

全基因组选择技术在玉米育种中的应用

曹士亮^{1,2}, 张学才^{3,4}, 张建国^{1,2}, 扈光辉^{1,2}, 于滔^{1,2},
曹靖生^{1,2}, 杨耿斌^{1,2}, 李文跃^{1,2}, 马雪娜^{1,2}

(1. 黑龙江省农业科学院玉米研究所, 哈尔滨 150086; 2. 玉米国家工程实验室(哈尔滨), 哈尔滨 150086;

3. 中国农业科学院作物科学研究所, 北京 100081; 4. 中国农业科学院国家南繁研究院, 海南 三亚 572024)

摘要: 全基因组选择(Genomic selection或Genome-wide selection, GS)是指利用覆盖各染色体范围的分子标记对作物的表型进行的预测和选择。随着分子标记技术、数理统计的发展和计算机运算能力的提升,全基因组选择逐渐从理论走向实践,在作物种质资源创新和品种选育中发挥着越来越重要的作用,通过与常规育种的有效结合带来了育种方式的新飞跃。全基因组选择能够降低育种成本和加速育种进程,实现作物育种由鉴定育种向预测育种的转变。本文主要介绍全基因组选择技术产生的背景、应用流程和影响全基因组选择预测精度的因素,回顾和展望该技术在玉米种质创新和杂交组合预测上应用的现状与前景,提出当前开展玉米全基因组选择工作面临的问题和建议。

关键词: 玉米; 全基因组选择; 影响因素; 展望

中图分类号: S513.035.3

文献标识码: A

Application of Genomic Selection in Maize Breeding

CAO Shi-liang^{1,2}, ZHANG Xue-cai^{3,4}, ZHANG Jian-guo^{1,2}, HU Guang-hui^{1,2}, YU Tao^{1,2},

CAO Jing-sheng^{1,2}, YANG Geng-bin^{1,2}, LI Wen-yue^{1,2}, MA Xue-na^{1,2}

(1. Maize Research Institute, Heilongjiang Academy of Agricultural Sciences, Harbin 150086; 2. National Engineering Laboratory for Maize(Harbin), Harbin 150086; 3. Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081; 4. Nanfan Research Institute, CAAS, Sanya 572024, China)

Abstract: Genomic selection or Genome-wide selection(GS) may be defined as the selection of traits with markers simultaneously, which cover the entire genome in a dense manner. With advancements in molecular marker technology, statistical methods, and computing power, GS has transitioned from theory to practice, playing an increasingly important role in crop germplasm innovation and breeding. Through effective integration with conventional breeding, GS has brought a significant leap in breeding methods. It can reduce the breeding cost, accelerate the breeding process, and realize the transformation from identification breeding to prediction breeding. This paper introduces the background, basic process, and factors that affect the prediction accuracy of GS. Using maize as an example, it reviews the current situation and prospects of the application of this technology in maize germplasm innovation and prediction of maize hybrid, and puts forward the problems and suggestions faced by the current implementation of maize genomic selection, hoping to provide some references for accelerating the application of this technology in maize breeding.

Key words: Maize; Genomic selection; Influencing factor; Prospect

录用日期: 2024-04-18

基金项目: 黑龙江省省属科研院所科研业务费(CZKYF2023-1-A003)、黑龙江省博士后科研启动资金项目、黑龙江省农业科学院创新跨越工程项目(CX23TS02)、黑龙江省头雁团队项目

作者简介: 曹士亮(1980-), 男, 哈尔滨人, 博士, 从事玉米分子育种研究。E-mail: caoshiliang2003@126.com
张学才为本文通信作者。

1 作物育种的预测与选择

植物育种是预测的科学, 预测是选择的基础^[1]。选择是最古老的育种技术, 也是迄今为止仍被最广泛应用的技术^[2]。植物育种中的预测主要包括利用高世代材料对后代的预测、利用亲本材料对杂交种的预测和利用某年份或地点的表现对其他年

份或地点的预测。世界作物育种发展大致经历了原始驯化选育(育种1.0)、常规育种(育种2.0)和分子育种(育种3.0)3个主要阶段^[3]并向智能育种(育种4.0)阶段迈进,预测与选择始终贯穿于植物育种发展的各个阶段。

1.1 驯化育种阶段的预测与选择

驯化育种始于约1万年前,主要包括野生资源向农作物的驯化和农家种的选育^[4-5]。在这一阶段预测与选择的特点是育种者(通常是耕种者)通过表型观察进行作物驯化与改良,有意识地选择那些果穗或者子粒较大的植物类型或个体,使栽培种相对于其祖先越来越能满足人类的需要。这个阶段的预测通常是一种主观意识的预测,选择的周期十分漫长。

1.2 常规育种阶段的预测与选择

孟德尔的研究为选择提供了可预测性框架,所揭示的遗传定律催生了遗传学的诞生并长期影响着育种的发展。常规育种阶段的预测与选择由驯化育种阶段的“潜意识”发展到基于遗传学和数量遗传学的“有意识”设计。随着孟德尔的遗传学基本原理的推广和扩展以及田间实验设计与统计分析技术的快速发展,对遗传学的理解及其在植物育种中的应用更加深入,育种的科学性越来越受到重视。

常规育种阶段形成的育种家方程成为数量性状改良和育种工作的重要指导原则。育种家方程可以用公式 $\Delta G = ih\sigma_A/L$ 来表示。其中, ΔG 代表遗传增益; i 为选择强度; h 为狭义遗传力的平方根; σ_A 为加性遗传方差的平方根; L 为育种周期间隔或世代数。因此育种家方程也被称为遗传增益公式^[6]。由以上公式可知,增加选择强度、提高狭义遗传力、增加优良等位基因的加性方差以及缩短育种周期都是提高遗传增益的重要手段。理论上,作物育种又被称作遗传改良,因此遗传增益公式应该是评价育种效率的高低的重要指标。在实际上,直接以遗传增益对育种效率进行评价在植物育种中相对较少,更习惯于以育成优良品种的数量、品种的推广面积和影响对育种工作进行评价。

