科学通报 2024 年 第 69 卷 第 30 期: 4432 ~ 4446
清华大学生物系复系40周年专辑 评 述
《中国科学》杂志社 SCIENCE CHINA PRESS

复杂疾病中多组学多模态数据的生物信息学研究进展

刘晓帆[1,2], 鲁志[1,2*]

1. 清华大学生命科学学院, 合成和系统生物学中心, 教育部生物信息学重点实验室, 北京 100084;
2. 清华大学精准医学研究所, 北京 100084
* 联系人, E-mail: zhilu@tsinghua.edu.cn

2024-04-16 收稿, 2024-06-11 修回, 2024-08-20 接受, 2024-08-27 网络版发表

摘要 随着高通量测序技术的发展, 多组学多模态数据的整合已成为复杂疾病研究的重要趋势, 为深入理解疾病的发生发展提供了新视角, 为实现复杂疾病的精准诊疗提供了重要支持. 本文首先介绍了复杂疾病研究中的不同组学类型, 如基因组学、转录组学、蛋白质组学、代谢组学、微生物组学、影像组学等, 以及相应的多组学数据库. 然后本文对多组学、多模态数据的整合方法进行了系统的分类, 详细阐述了基于关联分析和网络的方法, 以及基于数据矩阵和机器学习的方法中早期整合、中期整合和后期整合方法. 此外, 本文还讨论了多组学整合模型在疾病筛查、分型、预后和药物反应预测等方面的应用. 最后, 本文总结了当前多组学整合面临的挑战, 分为样本层面、数据层面和模型层面三类, 并展望了未来的发展方向. 本文为复杂疾病中多组学、多模态数据整合研究提供了系统的梳理, 对该领域的进一步发展具有重要意义.

关键词 多组学数据, 多模态数据, 生物信息, 复杂疾病, 整合模型

随着精准医疗的迅速发展, 越来越多研究人员开始利用不同维度的生物学数据深入解析复杂疾病的生物学机制. 复杂疾病, 如癌症、心血管疾病、神经退行性疾病和自身免疫病, 通常是由遗传变异、环境因素和生活方式等多个因素共同作用引起的. 近年来, 各类组学技术的应用为理解复杂疾病的机制提供了新的视角和工具. 例如, 基因组学通过分析基因序列揭示了疾病相关的遗传背景和基因变异; 转录组学通过研究表达模式揭示了疾病相关的基因调控关系; 蛋白质组学关注蛋白质的表达、修饰和相互作用; 代谢组学通过代谢物变化反映了患病前后代谢通路的调整; 放射组学通过医学影像展示了疾病引起的变化等. 对这些组学数据进行整合分析有利于弥补单一组学的信息缺失, 更加全面地理解疾病的发生、发展和病理过程. 复杂的多组学数据给其分析过程带来了许多困难, 例如维数灾难、异质性、数据缺失等[1]. 近年来, 国内外已有很多针对多组学整合的生物信息学方法研究, 尤其是针对癌症、阿尔兹海默病等复杂疾病的诊断、分型和预后. 本文综述了针对复杂疾病研究的多组学数据的生物信息学方法, 从组学分类、数据获取、整合方法、疾病应用、挑战和未来方向等方面依次进行阐述.

1 复杂疾病中的多组学

随着高通量技术的不断发展和完善, 我们可以通过各种实验手段获取到与复杂疾病的发生发展密切相关的各种组学数据(图1). 综合分析这些多组学数据有助于从多个角度深入理解复杂疾病所涉及的各种生物学过程, 从而在精准医疗领域发挥重要作用. 我们接下来从多组学数据类型和数据获取两方面来展开介绍.

引用格式: 刘晓帆, 鲁志. 复杂疾病中多组学多模态数据的生物信息学研究进展. 科学通报, 2024, 69: 4432–4446
Liu X F, Lu Z. Progress of bioinformatics studies for multi-omics and multi-modal data in complex diseases (in Chinese). Chin Sci Bull, 2024, 69: 4432–4446, doi: 10.1360/TB-2024-0416

