彭歆, 钱乾, 谭健韬, 等. 水稻遗传育种相关生物信息数据库和工具的研究进展 [J]. 华南农业大学学报, 2023, 44(6): 854-866.
PENG Xin, QIAN Qian, TAN Jiantao, et al. Research progress on bioinformatics databases and tools related to rice genetics and breeding[J]. Journal of South China Agricultural University, 2023, 44(6): 854-866.

特约综述

# 水稻遗传育种相关生物信息数据库和工具的研究进展

彭　歆✉，钱　乾，谭健韬，彭　波，甘玉立，王成睿，刘　琦，沈梦圆✉
(广东省农业科学院 水稻研究所/广东省水稻育种新技术重点实验室/广东省水稻工程实验室/
农业农村部华南优质稻遗传育种实验室 (部省共建), 广东 广州 510640)

**摘要:** 水稻 *Oryza sativa* L.是主要的粮食作物，也是单子叶植物研究的模式植物。面对日益严峻的环境和人口压力，培育高产、优质、环境适性强的水稻品种是解决当前粮食安全问题的有效途径。随着多组学技术的快速发展，积累了海量的水稻遗传育种相关的数据。生物信息数据库和在线分析工具是存储这些数据的载体，用以整合、可视化和共享数据，并为数据的深入挖掘和利用提供工具，从而为育种决策提供数据支撑。本综述系统梳理了近 20 年来开发的水稻生物信息数据库和在线分析工具，并基于内置数据集和功能对它们进行了分类和总结。最后，讨论了现有的水稻生物信息数据库和在线分析工具的问题与不足，并对它们在大数据和人工智能时代的发展方向进行了展望。

**关键词:** 水稻; 遗传育种; 生物信息数据库; 在线分析工具
**中图分类号:** S511; S32　　**文献标志码:** A　　**文章编号:** 1001-411X(2023)06-0854-13

# Research progress on bioinformatics databases and tools related to rice genetics and breeding

PENG Xin✉, QIAN Qian, TAN Jiantao, PENG Bo, GAN Yuli, WANG Chengrui, LIU Qi, SHEN Mengyuan✉
(Rice Research Institute, Guangdong Academy of Agricultural Sciences/ Guangdong Key Laboratory of New Technology in Rice Breeding/Guangdong Rice Engineering Laboratory/ Key Laboratory of Genetics and Breeding of High Quality Rice in Southern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Guangzhou 510642, China)

**Abstract:** Rice (*Oryza sativa* L.) is both a major staple food and a model crop plant for monocot studies. Facing the increasingly severe environmental and population problems, breeding varieties with high yield, high quality, and wide adaptability is the efficient way to solve the food security problems. With the rapid development of multi-omics technology, large volumes of data related to rice genetics and breeding have been accumulated. Bioinformatics databases and online analysis tools are developed to store, integrate, visualize, and share these datasets. In addition, some databases possess built-in tools for further mining and using datasets to provide data support for decision-making in breeding. In this review, we systematically sort out rice bioinformatics databases

**收稿日期:** 2023–07–10　　**网络首发时间:** 2023–09–12 11:23:41
**首发网址:** https://link.cnki.net/urlid/44.1110.s.20230908.1426.004
**作者简介:** 彭　歆，助理研究员，博士，主要从事水稻生物信息大数据挖掘利用和数据库的构建相关研究，E-mail: pengxin@gdaas.cn；通信作者：沈梦圆，助理研究员，博士，主要从事水稻 RNA 表观转录组学及相关数据库和软件的开发研究，E-mail: mengyuanshen@126.com；刘　琦，研究员，博士，主要从事水稻大数据育种及相关数据库和软件的开发研究，E-mail: qiliu@gdaas.cn
**基金项目:** 广东省农业科学院协同创新中心项目 (XTXM202203)；广东省农业科学院水稻研究所“优谷计划”(2023YG08)；省级乡村振兴战略专项“种业振兴项目”(2022NJS00004)；广东省水稻育种新技术重点实验室项目 (2020B1212060047)

and online analysis tools developed in the past two decades. Subsequently, we classified and summarized these resources based on their built-in datasets and features. Finally, the problems and deficiencies of the existing rice bioinformatics resources were discussed, and the development direction of bioinformatics resources in the era of big data and artificial intelligence was prospected.

**Key words:** Rice; Genetics and breeding; Bioinformatics database; Online analysis tool

水稻、小麦和玉米是全球三大粮食作物，合计约占全球粮食总产量的 87%。预计 2010—2050 年世界对主要粮食作物的需求将增加 60%[1]。水稻是单子叶植物的模式植物，也是第 1 个完成全基因组测序的谷类作物。水稻研究的主要目标是确定每个基因的功能并改善作物的农艺性状[2]。以基因组重测序技术为代表的高通量组学技术在水稻中得到广泛应用，促进了水稻基因功能研究和重要性状遗传改良[3]。

生物信息学是生物科学的一个跨学科分支，可以用来开发用于收集、处理和分析不同生物数据以了解生物功能的方法和工具[4]。在水稻基因组研究中，生物信息学可用于基因序列比对、基因组组装、基因组元件预测和群体遗传学分析等[4-5]。随着越来越多的水稻基因功能得到解析，借鉴水稻生物信息学的研究方法将有助于改善其他作物的农艺性状。此外，生物信息学与前沿基因编辑和合成生物学技术相结合，有望实现对农艺学性状的精准调控和品种的定向改良[3,6]。生物数据正在爆炸性增长，为了有效管理这些大数据，研究者已开发了超过 5 909 个数据库，涉及 1 525 个物种[7]。鉴于水稻生物信息数据库和在线分析工具的激增，对其进行系统的归纳和总结，有利于科研工作者和育种家更好地利用这些资源。

在这篇综述中，笔者系统总结了已报道的水稻相关生物信息学数据库和基于网络服务器开发的在线分析工具，并进行了分类。根据是否只包括水稻一个物种，将数据库分为水稻专门数据库和综合数据库；根据数据集的主要类型和功能，将数据库分为基因组数据库、转录和转录后调控数据库、基因网络数据库和种质资源信息数据库。此外，笔者还总结了可用于基因编辑和智能育种的在线分析工具和数据库。最后，讨论了现有的水稻生物信息数据库和在线分析平台的问题与不足，并对它们在大数据和人工智能时代的发展方向进行了展望。

## 1　水稻生物信息数据库

### 1.1　基因组数据库

高质量的参考基因组信息和准确的基因功能注释，是水稻功能基因组学研究的基础。在 21 世纪初期，为了推动稻属基因组研究的发展，Wing 等[8] 提出了 OMAP 计划 (*Oryza* map alignment project)。在这个框架下，已完成的超过 73 个稻属种质资源的参考基因组，涉及 17 个物种，包括所有的二倍体物种和 2 个异源四倍体 (CCDD 和 KKLL)。由于使用单一参考基因组会导致基因组分析出现严重的偏差，丢失大量群体尺度的遗传变异信息。泛基因组包括一个物种的核心基因和非必须基因，是解决上述问题的有效方法。Yao 等[9] 基于 1483 个栽培水稻种质的低覆盖度重测序构建了第 1 个水稻泛基因组。3KRG (3 000 Rice genomes project) 完成了 3 010 份国际水稻基因组测序，并获得了它们的线性泛基因组[10]。Zhao 等[11] 基于 67 份亚洲栽培稻和普通野生稻高深度二代测序数据，组装了亚洲栽培稻−普通野生稻的线性泛基因组。Qin 等[12] 构建了 33 个水稻品种的图形泛基因组。Shang 等[13] 构建了包括亚洲栽培稻、普通野生稻、非洲栽培稻和短舌野生稻的稻属超级泛基因组。RGI (Rice gene index) 基于同源基因簇构建了水稻泛基因组数据库，提供了丰富的模块和工具，支持对不同种质基因以及基因同源关系进行查询、分析和可视化[14]。Wang 等[15] 构建了 413 份国际籼稻的图形泛基因组。此外，已完成了超过 6 000 份水稻种质资源的基因组重测序，这些原始数据大部分储存在 SRA (Sequence read archive) 和 GSA (Genome sequence archive) 数据库。针对上述水稻参考基因组、泛基因组及海量的基因组重测序数据，已开发了多个数据库 (表 1)。根据是否只包含水稻的数据集，这些数据库可以分为综合数据库和水稻专门数据库。NCBI[16]、Ensembl[17] 和 Phytozome[18] 是比较著名的综合数据库，它们提供了大部分已测序物种的参考基因组和比较详细的基因组注释信息；RAP-DB[19] 和 MSU-RGAP[20] 数据库是 2 个著名的水稻专门数据库，为第 1 个水稻参考基因组‘日本晴’提供基因组注释资源；基于籼稻参考基因组‘珍汕 97’和‘明恢 63’，Song等[21] 开发了 RIGW 数据库，包括了基因组学、转录组学和蛋白质−蛋白质相互作用数据集；IC4R 提供了一个专门用于整合水

