中国农业信息 China Agricultural Informatics
第 30 卷第 5 期 2018 年 10 月 2018，30（5）：13-23 Vol.30，No.5 Oct.，2018

作物表型组数据库研究进展及展望*

王璟璐[1,2]，张 颖[1,2]，潘晓迪[1,2]，卢宪菊[1,2]，
马黎明[1,2]，郭新宇[1,2※]
（1. 北京市农林科学院北京农业信息技术研究中心，北京 100097；2. 数字植物北京重点实验室，北京 100097）

摘要：【目的】总结归纳作物表型数据库的研究进展，对作物表型数据库的构建进行展望。**【方法】**采用文献综述法，从 Web of Science、NCBI 的 PubMed 和中国知网等常用公共文献数据库中对已发表的作物表型组学相关研究文献进行检索，据此对国内外作物表型组学研究现状进行分析，并基于其中的数据库研究，对目前的作物表型相关数据库进行了介绍。最后对作物表型组数据库构建标准及要求进行了介绍。**【结果】**不同于基因组学已有许多大型的、公认的、成熟的公共数据库，有关作物表型组学的数据库虽已有一些，但综合性较强、普适性较广的通用标准数据库却不是很多。因此，构建综合性作物表型组标准数据库或构建特定作物的表型组数据库，将成为该领域相关研究人员的工作重点。**【结论】**农业信息化是现代农业的一个必然发展趋势，作物表型组数据库的构建也是顺应时代发展的产物。今后，在充分利用各种综合和专用数据库的基础上，研究人员应在实际研究中构建自己的作物表型组数据库，增强数据管理和共享。

关键词：作物；表型组；数据库；构建标准；数据管理
DOI：10.12105/j.issn.1672-0423.20180502

0 引言

作物及其相关领域科学研究与粮食问题息息相关。由于全球气候变化，作物生产面临着更频繁的极端天气，加之有限的水分及养分资源和可耕地面积，农业生产迫切需要新型气候适应性品种的繁育，以满足人们日益增长的粮食需求以及生物能源等其他工业用途的作物供应需求。

随着人类基因组计划（Human Genome Project，HGP）的完成，水稻[1-2]、玉米[3]、高粱[4]、大豆[5]和小麦[6]等主要农作物的基因组也相继被破译，作物研究随之进入

收稿日期：2018-09-15
第一作者简介：王璟璐（1990—），女，汉族，河北邯郸人，硕士、助理工程师。研究方向：生物信息学。Email：wangjl@nercita.org.cn
※ **通信作者简介：**郭新宇（1973—），男，汉族，内蒙古人，博士、研究员。研究方向：数字植物理论研究。Email：guoxy@nercita.org.cn
* **基金项目：**国家自然科学基金（No.31671577）；北京市自然科学基金青年项目（5174033）；北京市农林科学院数字植物科技创新团队（JKNYT201604）；北京市博士后工作经费资助项目（2016 ZZ-66）；北京市博士后科研活动经费资助（2018-ZZ-060）；院科技创新能力建设专项—基于组学的玉米维管束形成机理解析（KJCX20170404）

组学时代。计算机技术的快速发展为有效管理急速增多的生物学数据提供了可能，而生物信息学成为处理和挖掘高通量数据信息的主要手段。在生物信息学中，数据库作为其研究的主要载体出现在生命科学的众多领域。数据库管理系统（Database Management system，DBMs）可以实现数据的存储、检索、分析和维护，互联网技术为数据库的开发、维护、推广和应用提供了有效工具。如今，基因组学、蛋白质组学、代谢组学等各类组学数据库，不仅为该领域的研究和发展提供了丰富的数据信息，同时又加强了多组学间及与其他系统生物学分支间的联系，为学科间的交叉研究奠定了基础。

近年来，表型组学（Phenomics）日渐兴起并成为一门快速发展的数据密集型学科。表型组学相关技术和研究手段的高速发展，带来了数量巨大、尺度多维、数据多样的表型信息，如 RGB、高光谱、近红外、热和荧光成像等图像数据，植物生长过程中的各项生理指标数据等[7]。促使该领域的模型和数据管理系统随之发展，以便能够合理利用这些复杂的、动态的、大规模表型数据。