1.3 分子育种阶段的预测与选择

分子标记辅助选择、转基因和基因编辑共同构成了分子育种的核心技术^[7]。从预测和选择的角度来看,转基因和基因编辑主要针对功能基因本身,目标明确、效果直接,如外源抗虫、抗除草剂基因的导入,对普通玉米 *Waxy* 基因进行编辑创造糯玉米材料,这里的预测和选择几乎是等价的,即改变了基因就改变了性状^[8]。分子标记辅助选择通常是指对有

主效基因控制的性状进行的选择,操作对象是1个或少数几个分子标记,这些分子标记与性状是一种相关关系而非因果关系,其预测和选择效果往往不如转基因和基因编辑直接。全基因组选择是对复杂数量性状进行的选择,相比MAS方法,全基因组选择模型中包括了覆盖于全基因组的分子标记,能更好地解释表型变异。无论是标记辅助选择还是全基因组选择,如何量化分子标记与表型的数学关系是分子育种阶段预测与选择面临的关键问题。

1.4 智能育种阶段的预测与选择

育种4.0时代是“生物技术+信息技术+人工智能+大数据应用”育种,可以简单地概括为智能育种时代,是一个新型生物技术叠加图像识别、大数据、人工智能等信息技术融合的育种阶段。全基因组选择仍然是育种4.0阶段的核心技术之一,此阶段的预测和选择的对象不仅仅是基因组数据,而是扩展到了代谢组、蛋白组、表型组等多组学以及气象、土壤等环境条件数据。通过组学大数据与生物信息学、基因编辑技术与合成生物学、人工智能与机器学习技术等多学科、多领域的共同支撑,使基因组智能设计育种逐渐成为现实,将推动育种方向由鉴定优先向计算优先转变,育种的数字化、信息化、智能化、工程化的特点更加明显,育种智能数据决策系统的作用更加突出。

2 全基因组选择技术及其选择准确性的主要影响因素

2.1 全基因组选择的概念及基本流程

全基因组选择(Genomic selection, GS)的概念最早由Haley和Visscher于1998年提出^[10]。Meuwissen的研究提供了全基因组选择的具体方法,推动了全基因组选择技术的应用和发展^[11]。与分子标记辅助选择(MAS)不同,全基因组选择不需要事先确定控制性状的主效基因或QTL,而是利用训练群体(也叫建模群体)的表型和覆盖整个基因组的分子标记来建立基因组估计育种值(Genomic Estimated Breeding Value, GEBV)与分子标记间的回归模型,再利用这个模型和其他群体(候选群体)的基因型对其他群体的表型进行预测和选择。

在开展全基因组选择育种工作前,一般需要预测模型构建并对影响预测精度的因素进行优化,这一过程通常被称为全基因组预测(Genomic Prediction, GP)。用于模型构建的群体可以是现有的育种群体,也可以利用针对某一研究目的专门构建的遗传群体。全基因组预测的预测精度(Prediction accu-

racy)定义为验证群体基因组估计育种值和表型值的相关系数(通常是皮尔森相关系数)的均值,是评价模型预测效果的重要指标。用于模型预测精度评价的群体通常叫作测试群体或验证群体(Test population 或 Validation population),验证群体与训练群体合在一起通常叫作参考群体(Reference Population)。验证群体可以是训练群体的一部分,也可以是另外单独的群体。独立的验证群体,可以是不同年份的育种群体,也可以是选择晋级后的田间试验群体等。在全基因组预测精度评价时,通常采用5倍交叉验证(five-fold cross-validation)的方法,即抽

取参考群体的80%作为建模群体,剩余20%作为验证群体^[12]。当预测模型建立后,对于育种群体(动物育种中常称为候选群体)只需要进行基因型测定,再利用预测模型和育种群体基对育种群体进行表型预测,再筛选出预测表现优良的部分材料进行表型鉴定,这样会提高选择强度和减少表型鉴定的工作量。由于基因分型成本的不断下降,且表型鉴定成本的显著提升,全基因组选择技术已经逐渐成为作物育种的一种重要技术手段且越来越受到重视。全基因组选择的基本流程见图1。

