DOI: 10.13376/j.cbls/2018135
文章编号：1004-0374(2018)10-1120-09

周发松，博士，国家第六批“千人计划”专家，现任武汉双绿源创芯科技研究院董事长、院长。1998 年获丹麦皇家农牧大学植物分子生物学博士学位。1998—2010 年先后在英国 John Innes 研究中心、美国康奈尔大学、美国 Ceres 生物技术公司、美国 Mendel 生物技术公司从事植物生物技术研究。2010—2018 年回国参与筹建中国种子集团有限公司生命科学技术中心，历任基因组育种部总监，作物育种技术创新与集成国家重点实验室主任、首席科学家。主要研究领域包括基因组智能育种系统开发、全基因组育种基因芯片研制、高通量 DNA 测序及分子标记技术，植物功能基因克隆、植物抗病遗传和分子基础，能源作物品种资源研究和稻、麦作物新品种选育。

水稻全基因组选择育种技术平台构建与应用

邱树青[1,2]，陆　青[1,2]，喻辉辉[1]，倪雪梅[3,4]，张耕耘[3,4]，何　航[5]，谢为博[6]，周发松[1,2]*
(1 中国种子集团有限公司生命科学技术中心，武汉 430206；2 武汉双绿源创芯科技研究院，武汉 430056；
3 深圳华大生命科学研究院，深圳 518083；4 农业基因组学国家重点实验室，深圳 518083；
5 北京大学蛋白质与植物基因研究国家重点实验室，北京 100871；
6 华中农业大学作物遗传改良国家重点实验室，武汉 430070)

摘　要：随着 DNA 标记技术的发展和越来越多的功能基因被鉴定和克隆，水稻全基因组选择育种已经成为精准高效的新品种培育方法。该文重点介绍新创建的基于高通量测序和基因组育种芯片的全基因组选择技术平台，以及建立的水稻基因组序列变异数据库、受选择功能区段鉴定技术体系和品种系谱溯源技术体系。这些技术平台和体系将有助于水稻全基因组选择育种工作的开展，提高育种效率，更好地满足消费市场对品种培育提出的新要求。

关键词：DNA 测序；基因芯片；基因数据库；系谱溯源；基因组育种
中图分类号：Q819；S511　　**文献标志码**：A

The development and application of rice whole genome selection breeding platform

QIU Shu-Qing[1,2], LU Qing[1,2], YU Hui-Hui[1], NI Xue-Mei[3,4],
ZHANG Geng-Yun[3,4], HE Hang[5], XIE Wei-Bo[6], ZHOU Fa-Song[1,2]*
(1 Life Science and Technology Center, China National Seed Group Co., Ltd, Wuhan 430206, China;
2 Wuhan Greenfafa Institute of Novel Genechip R&D, Wuhan 430056, China;
3 BGI-Shenzhen, Shenzhen 518083, China; 4 State Key Laboratory of Agricultural Genomics, Shenzhen 518083, China;
5 State Key Laboratory of Protein and Plant Gene Research, Peking University, Beijing 100871, China;
6 National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China)

收稿日期：2018-10-18
基金项目：国家高技术研究发展计划(“863”计划)“绿色超级稻新品种选育”(2014AA10A602)
***通信作者**：E-mail: zhoufasong@greenfafa.com

Abstract: It is becoming a practical method to develop super green rice varieties efficiently through whole genome selections with DNA marker technology ready to practise as well as more and more rice functional genes being identified and cloned. In this paper, we presented two whole genome selection methods which were established based on high throughput sequencing technology and whole genome rice gene-chip. We also built RiceVerMap database, a functional haplotype identification system, and a rice pedigree analysis system. All of these may greatly help rice breeders to perform the whole genome selection method in their breeding programs to better meet consumers' novel demand for safety and high quality with more efficient development of new rice varieties.

Key words: DNA sequencing; gene-chip; gene database; pedigree analysis; whole genome breeding

在新品种培育的过程中，性状或者功能基因的准确鉴定和选择是关键。过去育种选择主要依赖于育种家对田间表型的评价，育成一个新品种需要反复多年的大田种植、表型观察和性状综合评价和选择，过程漫长。随着生物技术的发展，育种工作者逐渐认识到通过基因对性状进行选择不仅准确可靠，而且节约时间，特别是在植物表型性状难以准确鉴定的情况下，通过检测基因组中与功能基因紧密连锁的特异性 DNA 标记来进行选择，可以大幅度提高育种效率。建立单一目标性状的对应基因或者分子标记的连锁关系，在育种群体中对这些标记进行选择，能实现目标性状的有效改良。这种分子标记辅助选择方法已经成为普通的育种实践。但是，低通量的分子标记辅助育种只能鉴定选择少数的功能基因，很难解决目标基因转育过程中经常发生的遗传连锁累赘问题，更不能排除遗传背景的干扰。随着 DNA 测序技术的迅猛发展，测序成本大大降低，有效促进了水稻重测序的广泛开展，目前公共基因数据库中已经积累了海量的水稻基因组重测序数据，为大量开发分子标记提供了充分的信息。同时，分子标记检测技术也逐渐实现了高通量、低成本和自动化，为育种应用奠定了坚实基础。高通量分子标记主要有两大类：一类是基于新一代测序技术的分子标记，另一类是基因芯片技术。基于 DNA 测序的基因分型技术，以及基于基因芯片的分子标记检测技术，各有其特点和技术优势，适用于不用的育种环节和目的。