文章从 Web of Science（http：//apps.webofknowledge.com）、NCBI 的 PubMed（https：//www.ncbi.nlm.nih.gov/pubmed/）和中国知网（CNKI，http：//epub.cnki.net/kns/default.htm）等常用公共文献数据库中对已发表的作物表型组学相关研究文献进行检索，据此对国内外作物表型组学研究现状进行分析，并基于其中的数据库研究，对目前的作物表型相关数据库进行综述。最后，该文就作物表型组数据库构建的标准及要求进行了介绍，并将参照这些数据库构建原则在实际研究中设计自己的作物表型组数据库。

1 作物表型组学研究现状

表型组学这一概念于 1996 年由衰老研究中心主任 Steven A.Garan 在滑铁卢大学的一次应邀演讲上首次提出[8]。表型组学的定义类似于基因组学及其他组学，是指在基因组水平上系统地研究某一生物或细胞在各种不同环境条件下所有表型的学科。自 2009 年以来，随着植物表型无损获取方法以及大规模自动化高通量表型获取设施的建立[9]，表型组技术开始应用于基础植物研究和作物育种中，并有望打破育种中的表型瓶颈[10]。如今，表型组学在植物，尤其是作物研究中逐年增多。作物表型组学的研究基于高通量信息获取平台收集的大量作物表型数据，包括株高、叶面积、果实等形态特征，水分利用效率和光合作用等生理特征以及花青素含量等生化特征。因为作物表型本身具有很高的复杂性，且时常处于动态变化中，所以研究人员在实际研究过程中一般只关注少数几个表型，进行非动态的粗略研究。加之传统的作物表型获取效率低，表型研究技术也相对落后，使得表型组学在作物研究领域严重滞后于其他组学研究。截至目前，在单一表型或只关注少数几个表型层面的研究已有很多，而从组学出发对作物表型进行的研究才刚刚起步。

该文在常用文献检索数据库 Web of Science、PubMed 和中国知网上对已发表的作物表型组学相关研究进行检索。从表型组的概念提出至今，外文文献中以表型组学为主题的文献有 720 篇，其中限定为作物和常见作物名称（如水稻、玉米、小麦等）后的文献

数量为 288 篇。而以作物表型组学及常见作物名称为关键词在中国知网中进行检索，可得到中文期刊文献约 20 篇。由图 1 可以看出，近年来，作物研究领域中以表型组学为主题的文章数目逐年增多，且近 5 年来数量陡增，可见随着高通量作物表型获取手段的不断开发和完善，研究人员越来越关注表型组学的研究。