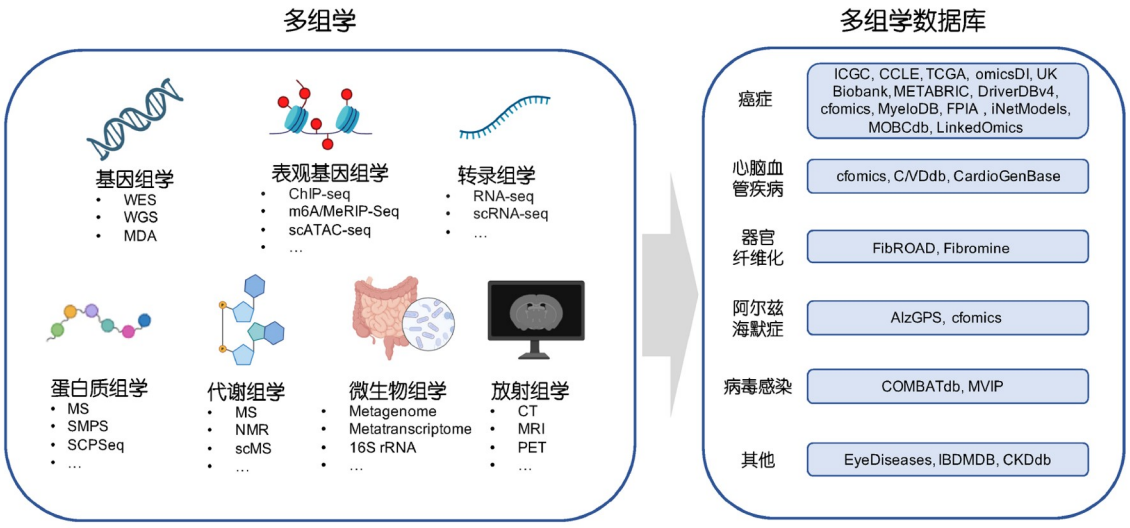

图 1 复杂疾病研究中多组学方法. 左侧展示了不同组学类型数据和其测序方法, 右侧展示了针对不同类型复杂疾病的多组学数据库. 图中部分元素来自Biorender.com

Figure 1 Multi-omics approaches in complex disease research. The left side displays different types of omics data and their sequencing methods, while the right side displays multi-omics databases for complex diseases. Some of the elements in the image are from Biorender.com

1.1 多组学类型

基因组学(genomics)旨在研究生物体内所有遗传信息, 可以通过全基因组测序(whole genome sequencing, WGS)和全外显子组测序(whole exome sequencing, WES)技术获取. 在复杂疾病研究中, 基因组学刻画了基因、遗传变异、环境和生命体之间的相互作用, 可应用于疾病预防、诊断和治疗, 解决疾病机制研究、遗传风险探索和新疗法开发等科学问题. 通过全基因组关联算法研究疾病表型和单核苷酸多态性(single nucleotide polymorphism, SNP)、拷贝数变异(copy number variation, CNV)等不同特征之间的关联, 我们可以识别复杂疾病相关的易感基因和调控元件[2]. 随着精准医疗的发展, 研究者开始不满足于仅在个体水平上研究各种疾病, 而是开始深入探索不同细胞亚群的内在机制. 研究者利用重置换扩增(multiple dilacement amplification, MDA)和全基因组扩增(whole genome amplification, WGA)等技术分析单细胞基因组变异, 发展出了单细胞基因组学, 可以揭示疾病中不同细胞亚型的异质性[3].

表观基因组学(epigenomics)旨在研究生物体内基因表达调控的表观遗传修饰, 包括DNA甲基化、染色质可及性和组蛋白修饰等. 表观基因组学也被证明与多种复杂疾病存在关联, 如心血管疾病[4]、癌症[5]等. 研究者可以通过全基因组亚硫酸盐测序(whole genome bisulfite sequencing, WGBS)、染色质可及性测序(assay for transposase-accessible chromatin with high throughput sequencing, ATAC-seq)、组蛋白修饰测序(chromatin immunoprecipitation sequencing, ChIP-seq)、单细胞全基因组亚硫酸盐测序(single-cell bisulfite sequencing, scBS-seq)和单细胞染色质可及性测序(single-cell assay for transposase-accessible chromatin sequencing, scATAC-seq)等技术定量分析生物体和细胞内的表观遗传调控机制及其在疾病状态下的变化.

