

特色专题

生物多样性保护的多组学数据资源进展 ——以国家基因组科学数据中心为例

唐碧霞^{1,2}, 王钊博^{1,2,3}, 张思思^{1,2}, 张陌尘^{1,2}, 康海龙^{1,2,3}, 王彦青^{1,2}, 陈梅丽^{1,2}, 田东梅^{1,2}, 徐添翼^{1,2}, 李茹姣^{1,2}, 郝丽丽^{1,2}, 肖景发^{1,2,3}, 宋述慧^{1,2,3}, 章张^{1,2,3}, 鲍一明^{1,2,3}, 赵文明^{1,2,3*}

摘要 基因组、转录组、表观组等多组学数据,因含有丰富的生物分子遗传信息,对生物遗传多样性的研究以及生物多样性的保护具有重要意义,但这些不同层面的组学数据不仅数据量大,而且分散在不同的数据资源网站甚至文献中,对数据的获取、共享、再整合利用等带来很大的挑战。为促进生命科学领域多组学数据的快速递交和及时共享,国家基因组科学数据中心(National Genomics Data Center, NGDC)构建了多个基础数据汇交平台。同时,为了充分整合和挖掘已有组学数据资源,构建了多个家养动植物的多组学整合资源库以及知识库。目前,NGDC形成了综合的多组学大数据资源体系,这将对生物多样性保护,包括物种鉴定、濒危物种保护、育种、生境变化监测等起到重要支撑作用。

关键词 多组学;生物多样性;数据汇交;多组学数据整合;国家基因组科学数据中心;国家生物信息中心

生物多样性是人类生存和发展的物种基础,但随着全球气候变化和人类活动的加剧,自然界中野生动植物多样性急剧降低^[1],全球物种灭绝速度加快,这对人类生存和发展提出重大挑战^[2]。中国是世界上生物多样性最为丰富的国家之一,也是最早加入《生物多样性公约》的国家之一^[3]。中国在生物多样性编目、自然保护区建设以及生态建设方面取得了很好的进展^[4],构建了生物多样性与生物安全大数据平台、生物多样性检测网络等,实现对生物多样性的动态监测^[5]。

在分子生物学、细胞生物学、遗传学等学科的推

动下,生物多样性研究已经发展成为一个跨生态系统、物种和遗传多个层面的综合研究领域^[6]。目前,基于组学大数据的生物多样性研究正迅速崛起,成为该领域的一股新兴力量^[7]。这些组学大数据包括基因组^[8-10]、转录组^[11-12]、蛋白质组^[13-14]和代谢组^[15-16]等,它们为生物多样性的研究提供了丰富的信息和新的视角。其中,基因组学技术^[2]尤其是下一代测序技术(next-generation sequencing, NGS)的发展,通过高通量全基因组测序、转录组测序^[17],产生的组学数据可以应用在生物多样性保护的多个研究领域,包括种群遗传学研究、物种鉴定和保护、克隆与保护濒危物种等。此外,蛋白质组和代谢组等组学数据,能够揭示特定环境条件下生物的生理生化状态,并展示其独特的分子特征,在生物多样性应用领域具有巨大的潜力

1. 国家基因组科学数据中心,国家生物信息中心,北京 100101

2. 中国科学院北京基因组研究所,北京 100101

3. 中国科学院大学,北京 100049

收稿日期:2025-02-07;修回日期:2025-06-17

基金项目:中国科学院战略性先导科技专项(XDC0200000);中国科学院前瞻战略科技A类先导专项(XDA0460401)

作者简介:唐碧霞,高级工程师,研究方向为组学大数据整合挖掘与可视化,电子信箱:tangbx@big.ac.cn;赵文明(通信作者),正高级工程师,研究方向为生物信息大数据与人工智能,电子信箱:zhaowm@big.ac.cn

引用格式:唐碧霞,王钊博,张思思,等.生物多样性保护的多组学数据资源进展——以国家基因组科学数据中心为例[J].科技导报,2025,43(13):52-62;

doi:10.3981/j.issn.1000-7857.2025.02.00195

和广泛的应用前景^[6]。同时,这些多组学大数据资源中也包含了丰富的遗传资源数字序列信息(digital sequence information on genetic resources, DSI)^[18],对于生物多样性描述识别、生物分类、生境变化监测等生物多样性的保护意义重大^[19-20]。

国际上有3个主要生物数据中心,分别是美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)、欧洲生物信息学研究所(European Bioinformatics Institute, EBI),以及日本国立遗传学研究所的日本DNA数据库(DNA Data Bank of Japan, DDBJ)。这3个中心形成了国际核酸序列数据联盟(International Nucleotide Sequence Database Collaboration, INSDC)^[21],并在联盟的统一约束下制定生物数据标准,进行数据交换。目前,这3个数据中心收录了大量的组学数据。然而,由于国际共享数据存在的网络访问速度限制、格式与存储标准差异等问题,中国的科研人员在数据的获取、共享、再整合利用等方面存在很大的挑战。一方面,数据类型的多样性,以及格式的不统一,对数据的存储标准、管理规范带来很高的要求;另一方面,数据的整合和挖掘分析需要高度专业化的技术手段和方法,来促进数据的有效利用和科学价值的最大化。因此,国家基因组科学数据中心^[22]应运而生,并围绕着生物数据的存储与管理、整合与挖掘,以及转化与应用,构建了一系列的组学数据资源。

1 国家基因组科学数据中心的目標与使命

作为国家生物信息中心的重要组成部分,国家基因组科学数据中心(National Genomics Data Center, NGDC)的使命是面向中国人口健康和社会可持续发展的重大战略需求,建立生物信息大数据汇交存储、安全管理、开放共享与整合挖掘研究体系,研发大数据前沿交叉与转化应用的新方法和新技术,支撑公益性科学研究和产业创新发展^[22-23]。经过8年的迭代发展,NGDC的数据资源不断扩大,服务能力不断增强,成为具有一定国际影响力的数据中心。截至2024年12月,NGDC数据库体系涵盖9大类100多个数据库系统,共享的组学数据量约59 PB,服务支撑项目约25000个,支撑的论文发表数约4900篇(分布在783种期刊中),独立用户访问数达到645万,数据下载量达到209亿条。同时,组学原始数据归档库(Genome Sequence Archive, GSA)^[24],入选全球核心生物数据资源(Global Core Biodata Resources, <https://globalbiodata.org>),也是目前中国唯一入选的独立建设与运行的数据库。