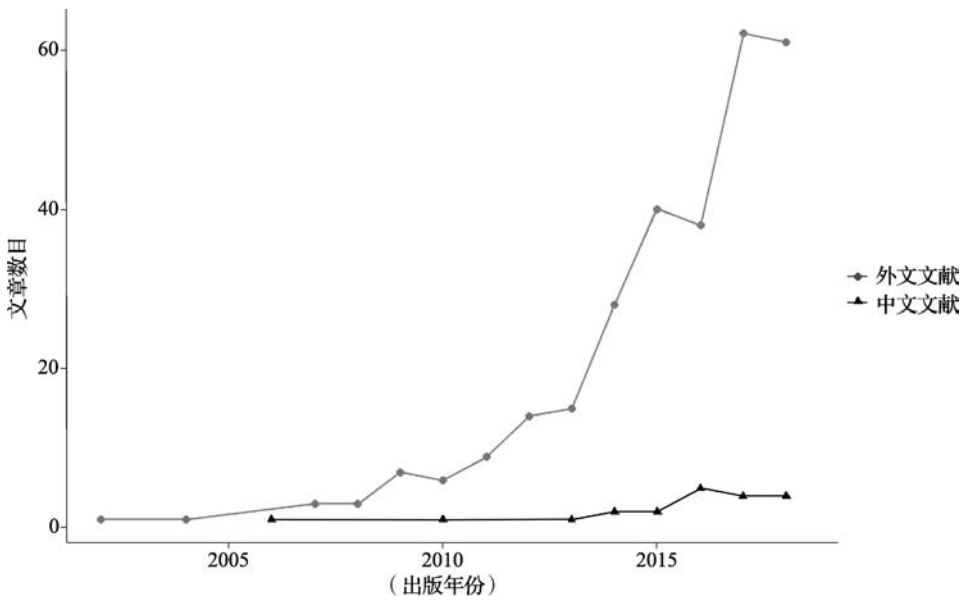

图 1　近年来作物表型组学研究文献数量及趋势

Fig.1　The number and trend of published papers focused on Crop Phenomics in recent years

作物表型组学的急速发展伴随着大量表型数据的产生，这就需要研究人员思考如何更好地对获得的表型数据进行管理。在数据管理中，建立标准数据库是一种十分便利且有效的方式。通过建立作物表型组数据库，可以对表型数据进行存储和分类，便于研究人员检索、分析并分享研究成果。

2　作物表型组数据库研究进展

不同于基因组学已有许多大型的、公认的、成熟的公共数据库，如人类基因组图谱数据库（The Genome Database，GDB）[11]、Ensembl 基因组注释数据库[12] 和 GenBank DNA 序列数据库[13] 等，作物表型组学数据库虽已有一些，但综合性较强、普适性较广的通用标准数据库却不是很多。在该文检索到的近 300 篇有关作物表型组学的研究中，关于表型组数据库的研究仅 20 余篇。这些作物表型组数据库大多以物种进行分类，其数据形式丰富多样，具体内容和访问网址详见表 1。

该文对 Planteome 数据库[14]、PGP 知识库[15] 和 OPTIMAS–DW 玉米资源库[16] 等主要作物表型相关数据库进行介绍，便于相关研究人员更好地使用，也为建立自己的作物表型组数据库提供借鉴。

表 1 主要作物表型数据库信息

Table 1 List of main crop phenotypic databases

数据库名称	简介	发布年份	PMID	URL
Planteome[14]	植物基因组和表型组数据共享平台	2018	29186578	http://www.planteome.org
PGP repository[15]	植物表型和基因组学数据发布基础平台	2016	27087305	http://edal.ipk-gatersleben.de/repos/pgp/
OPTIMAS-DW[16]	玉米的转录组学、代谢组学、离子组学、蛋白质组学和表型组学综合数据资源库	2012	23272737	https://apex.ipk-gatersleben.de/apex/f?p=270:1
BIOGEN BASE-CASSAVA[17]	木薯表型组和基因组信息资源库	2011	21904428	http://www.tnaugenomics.com/biogenbase/casava.php
TRIM[18]	台湾水稻插入突变体数据库	2007	28854617	http://www.trim.sinica.edu.tw
BreeDB	包含育种所需数量农艺性状	2012	—	https://www.wur.nl/en/show/BreeDB.htm
PhenoFront	LemnaTec 表型平台的网页服务器前端，包含实验数据和相关植物快照	—	—	https://github.com/danforthcenter/PhenoFront
Gramene[19]	植物基因组比较基因组学数据库	2011	20931385	http://www.gramene.org
MaizeGDB[20]	玉米遗传学和基因组数据库	2004	14681441	https://www.maizegdb.org

2.1 Planteome 数据库：植物基因组和表型组数据共享平台

Planteome 数据库[14]为特定物种的植物本体以及基因和表型注释提供了一套参考。本体用作大量且不断增长的植物基因组学、表型组学和遗传学数据语料库的语义整合的通用标准。参考本体包括植物本体论（Plant Ontology），植物性状本体论（Plant Trait Ontology），由 Planteome 开发的植物实验条件本体论（Plant Experimental Conditions Ontology），基因本体论（Gene Ontology），生物学兴趣的化学实体（Chemical Entities of Biological Interest），表型和属性本体论（Phenotype and Attribute Ontology）等。该项目还提供了来自世界各地的各种植物育种和研究团体开发的特定物种作物本体的途径。该数据库中提供了来自 95 种植物分类群的植物性状、表型、基因功能和表达的综合数据并以参考本体术语注释。Planteome 项目还开发了一个植物基因注释平台——Planteome Noctua，方便研究人员参与交流。所有 Planteome 本体都是公开可用的，并存放于 Planteome GitHub 站点，便于共享、跟踪修订和新请求。Planteome 数据库中所存储的数据均可免费访问。

