科学通报 2015 年 第 60 卷 第 4 期: 367 ~ 371
快 讯 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

水稻功能基因组育种数据库(RFGB): 3K 水稻 SNP 与 InDel 子数据库

郑天清[①*†], 余泓[②*], 张洪亮[③*], 吴志超[①*], 王文生[①*], 太帅帅[④], 迟璐[①], 阮珏[⑤], 韦朝春[⑥], 石建新[⑥], 高用明[①], 傅彬英[①], 周永力[①], 赵秀琴[①], 张帆[①], McNally Kenneth L.[⑦], 李自超[③], 张耕耘[④], 李家洋[②], 张大兵[⑥], 徐建龙[①⑤†], 黎志康[①⑤†]

① 中国农业科学院作物科学研究所, 国家作物基因资源与遗传改良重大科学工程, 北京 100081;
② 中国科学院遗传与发育生物学研究所, 北京 100101;
③ 北京市作物遗传改良重点实验室, 中国农业大学植物遗传育种系, 北京 100193;
④ 深圳华大基因研究院, 深圳 518083;
⑤ 中国农业科学院深圳农业基因组研究所, 深圳 518120;
⑥ 上海交通大学生命科学技术学院, 上海 200240;
⑦ T.T. Chang Genetic Resources Center, IRRI, Los Baños, Laguna 4031, Philippines
* 同等贡献
† 联系人, E-mail: tonyztq@163.com; xujlcaas@126.com; zhkli1953@126.com

2014-11-18 收稿, 2014-12-16 接受, 2015-01-23 网络版发表
国家高技术研究发展计划(2012AA101101, 2014AA10A601)、引进国际先进农业科学技术计划(2011-G2B)、国家国际合作专项(2012DFB32280)、比尔盖茨基金(OPP51587, OPPGD1393)和深圳市技术研究开发计划(JSGG20121026152117750)资助

摘要 水稻是全世界最重要的粮食作物之一. 对全球水稻品种资源进行基因组研究, 挖掘有利等位基因和指导基于全基因组信息的分子育种具有重大的理论和实践意义, 是关系国家乃至全球粮食安全的重大战略问题. 本研究通过全球 3000 份水稻资源(3K 水稻)的全基因组重测序, 获取了其中质量较好的 2859 份水稻基因组的单核苷酸多态性(SNP)及小片段插入缺失(InDel)的基因组变异数据, 建立了综合性的水稻功能基因组育种数据库的 SNP 与 InDel 多态性子数据库. 该子数据库包含了 3K 水稻多态性信息检索、基因组浏览器可视化系统、特定区段基因组数据导出系统等多项功能. 本数据库的建立将为研究水稻基因功能、指导水稻全基因组选择育种提供重要平台.

关键词
3K 水稻
单核苷酸多态性
插入缺失
基因组育种
数据库

水稻作为全世界最重要的主粮之一, 养活了超过二分之一的人口. 2005 年, 完整的水稻基因组图谱首次被绘制出来[1]. 此后, 水稻功能基因组的研究日新月异, 这些研究工作对水稻的遗传育种研究产生了巨大的影响[2]. 最近, 3000 份水稻重测序项目(3K Rice Genome Project, 3K 水稻基因组项目)收集了全球 2859 份水稻品种, 这些品种覆盖了全球不同遗传背景和丰富表型的水稻资源, 并通过二代基因组测序技术, 得到了平均 14 倍覆盖深度的基因组测序数据[3]. 如何高效地分析和整理这些海量的基因组数据, 对利用这些基因组信息研究水稻基因功能, 指导全基因组选择育种有重大的科学意义和实用价值[4].

单核苷酸多态性(single nucleotide polymorphism, SNP)是指在一个群体内, 由单个碱基的突变所引起的 DNA 序列多态性, 是可遗传的变异中最常见的一种, 占所有已知多态性的 90%以上[5]; 小片段的插入缺失(insertion and deletion, InDel)也是基因组中常见的变异. 通过 3K 水稻基因组项目[3], 我们获得了包含 3000 份不同水稻品种的基因组重测序原始数据, 共约 16T. 原始数据通过标准 GATK 分析流程进行处理[6], 通过和水稻日本晴基因组[7]进行比对, 最终得到 2859 份不同水稻

引用格式: 郑天清, 余泓, 张洪亮, 等. 水稻功能基因组育种数据库(RFGB): 3K 水稻 SNP 与 InDel 子数据库. 科学通报, 2015, 60: 367–371
Zheng T Q, Yu H, Zhang H L, et al. Rice functional genomics and breeding database (RFGB): 3K-rice SNP and InDel sub-database (in Chinese). Chin Sci Bull, 2015, 60: 367–371, doi: 10.1360/N972014-01231

品种的 SNP 和 InDel 信息.