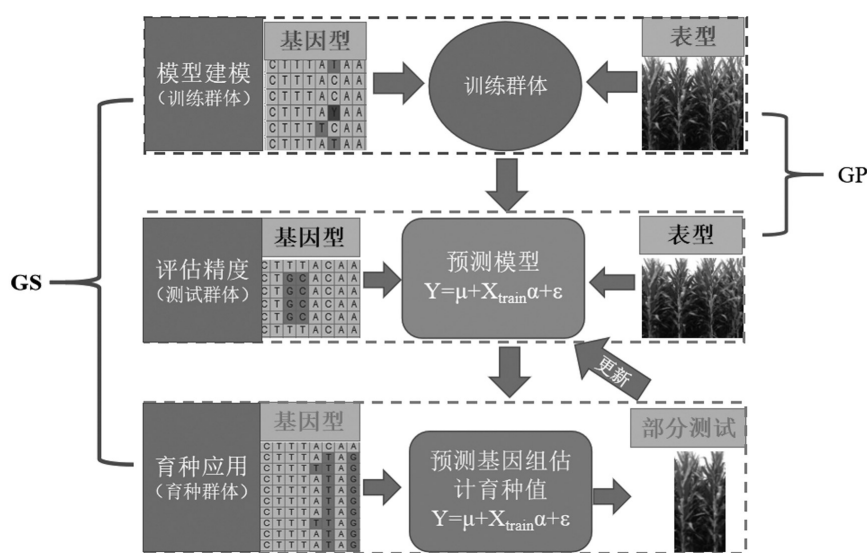


图1 全基因组选择的基本流程

Fig.1 Basic process of genomic selection

2.2 影响全基因组选择预测精度的因素

由全基因组选择的基本流程可知,全基因组选择至少受到建模群体大小、基因型、表型、预测模型等因素的影响。

2.2.1 建模群体的大小和组成对预测精度的影响及其优化

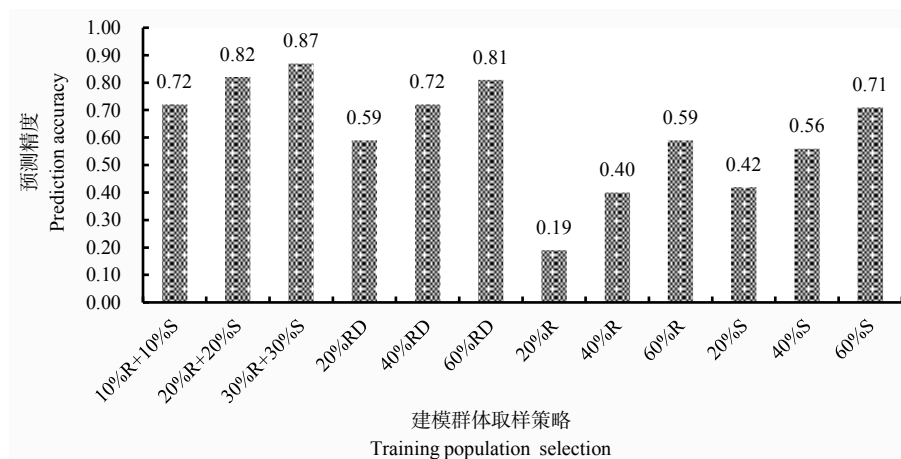
建模群体大小和组成不仅影响全基因组选择的预测精度还影响全基因组选择的成本。一般而言,预测精度会随着建模群体样本大小的增加而增大^[13]。建模群体的组成包括建模群体遗传多样性、个体的亲缘关系以及建模群体与验证群体的关系,这些因素单独或相互作用影响全基因组选择的预测精度。当建模群体规模较小时,GS预测精度下降的一个重要原因是较小的样本不足以捕捉遗传变异的整体信息。但当建模群体与验证群体亲缘关系较近时,相对较小的样本也会获得较高的预测精度^[14-15]。Riedelsheimer等研究了建模群体的遗传组成对预测

精度的影响,发现DH系全同胞家系要比半同胞家系预测精度提高42%^[16]。Zhao等研究发现,依据表型的双向选择由于具有更高的遗传多样性要比随机选择样本构建建模群体获得更高的预测精度^[17]。Cao等针对玉米TSC的全基因组预测研究发现,分别选择20%、40%和60%的样本预测剩余的样本时,同样比例包括抗感(R+S)的样本的预测精度要高于随机样本和仅选择抗(R)或感病(S)样本的预测精度(图2)。

确定建模群体的大小和组成已经成为全基因组选择的关键,这一过程常被称为建模群体优化。Technow等依据杂种优势群来优化建模群体的大小,发现把马齿和硬粒两个群整合在一起作为建模群体时会显著提高对每一个杂种优势群的玉米叶枯病的预测精度^[18]。Akdemir等不考虑材料的表型,而是根据材料的基因型通过一定的遗传算法筛选一部分样本作为建模群体,然后再对这部分材料进行表

型鉴定进而建立预测模型^[19]。Dnyaneshwar发现,采用预测误差方差(Prediction Error Variance, PEV)和决定系数(Coefficient of Determination, CD)进行训练群体优化可以提高预测精度^[20]。为了解决多世代数据预测带来的训练群体遗传关系的复杂性,CIM-MYT开发了稀疏选择指数(The sparse selection index, SSI)来进行训练群体优化来去除掉训练群体中不重要的个体,达到了提高预测精确性和减少预测

群体的目的。Isidro等研究了群体结构对预测精度的影响,并对群体结构存在的情况下建模群体构建的方法进行了优化,发现与分层抽样和随机抽样相比,基于平均预测误差 PEVmean 和平均决定系数 CDmean 的群体优化方法能够获得更高的预测精度,最大化标记变异、减少标记间的共线性以及均匀化采样是训练群体设计的3个重要目标^[22]。



注:R表示抗病材料,S表示感病材料,RD表示随机选择材料。

Note: R represents disease-resistant materials, S represents susceptible materials, and RD represents randomly selected materials.

图2 不同训练群体表型取样策略的预测精度

Fig.2 Prediction accuracy under different training population selection of phenotype

2.2.2 分子标记对全基因组选择预测精度的影响

快速发展的分子标记技术是推动全基因组选择从理论走向应用的重要因素。相对于分子标记辅助选择,全基因组选择对标记的数量和覆盖度有了更高的要求。如何权衡和优化标记类型、密度和质量是全基因组选择研究的重要内容。在不同类型的分子标记中,SNP标记因其具有共显性遗传、数量丰富和易于实现自动化及基因分型错误率相对较低的优点,已经成为当前全基因组预测研究和全基因组选择育种中应用最广泛的分子标记。

研究表明,高密度的覆盖各个染色体的分子标记会捕获更多的遗传信息,因此,在一定范围内,随着标记的增加全基因组预测精度呈增加的趋势^[23-24],但这种增加并非是线性的,而是到达一定的程度后就增加得很少或者基本不再增加,这可能与染色体上的标记本身存在效应差异有关。Meuwissen等的模拟研究表明,当标记的连锁不平衡 r^2 值从0.1提高至0.2时,GBEV的预测准确性可以提高10个百分点以上,但在1个单倍型内紧密连锁的标记的增加将不会提高预测精度,反而会增加预测的计算

