DOI: 10.13376/j.cbls/2018127
文章编号：1004-0374(2018)10-1051-09

徐辰武，扬州大学二级岗位教授、博士生导师，《生物统计与试验设计》国家精品课程和国家精品资源共享课程负责人，青岛农业大学和河南科技大学兼职教授。兼任 *Heredity* 和 *BMC Evolutionary Biology* 责任编委，江苏省遗传学会理事。主要研究领域为生物统计、应用数量遗传、统计基因组学和生物信息学。先后主持承担国家重点研发计划课题 1 项、国家“973”项目课题 2 项、国家“863”计划子课题 1 项和国家自然科学基金项目 7 项。先后发表学术论文 160 余篇，主编或副主编教材 3 部，参编英文专著 1 部、教材 2 部。作为第一作者或通讯作者，在国际权威刊物 *Trends in Plant Science*、*Plant Physiology*、*New Phytologist*、*Genetics*、*Heredity* 等发表 SCI 论文 50 余篇。先后获得教育部科技进步奖二等奖 (1998 年度，排名第 3) 和教育部自然科学奖二等奖 (2011 年度，排名第 1) 各 1 项。先后入选江苏省高校“青蓝工程”第二期计划省级中青年学术带头人 (2002 年)、教育部“新世纪优秀人才支持计划”(2005 年)、江苏省“333 高层次人才培养工程”中青年科学技术带头人 (2007 年)、江苏省高校“青蓝工程”科技创新团队带头人 (2012 年) 以及农业部全国农业科研杰出人才和创新团队带头人 (2015 年)。

绿色性状的基因定位和基因组选择的多变量方法

徐　扬，王　欣，徐辰武*
(扬州大学，江苏省作物遗传生理重点实验室/植物功能基因组学教育部重点实验室/江苏省作物基因组学和分子育种重点实验室，江苏省粮食作物现代产业技术协同创新中心，扬州 225009)

摘　要：绿色性状的遗传改良为绿色超级稻的培育奠定了坚实的基础。绿色性状，如高产、抗病、抗逆、氮磷高效利用等，大多是受多基因控制的复杂性状。关联分析和基因组选择是对植物复杂数量性状进行遗传解析和改良的重要方法。在绿色超级稻的育种实践中，需要同时改良多个绿色性状。然而，目前关联分析和基因组选择方法大多仍专注于对单个性状的分析，忽略了性状间的相关性。现分别提出关联分析和基因组选择的多性状方法，两者充分利用了性状之间的遗传相关和环境相关信息；而模拟研究和实证研究均表明，多性状方法能有效提高基因定位和表型预测的准确性，为绿色性状的遗传改良提供重要技术支撑。

关键词：绿色性状；多变量；关联分析；基因组选择；选择指数
中图分类号：S511；S513　　**文献标志码**：A

Multivariate approaches of gene mapping and genomic selection for green traits

XU Yang, WANG Xin, XU Chen-Wu*
(Jiangsu Key Laboratory of Crop Genetics and Physiology/ Key Laboratory of Plant Functional Genomics of the Ministry

收稿日期：2018-09-27
基金项目：国家高技术研究发展计划(“863”计划)(2014AA10A601)；杂交水稻国家重点实验室(武汉大学)开放课题基金(KF201701)；中国博士后科学基金(2018M630613)
***通信作者**：E-mail: cwxu@yzu.edu.cn

of Education/ Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding, Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou 225009, China)

Abstract: The genetic improvement of green traits has laid a solid foundation for the development of Green Super Rice. Green traits, such as high yield, disease resistance, stress resistance, nitrogen and phosphorus use efficiency, are mostly complex traits controlled by multiple genes. Association analysis and genomic selection are important methods for genetic analysis and improvement of complex quantitative traits in plants. In the breeding of Green Super Rice, it is necessary to improve many green traits simultaneously. However, most of the current association analysis and genomic selection methods are still focused on a single trait, ignoring the correlation between traits. Therefore, we proposed the multi-trait methods of association analysis and genomic selection. Both simulation and empirical studies show that the multi-trait methods effectively increase the accuracy of gene mapping and phenotype prediction. The proposed methods provide an important technical support for the genetic improvement of green traits.