表 1 已发表的水稻基因组数据库

Table 1 The published rice genomic databases

| 数据库 Database | 描述 Description | 参考文献 Reference |
|---|---|---|
| NCBI | 综合数据库、稻属16个物种参考基因组、基因组重测序数据，https://www.ncbi.nlm.nih.gov/ | [16] |
| Ensembl | 综合数据库、稻属10个物种参考基因组、基因组注释，http://plants.ensembl.org/ | [17] |
| Phytozome | 综合数据库、‘日本晴’和‘Kitaake’参考基因组、基因组注释，https://phytozome-next.jgi.doe.gov/ | [18] |
| RAP-DB | ‘日本晴’、IRGSP-1.0参考基因组、基因组注释，http://rapdb.dna.affrc.go.jp/ | [19] |
| MSU-RGAP | ‘日本晴’、MSU7.0参考基因组、基因组注释，http://rice.uga.edu/ | [20] |
| RIGW | ‘珍汕97’和‘明恢63’参考基因组、多组学数据、互作数据，http://rice.hzau.edu.cn/rice_rs3/ | [21] |
| IC4R | 参考基因组、基因组注释、基因表达谱，http://ic4r.org/ | [22] |
| Rice Genome Hub | 稻属10个物种的参考基因组(32个基因组信息)，https://rice-genome-hub.southgreen.fr/ | [23] |
| RPAN | 3KRG线性泛基因组、泛基因组浏览器，http://cgm.sjtu.edu.cn/3kricedb/ | [10] |
| RicePanGenome | 线性泛基因组、基因组变异、67个参考基因组，http://db.ncgr.ac.cn/RicePanGenome/ | [11] |
| RiceRc | 图形泛基因组、33个参考基因组，http://ricerc.sicau.edu.cn/ | [12] |
| RiceSuperPIRdb | 图形泛基因组、251个参考基因组，http://www.ricesuperpir.com/ | [13] |
| RGI | 基于同源基因簇的水稻泛基因组、16个水稻参考基因组，https://riceome.hzau.edu.cn | [14] |
| OryzaGenome | 稻属参考基因组，208个种质基因组信息，涉及19个野生稻和2个栽培稻物种，http://viewer.shigen.info/oryzagenome2detail/ | [24] |
| RiceRelativesGD | 水稻17个近缘物种基因组和单倍型信息，http://ibi.zju.edu.cn/ricerelativesgd/ | [25] |
| funRiceGenes | 基因功能数据库、IRGSP-1.0和MSU7.0基因注释，http://funricegenes.ncpgr.cn/ | [26] |

稻基因组序列、基因注释和表达水平信息的平台[22]；RPAN[10] 和 Rice-PanGenome[11] 是 2 个为水稻泛基因组分析提供资源和工具的数据库；Rice Genome Hub 提供了稻属 10 个物种的参考基因组信息，并提供了基因组浏览、基因检索和共线性分析等在线工具[23]；OryzaGenome[24] 和 RiceRelativesGD[25] 包含了水稻近缘物种的基因和种质资源信息；funRiceGenes[26] 是一个专门提供水稻中已克隆基因信息的数据库。综上所述，基因组数据库是水稻所有数据库的基础，也是资源最丰富和内容最完整的一类数据库。然而，由于数据可视化和信息检索方式的差异，容易给用户的使用造成困扰。因此，需要开发一个整合了所有的参考基因组和基因组重测序资源的水稻基因组综合数据库，从而更加规范管理和使用现有的数据资源。

### 1.2 转录和转录后调控数据库

通过微阵列或转录组测序，获得不同组织或细胞在各个发育阶段和处理下的基因表达信息，加深了对基因时空表达模型的理解，有利于基因功能的研究[27]。研究者们利用统一的标准流程，将海量的基因表达数据整合在一起，构建了一系列水稻基因表达数据库，包括 RiceXPro、CREP、RED、TENOR、eRice 和 RiceENCODE 等（表 2）。RiceXPro 是基于微阵列表达数据开发的基因转录数据库，提供不同组织或细胞在各个生长发育阶段、逆境胁迫及植物激素处理条件下的基因表达信息[27]；CREP 包含了‘珍汕 97’和‘明恢 63’不同组织的微阵列数据集 (Microarray data)，支持通过基因序列、基因名称或探针标识来检索基因表达信息[28]；RED 整合了水稻 9 个组织的 mRNA-seq 数据，提供了不同组织和胁迫条件下的基因表达谱[29]；TENOR 包含了‘日本晴’在不同环境胁迫和激素处理条件下的 mRNA-seq 数据集[30]；PPRD 收集了 2021 年之前发表的 11 726 个水稻 mRNA-seq 数据集，使用统一流程和最新的参考基因组进行分析和整合[31]；eRice 包含了‘日本晴’和‘9311’的 mRNA-seq 数据集，并整合了 DNA 甲基化和组蛋白修饰数据集[32]；RiceENCODE 整合了包括 ChIA-PET、Hi-C 和 mRNA-seq 等在内的 972 个数据集，用于检索调控 RNA 转录的 DNA 修饰、组蛋白修饰和染色质构象等表观调控[33]。非编码 RNA (ncRNA，包括 miRNA、lncRNA、siRNA、circRNA 等) 参与对基因的转录后调控，从而影响植物的生长发育和对环境的响应过程。水稻专门非编码 RNA 数据库有 RiceNCexp、ARMOUR、RiceLncPedia 和 RiceATM(表 2)。RiceNCexp 基于 mRNA-seq 和

**表 2　水稻转录和转录后调控相关数据库**

**Table 2　The transcriptional and posttranscriptional regulation related databases in rice**

| 数据库<br>Database | 描述<br>Description | 参考文献<br>Reference |
|---|---|---|
| RiceXPro | 微阵列数据集，自然条件下各个生长发育阶段、幼苗激素和胁迫处理的基因表达信息，https://ricexpro.dna.affrc.go.jp/ | [27] |
| CREP | ‘珍汕97’和‘明恢63’的39个组织的基因表达信息，http://crep.ncpgr.cn/crep-cgi/home.pl | [28] |
| RED | 水稻9个组织，在不同生长阶段和处理的基因表达谱和基因共表达网络，http://expression.ic4r.org | [29] |
| TENOR | 包括‘日本晴’在不同环境胁迫和激素处理条件下的140个mRNA-seq数据集，https://tenor.dna.affrc.go.jp/ | [30] |
| PPRD | 11 726个水稻mRNA-seq数据集，使用统一流程和最新的参考基因组进行分析和整合，http://ipf.sustech.edu.cn/pub/ricerna/ | [31] |
| eRice | ‘日本晴’和‘9311’的mRNA-seq、DNA甲基化和组蛋白修饰数据库，http://www.elabcaas.cn/rice/index.html | [32] |
| RiceENCODE | 综合调控RNA转录的DNA修饰、组蛋白修饰、染色质构象等表观调控元件，http://glab.hzau.edu.cn/RiceENCODE/ | [33] |
| RiceNCexp | 提供基于mRNA-seq和sRNA-seq的基因和sRNA转录水平和共表达网络信息，https://cbi.njau.edu.cn/RiceNCexp/ | [34] |
| ARMOUR | 7个水稻品种在不同发育时期、组织和胁迫下的miRNA和相应的靶标信息，https://www.icgeb.org/armour.html | [35] |
| RiceLncPedia | 包含了水稻ncRNAs的表达谱、变异位点、ncRNA之间和ncRNA与编码基因的共表达网络信息，http://3dgenome.hzau.edu.cn/RiceLncPedia | [36] |
| RiceATM | 挖掘miRNA与水稻农艺性状的关系，包括表型选择、样本分组、微阵列数据预处理、统计分析和靶基因预测等功能，http://syslab3.nchu.edu.tw/rice/ | [37] |
| CSRDB | 整合了水稻和玉米的sRNA和它们的靶基因信息，http://sundarlab.ucdavis.edu/smrnas/ | [38] |
| miRbase | 包含水稻已知和新的miRNA的序列和前体序列信息，是使用最广泛的miRNA综合数据库，http://mirbase.org/ | [39] |
| PceRBase | 包含水稻等26个物种的ceRNA、miRNA和它们的靶基因信息，http://bis.zju.edu.cn/pcernadb/index.jsp | [40] |
| GreeNC 2.0 | 水稻lncRNA数据库，http://greenc.sequentiabiotech.com/wiki2/Main_Page | [41] |
| PLncDB | 提供lncRNA的长度、类型、表达谱和表观遗传等信息，http://plncdb.tobaccodb.org/ | [42] |
| CANTATAdb | 提供lncRNA长度、类型和表达谱信息，http://cantata.amu.edu.pl，http://yeti.amu.edu.pl/CANTATA/ | [43] |
| PmiREN | 包含水稻miRNA及其前体序列、二级结构、表达模式、潜在靶点等信息，http://www.pmiren.com/ | [44] |
| PlantcircBase | 提供水稻circRNA的分类、表达谱信息，http://ibi.zju.edu.cn/plantcircbase/ | [45] |
| PlaASDB | 水稻和拟南芥在非生物和生物胁迫下的AS事件及AS与基因表达之间的联系，http://zzdlab.com/PlaASDB/ASDB/index.html | [46] |
| PlantAPAdb | 水稻和拟南芥等6个物种基因组范围内的APA位点及注释信息，http://www.bmibig.cn/plantAPAdb | [47] |

