焦宇馨,张宇翔,杨文艳,等. 结合辅助性状的玉米全基因组选择预测力评估[J].江苏农业学报,2023,39(2):313-320.
doi:10.3969/j.issn.1000-4440.2023.02.002

结合辅助性状的玉米全基因组选择预测力评估

焦宇馨[1,2], 张宇翔[1,2], 杨文艳[1,2], 经思宇[1,2], 尹玉琳[1,2], 刘 畅[1,2], 王 欣[1], 徐辰武[1,2], 徐 扬[1,2]

(1.江苏省作物基因组学和分子育种重点实验室/植物功能基因组学教育部重点实验室/江苏省作物遗传生理重点实验室,扬州大学农学院,江苏 扬州 225009; 2.江苏省粮食作物现代产业技术协同创新中心,扬州大学,江苏 扬州 225009)

摘要: 多性状联合全基因组选择能够有效利用性状间的遗传相关和环境相关,有望提高表型预测的准确性。本研究提出了结合辅助性状的全基因组选择策略,以来源广泛的342份玉米自交系为试验材料,对其进行基因分型测序(GBS)并分析其农艺性状,对每个目标性状均基于辅助性状及其组合进行预测,利用五倍交叉验证法评价其预测力。结果表明,利用与目标性状相关性较高的辅助性状可较大程度地提升预测力,尤其是对于低遗传力性状;随着辅助性状个数的增加,预测力也随之增加。进一步比较了5种统计模型结合辅助性状的全基因组选择的表型预测力,总体而言,再生核希尔伯特空间(RKHS)模型和贝叶斯B(BayesB)模型的预测效果较优,而极端梯度提升(XGBOOST)模型的预测效果较差。本研究结合辅助性状有效提高了玉米全基因组选择的预测准确性,为玉米的全基因组选择育种提供新的思路和参考。

关键词: 玉米; 全基因组选择; 辅助性状; 预测力

中图分类号: S513;Q943 **文献标识码:** A **文章编号:** 1000-4440(2023)02-0313-08

Predictability of maize genome-wide selection combined with auxiliary traits

JIAO Yu-xin[1,2], ZHANG Yu-xiang[1,2], YANG Wen-yan[1,2], JING Si-yu[1,2], YIN Yu-lin[1,2], LIU Chang[1,2], WANG Xin[1], XU Chen-wu[1,2], XU Yang[1,2]

(1.*Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Key Laboratory of Plant Functional Genomics of the Ministry of Education/ Jiangsu Key Laboratory of Crop Genetics and Physiology, Agricultural College of Yangzhou University, Yangzhou* 225009, *China*; 2.*Jiangsu Co-innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou* 225009, *China*)

Abstract: Multi-trait genomic selection can use genetic and environmental correlations between traits, which holds great promise to improve the prediction accuracy. This study proposed a genomic prediction strategy using auxiliary traits. A total of 342 maize inbred lines from a diversity panel were used as test materials. Genotyping by sequencing (GBS) was performed and six agronomic traits were measured in the field. Each target trait was predicted based on auxiliary traits and their combinations. The predictability was evaluated using five-fold cross-validation. The results showed that the use of auxiliary traits highly correlated with target traits greatly improved predictability and low-heritability traits could benefit more from auxiliary traits. As the number of auxiliary traits increased, the predictability also increased. We also compared the prediction performance of five different models combined with auxiliary traits. Overall, reproducing kernel Hilbert space (RKHS) model and BayesB model performed well, while extreme gradient boosting (XGBOOST) model performed worst. This study improves the accuracy of genomic prediction and provides new ideas and references for genomic selection breeding of maize.

Key words: maize; genomic selection; auxiliary traits; predictability

收稿日期:2023-02-13

基金项目:国家自然科学基金项目(32170636、32061143030);江苏省重点研发计划项目(BE2022343);江苏省种业振兴揭榜挂帅项目[JBGS(2021)009];江苏省高等学校大学生创新创业训练计划项目(202111117029Z)

作者简介:焦宇馨(2001-),女,黑龙江黑河人,本科,主要从事数量遗传学研究。(E-mail)191702106@stu.yzu.edu.cn

通讯作者:徐 扬,(E-mail)yangx@yzu.edu.cn

玉米是中国最重要的粮食作物之一,为保障国家粮食安全作出重要贡献。“十二五”以来,中国培育了一批优良的玉米品种,其丰产性和稳产性得到了明显提升[1]。然而与一些发达国家相比,中国玉米平均单产偏低,生产成本高,竞争力不强。中国玉米育种仍以常规技术为主,存在预见性差、周期长、效率低等突出问题[2]。随着高通量测序技术的不断发展,全基因组选择育种技术已成为玉米精准育种的重要手段和发展方向。

全基因组选择(GS)是根据训练群体基因型与表型间的关联构建统计模型,从而对未知表型的候选群体进行表型预测和选择[3]。GS 在获取样本基因型时就可对其育种值进行评估,能够大幅提升选择准确性和育种效率,缩短育种周期,实现从经验育种至精准育种的飞跃[4]。一些发达国家玉米商业化育种起步相对较早,全球种业企业如科迪华公司等已运用 GS 技术提高玉米品种选育效率。科迪华公司和先正达公司利用全基因组选择技术分别培育的抗旱玉米品种 AQUAmax 和 Artesian 已进入市场。国际玉米小麦改良中心在全球玉米育种计划中纳入全基因组选择[5]。GS 技术虽然为玉米育种提供了新的契机,但是对受环境影响较大的数量性状来说,其预测准确性仍较低[6-7]。GS 方法的改进一直是 GS 研究的重要课题,也是对品种进行精准选择的关键。