随着人们生活水平的提高，人们对稻米的品质和健康指标提出了更高的要求，利用全基因组选择技术进一步改良品质、增强抗病和抗虫能力以及提高肥料和水资源利用效率成为了育种家的主要任务。本课题根据绿色超级稻品种培育的需要，充分利用水稻基因组的最新研究成果，致力于建立一套水稻全基因组选择技术平台。本文将重点介绍以高通量测序和基因芯片为核心的水稻全基因组选择育种技术平台，同时介绍全基因组分子标记技术在水稻品种系谱溯源中的应用，以及利用水稻育种基因组数据库提高基因和种质资源的管理效率。

1 以育种芯片为核心的水稻全基因组选择育种技术平台

1.1 水稻育种芯片的设计

为了使水稻功能基因组研究成果得到利用，不断满足水稻大规模商业化育种需求，我们利用美国 Illumina Infinium 专利制造技术设计制作了三款水稻 SNP (单核苷酸多态性) 基因芯片，三款基因芯片的具体设计时间和设计思路如图 1。

第一款水稻全基因组育种芯片 RICE6K[1] 包含两种类型的探针：第一类探针包含从核心亲本基因组序列及 520 个水稻品种重测序数据比较分析中筛选出的 5 556 个 SNP 位点 [2]；第二类探针包含与 40 个水稻功能基因相关的 80 个 SNP/INDEL 位点 [3]。两类设计的探针共检测 5 636 个 SNP/INDEL 位点。经过测试，RICE6K 能够检测分布于水稻全基因组的 5 102 个 SNP/INDEL 位点，其中约有 4 500 个 SNP/INDEL 位点具有很好的基因分型能力 [1]。

第二款水稻全基因组育种芯片 RICE60K 的探针设计分两步：首先，对 801 份水稻品种的重测序数据中发掘的 1 000 万个 SNP 位点进行筛选；另外，RICE6K 水稻 SNP 芯片上的高质量的 SNP/INDEL 位点和 1 000 个位于已克隆的水稻重要功能基因上的 SNP 位点直接用于合成 RICE60K 育种芯片上的探针。RICE60K 共设计了 58 290 个 SNP 位点，成功合成了 51 478 个 SNP 位点，其中高质量 SNP 位点数约为 4.3 万个，该款芯片又称为 RiceSNP50[4]。

第三款水稻全基因组育种芯片 RICE90K 的设计分三步实现：第一步是在 RICE60K 的基础上增加 30K 的标记，研制成育种芯片 RICE60KAdd1；第二步是在 RICE60K 的基础上再增加 30K 的标记，研制成育种芯片 RICE60KAdd2；第三步是将前面

RICE6K (2011)：基于 **520** 份水稻品种设计，着眼于籼粳之间和籼稻内部的主要多态性位点，包含部分针对已克隆基因设计的探针。

RICE60K (2012)：基于 **801** 份水稻品种设计，着眼于普遍适用性，包含各水稻品种的主要多态性位点，针对基因区 SNP 进行优化，包含更多针对已克隆基因设计的探针，具有优秀的继承性。

RICE60KAdd1 (2013)：基于**>1600** 份水稻品种设计，在 60K 基础上增加 30K。着眼于骨干育种亲本包含的多态性，包含部分野生稻来源标记，及更多关键基因区域的标记。

RICE60KAdd2 (2014-2015)：基于**>5500** 份水稻品种设计，在 60K 基础上增加 30K。包含与性状紧密连锁的 GWAS 及受选择位点，功能标记，功能探针单倍型标记，GMO 标记，及基因区域大效应位点。

图1 系列育种芯片的整体设计思路

所有高质量 SNP 位点聚合成最终的 RICE90K 芯片。

育种芯片 RICE60KAdd1 新增的 30K 探针代表 27 781 个位点，包含以下类型探针：从已报道的 879 个功能基因区选择的关键基因区域标记 [5]；从已发表的 446 个野生稻品种中设计的均匀分布的野生稻来源的标记 (http://202.127.18.221/RiceHap3/index.php)；从 100 多个生产上广泛应用的杂交稻混合测序后开发的推广杂交种特有的标记；从 1 491 个水稻品种重测序中分析设计的品种代表性标记；5 个育种相关重要基因区段的基因区域标记。

育种芯片 RICE60KAdd2 新增的 30K 探针包括：3 000 份水稻品种重测序数据分析获得的种质资源代表性标记 [6]；超过 20 个农艺性状的 500 个全基因组关联分析 (genome-wide association study, GWAS) 效应位点的标记；转基因 (GMO) 成分检测的标记，包含 CaMV 35S 启动子、NOS 终止子、抗除草剂基因 *Bar*、抗虫 *Bt* 基因 *Cry1A* 等十多个基因 / 元件的序列，共 70 多条检测探针；新增功能基因区段单倍型标记，共约 20 个基因的 4 000 多个单倍型检测探针；新增功能基因标记。

新设计制造的育种芯片 RICE60KAdd1 和 RICE-60KAdd2 都经过了严格的测试。用于测试的水稻样品包括：21 份代表性水稻品种 (11 个籼稻、7 个粳稻、3 个中间类型) 两两双列杂交获得的 210 个杂交 F_1 组合中随机选取的 48 个组合；100 份水稻核心种质资源与 4 个两系不育系所配的 400 个杂种 F_1 中随机选取的 48 个组合；90 多份粳稻与 2 个两系籼型不育系所配的 96 个杂交 F_1 组合；96 个水稻核心种质资源 (籼稻)；从 400 多份粳稻中选取的代表性粳稻种质资源 100 多份。测试结果表明，育种芯片 RICE60KAdd1 上高质量的 SNP 标记位点数约为 6.5 万个，RICE60KAdd2 上高质量的 SNP 标记位点数目约 6.3 万个。将所有高质量的 SNP 位点聚合在一起，成为最终的 RICE90K 芯片，其中高质量的 SNP 标记位点数目达到 8.5 万个。三款芯片的具体参数指标如表 1 所示。