工作量,因此一些研究也尝试利用单倍型进行预测来提高预测精确度。

标记质量存在的问题包括基因分型错误、较高的标记缺失率以及较低的标记多态性(通常用小等位基因频率来表示)。这些因素都是造成全基因组选择预测精度较低的重要原因。除此之外,尽管全基因组选择不需要事先确定每一个标记与性状的关系,但是一些研究表明,与性状相关性不同的标记在预测效果上存在着差异。随机选择覆盖染色体上的标记和有目的选择与性状相关的标记或者不同效应的标记用于全基因组选择效果会存在很大差异,由此产生的标记选择问题也成了全基因组选择研究的一个热点^[24]。

2.2.3 统计模型对全基因组选择预测精度的影响

在育种的发展过程中,对于植物表型的预测是早于分子标记技术的应用的。数理统计和计算机技术推动了预测方法的进步。1976年,Henderson发明了能够同时计算随机效应和固定效应的最优线性无偏估计方法^[25]。1994年,Bernardo^[26]在Henderson的混合线性模型的基础上,将BLUP模型中基于系

谱的亲缘关系矩阵(A矩阵)替换为基于分子标记的关系矩阵(G矩阵),开辟了作物全基因组选择里程碑式的工作,由于采用了基因组标记信息的亲缘关系矩阵作为方差协方差矩阵,该模型被称为GBLUP模型^[27]。科学家们后续在GBLUP模型的基础上又发展了综合了标记亲缘关系矩阵和系谱的SSBLUP模型,以及结合了G矩阵和非线性高斯核(Gaussian kernel, GK)的再生希尔伯特空间回归模型(Reproducing Kernel Hilbert Spaces, RKHS)。预测模型不需要估计标记的效应,只需要利用标记或者系谱信息构建材料关系矩阵,因此这些方法被称为直接法全基因组选择^[28]。

与直接法相对应的即间接法全基因组选择,间接法先对标记效应进行估计,再对个体的基因组估计育种值进行预测。根据标记效应和方差的假设不同,这些模型主要包括岭回归最优线性无偏估计(rr-BLUP)、贝叶斯(Bayes)等方法。rr-BLUP的方法将标记效应作为随机效应,假设所有的标记具有相等的非零效应的方差,采用同等的压缩方法将预测子效应向0压缩,有效地解决了全基因组选择面临的“大p小n”问题。贝叶斯方法发展的各种模型区别主要在于对标记效应及方差分布的假设不同。BayesB模型认为,大部分标记无效应,只有少部分标记有效应,方差服从逆卡方分布,实际上是对标记进行了选择^[29]。整体看来,GBLUP和rr-BLUP相对更适合大量的微效基因的情况,贝叶斯方法则更适合于一些标记存在相对较大的效应的情况。全基因组选择的方法除了按照直接法和间接法进行分类外,还可以分为参数法和非参数法^[30]。近年来,机器学习的方法被广泛地应用于全基因组选择研究^[31-32]。除了上述一些因素,性状的遗传力、环境效应及其与基因型单独或共同影响着预测精度。

3 全基因组选择在玉米育种中的应用

3.1 全基因组选择研究在玉米遗传改良中的应用

全基因组选择在作物遗传改良上的应用最初是从对模拟数据的研究开始的。Piyasatian等通过模拟数据比较了10个世代的选择响应,发现在遗传力较高时,全基因组选择相对于标准MAS的优势较小;当遗传力较低时,全基因组选择通过增加具有不同效应的基因位点的频率,比仅基于显著标记位点的MAS有更大的选择响应^[33]。国际玉米小麦改良中心(CIMMYT)张学才博士设计了多亲群体的快速轮回全基因组选择(Rapid Cycling Genomic Selection, RCGS),采用18个优良的热带玉米自交系经过两次

杂交和1次自交,形成了循环选系的基础群体(C_0),通过4个点的表型鉴定选择表现最优的个体作为RCGS的 C_1 轮的亲本。通过全基因组预测形成 C_1 ,并选择了预测最佳的品系作为 C_2 的父、母本。每年达到两个循环,进行4轮;从 C_1 至 C_4 的实现产量达到每轮 0.225 t/hm^2 ,相当于在4.5年的育种期内产量每年提高 0.1 t/hm^2 ,证明了全基因组选择是一种有效的育种策略,可以保持遗传多样性并实现高遗传增益^[34]。

3.2 全基因组选择在玉米自交系或DH系评价和选择上的应用

对玉米自交系或DH系的评价是玉米育种的重要工作之一。研究表明,在早代测交测试中,通过采用低成本分子标记和基于基因组估计育种值的选择,全基因组选择可以部分代替多环境表型鉴定和表型选择^[35]。对玉米自交系群体(自然群体)或DH群体进行全基因组选择的研究可以从现有的关联分析和连锁分析所应用的群体入手。Guo等采用了一个此前用作QTL作图的巢式关联分析群体,采用全基因组选择的RR-BLUP模型对雌穗抽丝天数的预测精度达到0.43,较符合区间作图法(Composite Interval Mapping, CIM)法提高34%;雄穗抽丝天数的预测精确度较CIM提高30%(预测精度为0.43),雌雄间隔期较CIM高47%(预测精度为0.25),证明了全基因组选择方法的有效性^[36]。Gowda利用两个自然群体研究了玉米坏死病(maize lethal necrosis, MLND)的全基因组选择预测精度,采用RR-BLUP的预测方法,两个群体IMAS-AM和DTMA-AM的预测精度分别为0.56和0.36^[37]。利用同一套群体的基因型只要具有表型数据就可以针对某一性状进行预测,在建立好不同性状的预测模型后,对育种群体进行1次基因分型即完成不同性状的预测。如CIMMYT利用DTMA群体分别开展了TSC、MLND^[37]等不同性状的全基因组选择预测工作,为同时开展这些玉米病害的全基因组预测奠定了重要的基础。对于通过表型进行病害鉴定,很难同时对同一单株进行不同病害的评价,这也是全基因选择的突出优势之一。