转录组学(transcriptomics)旨在研究生物体内所有RNA的表达模式和功能, 解释疾病引起的分子动态变化. 研究者通过各种RNA-seq技术对各类RNA的表达值进行定量, 包括编码蛋白RNA(mRNA)、非编码RNA和环状RNA. 在复杂疾病研究中, 转录组学刻画了疾病相关基因的表达模式和调控网络, 使得研究者可以探索疾病中不同基因的功能和表达调控等科学问题. 作为遗传信息从DNA到蛋白质的传递者, mRNA不仅可以作为复杂疾病诊断和预后的生物标志物[6], 还可以作为疾病预防的疫苗[7]. 类似地, 非编码RNA也在复杂疾病中扮演关键角色[8]. 此外, RNA可变剪接和RNA可变多聚腺苷酸化等转录后调控事件是转录组的重要部分, 决定了mRNA的成熟、稳定性和翻译效率, 并且对疾病的变化产生影响. 例如, RNA异常剪接导致了肌

营养不良症中DMD基因编码的重要蛋白质-肌营养蛋白的缺失或功能受损[9]; 3′UTR区域的RNA可变多聚腺苷酸化也被认为可以影响肿瘤增殖[10]. 随着技术的发展, 单细胞转录组学和单细胞空间转录组学也越来越受研究者关注. 单细胞转录组学可以检测疾病中特定细胞类型的转录本, 单细胞空间转录组学可以在更高的空间分辨率上展示细胞在组织中的具体分布和相互间的空间关系. 在复杂疾病的研究领域, 他们也展现了巨大潜力, 例如在肿瘤内部的异质性、肿瘤的转移扩散及对治疗的耐药性方面[11].

蛋白质组学(proteomics)旨在研究生物体中所有的蛋白质, 通过质谱(mass spectrometry, MS)、单分子蛋白质测序(single-molecule protein sequencing, SMPS)、蛋白质芯片、单细胞质谱(single cell mass spectrometry, scMS)和单细胞蛋白质测序(single cell protein sequencing, SCPSeq)等技术来获取. 研究蛋白质组学可以帮助科学家理解复杂疾病. 例如, 在癌症研究中, 蛋白组学能揭示肿瘤标志物和潜在的治疗靶点, 如乳腺癌中的HER2蛋白[12]. 在阿尔茨海默病中, 异常的蛋白质聚集, 如Tau蛋白的异常沉积[13], 与病理过程密切相关. 此外, 各种翻译后修饰, 例如磷酸化、糖基化和亚硝基化等, 可以影响细胞内信号转导和蛋白质酶活性等过程, 也揭示了糖尿病[14]、癌症[15]等疾病的潜在机制.

代谢组学(metabolomics)主要关注生物体中参与代谢过程的各种小分子的变化, 可以通过核磁共振波谱(nuclear magnetic resonance, NMR)、质谱、单细胞荧光显微成像等高通量技术获取. 研究者通过研究代谢组学可以监控疾病进展. 例如, 胰岛素抵抗相关的代谢物变化和糖尿病发病直接相关[16]; 血脂代谢异常与心脏病发作风险增加有关[17]; 代谢物的失调既可用于早期癌症诊断也可作为癌症治疗靶点[18]; 在神经退行性疾病中, 脑部代谢物的变化是揭示阿尔茨海默病进程的标志物[19].

微生物组学(microbiomics)是针对包括人体在内的各种环境中的微生物群落进行研究的学科, 通常可以通过16S rRNA测序、宏基因组测序和宏转录组测序等技术获取微生物数据. 多种复杂疾病都与人体内微生物组密切相关, 例如肠道微生物组的失衡不仅和结肠直肠癌[20]、糖尿病[21]相关, 还可能通过与中枢神经的相关作用影响情绪和认知功能, 导致抑郁症、焦虑症和神经退行性疾病[22]. 此外, 口腔微生物组的变化可能会导致口腔癌[23]和心血管疾病[24]风险增加.

除了分子层面的生物数据, 人们开始关注更多不同类型的医学数据, 通过计算机断层扫描(computed tomography, CT)、磁共振成像(magnetic resonance imaging, MRI)、正电子发射断层扫描(positron emission tomography, PET)等技术获取的医学影像也成为研究复杂疾病的关键工具. 我们将结合放射学影像与生物信息学的学科称为放射组学(radiomics). 在癌症研究中, 放射组学可以帮助医生评估肿瘤的侵袭性、预测治疗响应或监测治疗后的复发[25]. 在心血管疾病中, 它有助于评估心脏组织的损伤程度和血流动力学变化[26].

