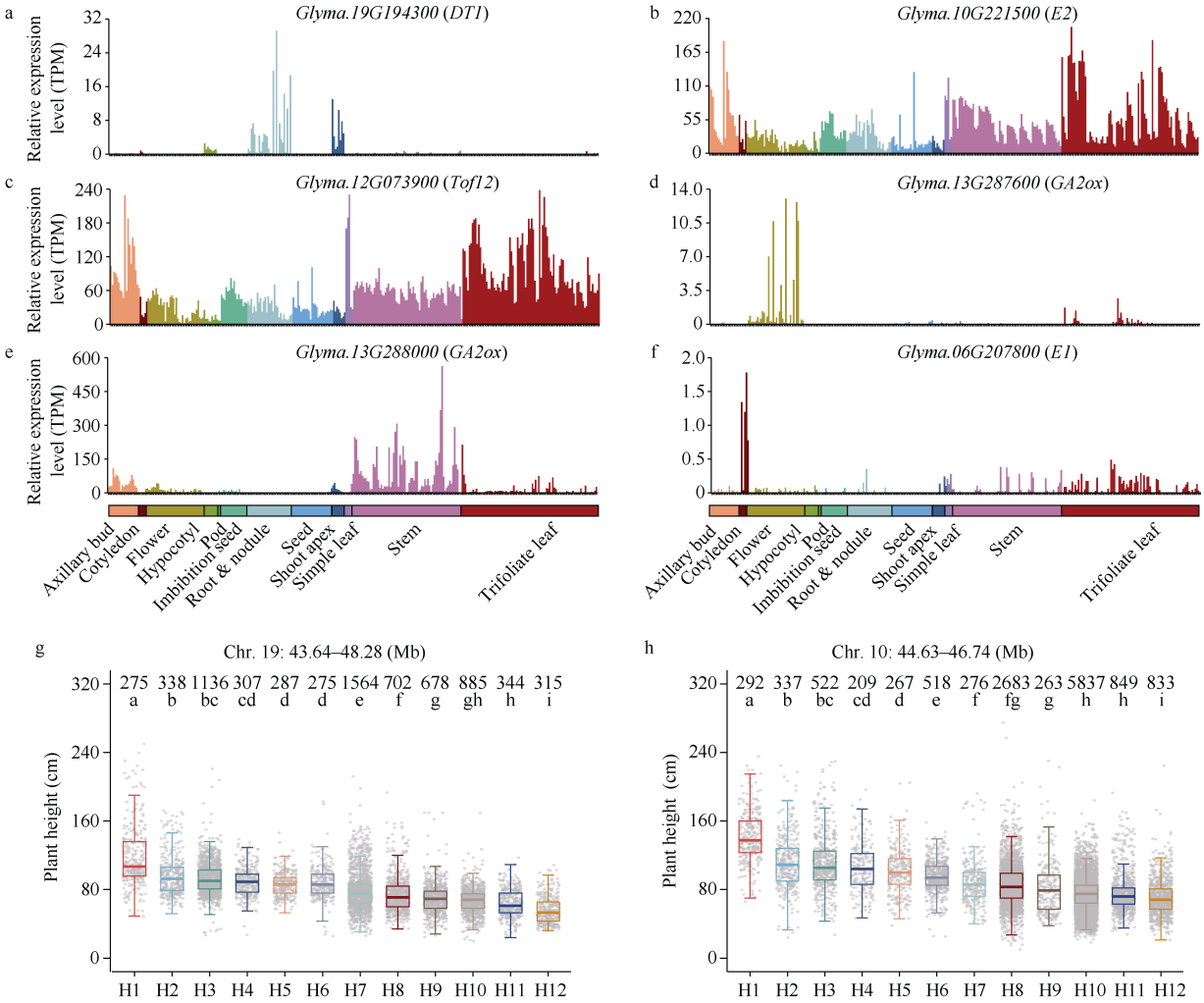
缩写同图 2。Abbreviations are the same as those given in Fig. 2.

其中 *Glyma.13G287600* 在花发育的各个时期表达量较高(图 3-d), 而 *Glyma.13G288000* 在茎发育的各个时期表达量较高(图 3-e)。该区间共有 5 个与株高显著关联的位点(表 1), 其中 H1 (116.61±39.44 cm)单倍型株高最高, 显著高于其他单倍型($P < 0.05$), 其次为 H2 (87.86±27.69 cm)、H3 (91.31±38.35 cm)、H4 (85.07±28.77 cm)和 H5 (75.81±25.63 cm), H5 单倍型显著低于其他单倍型($P < 0.05$) (图 3-j)。

6 号染色体 19.81~20.73 Mb 区间在 FarmCPU 和 Fast3VmrMLM 模型中被检测到(图 2), 该区间内的候选基因可能为 *Glyma.06G207800* (*E1*), 编码 AP2/B3-like 转录因子, 调控大豆株高和生育期。E1/_基因型通常开花期推迟, 株高较高, 而 e1/e1 基因型则表现为早花, 株高较矮[20]。*Glyma.06G207800* 在子叶中表达量较高(图 3-f), 可能在大豆生长发育的早期参与株高和花期的调控。该区间共有 3 个与株高显著关联的位点(表 1), 其 H1 (90.28±21.19 cm)、H2 (87.90±21.33 cm)、H3 (86.87±21.5 cm)、H4 (88.8±15.06 cm)和 H5 (87.61±31.82 cm)单倍型的株高较高, 而 H6 (76.95±32.43 cm)单倍型显著低于其他单倍型($P < 0.05$) (图 3-k)。

2.3　预测模型构建与评估

基于 17,148 份大豆株高表型数据和 35,215 个 SNP 标记, 分别构建了 rrBLUP、BayesA、SVR 和 DNN 四种基因组预测模型。模型性能评估结果显示, BayesA 模型表现最优, 其预测株高与观测株高的相关系数达 0.82 (图 4-a), 平均绝对百分比误差(MAPE)



