DOI: 10.13376/j.cbls/2023169
文章编号：1004-0374(2023)12-1553-08

李亦学，研究员，博士生导师。现为广州实验室研究员、国家生物数据中心体系粤港澳节点平台首席科学家兼主任，中国生物信息学会(筹)副理事长，上海生物信息学会理事长，上海交通大学教授，复旦大学遗传学教育部协同创新中心前沿生物技术部主任。主要研究领域为生物信息学、基因组学、精准医学、人工智能、大数据、知识图谱等。曾任国家“十五”863计划生物和农业技术领域生物信息技术主题专家组组长，国家“十一五”863计划生物医药技术领域专家组专家；国家蛋白质科学研究重大专项“模式生物和细胞等功能系统的系统生物学研究”“代谢生理活动与病理过程中信号转导网络的系统生物学研究”两任专项项目首席科学家。获得上海市自然科学奖一等奖、二等奖，教育部自然科学奖等，并荣获全国五一国际劳动奖章，上海市劳动模范，第一批上海市“科教兴市领军人才”等。

张国庆，研究员，博士生导师。现任中国科学院上海营养与健康研究所生物医学大数据中心副主任，中国科学院计算生物学重点实验室副主任，上海生物医学大数据工程技术研究中心执行主任，中国遗传学会生物大数据专业委员会委员，中国医药生物技术协会生物医学信息技术分会委员。主要研究方向是生物医学数据库与知识库，包括精准医学、自然及疾病人群队列、人类表型组、环境及病原及人体微生物组等领域的数据库和知识库的研发，致力于多维生命组学数据、文献数据、健康与医疗等真实世界数据的集成与管理，以及以知识库为代表的数据科学关键技术研究。

多组学大数据共享平台研究进展

凌鋆超[1#]，曹瑞芳[1#]，李亦学[1,2*]，张国庆[1*]
(1 中国科学院上海营养与健康研究所，生物医学大数据中心，上海 200032；2 广州实验室，广州 510005)

摘　要：高通量检测技术的快速发展催生了海量的多组学数据，数据驱动型研究规模正逐步超越传统假设型研究。不同层次组学数据的组合，通过对系统生物学和疾病发展更深入和全面的解读，持续改变生物医学研究方式。同时，多组学数据庞大的数据规模、异质的数据特性，以及强烈的数据共享内源性需求，都推动组学数据向规模化、平台化、标准化共享的方向发展。该文首先介绍了代表性的多组学平台和各组学数据的特点，接着以多维组学数据百科全书NODE为例，从多组学数据融合和多组学数据安全共享两方面对相应的方法和技术进行了细致的阐述，并展望了多组学数据平台未来的发展方向。

关键词：多组学；大数据；数据平台；数据共享
中图分类号：Q811.4　　**文献标志码：**A

收稿日期：2023-03-11；**修回日期：**2023-10-30
基金项目：国家重点研发计划(2018YFC2000205)
***通信作者：**E-mail: gqzhang@sinh.ac.cn；Tel: 13524783378

Advances in multi-omics big data sharing platform research

LING Yun-Chao[1], CAO Rui-Fang[1], LI Yi-Xue[1,2*], ZHANG Guo-Qing[1*]
(1 Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China; 2 Guangzhou Laboratory, Guangzhou 510005, China)

Abstract: High-throughput sequencing technologies have spurred a data-driven shift in biomedical research, producing vast quantities of multi-omics data that offer a comprehensive understanding of biological systems. The sheer volume and diversity of this data, along with the need for data sharing, call for standardized, scalable omics platforms. In this review, we first highlight key multi-omics platforms and their distinct features. Then, we explore methods for integrating and securely sharing multi-omics data using National Omics Data Encyclopedia (NODE), a multi-omics data repository. Lastly, we delve into the future prospects of multi-omics platforms.