1.2 复杂疾病相关的多组学数据库

随着越来越的多组学研究的出现, 有些研究者针对给定的生物医学问题, 收集大量多组学数据并汇总为数据库. 利用这些数据库, 研究者可以更快速地获取和分析多组学数据, 挖掘不同组学数据的相互作用和影响. 例如, 他们可以在线比较正常和病理条件下的数据, 来识别疾病中关键的生物标志物. 本文汇总了一些与复杂疾病相关的多组学数据库(表1). 这些数据库覆盖了癌症、心脑血管疾病、器官纤维化、慢性肾病、阿尔兹海默病、肠炎等复杂疾病, 主要涉及了基因组学、表观基因组学、转录组学、蛋白组学、代谢组学和微生物组学, 个别数据库还包含刻画药物反应的药物基因组学.

2 多组学数据的整合方法

多组学数据的主要优势在于利用不同组学的互补性帮助研究者更加深入地理解生物体信息. 正因如此, 越来越多的研究集中于探索多组学数据的整合策略, 旨在充分利用多组学数据的综合优势来解决复杂疾病中各种生物学问题. 我们根据计算方法的不同, 可以将多组学数据整合的方法分为两类: 基于关联和网络融合的方法和基于数据矩阵和机器学习的方法(图2). 基于关联和网络融合的方法旨在研究组学之间相互作用关系, 而基于数据矩阵和机器学习的方法旨在合并多组学数据, 以解决复杂疾病中的聚类或分类问题. 这两种方法是利用不同手段揭示数据间的潜在联系, 从而提供对复杂生物系统更深层次的认识.

2.1 基于关联和网络融合的多组学整合方法

基于关联和网络融合的多组学整合方法是指利用

表 1　现有多组学数据库的概括

Table 1　Summary of existing multi-omics databases

数据库名称	疾病类型	组学类型	网址	年份
DriverDBv4[27]	癌症	基因组学, 表观基因组学, 转录组学, 蛋白质组学	http://driverdb.bioinfomics.org	2024
cfomics[28]	癌症, 动脉粥样硬化等69种疾病	无细胞数据的基因组学, 表观基因组学, 转录组学, 蛋白质组学, 代谢组学	https://cfomics.ncRNAlab.org	2024
MyeloDB[29]	多发性骨髓瘤	基因组学, 表观基因组学, 转录组学	https://project.iith.ac.in/cgntlab/myelodb	2024
COMBATdb[30]	COVID-19	全血转录组学, 血浆蛋白质组学, 表观基因组学, c转录组学, 单细胞表观基因组学	https://db.combat.ox.ac.uk	2023
IsMOD[31]	癌症	基于图像的单细胞多组学(基因组、转录组和核蛋白质组)	https://www.i-smod.com	2023
FPIA[32]	癌症	基因组学, 蛋白质组学	http://bioinfo-sysu.com/fpia/	2022
FibROAD[33]	器官纤维化相关疾病	基因组学, 表观基因组学, 转录组学	https://www.fibroad.org	2022
iNetModels[34]	癌症	蛋白质组学、代谢组学, 基因组学, 微生物学	https://inetmodels.com	2021
AIzGPS[35]	阿兹海默病	基因组学, 转录组学, 蛋白质组学	https:/alzgps.lerner.ccf.org	2021
MVIP[36]	病毒感染	基因组学, 表观基因组学, 转录组学	https:/mvip.whu.edu.cn/	2021
Fibromine[37]	肺纤维化	转录组学, 蛋白质组学, 单细胞转录组学	http://www.fibromine.com/Fibromine	2021
EyeDiseases[38]	眼部疾病	基因组学, 表观基因组学, 转录组学	https:/eyediseases.bio-data.cn/	2021
IBDMDB[39]	肠炎	基因组学, 表观基因组学, 转录组学, 蛋白质组学, 代谢组学, 微生物组学	http://ibdmdb.org	2019
International Cancer Genomics Consortium (ICGC)[40]	癌症	基因组学, 表观基因组学和转录组学	https://dcc.icgc.org/repositories	2019
Cancer Cell Line Encyclopedia (CCLE)[41]	癌症	基因组学, 转录组学	https://depmap.org/portal/download/all/	2019
C/VDdb[42]	心脑血管疾病	基因组学, 转录组学, 蛋白质组学, 代谢组学	www.padb.org/cvd	2018
MOBCdb[43]	乳腺癌	基因组学, 表观基因组学, 转录组学, 药物基因组学	http://bigd.big.ac.cn/MOBCdb/	2018
LinkedOmics[44]	癌症	基因组学, 表观基因组学, 转录组学, 蛋白质组学	http://linkedomics.org	2017
cCKDdb[45]	慢性肾病	基因组学, 转录组学, 蛋白质组学, 代谢组学	www.padb.org/ckdbd	2017
OmicsDI[46]	糖尿病、癌症和精神疾病等	基因组学, 转录组学, 蛋白质组学和代谢组学	https://www.omicsdi.org/	2017
CardioGenBase[47]	心脑血管疾病	基因组学, 蛋白质组学	www.CardioGenBase.com	2015
UK Biobank[48]	癌症、心脏病、糖尿病和精神疾病等	基因组学, 转录组学, 蛋白质组学和放射组学	https://www.ukbiobank.ac.uk/	2014
The Cancer Genome Atlas (TCGA)[49]	癌症	基因组学, 转录组学, 蛋白质组学和放射组学	https://cancergenome.nih.gov/	2013
METABRIC[50]	乳腺癌	基因组学, 转录组学	http://molonc.bccrc.ca/aparicio-lab/research/metabric/	2012