2 面向生物多样性的多组学数据资源体系建设

为了满足生物多样性研究与应用的需求,NGDC构建并形成了集数据汇交、数据整合与知识挖掘的一体化多组学数据体系(图1)。表1列出了NGDC构建的部分数据库系统。



图1 面向生物多样性的多组学资源体系

表 1 NGDC 已构建的部分数据库系统

数据库类型	数据库名称	数据类型	物种覆盖范围
基础数据资源	组学原始数据归档库家族GSA Family	原始测序数据	物种不限
	基因组汇交数据库GWH	基因组拼接数据	物种不限
	基因序列汇交数据库GenBase	核酸片段序列以及小基因组	物种不限
	基因组变异数据库GVM	基因组变异	部分物种
	转录组图谱表达数据库GEN	转录	部分物种
	甲基化修饰谱数据库MethBank	表观	部分物种
	植物表型图像归档库OPIA	植物图像	部分物种
特色资源	家犬多组学数据资源库iDog	基因组、变异、转录、单细胞、表观、表型	特定物种
	绵羊多组学数据资源库iSheep	变异、表型	特定物种
	家鸡变异数据资源库ChickenSD	变异	特定物种
	大豆多组学数据资源库SoyOmics	基因组、变异、转录、表型	特定物种
	热带作物多组学数据资源库TCOD	基因组、变异、转录	特定物种
	高粱变异数据资源库SorgSD	变异、表型	特定物种
	水稻多组学数据资源库IC4R	变异、转录	特定物种
知识库	基因型-表型关联知识库GWAS Atlas	基因型-表型关联数据	部分物种
	转录-表型关联知识库TWAS Atlas	转录-表型关联数据	部分物种
	表观-表型的关联知识库EWAS Atlas	表观-表型关联数据	部分物种
	生物标志物知识库BioKA	生物标志物	部分物种
	同源基因数据库HGD	同源基因	部分物种
	植物RNA编辑体知识库PED	RNA编辑因子	部分物种

2.1 多组学数据汇交系统

在生命科学领域,组学原始数据、基因组数据、核酸序列数据、基因组变异数据、转录组数据以及脱氧核糖核酸(DNA)甲基化数据等不同类型的组学数据覆盖分子遗传信息的不同方面,对诠释复杂的生命调控机制具有重要的研究意义。同时,这些组学数据也是重要的遗传资源数字序列,构建标准规范以及安全共享的汇交存储资源管理体系,将加快促进生物多样性的遗传资源保护。

为了管理这些重要的生物数据,NGDC 规划并构建了 7 个基础数据资源库,即组学原始数据归档库家族(Genome Sequence Archive Family, GSA Family)^[24]、基因组数据汇交数据库(Genome Warehouse, GWH)^[25]、基因序列汇交数据库(GenBase)^[26]、基因组变异数据库(Genome Variation Map, GVM)^[27]、转录组图谱表达数据库(Gene Expression Nebulas, GEN)^[28]、甲基化修饰谱数据库(MethBank)^[29]以及植物表型图像归档库(Open Plant Image Archive, OPIA)^[30]。

截至 2024 年 12 月, GSA Family 包含生物研究项

目(BioProject)80 余万个,涉及生物学样本(BioSample)4343 万个,累计数据量达到 59.76 PB。GWH 覆盖的动植物等基因组完成图组装序列约 10464 条,染色体水平的基因组序列约 5376 条。GenBase 中汇交的动植物等的核酸序列 10 万条,蛋白序列 104 万条。GVM 归档了 53 万动植物样本的变异数据,注释了 23 个动物包括家犬(*Canis familiaris*)、家猪(*Sus scrofa*)等常见家养动物,以及 37 个植物包括水稻(*Oryza sativa*)、大豆(*Glycine max*)等农作物对应的变异数据。GEN 涵盖 28 个动物、10 种植物、2 个原生生物和 1 个真菌的转录组图谱数据,包括小鼠(*Mus musculus*)、家鸡(*Gallus gallus*)、水稻、小麦(*Triticum aestivum*)、拟南芥(*Arabidopsis thaliana*)、盘状网柄菌(*Dictyostelium discoideum*)、恶性疟原虫(*Plasmodium falciparum*)、酵母(*Saccharomyces cerevisiae*)等物种。MethBank 中自注释了 26 个物种,涵盖 3552 个动植物样本以及 384 个组织/细胞系对应的甲基化数据,包括人(*Homo Sapiens*)、家猪、家犬、玉米(*Zea mays*)、拟南芥等。OPIA 中审编了 11 个植物物种(水

稻、玉米、小麦等), 56 万多张表型图片信息。上述各数据库系统提供这些多样化组学数据的在线浏览、检索和下载, 极大地方便了科研人员对各类数据的使用。

2.2 动植物多组学数据系统

动植物的多组学数据可以全方位地解析物种多样性的遗传机制, 可为进一步培育优良品种提供理论基础^[5]。绵羊(*Ovis aries*)、家鸡等作为最早被人类驯化的动物, 由于其特有的经济价值, 在全世界范围内引起了广泛的研究。而水稻、高粱(*Sorghum bicolor*)等粮食作物, 以及大豆、木薯等经济作物, 更是由于其重要的经济价值和粮食安全贡献, 得到了全球科研人员的高度重视。这些研究产出了大量的组学数据, 但是数据分散在不同的文献、研究机构, 数据整合不足, 这给数据的使用带来极大不便。为了提供这些重要的家养动植物在组学研究上的工作进展, 促进组学数据的共享应用和及时更新, NGDC 构建了多个面向动物和植物的多组学大数据整合资源库, 包括 5 个动物资源库和 9 个植物资源库, 为研究动植物育种与遗传机制提供数据保障。