Key words: multi-omics; big data; data platform; data sharing

数据驱动型的研究规模已经逐步赶超传统假设型的研究，宇宙探索、环境气候、生物科学等领域都明显发展出数据密集的特性。在生命科学领域，高通量检测技术的高速发展，一方面使得在单项目检测数十万个实验样本成为可能，另一方面单一样本可以进行测序、质谱等不同组学的检测。实验技术的进步正在帮助研究者快速收集各种组学数据，其中包括大规模人群的基因组数据，也包括转录组学产生的基因表达数据和表观基因组学产生的基因调控数据，还可能包括对疾病终点事件有直接影响的蛋白质组数据和代谢组数据，以及与人存在共生关系的人源微生物组数据和赖以生存的环境微生物组数据。这些数据都用独特的方式细致刻画了人与环境的关联或复杂疾病的发病机制，涉及不同层次组学概念的若干级联事件，也为系统生物学研究和疾病发生发展的分子机制研究提供了更深入和更全面的理解，并正在逐步改变生物医学研究的方式。

由于生物系统和生物学问题的复杂性，数据驱动的假设验证往往需要从不止一个角度进行验证，而是从生物学中心法则出发，沿着基因组、转录组、蛋白质组、代谢组、影像组等不同过程和不同维度，使用不同类型的组学数据多方验证假设的正确性和精准性。在基于大规模人群队列的基础生物学研究和临床研究中，为了在海量的临床信息和组学信息中挖掘出特异性的分子标记，也需要来自第三方的、不同人群、不同特征的组学数据集作为对照。多组学数据作为最高通量、最高精度的高价值生物信息检测结果，也具有强烈的数据共享内源性需求。但另一方面，组学数据作为大数据的一种，具有数据规模差异大和数据异质性强两方面的特点。这在组学实验、测量仪器使用、数据生产、数据存储以及数据使用上对研究人员提出了更高的要求。这些因素的共同作用正是多组学数据共享的内在动力，并且推动组学数据向多组学联合研究以及大规模、平台式、标准化共享的方向发展，同时众多的组学类型以及海量的仪器平台、实验参数、样本类型，也对多组学平台的建设提出了重大的挑战。为此，我们研发了多维组学数据百科全书 (National Omics Data Encyclopedia, NODE)，作为公益性的开放的多组学大数据共享平台，提供公共服务。

1 代表性组学数据平台

国际上对生物大数据的研究，特别是临床、表型数据与组学数据的整合方面，形成了许多规模较大的项目和数据库，其中 TCGA、CPTAC、UK Biobank 等都是国际化大型项目和数据库的典范。不同数据库因其组学数据收录范围、组学数据类型、元数据标准、存储方式不同而差异化并形成各自的特色 (表 1)。

1.1 生命科学测序数据库

SRA 数据库是美国国家生物技术信息中心 (NCBI)2009 年建立的测序数据的存储库 [1]，由国际核酸序列数据库协会 (International Nucleotide Sequence Database Collaboration, INSDC)，即美国国家生物信息技术中心 (NCBI)、欧洲生物信息研究所 (EMBL-EBI)、日本 DDBJ 中心共同运行维护。SRA 收录的全球测序数据规模已经超过 70 PB[1]，并在此数据基数上仍以每年接近 50% 的增长率持续扩增规模 [2]。基因组序列库 (GSA)[3] 是由中国国家基因组科学数据中心建立及维护的二代测序数据的归档数据库群，还包括了存储人类遗传学相关数据的 GSA-Human 数据库。CNGBdb 是华大基因国家基因库

表1 代表性组学数据平台

序号	数据平台	国家	URL	主要数据类型
1	SRA	美国	https://www.ncbi.nlm.nih.gov/sra	测序数据
2	TCGA	美国	https://portal.gdc.cancer.gov	肿瘤基因组
3	CPTAC	美国	https://proteomics.cancer.gov	肿瘤蛋白质
4	UK Biobank	英国	https://www.ukbiobank.ac.uk	基因型及表型信息
5	CBioPortal	多机构团队	https://www.cbioportal.org	多维癌症基因组学数据
6	JGI	美国	https://jgi.doe.gov	基因组和宏基因组
7	GSA	中国	https://ngdc.cncb.ac.cn/gsa	测序数据
8	PRIDE	英国	https://www.ebi.ac.uk/pride	蛋白质组
9	iProX	中国	https://www.iprox.cn/	蛋白质组
10	Metabolomics Workbench	美国	https://www.metabolomicsworkbench.org/	代谢组
11	NODE	中国	https://www.biosino.org/node	多维组学数据
12	CNGBdb	中国	https://db.cngb.org	测序数据