small RNA-seq (sRNA-seq)，对基因和 sRNA 的转录水平和共表达网络进行了分析，并为用户提供了个性化分析工具[34]；ARMOUR 整合了 7 个水稻品种在不同发育时期、组织和胁迫下的 miRNA 和它们的靶标基因信息[35]；RiceLncPedia 基于2312 个共公数据库获取的 RNA-seq 数据集，提供了水稻 lncRNAs 的表达谱、变异位点、lncRNA 之间和 lncRNA 与编码基因的共表达信息[36]；RiceATM 是用来挖掘 miRNA 与水稻农艺性状关系的数据库，包括 187 个水稻品种的 8 个产量相关性状数据集和 193 个 miRNAs 在这些品种中的表达水平和靶基因数据集[37]。一些植物综合 ncRNA 数据库，也包含比较丰富的水稻 ncRNA 信息。CSRDB 是较早开发的 sRNA 数据库，整合了水稻和玉米的 sRNA 和

它们的靶基因信息[38]；miRbase 是广泛使用的 miRNA 数据库，包括已知的水稻 miRNA 的成熟序列、前体序列和靶基因信息[39]；PceRBase 整合了包含水稻在内的 26 个植物物种的竞争性内源 RNA (Competing endogenous RNA，ceRNA) 数据集，为 miRNA−靶基因调控网络的研究起到了补充作用[40]。此外，GreeNC[41]、PLncDB[42]、CANTATAdb[42-43]、PmiREN[44] 和 PlantcircBase[45] 等多物种数据库中也包含水稻 lncRNA、circRNA 和 miRNA 信息。

选择性多聚腺苷酸化 (Alternative cleavage and polyadenylation, APA) 是指具有多个多聚腺苷酸化信号位点 (Polyadenylation signal, PAS) 的基因在 mRNA 加工过程中，由于可变剪切 (Alternative splicing，AS) 和 APA 调控过程，形成不同的 mRNA 异构体的现象[48]。PlaASDB 综合分析了水稻和拟南芥在生物和非生物胁迫条件下的 AS 事件，并将 AS 与差异表达基因进行了关联[46]。PlantAPAdb 基于 3′-seq(3′ sequencing)，在基因组范围内详细鉴定了水稻和拟南芥等 6 个物种的 APA 位点，并提供了多样化的注释信息，包括 APA 在基因组位置、异质性切割位点、表达水平和样本信息等[47]。综上所述，转录和转录后调控过程中涉及的数据库，主要集中在基因表达谱和 ncRNA 方面，有关 AS、APA 和 RNA 修饰等转录后调控的数据库资源仍然较少。因此，亟需构建水稻 RNA 修饰等转录后调控数据库，为挖掘水稻优异性状和转录后调控之间的潜在联系提供数据基础。

### 1.3 基因网络数据库

在植物中，生物网络广泛用于功能基因的挖掘、基因功能和重要农艺性状遗传机制解析的研究中[49]。生物网络是以生物分子作为节点，分子之间的相互作用作为连接边构成的系统，比较常见的有蛋白质互作网络、基因调控网络、共表达网络、代谢网络等[50]。在植物中，基因网络推断主要依赖于表达数据，使用相关性分析和互信息算法推断 2 个基因之间的潜在关系[51]。近年来，利用机器学习算法，如图卷积神经网络 (Graph convolutional network, GCN)，能够有效降低网络的冗余，得到节点数尽可能少，并能最大化地表示原始图的子图，从而更真实地反映基因之间的互作网络[52]。目前已开发的水稻基因网络数据库主要是基于转录组和代谢组数据集构建的，涉及的网络推断方法主要有相关性分析、互信息算法、同源映射和机器学习 (表 3)。RiceFREND 和 OryzaExpresss 是较早建立的基因网络数据库，二者均是基于微阵列数据集构建的基因共表达网络，并提供了比较友好的网络交互界面和灵活的数据检索和查询功能[53-54]；RiceAntherNet 提供了水稻花粉和花药发育过程的基因网络信息，包括 57 个水稻花药组织微阵列数据集，有助于挖掘花药和花粉调控网络中的关键基因[55]；NetREx 是基于已发表的 mRNA-seq 数据集构建的不同胁迫条件下水稻基因共表达网络数据库，可以实现胁迫或组织特异性共表达模型的查看，并关联了基因功能注释和基因通路等数据集[56]。此外，NetREx 还提供了水稻、拟南芥、小麦、玉米、大麦和高粱的直系同源基因信息；PRIN 是基于同源映射法，通过对果蝇、线虫和拟南芥等模式物种蛋白互作数据进行同源推断获得的第 1 个水稻蛋白质互作网络数据库[57]；RiceNet 和 RicePPINet 是利用机器学习构建的基因网络数据库[58-59]。其中，RiceNet 基于多组学

**表 3 已发表的水稻基因网络数据库**

**Table 3 The publised gene network databases in rice**

| 数据库 Database | 描述 Description | 参考文献 Reference |
|---|---|---|
| RiceFREND | 基于不同组织不同生长发育阶段的微阵列数据构建的共表达网络，http://ricefrend.dna.affrc.go.jp/ | [53] |
| OryzaExpress | 基于微阵列数据构建的共表达网络，http://plantomics.mind.meiji.ac.jp/OryzaExpress/ | [54] |
| RiceAntherNet | 基于微阵列数据集构建的花粉和花药发育过程的共表达网络，https://www.cpib.ac.uk/anther/riceindex.html | [55] |
| NetREx | 基于同源映射和转录组数据集构建的基因在逆境和激素处理下的共表达网络，https://bioinf.iiit.ac.in/netrex/index.html | [56] |
| PRIN | 基于模式物种蛋白互作基因同源映射构建的水稻蛋白质互作网络，http://bis.zju.edu.cn/prin/ | [57] |
| RiceNetv2 | 基于蛋白互作、mRNA-seq等多种数据集，利用机器学习算法构建的基因网络，http://www.inetbio.org/ricenet | [58] |
| RicePPINet | 基于机器学习算法构建的蛋白质互作网络，http://netbio.sjtu.edu.cn/riceppinet | [59] |

数据集，利用贝叶斯框架、网络直接邻域法和机器学习整合微阵列、转录组、蛋白互作和不同物种的同源映射等数据集，构建了一个由 25 765 个基因结点和 1 775 000 个连接边组成的基因网络[58]。RicePPINet 包含了 16 895 个蛋白质的 708 819 个基于机器学习算法预测的蛋白质−蛋白质相互作用网络，并通过与已知蛋白互作信息和试验证明了预测结果的可靠性[59]。