目前 GS 通常针对单个性状进行预测和选择,而忽视了多个关联性状间的遗传基础[8]。多性状联合 GS 不仅能够获取性状间的遗传相关,还能获取性状间的环境相关,有望提升表型预测的准确性,尤其是一些低遗传力的性状[9-10]。在育种研究中,可能会面临一些性状难以测量或观测成本高昂的问题,可以考虑结合较易测量的性状去辅助预测较难鉴定的性状。本研究拟以来源广泛的 342 份玉米自交系为试验材料,对其进行基因分型测序(GBS)并分析产量相关性状,开展结合辅助性状的全基因组选择研究,利用交叉验证评估结合不同辅助性状的全基因组选择预测的效果,进一步比较 5 种不同统计模型对预测准确性的影响,以期为提高玉米全基因组选择的准确性提供技术支撑。

1 材料与方法

1.1 供试材料与试验设计

试验材料为 342 份来自热带、亚热带和温带的玉米自交系。试验材料于 2015 年、2016 年和 2017 年在海南省江苏南繁中心种植。田间试验采用随机区组设计,2 次重复,每份材料重复种植 2 行,行长 3.00 m,行距 0.50 m,株距为 0.25 m。

1.2 基因型分型与表型鉴定

在玉米成熟期,每份材料随机选取 6 株测量株高(*PH*),并选取 6 个长势一致的果穗,测定穗行数(*ERN*)、行粒数(*KNR*)、穗长(*EL*)、穗粗(*ED*)、穗粒质量(*KW*)。利用 R 语言 lme4 软件包,计算 3 个环境(2015 年、2016 年、2017 年)下表型数据的最佳线性无偏估计值用于后续分析。性状广义遗传力的计算公式为:$H^2=\sigma_g^2/(\sigma_g^2+\sigma_\varepsilon^2/l)$,式中,$\sigma_g^2$ 为遗传方差;σ_ε^2 为残差方差;l 为环境个数,方差由 R 语言 lme4 软件包估计。对全部自交系进行基因分型测序(GBS),根据质控标准最小等位基因频率(*MAF*)>0.05 及缺失率<10%,过滤后获得108 193 个单核苷酸多态性(SNP)标记用于后续分析。

1.3 结合辅助性状的全基因组选择预测模型

本研究所使用的基因组最佳线性无偏预测(Genomic best linear unbiased prediction,GBLUP)模型如下:

$$y=P\beta+\sum_{k=1}^{m}Z_k\gamma_k+\varepsilon$$

其中,y 为表型值向量;P 为辅助性状的表型值矩阵;β 为固定效应值;Z_k 为 n 个个体在第 k 个标记的基因型向量;γ_k 为第 k 个标记的效应,假定 $\gamma_k\sim N(0,\frac{1}{m}\phi^2)$;$m$ 为全部标记数目;ϕ^2 为多基因方差;ε 是随机误差,服从 $\varepsilon\sim N(0,I_n\sigma^2)$。则 y 的期望为 $E(y)=P\beta$,方差协方差矩阵为:

$$var(y)=\frac{1}{m}\sum_{k=1}^{m}Z_kZ_k^T\phi^2+I_n\sigma^2=K\phi^2+I_n\sigma^2$$

其中,m 为全部标记数目;Z_k 为 n 个个体在第 k 个标记的基因型向量;K 为亲缘关系矩阵;ϕ^2 为多基因方差,方差组分可由限制性极大似然法估计。

进行辅助性状的全基因组选择预测时,所有辅助性状的表型值均需要进行标准化和中心化,辅助性状为某一单一性状或不同性状的组合。

1.4 模型比较

本研究进一步比较了 5 种统计模型进行辅助性状全基因组选择预测的效果,5 种统计模型包括基因组最佳线性无偏预测(Genomic best linear unbi-

ased prediction,GBLUP)模型、再生核希尔伯特空间(Reproducing kernel Hilbert space,RKHS)模型、最小绝对收缩与选择算子(Least absolute shrinkage and selection operator,LASSO)模型、贝叶斯 B(BayesB)模型、极端梯度提升(Extreme gradient boosting, XGBOOST)模型,所有计算运用 R 语言包 predhy 实现。其中 XGBOOST 模型的学习率设置为 0.07,迭代次数设置为1 000;RKHS 模型采用多核模型,迭代次数设置为1 500,老化(BurnIn)设置为 500,其余参数均采用默认值。

1.5 预测力评估

本研究基于五倍交叉验证法评估预测力,即将样本随机划分为 5 等份,将 4 份作为训练集,1 份用作验证集。预测力采用预测值与实际观测值的决定系数表示。为了避免随机干扰,重复交叉验证过程 20 次,以预测力的平均值作为评价预测准确性的指标。

2 结果与分析

2.1 农艺性状的描述性统计及相关分析

对玉米自交系的 6 个农艺性状[穗行数(*ERN*)、行粒数(*KNR*)、穗粗(*ED*)、穗长(*EL*)、株高(*PH*)、穗粒质量(*KW*)]进行描述性统计分析,结果(表 1)表明,所有农艺性状在自交系间均有丰富变异,变异系数都高于 0.10,其中 *KW* 的变异系数最高,为 0.30,*ED* 的变异系数最低,为 0.11。遗传力分析结果表明,6 个农艺性状的遗传力变化范围为 0.33~0.66,其中 *KNR* 的遗传力最低,*ERN* 的遗传力最高。对 6 个农艺性状进行相关分析,表 2 显示,除 *ERN* 与 *EL*、*ED* 与 *KNR* 之间不存在显著相关外,其余性状间均存在显著正相关,其中 *KW* 与 *KNR* 的相关系数最高,达到 0.650。