(CNGB) 运行维护的一个公共的、非营利的、开放的生物数据平台，集成了“存储、读取和写入”海量生物资源的能力，即保存生物样本、生物信息和生物活体资源，破译和利用遗传信息。

在蛋白质组数据方面，成立于 2006 年的 ProteomeXchange 联盟 [4](以下简称 PX) 就是为实现蛋白质组公共资源库在原始数据、结果和元数据三个层面的统一共享而创立。当前 PX 联盟成员包括：PRIDE [5]、PeptideAtlas [6]、MassIVE、jPOST [7]、iProX [8]、Panorama Public [9]。

代谢组数据仓库管理代谢组原始数据及相关实验信息，例如实验设计、样品、用户、仪器等。目前国际上已有的代谢组数据仓库包括 MetaboLights [10]、Metabolimcs Workbench[11]、Metabolonote [12] 等。

1.2 临床组学数据库

癌症基因组图谱 (TCGA) 是目前国际上规模与影响最大的表型与基因组数据结合的项目，其目标是通过更好地了解疾病的遗传基础，提高诊断、治疗和预防癌症的能力 [13]。TCGA 收集了来自 86 513 个病例的超过 55 种癌症类型的大规模基因组、转录组测序数据、表型数据以及综合分析数据等。

cBioProtal 癌症基因组学网站 [14] 是一个开放的资源网站，用于交互探索多维癌症基因组学数据集，提供了来自 367 个癌症研究的 18 万多个肿瘤样本的数据访问。cBio 癌症基因组学门户网站可以协助研究人员便捷地获取大规模癌症基因组学项目的分子特征和临床属性，以便研究人员能够将这些丰富的数据集转化为生物学见解和临床应用。

临床蛋白质组学肿瘤分析联盟 (CPTAC) 主要通过应用定性和定量的蛋白质组学技术和分析流程，加快癌症分子的基础理解，提高诊断、治疗和预防癌症的能力。截至 2022 年底，CPTAC 共收录来自 14 个肿瘤部位的超 4 600 份样本。除蛋白质组数据外，CPTAC 也通过基因组数据共享中心 (GDC) 向公众提供全基因组测序 (WGS)、全外显子组测序 (WXS) 等组学数据。

1.3 人群队列数据库

英国生物样本数据库 (UK Biobank, UKB) 是目前世界上已建成最大的人类遗传队列生物样本库和生物医学数据库资源。该项目收集了来自英国各地，年龄在 40~69 岁之间，大约 50 万志愿者的基因型遗传数据以及生活方式、遗传信息等深入表型信息。该数据库定期更新数据，并在全球范围内供经批准的研究人员使用 [15-16]。

1.4 微生物组数据库

DOE JGI 成立于 1997 年，整合了三个美国国家实验室在 DNA 测序、信息学和技术开发方面的专业知识和资源 [17]，包括 Genome Portal、Phytozome、IMG 和 Genomes OnLine 数据库，在植物、真菌、微生物基因组和宏基因组这几个领域的生物体测序数量上处于世界领先地位。

1.5 多组学数据库

组学数据百科全书 (NODE) 是一个综合各类组学特征和样本类型的可比较及扩展的多组学资源汇交管理平台，由中国科学院上海营养与健康研究所运行维护。NODE 将项目、样本、实验、运行、原始数据和分析数据以层次化的结构进行组织，既支撑同一宿主、个人等的多样本的数据，又支持单样本多种组学的数据。