1.2 水稻芯片全基因组选择育种技术平台

以育种芯片为核心的水稻全基因组选择育种技术体系，具体包括利用高通量 SSR 标记技术鉴定筛选目标基因、利用 OpenArray 芯片技术鉴定筛选染色体区段单倍型、利用全基因组育种芯片技术鉴定筛选遗传背景。应用全基因组育种技术，我们对目前生产上大面积种植的水稻骨干亲本进行了抗病虫定向改良，实现了稻瘟病、褐飞虱等抗性基因的精准导入，在不改变水稻亲本综合农艺性状的前提下，显著提高了亲本的病虫害抗性；将多个抗稻瘟病基因分别导入同一亲本或品种，培育出了遗传背景高度一致的近等基因系；将带有不同抗稻瘟病基因的

表1 水稻全基因组育种芯片技术参数和应用范围

芯片类型	微珠总数	高质量标记数	标记来源品种数	推荐使用范围
RICE6K	6 000	4 473	520	品种真实性检测、籼粳成分分析、籼稻基因分型、基因初步定位
RICE60K	60 000	43 386	801	遗传背景分析、粳稻基因分型、基因精细定位
RICE90K	90 000	~85 000	>5 500	基因指纹分析、功能基因单倍型分析、全基因组关联分析、设计育种

多个近等基因系组合在一起，可以构成农艺性状整齐一致、抗性水平高且持续稳定的多系品种。

水稻全基因组育种技术可用于商业化育种流程各环节，加速商业化育种进程，如图 2 所示。Chen 等 [4] 利用育种芯片 RiceSNP50 对 195 个不同来源的水稻资源进行了分析，通过对 43 386 个高质量的 SNP 标记多态性的聚类，将 195 份水稻资料清晰地分成 3 个大类。还可以利用育种芯片，结合单倍型分析，对育种资源的特异功能基因进行分析，对部分表现型进行预判。在亲本创制中，可以利用基因组育种技术，通过定向改良，将优异的基因定向导入受体亲本中，创建新的种质资源。Mi 等 [7] 通过基因组育种技术，成功将来源于 Dular 的广亲和基因定向导入 9311 中，创制了具有广亲性的 9311，且与典型粳稻品种 Balilla 杂交后，杂种 F_1 代结实率显著提高。通过育种芯片对测配亲本所含功能基因的分析，排除明显不含主效抗性基因的组合，或者排除一些近似姊妹系等，减少组合的数量。在测试、测评中，利用育种芯片可以保证参试组合的准确性，避免多年次参试组合的差异，也可以为晋升的组合提供基因指纹。在种子生产和产品销售中，利用育种芯片结合 SSR 标记，可以精确检测种子纯度和品种真实性。利用育种芯片检测结果可以建立高分辨率的品种基因指纹身份证，用于品种权保护和品种系谱溯源。

2　基于GBS的水稻全基因组选择育种技术平台

简化基因组测序 (reduced-representation genome sequencing, RRGS)，现在常用的包括 RAD-seq (restriction site associated DNA sequencing) 和 GBS (genotyping by sequencing)。通过对酶切获得的酶切片段进行高通量测序，能够降低基因组的复杂度，操作简便，同时不受参考基因组的限制，可快速鉴定出高密度的 SNPs，用于遗传图谱构建、群体进化分析、QTL 定位、辅助 scaffold 组装到染色体等。

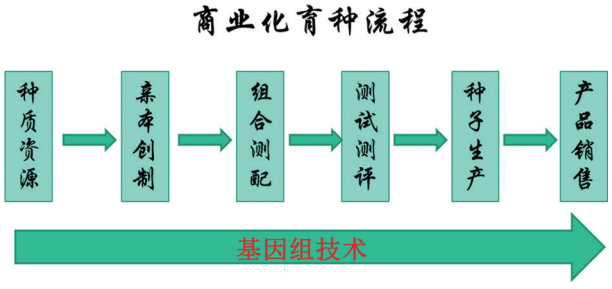


图2　以育种芯片为核心的水稻全基因组育种技术在商业化育种流程各环节中的应用

GBS 是一种高效而经济的 SNP 发掘和基因分型技术。近年来，随着高通量测序技术的发展，基因分型的成本持续降低，GBS 技术作为第二代深度测序基础上发展起来的简化基因组测序技术，通过采用酶切加标签的方法，使多样本高通量平行测序得以实现 [8]。这不仅大大降低了基因测序的成本，也使对大样本全基因组的基因分型成为可能，对深入了解种质资源的遗传背景和系统演化具有重要意义 [9-10]。同时，GBS 获得的短读序列可通过有参考或无参考基因组的形式进行拼接组装，进而获得高密度的 SNP 标记，利用这些 SNP 标记或者开发的 bin 标记可进行遗传图谱构建 [11]、遗传图谱加密 [12-13]、全基因组关联分析 (GWAS)[14] 及基因组的辅助组装等研究。

目前，GBS 技术作为基因分型的重要手段，已经在遗传图谱构建 [15]、遗传选择 [10]、基因多样性研究 [16]、种质鉴定 [17] 及品种识别 [18-19] 等领域得到广泛的应用。