关联分析、网络融合找到不同组学不同分子之间的相互作用关系, 旨在找到复杂疾病的生物标志物. 早年多组学整合最常见的是基于关联分析或映射的方法, 找到两个不同组学之间的关联关系或因果关系, 从不同角度描述生物事件的发展规律. 例如, 研究复杂疾病中基因突变对转录水平的影响[51], 研究阿尔茨海默病中转录后调控对蛋白质表达的影响[52]等. 随着研究的深入, 多组学整合研究开始关注于分析多组学数据间的网络关系, 以描述生物系统内的复杂相互作用和调控机制. 在基于网络融合的整合方法中, 研究者首先将各组学中不同分子之间相互作用转化为网络连接关系, 接着利用随机游走等网络扩散算法来扩展和融合多组学网络, 最后根据特定的表型识别代表重要调控关系的子网络. 例如, OmicsNet算法是基于公开数据库构建基因、转录因子和代谢物之间相互作用进行整合[53]. 类似地, HENA算法也是基于图卷积神经网络整合阿尔

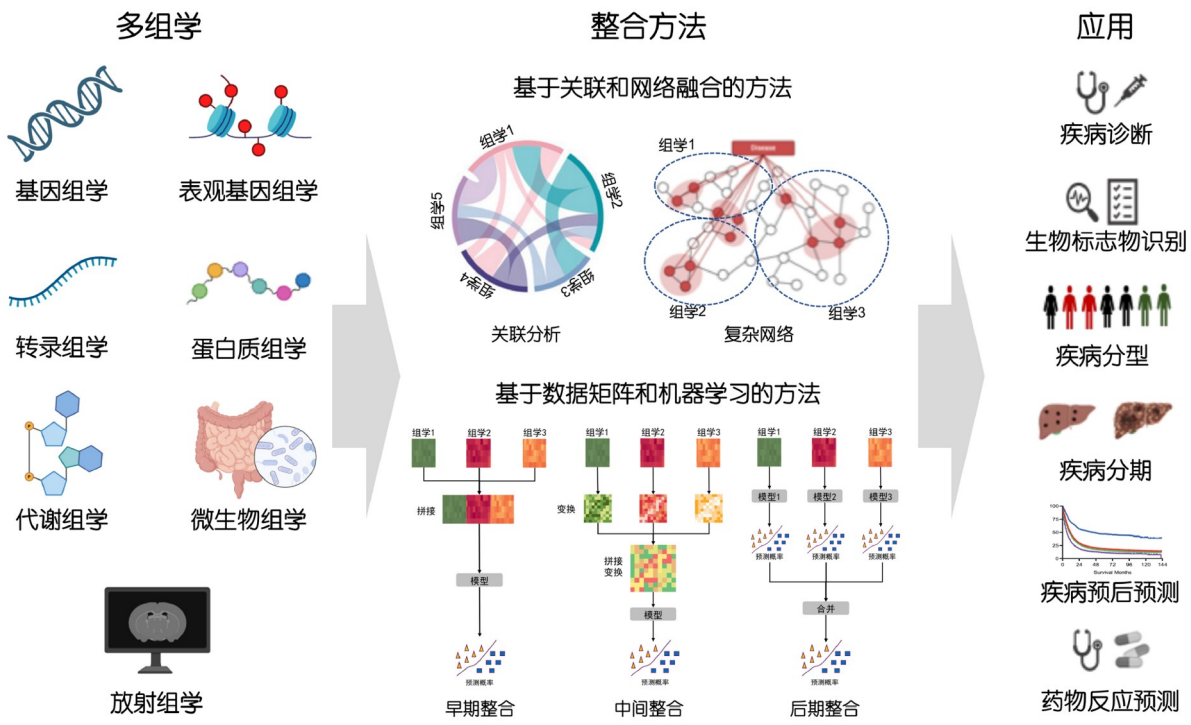

图 2 多组学整合工作流程. 图中部分元素来自Biorender.com
Figure 2 Workflow of multi-omics integration. Some of the elements in the image are from Biorender.com