表 1 玉米自交系农艺性状的描述性统计和遗传力

Table 1 Descriptive statistics and heritability for agronomic traits of maize inbred lines

农艺性状	均值	变异范围	标准差	农艺性状	变异系数	偏度系数	峰度系数	遗传力
穗行数	13.52	8.00~20.00	1.84	穗行数	0.14	0.59	0.64	0.66
行粒数	20.77	10.00~30.00	3.81	行粒数	0.18	−0.14	−0.30	0.33
穗粗(cm)	3.90	2.31~5.38	0.41	穗粗	0.11	−0.23	1.51	0.62
穗长(cm)	12.03	7.50~16.46	1.61	穗长	0.13	−0.07	−0.15	0.42
株高(cm)	167.59	118.74~231.50	20.02	株高	0.12	0.08	−0.17	0.58
穗粒质量(g)	61.06	21.68~119.24	18.19	穗粒质量	0.30	0.31	−0.34	0.36

表 2 玉米自交系 6 个农艺性状间的相关分析

Table 2 Correlation coefficients between six agronomic traits of maize inbred lines

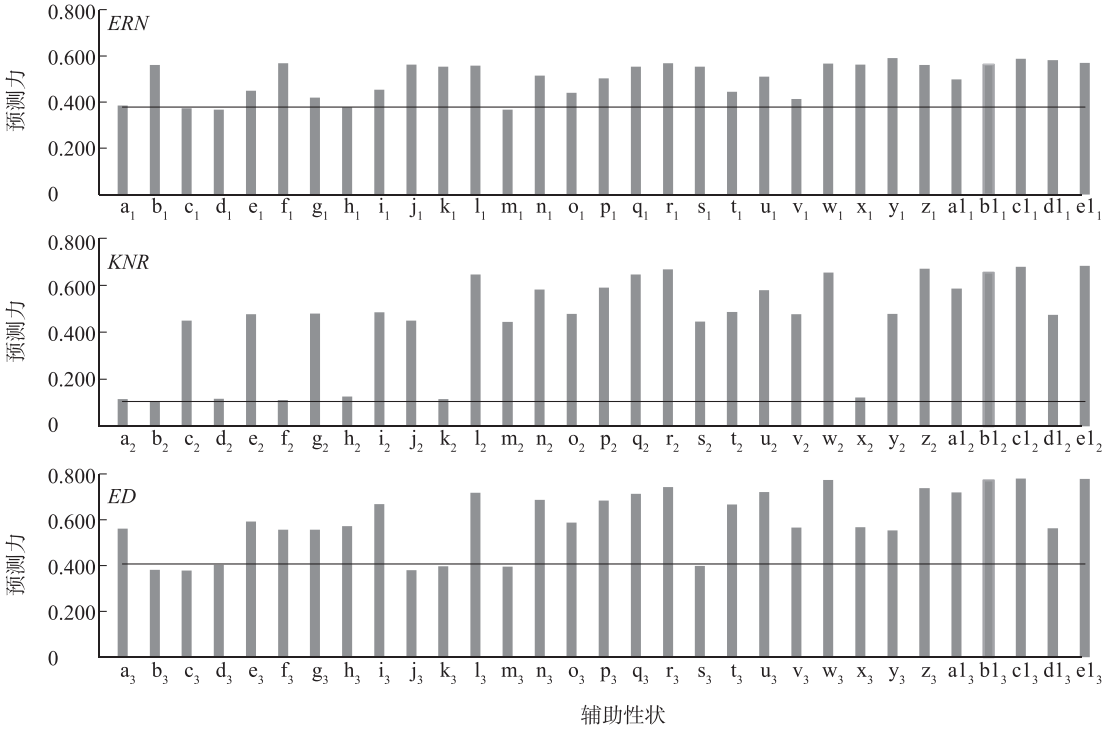
农艺性状	穗行数	行粒数	穗粗	穗长	株高
行粒数	0.115*				
穗粗	0.639***	0.065			
穗长	−0.042	0.582***	0.136*		
株高	0.297***	0.274***	0.296***	0.352***	
穗粒质量	0.416***	0.650***	0.642***	0.566***	0.470***

*、**、*** 分别表示在 0.050、0.010、0.001 水平显著相关。

2.2 结合辅助性状的全基因组选择预测力

为了研究结合辅助性状的玉米全基因组选择的预测效果,本研究基于 GBLUP 模型,对每个目标性状采用辅助性状及其组合的全基因选择的预测,利用五倍交叉验证法评价预测力,并与目标性状自身全基因组选择的预测效果进行比较。结果(图 1、图 2)表明,大部分辅助性状及其组合均可提高目标性状的预测力。预测 *ERN* 时,单一辅助性状中,以 *ED* 为辅助性状时预测力最高,为 0.560,以 *PH* 为辅助性状时预测力最低,仅为 0.367;多性状辅助预测时,以 *ED*、*EL* 和 *KNR* 联合辅助时的预测力最高,为 0.590。预测 *KNR* 时,单一辅助性状中,以 *KW* 为辅助性状时预测力最高,为 0.477,以 *ED* 为辅助性状预测力最低,仅为 0.105;多性状辅助预测时,以 *ERN*、*ED*、*EL*、*PH*、*KW* 联合辅助时的预测力最高,为 0.683。预测 *ED* 时,单一辅助性状中,以 *KW* 为辅助性状时预测力最高,为 0.591,以 *EL* 为辅助性状时预测力最低,为 0.377;多性状辅助预测时,以 *ERN*、*KNR*、*EL* 和 *KW* 联合辅助时的预测力最高,为 0.779。预测 *EL* 时,单一辅助性状中,以 *KNR* 为辅助性状时预测力最高,为 0.581,以 *ED* 为辅助性状时预测力最低,为 0.280;多性状辅助预测时,以 *ERN*、*KNR*、*ED*、*PH*、*KW* 联合辅助时的预测