这些多组学大数据资源库通过使用生物信息领域标准化的组学数据分析与注释流程, 确保组学分析结果数据与注释数据的可靠性。同时, 通过使用基因、表型等信息, 将不同的组学注释信息进行关联, 从而可以全方面地了解目标基因或表型在基因组、转录组、表观组等层面上的信息, 对于进一步的育种等研究领域具有极大的促进作用。

2.2.1 家养动物多组学资源库

(1) 家犬多组学数据资源库(iDog)。

家犬作为人类历史上最早被驯化的家养动物之一, 其驯化历史已经有很多科学家进行了研究^[31]。家犬由于其特有的种群结构、基因组结构以及和人类共享的生活环境, 使其可以作为人类疾病研究的自然模型, 在人类疾病研究上做出了很大的贡献^[32]。随着对家犬研究的深入, 产出了大量的组学数据, 但国际上缺少家犬的多组学资源平台, 这些数据都分散在原始序列文件中, 不利于数据的共享。为此, NGDC 在 2019 年构建了家犬多组学数据资源库 iDog^[33], 并在 2024 年更新到了 iDog 2.0^[34], 广泛收集家犬相关的基因组、变异、表达、表观以及表型数据, 并以基因为中心, 关联了不同的组学数据。

iDog(<https://ngdc.cncb.ac.cn/idog>)收录了 2 个从头拼接基因组及其注释数据, 即豺(*Cuon alpinus*)和灰狼(*Canis lupus*)。豺的数据中, 包含了 749 个片段框架(scaffold), 18077 个预测基因, 以及其对应的基因注释信息。灰狼数据中, 包含了 581 个 scaffolds, 以及 18172 个预测基因。由于环境不断变化以及多种因素的影响, 目前全球野生豺的数量可能不足 2500 只, 中国的豺也濒临绝迹, 目前豺是中国的一级保护野生动物, 狼是二级保护野生动物。同时, 狼也是家犬驯化历史研究中的重要物种。因此, 收录和共享豺和灰狼的基因组和注释数据, 对于保护分子遗传资源、研究生物多样性, 意义重大。

iDog 2.0 全面整合分析了犬类的多组学数据资源。在基因组变异层面, iDog 2.0 收集了来自 1929 个现代样本约 2955 万个单核苷酸多态性位点(single nucleotide polymorphisms, SNPs)和 1654 万个插入/缺失(insertion-deletion, InDels), 以及来自 111 个古代样本约 2909 万个 SNPs, 同时 iDog 2.0 还鉴定并整合了 43487 个特定品种的 SNPs 和 530 个与疾病/性状相关的变异, 为全球犬类研究提供了宝贵的遗传信息。在基因表达层面, 整合分析了来自 2947 个样本的基因表达情况, 提供涉及不同犬类品种和组织的基因表达图景和不同犬类疾病的差异基因表达信息, 以及 105057 个比格犬(Beagle)海马体细胞的单细胞转录组分析结果数据。在表观基因组学层面, iDog 2.0 提供了来自 547 个样本的 DNA 甲基化信息, 以及来自 87 个样本的染色质可及性信息, 为分析犬类基因表达调控提供支持。此外, iDog 2.0 提供了 897 个犬类疾病的表型数据、3207 个基因型-表型对以及与犬类疾病关联的 349 个基因信息, 并构建了犬类品种本体和犬类疾病本体, 以实现品种和疾病标准化。iDog 2.0 还提供了 13 种在线分析工具, 包括基于检索增强生成的问答工具(dog retrieval-augmented generation, DogRAG)和基于图像的品种分类工具(dog visual classification, DogVC)^[35]等, 为犬类研究者提供便利。

(2) 绵羊多组学数据资源库(iSheep)。

绵羊是最古老的家畜之一, 其家养化历史发生在距今 8000~12000 年。作为与人类生活接触最近的动物之一, 通过分析绵羊的基因变异模式, 有助于揭示人类定居和扩张的历史。随着绵羊参考基因组的发

布,产生许多相关的组学数据,但国际上仍缺少绵羊相关的多组学数据资源,NGDC在2019年构建并发布了绵羊多组学数据资源库 iSheep^[36]。

iSheep(<https://ngdc.cncb.ac.cn/isheep>)收录了来自355个全基因组测序样本约7037万个SNPs和1231万个InDels,以及2423个芯片的变异信息。该数据资源库还收集整理26802个基因及其功能注释信息,1417个绵羊品种的表型信息,以及审编了922条基因型-表型关联信息。此外,iSheep中还提供用于品种变异数据的比较分析的在线工具。

(3) 家鸡变异数据资源库(ChickenSD)。

家鸡是一种重要的经济动物,借助组学技术(如基因组学、转录组学、蛋白组学等)可以为家鸡生长发育机制解析、肉蛋品质性状改良和疾病发生机理研究提供重要的技术支撑和研究思路。随着家鸡研究领域的深入,产出了许多的基因组学数据,但却分散在不同的文献和研究机构。为了提供一个集中整合的家鸡遗传变异数据平台,方便数据的浏览、检索和下载,2018年,NGDC构建了家鸡变异数据资源库 ChickenSD。

ChickenSD(<https://ngdc.cncb.ac.cn/chickensd>)收录了来自865个家鸡个体的基因组变异位点信息,包含约3300万个SNPs,其中约1791万个SNP位于基因区域。同时,ChickenSD还整合了从文献中审编得到的1568条基因型-表型关联信息。ChickenSD支持以染色体为单位,浏览相应的变异位点及其基因注释信息,同时还提供查询不同个体的变异分布和基因注释信息。在变异位点的详细信息页面,该数据库支持查看变异在转录本上的注释情况和蛋白质上的氨基酸变化情况,以及群体基因型频率统计信息。此外,ChickenSD还提供在线基因组浏览器用于可视化显示变异位点、基因结构以及相关的注释信息。

2.2.2 植物多组学数据资源库

(1) 大豆多组学数据资源库(SoyOmics)。

大豆是一种重要的豆类作物,大约5000年前在中国由野生大豆驯化而来。大豆含有动物所需的所有必需氨基酸,并且是食用油的重要原料。为加速大豆功能研究和分子育种,2023年,NGDC构建了大豆的多组学资源库 SoyOmics^[37],整合了多维度的基因组、变异组、转录组、表型组等数据。