2 NODE集成的多组学数据

不同组学的数据实质上代表的是研究领域和实验技术的差异，表 2 展示了常见的组学数据及其特征。基因组、转录组、表观基因组、宏基因组、宏转录组等数据的核心都是由高通量测序仪生产的海量序列数据，由于仪器本身的检测精度高、通量大，且成本日趋下降，逐渐成为多组学数据中使用最为广泛、数据量最大的一种数据类型。随着技术的不断发展，逐渐衍生出单细胞测序、空间转录组测序等新方法，可以获得特定微环境下的细胞序列差异以方便研究其功能差异 [18-19]。蛋白质组 [20] 和代谢组数据 [21-22] 通常由质谱平台产出原始格式 (RAW) 的质谱数据，并经过后续的定性和定量分析，结合不同的参考数据库鉴定不同蛋白质 / 代谢物的组成及其丰度，并结合下游生物信息学方法或同级分析方法分析蛋白质组 / 代谢组数据与疾病、表型等的关联。暴露组从概念上并没有对其数据类型进行精细化的定义，因此暴露组学的研究通常是对广义暴露因素及其对应数据之间采取的关联分析研究 [23-24]，常见的数据类型为甲基化芯片数据或是基因组芯片数据等。不同的组学类型在采样方法、实验参数、实验仪器、产出文件格式等方面差异极大，这也造成了不同组学数据存储和共享标准上的融合困难。目前 NODE 支持包括基因组、转录组、表观基因组、宏基因组、宏转录组等的组学测试数据，基于质谱技术的蛋白质组、代谢组及芯片数据的共享和发布。

3 NODE的多组学数据整合策略

单一的组学技术只能为相关机制研究提供特定类型分子的检测和发现能力。来自遗传学、蛋白质组学和代谢组学等各种信息源的数据的整合和分析，可以帮助研究人员获得对生物系统的更全面的检测，从而发现新的生物标志物。这些生物标志物有可能帮助进行准确的疾病预测、病人分层和精准医疗 [25-27]。多组学研究指的是从中心法则出发，从生物学过程的不同维度，通过设计不同组学的实验去验证同一个生物学假设，从而确保假设的验证不因组学维度的不同而造成认识的偏差。同时，沿着中心法则中生物学过程发生的先后关系，更能有效地发现数据关联之间的因果关系。因此，从多组学数据研究的角度来看，多组学的融合体现的是生物学过程的融合；从多组学数据生产的角度来看，多组学的融合体现的是同一样本或同一个体不同样本在不同组学数据上的融合；从多组学数据平台融合技术来看，多组学数据的融合方法主要体现在数据标准的融合和数据结构的融合。

3.1 数据标准的整合

不同类型组学的数据标准，不论是原始数据格式还是元数据标准都存在很强的异质性。NODE 进行多组学数据融合时，在领域数据格式和元数据标准上尽量与传统单组学数据库标准达成统一和兼容。如全基因组测序数据、全外显子测序数据、转录组测序数据等，同 SRA、TCGA、GEO 等数据库标准保持一致；蛋白质组实验则是跟随蛋白质组学公共资源库通用框架 PX 联盟的数据共享标准 [28]；代谢组数据兼容了 Metabolomics-Workbench 数据标准；流式细胞数据可参考 FlowRepository 库数据标准 [29] 等。对专业领域特异的数据格式和元数据标准，无须强行进行组学之间的标准映射。同时，针对其中每一种组学类型，在样本和实验元数据上，NODE 也参考了各类型非冗余的最小化信息标准，确保元数据的完整性满足后续数据分析的需要。例如，微生物组学相关元数据信息可参考基因组标准联盟 (GSC) 的最低信息检查表和环境包，即 MIxS 标准 [30]。使用领域内已成型的数据标准，保证数据汇交过程中可使用标准化的术语对数据进行描述，

表2 组学数据特征表

序号	组学	数据生产方法	常见数据格式	单样本数据量
1	基因组	高通量测序等	Fastq、bam、siff等	10 GB级
2	转录组	高通量测序等	Fastq、bam等	GB级
3	蛋白质组	鸟枪法质谱等	Raw、 wiff/wiff2等	100 MB-GB级
4	宏基因组	高通量测序等	Fastq、 bam等	10 MB-GB级
5	宏转录组	高通量测序等	Fastq、 bam等	10 MB-GB级
6	代谢组	质谱或核磁影像等	Raw、wiff/wiff2、dcm等	10 MB-GB级
7	暴露组	甲基化芯片等	Idat、bam等	10 MB-GB级
8	影像组学	核磁影像等	dcm等	