在植物中，基因网络推断主要依赖于表达数据，使用相关性分析和互信息算法推断两个基因之间的潜在调控关系。严格来说，基于表达数据推断出的基因与基因之间的联系并不一定是客观存在的。机器学习算法可以用来整合不同类型的基因网络数据集，如基因共表达网络、蛋白互作网络和代谢网络，尽可能准确地推断基因之间的关系。此外，利用 GCN 算法可以降低网络的冗余，得到节点数尽可能少并能最大化表示原始图的子图。综上所述，基于多组学数据和机器学习算法构建的基因网络，将更加准确地反映基因之间的相互作用关系。

### 1.4　种质资源信息数据库

随着大规模种质资源测序项目的开展，积累了海量的数据集。研究者通过构建种质资源信息数据库整合和共享这些数据集，并为数据的进一步挖掘和利用提供了在线分析工具。在水稻中，已有多个种质资源信息数据库，如 MBKBASE、RiceVarMap 和 SNP-Seek，它们为水稻分子育种提供了丰富的基因组变异、单倍型以及变异的遗传力等信息 (表 4)。MBKBASE 收录了 130 578 份水稻种质资源信息，并以‘R498’和‘日本晴’为参考基因组，进行了单倍型鉴定，并提供了单倍型与性状和基因表达水平的关联信息[60]；RiceVarMap 收录了全球 140 个国家和地区的 4 726 份水稻种质资源及其遗传变异信息，包括 14 541 446 个 SNP 和 2 855 580 个 InDel[61]；SNP-Seek 包含了来源于 119 个国家的4 036 份水稻品种资源信息，包括种质资源群体结构、遗传变异和 57 个表型等数据集[62]；SR4R 提供了 5 152 份水稻资源的变异图谱，通过统一的流程整合遗传变异数据，有效地减少了 SNP 冗余[63]；HapRice 包括 76 份世界水稻品种的 3 334 个 SNP 和 177 份日本水稻品种的 3 252 个 SNP，并构建了 SNP 单倍型图谱[64]；RFGB 基于 3KRG 数据集，整合了表型、单倍型、SNP 和 InDel 等信息[65]；中国水稻品种及其系谱数据库是国家水稻数据中心 (http://www.ricedata.cn/) 的一个子平台，主要收录省级以上审定品种、大面积推广品种、外引品种以及地方性农家品种，具有品种检索和系谱追溯 2 个核心功能[66]。转座子是重复序列的一种，也是一种基因组变异，在真核生物中广泛存在。水稻种质资源转座子注释数据库有 RiTE DB[67] 和 RTRIP[68]。RiTE DB 是第 1 个水稻转座子数据库，包括从 266 份水稻品种中鉴定到的 54 911 个转座子信息[67]；RTRIP 包含基于 3KRG 的基因组重测序数据构建的转座子序列位点

**表 4　已发表的水稻种质资源信息数据库**

**Table 4　The published germplasm information resources in rice**

| 数据库 Database | 描述 Description | 参考文献 Reference |
|---|---|---|
| MBKBASE | 以‘R498’和‘日本晴’为参考基因组，包含130 578份种质的表型、群体结构和单倍型信息，https://www.mbkbase.org/rice/ | [60] |
| RiceVarMap | 4 726份水稻种质资源的基因组变异、基因型和表型数据，http://ricevarmap.ncpgr.cn/v2/ | [61] |
| SNP-Seek | 4 036份水稻的SNP图谱、表型和全基因组关联分析数据集，https://snp-seek.irri.org/ | [62] |
| SR4R | 包含5 152份种质资源的遗传变异、群体遗传学和进化基因组学数据集，http://sr4r.ic4r.org/ | [63] |
| HapRice | 253份国际和日本来源的水稻种质的SNP和单倍型图谱，http://qtaro.abr.affrc.go.jp/index.html | [64] |
| RFGB | 3KRG的表型、遗传变异和基因单倍型信息，http://www.rmbreeding.cn/ | [65] |
| RiceData | 包含省级以上审定品种、大面积推广品种、外引品种以及地方性农家品种信息，http://www.ricedata.cn/variety/ | [66] |
| RiTE DB | 包含266份水稻品种中鉴定到的54 911个转座子信息数据集，https://www.genome.arizona.edu/cgi-bin/rite/index.cgi | [67] |
| RTRIP | 3KRG的基因组转座子信息，提供转座子序列位点图谱、遗传多样性、基因组进化和分子标记等信息，http://ibi.zju.edu.cn/Rtrip/index.html | [68] |

图谱[68]。

在作物驯化和遗传改良过程中，遗传多样性不断变窄。国家/国际基因库 (种质库) 提供了丰富来源的等位基因，这些等位基因可能对作物遗传改良至关重要。因此，在基因组和表型组水平上对尽可能多的种质资源进行鉴定，为确定等位基因与表型的联系提供了信息，从而有助于育种决策。综上所述，对基因库中存在的全部种质进行测序和表型鉴定，获得所有种质的遗传变异、等位基因以及它们与表型的关联信息，构建水稻种质资源综合数据库是未来育种计划的重要组成部分。

## 2 水稻基因编辑生物信息工具

### 2.1 基因编辑系统概述

基因编辑能在不引入 DNA 双链断裂 (Double-strand DNA breaks, DSBs) 的情况下，对靶位点一定范围内的碱基进行编辑。CRISPR/Cas 系统作为第 3 代基因组编辑技术，具有操作简单灵活、精准靶向和效率高等优点，在基础研究和生物育种中具有广泛的应用前景[69-70]。

如表 5 所示，主要的基因编辑系统包括胞嘧啶碱基编辑器 (Cytosine base editor, CBE)、腺嘌呤碱基编辑器 (Adenine base editor, ABE)、胞嘧啶−鸟嘌呤碱基编辑器 (Cytosine-to-guanine base editor, CGBE)、引导编辑器 (Prime editor, PE) 等，已广泛应用在人类、动植物和微生物的相关研究中[71]。近年来，研究者将 Cas9 蛋白和反转录酶蛋白 (M-MLV) 融合，以及将 sgRNA 改造成可提供修复模板的 pegRNA，构建了 PE 系统，实现了 A、T、C、G 碱基的自由替换和小 DNA 片段的替换或插入，可编辑远离 PAM(Protospacer adjacent motif) 的序列，显著提高了基因编辑的精准度[72]。

**表 5 常用的基因编辑系统**

**Table 5 Commonly used gene editing systems**

| 编辑系统 Editing system | 编辑作用 Editing function | 用途 Application |
|---|---|---|
| CRISPR/Cas | 靶点切割/突变 | 基因敲除、遗传改良 |
| CRISPR/Cas/Donor | DNA片段插入/替换 | DNA片段突变、遗传改良 |
| CRISPR/Cas-CBE | 单碱基转换：C > T(G > A) | 单碱基转换、遗传改良 |
| CRISPR/Cas-ABE | 单碱基转换：A > G(T > C) | 单碱基转换、遗传改良 |
| CRISPR/Cas-CGBE | 单碱基转换：C > G(G > C) | 单碱基转换、遗传改良 |
| CRISPR/Prime editors | 小片段插入、替换 | 单碱基的任意转换和小DNA片段的替换或插入、遗传改良 |

### 2.2 基因编辑靶标设计生物信息工具

基因编辑中目标基因靶位点的选择、sgRNA (Single guide RNA) 表达盒的设计和组装、靶位点片段的扩增与测序、脱靶 (Off-target) 及自靶向 (Self-target) 分析等过程仍会耗费大量的人力物力。近年来，利用生物信息工具辅助基因编辑的设计和分析流程，可在减少人为错误的同时也极大地节省时间和人力成本，使基因编辑高通量化[73-75]。

表 6 列出了已开发的水稻基因编辑数据库或在线工具。CRISPR-GE 提供了“一站式”基因组编辑便捷工具包，包含了靶点筛选、引物设计、结果鉴定、脱靶分析等流程的在线分析软件[73]; CRISPR-GE 持续更新，在后续的开发中添加了利用微同源末端连接介导基因组大片段删除的靶点设计工具 MMEJ-KO、辅助碱基编辑靶点选择工具 BEtarget、基于多个高质量水稻基因组整合的数据库 GeneCat、以及基于二代测序方法高通量分析靶位点编辑情况的 HiDecode[73-74,76-78]; 此外，CRISPR REGN 提供基因编辑在线工具包，如 sgRNA 设计工具 Cas-Designers[79]、脱靶预测工具 Cas-Offinder[80] 和预测 MMEJ(Microhomology-mediated end joining) 介导片段删除的效率分析工具 Microhomology-Predictor 等[75]; CRISPR-P 支持 49 个物种基因组的 sgRNA 设计，开发了靶向效率和脱靶概率评估系统，使预测结果更为准确，并支持多种 CRISPR/Cas 系统的设计[81-82]; CRISPRbase收集了包括水稻在内的 17 个物种多达 120 万条碱基编辑信息，整理分析了多个编辑系统在不同物种中的编辑效率、靶点偏好性和精准度，并提供了在线分析工具[83]; PGED 收集了水稻等 8 个物种的 CRISPR/Cas9 编辑植株信息，包括 sgRNA 序列、序列变异和表型等信息，用户可以通过物种或基因名查找已有的 CRISPR/Cas9 突