H1: AAATCTACTCGTTACAATCGTCTGTTCCG
H2: AAATCTACTCGTTACAATCGTCTGTCCCG
H3: AAACATACTTACCAACATCACTCATTATA
H4: AGGTCTACCCGTTACAATCGTCTGTTATA
H5: AGGTCTACCCGTTACAATCGTCTGTTACG
H6: AAACATACTTACCAACATCACTCATTACG
H7: AGGTAGGCCCGCTCACGCTGTCTGTTCCG
H8: AGGTAGGCCCGCTCACGCTGTCTGTCCCG
H9: AGGTAGGCCCGCTCACGCTGTCTGTTACG
H10: AGGTAGGCCCGCTCACGCTGTCTGTCACG
H11: AGGTAGGCCCGCTCACGCTACCCGCCACG
H12: GGGTAGGCCCGCTCACGCTGTCTGTCACG

H1: GGAGGATAGTAA
H2: AAGTAGCAGTAA
H3: GGAGGATGGTAA
H4: AGAGGATGGTAA
H5: AGGTAATAGTAA
H6: AGGTAGCAGCAA
H7: GGGTAATAGCAA
H8: AGGTAATAGCAA
H9: AGGTAATAACAG
H10: AGGTAATAGCAG
H11: GGGTAATGGCAG
H12: GGGTAATAGCAG

(图 3)

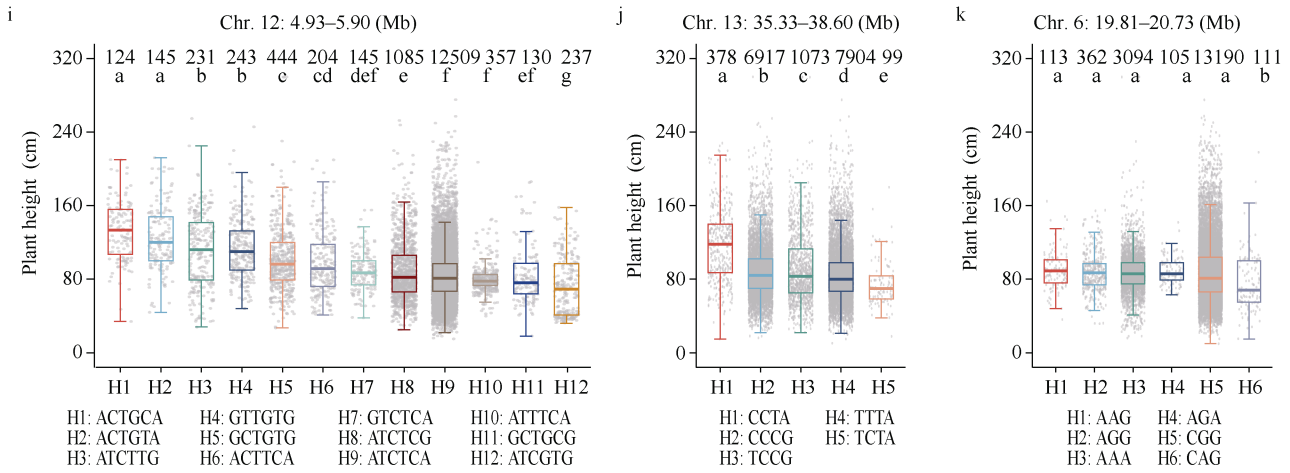


图 3 基因表达及单倍型分析

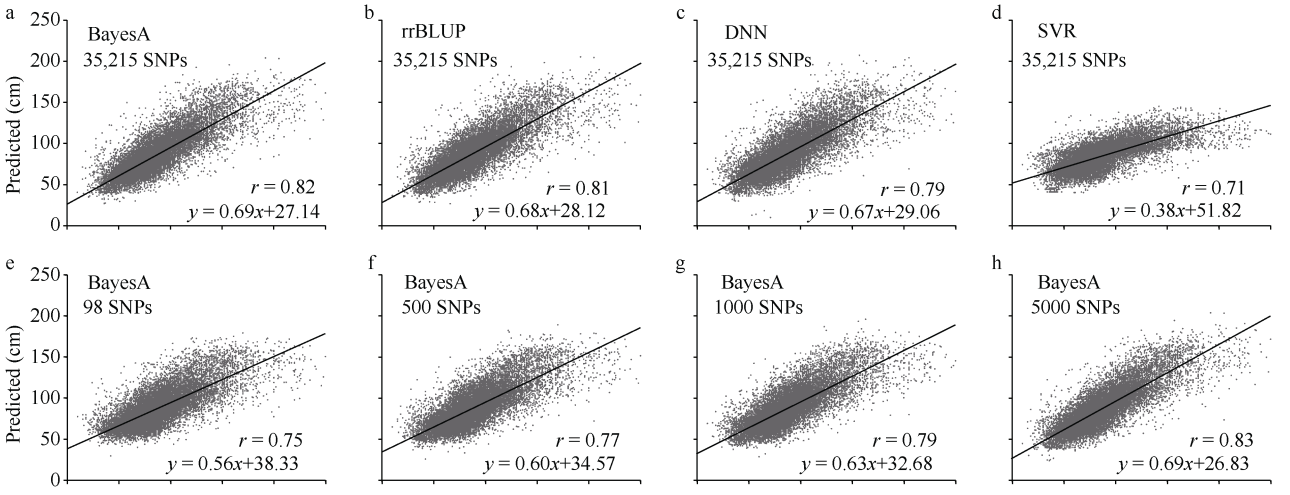
Fig. 3 Gene expression and haplotype analysis

a–f: *Glyma.19G194300*、*Glyma.10G221500*、*Glyma.12G073900*、*Glyma.13G287600*、*Glyma.13G288000* 和 *Glyma.06G207800* 在不同器官不同时期的表达情况。g–k: 19 号染色体 43.64–48.28 Mb、10 号染色体 44.63–46.74 Mb、12 号染色体 4.93–5.90 Mb、13 号染色体 35.33–38.60 Mb 和 6 号染色体 19.81–20.73 Mb 区间内不同单倍型的株高表型分析。各单倍型序列以窄字体标注于对应图下方。TPM: 每百万读取次数中某个转录本的占比。

a–f: expression patterns of *Glyma.19G194300*, *Glyma.10G221500*, *Glyma.12G073900*, *Glyma.13G287600*, *Glyma.13G288000* and *Glyma.06G207800* across organs and developmental stages. g–k: phenotypic analysis of plant height for different haplotypes in the genomic regions of 43.64–48.28 Mb on chromosome 19, 44.63–46.74 Mb on chromosome 10, 4.93–5.90 Mb on chromosome 12, 35.33–38.60 Mb on chromosome 13 and 19.81–20.73 Mb on chromosome 6. Haplotype sequences are shown below the corresponding panels in a narrow font. TPM: transcripts per million.