Key words: green traits; multivariate; association analysis; genomic selection; selection index

水稻 (*Oryza sativa*) 为全世界超过 50% 的人口提供了稳定的食物来源，水稻增产对保障我国乃至全球粮食安全做出卓越的贡献。20 世纪 60、70 年代矮秆品种和杂交稻的培育和应用，使我国水稻产量实现了两次飞跃；但大量高产品种的培育和大面积推广，引发化肥、农药、水以及劳动力的投入大幅增长，导致水稻生产与资源环境的矛盾激增 [1]。针对我国水稻生产中病虫害严重、农药化肥使用过量、水资源短缺、产量持续徘徊不前等问题，张启发院士提出了培育“少打农药、少施化肥、节水抗旱、优质高产”的绿色超级稻育种目标 [2]。绿色超级稻的绿色性状，如抗病、抗旱、抗逆、氮磷高效利用等，大多是受多基因控制的复杂性状，受环境影响较大，常规选育的效果不太理想。随着高通量测序技术的发展，关联分析和基因组选择已成为对植物复杂数量性状进行遗传解析和改良的重要方法。

在关联分析和基因组选择模型中，标记的数量经常远远超过样本量，会产生自由度不足和多重共线性等问题。为了解决这些问题，前人发展出了许多统计方法 [3-4]。然而，目前的关联分析和基因组选择方法仍专注于对单一性状的分析，忽略了性状之间的遗传相关和环境相关。在绿色超级稻的育种工作中，需要培育同时具有高产优质、节水抗旱、抗病抗逆等绿色性状的品种，而不是追求单一目标性状的改良。为了满足绿色超级稻的育种需求，本文提出了关联分析和基因组选择的多性状方法。

1 关联分析的多性状方法

关联分析，是一种基于连锁不平衡，鉴定某一群体内目标性状与遗传标记或候选基因关系的分析方法。与传统的连锁作图相比，关联分析利用的是自然群体在长期进化中所积累的重组信息，解析精度较高，可实现对数量性状位点 (quantitative trait locus, QTL) 的精细定位，是解析数量性状的强有力的工具。自 2001 年以来，应用关联分析方法发掘植物数量性状基因就备受关注 [5]，大量的统计方法以及软件包应运而生。

目前应用最广泛的是基于单位点扫描的混合线性模型方法，最早由 Yu 等 [6] 提出。该方法能够有效控制群体结构和多基因背景，从而降低关联分析的假阳性。为了处理较大样本并且提高运算效率，一系列基于混合线性模型的改进方法相继被提出，如 EMMA[7]、GEMMA[8] 等。为了获得更快的运算速度，一些近似算法也被提出，如 EMMAX [9]、P3D [10]、CMLM [10]。虽然已被广泛应用，但单位点扫描方法仍存在一定的局限性，如要进行多重测验矫正，不能同时考虑所有位点信息等。因此，多位点的关联分析方法逐渐受到关注。在进行多位点关联分析时，样本量远大于标记数目 (n)，许多方法被提出处理此类“大 p 小 n”的问题，如逐步线性回归方法、岭回归分析、LASSO、主成分回归、偏最小二乘法 (partial least squares, PLS) 及贝叶斯方法等。为了更有效地控制群体结构，基于混合线性模型的多位点方法也逐渐被提出，如 MLMM[11]、LMM-Lasso[12] 和 BSLMM[13] 等。

然而，上述的关联分析方法仍停留在单个目标性状上，而在绿色超级稻的育种工作中，需要培育出同时具备多个绿色性状的新品种。多性状关联分析既可以同时对多个目标性状进行联合分析，实现多个目标性状的协调发展，也可利用性状之间的遗

传相关和环境相关信息，提高关联分析的功效。要想获取绿色性状基因，必须准确鉴定出与目标性状关联的基因变异位点，并精确估计出该位点对目标性状的遗传效应和贡献率。因此，本文拟介绍一种基于偏最小二乘的多性状联合关联分析方法，并利用模拟实验验证该方法的可行性。

1.1　分析方法

本文采用以下策略进行基于偏最小二乘的多性状联合关联分析，步骤如下。(1) 采用非线性迭代偏最小二乘法求解多变量模型的回归系数。(2) 计算变量投影重要性 (variable importance in projection, VIP)，分别计算 M 个自变量的 VIP 指标并从大到小排序，即：$VIP_{(1)} \geqslant VIP_{(2)} \geqslant ... VIP_{(M)}$。(3) 根据回归系数选择对应每个表型性状的重要变量 S_k^v，然后根据 VIP 指标同样选择重要变量子集 S_{VIP}，均选择前 $\frac{n}{2}-1$ (n 为偶) 或 $\frac{n\text{-}1}{2}$ (n 为奇) 个效应变量，从而综合对各个性状都相对重要的自变量集合。整合两类重要效应变量集合，生成对应每个表型性状的重要效应变量的预选择 $S_k = S_k^V \cup S_{VIP}$。其中 S_k 中的变量个数小于 n–1，经过变量压缩，线性模型由过饱和变为普通的线性模型。(4) 采用双向筛选的逐步回归进行重要效应变量的选择，并用 BIC 信息准则作为增加或删除变量的标准。最终根据最优模型中的效应变量，以普通线性回归模型计算效应大小及相应的 p 值，具体算法参见 [14]。