图 1(a)和(b)显示了这些样本的测序深度及其中 SNP 和小片段 InDel 的分布. 同时, 我们对 2859 份水稻品种的基本信息也进行了统一的收集和整理, 这些信息包括水稻品种的样品名称、来源、亚种类型、种子保存地、聚类组群信息等; 图 1(c)显示了数据库所包含的 2859 份样品的几种主要类型. 根据这些 SNP 和 InDel 数据和品种基本信息, 我们建立了以 MySQL 为基础的数据库管理系统, 并利用 php 语言开发了数据检索页面. 在此基础上, 利用 JBrowser 工具[8], 开发并建立了可以进行可视化显示和多品种对比的基因组 SNP 及 InDel 浏览器页面.

图 2 展示了全球水稻单核苷酸多态性数据库的结构和主要内容. 数据库共包含了 2859 份水稻资源的共计超过 1890 万条 SNP 和 InDel 信息, 针对不同类型的用户, 设计了 2 个主要的用户入口, 浏览或者检索 3K 水稻样品, 其中浏览入口可以进行多品种的比对; 检索入口可以方便地对特定品种进行分析、全基因组单核苷酸多态性数据展示和特定区段基因组序列下载.

图 3 展示了品种信息检索结果. 该表格记录了水稻品种样品的名称、来源、亚种类型、种子保存地、聚类组群信息, 并且可以通过与 JBrowser 界面的连接, 实现上文中提到的 JBrower 功能. 除此之外, 为了方便遗传育种工作者进行序列下载和引物设计, 数据库根据单核苷酸多态性信息对参考基因组序列的相应位置进行序列替换, 从而得到了 2859 份不同水稻品种的基因组还原序列信息. 通过此页面, 用户可以对 2859 份水稻品种的多态性替换后的基因组序列进行下载, 导出的基因组序列为国际通用的 fasta 格式, 进而可以直接进行多重序列比对或 PCR 引物设计. 这项功能的实现极大地方便了育种用户对 SNP 及 InDel 变异信息的使用.

图 4 展示了利用 JBrowser 建立的多品种对比单核苷酸多态性基因组浏览器页面. 数据库将每个品种的每条染色体建立了一条 track, 通过 select track 栏, 用户可以根据染色体(Chr)、样品名称(Designation)、聚类组群(Group)、样品来源(Origin)、种子保存地(Seed provider)以及亚种类型(Type)等信息选择需要浏览的样本基因组.