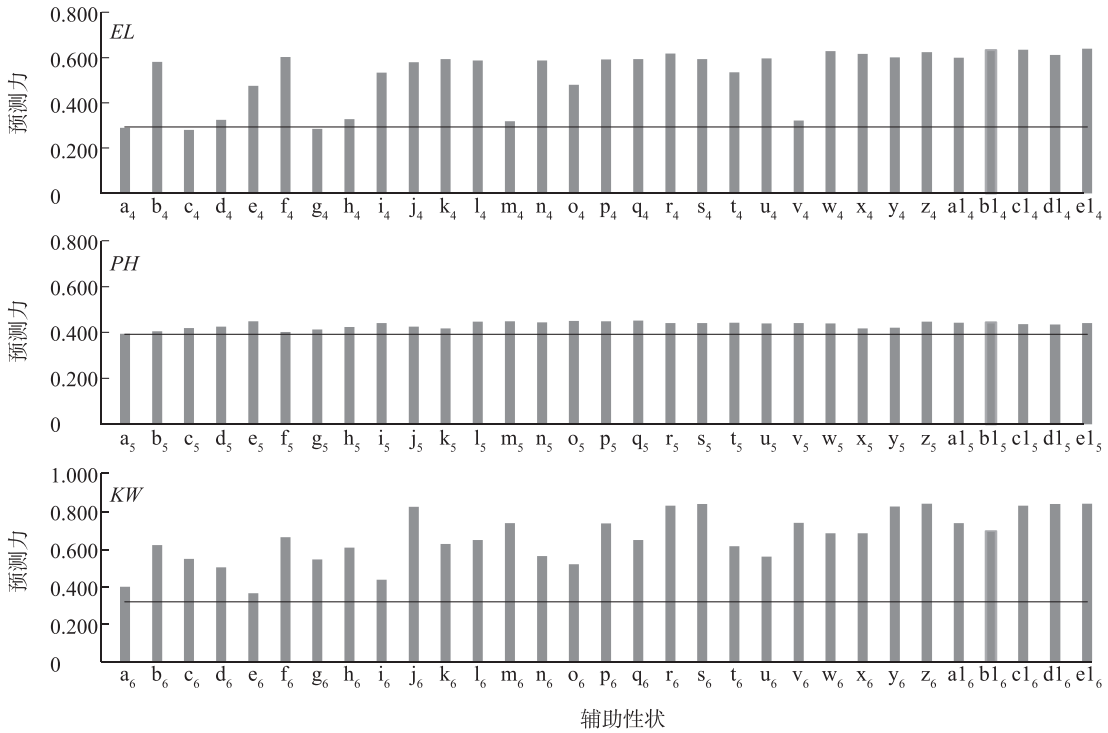
力最高,为 0.639。预测 *PH* 时,单一辅助性状中,以 *KW* 为辅助性状时预测力最高,为 0.449,以 *ERN* 为辅助性状时预测力最低,为 0.394;多性状辅助预测时,以 *KNR*、*EL* 和 *KW* 联合辅助时的预测力最高,为 0.452。预测 *KW* 时,单一辅助性状中,以 *KNR* 为辅助性状时预测力最高,为 0.625,以 *PH* 为辅助性状时预测力最低,为 0.366;多性状辅助预测时,以 *KNR*、*ED*、*EL* 和 *PH* 联合辅助时的预测力最高,为 0.848。对于 *ERN*、*KNR*、*ED*、*EL*、*PH* 和 *KW*,与未结合辅助性状的目标性状本身预测力相比,采取最佳辅助性状组合预测时,预测力分别提高了 0.212、0.577、0.373、0.345、0.060 和 0.526。从结果中可以发现,基于那些与目标性状相关性较高的辅助性状,可在较大程度上提高预测力。例如,*KW* 与 *KNR* 的相关系数最高,基于单一辅助性状预测 *KW* 时,以 *KNR* 为辅助性状的预测力最高,而预测 *KNR* 时,以 *KW* 为辅助性状的预测力最高;*ED* 与 *KNR* 不存在显著相关,基于单一辅助性状预测 *ED* 时,以 *KNR* 为辅助性状的预测力低于 *ED* 自身预测力,预测 *KNR* 时,以 *ED* 为辅助性状的预测力也低于 *KNR* 自身预测力。