表 6　水稻基因编辑生物信息工具与数据库

Table 6　The bioinformatics tools and databases for gene editing in rice

| 数据库 Database | 描述 Description | 参考文献 Reference |
|---|---|---|
| CRISPR-GE | “一站式”基因编辑设计工具包，涵盖靶点筛选、引物设计、编辑结果鉴定、脱靶分析等流程，http://skl.scau.edu.cn/ | [73-74] |
| MMEJ预测工具 | 评估MMEJ介导片段删除效率，http://www.rgenome.net/ | [75] |
| BEtarget | 提供候选靶点在基因上的位置、GC含量、潜在脱靶值和脱靶位点、编辑窗口内的碱基变化及对应的氨基酸变化等信息，http://skl.scau.edu.cn/betarget/ | [76] |
| MMEJ-KO | 基于微同源删除基因组片段的靶点设计工具，http://skl.scau.edu.cn/mmejko/ | [77] |
| GeneCat | 快速提取基因组序列工具，http://skl.scau.edu.cn/genecat/ | [78] |
| Cas-Designers | sgRNA设计线上工具，http://www.rgenome.net/cas-designer/ | [79] |
| Cas-Offinder | sgRNA脱靶预测线上工具，http://www.rgenome.net/cas-offinder/ | [80] |
| CRISPR-P | 多功能的sgRNA设计工具，随后升级为CRISPR-P 2.0，支持49个物种基因组的sgRNA设计，http://crispr.hzau.edu.cn/CRISPR2/ | [81-82] |
| CRISPRbase | 碱基编辑综合知识平台，统计了多个物种的编辑效率、靶点偏好性和精准度，并进行功能注释，http://crisprbase.maolab.org/ | [83] |
| PGED | 植物基因组编辑数据库，可供用户查阅或上存基因编辑靶位点、突变情况、表型信息等，http://plantcrispr.org | [84] |
| CAFRI-Rice | 水稻冗余基因数据库，为基因编辑候选靶标的选择提供参考，http://pcafri-rice.khu.ac.kr | [85] |
| Ribo-uORF | uORF综合数据库，收集了6个物种的高可信度uORF和TIS位点信息，为uORF编辑提供靶标，http://rnainformatics.org.cn/RiboUORF | [86] |

变体[84]。

CAFRI-Rice 基于系统发育分析和基因表达模式，建立了水稻冗余基因数据库，为基因编辑候选靶标的选择提供了参考[85]。uORF (Upstream open reading frame, uORF) 定义为位于 5′UTR 上游的开放读码框，对主开放读码框 (Primary ORF, pORF) 的翻译具有抑制作用[86]。通过对 uORF 区域进行基因编辑，能够实现对目标蛋白表达量的精准调控[87]。Liu 等[86] 基于深度测序 Ribo-seq(Ribosome profiling) 和 QTI-seq(Quantitative translation initiation sequencing)，构建了 Ribo-uORF 数据平台，包含了 6 个物种的 501 554 条高可信度的 uORF和 107 914 条翻译起始位点 (Translation initiation site, TIS) 信息。Ribo-uORF 提供了丰富的 uORF 信息，为通过基因编辑精准调控蛋白质的表达水平提供了候选靶标[86]。

综上所述，生物信息工具辅助基因编辑可以有效提高编辑精准度 (降低潜在脱靶风险)、预测编辑效果和提供新型编辑靶点。然而，全基因组测序结果发现了编辑植株中存在许多与序列相似度无关的脱靶位点。开发生物信息工具，对与序列相似度无关的脱靶事件进行预测，有助于进一步提高编辑精准度。此外，结合 uORF、dORF(Downstream ORF) 和 $m^6A$ 修饰等数据库，将有助于指导新型编辑策略的开发与应用。

## 3　水稻智能育种生物信息工具

随着人类社会步入互联网、大数据和人工智能“三位一体”时代，对育种提出了革命性理念，即强调生命科学、信息科学与育种科学的深度融合[2-3,88]。在智能育种阶段，育种专家将综合多层面生物技术与信息技术推动育种向着智能化的方向快速发展：1) 利用高通量低成本的基因组测序技术和表型鉴定技术，结合人工智能图像识别技术，通过基因型与表型数据的自动化获取与解析，实现组学大数据的快速积累；2) 利用生物信息学与机器学习的方法和手段，整合遗传变异和各类组学数据、杂交育种数据，实现控制作物性状关键调控基因的快速挖掘与表型的精准预测；3) 利用基因编辑与合成生物学技术，通过人工改造基因元器件与人工合成基因网络，实现作物具备新的抗逆、高效和优良生物学性状；4) 利用作物组学大数据与人工智能技术，建立机器学习预测模型，在全基因组层面上建立智能组合优良等位基因的自然变异、人工变异、数量性状位点的育种设计方案，实现智能、高效、定向培育新品种。

由于海量的植物育种相关的大数据必须借助于人工智能进行数据分析、处理和预测，机器学习

模型在智能育种中起重要作用。表 7 列出了基于机器学习进行组学数据处理和辅助育种决策的软件和算法，可分为线性和非线性的方法。线性模型以基于混合线性模型的方法为代表，如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)、贝叶斯算法 (BayesA/B) 等；非线性模型主要以随机森林 (Random forest, RF)、支持向量机 (Support vector machine, SVM) 和神经网络 (Neural network) 为主。目前存在 3 类基于不同的统计模型的智能育种工具：第 1 类是免费开源的包/库，如 rrBLUP 和 sommer 等 R 包，以及基于 Python 的机器学习模块，如 sklearn 模块中的 RF 和 SVM 算法；第 2 类是成熟的遗传评估软件，主要用于动物育种，但也有少量功能适用于植物育种，如 JWAS；第 3 类是基于网页或图形界面工具，例如 solGS 和 IPAT 等[106]。在智能育种设计方案方面，Xu 等[107]提出了一种名为 iGEP 的智能化育种方案，综合运用多组学信息、大数据技术和人工智能算法，通过利用从多个来源获取的数据 (包括时空组学，基因组学、表型组学和环境组学)，建立智能育种机器学习模型，并通过合成生物学的方法从头培育新品种，实现理想作物的智能化育种。

**表 7 可用于水稻智能育种的机器学习软件和算法**

**Table 7 The machine learning software and algorithms for intelligent breeding in rice**

| 软件 Software | 模型 Model | 描述 Description | 参考文献 Reference |
|---|---|---|---|
| BGLR | BL、BR、BayesA/B | 基于基因–环境互作和多性状的基因组选择模型的构建，https://github.com/gdlc/BGLR-R | [89] |
| BRNN | brnn | 基于双向循环神经网络进行基因组选择和表型预测，https://cran.r-project.org/web/packages/brnn/ | [90] |
| BWGS | BayesA/B、BL、BRR | 基于R语言开发进行基因组选择和表型预测，https://cran.r-project.org/web/packages/BWGS/ | [91] |
| CropGBM | LightGBM | 基因型和表型数据预处理、群体结构分析、SNP 特征选择、表型预测和数据可视化，https://github.com/YuetongXU/CropGBM | [92] |
| DeepGS | CNN | 基于深度学习整合多组学数据进行基因组选择，https://github.com/cma2015/DeepGS/ | [93] |
| DNNGP | DNN | 基于深度神经网络整合多组学数据进行基因组选择，http://github.com/AIBreeding/DNNGP/ | [94] |
| HIBLUP | BLUP | 利用谱系、基因组和表型信息，评估个体的遗传价值，https://hiblup.github.io/ | [95] |
| KAML | KAML、GBLUP | 控制质量性状和数量性状的关键基因挖掘，https://github.com/YinLiLin/KAML | [96] |
| PopVar | RRBLUP、 BayesA/B/C、BL、BRR | 利用基因型和表型数据预测双亲后代的遗传方差和表型值，https://cran.r-project.org/web/packages/PopVar/ | [97] |
| rrBLUP | RRBLUP | 基因组分子标记遗传效应估计与表型预测，https://cran.r-project.org/web/packages/rrBLUP/ | [98] |
| sommer | GBLUP、RRBLUP | 加性效应、显性效应、上位性效应评估和遗传力的计算，https://cran.r-project.org/web/packages/sommer/ | [99] |
| STGS | ANN、BLUP、LASSO、RF、RR、SVM | 基于分子标记对单一性状进行基因组选择，https://cran.r-project.org/web/packages/STGS/ | [100] |
| GCTA | BLUP | SNP 遗传力评估和全基因组关联分析，http://cnsgenomics.com/software/gcta/ | [101] |
| JWAS | Bayes | 基因组选择和全基因组关联分析，http://reworkhow.github.io/JWAS.jl/latest/ | [102] |
| PIBULP | BLUP | 遗传参数评估与育种值估计，https://github.com/huiminkang/PIBLUP | [103] |
| solGS | RRBLUP | 基于组学数据进行复杂性状表型预测，http://cassavabase.org/solgs | [104] |
| IPAT | GBLUP、RRBLUP、BayesB | 全基因组关联分析与育种值估计的在线图形化界面工具，http://poissonfish.github.io/iPat/index.html | [105] |