3.3 全基因组选择在玉米杂交组合预测上的应用

单倍体技术的应用使育种单位每年产生纯系的数量大幅上升,带来杂交组合选配和鉴定的压力越来越大,杂交组合鉴定的成本占玉米育种成本的比重日益增加。利用全基因组选择开展玉米杂交组合的预测已经成为国际种业竞争的核心技术之一。

3.3.1 利用基因型进行杂交组合表现的预测

最初人们对杂交组合的预测是由自交系本身出

发的。但大量的研究表明,由于非加性基因效应的存在,自交系自身表现与测交组合表现的相关性并不高^[38]。分子标记技术产生以后,人们陆续开展利用自交系分子标记表征的多样性进行杂交组合表现的预测,但研究发现杂交种的表现与亲本分子标记多样性间的相关性仍比较低,预测的效果并不理想。Bernardo认为,标记的染色体覆盖度不足、非性状关联标记的随机分布以及杂交种间显性效应水平的不同可能是造成这种相关性较低的主要原因^[39]。1994年,Bernardo将动物育种常用的最优线性无偏预测与RFLP标记结合,开创了利用GBLUP进行玉米杂交组合预测的先河。研究发现,利用RFLP(restriction fragment length polymorphism, RFLP)标记计算共祖系数的预测效果要略优于利用系谱计算共祖系数的预测,验证了采用GBLUP方法进行杂交组合预测的可行性^[40]。尽管全基因组选择的概念和应用一般被认为是从2001年开始的,但是Bernardo的研究实际上已经开始了全基因组选择在玉米杂交组合预测上的应用,因此这一研究也被称为作物全基因组选择研究上里程碑式的工作。GUO等采用261个多态SSR标记预测了来自玉米品种Yuyu22重组自交系组成的永久F₂群体的杂种优势,研究表明,育种值和全基因组预测模型相结合可获得最高的预测精度,预测出表现超过原始品种的单交组合114个^[41]。

近年来,基于SNP标记预测DH系测交组合的表现逐渐成为玉米全基因组选择研究和应用的热点^[42]。Albrecht等采用1152个SNP标记对来自36个杂交组合的1380个DH系组成测交组合进行了预测,使用最佳线性无偏预测和分层交叉验证来进行模型构建,并比较了自交系间亲缘关系建模和基于基因组的相似性系数建模预测模型的性能,研究表明,相似系数并不影响预测精度,且当建模群体与预测群体相关程度较高时,对产量这样复杂的性状,预测精度也很高,达到了0.72~0.74,该研究被认为是第1个玉米高世代育种群体进行测交组合预测的研究^[43]。Zhang等进行了3个玉米测交群体杂交种表现、一般配合力和特殊配合力的全基因组预测,分子标记为基于DARt-seq平台的SNP标记,结果表明,当将自交系和测验种的加性效应包括到模型中时,杂交种表现的预测精度为0.59~0.81;当将加性和非加性效应都包括到模型中时,预测精度为0.64~0.86^[44]。Luo等用34个优良玉米自交系,以部分双列杂交设计产生的285个玉米杂交组合并进行了多环境表型鉴定,同时采用45K玉米SNP芯片对

亲本自交系进行基因分型,构建了适宜我国黄淮海夏播区(SUS)和东华北春播区(SPS)的玉米产量基因组预测模型,未经测试的杂交组合表现的预测精度分别达到0.51和0.46,产量相关性的预测精度达到了0.8以上^[45]。

高质量的玉米杂交种历史数据构成了预测模型构建的重要数据来源。Technow等收集了德国霍恩海姆大学玉米项目组1999–2012年14年间1254个杂交种的基因型和表型数据,这些杂交种表型鉴定遍布德国南部20个试验点共计131个环境,通过研究亲本基因组特性(包括等位基因频率、连锁不平衡状态)和杂交种全基因组选择预测的前景,发现采用GBLUP和贝叶斯B预测方法,产量的交叉验证预测精度在0.75~0.92,含水量的预测精度在0.59~0.95。如果亲本都包含在训练群体中,则对未经测试的杂交种的预测准确率最高;如果亲本都未参与任何训练集杂交种组配,则预测准确率最低,因此,优化训练群体亲本自交系和杂交种数量可以进一步提高预测精度,认为全基因组选择实现了杂交育种的方式由关注亲本系在测试杂交中的表现向关注测试杂交种的表现进行转变^[46]。Schrage等收集了一个公共玉米项目组1970个杂交种19年的表型数据库,采用50K SNP芯片进行基因分型,通过全基因组选择模型的构建研究发现,GBLUP模型的预测结果要优于基于系谱预测的结果^[47]。先正达公司举办了一场基于自交系和测验种不同环境下的历史产量数据来预测潜在杂交组合表现的挑战赛,总体数据是593个自交系和496个测验种组配的294128个杂交种的数据,训练群体为10919个杂交种,这些杂交种植于2016–2018年的280个地点,同时考虑年份和天气因素形成杂交种-地点-天气组合数据155765个,再加上部分杂交种的多重复数据合计产量记录199476个,占有杂交种-地点-天气组合数据点24706752(593×496×280×3)的0.08%。由于试验涉及企业的技术秘密和知识产权,最终的预测精度等结果并未公布。Ansarifar等采用的整合随机森林加基因型环境互作检测的模型预测精度最高,相对均方根误差为0.0869,最终赢得了挑战赛^[48]。

4 玉米全基因组选择应用思考与展望

当前,世界玉米育种已进入4.0阶段,全基因组选择育种技术已经成为国际商业化育种体系中的通用技术底盘。育种的竞争很大程度上是规模的竞争,全基因组选择技术通过预测和淘汰可以实现同