Planteome 数据库拥有 8 种特定种类的作物本体（Crop Ontologies）[14]，其中对性状和表型评分标准的描述已被国际育种项目 maize（玉米），sweet potato（甘薯），soybean（大豆），pigeon pea（木豆），rice（水稻），cassava（木薯），lentil（小扁豆）和 wheat（小麦）采用。此外，该数据库还提供了 Planteome Noctua 基因注释工具，用于将研究社区与植物基因的功能注释相结合。

Planteome 数据库具有本体浏览器和分面搜索选项，可访问各种生物实体的本体和基于本体的注释。所有数据和本体都存储在一个索引系统中，该索引系统允许通过本体浏

览器进行全文搜索。GitHub 存储库（https：//github.com/Planteome/amigo）提供了数据存储设计的模式和索引文件。在目前的 Planteome 2.0 Release 中，Planteome 数据库囊括了大约 200 万生物或数据对象的访问，包括蛋白质、基因、RNA 转录、基因模型、种质和数量性状基因座。生物实体注释通常使用来自同一或多个引用本体类的多个本体术语。目前，这 200 万个实体大约有 2 100 万个注释。此外，该数据库还提供了转至多个参考本体的链接（表 2）。

表 2　Planteome 参考本体和词汇

Table 2　Planteome reference ontologies and vocabularies

本体名称	核心内容	URL
Ontology name Plant Ontology（PO）	植物结构和发育阶段	http：//browser.planteome.org/amigohttps：//github.com/Planteome/plant- ontology
Plant Trait Ontology（TO）	植物性状	http：//browser.planteome.org/amigohttps：//github.com/Planteome/plant- trait- ontology
Plant Experimental Conditions Ontology（PECO）	植物科学实验中使用的处理和生长条件	http：//browser.planteome.org/amigohttps：//github.com/Planteome/plant- experimental- conditions- ontology
Gene Ontology（GO）	分子功能，生物过程，细胞成分	http：//www.geneontology.org/
Phenotypic Qualities Ontology（PATO） Chemical Entities of Biological Interest（ChEBI）	品质和属性	https：//github.com/pato- ontology/pato https：//www.ebi.ac.uk/chebi/
Evidence and Conclusion Ontology（ECO）	侧重于小化合物的分子实体，用于支持科学研究结论	http：//www.evidenceontology.org/
Planteome NCBI Taxonomy*	生物分类层次	https：//github.com/Planteome/planteome-ncbi- taxonomy

2.2　PGP 知识库：植物表型和基因组学数据发布基础平台

PGP 知识库[15]（Plant Genomics and Phenomics Research Data Repository）是由莱布尼茨植物遗传与作物植物研究所和德国植物表型分析网络联合发起的植物基因组学和表型组学研究数据库，目的在于分享源自植物基因组学和表型组学的研究数据。PGP 中涵盖了因数量或数据范围不被支持而未在中央存储库中发布的跨域数据集，如来自植物表型和显微镜的图像集，未完成的基因组、基因型数据，形态植物模型的可视化，来自质谱以及软件和文档的数据等。该存储库由莱布尼茨植物遗传学和作物植物研究所托管，使用 e!DAL 作为软件基础平台，并使用分层存储管理系统作为数据存档后端。PGP 知识库具有成熟的数据提交工具，该工具高度自动化，可降低数据发布的障碍。经过内部审核流程之后，数据将作为可引用的数字对象标识符发布，并在 DataCite 中注册一组核心技术元数据。e!DAL 嵌入式网页前端为每个数据集生成登录页面并支持交互式探索。PGP 作为有效的 EU Horizon 2020 开放数据存档，在 BioSharing.org、re3data.org 和 OpenAIRE 已注册为研究数据存储库。在上述功能中，编程接口和标准元数据格式的支持使 PGP 能够实现 FAIR 数据原则——可查找、可访问、可互操作和可重用。