兹海默病相关异构网络, 从而探索基因、蛋白质、SNP和基因探针之间的生物相互作用[54]; Mergeomics 2.0是将疾病关联的多组学数据进行汇总, 利用标记集富集分析和关键驱动分析识别疾病相关通路和网络的重要调节因子[55]; PaintOmics3算法是将多种组学数据集成到统一的通路图中[56]. 此外, 还有些研究者将时间维度也考虑到网络融合中, 例如Bodein等人提出的netOmics算法, 使用随机游走算法整合来自不同时间点的纵向多组学数据, 并回答疾病亚型、生物标志物发现等问题[57]. 这类方法的主要挑战在于面对不同的生物学问题如何对各种组学数据进行有效处理, 以提取多维度的信息, 从而在含有噪音的数据中识别出真正的生物调控关系, 而不是仅仅是数据在统计学上关联性.

2.2 基于数据矩阵和机器学习的多组学整合方法

随着高通量测序技术的不断进步产生了大量高维且稀疏的组学数据, 迫切需要更有效的算法来捕捉多组学数据之间的依赖关系. 基于数据矩阵和机器学习的多组学整合方法是指利用矩阵分解、机器学习(如深度学习)等模型进行数据融合, 旨在实现复杂疾病中的聚类或分类任务, 并在过程中揭示不同组学的内在关系和疾病相关的生物标志物. 从模型类型的角度来看, 我们也可以将生物多模态整合方法分为早期整合、中期整合和后期整合三类[58](图3).

2.2.1 早期整合方法(early integration method)

早期整合方法(early integration method)是指将多组学数据合并成一个联合矩阵, 然后利用有监督或无监督方法进行分析. 我们根据研究目的可以将多组学研究分为有监督问题和无监督问题, 有监督问题包含疾病诊断和疾病表型分类等, 无监督问题包括患者聚类、细胞注释和调控网络重建等. 在有监督问题中, 早期整合方法通常是将联合矩阵作为各种机器学习分类器的输入, 例如通过随机森林(random forests, RF)和支持向量机(support vector machines, SVM)模型来整合转录组和基因组数据从而预测抗癌药物反应[59]; 通过最小绝对收缩和选择算子模型(least absolute shrinkage and selection operator, LASSO)整合RNA表达、DNA甲基化和DNA拷贝数来预测卵巢癌预后[60]. 早期整合方法还可以解决无监督问题, 例如Fridley等人[61]通过贝叶斯路径分析处理RNA表达和SNP的联合矩阵来识别基因组对表型的直接和间接影响.