降低和避免因跨领域研究造成的数据汇聚质量问题。

各组学领域数据标准的研究并不意味着组学之间相互独立不进行关联。为了促进不同组学数据之间的融合，方便研究人员从自身研究关心的角度搜索和获取数据，NODE 也从数据的产生过程、检测目标、检测方法等角度为不同组学数据赋予共同的标签。常见的标签可分为 9 类：(1) 组学标签用于记录数据对应的组学类型，如基因组、蛋白质组等；(2) 检测平台标签用于记录数据产生的平台，如 Illuminna HiSeq 4000、BGISEQ-500 等；(3) 数据产生方法标签用于记录实验的方法，如 WGS、ChIP-Seq 等；(4) 数据格式标签记录数据格式，如 fastq.gz、raw、vcf 等；(5) 物种标签记录采样个体对应的物种信息，如人类、小鼠等；(6) 样本信息用于记录采样对应的组织 / 器官，如血液、心脏、肝脏等；(7) 疾病标签用于记录采样源的疾病信息，如肝癌、新冠肺炎等；(8) 数据源标签用于记录组学数据的来源信息，如 SRA、TCGA 等；(9) 自定义标签信息用于记录常见分类，如精准医学、环境微生物组等。将相同标签的多组学数据进行融合，结合关键词搜索，基本能够满足研究者对数据的筛选过滤和统计使用需求，为不同领域建立沟通的桥梁，优化组学数据的跨领域应用。

3.2　数据结构的融合

除了在内容层面对多组学数据标准进行融合以外，NODE 在多组学数据平台系统架构设计层面，根据多组学数据之间的关联和依赖关系，设计了各层级相对独立的数据架构，在对 SRA、TCGA、PX 联盟等国际知名的组学数据库兼容的同时，兼顾标准化存储。对不同组学数据及数据库的底层存储方式进行汇总后我们发现，在总体存储架构上，不论何种组学数据，都能够按照数据类型分为元数据层、原始数据层、分析数据层三个不同的层次，顺应组学数据的产生、使用和安全共享问题，在尽可能保留数据信息的基础上，实现低冗余的数据存储 (图 1)。三个层次之间互相依赖，以组学数据仪器直接生产的原始数据为核心，向上对接元数据层，记录与原始数据的产生直接相关的项目、实验、研究对象、样本、仪器上机批次等信息；向下对接分析数据层，记录使用一种或多种组学原始数据，经过数据分析后的中间数据和分析结果数据。在元数据的维度上，由于多组学不仅代表了多种组学数据在同一项目中的样本再组合，更体现了相同研究对象甚至同一样本在不同组学实验中的复用，因此在数据结构的设计上应将研究对象个体维度及样本维度的元信息作为一级结构及其子结构独立出来，以多元分类结构替代原有的线性层级结构，满足多组学数据融合时的管理和搜索需求。

基于此融合后的 NODE 多组学数据结构，用户可以根据应用切换数据视图：从项目角度，用户能够总览项目产出组学数据，评估项目数据产出的进程和规模；从多组学融合角度，用户还可以按样品 (例如组织、器官等) 发现多组学实验之间的关系，或按实验类型来揭示样品之间的差异；也可以按宿主或个体组织，以显示来自不同器官、组织或细胞的多个样本；从研究成果的汇总角度，用户可以将下机原始数据和分析结果数据进行关联和归类，加强数据的分层分级和二次利用。