SoyOmics(<https://ngdc.cncb.ac.cn/soyomics/index>)收录了27个大豆品系的从头组装基因组数据,并进行了系统的基因组注释。在遗传变异方面,SoyOmics整合了2898个样本约3800万条SNPs/InDels变异数据,以及基于大豆泛基因组分析获得的约55万条结构变异数据。在转录组层面,该资源库整合了来自大豆ZH13和Williams82 2个参考基因组的27个组织时期的表达数据,以及其他26个品系9个组织时期的表达数据,并提供了不同品系间同源基因的差异表达分析图谱。表型组数据方面,SoyOmics对115个表型多年多点测定的约2.7万条表型记录进行了标准化的本体注释和系统归类,并实现了表型数据与变异数据的数据关联。此外,SoyOmics还提供了部分种质资源的全基因组甲基化测序数据,以及Soy40K大豆芯片数据。这些多组学数据的整合为大豆功能基因组研究和分子育种提供了重要的数据支撑和分析平台。

(2) 热带作物多组学数据资源库(TCOD)。

天然橡胶、油棕、甘蔗、木薯等热带作物是全球重要的工业原料来源,而咖啡、可可、胡椒和香蕉等则是重要消费品。目前,这些作物的多组学数据分散在不同的数据中心或网站,导致数据的整合和使用困难。为此,NGDC在2023年构建了热带作物的多组学资源平台TCOD^[38],整合了基因组、转录组及表型数据,并支持在线分析工具。

TCOD(<https://ngdc.cncb.ac.cn/tcod>)整合了15种热带作物的多组学数据资源,涵盖木薯(*Manihot esculenta*)、咖啡(*Coffea canephora*)、可可(*Theobroma cacao*)、香蕉(*Musa acuminata*)等物种。该数据资源库目前收录了34个染色体级别基因组组装、125万个基因的功能注释信息、基于2048个全基因组重测序样本鉴定的2824万个变异位点、88个转录组测序项目的基因表达谱数据以及13381份种质资源的描述信息。通过以基因为桥梁整合多种组学数据,TCOD提供了一站式的数据浏览、检索和下载等服务。此外,TCOD还构建了不同物种基因组间的同源基因关系,支持研究人员基于基因功能、通路注释、基因组变异和基因表达等多维度数据,深入探索不同物种之间共有的生物学特征,促进跨物种研究开展。

(3) 高粱变异数据资源库(SorgSD)。

高粱是世界第五大粮食作物,其谷物产量和种植面积仅次于玉米、水稻、小麦和大麦,广泛分布于美洲、非洲、亚洲和大洋洲等地区。高粱具有优异的农艺学和生物学特性,如耐热性和耐旱性,成为贫瘠地区重要的粮食作物。随着高粱参考基因组(BTx623)的不断完善和高通量测序技术的发展,关于高粱驯化过程及其独特表型遗传机制的研究取得了显著进展,并积累了大量的组学数据。为系统整合这些数据资源,NGDC于2016年构建了首个高粱变异数据库SorgSD^[39],并于2021年完成了版本升级^[40]。

SorGSD(<https://ngdc.cncb.ac.cn/sorgsd>)整合了来自289个高粱品系的大规模基因组变异数据,包含约3382万个SNPs和572万个InDels,并提供了详细的变异注释信息。此外,SorgSD还收录了357个高粱样本对应的表型数据,为高粱功能基因组研究提供重要资源。在功能服务方面,该资源库提供基因标识符在线转换、同源基因检索以及序列比对的在线服务,支持高粱基因组学研究的深入开展。

SorGSD全面整合了大规模基因组变异和表型信息,并集成了数据挖掘、基因组导航及在线分析工具。该资源库为高粱功能基因组学研究、分子育种及种质资源保护提供了重要数据支撑,对推动高粱遗传改良和品种选育具有重要意义。

(4) 水稻多组学数据资源库(IC4R)。

水稻是单子叶植物和谷类作物的重要模式生物,也是全球大部分人口的主要粮食作物,具有重要的经济价值。近年来,多组学技术(如基因组学、转录组学和蛋白质组学)的快速发展,为深入解析水稻重要农艺性状的遗传调控机制、种质资源多样性及驯化过程奠定了重要基础。为促进相关研究,NGDC于2016年构建并发布了水稻的多组学数据资源库IC4R^[41]。

IC4R(<https://ngdc.cncb.ac.cn/ic4r/>)收录了来自5524个样本的约854万个SNPs信息,以及17套覆盖8个组织的RNA-seq测序数据,包含1002万个基因的表达谱和1145万个转录本的表达特征。该资源库还系统整理了17个水稻同源物种的14739组同源基因信息,1676个真核生物蛋白激酶(PKs)和1814个泛素相关酶的注释数据,以及3746个蛋白质磷酸化位点。通过社区协作,该资源库已完成1005个基因的

功能注释,并整合了35717篇相关研究文献。为方便研究人员使用,IC4R提供了BLAST序列比对、基因标识符转换和特异性基因检索等在线分析工具,实现了数据的便捷访问和分析。

IC4R包含的遗传变异信息将有助于水稻种群的遗传多样性评估,了解不同品种或种群之间的遗传变异和亲缘关系,从而为保护具有独特遗传特性的水稻品种和种群提供科学依据。同时,IC4R中包含的转录表达谱信息,可以用于了解水稻的适应性,有助于培育更适应环境变化的水稻品种。