2.1　水稻全基因组选择育种测序技术平台的建立

基于 Hiseq 4000、×Ten、BGISEQ-500 测序平台，已建成了高通量、低成本的全基因组选择育种技术平台。该平台包含样品 DNA 提取及检测、文库构建、样本测序及信息分析四大模块，年通量达到 6 万样品，单个水稻样品进行基因分型的成本低廉，能够完成不同样品的基因分型工作，并构建遗传连锁图谱、遗传多样性研究、种质鉴定、基因定位等。

基本的技术流程为：提取基因组 DNA，利用 *EcoR*I 或 *Apek*I 酶进行酶切，通过 GBS 建库流程构建文库，最后选择适宜的测序策略对酶切位点相关片段进行测序，得到的原始数据经过质控和数据过滤，进行生物信息分析。

该技术平台操作物种多样，不受参考基因组限制，有无参考基因组均可，不仅适用于二倍体，多倍体也可使用。与芯片方法相比较，可以检测到新的变异序列。每个酶切 tag 的平均覆盖深度至少 10×，可以保证分子标记的高准确性。如果只测序酶切简化的基因组 DNA，数据量要求低；建库也可以选择样本加 barcode 混合测序，降低建库成本；只需 2~3 个月即可完成大样本的群体研究。另外，该技术平台涉及领域广，涵盖进化、遗传图谱构建与 QTL 定位、辅助基因组组装、GWAS、BSA 等多个领域。

2.2 水稻全基因组选择测序技术平台的应用

2.2.1 获取骨干亲本基因型数据

目前国内籼稻两系不育系不育来源主要为农垦58S和安农S-1，选取这两支来源的代表不育系亲本进行基因型检测，完成Y58S、P88S、宣69S、新安S、深08S、广占63-4S、广占63-2S、C815S、03S等超级稻温敏两系骨干不育系亲本的基因型分析，并进行多样性分析，绘制了系谱关系图和亲缘关系图。结果显示，广占63S系列的不育系与培矮64S系列的不育系在基因组上有比较大的差别，广占63S系列的不育系彼此之间差别比较小。培矮64S和安农S-1系列的不育系之间差异相对大一些，其中Y58S和深08S比较近。

2.2.2 完成整理10种以上绿色超级稻重要绿色性状信息

完成10个重要绿色性状基因信息的整理：*sd1*（株高），*gn1*、*IPA1*（穗粒数），*Hd1*（生育期），*pi1*、*pi2*（抗稻瘟），*Xa21*（抗白叶枯），*qsW5*、*GS3*和*GW2*（粒重），包括基因序列、基因位点、主要农艺性状等。完成*qsW5*、*GS3*、*GW2*这3个粒重相关基因的等位基因型分析，得到5个主要等位基因型及数十个稀有等位基因型。

对92份主栽水稻亲本材料*GS3*、*GW2*、*qSW5*这3个基因区段进行分析，利用其SNP信息，进行等位基因型区分及最终等位基因型确定。在92份主栽水稻亲本中，共得到11种*qsw5*等位基因型、10种*GS3*等位基因型和8种*GW2*等位基因型，通过与其关联对应，确立各类等位基因型代表种质及其等位基因型表型。

2.2.3 挖掘抗病虫、高产优质性状QTL位点

基于表型数据，开展大规模的GWAS和QTL定位分析，精确定位水稻抗白背飞虱、抗稻瘟病、粒型、株高、分蘖数、直链淀粉含量等重要性状QTL位点。

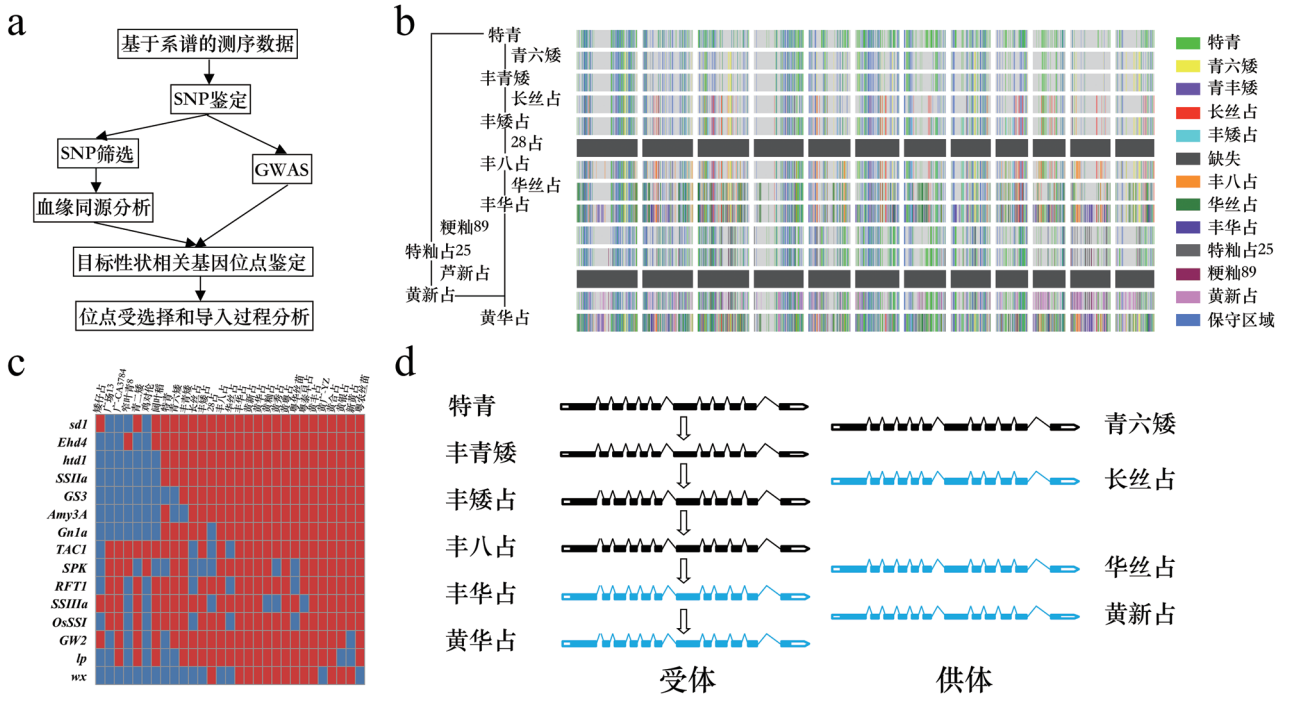
以“巨穗稻”R1128和日本晴(Nipponbare, NIP)为亲本，构建了含1 200株单株的F_2分离群体，对分离群体每个单株进行穗粒数、落粒性、米粒长宽、一级梗数目、二级梗数目、分蘖数、株高、生育期、结实率等性状进行表型鉴定。开发了74 329个SNP分子标记，结合表型鉴定信息，采用复合区间作图法定位穗粒数、落粒性、结实率等基因/QTL位点51个，挖掘高产、优质性状基因，为绿色超级稻品种选育提供良好的基因来源。