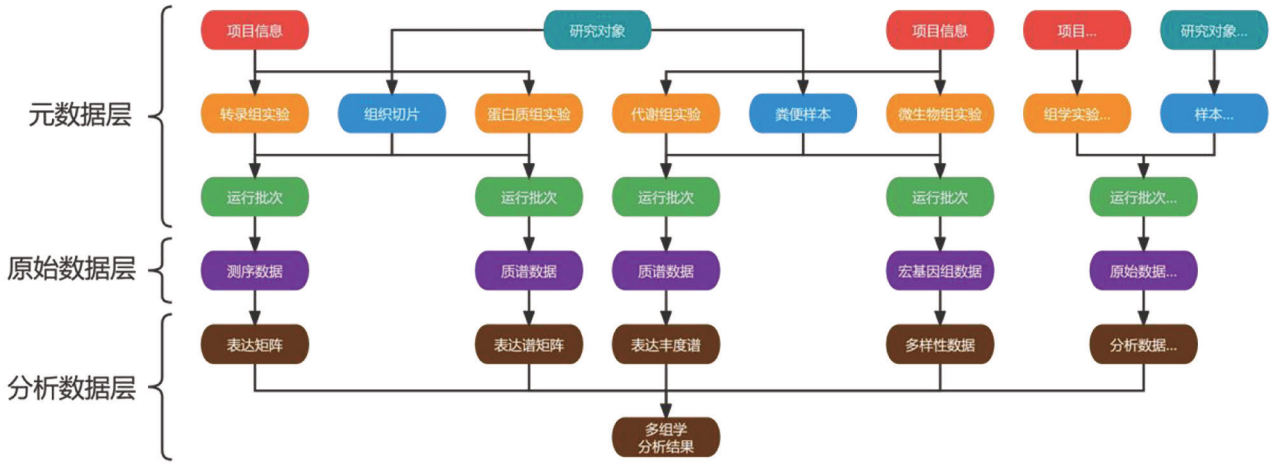

图1　多组学数据结构融合框架图

4 多组学数据安全共享

由于多组学研究样本数量多、领域跨度大、生产成本高、可复用性强等基础特性，决定了对于多组学数据平台中存储的海量数据，除了为数据所有者提供数据归档管理服务外，对于其他平台用户，在不同应用场景下，都有较强的数据共享需求。因此，多组学数据平台为了在合法合规的前提下达到数据的最大化利用，在平台建设时，需要采取一系列技术手段，来保护多组学数据的隐私和机密性，保护用户数据权益不受侵害，保护国家数据安全的前提下，实现数据的安全共享。

4.1 数据的分层分级保护和元数据的开放共享

上节中提到多组学数据从结构上分为元数据、原始数据和分析数据三类，针对这三类数据，由于保护目的不同，NODE 采用了不同的安全策略。对于高度结构化的元数据，如项目、实验、样本等元信息，在数据脱敏的前提下，需要尽量满足信息的可获得性，实现元数据的开放共享，保证用户在申请原始或分析数据时有足够的信息判断数据共享需求的必要性。对于多组学原始数据和分析数据，由于数据产生的方法、测量的精度、分析目的的不同，NODE 支持分层分级的安全策略，分层即是对原始数据和经过加工的分析数据，或是同级数据不同样本，允许用户设置独立的安全保护和共享接口；分级即是允许用户对多组学原始数据和分析数据根据自身研究或者发布需要设置不同的安全等级 (公开、受限、私有)，以保证数据所有者在评估共享需求时，既能满足用户的应用需求，又不过度释放过多的数据权限，以保证共享双方的利益。

4.2 伴随数据生命周期的数据共享

在多组学数据平台的建设上，考虑到用户数据发布和共享的全流程中对数据安全的需求，NODE 设计了伴随数据生命周期的数据发布和共享功能。现有的组学数据平台一般至少将数据保护状态分为公开数据 (public) 和受限数据 (restricted) 两种安全等级，部分平台根据元数据的可见与否 (数据的可搜索性) 进行分类控制。NODE 在此基础上进一步将受限数据在平台内部细分为受限数据 (restricted) 和私有数据 (private)。在数据共享功能上，NODE 支持根据数据共享发起人和目标人的情况，进行数据请求 (request)、数据共享 (share)、数据审阅 (review) 三类操作。通过对数据保护状态和数据共享功能的有机组合，NODE 可以满足用户在数据生产期、数据分析期、数据发布期、数据共享期等数据生命周期的不同阶段，在保证数据安全性的前提下实现数据的有效共享。