1.2　模拟研究

为了探究该方法的可行性，本团队在不同的条件下比较了该方法和两个单性状的多位点方法 LASSO[15] 和基于偏最小二乘的多位点关联分析 (PLS-based MLAS)[16] 的检验功效。本文以两个相关数量性状的联合分析为例进行。假设基因组中存在 10 000 个位点，每个位点均有 2 个等位基因，其中 10 个位点具有遗传效应，分别为 Q1~Q10。Q1~Q4 仅控制性状 1 的表达，Q7~Q10 仅控制性状 2 的表达，Q5 和 Q6 设为一因多效基因；为了更加符合实际情况，对不同位点设置不同水平的多态信息含量 (PIC)。10 个位点的效应值及 PIC 设置如表 1 所示。同时，考察样本容量 n 和候选基因遗传力 h^2 两个因素，其中样本容量设定为 3 个水平，分别为 100、300 和 500；遗传力设定 3 个水平，分别为 30%、50% 和 70%。性状 1 和性状 2 的平均值均设为 10，两个性状间的剩余误差相关系数 r_e 设定为 0.5。

结果表明，样本容量和遗传力均对检测功效具有明显影响 (图 1)。当样本容量一定时，随着遗传力的增加，检测功效逐步提高；与此类似，当遗传力一定时，随着样本容量的增加，检测功效也相应提高。从表 2 可以看出，PIC 值对其检测功效具有较大的影响：位点的 PIC 值越大，其被检测到的概率越高；PIC 值较小的位点在样本较少和遗传力较低时被检测到的概率较低，甚至检测不到。由表 2 还可以看出，效应的大小对检测功效有较大影响：

表1　模拟中各QTL的效应及PIC值

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Effect value	Trait 1	2	2	2	2	2	2	0	0	0	0
	Trait 2	0	0	0	0	1.5	–1.5	2.0	–2.0	2.5	–2.5
PIC value		0.07	0.22	0.3	0.35	0.37	0.37	0.37	0.37	0.37	0.37

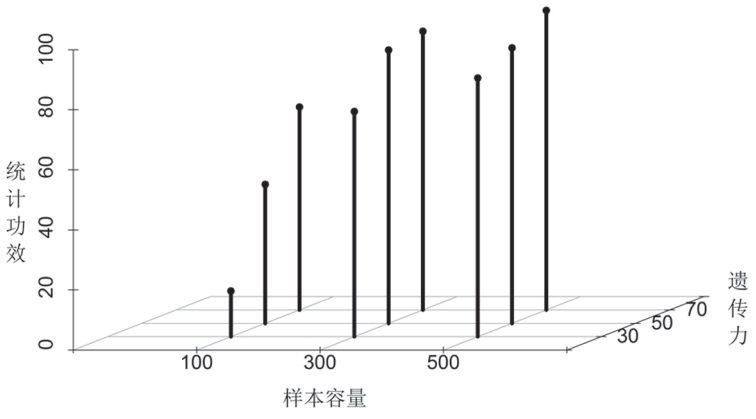

图1　多性状联合关联分析时样本容量和遗传力(%)对QTL统计功效(%)的影响

在 PIC 保持恒定的情况下，QTL 的效应值的绝对值愈大，其被检测到的概率愈高，然而效应真值的正负对检测功效影响不大。样本容量和遗传力对效应估计的准确度和精确度均具有一定影响，随着样本容量和遗传力的增加，位点效应的估计值越趋近于真值。

采用多性状联合分析、LASSO 以及 PLS-based MLAS 方法获得的统计功效均列于表 2，从中可以看出，多性状联合分析时，一因多效位点的统计功效与单性状分析相比有了大幅度的提高。此外，大多数情况下，即使一个位点仅控制一个性状，联合分析仍具有一定优势。但是对于 PIC 值极低的位点，联合分析似乎没有优势。

2 基因组选择的多性状方法

基因组选择 (genomic selection, GS) 能够估计全基因组上所有标记的遗传效应，进而实现对品种更加可靠的选择 [17-18]。特别是在水稻等作物的杂种育种中，杂交种的基因型可以由亲本基因型进行推断，GS 的优势更加突出 [19-20]。然而传统的 GS 方法专注于单一环境下单个性状的预测，忽视了性状之间的遗传相关和不同环境之间的关联，而相互关联的性状或者不同环境下的作物品种之间可能拥有共同的遗传或生物学基础 [21]，从而使多性状联合分析拥有更高的预测能力 [22]。在绿色超级稻的育种实践中，一些性状可能难以度量或者观测的成本过于昂贵 (小区产量、抗逆性状、肥料的吸收利用和根系相关性状等)，多性状联合分析提供了一种新的策略，即利用更易鉴定的相关表型预测难以获取的重要性状 [23]。

此外，水稻等作物的产量遗传力较低，所以基因组预测的效果不佳。如 Xu 等 [24] 利用 GBLUP 模型预测了水稻杂交种的 4 个性状，其中产量的预测能力最低，只有 0.13。另一方面，育种家真正关心的是水稻等作物多个农艺性状的综合表现，千粒重等农艺性状的遗传力较高，相应的基因组预测精度也较高。绿色超级稻的育种目标有多个，要求“少打农药、少施化肥、节水抗旱、优质高产”等，实际的 GS 工作中可以利用一些较高遗传力的性状构建选择指数，实现对品种的综合预测。而一般的多性状预测方法往往是基于多变量模型的建立，未能与选择指数理论 [25] 相结合，从而无法利用与目标性状相关的多个其他性状构建选择指数，对作物进行更加全面的选择。