样土地上测试更多的品种,在同等条件下扩大了育种的规模。通过种质改良和品种选育过程中预测的实施很大程度地减少了传统育种的盲目性,加快育种进程,提高了效率,最终实现了规模和效益的统一。全基因组选择改变了传统表型鉴定的作用并对表型鉴定的精准性提出了更高的要求,让育种重新回归到数量遗传学和统计学,推动现代育种向精准化和高效化方向发展。

4.1 全基因组选择对数量遗传学和生物统计与实验设计提出更高的要求

数量遗传学、生物统计与试验设计是常规育种的技术基础,通过预先的试验设计育种家实现了作物改良和新品种筛选,全基因组选择技术的应用,对这些传统技术提出了新的要求。传统的数量遗传学重视群体性状的描述,通常以群体的均值和方差作为选择的重要指标。得益于基因分型技术的进步和成本的下降,全基因组选择对群体中的个体性状表征成为可能。数量遗传学中的经典假设在植物育种中没有得到满足,特别是在玉米育种上所应用的基础群体并非随机交配群体。Bernardo认为,随着基因组学和计算机科学的进步,应该重新思考和评估植物育种中的数量遗传学,植物育种将受益于借鉴其他学科中有用的方法。如将可靠性作为衡量遗传与非遗传效应影响的一种新方法,以及设计育种方案的运筹学和模拟方法^[49],分子标记技术的产生和计算机运算能力的进步将重塑传统的数量遗传学,大量分子标记数据的应用使育种对运算能力的要求空前提高。

1994年,Bernardo开发 Genomic Best Linear Unbiased Prediction(GBLUP)的方法是在具有64兆内存的奔腾90机器上实现的。全基因组选择特别是贝叶斯、机器学习等模型的使用需要迭代计算来优化一些参数,所需要的运算能力相较于奔腾90已经发生了翻天覆地的变化。在训练群体的优化上,选择哪些材料来构建训练群体,如何实现加性、显性和上位性效应及其与环境互作效应的分解是从试验设计阶段就应该认真考虑的问题。传统的生物统计在育种上的应用重点已经从推断析因向预测未来的方向转变,预测因子从基因型扩展到太阳辐射、降水、地理纬度、年份、生物逆境等诸多因子,通过生物信息、环境信息等因素的应用传统育种已经进入到数字化、信息化和智能化阶段。现代育种要求育种家在掌握遗传学和传统统计学的基础上,应更加注重生物信息学、计算机科学甚至是机器学习等知识和技能的储备和持续更新。

4.2 全基因组选择使表型鉴定的作用发生了变化并且越来越重要

在传统育种中表型鉴定的作用是筛选出优秀的杂交种供生产使用,在初级鉴定中常常只对那些表现优秀的组合进行测产和升级,对那些肉眼观察具有明显缺陷的组合并不进行测产而直接淘汰。全基因组选择训练群体构建需要有相对完整的表型数据,至少训练群体中每一个个体的表型应该是完整的。从Technow、Schrage等的研究不难看出,这些研究之所以能够收集不同年份的玉米育种历史数据进行训练群体构建,一个重要的前提是这些杂交种表型数据是相对完整的。实际上,近年来带有测产功能的玉米小区收割机的应用极大地促进了表型数据的积累,即使是初级鉴定试验也可以完成至少一年的产量和含水量等表型数据的采集。这些数据的采集和保存将为全基因组选择训练群体的构建奠定重要基础。全基因组选择对表型鉴定的另一个要求就是准确性,传统育种表型鉴定随着试验级别的提升入选组合数会越来越少,性状鉴定上更强调的定性的筛选,区分“优”与“差”。如在抗病性上可能感病的品种都会淘汰掉,而全基因组选择则需要对这些性状进行量化,并要求表型具有丰富的多样性。另外,对于抽丝期、散粉期等性状随着鉴定组合数量的增大性状调查的人为误差往往会增大,严重影响数据质量。高通量的表型组学技术的应用将为全基因组选择表型数据的采集提供新的手段。

4.3 基因分型成本是当前全基因组选择技术应用的重要限制性因素之一

基因分型的成本下降是驱动育种企业应用全基因组选择的重要动力。尽管育种家公式中并没有成本的考虑,但是研发投入是比较不同育种方法的效率必须要考虑的因素^[50]。相对于常规育种,全基因组选择多了1个基因分型的成本,如果总投资相同,基因分型的成本必须通过其他育种方案来补偿。全基因组选择通过预测的方法可以通过仅对部分材料进行表型测试来降低成本。随着人力和土地成本的上涨,基因分型的成本已经逐渐低于表型鉴定的成本。此外,近年来,一些简化测序技术如GBS-SNP(genotyping-by-sequencing-SNP)、Dart-SNP(Diversity Arrays Technology)、靶向测序基因型检测(genotyping by target sequencing, GBTS)等技术被陆续开发出来应用于全基因组选择研究。当前,检测1个玉米样品1000个SNP位点的费用已经低于1个杂交组合进行组配和一年三点的鉴定的费用。尽管如此,在育种的哪个环节采用全基因组选择技术,以及如

何优化建模群体提高全基因组选择预测精度是有效降低全基因组选择成本需要认真考虑和研究的问题。对玉米育种而言,对DH系进行基因分型,再通过DH系的基因型两两结合推导出杂交种的基因型,既可以实现对DH系自身性状的预测,又可以实现对可能组配杂交组合的预测,因此结合玉米DH系开展全基因组选择工作是提高玉米育种效率的一个重要工作内容。