为(15.08±0.27)%、对称平均绝对百分比误差(SMAPE)为(14.22±0.17)%、均方误差(MSE)为(294.30±11.85), 均优于其他模型。rrBLUP 模型表现次之, 而 DNN 和 SVR 模型则表现较差(表 2 和图 4-a~d)。

在育种实践中, 针对大规模群体实施高通量标记测序面临较高的经济成本。鉴于此, 本研究从 Fast3VmrMLM 关联分析结果中, 按关联信号强度选取与株高紧密关联的前 98、500、1000 和 5000 个 SNP 标记, 采用表现较好的 BayesA 和 rrBLUP 两种模型进行基因组预测建模。结果显示, 前 98 个 SNP 构建的 BayesA 和 rrBLUP 模型, 其相关系数均达到 0.75; 前 500 个 SNP 构建的模型, 其相关系数均为 0.77; 前 1000 个 SNP 构建的模型, 其相关系数均为 0.79; 前 5000 个 SNP 构建的模型, 其相关系数分别为 0.83 和 0.82 (图 4-e~l)。然而, 选用全部标记(35,215 个 SNP)构建的 BayesA 和 rrBLUP 模型, 其相关系数分别为 0.82 和 0.81 (图 4-a~b)。上述结果表明, 随着标记数目的增加, 预测株高与观测株高的相关性系数呈现上升趋势, 但标记数超过 5000 时, 预测效果并未显著提升(图 4-e~l)。



(图 4)

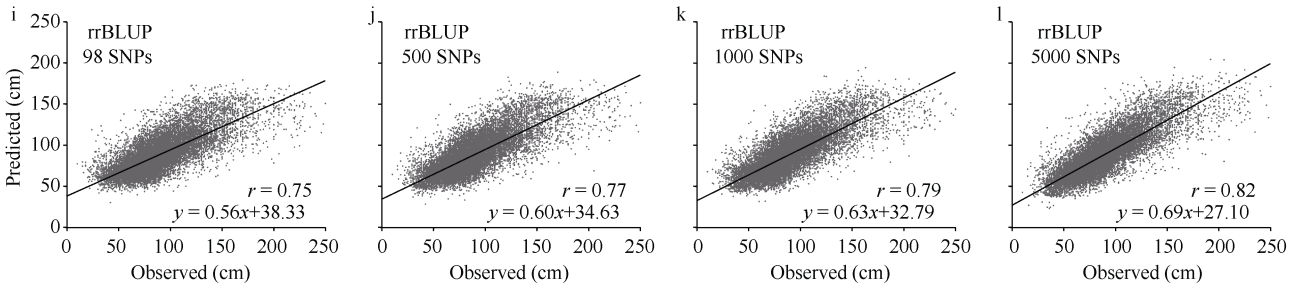


图 4　不同模型的株高表型预测表现

Fig. 4　Prediction performance of plant height using different models

a–d: BayesA 模型、rrBLUP 模型、DNN 模型和 SVR 模型预测株高与实测株高的比较; e–h: 前 98、500、1000 和 5000 个与株高紧密关联的 SNP 位点构建的 BayesA 模型, 其预测株高与实测株高的比较; i–l: 前 98、500、1000 和 5000 个与株高紧密关联的 SNP 位点构建的 rrBLUP 模型, 其预测株高与实测株高的比较。rrBLUP: 岭回归最佳线性无偏预测; BayesA: 贝叶斯 A; SVR: 支持向量回归; DNN: 全连接神经网络。

a–d: comparisons between predicted and observed plant height using BayesA, rrBLUP, DNN, and SVR; e–h: comparisons between predicted and observed plant height using BayesA models built with the top 98, 500, 1000, and 5000 SNPs associated with plant height; i–l: comparisons between predicted and observed plant height using rrBLUP models built with the top 98, 500, 1000, and 5000 SNPs associated with plant height. rrBLUP: ridge regression best linear unbiased prediction; SVR: support vector regression; DNN: deep neural network.

表 2　不同预测模型的性能比较

Table 2　Performance comparison of prediction models

指标 Metric	贝叶斯 A BayesA	岭回归最佳线性无偏预测 rrBLUP	深度神经网络 DNN	支持向量回归 SVR
平均绝对百分比误差 MAPE (%)	15.08±0.27	15.37±0.25	16.29±2.35	18.53±2.97
对称平均绝对百分比误差 SMAPE (%)	14.22±0.17	14.49±0.16	15.25±1.77	17.29±2.63
平均绝对误差 MAE	12.18±0.15	12.41±0.12	13.01±2.33	15.07±3.51
均方误差 MSE	294.30±11.85	305.69±11.50	333.41±134.86	476.21±258.61
皮尔逊相关系数 *r*	0.82±0.01	0.81±0.01	0.79±0.04	0.71±0.07

数据表示平均值±标准差。缩写同图 4。

Data are presented as means ± SD. Abbreviations are the same as those given in Fig. 4. MAPE: mean absolute percentage error; SMAPE: symmetric mean absolute percentage error; MAE: mean absolute error; MSE: mean squared error; *r*: Pearson correlation coefficient.