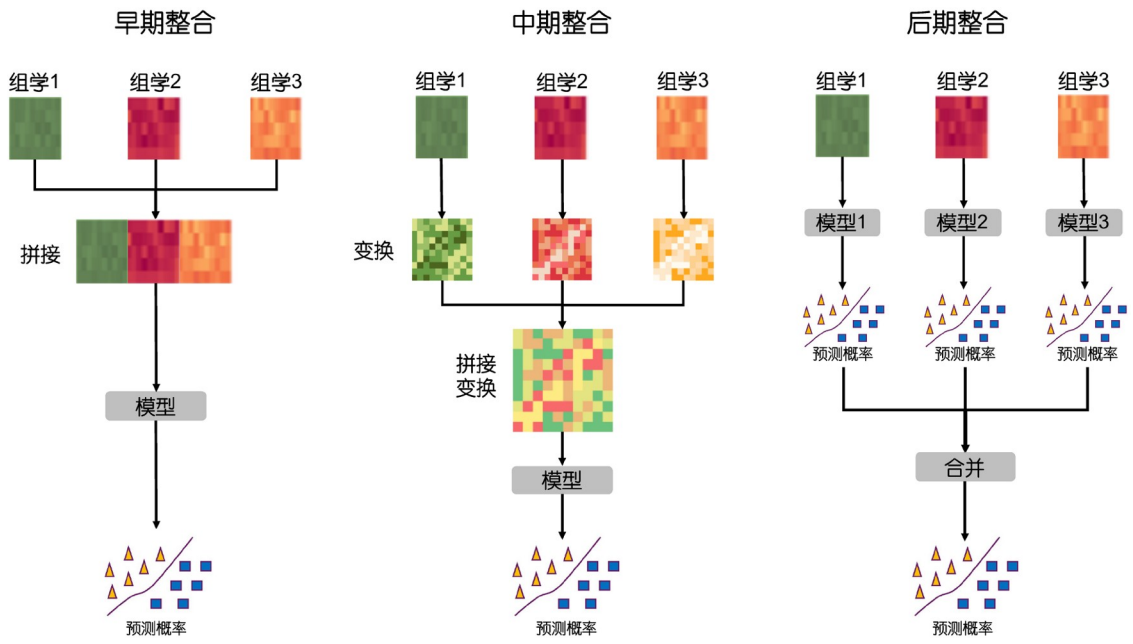

图 3 早、中和后期整合方法的工作流程
Figure 3 Workflows of early, intermediate, and late integration methods

早期整合方法的优点在于简单、易于实现且具有高度适应性，但是由于其将多组学数据合并为更复杂的高维数据，往往会导致数据的复杂性变高，不利于进行有效学习．灵活且强大的深度学习一定程度上可以解决这些问题，例如谢刚才等人提出癌症预测组正则化深度学习模型，将深度神经网络(deep neural networks, DNN)和Cox比例风险模型相结合来进行生存预测[62]．然而深度学习模型面临着黑盒问题，往往缺乏可解释性．因此有些研究者将生物通路等先验知识加入神经网络结构，使得先验生物知识直接影响信息提取过程，提高模型的可解释性．基于先验知识的深度学习方法大多利用生物实体之间的关系来定义稀疏神经网络，其中输入层代表不同分子(基因、突变和蛋白质等)，中间层代表生物通路的不同层级，每个节点都有明确的生物学意义并且仅在有关系的节点之间存在链接．例如，DeepOmix算法是将多组学串联矩阵作为输入，并利用基因-通路之间的关系构建稀疏神经网络进行癌症的生存预测[63]；Deng等人[64]利用通路引导的深度神经网络整合基因组数据和药物靶标数据来进行药物敏感性预测；Elmarakeby等人[65]在*Nature*上发表的P-NET(多层次稀疏神经网)，不仅被证明可以预测前列腺癌的转移状态，而且通过生物验证实验证明了模型所识别的生物标志物的有效性；Liu等人[66]提出的Pathformer (基于生物通路的Tranformer模型)是将多组学数据并联后作为每个基因的嵌入向量，然后利用稀疏神经网络将其转换为通路嵌入向量，并利用交叉注意力机制实现通路嵌入向量和通路串扰网络的信息融合，从而应用于癌症诊断和预后等各种任务．

2.2.2 中期整合方法(intermediate integration method)

中期整合方法(intermediate integration method)是基于集成思想，首先对不同组学分别进行单独建模，然后对转换后的矩阵或模型进行集成．根据模型类型，我们可以将中期整合方法分为基于回归的方法、基于贝叶斯的方法、基于多核学习的方法、基于相似网络的方法、基于联合降维的方法和基于深度学习的方法等，这些方法大多针对无监督问题，针对有监督问题的方法多集中于基于回归或深度学习的方法．

基于回归的方法的典型代表是mixOmics[67]，是同时适用于有监督和无监督问题的方法．mixOmics主要通过扩展潜在结构投影来识别不同组学的生物标志物，从而寻找多组学中高度相关的共信息，然后利用最小二乘法判别法进行有监督分类或利用主成分进行无监督聚类．它是有监督多组学整合问题中最经典的方法，但处理异质性较大的复杂疾病数据时精度有限．类似地，基于回归的方法还有sMB-PLS[68]、integrOmics[69]等．

基于贝叶斯的整合方法又分为参数整合和模型整

合. 参数整合是指在同一贝叶斯模型中对不同组学给出不同参数. 例如, Suter等人[70]提出的bnClustOmics是利用贝叶斯网络进行肝细胞癌患者的亚型聚类; Bao等人[71]提出的SBFA是一种结构贝叶斯因子分析框架, 可以整合基因组学、影像学和生物网络知识来研究阿尔兹海默病; 在单细胞聚类中, 基于贝叶斯概率模型的BREM-SC算法旨在找到多组学间的关联[72]. 模型整合是指将不同组学对应的贝叶斯模型进行集成, 例如Eric等人提出的贝叶斯共识聚类算法(Bayesian biclustering model, BBC), 整合了乳腺癌转录组、基因组和蛋白质组以进行患者聚类[73]; 在单细胞数据整合中, 基于变分贝叶斯方法的Clonealign算法可以实现肿瘤细胞中不匹配的scRNA-seq和scDNA-seq数据的对齐, 从而找到表达和癌症克隆之间的映射[74].