选择指数是农作物多目标育种选择的常用方法，其效率高于对所有性状的逐一选择，能够被用来同时改良多个数量性状。GS 的快速发展为选择指数带来了新的前景。Dekkers [26] 利用全基因组高密度标记估计的育种值构建选择指数；Jesus Ceron-Rojas 等 [27] 利用 GS 方法和选择指数对多个性状同时进行选择；Schulthess 等 [28] 使用黑麦中的两个性状建立选择指数，并将其看做单一性状用 GS 法进行预测；Lyra 等 [29] 将玉米杂交种在不同氮胁迫下的性状组合以构建选择指数，然后用 GS 方法进行预测，结果表明其方法是有效的。但是，已有的选

表2 不同处理和分析方法下10个QTL统计功效的比较(以样本容量100为例)

遗传力(%)	方法	统计功效 (%)									
		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
30	S1(PLS)	5.5	6.5	2.5	6	12	16.5				
	S2(PLS)					1.5	0	6.5	12.5	20	21
	S1(LASSO)	0	5.5	5.5	1	6.5	17.5				
	S2(LASSO)					0.5	0	5	4.5	14	17.5
	J12	0	5.5	9	9.5	33.5	30.5	7	13.5	23.5	21.5
50	S1(PLS)	8.5	7.5	10.5	45.5	32	68				
	S2(PLS)					4	1	22	31	69	61
	S1(LASSO)	4.5	6.5	12	62	37.5	54.5				
	S2(LASSO)					1.5	1.5	5	14.5	28	85.5
	J12	0	8.5	16	64.5	84.5	78.5	24	34.5	65	87.5
70	S1(PLS)	9.5	13.5	50	82	61.5	94.5				
	S2(PLS)					6.5	0	46	51	94.5	85.5
	S1(LASSO)	1	8.5	64	82	11.5	93.5				
	S2(LASSO)					8.5	0	36.5	87.5	89.5	93.5
	J12	10	17.5	72	74	93	99.5	60	71.5	92	96

择指数研究大多专注于对动植物多个目标性状的加权选择，在将选择指数和 GS 结合时，指数本身的预测精度往往成为关注的对象，而忽视了利用选择指数聚集辅助性状信息以预测目标性状的能力。在对复杂目标性状进行预测的过程中，与目标性状相关的辅助性状能够提供大量有用的信息，将选择指数和 GS 方法相结合，有助于充分利用这些信息，帮助育种家对目标性状进行更加准确的预测，从而有效开展选种和配种工作。

因此，本研究开展了多性状联合预测的 GBLUP 模型研究，并利用与目标性状相关的多个辅助性状建立选择指数，实现对水稻性状的综合选择，提高预测精度。本研究使用的水稻数据集来自武汉大学 [20]，以 115 份水稻纯系品种为父本，5 份不育系作母本，采用 NCII 交配设计配制 575 份杂交种。考察了 8 个性状，包括单株产量 (grain yield per plant, GY)、千粒重 (thousand-grain weight, TGW)、有效穗 (productive panicle number per plant, PN)、株高 (plant height, PH)、一次枝梗 (primary branch number, PB)、二次枝梗 (secondary branch number, SB)、主穗实粒数 (grain number per panicle, GN) 和穗长 (panicle length, PL)。本研究模型中使用到的水稻表型数据是两种环境下两次重复的平均值。同时，对 120 份水稻亲本进行高通量测序，共获得 3 299 150 个 SNP 标记。

2.1　基因组选择方法的比较

自从 Meuwissen 等 [30] 首次提出 GS 的概念后，大量的方法涌现，包括各种参数、半参数和非参数的方法。参数方法主要有 GBLUP、LASSO、PLS 以及贝叶斯方法，非参数方法有随机森林、SVM 和 RKHS 等。比较各种 GS 方法的异同，了解每种方法所适用的条件，以更好地服务于育种工作，是 GS 研究中的重要内容。已有一些研究对这些方法进行了比较。de los Campos 等 [31] 对参数方法的预测准确性进行了比较，发现对于大多数性状 GBLUP 都有较好的预测准确性，BayesB 对于大效应 QTL 控制的性状预测效果较好。Riedelsheimer 等 [32] 用 5 种不同预测方法对玉米 3 个农艺性状和 3 个代谢性状进行了预测，得出的结论是这些方法在预测准确性上差异不大。Heslot 等 [33] 利用了 10 种不同的参数和非参数方法对从不同物种中测量的 18 个性状进了预测准确性比较，发现 RKHS 在不同的性状和物种中都有较好的预测表现。Howard 等 [34] 利用模拟数据进行了参数方法和非参数方法的预测准确性比较，发现参数方法预测受加性效应控制的性状时效果比非参数方法要稍好一些。然而在实际水稻数据中，这些基因组预测方法的比较尚少有报道。