为进一步验证所构建预测模型的跨群体适用性, 本研究以 2014 年播种于中国山西的 914 份栽培大豆株高表型数据和 2013 年播种于中国北京的 932 份栽培大豆株高表型数据[30-31]为预测数据集。对播种于山西的 914 份栽培大豆株高表型进行预测时, 使用前 98 个 SNP 构建的 2 种模型, 其预测值与实测值的相关系数均为 0.52 (图 5-a, f); 使用前 500 个 SNP 时, 相关系数提升至 0.58 (图 5-b, g); 使用前 1000 个 SNP 时, 相关系数仅为 0.59 (图 5-c, h); 使用前 5000 个 SNP 时, 相关系数维持在 0.58 (图 5-d, i); 而使用全部 35,215 个 SNP 时, 相关系数仅为 0.61 和 0.60 (图 5-e, j)。对播种于北京的 932 份栽培大豆株高表型进行预测时, 使用前 98 个 SNP 构建的 2 种模型, 其预测值与实测值的相关系数均为 0.56 (图 5-k, p); 使用前 500 个 SNP 时, 相关系数提升至 0.63 和 0.64 (图 5-l, q); 使用前 1000 个 SNP 时, 相关系数仅为 0.66 (图 5-m, r); 使用前 5000 个 SNP 时, 相关系数跌落至 0.64 和 0.63 (图 5-n, s); 而使用全部 35,215 个 SNP 时, 相关系数仅为 0.68 和 0.66 (图 5-o, t)。上述数据表明, 随着标记数目的增加, 模型预测性能逐步提升, 但当标记数量超过 500 个后, 性能提升幅度显著减弱。因此, 在实际育种应用中, 选用 500~1000 个 SNP 标记已能满足需求。

3　讨论

大豆株高是产量性状和理想株型育种的重要目标之一, 由多基因调控, 其表型易受环境因素影响[75]。大豆株高遗传机制的解析, 对大豆产量提升及我国粮食安全均具有重要意义。本研究以全球范围内收集的大豆种质资源为研究材料, 通过全基因组关联分析共鉴定到 244 个与大豆株高显著关联的遗传位点。其中, 在 19 号染色体上鉴定到的与株高显著关联的候选基因为 *Dt1*, 与拟南芥 *TFL1* 基因同源, 在豆科植物中具有高度保守性, 其核心功能是调控大豆茎顶端分生组织发育[10,47,76-77]。其显性(Dt1/_)基因型通常使大豆表现为无限或亚有限生长习性, 株高