智能育种技术具有广阔的应用前景，是未来水稻育种的发展方向之一。目前智能育种发展仍处于起步阶段，面临多组学大数据积累不足、基因分型成本较高、基因组预测育种模型原创性不足、种质资源较为分散等众多挑战[108]。因此，制定多维度数据采集、分析、存储与管理的标准与规范，协同建立通用的育种大数据平台，实现育种信息的充分共享与利用，是现阶段促进生物信息学在智能育种中应用的首要目标。

## 4　总结与展望

自从第 1 个水稻‘日本晴’参考基因组公布以来，生物信息学在组学数据分析与整合、基因挖掘、基因网络推断和重要性状遗传解析等方面起到了关键性作用。基于海量的数据集，研究者们开发了许多种类丰富的生物信息数据库和在线工具。由于数据来源、试验设计和分析方法的不同，不同数据库资源很难整合。因此，需要系统整合不同类型数据库，开发大型水稻综合数据库平台，以供不具有高级生物信息学技能的生物学家有效使用。此外，如何将多组学数据与表型数据进行整合，指导基因组选择和全基因组设计育种，提高水稻产量、品质和环境适应性，是从事相关科研工作的研究者需要考虑的问题之一。在数据库类型方面，涉及基因转录后调控、基因编辑和智能育种的资源仍然缺乏。此外，修饰组和翻译组，如 RNA 甲基化、蛋白质泛素化和 uORF 等可以实现基因表达水平的精准调控，开发相应的数据库和分析工具，将有利于分子设计育种和品种的定向改良。最后，鉴于大部分数据库更新周期较长或没有更新，应用机器学习的方法，如文本挖掘、随机森林和图卷积神经网络，对最新发布的数据集进行自动下载，并整合到后台数据库，将有望实现数据库的实时动态更新。

**参考文献：**


[1] VAN ITTERSUM M K. Crop yields and global food security. Will yield increase continue to feed the world?[J]. European Review of Agricultural Economics, 2016, 43(1): 191-192.

[2] 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(10): 1356-1365.

[3] WALLACE J G, RODGERS-MELNICK E, BUCKLER E S. On the road to breeding 4.0: Unraveling the good, the bad, and the boring of crop quantitative genomics[J]. Annual Review of Genetics, 2018, 52: 421-444.

[4] JIA L, XIE L J, LAO S T, et al. Rice bioinformatics in the genomic era: Status and perspectives[J]. The Crop Journal, 2021, 9(3): 609-621.

[5] 彭歆, 罗立新, 张力, 等. 重离子诱发的 2 个水稻突变体表型鉴定及遗传分析[J]. 华南农业大学学报, 2018, 39(1): 12-17.

[6] 程式华. 中国水稻育种百年发展与展望[J]. 中国稻米, 2021, 27(4): 1-6.

[7] WREN J D, GEORGESCU C, GILES C B, et al. Use it or lose it: Citations predict the continued online availability of published bioinformatics resources[J]. Nucleic Acids Research, 2017, 45(7): 3627-3633.

[8] WING R A, AMMIRAJU J S S, LUO M, et al. The *Oryza* map alignment project: The golden path to unlocking the genetic potential of wild rice species[J]. Plant Molecular Biology, 2005, 59(1): 53-62.

[9] YAO W, LI G, ZHAO H, et al. Exploring the rice dispensable genome using a metagenome-like assembly strategy[J]. Genome Biology, 2015, 16: 187.

[10] SUN C, HU Z Q, ZHENG T Q, et al. RPAN: Rice pan-genome browser for ～3000 rice genomes[J]. Nucleic Acids Research, 2017, 45(2): 597-605.

[11] ZHAO Q, FENG Q, LU H, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice[J]. Nature Genetics, 2018, 50(2): 278-284.

[12] QIN P, LU H W, DU H L, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations[J]. Cell, 2021, 184(13): 3542-3558.

[13] SHANG L, LI X, HE H, et al. A super pan-genomic landscape of rice[J]. Cell Research, 2022, 32(10): 878-896.

[14] YU Z, CHEN Y, ZHOU Y, et al. Rice Gene Index: A comprehensive pan-genome database for comparative and functional genomics of Asian rice[J]. Molecular Plant, 2023, 16(5): 798-801.

[15] WANG J, YANG W, ZHANG S, et al. A pangenome analysis pipeline provides insights into functional gene identification in rice[J]. Genome Biology, 2023, 24(1): 19.

[16] PRUITT K D, TATUSOVA T, MAGLOTT D R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins[J]. Nucleic Acids Research, 2007, 35(Suppl_1): D61-D65.

[17] HUBBARD T, BARKER D, BIRNEY E, et al. The Ensembl genome database project[J]. Nucleic Acids Research, 2002, 30(1): 38-41.

[18] GOODSTEIN D M, SHU S, HOWSON R, et al. Phytozome: A comparative platform for green plant genomics[J]. Nucleic Acids Research, 2012, 40(DI): D1178-D1186.

[19] SAKAI H, LEE S S, TANAKA T, et al. Rice annotation project database (RAP-DB): An integrative and interactive database for rice genomics[J]. Plant and Cell Physiology, 2013, 54(2): e6.

[20] KAWAHARA Y, DE LA BASTIDE M, HAMILTON J P, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data[J]. Rice, 2013, 6(1): 4.

[21] SONG J M, LEI Y, SHU C C, et al. Rice Information GateWay: A comprehensive bioinformatics platform for

*indica* rice genomes[J]. Molecular Plant, 2018, 11(3): 505-507.

[22] SANG J, ZOU D, WANG Z, et al. IC4R-2.0: Rice genome reannotation using massive RNA-seq data[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 161-172.

[23] AGRET C, GOTTIN C, DEREEPER A, et al. South green resources to manage rice big genomics data[C]// The Plant & Animal Genome Conference (PAG). San Diego: Scherago International, 2020.

[24] OHYANAGI H, EBATA T, HUANG X, et al. Oryza-Genome: Genome diversity database of wild Oryza species[J]. Plant and Cell Physiology, 2016, 57(1): e1.

[25] MAO L, CHEN M, CHU Q, et al. RiceRelativesGD: A genomic database of rice relatives for rice research[J]. Database, 2019, 2019: baz110.

[26] YAO W, LI G, YU Y, et al. funRiceGenes dataset for comprehensive understanding and application of rice functional genes[J]. GigaScience, 2018, 7(1): gix119.

[27] SATO Y, TAKEHISA H, KAMATSUKI K, et al. RiceXPro version 3.0: Expanding the informatics resource for rice transcriptome[J]. Nucleic Acids Research, 2013, 41(D1): D1206-D1213.

[28] WANG L, XIE W, CHEN Y, et al. A dynamic gene expression atlas covering the entire life cycle of rice[J]. The Plant Journal, 2010, 61(5): 752-766.

[29] XIA L, ZOU D, SANG J, et al. Rice Expression Database (RED): An integrated RNA-Seq-derived gene expression database for rice[J]. Journal of Genetics and Genomics, 2017, 44(5): 235-241.