本课题组基于上述杂交水稻数据集，对 6 种常用的预测方法进行比较 [35]，包括 4 种参数方法 LASSO、GBLUP、BayesB 和 PLS 以及 2 种非参数方法 RKHS 和 SVM。通过对 6 种预测方法的预测准确性进行方差分析发现，不同方法的预测准确性之间存在极显著差异，而且预测准确性在方法、性状之间的互作也存在极显著差异。进一步进行多重比较发现，6 种方法中，GBLUP 方法的预测准确性最高，而 PLS 和 SVM 的预测准确性最低，其余方法的准确性介于两者之间。此外，不同的方法适用于预测不用的性状，如 GBLUP 方法对于主穗实粒数、株高和千粒重的预测准确性最高，LASSO 对于一次枝梗和二次枝梗的预测准确性最高，SVM 是穗长的最佳预测方法。尽管没有一种方法能适用于所有性状，但 GBLUP 方法在所有性状的预测中表现最稳健。因此，后续的多性状的 GS 研究将基于 GBLUP 模型开展。

2.2　基因组选择模型的多变量方法

传统的 GS 方法主要用于单一环境下单个性状的预测。虽然迄今为止已经发表的大部分 GS 研究结果使用的都是单变量模型，但是一些研究建议使用多变量模型 [28, 36]。本课题组将单性状加 - 显模型扩展为多性状加 - 显模型，并且把非遗传的残差项分解为完全相关的部分和完全独立的部分，得到了多性状加 - 显模型 MV-AD 以及包含共同环境效应的多性状加 - 显模型 MV-ADE。此外，本研究利用辅助变量构造的关系矩阵开发了一种高效的多变量模型 MV-ADV[20]。各种模型的预测能力使用交叉验证的方法衡量。

与单性状模型相比，多性状模型能够利用到群体更多的性状信息。以 GY 和 TGW 分别为目标性状，使用 MV-AD、MV-ADE 和 MV-ADV 对水稻杂交种进行的两性状联合预测结果见表 3。对于 GY，MV-ADE 和 MV-ADV 的预测能力明显优于 MV-AD 和 UV-AD。虽然 MV-ADV 相较 MV-ADE 的优势绝对值较小，但是成对比较表明 MV-ADV 的优势是显著的。然而无论以哪个性状为辅，都无法提高对 TGW 的预测能力。表明多性状模型能够提高对 GY 等低遗传力性状的预测能力，对于 TGW 等高遗传力的性状，多性状模型的预测效果并无明显改进，此时只需应用单性状模型进行预测即可。

本研究不仅考察了两性状联合预测的情况，还

将所有的 8 个性状放在一起进行了联合预测。图 2 对比了单性状 UV-AD、两性状 MV-ADV 和八性状 MV-ADV 的预测能力。其中八性状联合的预测能力显著高于单性状预测和两性状预测，特别是 GY 和 PN 的预测能力提高最大，TGW 和 PH 的预测能力提高较少，再次说明多性状联合分析的策略更适用于低遗传力的性状。两性状的平均预测能力比单性状预测要高 10.2%，而八性状的平均预测能力比单性状预测要高 50.6%。很明显，八性状的预测能力相对于两性状和单性状预测获得了更为显著的增加，说明结合更多的辅助性状信息，可以较大程度提高目标性状的预测能力。

2.3 综合性状的基因组选择方法

本课题组以杂交种的 SNP 数据为基础，针对基于选择指数的多性状 GS 方法开展了大量模拟研究，随机模拟了具有 100 个 QTL 的多个性状。其中遗传力为 0.3 的表型性状 T3 为待预测的目标性状，遗传力为 0.7 和 0.4 的多个性状为辅助性状，用以构建选择指数。本研究同时以模拟表型和实际表型数据为例说明多性状选择指数 GS 方法的应用，并对其效果进行评估。

本研究针对基于选择指数的多性状 GS 方法，选定某一目标性状 (如 GY)，对与之相关的 *s* 个辅助性状构建了选择指数 *SI*。选择指数能够综合多个辅助性状与目标性状的相关遗传信息，具有较强的预测价值，能够实现对水稻多个性状的综合选择。此外，如果预测的目标是产量等单一性状，可以将选择指数的预测结果与传统 GS 的预测结果相结合，