2.3 面向生物多样性应用的知识库体系

2.3.1 表型关联知识库

表型是刻画物种遗传性状和环境相互作用结果的重要指标^[42],然而越来越多的表型相关的研究成果却分散在已公开发表的文献中,不利于知识的检索和快速分享。为此,NGDC通过文献审编的方式,构建了基因型-表型关联知识库GWAS Atlas(<https://ngdc.cncb.ac.cn/gwas>)^[43],转录-表型关联知识库TWAS Atlas(<https://ngdc.cncb.ac.cn/twas>)^[44],表观-表型的关联知识库EWAS Atlas(<https://ngdc.cncb.ac.cn/ewas>)^[45],形成覆盖较全面的表型关联知识体系。目前,GWAS Atlas中收录了5个家养动物如家猪、家鸡等,26个家养植物如苹果(*Malus domestica*)、棉花(*Gossypium hirsutum*)等,总共302295条基因型-表型知识对。TWAS Atlas中收录了来自人的135个组织401266条转录关联表型知识对。EWAS Atlas中收录了来自人的3767个队列的752935条表观关联表型知识对。这些知识信息,对于重要性状的遗传研究保护和育种应用以及人类健康和疾病研究具有重要价值。如Xu等^[46]利用GWAS Atlas中水稻对应的基因型-表型数据作为训练数据集,利用深度学习方法构建了用于预测水稻生物胁迫性状相关SNPs的工具,促进水稻研究和抗性品种培育。Zaghlool等^[47]利用EWAS Atlas中审编整合的甲基化位点数据,从研究数据中发现了在众多研究中报道过的215种疾病关联,经过进一步的研究测试,识别出一些重要网络,这些网络可能最终成为针对慢性低度炎症的个性化治疗靶点。

2.3.2 异常表型相关联的生物标志物知识库

生物标志物(Biomarker)是具有被客观测量并评价等特点的指标,可用于指示生理/病理过程以及治疗

干预的药理反应^[48]。当前生物标志物的研究成果分散于已公开发表的文献中,且现有的生物标志物知识库都聚焦在人类疾病相关的生物标志物中,在非人类物种疾病和分子设计育种领域仍是空白。为此,2023年,NGDC通过文献审编的方式,构建了生物标志物数据库 BioKA^[49],不仅扩充了人类生物标志物信息,更填补了已有的生物标志物数据资源在非人动物疾病和分子育种方面的空缺。

BioKA(<https://ngdc.cncb.ac.cn/bioka>)通过人工审编 4747 篇文献和资源整合,收录涵盖人和 30 个动物物种,包括 9 种模式动物如小鼠、斑马鱼(*Danio rerio*),以及 16 种家养动物如猫(*Felis catus*)、牛(*Bos taurus*)的 16296 个生物标志物,涉及 951 个疾病/性状,包括 11925 个基因/蛋白、1784 个 miRNA、1043 个突变、773 个代谢物、357 个 circRNA 和 127 个 lncRNA,并提供了经过标准化处理的 308 个品种及其对应生物标志物信息。这些知识信息被用在了癌症能量代谢^[50]、山羊奶产量^[51]等研究,为人类疾病机制解析、动物疾病研究及分子育种提供了重要数据支持。

2.3.3 跨物种比较的同源基因知识库

同源基因是指来源于共同祖先的基因,常用于研究基因进化过程以及推断基因的潜在功能,在进化基因组研究以及系统生物学的功能研究中具有重要价值^[52]。由于目前已有的同源基因数据库中,同源关系和同源基因标识方面存在差异、同源推断方法多样以及部分数据库同源基因鉴定结果不一致,制约了数据的有效利用。为此,NGDC 于 2022 年整合了多个高质量的同源基因数据集,构建了同源基因数据库 HGD^[53],提供一个综合全面且可对数据来源进行溯源的同源基因整合结果。

HGD(<https://ngdc.cncb.ac.cn/hgd>)整合了人、常见动植物以及微生物共 37 个物种的同源基因以及相应的注释信息,包括 10 种重要模式生物如小鼠、斑马鱼、果蝇(*Drosophila melanogaster*)等。该知识库共收录了 11238 万个同源对,其中包括 16909 个同源基因关联表型数据,276670 个同源基因关联变异数据,398573 个同源基因关联表达数据,536852 个同源基因关联 GO 数据。该知识库支持通过物种名、基因 ID、基因功能、蛋白 ID、蛋白名以及相关注释等多条件检索。HGD 提供的同源基因信息已被用于单细胞

转录组下游功能分析的参考数据,在胃部炎症-癌变转化研究^[54]、肾脏异种移植^[55]等有相应的研究应用。

此外,HGD 中所包含的丰富的跨物种同源基因信息,可以助力濒危物种的保护,可通过已知物种的同源基因功能推测濒危物种的关键基因功能,为物种保护工作提供科学依据。同时,HGD 还可以辅助入侵物种的监测和防控,通过比较入侵物种与本地物种的同源基因,了解其适应性和竞争力,制定更有效的监测和防控措施,防止入侵物种的进一步扩散等。

2.3.4 植物 RNA 编辑体知识库

核糖核酸(RNA)编辑在植物的生长、开花、细胞器发育和应激反应等过程中发挥重要作用。RNA 编辑因子通过组装形成复杂多样的编辑复合体(RNA editosome)并对特异的 RNA 位点行使编辑功能。目前植物 RNA 编辑因子研究仍缺乏系统性的知识库,为此,NGDC 通过文献审编结合植物细胞器基因组注释的方法,构建了植物的 RNA 编辑体知识库 PED^[56]。该知识库全面整合了 RNA 编辑因子及相关数据,为研究人员提供植物 RNA 编辑领域的最新研究资源。

PED(<https://ngdc.cncb.ac.cn/ped>)整合了 8 种模式植物(拟南芥、水稻、玉米等)中鉴定的 98 个 RNA 编辑因子,提供 RNA 编辑因子注释、编辑因子与编辑位点相互作用、编辑因子对植物功能和表型影响等关键信息,并附相关实验证据。此外,该知识库还系统收录了来自 1621 个植物物种的 20836 个 RNA 编辑事件,这些位点分布于 203 种细胞器基因,覆盖 1673 个植物细胞器。通过对 RNA 编辑因子、RNA 编辑事件及其互作关系的标准化处理,PED 为解析 RNA 编辑的分子机制及系统进化研究奠定了重要数据基础。

PED 知识库的建立将为植物遗传资源的精准保护与利用提供重要依据,能够有效支持关键遗传元件的鉴定工作,包括调控植物生长发育和环境适应的功能基因;同时为评估遗传资源对环境压力^[57]、病虫害等的适应性提供数据支撑。