3 品种系谱溯源技术平台

在群体遗传分析中，稀有变异往往发挥着重要的作用[20-21]，而传统的全基因组关联分析(GWAS)往往难以准确鉴定稀有变异[22]。人类中的家谱或作物中的系谱信息则能够通过富集稀有变异，从而对这些变异做出精确高效的鉴定[23]。在农作物改良过程中，系谱的选育可导致粒型、株型、产量、抗虫等重要农艺性状的明显差异[24-25]。因此，通过引入品种的系谱信息，能够更加准确地挖掘出育种过程中与农艺性状相关的基因，对作物的育种改良具有重要价值[26]。

另外，通过谱系中受选择区域的分析，发现育种过程中的关键基因结构，进而推断育种材料中重要性状的表型，也逐渐成为基因组育种的重要手段。基于系谱的基因型数据分析一般包括如下几个步骤：首先，通过鉴定筛选多态性SNP，寻找系谱样品中存在的变异位点；随后，对这些突变位点进行筛选，去除在亲代和子代之间无遗传关系的变异；对这些经过筛选后的变异位点进行血缘同源(identical by descent, IBD)分析，即可溯源后代血缘，并鉴定与系谱育种中受选择区域关联的位点；最后，结合GWAS等分析，确定受选择的关联位点[27-29]（图3a）。在作物中，随着系谱信息的不断完善和测序、芯片技术的发展，系谱溯源成为有力的遗传分析手段，如对玉米自交系郑58、5003和478的分析发现了父辈和祖辈中不同亲本对子代的基因组贡献基本比例存在较大差异[30]，对水稻系谱桂朝2号、黄华占、蜀恢527的分析发现了骨干品种中控制优异农艺性状表现的关键基因[31]。

利用水稻高通量SNP芯片，我们建立了水稻品种系谱溯源技术平台。该平台利用SNP芯片快速获得系谱中材料的基因型信息，鉴定育种过程中水稻基因组的受选择区域和在系谱中导入并逐代传递的区域，并结合GWAS和QTL分析结果，发掘一系列与重要农艺性状有关的基因变异位点，根据这些位点在系谱中的变化情况，揭示系谱在育种历史中性状改良的分子过程。利用该平台对优质水稻品种黄华占系谱中99个材料进行RiceSNP50芯片基因型测定，并进行溯源分析[32-33]。结果显示，黄华占的血缘中有18.21%来自于祖先特青，其余来自于后续引入的青六矮、丰青矮、长丝占等（图3b）。具体到黄华占65年育种谱系中受到选择的基因组区域，能看到在不同年代中有显著特异的基因



a：系谱分析的基本流程；b：黄华占育种过程中基因组来源分布情况；c：13个重要农艺性状相关农艺位点基因型(红色表示与黄华占基因型一致，蓝色表示不一致)；d：基因*wx*在选择过程中的动态变化情况(蓝色表示优势等位基因)

图3　系谱溯源分析

类型受到选择。在 1997 年以前，偏向于选择包括 *OsMSH5* 和 *rFCA* 在内的与开花和不育相关的基因；在 1997—2004 年，偏向于选择 *Pib* 等与广谱抗性有关的基因；在 2004 年以后，偏向于选择其他生物胁迫相关的基因，如抗菌基因 *NH1*、抗虫基因 *OsSUT1*(图 3c)。这些结果表明黄华占育成是由几方面选育综合作用的结果，包括早期的株型和适应性改良 (株高、花期) 及后期的抗性改良 (生物及非生物抗性)。以黄华占中米质相关基因 *wx* 为例，原本特青中因未含有优势等位基因而口感较差，而长丝占和华丝占两个可能的变异为黄华占提供了优势等位基因，且在随后的系谱育种过程中该变异被固定 (图 3d)。