4.4 全基因组选择模型具有种质和生态区域特殊性

提高作物改良的速度对于实现持续的粮食生产和不断满足增加人口的需要是必不可少的。全基因组选择(GS)作为重要的育种工具,已成功应用于动物育种,并正在纳入植物育种。GS有望通过快速选择优良基因型来加快育种周期。全基因组选择是生物育种的重要内容之一,相对于基因编辑、转基因和生物合成技术,全基因组选择是优势传统育种单位最容易掌握和转换的赛道。全基因组选择的预测模型以种质资源和特殊的育种目标为出发点,目前来看,即使针对同一个性状温带玉米与热带玉米,甚至是不同熟期的玉米杂交组合预测都难以应用一个统一的模型来实现。那些传统育种优势单位往往具有种质资源和表型鉴定的优势,对当地的育种目标、气候特点、生产条件等具有清晰的理解和深刻的认识,可以进一步通过提高表型鉴定的准确性积累大量的表型数据。由于全基因组选择技术的应用和发展,基因分型工作和表型鉴定工作的分工将越来越专业化和细化,考虑到我国在基因分型设备和技术上的优势,基因分型的成本将会不断降低。在当前阶段,常规育种单位可以通过基因分型工作外包的形式积累种质的基因型,整理那些数据质量高的育种资源开展全基因组研究,不断积累经验、完善预测模型,在基因分型工作中应该注意采用技术兼容性。

参考文献:

- [1] DESTA Z A, ORTIZ R. Genomic selection: genome-wide prediction in plant improvement[J]. Trends in Plant Science, 2014, 19(9): 592–601.
- [2] AL-KHAYRI J M, JAIN S M, JOHNSON D V. Advances in plant breeding strategies: breeding, biotechnology and molecular tools[M]. Springer International Publishing, 2015: 69–91.
- [3] 景海春, 田志喜, 种康, 等. 分子设计育种的科技问题及其展望概论[J]. 中国科学: 生命科学, 2021(51): 1356–1365.
JING H C, TIAN Z X, ZHONG K. Progress and perspective of molecular design breeding[J]. Scientia Sinica Vitae, 2021(51): 1356–1365. (in Chinese)
- [4] GEPTS P. A comparison between crop domestication, classical plant breeding, and genetic engineering[J]. Crop Science, 2002, 42(6): 1780–1790.
- [5] 王向峰, 才卓. 中国种业科技创新的智能时代——“玉米育种4.0”[J]. 玉米科学, 2019, 27(1): 1–9.
WANG X F, CAI Z. Era of maize breeding 4.0[J]. Journal of Maize Sciences, 2019, 27(1): 1–9. (in Chinese)
- [6] RAY D K, RAMANKUTTY N, MUELLER N D, et al. Recent patterns of crop yield growth and stagnation[J]. Nat Commun, 2012(3): 1293.
- [7] 徐云碧, 杨泉女, 郑洪建, 等. 靶向测序基因型检测(GBTS)技术及其应用[J]. 中国农业科学, 2020, 53(15): 2983–3004.
XU Y B, YANG Q N, ZHENG H J. Genotyping by target sequencing (GBTS) and its applications[J]. Scientia Agricultura Sinica, 2020, 53(15): 2983–3004. (in Chinese)
- [8] GAO H, GADLAGE M J, LAFITTE H R, et al. Superior field performance of waxy corn engineered using CRISPR – Cas9[J]. Nat Biotechnol, 2020, 38(5): 579–581.
- [9] BERNARDO R, YU J. Prospects for genomewide selection for quantitative traits in maize[J]. Crop Science, 2007, 47(3): 1082–1090.
- [10] MEUWISSEN T. Genomic selection: marker assisted selection on a genome wide scale[J]. Journal of Animal Breeding & Genetics, 2007, 124(6): 321–322.
- [11] MEUWISSEN T, HAYES B J, GODDARD M E M. Prediction of total genetic value using genome-wide dense marker maps[J]. Genetics, 2001, 157(4): 1819–1829.
- [12] SARINELLI J M, MURPHY J P, TYAGI P, et al. Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel[J]. Theoretical and Applied Genetics, 2019, 132(4): 1247–1261.
- [13] ZHANG A, WANG H, BEYENE Y, et al. Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations [J]. Frontiers in Plant Science, 2017, 8: 1916.
- [14] ASORO F G, NEWELL M A, BEAVIS W D, et al. Accuracy and training population design for genomic selection on quantitative traits in elite north American oats[J]. The Plant Genome, 2011, 4(2): 132–144.
- [15] CAO S, SONG J, YUAN Y, et al. Genomic prediction of resistance to tar spot complex of maize in multiple populations using Genotyping-by-sequencing SNPs[J]. Frontiers in Plant Science, 2021, 12: 1438.
- [16] RIEDELSHEIMER C, ENDELMAN J B, STANGE M, et al. Genomic predictability of interconnected biparental maize populations[J]. Genetics, 2013, 194(2): 493.
- [17] ZHAO Y, GOWDA M, LONGIN F H, et al. Impact of selective genotyping in the training population on accuracy and bias of genomic selection[J]. Theoretical & Applied Genetics, 2012, 125(4): 707–713.
- [18] TECHNOW F, BURGER A, MELCHINGER A E. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups[J]. G3: Genes, Genomes, Genetics, 2013, 3(2): 197–203.
- [19] AKDEMIR D, SANCHEZ J I, JANNINK J L. Optimization of genomic selection training populations with a genetic algorithm[J].