3 结论

多组学数据蕴含丰富的分子遗传信息,是解析生物遗传多样性的重要资源。目前,NGDC 通过标准化数据汇交、多组学数据整合分析以及文献审编,构建了

覆盖基因组、转录组等维度的 100 多个数据库系统(截至 2024 年),其中包含家犬、大豆等 14 个特色数据资源库。目前共享的组学数据总量达 59 PB(截至 2024 年),形成了一体化多组学数据资源体系。这些资源不仅为物种鉴定、濒危物种保护及分子育种等提供了良好的数据基础,也为中国生物数据库的标准化建设提供了数据质量控制、跨库关联分析等关键技术经验。

在物种鉴定与分类研究方面,GWH 库收录的超过 1 万个基因组完成图组装序列数据(截至 2024 年),为分子鉴定提供了权威参考序列。与传统形态学鉴定方法相比,这些基因组数据可有效区分形态相似但遗传分化显著的物种。在进化生物学领域,GVM 提供的 SNPs 和 InDels 数据,有助于分析种群遗传结构、基因流模式和历史种群动态。在濒危物种保护领域,GWH 和 GenBase 通过汇交和镜像收录了华南虎(*Panthera tigris amoyensis*)、中华穿山甲(*Manis pentadactyla*)等中国特有物种的基因组资源,其变异模式分析,可为评估遗传多样性、近交程度及有害突变负荷提供依据。农业分子育种中,GWAS Atlas、BioKA、PED 等知识库提供的变异、生物标志物以及 RNA 编辑因子,将加速优良品种的选育过程,实现从“经验育种”到“精准育种”的转变。

NGDC 在建设过程中坚持自主创新与国际标准协同推进。通过自主研发构建了 GSA、GWH 等核心资源库,实现数据本地化存储,提升国内科研人员的访问效率。同时,在数据库的建设过程中,积极对接国际标准,如 GSA 与 INSDC 的数据交换,确保数据的互操作性和国际认可度。为提高数据应用效率,NGDC 开发了生物信息在线工具平台 BIT(<https://ngdc.cncb.ac.cn/bit/>),提供如 Blast、RNA-seq 分析等工具,方便用户使用数据。此外,GSA 提供 FTP、Aspera 高速传输工具,配合 NCBI 镜像数据服务,显著提高国内用户的下载速度。

然而,在多组学资源体系的建设和发展过程中,也有一些需要改进的地方。首先,对于数据汇交和共享来说,NGDC 目前汇交的数据大部分都是来源于发表文章的需求,有大量的序列数据仍然保留在研究人员或者机构手中,数据汇交共享的动力不足。其次,随着人工智能在生物信息领域的应用,尤其是 Gene Compass 等大语言基座模型的出现^[58],对于数据的存

储格式、数据质量等都提出了更高的要求。在兼顾传统的数据使用需求时,如何紧跟时代步伐,促进已构建的多组学数据体系在 AI for Science 生命科学领域的应用,是发展过程中需要思考的问题。此外,在组学数据的整合和知识挖掘中,虽然已有比较完善的组学数据分析流程和方法,但在数据展示上还主要是以组学数据解析结果的浏览和检索为主,缺少对多组学数据的深层次整合和应用挖掘。

未来,NGDC 将会继续扩大宣传力度,研究数据共享应用机制,提高数据汇交共享动力,以及进一步扩充更多的物种和数据资源,提供更加丰富的数据服务,来促进多组学资源体系在生物多样性领域的应用。同时,NGDC 还将紧跟时代步伐,在 AI+时代,围绕数据的汇交、整合分析与挖掘应用,研制自动化以及智能化的审编工具,提高数据汇聚效率和数据质量。如研发 GSA smart 工具,对单细胞组学分析提交的元数据信息进行智能审核、纠错、标注,提高元数据信息质量。NGDC 还将以已有数据资源为基石,构建高质量的 AI-ready 语料数据集,以赋能生命科学领域的大语言模型,从而推动该领域的研究和应用。如针对基因序列库 GenBase,将构建基于序列的原始语料和序列注释坐标位置信息的序列结构语料,为预训练模型的研发提供生物数据支持。进一步地,NGDC 还将开发基于深度学习以及大语言模型的数据分析挖掘方法,如开发自动审编工具,提高知识发现和知识库更新的效率以及智能检索工具,实现语义检索功能。此外,NGDC 还将考虑与已有的生物多样性平台进行数据整合和共享,构建全面的从宏观的生物多样性动态检测,到微观的生物多样性遗传信息的数据访问通道,这将对我国生物多样性保护具有重要意义。同时,NGDC 还将积极跟进我国发布的 DSI 相关的管理法律法规体系,提供遗传资源数字序列信息的安全共享,在促进生物遗传资源的互惠互利上^[59]发挥积极作用。

参考文献(References)

- [1] Jing Z Y, Cheng K G, Shu H, et al. Whole genome resequencing approach for conservation biology of endangered plants[J]. Biodiversity Science, 2023, 31(5): 22679.
- [2] 李律瑄, 李招应. 基因组学技术在生物多样性保护研究应用