PGP 主要着眼于发布和共享涵盖各种数据领域的主要实验数据，如高通量植物表型分类的图像收集、序列组装、基因分型数据、形态植物模型的可视化和质谱数据，甚至软件。PGP 存储库中的数据集被分配给在 DataCite 上注册的可用 DOI，其中包含一组标准化的技术元数据。截至 2015 年 12 月，PGP 中已有 54 个数据集作为 DOI 发布，并在 DataCite 研究数据目录中注册。其中，每个数据集中都包括与特定实验或科学论文相关的所有记录。PGP 存储库目前拥有 21 157 个数据实体，总体容量为 65.4 GB。

2.3 OPTIMAS–DW：玉米的转录组学、代谢组学、离子组学、蛋白质组学和表型组学综合数据资源库

OPTIMAS–DW(OPTIMAS Data Warehouse) 数据库[16]是有关玉米研究的综合数据集。该数据库整合了来自不同数据域的数据，如转录组学、代谢组学、离子组学、蛋白质组学和表型组学。OPTIMAS 项目中设计并注释了 44 K 寡核苷酸芯片，以描述所选 unigenes 的功能。该项目进行了几个处理和植物生长阶段实验，并将测量数据填充到数据模板中。数据模板中的数据通过基于 Java 的导入工具导入数据库中。Web 界面允许用户浏览 OPTIMAS–DW 中所有数据域的存储实验数据。此外，用户可以过滤数据以提取自己感兴趣的信息。数据库中的所有数据可以导出为不同的文件格式，以进行进一步的数据分析和可视化。数据分析集成了来自不同数据领域的数据，使用户能够找到不同系统生物学问题的答案。此外，OPTIMAS–DW 数据库中还给出了玉米特异性通路信息。该数据库的特点是能够处理不同的数据领域，还包含了几项数据分析结果，这些都对相关研究人员的工作起到支持作用，特别是系统生物学研究领域。

2.4 BIOGEN BASE–CASSAVA：木薯表型组和基因组信息资源库

BIOGEN BASE–CASSAVA 是用于研究木薯表型组学和基因组学信息的网络可访问资源库[17]，该数据库中展示了农作物木薯（Casssava）的研究成果。其中，木薯表型检索板块中，每种种质都有包括定量和定性性状在内的约 28 个表型特征。CASSAVA 数据库使用 PHP 和 MySQL 设计，并配备了广泛的搜索选项。它通过开放、通用和全球性的论坛为所有对该领域感兴趣的个人提供丰富的遗传学和基因组学数据。该数据库界面友好，所有数据均公开发布，有助于相关研究者对木薯的研究和开发。BIOGEN BASE 资源库由泰米尔纳德邦农业大学的两个研究站（Tapioca 和 Castor）维护。除木薯外，BIOGEN BASE 资源库还拥有水稻和玉米资源库以及其他数据库资源。

2.5 其他作物表型相关数据库

除以上作物组学数据库外，还有一些数据库中也包含了特有的作物表型信息。TRIM 数据库[18]，即台湾水稻插入突变体数据库，包含了有关突变体系的整合位点和表型信息，为水稻表型组学研究提供了良好资源。Gramene[19]是一个植物基因组比较基因组学数据库，提供了多种作物（如水稻、高粱和玉米等大田作物）的公开数据来源，除作物基因组学数据（如遗传标记、基因、蛋白、信号通路等）外，还包含了部分作物表型信息。Grain Genes 作为小麦家族作物信息的专门数据库，包含了小麦等麦类的分子和表型信息数据。