[30] KAWAHARA Y, OONO Y, WAKIMOTO H, et al. TENOR: Database for comprehensive mRNA-seq experiments in rice[J]. Plant and Cell Physiology, 2016, 57(1): e7.

[31] YU Y, ZHANG H, LONG Y, et al. Plant Public RNA-seq Database: A comprehensive online database for expression analysis of ~45 000 plant public RNA-Seq libraries[J]. Plant Biotechnology Journal, 2022, 20(5): 806-808.

[32] ZHANG P, WANG Y, CHACHAR S, et al. eRice: A refined epigenomic platform for *japonica* and *indica* rice[J]. Plant Biotechnology Journal, 2020, 18(8): 1642-1644.

[33] XIE L, LIU M, ZHAO L, et al. RiceENCODE: A comprehensive epigenomic database as a rice Encyclopedia of DNA elements[J]. Molecular Plant, 2021, 14(10): 1604-1606.

[34] ZHANG B, FEI Y, FENG J, et al. RiceNCexp: A rice non-coding RNA co-expression atlas based on massive RNA-seq and small-RNA seq data[J]. Journal of Experimental Botany, 2022, 73(18): 6068-6077.

[35] SANAN-MISHRA N, TRIPATHI A, GOSWAMI K, et al. ARMOUR - A rice miRNA: mRNA interaction resource[J]. Frontiers in Plant Science, 2018, 9: 602.

[36] ZHANG Z, XU Y, YANG F, et al. RiceLncPedia: A comprehensive database of rice long non-coding RNAs[J]. Plant Biotechnology Journal, 2021, 19(8): 1492-1494.

[37] LIU W T, YANG C C, CHEN R K, et al. RiceATM: A platform for identifying the association between rice agronomic traits and miRNA expression[J]. Database, 2016, 2016: baw151.

[38] JOHNSON C, BOWMAN L, ADAI A T, et al. CSRDB: A small RNA integrated database and browser resource for cereals[J]. Nucleic Acids Research, 2007, 35 (Suppl_1): D829-D833.

[39] KOZOMARA A, BIRGAOANU M, GRIFFITHS-JONES S. miRBase: From microRNA sequences to function[J]. Nucleic Acids Research, 2019, 47(D1): D155-D162.

[40] YUAN C, MENG X, LI X, et al. PceRBase: A database of plant competing endogenous RNA[J]. Nucleic Acids Research, 2017, 45(D1): D1009-D1014.

[41] MARSICO M D, PAYTUVI GALLART A, SAN-SEVERINO W, et al. GreeNC 2.0: A comprehensive database of plant long non-coding RNAs[J]. Nucleic Acids Research, 2022, 50(D1): D1442-D1447.

[42] JIN J, LU P, XU Y, et al. PLncDB V2.0: A comprehensive encyclopedia of plant long noncoding RNAs[J]. Nucleic Acids Research, 2021, 49(D1): D1489-D1495.

[43] SZCZEŚNIAK M W, BRYZGHALOV O, CIOMBOROWSKA-BASHEER J, et al. CANTATAdb 2.0: Expanding the collection of plant long noncoding RNAs[J]. Methods in Molecular Biology, 2019, 1933: 415-429.

[44] GUO Z, KUANG Z, WANG Y, et al. PmiREN: A comprehensive encyclopedia of plant miRNAs[J]. Nucleic Acids Research, 2020, 48(D1): D1114-D1121.

[45] XU X, DU T, MAO W, et al. PlantcircBase 7.0: Full-length transcripts and conservation of plant circRNAs[J]. Plant Communications, 2022, 3(4): 100343.

[46] GUO X, WANG T, JIANG L, et al. PlaASDB: A comprehensive database of plant alternative splicing events in response to stress[J]. BMC Plant Biology, 2023, 23(1): 225.

[47] ZHU S, YE W, YE L, et al. PlantAPAdb: A comprehensive database for alternative polyadenylation sites in plants[J]. Plant Physiology, 2020, 182(1): 228-242.

[48] 徐海冬, 宁博林, 牟芳, 等. 选择性多聚腺苷酸化的生物学效应及其调控机制研究进展[J]. 遗传, 2021, 43(1): 4-15.

[49] HAN L, ZHONG W, QIAN J, et al. A multi-omics integrative network map of maize[J]. Nature Genetics, 2023, 55(1): 144-153.

[50] LIU C, MA Y, ZHAO J, et al. Computational network biology: Data, models, and applications[J]. Physics Reports, 2020, 846: 1-66.

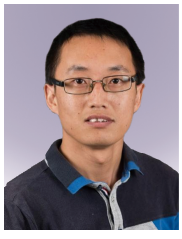
[51] HAQUE S, AHMAD J S, CLARK N M, et al. Computational prediction of gene regulatory networks in plant growth and development[J]. Current Opinion in Plant Biology, 2019, 47: 96-105.

[52] YAN J, WANG X. Machine learning bridges omics sciences and plant breeding[J]. Trends in Plant Science, 2023, 28(2): 199-210.

[53] SATO Y, NAMIKI N, TAKEHISA H, et al. RiceFREND: A platform for retrieving coexpressed gene

networks in rice[J]. Nucleic Acids Research, 2013, 41(Database issue): D1214-D1221.
[54] HAMADA K, HONGO K, SUWABE K, et al. OryzaExpress: An integrated database of gene expression networks and omics annotations in rice[J]. Plant and Cell Physiology, 2011, 52(2): 220-229.
[55] LIN H, YU J, PEARCE S P, et al. RiceAntherNet: A gene co-expression network for identifying anther and pollen development genes[J]. The Plant Journal, 2017, 92(6): 1076-1091.
[56] SIRCAR S, MUSADDI M, PAREKH N. NetREx: Network-based rice expression analysis server for abiotic stress conditions[J]. Database, 2022, 2022: baac060.
[57] GU H, ZHU P, JIAO Y, et al. PRIN: A predicted rice interactome network[J]. BMC Bioinformatics, 2011, 12: 161.
[58] LEE T, OH T, YANG S, et al. RiceNet v2: An improved network prioritization server for rice genes[J]. Nucleic Acids Research, 2015, 43(W1): W122-W127.
[59] LIU S, LIU Y, ZHAO J, et al. A computational interactome for prioritizing genes associated with complex agronomic traits in rice (*Oryza sativa*)[J]. The Plant Journal, 2017, 90(1): 177-188.
[60] PENG H, WANG K, CHEN Z, et al. MBKbase for rice: An integrated omics knowledgebase for molecular breeding in rice[J]. Nucleic Acids Research, 2020, 48(D1): D1085-D1092.
[61] ZHAO H, YAO W, OUYANG Y, et al. RiceVarMap: A comprehensive database of rice genomic variations[J]. Nucleic Acids Research, 2015, 43(Database issue): D1018-D1022.
[62] MANSUETO L, FUENTES R R, BORJA F N, et al. Rice SNP-seek database update: New SNPs, indels, and queries[J]. Nucleic Acids Research, 2017, 45(D1): D1075-D1081.
[63] YAN J, ZOU D, LI C, et al. SR4R: An integrative SNP resource for genomic breeding and population research in rice[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 173-185.
[64] YONEMARU J, EBANA K, YANO M. HapRice, an SNP haplotype database and a web tool for rice[J]. Plant and Cell Physiology, 2014, 55(1): e9.
[65] WANG C, YU H, HUANG J, et al. Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0[J]. Plant Biotechnology Journal, 2020, 18(1): 14-16.
[66] 鄂志国, 王磊. 中国水稻品种及其系谱数据库[J]. 中国水稻科学, 2011, 25(5): 565-566.
[67] COPETTI D, ZHANG J, EL BAIDOURI M, et al. RiTE database: A resource database for genus-wide rice genomics and evolutionary biology[J]. BMC Genomics, 2015, 16(1): 538.
[68] LIU Z, WANG T, WANG L, et al. RTRIP: A comprehensive profile of transposon insertion polymorphisms in rice[J]. Plant Biotechnology Journal, 2020, 18(12): 2379-2381.
[69] 刘耀光, 李构思, 张雅玲, 等. CRISPR/Cas 植物基因组编辑技术研究进展[J]. 华南农业大学学报, 2019, 40(5): 38-49.
[70] 李文龙, 栾鑫, 张强, 等. 基于 CRISPR/Cas9 基因编辑技术的水稻定向改良研究进展[J]. 广东农业科学, 2022, 49(9): 114-124.
[71] 何晓玲, 刘鹏程, 马伯军, 等. 基于 CRISPR/Cas9 的基因编辑技术研究进展及其在植物中的应用[J]. 植物学报, 2022, 57(4): 508-531.
[72] ANZALONE A V, RANDOLPH P B, DAVIS J R, et al. Search-and-replace genome editing without double-strand breaks or donor DNA[J]. Nature, 2019, 576(7785): 149-157.
[73] XIE X, MA X, ZHU Q, et al. CRISPR-GE: A convenient software toolkit for CRISPR-based genome editing[J]. Molecular Plant, 2017, 10(9): 1246-1249.
[74] LIU W, XIE X, MA X, et al. DSDecode: A web-based tool for decoding of sequencing chromatograms for genotyping of targeted mutations[J]. Molecular Plant, 2015, 8(9): 1431-1433.
[75] BAE S, KWEON J, KIM H S, et al. Microhomology-based choice of Cas9 nuclease target sites[J]. Nature Methods, 2014, 11(7): 705-706.
[76] XIE X, LI F, TAN X, et al. BEtarget: A versatile web-based tool to design guide RNAs for base editing in plants[J]. Computational and Structural Biotechnology Journal, 2022, 20: 4009-4014.
[77] XIE X, LIU W, DONG G, et al. MMEJ-KO: A web tool for designing paired CRISPR guide RNAs for microhomology-mediated end joining fragment deletion[J]. Science China Life Sciences, 2021, 64(6): 1021-1024.
[78] MUTWIL M, OBRO J, WILLATS W G T, et al. GeneCAT: Novel webtools that combine BLAST and co-expression analyses[J]. Nucleic Acids Research, 2008, 36(Suppl_2): W320-W326.
[79] PARK J, BAE S, KIM J S. Cas-Designer: A web-based tool for choice of CRISPR-Cas9 target sites[J]. Bioinformatics, 2015, 31(24): 4014-4016.
[80] BAE S, PARK J, KIM J S. Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases[J]. Bioinformatics, 2014, 30(10): 1473-1475.
[81] LEI Y, LU L, LIU H Y, et al. CRISPR-P: A web tool for synthetic single-guide RNA design of CRISPR-system in plants[J]. Molecular Plant, 2014, 7(9): 1494-1496.
[82] LIU H, DING Y, ZHOU Y, et al. CRISPR-P 2.0: An improved CRISPR-Cas9 tool for genome editing in plants[J]. Molecular Plant, 2017, 10(3): 530-532.
[83] FAN J, SHI L, LIU Q, et al. Annotation and evaluation of base editing outcomes in multiple cell types using CRISPRbase[J]. Nucleic Acids Research, 2023, 51(D1): D1249-D1256.
[84] ZHENG Y, ZHANG N, MARTIN G B, et al. Plant genome editing database (PGED): A call for submission of information about genome-edited plant mutants[J]. Molecular Plant, 2019, 12(2): 127-129.
[85] HONG W J, KIM Y J, KIM E J, et al. CAFRI-Rice: CRISPR applicable functional redundancy inspector to accelerate functional genomics in rice[J]. The Plant Journal, 2020, 104(2): 532-545.