- 中面临的挑战及应对策略[J]. *China-Arab States Science and Technology*, 2024, 1: 89–93.
- [3] 刘张璐, 赵兰勇, 朱秀芹. 中国生物多样性及其保护规划发展研究现状[J]. *中国园林*, 2010, 26(1): 81–83.
- [4] 王庆刚, 池秀莲, 程蒙, 等. 中国生物多样性保护与利用现状及挑战[J]. *中国食品药品监管*, 2023(3): 118–123.
- [5] 米湘成, 冯刚, 张健, 等. 中国生物多样性科学研究进展评述[J]. *中国科学院院刊*, 2021, 36(4): 384–398.
- [6] Liu S L, Qiu N, Zhang S Y, et al. Application of genomics technology in biodiversity conservation research[J]. *Biodiversity Science*, 2022, 30(10): 22441.
- [7] Seymour M. Rapid progression and future of environmental DNA research[J]. *Communications Biology*, 2019, 2: 80.
- [8] Stein J C, Yu Y, Copetti D, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*[J]. *Nature Genetics*, 2018, 50(2): 285–296.
- [9] Schreiber M, Jayakodi M, Stein N, et al. Plant pangenomes for crop improvement, biodiversity and evolution[J]. *Nature Reviews Genetics*, 2024, 25(8): 563–577.
- [10] Appels R, Nystrom-Persson J, Keeble-Gagnere G. Advances in genome studies in plants and animals[J]. *Functional & Integrative Genomics*, 2014, 14(1): 1–9.
- [11] Wang Y, Huan Q, Li K, et al. Single-cell transcriptome atlas of the leaf and root of rice seedlings[J]. *Journal of Genetics and Genomics*, 2021, 48(10): 881–898.
- [12] Shaw R, Tian X, Xu J. Single-cell transcriptome analysis in plants: Advances and challenges[J]. *Molecular Plant*, 2021, 14(1): 115–126.
- [13] Dargue R, Grant I, Nye L C, et al. The analysis of acetaminophen (paracetamol) and seven metabolites in rat, pig and human plasma by U(H)PLC-MS[J]. *Bioanalysis*, 2020, 12(7): 485–500.
- [14] Adak T, Mahanty A, Sarkar S, et al. Development and validation of HS-SPME-GCMS/MS method for quantification of 2-acetyl-1-pyrroline in rice cultivars[J]. *Journal of Food Science and Technology*, 2023, 60(3): 1185–1194.
- [15] Li Z L, Sun J S, Li K B, et al. Metabolomic analysis reveals spermatozoa and seminal plasma differences between Duroc and Liang Guang Small-spotted pig[J]. *Frontiers in Veterinary Science*, 2022, 9: 1078928.
- [16] Guo Z L, Wang S C, Zhang F, et al. Common and specific genetic basis of metabolite-mediated drought responses in rice[J]. *Stress Biology*, 2024, 4(1): 6.
- [17] Sharma R, Patil C, Majeed J, et al. Next-generation sequencing in the biodiversity conservation of endangered medicinal plants[J]. *Environmental Science and Pollution Research International*, 2022, 29(49): 73795–73808.
- [18] 吴林寰, 石蕾, 高孟绪, 等. 浅析遗传资源数字序列信息的发展现状与建议[J]. *中国科技资源导刊*, 2021, 53(2): 36–43.
- [19] Li B P, Xue D Y. Application of digital sequence information in biodiversity research and its potential impact on benefit sharing[J]. *Biodiversity Science*, 2019, 27(12): 1379–1385.
- [20] 郝丽丽, 鲍一明. 专家解读 | 生物多样性保护与遗传资源数字序列信息[EB/OL]. (2024–04–17) [2025–06–16]. https://www.mee.gov.cn/zcwj/zcjd/202404/t20240417_1070941.shtml.
- [21] Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration[J]. *Nucleic Acids Research*, 2021, 49(D1): 121–124.
- [22] 陈梅丽, 王彦青, 李茹姣, 等. 国家基因组科学数据中心: 现状及展望[J]. *中国科学*, 2024, 54: 1–13.
- [23] BIG Data Center Members. The BIG Data Center: From deposition to integration to translation[J]. *Nucleic Acids Research*, 2017, 45(D1): 18–24.
- [24] Chen T T, Chen X, Zhang S S, et al. The genome sequence archive family: Toward explosive data growth and diverse data types[J]. *Genomics, Proteomics & Bioinformatics*, 2021, 19(4): 578–583.
- [25] Chen M L, Ma Y K, Wu S, et al. Genome warehouse: A public repository housing genome-scale data[J]. *Genomics, Proteomics & Bioinformatics*, 2021, 19(4): 584–589.
- [26] Bu C F, Zheng X C, Zhao X T, et al. GenBase: A nucleotide sequence database[J]. *Genomics, Proteomics & Bioinformatics*, 2024, 22(3): qzae047.
- [27] Li C P, Tian D M, Tang B X, et al. Genome Variation Map: A worldwide collection of genome variations across multiple species[J]. *Nucleic Acids Research*, 2021, 49(D1): 1186–1191.
- [28] Zhang Y S, Zou D, Zhu T T, et al. Gene expression nebulas (GEN): A comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels[J]. *Nucleic Acids Research*, 2022, 50(D1): 1016–1024.
- [29] Zhang M C, Zong W T, Zou D, et al. MethBank 4.0: An updated database of DNA methylation across a variety of species[J]. *Nucleic Acids Research*, 2023, 51(D1): 208–216.
- [30] Cao Y R, Tian D M, Tang Z X, et al. OPIA: An open archive of plant images and related phenotypic traits[J]. *Nucleic Acids Research*, 2024, 52(D1): 1530–1537.
- [31] Wang G D, Zhai W W, Yang H C, et al. Out of southern east Asia: The natural history of domestic dogs across the