3 作物表型组数据库构建标准及要求

数据管理是管理、存储和共享研究数据的过程[7]。当数据研究涉及多个研究人员或在复杂环境中进行研究时，这项工作将非常具有挑战性[21]。数据的管理方法取决于整个研究过程中所涉及的数据类型、数据收集和存储方式以及数据的利用。而数据的管理情况也在一定程度上影响着研究结果。对数据进行管理有助于研究人员在后续研究中进行更好地分析和利用，确保研究质量。如果数据管理得当，研究人员可以轻松查找信息，并有助于他们得到预期结果。

如今，随着高通量植物表型获取技术的开发和应用，大规模作物表型数据相伴而生，作物表型数据量也呈指数级增长。因此，这就需要研究人员在研究期间及获取数据后对表型数据进行妥善管理。需要对从各种表型平台中获得的大量原始表型数据进行分析，而拥有最优数据管理才能实现最佳应用，从而完成对数据的深度挖掘。针对与日俱增的作物表型数据，构建作物表型组学数据库便是一项有效的数据管理措施。

3.1 表型数据的标准化和存储

通过现有的高通量作物表型信息获取平台和技术，研究人员获得的表型数据量通常高达 GB 甚至 PB，而且这些非结构化的“大数据”，通常包含大量复杂的图像、光谱和环境数据。因此，表型数据的有效存储、管理和检索成为目前研究人员需要考虑的重要问题[22]。

当前普遍接受的信息标准化原则包括 3 个方面：（1）最小信息（minimum information，MI），建议利用最小信息法来定义数据集的内容；（2）本体术语（ontology terms），采用本体术语作为数据的唯一和可重复性注释，有利于数据共享和荟萃分析；（3）数据格式（data format），选择适当的数据格式来构建数据集，如 CSV，XML，RDF 和 MAGE-TAB 等。

组织文件是数据存储的重要组成部分。在数据集中，跟踪文档及其版本至关重要，例如目录结构命名和文件命名约定。对于多站点项目，原始数据将上传并存储在文件服务器上。在通过脚本处理之后，输出文件存储在文件服务器上，研究人员可以从该文件服务器下载副本。从数据库数据标准化和存储的角度来看，基于“云技术”的存储方案正在成为植物表型数据存储发展的趋势。云存储系统可以优化作物表型平台系统架构、文件结构和高速缓存等设计。目前，各种表型数据采集平台仍然相对独立，尚未在地区、国家或大陆层面建立。通过人工智能的先进技术，建立基于多层表型信息的典型作物表型数据库，例如 GDB 人类基因组数据库，将引起相关研究人员的极大关注。

3.2 表型数据的科学管理

对于任何科学数据管理系统，都需要满足多项必要的要求[7]。

（1）数据存储和管理

数据密集型学科（如组学）中的研究活动通常会产生大量数据。有效获取、存储和管理大量数据的能力至关重要。

（2）数据背景化

需要拥有足够的上下文信息，以便更有效地组织、理解和挖掘原始数据。背景信息包括概念域模型（如研究活动如何组织和实施）和元数据（如出处信息）。

（3）数据安全

数据安全包括许多方面，如访问控制和存档。有效的数据管理系统需要通过使用身份验证和授权以及声音版本控制和备份解决方案来确保数据安全。

（4）数据识别和使用寿命

为了支持科学发现的传播，数据库中的数据需要在发布后可以公开访问，因而需要持久且唯一的命名方案。此外，有价值的科学数据也需要永久存储。

（5）数据重用和集成

上下文信息有助于理解原始数据。此外，还需要通过全文搜索、分面浏览和复杂查询应答等机制使数据可被发现，以允许集成和重用原始数据。

（6）模型可扩展性

数据管理系统可能需要管理各种各样的数据，这些数据可以由不同软件生成并由不同平台捕获。因此，表达和可扩展的域模型对于满足域概念的修改、添加和删除至关重要。此外，还需要设计数据管理系统，以便在发生此类模型更改时最大限度地减少服务中断。