我们同样利用该平台分析了黄华占衍生系谱 [33]。考察了品质改良和产量提高两条衍生系谱，鉴定出 1 113 个保守且可追溯的染色体区域 (conserved Huanghuazhan traceable blocks, cHTBs)。在两条系谱中均高度保守的 cHTBs 中包括了许多重要的农艺性状相关基因，如控制株高的 *sd1*、控制抽穗期的 *Ehd4*、控制分蘖高度和矮化的 *htd1*、控制可溶淀粉合成的 *SSIIa*、控制籽粒大小的 *GS3*、控制 α 淀粉酶的 *Amy3A*、控制籽粒数目的 *Gn1a*、控制分蘖角度的 *TAC1* 等，表明这些等位基因在育种中已固定。而 111 个特异存在于品质改良系谱的 cHTBs 中的基因，如 *SPK43*、*RFT144*、*SSIIIa45* 及 *OsSSI4* 等与株型、开花、淀粉合成等过程有关；201 个特异存在于产量提高系谱的 cHTBs 中的基因，如 *GW2*、*lp*、*wx* 等与籽粒长宽、穗大小、食用品质等性状相关。这些 cHTBs 的变异与其对应性状的改良有关，这些 cHTBs 的鉴定对黄华占分子育种有重要价值。

4　水稻基因组序列变异数据库

基因组序列变异是开展全基因组选择育种应用的基础。完善的基因组序列变异数据库为分子标记设计、品种亲缘关系鉴定、受选择区域分析、基因功能分析以及全基因组关联分析等研究提供了便捷可靠的资源。

水稻基因组序列变异数据库 RiceVarMap v2.0 (http://ricevarmap.ncpgr.cn/v2/) 是一个功能完备的水稻序列变异数据库，整合了基因组变异数据、变异功能注释数据、表型数据以及全基因组关联分析数据。该数据库通过对 4 507 份水稻重测序数据进行分析 [34-36]，并基于水稻日本晴参考基因组 [37]，鉴定出 17 397 026 个基因组变异位点 (包含 14 541 446

个 SNP 位点以及 2 855 580 个 INDEL 位点)。由于大部分品种测序覆盖度较低，原始基因型的缺失率高达 33.4%；采用 LD-KNN 算法对基因型进行补缺后，平均缺失率降为 2.32%，经评估补缺后准确率达到 99% 以上。

该数据库通过整合多组学的数据获得了变异的精准注释：(1) 使用 snpEff[38]、CooVar[39] 与 PolyPhen-2[40] 对编码区的变异及错义突变进行效应评估，CooVar 可以考虑多个变异的共同影响 (如有 INDEL 造成阅读框发生改变)，而 PolyPhen-2 基于蛋白质局部序列的保守性定量评估变异效应；(2) 整合染色质开放区数据对非编码区变异潜在影响进行评估，目前提供水稻愈伤与幼苗的染色质开放区数据[41]；(3) 整合 GWAS 结果，提供与序列变异显著关联的表型的信息，目前整合了 13 个农艺性状 (包含抽穗期、株高与粒重等) 和苗期 840 个代谢性状的全基因组关联分析的结果。

同时，为了方便研究者使用，数据库提供了丰富的查询界面以及实用工具。基于 bokeh 库，RiceVarMap v2.0 提供了多种可视化展示模式。数据库主要包含 3 个模块：(1) Genomic Variation 界面，包含多种变异数据查询功能；(2) Cultivar & Phenotype 查询界面，包含品种信息查询、品种表型以及 GWAS 结果查询；(3) Tools 界面，包含多种实用工具，如比对工具 blast、引物设计工具、单倍型网络分析工具，以及不同版本的变异位置转换工具与基因组变异可视化工具 GBrowse。通过这些功能模块的查询，研究者可以轻而易举地获取目标区段的变异数据及注释结果，为下一步的实验设计提供帮助。

5 受选择功能区段鉴定技术平台

育种是对原有品种中的遗传变异进行选择和重新组装，以获得目标性状改良的新品种的技术。长期育种实践中，众多的育种工作者为了相同或相似的育种目标而努力，对大量品种资源进行密集选择，并由此改变了有利基因型在品种群体中的频率分