- world[J]. Cell Research, 2016, 26(1): 21–33.
- [32] Cadieu E, Ostrander E A. Canine genetics offers new mechanisms for the study of human cancer[J]. Cancer Epidemiology, Biomarkers & Prevention, 2007, 16(11): 2181–2183.
- [33] Tang B X, Zhou Q, Dong L L, et al. iDog: An integrated resource for domestic dogs and wild canids[J]. Nucleic Acids Research, 2019, 47(D1): 793–800.
- [34] Liu Y H, Wang Y B, Sun J N, et al. iDog: A multi-omics resource for canids study[J]. Nucleic Acids Research, 2025, 53(D1): 1039–1046.
- [35] Cui Y, Tang B X, Wu G G, et al. Classification of dog breeds using convolutional neural network models and support vector machine[J]. Bioengineering, 2024, 11(11): 1157.
- [36] Wang Z H, Zhu Q H, Li X, et al. iSheep: An integrated resource for sheep genome, variant and phenotype[J]. Frontiers in Genetics, 2021, 12: 714852.
- [37] Liu Y C, Zhang Y, Liu X N, et al. SoyOmics: A deeply integrated database on soybean multi-omics[J]. Molecular Plant, 2023, 16(5): 794–797.
- [38] Kang H L, Huang T H, Duan G Y, et al. TCOd: An integrated resource for tropical crops[J]. Nucleic Acids Research, 2024, 52(D1): 1651–1660.
- [39] Luo H, Zhao W M, Wang Y Q, et al. SorGSD: A *Sorghum* genome SNP database[J]. Biotechnology for Biofuels, 2016, 9: 6.
- [40] Liu Y M, Wang Z H, Wu X Y, et al. SorGSD: Updating and expanding the *Sorghum* genome science database with new contents and tools[J]. Biotechnology for Biofuels, 2021, 14(1): 165.
- [41] Sang J, Zou D, Wang Z N, et al. IC4R-2.0: Rice genome reannotation using massive RNA-seq data[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 161–172.
- [42] 蒋方圆, 王丽娟, 孙静, 等. 全表型组关联研究方法学研究进展[J]. 中华流行病学杂志, 2022, 43(7): 1154–1161.
- [43] Tian D M, Wang P, Tang B X, et al. GWAS Atlas: A curated resource of genome-wide variant-trait associations in plants and animals[J]. Nucleic Acids Research, 2020, 48(D1): 927–932.
- [44] Lu M M, Zhang Y D, Yang F C, et al. TWAS Atlas: A curated knowledgebase of transcriptome-wide association studies[J]. Nucleic Acids Research, 2023, 51(D1): 1179–1187.
- [45] Xiong Z, Yang F, Li M W, et al. EWAS Open Platform: Integrated data, knowledge and toolkit for epigenome-wide association study[J]. Nucleic Acids Research, 2022, 50(D1): 1004–1009.
- [46] Xu J J, Gao Y J, Lu Q, et al. RiceSNP-BST: A deep learning framework for predicting biotic stress-associated SNPs in rice[J]. Briefings in Bioinformatics, 2024, 25(6): bbae599.
- [47] Zaghlool S B, Kühnel B, Elhadad M A, et al. Epigenetics meets proteomics in an epigenome-wide association study with circulating blood plasma protein traits[J]. Nature Communications, 2020, 11(1): 15.
- [48] Califf R M. Biomarker definitions and their applications[J]. Experimental Biology and Medicine, 2018, 243(3): 213–221.
- [49] Wang Y B, Lin Y H, Wu S C, et al. BioKA: A curated and integrated biomarker knowledgebase for animals[J]. Nucleic Acids Research, 2024, 52(D1): 1121–1130.
- [50] Zhang X, Zhang Y C, Liu Q, et al. Glycolysis-associated lncRNAs in cancer energy metabolism and immune microenvironment: A magic key[J]. Frontiers in Immunology, 2024, 15: 1456636.
- [51] Pozovnikova M V, Leibova V B. Analysis of the Associations between the miR-767, miR-335-3p, and miR-106b-5p microRNA Expression Patterns and Milk and Serum Metabolites in Goat (*Capra hircus*)[J]. Russian Journal of Genetics, 2025, 61(2): 213–227.
- [52] Koonin E V. Orthologs, paralogs, and evolutionary genomics[J]. Annual Review of Genetics, 2005, 39: 309–338.
- [53] Duan G Y, Wu G G, Chen X N, et al. HGD an integrated homologous gene database across multiple species[J]. Nucleic Acids Research, 2023, 51(D1): 994–1002.
- [54] Hu M M, Xu S Y, Xu R F, et al. Analysis of single-cell RNA sequencing data to examine the gastric inflammation-to-cancer transition and evaluation of the effect of probiotic on precancerous lesions[J]. Engineering Microbiology, 2025, 5(3): 100208.
- [55] Pan W Q, Zhang W M, Zheng B H, et al. Cellular dynamics in pig-to-human kidney xenotransplantation[J]. Med, 2024, 5(8): 1016–1029.
- [56] Li M, Xia L, Zhang Y S, et al. Plant editosome database: A curated database of RNA editosome in plants[J]. Nucleic Acids Research, 2019, 47(D1): 170–174.
- [57] Hong Z H, Zhu L Y, Gao L L, et al. Chloroplast precursor protein preClpD overaccumulation triggers multilevel reprogramming of gene expression and a heat shock-like response[J]. Nature Communications, 2025, 16(1): 3777.
- [58] Yang X D, Liu G L, Feng G H, et al. GeneCompass: Deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model[J]. Cell Research, 2024, 34(12): 830–845.
- [59] Sett S, Kress W J, Halewood M, et al. Harmonize rules for digital sequence information benefit-sharing across UN frameworks[J]. Nature Communications, 2024, 15(1): 8745.

National Genomics Data Center: Multi-omics data resources to support biodiversity conservation

TANG Bixia^{1,2}, WANG Yibo^{1,2,3}, ZHANG Sisi^{1,2}, ZHANG Mochen^{1,2}, KANG Hailong^{1,2,3}, WANG Yanqing^{1,2}, CHEN Meili^{1,2},
TIAN Dongmei^{1,2}, XU Tianyi^{1,2}, LI Rujiao^{1,2}, HAO Lili^{1,2}, XIAO Jingfa^{1,2,3}, SONG Shuhui^{1,2,3}, ZHANG Zhang^{1,2,3},
BAO Yiming^{1,2,3}, ZHAO Wenming^{1,2,3*}

1. National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China

2. Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

3. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Multi-omics data, including genomics, transcriptomics, and epigenomics, contain a wealth of genetic information and are crucial for the study of biological genetic diversity and the conservation of biodiversity. However, the integration and reuse of multi-omics data remain challenging due to their massive volume and fragmented distribution across disparate databases and published literature. To facilitate timely submission and sharing of multi-omics data in life science, National Genomics Data Center (NGDC) has established multiple foundational data submission platforms. Additionally, to fully integrate and explore existing omics data resources, NGDC has developed several multi-omics integration resources and knowledge bases for domesticated animals and plants. Currently, NGDC has preliminarily formed a comprehensive multi-omics big data framework, which will significantly contribute to the conservation of biodiversity, including species identification, the protection of endangered species, breeding programs, and monitoring changes in habitats.

Keywords multi-omics; biodiversity; data submission; multi-omics data integration; National Genomics Data Center; China National Center for Bioinformation ●



(责任编辑 徐丽娇)