3.3 表型数据库的构建规划

一个数据库的构建规划由许多元素组成，这些元素涵盖了描述、文档、过程和存档等多方面内容，因此表型数据库的构建规划中也必须具备以下几个方面。

（1）数据描述

数据的描述主要包括研究目的、数据及数据内容、数据来源、数据收集方式及形式、数据收集耗时及变化频率以及管理人员信息等。

（2）说明文档

说明文档涵盖的范围较广，主要有①创建的便于其他研究人员理解数据的文档；②元数据标准化、管理和存储方式；③文件格式及其标准；④文件命名、存储、安全和备份程序；⑤阅读或查看数据等需要的工具或软件。

（3）数据处理

诸如数据的访问、共享和重用等，都需要明确以下信息：①数据版权；②数据分享内容、时间和方式；③数据及其他信息的知识产权；④数据共享专利；⑤允许重用、再开发，或创建新工具、服务、数据集或产品等。

（4）存档

在数据的存档中，需规定：①数据归档方式；②数据存档期限及访问权限；③数据提交方式及要求；④数据保留时间等。

3.4 表型数据的共享和发布

生物技术和生物科学研究委员会（BBSRC）已实施数据共享政策。根据 BBSRC 要求，数据共享应包括以下细节：数据区域和数据类型，标准和元数据，与公共存储库中

可用的其他数据的关系，二次使用—已完成数据集的进一步预期或可预见的研究用途、数据共享方法、专有数据、时限以及数据集最终格式[23]。

4 展望

作物表型组学是一个快速发展的领域，新的表型获取手段和研究方法不断出现，只会催生越来越庞大复杂的作物表型组数据。因此，构建综合性作物表型组标准数据库，或构建特定作物的表型组数据库，将成为该领域相关研究人员的工作重点。

在形式上，理想的作物表型组数据库应具备界面友好、图文并茂、操作简单和更新及时等特征，不仅要具有多维度、多生境表型信息的存储能力，还要便于用户检索和查阅，增强数据资源的信息共享，提高来之不易的作物表型数据的利用效率。在内容上，作物表型组数据库应涵盖从微观到宏观，从显微到器官再到个体乃至群体的多维度数据，应包含作物相关的生理生化和颜色纹理等多种信息。

农业信息化是现代农业的必然发展趋势，作物表型组数据库的构建也是顺应时代发展的产物。今后，应持续关注作物表型组研究领域内的数据库相关研究，充分利用各种综合和专用数据库，并在实际研究中着力构建自己的作物表型组数据库。

参考文献

[1] Yu J., Hu S., Wang J., et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, 2002, 296(5565): 79~92.
[2] Goff S. A., Ricke D., Lan T. H., et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, 2002, 296(5565): 92~100.
[3] Schnable P. S., Ware D., Fulton R. S., et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 2009, 326(5956): 1112~1115.
[4] Paterson A. H., Bowers J. E., Bruggmann R., et al. The Sorghum bicolor genome and the diversification of grasses. *Nature*, 2009, 457(7229): 551~556.
[5] Schmutz J., Cannon S. B., Schlueter J., et al. Genome sequence of the palaeopolyploid soybean. *Nature*, 2010, 463(7278): 178~183.
[6] Ling H. Q., Zhao S., Liu D., et al. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, 2013, 496(7443): 87~90.
[7] Kim S. -L., Solehati N., Choi I. -C., et al. Data management for plant phenomics. *Journal of Plant Biology*, 2017, 60(4): 285~297.
[8] 玉光惠 , 方宣钧 . 表型组学的概念及植物表型组学的发展 . 分子植物育种 , 2009, 7(4): 639~645.
[9] Finkel E. Imaging. With 'phenomics,' plant scientists hope to shift breeding into overdrive. *Science*, 2009, 325(5939): 380~381.
[10] Furbank R. T., Tester M. Phenomics--technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, 2011, 16(12): 635~644.
[11] Letovsky S. I., Cottingham R. W., Porter C. J., et al. GDB: the human genome database. *Nucleic Acids Research*, 1998, 26(1): 94–99.
[12] Cunningham F., Achuthan P., Akanni W., et al. Ensembl 2019. *Nucleic Acids Research*, 2018.
[13] Benson D. A., Boguski M., Lipman D. J., et al. GenBank. *Nucleic Acids Research*, 1994, 22(17): 3441~3444.
[14] Cooper L., Meier A., Laporte M. A., et al. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, 2018, 46(D1): D1168~D1180.
[15] Arend D., Junker A., Scholz U., et al. PGP repository: a plant phenomics and genomics data publication infrastructure.