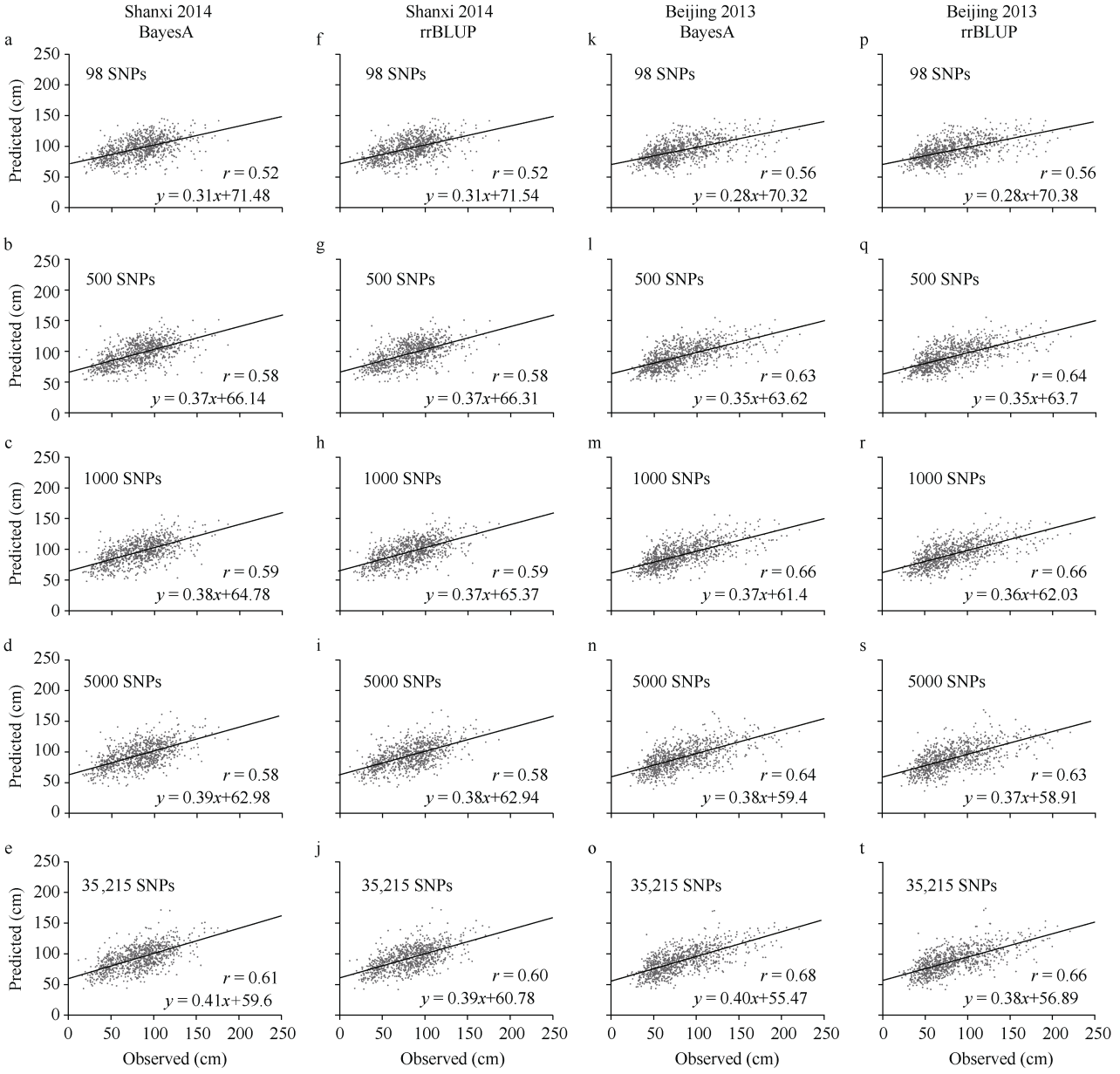


图 5 基于不同数量 SNP 标记的株高表型预测

Fig. 5 Plant height prediction using different numbers of SNP markers

a–j: 基于 GWAS 筛选的前 98, 500, 1000 和 5000 个显著关联 SNP 位点及全基因组 SNP 位点，分别构建 BayesA 模型(a–e)和 rrBLUP 模型(f–j)，对 2014 年种植于山西的 914 份大豆材料株高进行预测，并与实测株高结果比较；k–t: 基于 GWAS 筛选的前 98、500、1000 和 5000 个显著关联 SNP 位点及全基因组 SNP 位点，分别构建 BayesA 模型(k–o)和 rrBLUP 模型(p–t)，对 2013 年种植于北京的 932 份大豆材料株高进行预测，并与实测株高结果比较。

a–j: BayesA (a–e) and rrBLUP (f–j) models built using the top 98, 500, 1000, or 5000 GWAS-significant SNPs, or all genome-wide SNPs, to predict plant height for 914 soybean accessions grown in Shanxi in 2014, predicted values were compared with field-measured plant height. k–t: BayesA (k–o) and rrBLUP (p–t) models built using the same SNP sets to predict plant height for 932 soybean accessions grown in Beijing in 2013, predicted values were compared with field-measured plant height.

较高，隐性纯和基因型(dt1/dt1)则表现为有限生长习性，株高显著降低[6,10,12,47,49]。尽管 *Dt1* 基因是调控大豆株高的关键位点，但存在明显的遗传多效性，其隐性基因型在降低株高的同时会减少主茎节数，而节数直接决定果荚数量；与之相反，高秆品种虽主茎节数较多，却易发生倒伏，且不适宜密植栽培[6]。因此，仅依赖 *Dt1* 基因进行株高改良，难以实现产量提升与抗倒伏性增强的协同平衡，需采用多基因协同调控策略，才能实现大豆株高的精准改良与优良农艺性状的综合整合。

大豆株高性状的形成与光周期调控存在密切关联[11]。*E1* 和 *E2* 基因作为大豆光周期和纬度适应性调控的核心基因[9,20,78-79], 其多效性还包括通过复杂的遗传网络显著影响植株形态结构。已有研究表明, *E1* 基因的敲除会导致大豆分枝数减少、结荚习性改变[80]。*E2* 基因则与主茎节数调控相关, 其表达模式受光周期条件影响, 在弱光下短日照表达量更高, 强光下长日照表达量更高[48]。综上, *E1* 和 *E2* 基因通过光周期响应与植株形态建成的交叉调控网络, 共同决定大豆的株型特征与产量潜力[12,81], 这也为后续通过调控光周期相关基因改良大豆株高提供了重要的基因资源与理论依据。

赤霉素在作物株高调控中扮演着关键角色。上世纪 60 至 70 年代“绿色革命”, 小麦中的矮秆基因(*Reduced height*, *Rht*)与水稻中的半矮秆基因(*semi-dwarf1*, *sd1*), 正是通过调控作物的赤霉素信号通路来控制株高, 结合矮化密植栽培技术, 使当时作物产量提升了 20%~30%, 被称为“绿色革命基因” [5,6,82-84]。相比之下, 大豆的“绿色革命”进程相对滞后。近年来研究表明, 赤霉素在大豆株高调控中同样起重要作用, 如 *RIN1* [14]、*GmILPA1* [15]、*GA2ox* [8]、*GmDW1* [16]、*miR166* [17]、*GmFULc* [18]和 *GmRAV* [19]等基因均通过调控赤霉素通路来实现对大豆株高的调控。因此, 整合参与株高调控的多通路基因资源, 挖掘赤霉素信号通路及其他相关通路中的关键调控基因, 是选育理想株高大豆新品系的重要途径。

不同全基因组关联分析模型的算法特性, 对关联位点的鉴定效率与准确性影响较大。MLM 模型作为 GWAS 分析的经典算法, 通过引入群体结构和亲缘关系作为随机效应, 可显著降低假阳性位点的检出率, 但由于其矫正过于严格, 导致本研究中仅检测到 27 个与株高关联的 SNP 位点, 假阴性率较高, 难以有效挖掘调控株高的微效位点。相比之下, FarmCPU 和 Fast3VmrMLM 模型有效克服了 MLM 模型过于保守的缺陷, 分别检测到 154 个和 98 个关联 SNP 位点, 在挖掘微效关联位点方面表现更具优势[27,85]。因此, 在大豆株高这类复杂数量性状的遗传解析中, 采用多模型相互验证的策略, 可有效提高关键关联位点的鉴定准确性, 为后续候选基因的筛选与功能验证奠定坚实基础。

在基因组选择模型中, BayesA 模型在 4 种算法中预测能力最好, 相关系数达 0.82。尽管深度学习框架在多种作物性状预测中表现较好, 如 SoyDNGP 模型在大豆粒重、株高等关键性状的预测中相关系数均达到了 0.70 以上, 其中粒重预测更是高达 0.84, 在玉米、小麦和水稻等作物中也表现出良好的鲁棒性和泛化能力[86]。然而本研究中采用的深度学习模型的预测性能稍差于 BayesA 和 rrBLUP 模型, 这可能与深度学习模型的参数设置有一定关系。尽管深度学习模型在捕捉复杂非线性关系方面具有潜力, 但贝叶斯方法在中等规模数据集上往往展现出更稳健的性能[87], 能够灵活估计标记效应分布, 特别适用于存在主效 QTL 的数量性状预测[24], 这也与本研究中 BayesA 模型的最优表现相一致。