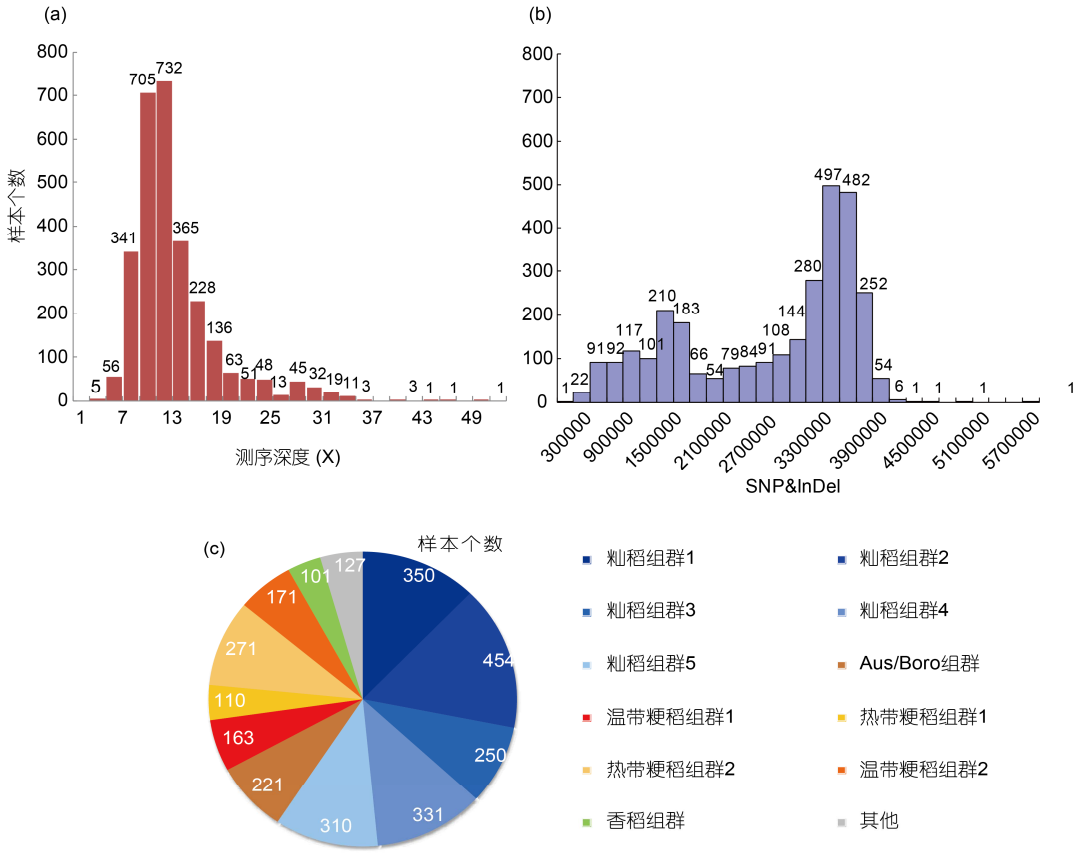


图 1 (网络版彩色)水稻功能基因组育种数据库(RFGB)——3K 水稻 SNP 与 InDel 子数据库所包含的数据内容. (a) 样本测序深度的分布图; (b) 样本包含的 SNP 和 InDel 信息条数分布图; (c) 样本的构成类型

Figure 1 The data content of Rice Functional Genomics and Breeding Database (RFGB) —3K-rice SNP and InDel sub-database. (a) The sample distribution of depths of sequencing. (b) The sample distribution of SNP & InDel. (c) The composition of types of samples

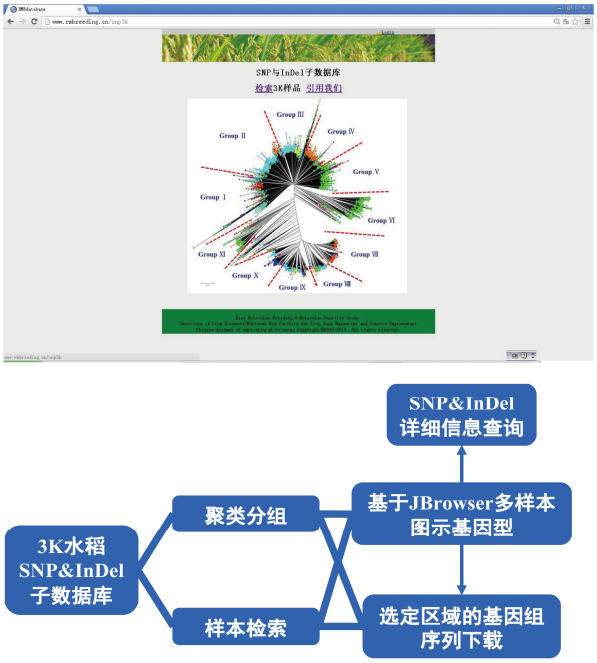


图 2 (网络版彩色)3K 水稻 SNP 与 InDel 子数据库的主页面及数据库结构

Figure 2 The homepage and the structure of the 3K-rice SNP and InDel Sub-database