4.3 数据的加密和安全传输

数据共享即意味着访问权限控制和数据的安全传输。由于用户在平台上被授权访问的数据范围是高度特异的，NODE 在操作系统级权限控制的基础上，叠加独立的软件级用户模块和权限控制模块，允许数据所有者控制和调整其数据的访问权限。在数据的制备和传输方面，由于组学数据在数据容量上的特殊性，需要平衡计算资源和网络资源，或采用数据加密方式为每次共享需求快速制备独立的加密数据，并为终端用户提供高效的解密工具；或在数据传输的过程中实时加密传输的数据流，确保在整个传输过程中不会因为网络广播的监听造成数据泄露。NODE 针对数据的公开状态和共享情况，在软件层为每位注册用户建立了与数据共享状态匹配的数据空间，并使用底层 HTTPS 协议、SFTP 协议实现加密数据的流式安全传输。

4.4 数据共享协议、伦理和规范审查

除了从技术上对数据共享进行规范外，在数据共享行为发生前，应由数据收集或产出方的伦理委员会进行伦理审查。这涉及到对人类或动物研究的伦理问题进行评估和解决。伦理审查委员会可以审核和批准多组学数据共享计划，并确保其符合伦理标准和道德原则。同时，作为人类遗传资源的一部分，NODE 提示用户在对人类相关组学数据进行共享时，需要根据《人类遗传资源管理条例》(http://www.gov.cn/zhengce/content/2019-06/10/content_5398829.htm) 的相关要求完成备案。对于大规模、高价值组学数据的共享，应预先签署数据共享协议，明确数据共享的目的、范围、访问条件、保密条款等内容。数据共享协议需要明确保护数据隐私和机密性的方法，以及对协议违反者的惩罚措施。

5 结论与展望

NODE建设的初衷是在一个平台上集成不同类型的组学数据，通过跨领域的研究加强成果的可信度，促进研究者之间的数据共享和协作，提高高质量组学数据的复用率，避免数据的重复采集。在组学技术快速革新的过程中，以高通量二代测序数据，SRA、TCGA 等数据库为代表，已经建立了基因组、转录组等测序组学数据的通用标准，累积了 PB 级规模的数据，并实现了基本的数据共享；在多组学

的数据整合方面，以 OMICSDI[31] 为代表的数据搜索引擎也以数据集为基本单位，从数据分类的角度实现了跨组学，甚至多组学的数据融合，基本满足了用户对数据的搜索需求。然而，通过对现有代表性多组学数据平台、多组学数据、数据融合、数据安全共享等方面的综合分析，也可以看到当前多组学数据共享平台的研究和发展还处于初级阶段。在新型数据生产方法 (如单细胞测序技术、时空组学技术) 以及异构组学数据 (如暴露组数据、影像组数据) 的一致性数据标准建设上还远不能达到完全的统一；在多组学数据的融合上，也由于不同平台对个体、样本等水平的存储方式上存在差异，而对元数据层和数据层的整合提出了更高的要求；在多组学数据共享时，更应着重考虑数据贡献者的贡献和权益，以技术手段而非强制性方法，在打消用户对数据安全顾虑的同时，维护国家数据安全，促进多组学数据共享。对于生物医学这个数据驱动型复合研究方向，大数据技术正不断发挥其作用：使用去中心化的平台建设方法体现数据共享者的贡献价值，并等量兑换为其可享受的权益；使用多维度、跨尺度的数据标准规范化多组学数据的生产和治理，融合更多类型的组学数据；在数据平台以外提供更多类型的数据分析流程、工具和资源，以多组学的方法促进组学数据的共享。总之，多组学数据共享平台的发展将持续帮助研究人员更好地理解生物学和疾病，并促进不同领域之间的协作和数据共享。

[参 考 文 献]

[1] Katz K, Shutov O, Lapoint R, et al. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res, 2022, 50: D387-90