- Genet Sel Evol, 2015, 47(1): 38.
- [20] KADAM D C, RODRIGUEZ O R, LORENZ A J. Optimization of training sets for genomic prediction of early-stage single crosses in maize[J]. Theoretical and Applied Genetics, 2021, 134(2): 687–699.
- [21] LOPEZ–CRUZ M, DREISIGACKER S, CRESPO–HERRERA L, et al. Sparse kernel models provide optimization of training set design for genomic prediction in multiyear wheat breeding data[J]. Plant Genome, 2022, 15(4): e20254.
- [22] ISIDRO J, JANNINK J L, AKDEMIR D, et al. Training set optimization under population structure in genomic selection[J]. Theor Appl Genet, 2015, 128(1): 145–58.
- [23] LORENZANA R E, BERNARDO R. Accuracy of genotypic value predictions for marker–based selection in biparental plant populations[J]. Theoretical & Applied Genetics, 2009, 120(1): 151–161.
- [24] E SOUSA M B, GALLI G, LYRA D H, et al. Increasing accuracy and reducing costs of genomic prediction by marker selection[J]. Euphytica, 2019, 215(2): 18.
- [25] HENDERSON C R. Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values [J]. Biometrics, 1976, 32(1): 69–83.
- [26] BERNARDO R. Prediction of maize single–cross performance using RFLPs and information from related hybrids[J]. Crop Science, 1994, 34(1): 20–25.
- [27] VANRADEN P M. Efficient methods to compute genomic predictions[J]. Journal of Dairy Science, 2008, 91(11): 4414–4423.
- [28] MCGOWAN M, WANG J, DONG H, et al. Ideas in genomic selection with the potential to transform plant molecular breeding: A review[J]. Plant Breeding Reviews, 2022(45): 45.
- [29] SHAMSHAD M, SHARMA A. The usage of genomic selection strategy in plant breeding[J/OL]. 2018. DOI:10.5772/intechopen.76247.
- [30] WANG X, XU Y, HU Z H, et al. Genomic selection methods for crop improvement: Current status and prospects[J/OL]. Crop Journal, 2018. DOI:CNKI:SUN:CROP.0.2018–04–003.
- [31] WANG K, ABID M A, RASHEED A, et al. DNNP, a deep neural network–based method for genomic prediction using multi–omics data in plants[J]. Mol Plant, 2023, 16(1): 279–293.
- [32] CROSSA J, P PÉREZ–RODRÍGUEZ, CUEVAS J, et al. Genomic selection in plant breeding: methods, models, and perspectives[J]. Trends in Plant Science, 2017, 8: 961–975.
- [33] PIYASATIAN N, FERNANDO R L, DEKKERS J C. Genomic selection for marker–assisted improvement in line crosses[J]. Theor Appl Genet, 2007, 115(5): 665–674.
- [34] ZHANG X, PÉREZ–RODRÍGUEZ P, BURGUEÑO J, et al. Rapid cycling genomic selection in a multiparental tropical maize population[J]. G3(Bethesda), 2017, 7(7): 2315–2326.
- [35] WANG N, WANG H, ZHANG A, et al. Genomic prediction across years in a maize doubled haploid breeding program to accelerate early–stage test cross testing[J]. Theor Appl Genet, 2020, 133(10): 2869–2879.
- [36] GUO Z, TUCKER D M, LU J, et al. Evaluation of genome–wide selection efficiency in maize nested association mapping populations [J]. Theoretical & Applied Genetics, 2012, 124(2): 261–275.
- [37] GOWDA M, DAS B, MAKUMBI D, et al. Genome–wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm[J]. Theor Appl Genet, 2015, 128(10): 1957–68.
- [38] HALLAUER A R, LOPEZ–PEREZ E. Comparisons among testers for evaluating lines of corn[J]. Proc. Corn Sorghum Industry Res Conf, 1979, 34: 57–75.
- [39] BERNARDO R. Relationship between single–cross performance and molecular marker heterozygosity[J]. Theor Appl Genet, 1992, 83(5): 628–634.
- [40] BERNARDO R. Prediction of maize single–cross performance using RFLPs and information from related hybrids[J/OL]. Crop Science, 1994, 34(1). DOI:10.2135/cropsci1994.0011183X003400010-0003x.
- [41] GUO T, LI H, YAN J, et al. Performance prediction of F1 hybrids between recombinant inbred lines derived from two elite maize inbred lines[J]. Theoretical & Applied Genetics, 2013, 126(1): 189–201.
- [42] COOPER M, MESSINA C D, PODLICH D, et al. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction[J]. Crop & Pasture Science, 2014, 65(4): 311.
- [43] ALBRECHT T, WIMMER V, AUINGER H J, et al. Genome–based prediction of testcross values in maize[J]. Theor Appl Genet, 2011, 23(2): 339–350.
- [44] ZHANG A, PAULINO PÉREZ–RODRÍGUEZ, VICENTE F S, et al. Genomic prediction of the performance of hybrids and the combining abilities for line by tester trials in maize[J]. The Crop Journal, 2021, 10: 109–116.
- [45] LUO P, WANG H, NI Z H, et al. Genomic prediction of yield performance among single–cross hybrids using a partial diallel cross design[J]. The Crop Journal, 2023, 11(6): 1884–1892.
- [46] TECHNOW F, SCHRAG T A, SCHIPPRACK W, et al. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize[J]. Genetics, 2014, 197(4): 1343–1355.
- [47] SCHRAG T A, SCHRAG T A, SCHIPPRACK W, et al. Across–years prediction of hybrid performance in maize using genomics[J]. Theoretical and Applied Genetics, 2019, 132(4): 933–946.
- [48] ANSARIFAR J, AKHAVIZADEGAN F, WANG L. Performance prediction of crosses in plant breeding through genotype by environment interactions[J]. Sci. Rep., 2020, 10(1): 11533.
- [49] BERNARDO R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE[J/OL]. Heredity, 2020, DOI:10.1038/s41437–020–0312–1.
- [50] FUGERAY–SCARBEL A, BEN–SADOUN S, BOUCHET S, et al. Analyzing the economic effectiveness of genomic selection relative to conventional breeding approaches[J]. Methods Mol Biol, 2022, 2467: 619–644.

(责任编辑:朴红梅)