ERN:穗行数;*KNR*:行粒数;*ED*:穗粗;*EL*:穗长;*PH*:株高;*KW*:穗粒质量。a_1:*KNR*;b_1:*ED*;c_1:*EL*;d_1:*PH*;e_1:*KW*;f_1:*KNR*+*ED*;g_1:*KNR*+*EL*;h_1:*KNR*+*PH*;i_1:*KNR*+*KW*;j_1:*ED*+*EL*;k_1:*ED*+*PH*;l_1:*ED*+*KW*;m_1:*EL*+*PH*;n_1:*EL*+*KW*;o_1:*PH*+*KW*;p_1:*EL*+*PH*+*KW*;q_1:*ED*+*PH*+*KW*;r_1:*ED*+*EL*+*KW*;s_1:*ED*+*EL*+*PH*;t_1:*KNR*+*PH*+*KW*;u_1:*KNR*+*EL*+*KW*;v_1:*KNR*+*EL*+*PH*;w_1:*KNR*+*ED*+*KW*;x_1:*KNR*+*ED*+*PH*;y_1:*KNR*+*ED*+*EL*;z_1:*ED*+*EL*+*PH*+*KW*;$a1_1$:*KNR*+*EL*+*PH*+*KW*;$b1_1$:*KNR*+*ED*+*PH*+*KW*;$c1_1$:*KNR*+*ED*+*EL*+*KW*;$d1_1$:*KNR*+*ED*+*EL*+*PH*;$e1_1$:*KNR*+*ED*+*EL*+*PH*+*KW*。a_2:*ERN*;b_2:*ED*;c_2:*EL*;d_2:*PH*;e_2:*KW*;f_2:*ERN*+*ED*;g_2:*ERN*+*EL*;h_2:*ERN*+*PH*;i_2:*ERN*+*KW*;j_2:*ED*+*EL*;k_2:*ED*+*PH*;l_2:*ED*+*KW*;m_2:*EL*+*PH*;n_2:*EL*+*KW*;o_2:*PH*+*KW*;p_2:*EL*+*PH*+*KW*;q_2:*ED*+*PH*+*KW*;r_2:*ED*+*EL*+*KW*;s_2:*ED*+*EL*+*PH*;t_2:*ERN*+*PH*+*KW*;u_2:*ERN*+*EL*+*KW*;v_2:*ERN*+*EL*+*PH*;w_2:*ERN*+*ED*+*KW*;x_2:*ERN*+*ED*+*PH*;y_2:*ERN*+*ED*+*EL*;z_2:*ED*+*EL*+*PH*+*KW*;$a1_2$:*ERN*+*EL*+*PH*+*KW*;$b1_2$:*ERN*+*ED*+*PH*+*KW*;$c1_2$:*ERN*+*ED*+*EL*+*KW*;$d1_2$:*ERN*+*ED*+*EL*+*PH*;$e1_2$:*ERN*+*ED*+*EL*+*PH*+*KW*。a_3:*ERN*;b_3:*KNR*;c_3:*EL*;d_3:*PH*;e_3:*KW*;f_3:*ERN*+*KNR*;g_3:*ERN*+*EL*;h_3:*ERN*+*PH*;i_3:*ERN*+*KW*;j_3:*KNR*+*EL*;k_3:*KNR*+*PH*;l_3:*KNR*+*KW*;m_3:*EL*+*PH*;n_3:*EL*+*KW*;o_3:*PH*+*KW*;p_3:*EL*+*PH*+*KW*;q_3:*KNR*+*PH*+*KW*;r_3:*KNR*+*EL*+*KW*;s_3:*KNR*+*EL*+*PH*;t_3:*ERN*+*PH*+*KW*;u_3:*ERN*+*EL*+*KW*;v_3:*ERN*+*EL*+*PH*;w_3:*ERN*+*KNR*+*KW*;x_3:*ERN*+*KNR*+*PH*;y_3:*ERN*+*KNR*+*EL*;z_3:*KNR*+*EL*+*PH*+*KW*;$a1_3$:*ERN*+*EL*+*PH*+*KW*;$b1_3$:*ERN*+*KNR*+*PH*+*KW*;$c1_3$:*ERN*+*KNR*+*EL*+*KW*;$d1_3$:*ERN*+*KNR*+*EL*+*PH*;$e1_3$:*ERN*+*KNR*+*EL*+*PH*+*KW*。横线表示未结合辅助性状的目标性状自身预测力。

图 1 基于 GBLUP 模型结合辅助性状的全基因组选择目标性状(穗行数、行粒数、穗粗)预测力

Fig.1 Predictability of target traits (row number per ear, grain number per row, ear diameter) based on GBLUP model combined with auxiliary traits