[2] Sayers EW, Bolton EE, Brister JR, et al. Database resources of the National Center for Biotechnology Information in 2023. Nucleic Acids Res, 2023, 51: D29-38

[3] Members CN, Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. Nucleic Acids Res, 2023, 51: D18-28

[4] Deutsch EW, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic Acids Res, 2020, 48: D1145-52

[5] Perez-Riverol Y, Bai J, Bandla C, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res, 2022, 50: D543-52

[6] Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep, 2008, 9: 429-34

[7] Moriya Y, Kawano S, Okuda S, et al. The jPOST environment: an integrated proteomics data repository and database. Nucleic Acids Res, 2019, 47: D1218-24

[8] Ma J, Chen T, Wu S, et al. iProX: an integrated proteome resource. Nucleic Acids Res, 2019, 47: D1211-7

[9] Sharma V, Eckels J, Schilling B, et al. Panorama public: a public repository for quantitative data sets processed in skyline. Mol Cell Proteomics, 2018, 17: 1239-44

[10] Steinbeck C, Conesa P, Haug K, et al. MetaboLights: towards a new COSMOS of metabolomics data management. Metabolomics, 2012, 8: 757-60

[11] Sud M, Fahy E, Cotter D, et al. Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res, 2016, 44: D463-70

[12] Ara T, Enomoto M, Arita M, et al. Metabolonote: a wiki-based database for managing hierarchical metadata of metabolome analyses. Front Bioeng Biotechnol, 2015, 3: 38

[13] Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet, 2013, 45: 1113-20

[14] Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov, 2012, 2: 401-4

[15] Schaefer SM, Kaiser A, Behrendt I, et al. Association of alcohol types, coffee and tea intake with mortality: prospective cohort study of UK Biobank participants. Br J Nutr, 2022, 129: 115-25

[16] Yuan S, Larsson SC. Adiposity, diabetes, lifestyle factors and risk of gastroesophageal reflux disease: a Mendelian randomization study. Eur J Epidemiol, 2022, 37: 747-54

[17] Nordberg H, Cantor M, Dusheyko S, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res, 2014, 42: D26-31

[18] Eberwine J, Sul JY, Bartfai T, et al. The promise of single-cell sequencing. Nat Methods, 2014, 11: 25-7

[19] Rao A, Barkley D, Franca GS, et al. Exploring tissue architecture using spatial transcriptomics. Nature, 2021, 596: 211-20

[20] Anderson JD, Johansson HJ, Graham CS, et al. Comprehensive proteomic analysis of mesenchymal stem cell exosomes reveals modulation of angiogenesis via nuclear factor-κB signaling. Stem Cells, 2016, 34: 601-3

[21] Jordan KW, Nordenstam J, Lauwers GY, et al. Metabolomic characterization of human rectal adenocarcinoma with intact tissue magnetic resonance spectroscopy. Dis Colon Rectum, 2009, 52: 520-5

[22] Niu Y, Jiang Y, Xu C, et al. [Preliminary results of metabolite in serum and urine of lung cancer patients detected by metabolomics]. Zhongguo Fei Ai Za Zhi,

2012, 15: 195-201
[23] Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev, 2005, 14: 1847-50
[24] Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. Science, 2010, 330: 460-1
[25] Gao Q, Zhu H, Dong L, et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell, 2019, 179: 561-77.e22
[26] Jiang YZ, Ma D, Suo C, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. Cancer Cell, 2019, 35: 428-40.e5
[27] Li C, Sun YD, Yu GY, et al. Integrated omics of metastatic colorectal cancer. Cancer Cell, 2020, 38: 734-47.e9
[28] Deutsch EW, Bandeira N, Perez-Riverol Y, et al. The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res, 2023, 51: D1539-48
[29] Spidlen J, Breuer K, Rosenberg C, et al. FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. Cytometry A, 2012, 81: 727-31
[30] Yilmaz P, Kottmann R, Field D, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol, 2011, 29: 415-20
[31] Perez-Riverol Y, Bai M, da Veiga Leprevost F, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol, 2017, 35: 406-9