本研究利用 GWAS 筛选获得的显著性 SNP 位点作为预测因子, 构建了针对目标表型的基因组预测模型。通过 GWAS 先验信息对标记进行降维处理, 显著提升了计算效率, 降低了对高密度芯片的依赖。在不同标记梯度的预测精度评估中发现, 在训练集(图 4)和独立预测集(图 5)中, 预测能力均未随标记数量的增加而呈线性持续增长, 而是表现出明显的边际递减效应。当 SNP 标记数量为 500~1000 个时, 模型对表型的预测准确率趋于平稳, 且能够捕捉到控制目标性状的主要遗传效应(图 4 和图 5)。相较于全基因组水平的数万个标记, 精简后的分子标记集在维持高预测精度的同时, 降低了数据处理的计算开销, 在预测效力与检测成本之间达到了平衡点, 这对开发低密度育种芯片、实现大规模低成本的早期表型预测提供了重要的理论依据, 也证明了基于显著位点驱动的精准预测模型在分子设计育种中的高效性与经济可行性。

在实际生产应用中, 有效整合外部表型与基因型数据可显著提升本地性状的预测精度[25]。本研究基于美国农业部公布的大豆株高表型及基因型数据构建的预测模型, 对 2014 年山西试点和 2013 年北京试点的大豆株高性状均表现出良好的预测效果(图 5)。对 2015 年河南试点的大豆株高表型进行预测时, 选取前 98、500、1000、5000 个 SNP 标记及全基因组 SNP 标记分析显示, BayesA 模型的预测值与实测值相关系数依次为 0.37、0.40、0.47、0.44 和 0.48, rrBLUP 模型依次为 0.36、0.40、0.46、0.43 和 0.44, 该试点的预测效果略低于山西与北京试点(数据未展示)。以上 3 个试点的分析结果均表明, 利用外部数据构建预测模型开展本地表型预测具备一定的可行性。在实际育种工作中, 可先借助外部数据对本地育种材料的目标表型进行预测, 以此加快育

种进程，降低田间检测与表型鉴定成本。经田间鉴定获得的基因型与表型实测数据，还可重新纳入模型训练集，持续优化模型，强化本地育种群体的性状预测能力。此外，研究表明，将基因共表达网络信息[88]或环境变量[89]纳入预测模型，均能有效提升基因组预测的准确性。整合多组学数据和多环境变量是未来提高模型预测能力的重要方向。

本研究鉴定的显著关联位点(如前 98、500 和 1000 个显著关联的 SNP)是株高性状分子标记辅助选择的重要靶点，为大豆株高的分子设计育种奠定了重要基础。可利用基因编辑等技术对这些候选基因进行精准功能验证，明确其调控株高、节数及抗倒伏性的分子机理。随着基因组选择算法的优化和基因型鉴定成本的降低，未来研究可整合多环境因子和多组学数据，构建更精准的株高预测模型，加速大豆株高性状的遗传改良，为适应不同生态区(如高纬度短日照地区)的品种设计提供理论支撑，实现大豆株高性状和产量潜力的协同提升。

4 结论

本研究共鉴定出 244 个调控大豆株高的关键遗传位点，为解析株高性状的遗传机制提供了重要靶点。在全基因组选择模型中，BayesA 和 rrBLUP 模型表现出较好的预测效果，前 500~1000 个与株高显著关联的 SNP 位点在表型预测中具有较高的性价比，可作为分子标记辅助选择或株高分子设计育种的优先候选位点。

References

[1] Du H P, Fang C, Li Y R, et al. Understandings and future challenges in soybean functional genomics and molecular breeding. J Integr Plant Biol, 2023, 65: 468–495.

[2] SoyStat. 2025, http://www.soystats.com.

[3] Zhu Z, Wang Y L, Liu S L, et al. Genomic atlas of 8105 accessions reveals stepwise domestication, global dissemination, and improvement trajectories in soybean. Cell, 2025, 188: 6519–6535.e15.

[4] Assefa T, Otyama P I, Brown A V, et al. Genome-wide associations and epistatic interactions for internode number, plant height, seed weight and seed yield in soybean. BMC Genomics, 2019, 20: 527.

[5] Fang C, Dong L D, Zhou J C, et al. Toward a Green Revolution in soybean: the role of ultra-high-density planting. J Integr Plant Biol, 2026, 68: 297–301.

[6] Liu S L, Zhang M, Feng F, et al. Toward a “Green Revolution” for soybean. Mol Plant, 2020, 13: 688–697.

[7] Lipka A E, Tian F, Wang Q S, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics, 2012, 28: 2397–2399.

[8] Wang X, Li M W, Wong F L, et al. Increased copy number of gibberellin 2-oxidase 8 genes reduced trailing growth and shoot length during soybean domestication. Plant J, 2021, 107: 1739–1755.

[9] Qin C, Li Y H, Li D L, et al. *PH13* improves soybean shade traits and enhances yield for high-density planting at high latitudes. Nat Commun, 2023, 14: 6813.

[10] Ping J Q, Liu Y F, Sun L J, et al. *Dt2* is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell, 2014, 26: 2831–2842.

[11] Zhou Z K, Jiang Y, Wang Z, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol, 2015, 33: 408–414.

[12] Fang C, Ma Y M, Wu S W, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol, 2017, 18: 161.

[13] Kou K, Su T, Wang Y P, et al. Natural variation of the *Dt2* promoter controls plant height and node number in semi-determinant soybean. Mol Breed, 2021, 41: 40.

[14] Li S C, Sun Z H, Sang Q, et al. Soybean reduced internode 1 determines internode length and improves grain yield at dense planting. Nat Commun, 2023, 14: 7939.

[15] Sun J Q, Huang S Y, Lu Q, et al. UV-B irradiation-activated E3 ligase GmILPA1 modulates gibberellin catabolism to increase plant height in soybean. Nat Commun, 2023, 14: 6262.

[16] Li Z F, Guo Y, Ou L, et al. Identification of the dwarf gene *GmDW1* in soybean (*Glycine max* L.) by combining mapping-by-sequencing and linkage analysis. Theor Appl Genet, 2018, 131: 1001–1016.

[17] Zhao C, Ma J J, Zhang Y H, et al. The *miR166* mediated regulatory module controls plant height by regulating gibberellic acid biosynthesis and catabolism in soybean. J Integr Plant Biol, 2022, 64: 995–1006.