ERN:穗行数;*KNR*:行粒数;*ED*:穗粗;*EL*:穗长;*PH*:株高;*KW*:穗粒质量。a_4:*ERN*;b_4:*KNR*;c_4:*ED*;d_4:*PH*;e_4:*KW*;f_4:*ERN+KNR*;g_4:*ERN+ED*;h_4:*ERN+PH*;i_4:*ERN+KW*;j_4:*KNR+ED*;k_4:*KNR+PH*;l_4:*KNR+KW*;m_4:*ED+PH*;n_4:*ED+KW*;o_4:*PH+KW*;p_4:*ED+PH+KW*;q_4:*KNR+PH+KW*;r_4:*KNR+ED+KW*;s_4:*KNR+ED+PH*;t_4:*ERN+PH+KW*;u_4:*ERN+ED+KW*;v_4:*ERN+ED+PH*;w_4:*ERN+KNR+KW*;x_4:*ERN+KNR+PH*;y_4:*ERN+KNR+ED*;z_4:*KNR+ED+PH+KW*;$a1_4$:*ERN+ED+PH+KW*;$b1_4$:*ERN+KNR+PH+KW*;$c1_4$:*ERN+KNR+ED+KW*;$d1_4$:*ERN+KNR+ED+PH*;$e1_4$:*ERN+KNR+ED+PH+KW*。a_5:*ERN*;b_5:*KNR*;c_5:*ED*;d_5:*EL*;e_5:*KW*;f_5:*ERN+KNR*;g_5:*ERN+ED*;h_5:*ERN+EL*;i_5:*ERN+KW*;j_5:*KNR+ED*;k_5:*KNR+EL*;l_5:*KNR+KW*;m_5:*ED+EL*;n_5:*ED+KW*;o_5:*EL+KW*;p_5:*ED+EL+KW*;q_5:*KNR+EL+KW*;r_5:*KNR+ED+KW*;s_5:*KNR+ED+EL*;t_5:*ERN+EL+KW*;u_5:*ERN+ED+KW*;v_5:*ERN+ED+EL*;w_5:*ERN+KNR+KW*;x_5:*ERN+KNR+EL*;y_5:*ERN+KNR+ED*;z_5:*KNR+ED+EL+KW*;$a1_5$:*ERN+ED+EL+KW*;$b1_5$:*ERN+KNR+EL+KW*;$c1_5$:*ERN+KNR+ED+KW*;$d1_5$:*ERN+KNR+ED+EL*;$e1_5$:*ERN+KNR+ED+EL+KW*。a_6:*ERN*;b_6:*KNR*;c_6:*ED*;d_6:*EL*;e_6:*PH*;f_6:*ERN+KNR*;g_6:*ERN+ED*;h_6:*ERN+EL*;i_6:*ERN+PH*;j_6:*KNR+ED*;k_6:*KNR+EL*;l_6:*KNR+PH*;m_6:*ED+EL*;n_6:*ED+PH*;o_6:*EL+PH*;p_6:*ED+EL+PH*;q_6:*KNR+EL+PH*;r_6:*KNR+ED+PH*;s_6:*KNR+ED+EL*;t_6:*ERN+EL+PH*;u_6:*ERN+ED+PH*;v_6:*ERN+ED+EL*;w_6:*ERN+KNR+PH*;x_6:*ERN+KNR+EL*;y_6:*ERN+KNR+ED*;z_6:*KNR+ED+EL+PH*;$a1_6$:*ERN+ED+EL+PH*;$b1_6$:*ERN+KNR+EL+PH*;$c1_6$:*ERN+KNR+ED+PH*;$d1_6$:*ERN+KNR+ED+EL*;$e1_6$:*ERN+KNR+ED+EL+KW*。横线表示未结合辅助性状的目标性状自身预测力。

图 2 基于 GBLUP 模型结合辅助性状的全基因组选择目标性状(穗长、株高、穗粒质量)预测力

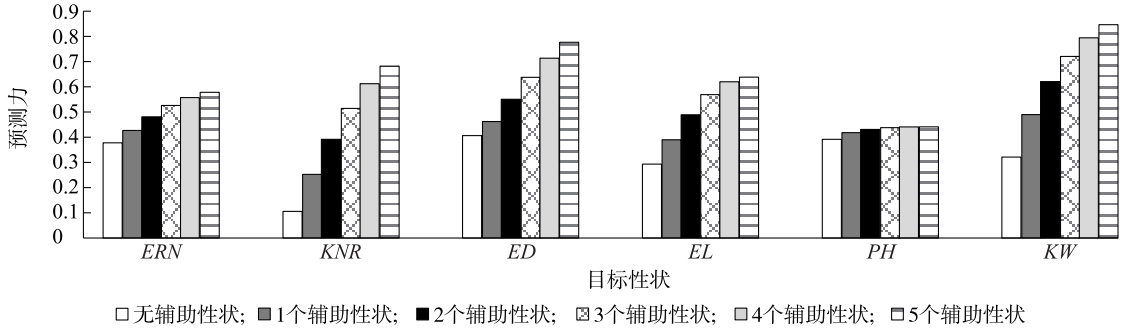
Fig.2 Predictability of target traits (panicle length, plant height, grain weight per panicle) based on GBLUP model combined with auxiliary traits

2.3 辅助性状数目对全基因组选择预测力的影响

为了了解辅助性状数目对目标性状全基因组选择预测力的影响,本研究评估了采用 1 至 5 个辅助性状全基因组选择的预测力,结果(图 3)表明,对于所有目标性状,随着辅助性状数目的增加,预测力提高。对于 *ERN*、*KNR*、*ED*、*EL*、*PH* 和 *KW*,相比目标性状自身全基因组选择预测力,采用 1 个辅助性状时,平均预测力分别提高了 12.95%、139.36%、13.74%、32.82%、6.76%和 52.53%;采用 2 个辅助性状时,平均预测力分别提高了 27.29%、271.27%、35.54%、66.64%、9.98%和 93.39%;采用 3 个辅助性状时,平均预测力分别提高了 39.22%、387.38%、56.93%、93.96%、11.80%和 124.53%;采用 4 个辅助性状时,平均预测力分别提高了 47.51%、480.42%、75.60%、111.30%、12.48%和 147.44%;采用 5 个辅助性状时,平均预测力分别提高了 53.03%、546.76%、91.12%、117.56%、12.63% 和 163.67%。