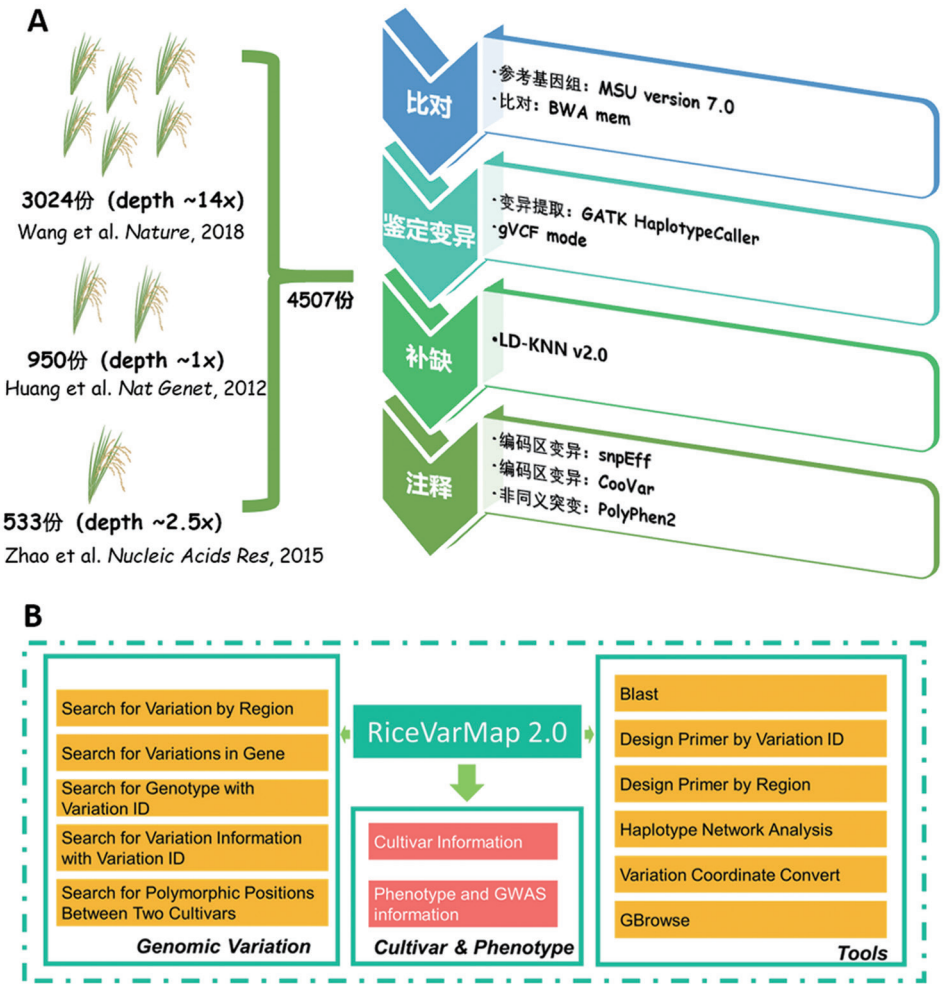


图4 水稻基因组序列变异数据库RiceVarMap数据处理流程及主要功能

布，在基因组中留下了相关的“印迹”。因此，通过对大量品种的序列进行分析，可以鉴定出育种过程中受选择的基因组位点。相比于 GWAS 需要基于特定的、准确的表型性状鉴定优良单倍型，受选择位点分析提供了独立于具体表型鉴定功能区段和优良单倍型的途径，这些位点或能对进一步的品种改良提供指导。

鉴定受选择位点和受选择单倍型一般包括如下几个步骤：首先，获得同一物种大量品种的多态性位点的基因型数据；随后，进行群体遗传学分析，划分亚群，并鉴定亚群间基因型频率差异巨大的区段；同时考虑各区域的重组率，结合重排 (Permutation) 分析，最终确定显著的受选择区段；最后，对获得的受选择区段，选一个差异最大的亚群作为外群，鉴定每个位点的原始基因型和衍生基因型；衍生基因型比例大于一定阈值的单倍型为该区段受选择的单倍型 [34]。

我们利用上述方法系统分析了籼稻品种中的受选择位点。首先基于高质量的 SNP 数据系统分析了水稻品种的群体结构，鉴别出了籼稻中的两大主要的亚群籼 I 和籼 II，它们具有不同的地理起源，其形成可能与“绿色革命”早期在中国及国际水稻研究所独立的育种工作有关。其中，中国传统农家种和三系水稻中的保持系属于籼 I，为中国南方血缘，而恢复系和大部分改良的水稻品种属于籼 II，带有东南亚血缘。三系杂交水稻中的保持系和恢复系分别属于这两个亚群，对应于籼稻中的两个杂种优势群。通过群体遗传学分析，我们鉴定了这两个亚群之间受到不同选择的 200 个基因组区段。这些区段包括了与产量 (如 *Gn1a*、*Rf1*、*sui1* 和 *LP* 等)、株型 (如 *sd1*、*SLR1*、*OsBRI1* 等)、抗性 (如 *Xa4*、*Xa24*、*Xa26* 和 *Xa27* 等) 以及营养吸收 (如 *OsGS1*、*OsNRT2.3*、*OsNAR2.2* 和 *OsAMT1;1* 等) 等绿色性状相关的许多已知功能基因和大量功能未知的位点。这些受选择位点为进一步改良水稻提供了重要靶点。

进一步研究发现，随着一个品种中受选择单倍型的累积，品种的产量得到稳步提高，表明一个品种中受选择单倍型的数量一定程度上可用于预测该品种的育种价值。定义不同品种包含的具有受选择单倍型或受选择区段的数目，作为受选择位点指数。受选择位点指数可应用于以下几方面：(1) 背景选择，当有若干材料能同时满足育种目标时，可优选具有较大受选择位点指数的材料；(2) 品种改良，对于现有优良品种，可通过导入其不具有的受选择单倍型，使其具有更高的受选择位点指数；(3) 杂交组合，如果两个亲本组合在一起包含更多的受选择单倍型，其杂种可能有更强的杂种优势。以上受选择区段鉴定技术平台可结合基因组选择育种实现快速培育新品种或杂交组合。

6　展望

将近一个世纪的水稻杂交育种研究使品种的产量潜力和综合农艺性状达到了相当高的水平，依靠传统的育种方法进一步培育出突破性的大品种越来越难。全基因组选择育种技术的逐渐成熟和广泛应用为育种研究提供了新的机遇，通过重要目标性状基因的精准鉴定和优化组合，有望培育出更受农民和消费者欢迎的绿色水稻新品种。选择综合性状优良的品种，针对其缺点，精准导入目标性状基因，有望实现高产与优质的统一，以及高效农业与环境友好的兼顾，从而支撑农业可持续性发展。

致谢：感谢中化集团和中国种子集团有限公司提供研发条件和项目配套经费；雷昉等基因组育种团队成员参与了水稻育种芯片研制和育种应用研究；赵虎完成了水稻基因组序列变异数据库的部分工作。

[参　考　文　献]

[1] Yu H, Xie W, Li J, et al. A whole-genome SNP array (RICE6K) for genomic breeding in rice. Plant Biotechnol J, 2014, 12: 28-37

[2] Huang X, Wei X, Sang T, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet, 2010, 42: 961-7

[3] Jiang Y, Cai Z, Xie W, et al. Rice functional genomics research: progress and implications for crop genetic improvement. Biotechnol Adv, 2011, 30: 1059-70