[18] Sun J Z, Zhang X M, Feng J H, et al. The transcription factor *GmFULc* regulates soybean plant height by binding the promoter of a gibberellin-responsive gene. Plant Physiol, 2025, 197: kiaf021.

[19] Xue Y G, Zhang Y T, Shan J M, et al. Growth repressor GmRAV binds to the *GmGA3ox* promoter to negatively regulate plant height development in soybean. Int J Mol Sci, 2022, 23: 1721.

[20] Xia Z J, Watanabe S, Yamada T, et al. Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. Proc Natl Acad Sci USA, 2012, 109: E2155–E2164.

[21] Chen L Y, Nan H Y, Kong L P, et al. Soybean AP1 homologs control flowering time and plant height. J Integr Plant Biol, 2020, 62: 1868–1879.

[22] Zhang J L, Chen L, Cai Y P, et al. A novel MORN-motif type gene *GmMRF2* controls flowering time and plant height of soybean. Int J Biol Macromol, 2023, 245: 125464.

[23] Brown A V, Conners S I, Huang W, et al. A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res, 2021, 49: D1496–D1501.

[24] Crossa J, Pérez-Rodríguez P, Cuevas J, et al. Genomic selection

in plant breeding: methods, models, and perspectives. Trends Plant Sci, 2017, 22: 961–975.
[25] Alemu A, Åstrand J, Montesinos-López O A, et al. Genomic selection in plant breeding: key factors shaping two decades of progress. Mol Plant, 2024, 17: 552–578.
[26] Li M, Zhang Y W, Zhang Z C, et al. A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. Mol Plant, 2022, 15: 630–650.
[27] Wang J T, Chen Y, Shu G P, et al. Fast3VmrMLM: a fast algorithm that integrates genome-wide scanning with machine lear_ning to accelerate gene mining and breeding by design for polygenic traits in large-scale GWAS datasets. Plant Commun, 2025, 6: 101385.
[28] Song Q J, Hyten D L, Jia G F, et al. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One, 2013, 8: e54985.
[29] Song Q J, Hyten D L, Jia G F, et al. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes Genomes Genet, 2015, 5: 1999–2006.
[30] Liu Y C, Du H L, Li P C, et al. Pan-genome of wild and cultivated soybeans. Cell, 2020, 182: 162–176.e13.
[31] Liu Y C, Zhang Y, Liu X N, et al. SoyOmics: a deeply integrated database on soybean multi-omics. Mol Plant, 2023, 16: 794–797.
[32] Chen S F. Fastp 1.0: an ultra-fast all-round tool for FASTQ data quality control and preprocessing. iMeta, 2025, 4: e70078.
[33] Van der Auwera G, O'Connor B. Genomics in the cloud: using Docker, GATK, and WDL in Terra. California, CA: O'Reilly Media, Inc., 2020.
[34] Alexander D H, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res, 2009, 19: 1655–1664.
[35] Price M N, Dehal P S, Arkin A P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol, 2009, 26: 1641–1650.
[36] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 2007, 81: 559–575.
[37] Wang J B, Zhang Z W. GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genom Proteom Bioinform, 2021, 19: 629–640.
[38] Fan J W, Shen Y T, Chen C, et al. A large-scale integrated transcriptomic atlas for soybean organ development. Mol Plant, 2025, 18: 669–689.
[39] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet, 2012, 44: 821–824.
[40] Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics, 2014, 198: 483–495.
[41] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res, 2011, 12: 2825–2830.
[42] Liu Y, Wang D L, He F, et al. Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Front Genet, 2019, 10: 1091.
[43] Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer New York, 2009.
[44] Mansur L M, Orf J H, Chase K, et al. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci, 1996, 36: 1327–1336.
[45] Lee S H, Bailey M A, Mian M A, et al. Identification of quantitative trait loci for plant height, lodging, and maturity in a soybean population segregating for growth habit. Theor Appl Genet, 1996, 92: 516–523.
[46] Sonah H, O'Donoughue L, Cober E, et al. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J, 2015, 13: 211–221.
[47] Liu B H, Watanabe S, Uchiyama T, et al. The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis TERMINAL FLOWER1*. Plant Physiol, 2010, 153: 198–210.
[48] 刘路平, 胡雪洁, 祁金, 等. 大豆生育期基因 *E1* 和 *E2* 的启动子克隆及其表达模式分析. 中国农业科学, 2025, 58: 840–850.
Liu L P, Hu X J, Qi J, et al. Cloning of the promoters and analysis of expression patterns of maturity genes *E1* and *E2* in soybean. Sci Agric Sin, 2025, 58: 840–850 (in Chinese with English abstract).
[49] Yu H, Bhat J A, Li C D, et al. Genome-wide survey identified superior and rare haplotypes for plant height in the north-eastern soybean germplasm of China. Mol Breed, 2023, 43: 22.
[50] Wang D, Graef G L, Procopiuk A M, et al. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theor Appl Genet, 2004, 108: 458–467.
[51] Li W X, Zheng D H, Van K, et al. QTL mapping for major agronomic traits across two years in soybean (*Glycine max* L. Merr.). J Crop Sci Biotechnol, 2008, 136: 125.
[52] Bhat J A, Karikari B, Adeboye K A, et al. Identification of superior haplotypes in a diverse natural population for breeding desirable plant height in soybean. Theor Appl Genet, 2022, 135: 2407–2422.
[53] Watanabe S, Xia Z J, Hideshima R, et al. A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering. Genetics, 2011, 188: 395–407.
[54] Li Y Q, Dong Y S, Wu H Y, et al. Positional cloning of the flowering time QTL *qFT12-1* reveals the link between the clock related PRR homolog with photoperiodic response in soybeans. Front Plant Sci, 2019, 10: 1303.
[55] Lu S J, Dong L D, Fang C, et al. Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. Nat Genet, 2020, 52: 428–436.
[56] Zhang J P, Song Q J, Cregan P B, et al. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genom, 2015, 16: 217.
[57] Keim P, Diers B W, Olson T C, et al. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics, 1990, 126: 735–742.
[58] Reinprecht Y, Poysa V W, Yu K F, et al. Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. Genome, 2006, 49: 1510–1527.
[59] Kabelka E A, Diers B W, Fehr W R, et al. Putative alleles for

increased yield from soybean plant introductions. Crop Sci, 2004, 44: 784–791.