2.4 不同 GS 模型对全基因组选择预测力的影响

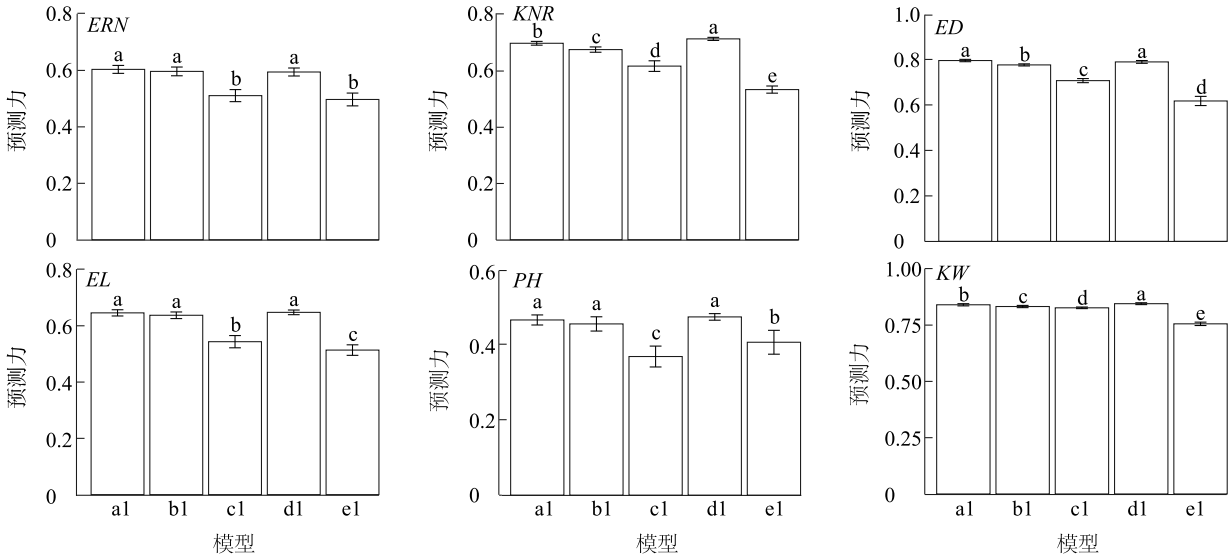
上述研究结果表明,采用 GBLUP 模型结合辅助性状全基因组选择时,对目标性状预测力有较大幅度提升,为了探究合适的预测模型,本研究进一步比较了 GBLUP、BayesB、LASSO、RKHS 和 XGBOOST 这 5 种 GS 模型对于结合全部辅助性状全基因组选择对目标性状的预测力。结果(图 4)表明,预测 *ERN* 时,BayesB、GBLUP、RKHS 模型预测效果最优,预测力分别为 0.604、0.596 和 0.594,LASSO、XGBOOST 模型预测效果较差,预测力分别为 0.510 和 0.497;预测 *KNR* 时,RKHS 模型预测效果最优,预测力为 0.714,XGBOOST 模型预测效果较差,预测力为 0.534;预测 *ED* 时,BayesB、RKHS 模型预测效果最优,预测力分别为 0.795 和 0.789,XGBOOST 模型预测效果最差,预测力为 0.617;预测 *EL* 时,RKHS、BayesB、GBLUP 模型预测效果最优,预测力分别为 0.648、0.646 和 0.637,XGBOOST 模型预测效果最差,预测力为 0.513;预测 *PH* 时,RKHS、BayesB、GBLUP 模型预测效果最优,预测力分别为 0.472、0.464 和 0.453,LASSO 模型预测效果最差,预测力为 0.367;预测 *KW* 时,RKHS 模型预测效果最优,预测力为 0.845,XGBOOST 模型预测效果最差,预测力为 0.756。总体而言,RKHS 模型和 BayesB 模型的预测效果较优,XGBOOST 模型的预测效果较差。



ERN:穗行数;*KNR*:行粒数;*ED*:穗粗;*EL*:穗长;*PH*:株高;*KW*:穗粒质量。

图 3 辅助性状数目对目标性状全基因组选择预测力的影响

Fig.3 Effects of the number of auxiliary traits on the predictability of target traits



ERN:穗行数;*KNR*:行粒数;*ED*:穗粗;*EL*:穗长;*PH*:株高;*KW*:穗粒质量。a1:BayesB 模型;b1:GBLUP 模型;c1:LASSO 模型;d1:RKHS 模型;e1:XGBOOST 模型。

图 4 5 种 GS 模型结合全部辅助性状的全基因组选择预测力比较

Fig.4 Comparison of predictability of five GS models combined with all auxiliary traits

3 讨 论

本研究开展了结合辅助性状的玉米 GS 预测力研究,有效提升了全基因组选择的准确性,主要原因是该策略能够利用性状间的相关性,因此预测效果与性状间的相关性存在一定的关联。利用与目标性状相关性较高的辅助性状可以最大程度地提高全基因组选择预测力,利用与目标性状不相关的辅助性状可能反而会降低全基因组选择对目标性状的预测力。在作物育种研究中,一些性状较难获取,如产量、抗逆、根系性状等,如果采用同一群体的其他相关性状进行辅助预测,能够有效降低育种成本,具有重要的实际应用价值。在利用辅助性状全基因组选择时,低遗传力性状预测力的提升相比于高遗传力性状更为明显,穗行数、穗粗和株高的遗传力均大于0.5,采用全部辅助性状全基因组选择预测时,它们的预测力提升幅度均小于 100%,而行粒数、穗长和穗粒质量的遗传力均小于 0.5,采用全部辅助性状全基因组选择预测时,它们的预测力提升幅度均大于100%。这可能是因为低遗传力性状更易被环境因素影响,而结合辅助性状全基因组选择时,有效借助了性状间的相关环境信息。