表3 杂交种多性状和单性状的平均预测能力

目标性状	辅助性状与多性状预测结果										UV-AD
	方法	GY	TGW	PN	PH	PB	SB	GN	PL	平均值	
GY	MV-AD		0.1422	0.2252	0.1538	0.1724	0.1915	0.2301	0.1486	0.1805	0.1558
	MV-ADE		0.1441	0.5779	0.1575	0.2110	0.2737	0.4494	0.1702	0.2834	
	MV-ADV		0.1670	0.5761	0.1544	0.2163	0.2744	0.4441	0.1903	0.2889	
TGW	MV-AD	0.7746		0.7732	0.7767	0.7742	0.7742	0.7725	0.7788	0.7749	0.7744
	MV-ADE	0.7755		0.7732	0.7786	0.7755	0.7749	0.7707	0.7804	0.7756	
	MV-ADV	0.7769		0.7733	0.7811	0.7756	0.7746	0.7735	0.7822	0.7767	

注：表中第一列为目标性状，第二行为辅助性状

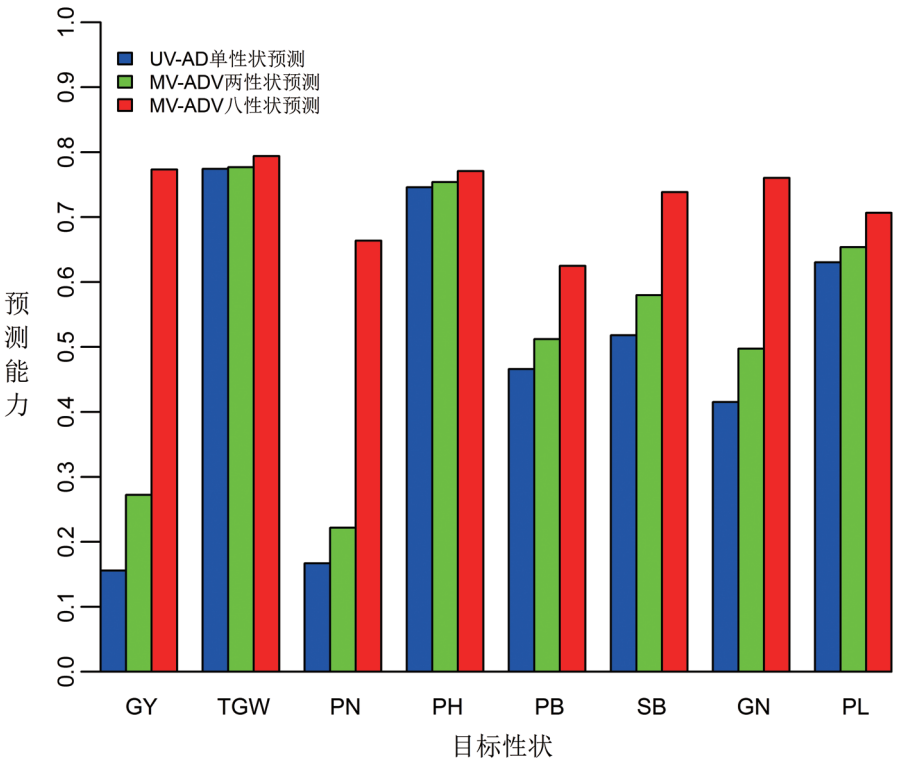

图2 杂交种单性状预测、两性状预测和八性状预测的平均预测能力

以获得更高的精度。

为了衡量利用选择指数进行预测的效果，本研究定义预测值与遗传效应值之间的决定系数为预测精度，与表型值之间的决定系数为预测能力。研究中使用 5 倍交叉验证的方法衡量各种模型的预测效果。在模拟研究中，每次以 T3 为目标性状，可选取不同组合的其他性状为辅助性状，然后建立选择指数。本研究对比 11 个不同的辅助性状组合，仅利用训练集数据构造选择指数 SI_{train}。指数预测能力反映了选择指数本身所具备的预测能力，在模拟的情形下，11 个组合的平均预测能力为 0.7035，远高于 T3 预测能力 (0.2470)，说明利用选择指数进行预测能实现更加精确的综合选择。而且，辅助性状组合与目标性状遗传相关程度越高，指数预测能力越高。此外，辅助性状的遗传力对指数预测能力也存在较大影响。因为辅助性状的遗传力越高，其中包含的遗传信息比重越大，以其为基础构建的选择指数也必然获得更高的遗传力，从而拥有较高的指数预测能力。水稻实际表型数据的研究中，指数预测能力的平均值为 0.5479，也远远高于 GY 的预测能力 (0.1344)。

当目标性状缺失时，可以用构建的选择指数直接预测目标性状。若选取不同的辅助性状组合构建选择指数，并直接利用指数预测 T3，在大多数情况下，预测的精度无法超越目标性状的预测精度，但是可以十分接近这一水平，这在目标性状缺失的情况下不失为一种有效的方法。但是，如果已经拥有了目标性状的训练群体，就没有必要直接用选择指数进行预测。选择指数能够综合多个辅助性状与目标性状的相关遗传信息，可以将选择指数的预测结果与传统 GS 的预测结果相结合，提高对目标性状的预测精度和预测能力。