[4] Chen H, Xie W, He H, et al. A high-density SNP genotyping array for rice biology and molecular breeding. Mol Plant, 2014, 7: 541-53

[5] 肖景华, 吴昌银, 袁猛等. 中国水稻功能基因组研究进展与展望. 科学通报, 2015, 60: 1711-22

[6] 3000 rice genomes project. The 3000 rice genomes project. Gigascience, 2014, 3: 7

[7] Mi J, Li G, Huang J, et al. Stacking S5-n and f5-n to overcome sterility in indica-japonica hybrid rice. Theor Appl Genet, 2016, 129: 563-75

[8] Sonah H, Bastien M, Iquira E, et al. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One, 2013, 8: e54603

[9] Glaubitz JC, Casstevens TM, Lu F, et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One, 2014, 9: e90346

[10] Poland J, Endelman J, Dawson J, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome, 2012, 5: 103-13
[11] Ward JA, Bhangoo J, Fernandez F, et al. Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. BMC Genomics, 2013, 14: 2
[12] Spindel J, Wright M, Chen C, et al. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor Appl Genet, 2013, 126: 2699-716
[13] Zhou Z, Zhang C, Zhou Y, et al. Genetic dissection of maize plant architecture with an ultra-high density bin map based on recombinant inbred lines. BMC Genomics, 2016, 17: 178
[14] Li X, Li X, Fridaman E, et al. Dissecting repulsion linkage in the dwarfing gene *Dw3* region for sorghum plant height provides insights into heterosis. Proc Natl Acad Sci USA, 2015, 112: 11823-8
[15] Guajardo V, Solis S, Sagredo B, et al. Construction of high density sweet cherry (*Prunus avium* L.) linkage maps using microsatellite markers and SNPs detected by genotyping-by-sequencing (GBS). PLoS One, 2015, 10: e0127750
[16] Lin M, Cai S, Wang S, et al. Genotyping-by-sequencing (GBS) identified SNP tightly linked to QTL for pre-harvest sprouting resistance. Theor Appl Genet, 2015, 128: 1385-95
[17] Wong MM, Gujaria-Verma N, Ramsay L, et al. Classification and characterization of species within the genus lens using genotyping-by-sequencing (GBS). PLoS One, 2015, 10: e0122025
[18] Lombardi M, Materne M, Cogan NO, et al. Assessment of genetic variation within a global collection of lentil (*Lens culinaris* Medik.) cultivars and landraces using SNP markers. BMC Genet, 2014, 15: 150
[19] Cabezas JA, Ibanez J, Lijavetzky D, et al. A 48 SNP set for grapevine cultivar identification. BMC Plant Biol, 2011, 11: 153
[20] Li B, Chen W, Zhan X, et al. A likelihood-based framework for variant calling and *de novo* mutation detection in families. PLoS Genet, 2012, 8: e1002944
[21] Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet, 2010, 11: 415-25
[22] Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. Nat Rev Genet, 2011, 12: 465-74
[23] Peng G, Fan Y, Palculict TB, et al. Rare variant detection using family-based sequencing analysis. Proc Natl Acad Sci USA, 2013, 110: 3985-90
[24] Song XJ, Kuroha T, Ayano M, et al. Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. Proc Natl Acad Sci USA, 2015, 112: 76-81
[25] Zhang X, Wang J, Huang J, et al. Rare allele of *OsPPKL1* associated with grain length causes extra-large grain and a significant yield increase in rice. Proc Natl Acad Sci USA, 2012, 109: 21534-9
[26] Huang J, Li J, Zhou J, et al. Identifying a large number of high-yield genes in rice by pedigree analysis, whole-genome sequencing, and CRISPR-Cas9 gene knockout. Proc Natl Acad Sci USA, 2018, 115: E7559-67
[27] Roach JC, Glusman G, Smit AFA, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science, 2010, 328: 636-9
[28] Bahlo M, Tankard R, Lukic V, et al. Using familial information for variant filtering in high-throughput sequencing studies. Hum Genet, 2014, 133: 1331-41
[29] Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet, 2015, 16: 275-84
[30] Lai J, Li R, Xu X, et al. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet, 2010, 42: 1027-30
[31] Zheng X, Li L, Liang F, et al. Pedigree-based genome re-sequencing reveals genetic variation patterns of elite backbone varieties during modern rice improvement. Sci Rep, 2017, 7: 292
[32] Chen S, Lin Z, Zhou D, et al. Genome-wide study of an elite rice pedigree reveals a complex history of genetic architecture for breeding improvement. Sci Rep, 2017, 7: 45685
[33] Zhou D, Chen W, Lin Z, et al. Pedigree-based analysis of derivation of genome segments of an elite rice reveals key regions during its breeding. Plant Biotechnol J, 2016, 14: 638-48
[34] Xie W, Wang G, Yuan M, et al. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc Natl Acad Sci USA, 2015, 112: E5411-9
[35] Huang X, Zhao Y, Wei X, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet, 2012, 44: 32-9
[36] Wang W, Mauleon R, Hu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature, 2018, 557: 43-9
[37] Kawahara Y, de la Bastide M, Hamilton JP, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice, 2013, 6: 4
[38] Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, 2012, 6: 80-92
[39] Vergara IA, Frech C, Chen N. CooVar: co-occurring variant analyzer. BMC Res Notes, 2012, 5: 615
[40] Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods, 2010, 7: 248-9
[41] Zhang T, Marand AP, Jiang J. PlantDHS: a database for DNase I hypersensitive sites in plants. Nucleic Acids Res, 2016, 44: D1148-53