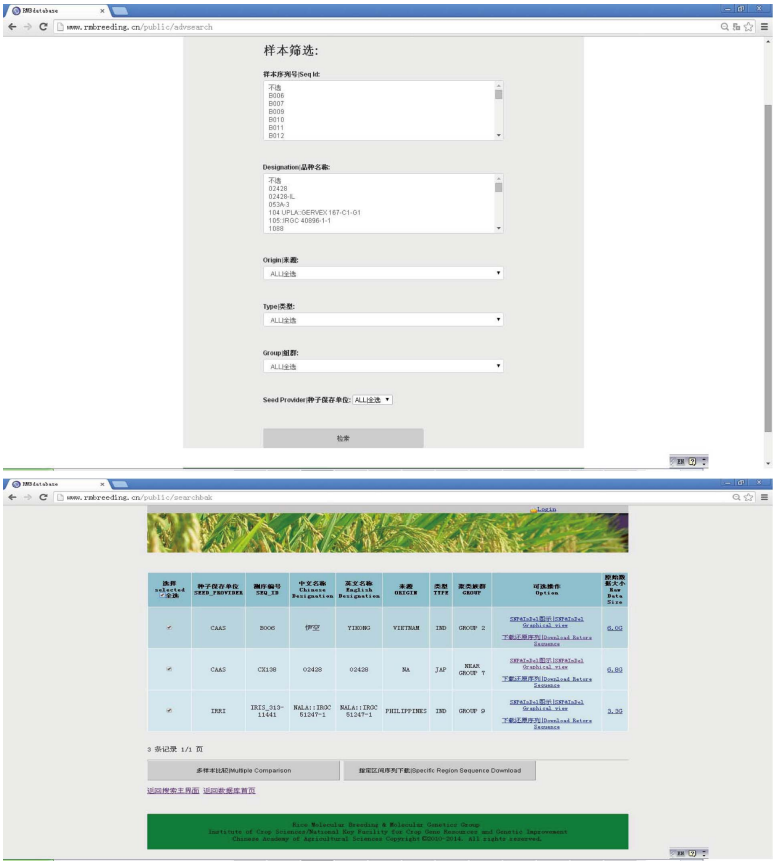


图 3 (网络版彩色)3K 水稻 SNP 与 InDel 子数据库的检索页面使用范例

Figure 3 A case for searching in the 3K-rice SNP and InDel Sub-database

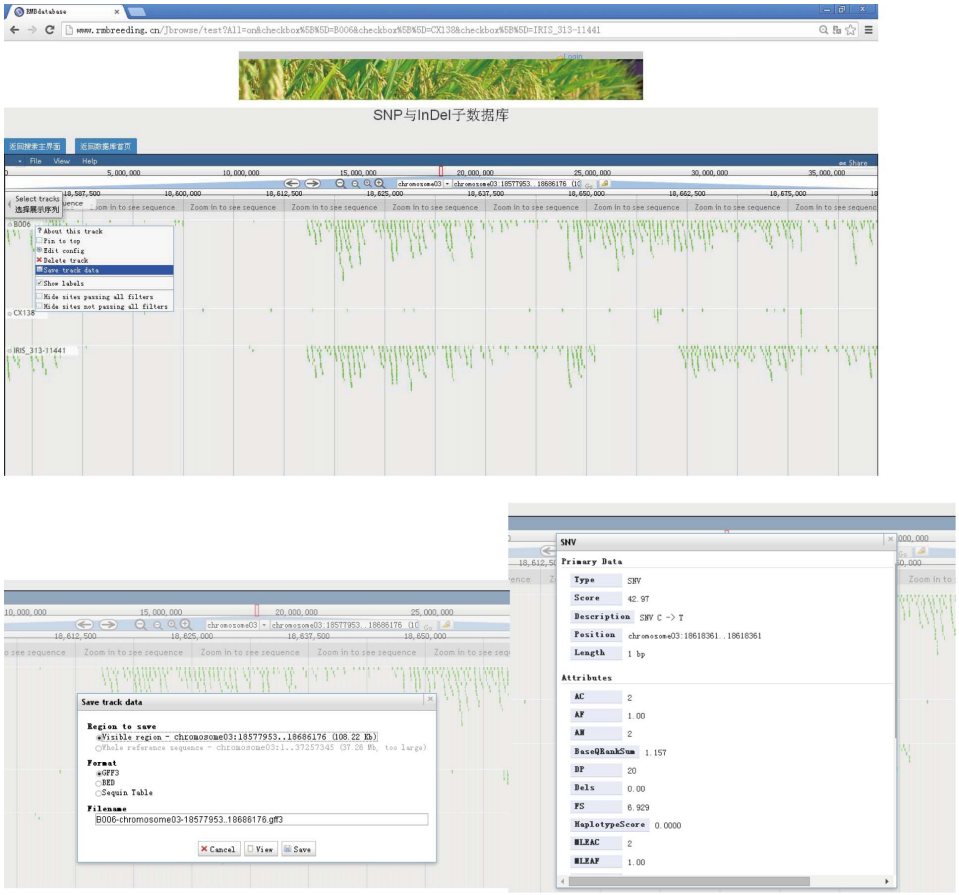


图 4 (网络版彩色)3K 水稻 SNP 与 InDel 子数据库的功能页面使用范例

Figure 4 A case for functional pages of the 3K-rice SNP and InDel Sub-database

在可视化基因组浏览器页面可以同时显示多条 track 并进行比较. 如图 4 所示, 通过在 select track 栏选取了若干个(考虑到服务器承载量, 同时展示限制不超过 50 个)品种, JBrowser 同时显示了第 3 染色体的 18577953~18686176 bp 的 108.22 kb 区段内这几个品种的所有单核苷酸多态性信息, 该结果直观地体现在 18618361 bp 附近可能存在用户感兴趣的多态位点, 点击该多态位点可以显示出该多态位点的详细信息. 用户可以结合不同的表型和品种, 对感兴趣的特异区段进行针对性的研究. 每条 track 都可以导出选定区域内全部碱基序列信息.