[86] LIU Q, PENG X, SHEN M, et al. Ribo-uORF: A comprehensive data resource of upstream open reading frames (uORFs) based on ribosome profiling[J]. Nucleic Acids Research, 2023, 51(D1): D248-D261.
[87] XUE C, QIU F, WANG Y, et al. Tuning plant phenotypes by precise, graded downregulation of gene expression[J]. Nature Biotechnology, 2023. doi:10.1038/s41587-023-01707-w.
[88] 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.
[89] PéREZ P, DE LOS CAMPOS G. Genome-wide regression and prediction with the BGLR statistical package[J]. Genetics, 2014, 198(2): 483-495.
[90] PÉREZ-RODRÍGUEZ P, GIANOLA D, GONZÁLEZ-CAMACHO J M, et al. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat[J]. G3:Genes| Genomes| Genetics, 2012, 2(12): 1595-1605.
[91] CHARMET G, TRAN L G, AUZANNEAU J, et al. BWGS: A R package for genomic selection and its application to a wheat breeding programme[J]. PLoS One, 2020, 15(4): e0222733.
[92] YAN J, XU Y, CHENG Q, et al. LightGBM: Accelerated genomically designed crop breeding through ensemble learning[J]. Genome Biology, 2021, 22(1): 271.
[93] MA W, QIU Z, SONG J, et al. A deep convolutional neural network approach for predicting phenotypes from genotypes[J]. Planta, 2018, 248(5): 1307-1318.
[94] WANG K, ABID M A, RASHEED A, et al. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants[J]. Molecular Plant, 2023, 16(1): 279-293.
[95] YIN L, ZHANG H, TANG Z, et al. HIBLUP: An integration of statistical models on the BLUP framework for efficient genetic evaluation using big genomic data[J]. Nucleic Acids Research, 2023, 51(8): 3501-3512.
[96] YIN L, ZHANG H, ZHOU X, et al. KAML: Improving genomic prediction accuracy of complex traits using machine learning determined parameters[J]. Genome Biology, 2020, 21(1): 146.
[97] MOHAMMADI M, TIEDE T, SMITH K P. PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations[J]. Crop Science, 2015, 55(5): 2068-2077.
[98] ENDELMAN J B. Ridge regression and other kernels for genomic selection with R package rrBLUP[J]. The Plant Genome, 2011, 4(3): 250-255.
[99] COVARRUBIAS-PAZARAN G. Genome-assisted prediction of quantitative traits using the R package sommer[J]. PLoS One, 2016, 11(6): e0156744.
[100] BUDHLAKOTI N, MISHRA D C, RAI A, et al. STGS: Genomic selection using single trait [EB/OL]. [2023-06-30]. https://cran.r-project.org/web/packages/STGS.
[101] YANG J, LEE S H, GODDARD M E, et al. GCTA: A tool for genome-wide complex trait analysis[J]. American Journal of Human Genetics, 2011, 88(1): 76-82.
[102] CHENG H, FERNANDO R, GARRICK D, et al. JWAS: Julia implementation of whole-genome analysis software[C]//Proceedings of the World Congress on Genetics Applied to Livestock Production. Auckland, New Zealand: World Congress on Genetics Applied to Livestock Production, 2018.
[103] KANG H, NING C, ZHOU L, et al. PIBLUP: High-performance software for large-scale genetic evaluation of animals and plants[J]. Frontiers in Genetics, 2018, 9: 226.
[104] TECLE I Y, EDWARDS J D, MENDA N, et al. solGS: A web-based tool for genomic selection[J]. BMC Bioinformatics, 2014, 15(1): 398.
[105] CHEN C J, ZHANG Z. iPat: Intelligent prediction and association tool for genomic research[J]. Bioinformatics, 2018, 34(11): 1925-1927.
[106] XU Y, MA K, ZHAO Y, et al. Genomic selection: A breakthrough technology in rice breeding[J]. The Crop Journal, 2021, 9(3): 669-677.
[107] XU Y, ZHANG X, LI H, et al. Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction[J]. Molecular Plant, 2022, 15(11): 1664-1695.
[108] 蒋金金, 苏汉东, 洪登峰, 等. 植物生物技术研究进展[J]. 植物生理学报, 2023, 59(8): 1436-1462.



刘 琦，研究员，广东省农业科学院水稻研究所生物信息与大数据育种研究室主任，博士毕业于中国林业科学研究院林木遗传育种专业，随后在哈佛大学医学院从事博士后研究工作，主要研究领域是生物信息学、比较基因组学与系统生物学。作为第一作者和通信作者在《Molecular Cell》《Nucleic Acids Research》《Journal of Clinical Investigation》《Briefings in Bioinformatics》《EMBO Reports》《Bioinformatics》《PLoS Computational Biology》和《Human Mutation》等期刊上发表多篇高水平论文，作为主要生物信息学作者在《Nature》《Nature Protocols》《Nature Communications》等上发表转录后调控研究相关论文。建立了多种生物信息学算法、软件和数据库，为基因调控研究提供重要的生物信息学支持。目前开展的研究主要为开发新的软件、技术和方法在个体和群体水平研究水稻生长发育、杂种优势、抗逆的系统调控机制，挖掘水稻优势性状的分子标记，为基于大数据的精准育种提供重要理论基础和生物信息学支持。

【责任编辑 庄 延】