Database (*Oxford*), 2016: 1~10.
[16] Colmsee C., Mascher M., Czauderna T., et al. OPTIMAS-DW: a comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize. *BMC Plant Biology*, 2012, 12: 245.
[17] Jayakodi M., Selvan S. G., Natesan S., et al. A web accessible resource for investigating cassava phenomics and genomics information: BIOGEN BASE. *Bioinformation*, 2011, 6(10): 391~392.
[18] Wu H. P., Wei F. J., Wu C. C., et al. Large-scale phenomics analysis of a T-DNA tagged mutant population. *Gigascience*, 2017, 6(8): 1~7.
[19] Jaiswal P. Gramene database: a hub for comparative plant genomics. *Methods in Molecular Biology*, 2011, 678: 247~275.
[20] Lawrence C. J., Dong Q., Polacco M. L., et al. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Research*, 2004, 32(Database issue): 393~397.
[21] Yang W., Duan L., Chen G., et al. Plant phenomics and high-throughput phenotyping: accelerating rice functional genomics using multidisciplinary technologies. *Current Opinion Plant Biology*, 2013, 16(2): 180~187.
[22] Ansell P., Furbank R., Gunasekera K., et al. Flexible scientific data management for plant phenomics research. *Eswc Semantics & Big Data*, 2013.
[23] Tenopir C., Allard S., Douglass K., et al. Data sharing by scientists: practices and perceptions. *Plos One*, 2011, 6(6): e21101.

Research progress and prospect on crop phenomics database

Wang Jinglu[1, 2], Zhang Ying[1, 2], Pan Xiaodi[1, 2], Lu Xianju[1, 2], Ma Liming[1, 2], Guo Xinyu[1, 2※]
(1. Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China; 2. Beijing Key Lab of Digital Plant, Beijing 100097, China)

Abstract: [**Purpose**] During the past few years, crop phenomics has emerged as a fast growing and data intensive discipline, under which the related research techniques and analysis methods have resulted in a large number of phenotypic information. The phenotypic data are usually multiple dimensional and in different data types. For example, the image data covers a wide range of spectral information, including RGB, hyperspectral, near-infrared, thermal and fluorescent. Moreover, the plant physiological data contains various physiological indicators and other data during plant growth. Consequently, the development of biological models and data management systems in this field requires a rational use of these complex, dynamic and large-scale phenotypic data. [**Methods**] The article reviews the published studies on crop phenomics in recent years, by consulting the public literature databases including Web of Science, PubMed of NCBI and China National Knowledge Infrastructure (CNKI). According to the search results, the research status of crop phenomics both in China and overseas are analyzed. At the same time, the current crop phenomics related databases obtained from the above search results are categorized into planteome, plant genomics and phenomics research data repository, and OPTIMAS-DW. At the end of the paper, the standards and requirements for constructing the crop phenomics related database are proposed. [**Results**] In the genomics research field, there are many large, well-recognized and mature public databases, including GDB, Ensembl, and GenBank. While in the crop phenomics, although there are already some available databases, the general-purpose standardized databases with strong comprehensiveness and wide universality are

still lacking. Despite a total of 300 case studies on crop phenomics have been coded, there are only about 20 studies relevant to database construction. This number is much smaller than the number of existing genomic databases. Given most of these crop phenomics databases can be classified by species, constructing a comprehensive and standardized crop phenomics database, or building a phenomics database for a specific crop, would be of research interests in future studies. [**Conclusion**] Agricultural informatization accelerates the development of modern agriculture, which also provides opportunities for the construction of crop phenomics databases. In the future, by fully using the integrated and specialized databases, researchers should build their own crop phenomics databases in their specific studies. Building a database can not only help researchers better manage their phenotypic data, but also can benefit the data sharing among researchers.

Key words: crop; phenomics; database; construction standard; data management