[60] Kim K S, Diers B W, Hyten D L, et al. Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. Theor Appl Genet, 2012, 125: 1353–1369.

[61] Sun D S, Li W B, Zhang Z C, et al. Quantitative trait loci analysis for the developmental behavior of Soybean (*Glycine max* L. Merr.). Theor Appl Genet, 2006, 112: 665–673.

[62] Chen Q S, Zhang Z C, Liu C Y, et al. QTL analysis of major agronomic traits in soybean. Agric Sci China, 2007, 6: 399–405.

[63] Alcivar A, Jacobson J, Rainho J, et al. Genetic analysis of soybean plant height, hypocotyl and internode lengths. J Agric Food Environ Sci, 2007, 1: 1–20.

[64] Specht J E, Chase K, Macrander M, et al. Soybean response to water: a QTL analysis of drought tolerance. Crop Sci, 2001, 41: 493–509.

[65] Prince S J, Vuong T D, Wu X L, et al. Mapping quantitative trait loci for soybean seedling shoot and root architecture traits in an inter-specific genetic population. Front Plant Sci, 2020, 11: 1284.

[66] Mansur L M, Lark K G, Kross H, et al. Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). Theor Appl Genet, 1993, 86: 907–913.

[67] Gai J Y, Wang Y J, Wu X L, et al. A comparative study on segregation analysis and QTL mapping of quantitative traits in plants: with a case in soybean. Front Agric China, 2007, 1: 1–7.

[68] Contreras-Soto R I, Mora F, de Oliveira M A R, et al. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PLoS One, 2017, 12: e0171105.

[69] Lee S H, Bailey M A, Mian M A R, et al. Molecular markers associated with soybean plant height, lodging, and maturity across locations. Crop Sci, 1996, 36: 728–735.

[70] Mansur L M, Orf J, Lark K G. Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (*Glycine max* L. Merr.). Theor Appl Genet, 1993, 86: 914–918.

[71] Lark K G, Chase K, Adler F, et al. Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. Proc Natl Acad Sci USA, 1995, 92: 4656–4660.

[72] Guzman P S, Diers B W, Neece D J, et al. QTL associated with yield in three backcross-derived populations of soybean. Crop Sci, 2007, 47: 111–122.

[73] Li R C, Jiang H W, Zhang Z G, et al. Combined linkage mapping and BSA to identify QTL and candidate genes for plant height and the number of nodes on the main stem in soybean. Int J Mol Sci, 2020, 21: 42.

[74] Orf J H, Chase K, Jarvik T, et al. Genetics of soybean agronomic traits: I. comparison of three related recombinant inbred populations. Crop Sci, 1999, 39: 1642–1651.

[75] Liu C, Tian Y, Liu Z X, et al. Identification and characterization of long-InDels through whole genome resequencing to facilitate fine-mapping of a QTL for plant height in soybean (*Glycine max* L. Merr.). J Integr Agric, 2022, 21: 1903–1912.

[76] Niu M R, Tian K W, Chen Q, et al. A multi-trait GWAS-based genetic association network controlling soybean architecture and seed traits. Front Plant Sci, 2024, 14: 1302359.

[77] Clark C B, Zhang D J, Wang W D, et al. Identification and mapping of a recessive allele, *dt3*, specifying semideterminate stem growth habit in soybean. Theor Appl Genet, 2023, 136: 258.

[78] Chang F G, Guo C Y, Sun F L, et al. Genome-wide association studies for dynamic plant height and number of nodes on the main stem in summer sowing soybeans. Front Plant Sci, 2018, 9: 1184.

[79] Wang L S, Li H Y, He M L, et al. GIGANTEA orthologs, *E2* members, redundantly determine photoperiodic flowering and yield in soybean. J Integr Plant Biol, 2023, 65: 188–202.

[80] Wan Z, Liu Y X, Guo D D, et al. CRISPR/Cas9-mediated targeted mutation of the *E1* decreases photoperiod sensitivity, alters stem growth habits, and decreases branch number in soybean. Front Plant Sci, 2022, 13: 1066820.

[81] Jing Y, Zhao X, Wang J, et al. Identification of loci and candidate genes for plant height in soybean (*Glycine max*) via genome-wide association study. Plant Breed, 2019, 138: 721–732.

[82] Spielmeyer W, Ellis M H, Chandler P M. Semidwarf (*sd-1*), "green revolution" rice, contains a defective gibberellin 20-oxidase gene. Proc Natl Acad Sci USA, 2002, 99: 9043–9048.

[83] Sasaki A, Ashikari M, Ueguchi-Tanaka M, et al. A mutant gibberellin-synthesis gene in rice. Nature, 2002, 416: 701–702.

[84] Peng J R, Richards D E, Hartley N M, et al. 'Green revolution' genes encode mutant gibberellin response modulators. Nature, 1999, 400: 256–261.

[85] Liu X L, Huang M, Fan B, et al. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet, 2016, 12: e1005767.

[86] Gao P F, Zhao H N, Luo Z, et al. SoyDNGP: a web-accessible deep learning framework for genomic prediction in soybean breeding. Brief Bioinform, 2023, 24: bbad349.

[87] Montesinos-López O A, Montesinos-López A, Pérez-Rodríguez P, et al. A review of deep learning applications for genomic selection. BMC Genom, 2021, 22: 19.

[88] Salesse-Smith C E, Sharwood R E, Busch F A, et al. Increased Rubisco content in maize mitigates chilling stress and speeds recovery. Plant Biotechnol J, 2020, 18: 1409–1420.

[89] Jarquín D, Crossa J, Lacaze X, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet, 2014, 127: 595–607.