本研究分析了不同辅助性状数目和统计模型对预测力的影响。有研究结果表明,在多性状联合分析中,辅助性状数目达到一定数量后,继续增加的辅助性状对单个特定性状预测力提升的贡献较低,并且随着辅助性状数目的增加,运算复杂度会大大增加[11]。在本研究中,尽管个别单一辅助性状也能较大程度提升预测力,但总体而言随着辅助性状数目的增加,预测准确性也随之增加,采用更多辅助性状能够更大幅度提升对目标性状预测的准确性。本研究的优势在于将辅助性状视为固定效应,因而增加辅助性状几乎不影响模型运算效率。在全基因组选择中,通过获取更多的相关表型信息辅助预测目标性状,有望进一步提高预测力。统计模型是影响 GS 准确性的关键因素[12],本研究比较了 GBLUP、BayesB、LASSO、RKHS 和 XGBOOST 这 5 种 GS 模型结合辅助性状全基因组选择的预测效果,整体而言,RKHS 模型和 BayesB 模型的预测效果较优,而 XGBOOST 模型的预测效果较差。BayesB 模型能够对大部分位点的效应进行压缩,因此更适于捕获显著位点效应[13-14]。有研究结果表明,BayesB 模型对基因的数量较为敏感,当性状由少数效应较大的基因控制时,预测力较高,当性状由许多微效基因控制时,预测力有所降低[15]。RKHS 模型的主要优势是擅于捕获一些非加性效应[16]。XGBOOST 是经典的机器学习算法,其预测力较低的原因可能是计算复杂度较高且调参数难度较大,易造成过拟合。

4 结 论

本研究提出了结合辅助性状的玉米 GS 育种新策略,以来源广泛的 342 份玉米自交系为试验材料,对其进行 GBS 并鉴定 6 个农艺性状,对每个目标性状均基于辅助性状及其组合进行预测,利用五倍交叉验证法评价预测力。结果表明,利用与目标性状相关性较高的辅助性状可较大程度地提高预测力;低遗传力性状的预测力提升相比高遗传力性状更为明显;随着辅助性状个数的增加,目标性状的预测准确性也随之增加。本研究进一步比较了 5 种 GS 模型结合辅助性状的全基因组选择的预测力,总体而言,RKHS 模型和 BayesB 模型预测效果较优,而 XGBOOST 模型预测效果较差。本研究有效提升了玉米表型预测的准确性,尤其对于一些低遗传力性状,研究结果能为玉米的 GS 育种提供重要支撑。

参考文献:

[1] 王振华,刘文国,高世斌,等. 玉米种业的昨天、今天和明天[J]. 中国畜牧业, 2021(19): 26-32.

[2] 黎 裕,徐辰武,秦 峰,等. 玉米生物育种:现状与展望[J]. 中国基础科学, 2022, 24(4): 18-28.

[3] MEUWISSEN T H, HAYES B J, GODDARD M E. Prediction of total genetic value using genome-wide dense marker maps[J]. Genetics, 2001, 157(4): 1819-1829.

[4] XU Y, LIU X, FU J, et al. Enhancing genetic gain through genomic selection: from livestock to plants[J]. Plant Communications, 2020, 1(1). DOI:10.1016/j.xplc.2019.100005.

[5] ZHANG X, PÉREZ-RODRÍGUEZ P, BURGUEÑO J, et al. Rapid cycling genomic selection in a multiparental tropical maize population[J]. G3, 2017, 7(7): 2315-2326.

[6] MILLET E J, KRUIJER W, COUPEL-LEDRU A, et al. Genomic prediction of maize yield across European environmental conditions[J]. Nat Genet, 2019, 51: 952-956.

[7] ALLIER A, TEYSSEDRE S, LEHERMEIER C, et al. Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs[J]. Theoretical and Applied Genetics, 2020, 133(1): 201-215.

[8] SCUTARI M, HOWELL P, BALDING D J, et al. Multiple quanti-

tative trait analysis using bayesian networks[J]. Genetics, 2014, 198(1): 129-137.

[9] HENDERSON C, QUAAS R. Multiple trait evaluation using relatives' records[J]. Journal of Animal Science, 1976, 43(6): 1188-1197.

[10] HAYASHI T, IWATA H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits[J]. BMC Bioinformatics, 2013, 14. DOI: 10.1186/1471-2105-14-34.

[11] SCHULTHESS A W, WANG Y, MIEDANER T, et al. Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes[J]. Theoretical and Applied Genetics, 2016, 129(2): 273-287.

[12] CROSSA J, PEREZ-RODRIGUEZ P, CUEVAS J, et al. Genomic selection in plant breeding: methods, models, and perspectives [J]. Trends Plant Science, 2017, 22(11): 961-975.

[13] GONZÁLEZ-RECIO O, FORNI S. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning[J]. Genetics Selection Evolution, 2011, 43(1). DOI: 10.1186/1297-9686-43-7.

[14] PÉREZ P, CAMPOS G D L. Genome-wide regression and prediction with the BGLR statistical package[J]. Genetics, 2014, 198(2): 483-495.

[15] WANG X, YANG Z F, XU C W. A comparison of genomic selection methods for breeding value prediction[J]. Science Bulletin, 2015, 60(10): 925-935.

[16] DE LOS CAMPOS G, GIANOLA D, ROSA G J, et al. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods[J]. Genetics Research, 2010, 92(4): 295-308.

(责任编辑:王 妮)