中国是水稻育种和基础研究领先的国家, 如何为广大的水稻育种和基础研究工作者提供直观便捷的基础数据访问方式是建立全球水稻功能基因组育种数据库的出发点. 其中 SNP 和 InDel 子数据库的建立, 为用户特别是中国的广大水稻育种工作者提供了一个高效检索 3K 水稻基因组项目中水稻单核苷酸多态性数据, 开展高效的信息挖掘和下载的平台. 3000 份水稻重测序项目共产出了超过 16 T (gz 格式)的原始数据, 以及超过 1890 万条单核苷酸多态性数据等变异信息. 面对如此大规模的基因组数据, 如何高效方便地进行数据分析处理、数据检索挖掘和数据可视化, 对所有的水稻研究人员和育种工作者有着重要意义. 通过数据检索, 可以高效地获得特定水稻品种内存在的基因组变异情况; 通过多品种比对, 可以得到特定性状基因在不同水稻品种间的变化和分类, 进行系统的研究; 通过特定区间序列的批量下载, 可以高效地对特定的多态性进行全基因组连接育种引物的设计, 从而加快分子辅助育种的进程. 通过 3K 水稻 SNP 与 InDel 子数据库这些功能, 可以极大地提高数据利用效率和水稻功能基因组的研究进度, 为基于水稻基因组信息的水稻育种工作提供便利. 今后, 根据广大水稻研究人员和育种工作者的需求, 对水稻功能基因组育种数据库(RFGB)的 3K 水稻 SNP 与 InDel 子数据库功能进行不断的提升和完善.

水稻功能基因组育种数据库(RFGB)的 3K 水稻 SNP 与 InDel 子数据库利用 MySQL 建立起数据库管理系统, 并通过 php 语言和 JBrowser 开发出了基于网页的交互界面, 实现数据的交换. 用户可以通过以下地址进行访问 http://www.rmbreeding.cn/snp3k.

参考文献

1 Matsumoto T, Wu J Z, Kanamori H, et al. Nature, 2005, 436: 793–800
2 Huang X, Kurata N, Wei X, et al. Nature, 2012, 490: 497–501
3 The 3000 rice genomes project. GigaScience, 2014, 3: 7
4 Li J Y, Wang J, Zeigler R S. GigaScience, 2014, 3: 8
5 Barreiro L B, Laval G, Quach H, et al. Nat Genetics, 2008, 40: 340–345
6 DePristo M A, Banks E, Poplin R, et al. Nat Genetics, 2011, 43: 491–498
7 Kawahara Y, de la Bastide M, Hamilton J P, et al. Rice (N Y), 2013, 6: 4
8 Skinner M E, Uzilov A V, Stein L D, et al. Genome Res, 2009, 19: 1630–1638

Rice functional genomics and breeding database (RFGB): 3K-rice SNP and InDel sub-database

ZHENG TianQing[1], YU Hong[2], ZHANG HongLiang[3], WU ZhiChao[1], WANG WenSheng[1], TAI ShuaiShuai[4], CHI Lu[1], RUAN Jue[5], WEI ChaoChun[6], SHI JianXin[6], GAO YongMing[1], FU BinYing[1], ZHOU YongLi[1], ZHAO XiuQin[1], ZHANG Fan[1], McNALLY Kenneth L.[7], LI ZiChao[3], ZHANG GengYun[4], LI JiaYang[2], ZHANG DaBing[6], XU JianLong[1,5] & LI ZhiKang[1,5]

[1] *Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing 100081, China;*
[2] *Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China;*
[3] *Beijing Key Laboratory of Crop Genetic Improvement/College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China;*
[4] *Beijing Genome Institute-Shenzhen, Shenzhen 518083, China;*
[5] *Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China;*
[6] *School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China;*
[7] *T. T. Chang Genetic Resources Center, IRRI, Los Baños, Laguna 4031, Philippines*

Rice is an important food crop worldwide. Genomic research on global germplasm has theoretical and practical significances in the mining of favorable alleles and on genome-based molecular breeding, which are related to Chinese and global food security. Through the 3000 (3K) Rice Genome Project, we collected single nucleotide polymorphism (SNP) and insertion and deletion (InDel) genomic variation data for the 2859 rice genomes, and established a comprehensive SNP and InDel sub-database for the Rice Functional Genomics-based Breeding (RFGB) Database. This sub-database is a global resource containing a polymorphism information retrieval function, a genome browser visualization system and a data export system for specific genomic regions, along with other tools. This study establishes an important platform for rice gene functional research and rice molecular breeding by genomic selection.

3K-rice, SNP, InDel, genome-based breeding, database

doi: 10.1360/N972014-01231