对本研究选取的多个辅助性状组合，其预测结果见表 4。虽然大部分组合的指数直接预测精度低于 T3 预测精度，但是，所有组合的指数辅助预测精度都优于 T3 预测精度，提高的百分比为 0.6%~8.3%。同时，辅助性状与 T3 的遗传相关越高，指数辅助预测精度越高。指数辅助预测能力也有类似的特点，只是由于误差项的干扰而波动较大。

除了模拟的情形，本课题组还以水稻杂交种的 GY 为目标性状，TGW、PN、PH、PB、SB、GN 和 PL 为辅助性状，考察了选择指数辅助预测的情况。因为实际性状的育种值是未知的，所以无法得到真实的预测精度，只能用指数辅助预测能力来衡量指数的优劣。为了找到最佳的辅助性状组合，本研究采用逐步剔除的方法，从全部辅助性状组合出发，每次去除一个辅助性状，然后将去除前和去除后的所有组合进行对比，找到指数辅助预测能力最大的组合。结果表明，各组合所构建选择指数的辅助预测能力均高于 GY 的预测能力，说明指数辅助预测是有效的。其中，TGW、PN、PH、PB 和 SB 等 5 个性状构建的选择指数辅助预测能力为 0.1461，是所有组合中最大的，比 GY 的预测能力高 8.7%。

表4 多个辅助性状组合所构建指数辅助预测目标性状的精度和能力

辅助性状组合	指数辅助预测精度	指数辅助预测能力
C1	0.8848	0.2672
C2	0.8545	0.2541
C3	0.8726	0.2559
C4	0.8220	0.2475
C5	0.8646	0.2534
C6	0.8556	0.2522
C7	0.8543	0.2514
C8	0.8555	0.2474
C9	0.8511	0.2578
C10	0.8441	0.2498
C11	0.8319	0.2526

3　结语与展望

针对绿色超级稻育种实践中多个绿色性状的遗传改良问题，本文提出了关联分析和基因组选择的多性状方法。

首先，基于多元偏最小二乘和两阶段变量选择策略，提出了用于优异等位基因挖掘和遗传效应估计的多性状关联分析方法，并通过大量的模拟研究验证了该方法的可行性。研究表明，随着样本容量、候选基因遗传力以及 PIC 的增加，检测功效以及效应估计的准确度和精确度均明显提高，由于本文提出的方法充分利用了性状之间的遗传相关与环境相关信息，并且对所有的基因位点同时分析，因此在进行关联分析时具有明显的优势。与单性状分析相比，不论候选基因是同时控制多个性状的表达，还是仅控制其中一个性状的表达，绝大多数情况下具有更高的检测功效和更准确的效应估计。对于一因多效基因，多性状联合分析的优势更为明显，这与 Xiao 等 [37] 对主基因的分离分析、Jiang 等 [38] 对多性状联合 QTL 定位的研究结论一致。然而对于 PIC 值很小的位点，例如 Q1，基于偏最小二乘的多性状联合分析的检测功效较单性状分析差，这可能是

偏最小二乘对成分的提取损失了部分遗传变异方差信息的缘故。本文提出的模型目前仅适用于群体结构不明显的自然群体，下一步的工作设想是对模型作进一步改进，考虑消除群体结构对分析结果的影响，将群体结构纳入协变量，使之适用于任何的自然群体，进一步提高模型的适用范围。此外，本研究的模型仅涉及到候选基因的主效应，该分析方法可以进一步扩展到主效应加互作效应的模型，模型的构建可以参考 Zeng 等 [39] 的方法进行。

其次，本文提出的多性状基因组选择方法，特别有利于对一些低遗传力或者难以观测的重要性状进行选择 [40]。利用这一策略能够提高预测的精度，特别是性状之间存在较强的关联时 [41]。在绿色超级稻育种中，很多绿色性状 (如氮磷高效利用、抗逆性状和小区产量) 较难测定，此时利用目标群体的其他辅助性状进行预测具有十分重要的现实意义。但是如果所预测的群体未得到种植，即没有任何性状得到鉴定时，可以使用基于选择指数的 GS 方法进行预测。另一方面，传统的水稻育种关注产量水平的提高，对绿色性状的重视不够。目前绿色超级稻育种要求多个绿色性状同时得到发展，这就给全基因组选择带来新的挑战。基于选择指数的 GS 技术适应了这一需要，它能实现对水稻多个性状的同时选择。不论是模拟表型数据还是实际表型数据，选择指数的预测能力都大大高于传统 GS 的预测能力，充分说明选择指数具有更加精确的综合选择能力，对于绿色超级稻等项目的多目标育种具有重要的意义。此外，基于选择指数的 GS 方法能够利用与目标性状相关的多个辅助性状建立选择指数，实现对某个目标性状的辅助预测，提高对目标性状的预测精度和预测能力。本研究中选择指数的建立都是基于不同性状间的线性关系，而实际中这种关系可能是非线性的，在今后的研究中考虑性状间更加复杂的非线性联系，可进一步提高选择指数预测的效果。对于水稻杂交种实际数据选择指数的研究，虽然表型性状局限于 GY、TGW 等性状，分析结果对水稻其他性状的应用可能有一些限制，但是本研究的方法和结果为水稻绿色性状选择指数的构建，以及利用选择指数预测目标性状，提供了重要的技术参考。

[参 考 文 献]

[1] Zhang QF. Strategies for developing green super rice. Proc Natl Acad Sci USA, 2007, 104: 16402-9

[2] 张启发. 绿色超级稻的构想与实践[M]. 北京: 科学出版社, 2009

[3] Crossa J, Perez-Rodriguez P, Cuevas J, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci, 2017, 22: 961-75

[4] Xu Y, Li PC, Yang ZF, et al. Genetic mapping of quantitative trait loci in crops. Crop J, 2017, 5: 175-84

[5] Thornsberry JM, Goodman MM, Doebley J, et al. *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet, 2001, 28: 286-9

[6] Yu J, Pressoir G, Briggs WH, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet, 2006, 38: 203-8

[7] Kang HM, Zaitlen NA, Wade CM, et al. Efficient control of population structure in model organism association mapping. Genetics, 2008, 178: 1709-23

[8] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet, 2012, 44: 821-4

[9] Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet, 2010, 42: 348-54

[10] Zhang Z, Ersoz E, Lai CQ, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet, 2010, 42: 355-60

[11] Segura V, Vilhjálmsson BJ, Platt A, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet, 2012, 44: 825-30

[12] Rakitsch B, Lippert C, Stegle O, et al. A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics, 2013, 29: 206-14

[13] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet, 2013, 9: e1003264

[14] Xu Y, Hu WM, Yang ZF, et al. A multivariate partial least squares approach to joint association analysis for multiple correlated traits. Crop J, 2016, 4: 21-9

[15] Robert T. Regression shrinkage and selection via the lasso: a retrospective. J R Statist Soc B, 2011, 73: 273-82

[16] Zhang F, Guo X, Deng HW. Multilocus association testing of quantitative traits based on partial least-squares analysis. PLoS One, 2011, 6: e16739

[17] Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 2001, 157: 1819-29

[18] Wang X, Xu Y, Hu Z, et al. Genomic selection methods for crop improvement: current status and prospects. Crop J, 2018, 6: 330-40

[19] Beukert U, Li Z, Liu G, et al. Genome-based identification of heterotic patterns in rice. Rice (N Y), 2017, 10: 22

[20] Wang X, Li L, Yang Z, et al. Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity, 2017, 118: 302-10

[21] Scutari M, Howell P, Balding DJ, et al. Multiple quantitative trait analysis using bayesian networks. Genetics, 2014, 198: 129-37
[22] Hayashi T, Iwata H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics, 2013, 14: 34
[23] Calus MP, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol, 2011, 43: 1-14
[24] Xu S, Zhu D, Zhang Q. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc Natl Acad Sci USA, 2014, 11: 12456-61
[25] Hazel L, Lush JL. The efficiency of three methods of selection. J Heredity, 1942, 33: 393-9
[26] Dekkers JC. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet, 2007, 124: 331-41
[27] Jesus Ceron-Rojas J, Crossa J, Arief VN, et al. A genomic selection index applied to simulated and real data. G3 (Bethesda), 2015, 5: 2155-64
[28] Schulthess AW, Wang Y, Miedaner T, et al. Multiple-trait and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theoret Appl Genet, 2016, 129: 273-87
[29] Lyra DH, de Freitas Mendonça L, Galli G, et al. Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. Mol Breeding, 2017, 37: 80
[30] Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 2001, 157: 1819-29
[31] de los Campos G, Hickey JM, Pong-Wong R, et al. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 2013, 193: 327-45
[32] Riedelsheimer C, Technow F, Melchinger AE. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics, 2012, 13: 452
[33] Heslot N, Yang HP, Sorrells ME, et al. Genomic selection in plant breeding: a comparison of models. Crop Sci, 2012, 52: 146-60
[34] Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 (Bethesda), 2014, 4: 1027-46
[35] Xu Y, Wang X, Ding XW, et al. Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice, 2018, 11: 32
[36] Jia Y, Jannink JL. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics, 2012, 192: 1513-2
[37] Xiao J, Wang X, Hu Z, et al. Multivariate segregation analysis for quantitative traits in line crosses. Heredity, 2007, 98: 427-35
[38] Jiang CJ, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics, 1995, 140: 1111-27
[39] Zeng ZB, Kao CH, Basten CJ. Estimating the genetic architecture of quantitative traits. Genet Res, 1999, 74: 279-89
[40] Alimi NA, Bink MC, Dieleman JA, et al. Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. Theor Appl Genet, 2013, 126: 2597-625
[41] Piepho HP, Möhring J, Melchinger AE, et al. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica, 2008, 161: 209-28