基于多核学习的方法的核心思想是利用不同核函数对不同组学进行特征映射, 然后根据映射后的特征或者核空间进行整合, 从而提高数据的表征能力. 例如, 特征选择多核学习(feature selection multi-core learning, FSMKL)是经典的基于多核学习的有监督方法, 它使用多个内核来捕获数据集之间的相似性以预测癌症死亡风险[70]. 更多基于多核学习的方法是针对无监督问题, 利用整合后的特征空间实现患者聚类, 例如rMKL-LPP[75]、hMKL[76]和KPCA[77].

基于相似网络的方法的核心思想是针对每个组学构建样本相似网络, 然后整合这些网络来揭示多组学之间的关系. 大部分基于相似网络的方法都聚焦于无监督问题, 例如Wang等人[78]利用相似网络融合算法(similarity network fusion, SNF)整合RNA表达和DNA甲基化从而实现癌症亚型分类; Nguyen等人[79]提出的PINSPlus通过扰动分析融合不用组学的子图实现患者聚类; Scott等人[80]针对哮喘患者构建亲和力网络关联聚类算法(merged affinity network association clustering, MANAclust算法); 还有为了解决组学不完全匹配问题构建的基于邻域的多组学聚类(neighborhood based multi-omics clustering, NEMO算法)[81].

基于联合降维的方法是应用最广泛的一类中期整合方法, 其核心思想是利用因子分解、非负矩阵分解和主成分分析等统计方法将高维的多组学数据分解为因子矩阵(代表样本状态)和权重矩阵(代表特征重要性), 进行无监督的样本聚类和生物标志物识别. 研究人员对9个代表性的基于联合降维的方法进行了全面评估, 包括iCluster[82]、JIVE[83]、intNMF[84]、MOFA[85]、MCIA[86]、MSFA[87]、RGCCA[88]、tICA[89]和scikit-fusion[90], 发现MCIA在大多数数据集中表现最佳, intNMF在患者聚类中表现较好[91]. 此外, 基于联合降维的方法也被广泛应用于单细胞多组学中, 例如Seurat3[92]、bindSC[93]和LIGER[94]等.

随着深度学习的快速发展, 越来越多的研究者发现在多组学整合中使用深度学习算法可以捕捉不同组学之间的非线性关系, 进行更有效的下游分析. 基于深度学习的中期整合方法的核心思想是利用不同的深度学习模型分别对不同的组学进行变换, 然后在潜层空间进行多组学整合, 最后将融合层用于有监督分类或无监督聚类. 我们可以根据深度学习模型的类型进行分类, 分为基于前馈网络、自编码器、卷积神经网络、图神经网络和注意力机制的方法. 基于前馈网络的方法的典型代表是MOLI算法, 利用前馈神经网络分别编码体细胞突变、拷贝数变异和基因表达, 然后将其串联以预测药物反应[95]. 基于自编码器(autoencoder, AE)的方法是目前最常见的基于深度学习的方法, 例如Chaudhary等人[96]利用AE、ANOVA和SVM实现肺癌亚型预测模型; Tianle等人[97]提出视图分解自动编码器(multi view autoencoder, MAE)来整合多组学和领域知识, 以挖掘生物分子与临床指标之间的关系; 还有用于无监督的癌症患者聚类的LSTM-VAE[98]和AE-k-means[99]等. 在单细胞注释中, 基于自编码器的方法也有广泛的应用, 例如totalVI[100]和scMVAE[101]等. 基于图神经网络的方法大多将生物学网络和深度学习方法相结合进行数据整合, 例如Limeng等人[102]提出的CancerOmicsNet是利用基于注意力传播机制的图神经网络来预测激酶抑制剂对肿瘤的治疗效果; Li等人[103]提出的MoGCN是将样本相似网络、图神经网络和自编码器相结合来进行癌症亚型分类; Li等人[103]提出的MODIG是基于图注意力网络整合多组学进行癌基因识别. 除了上述方法, 卷积神经网络、对抗生成网络和注意力机制等方法也被用于多组学整合; 如Fatima等人[104]提出的iSOM-GSN是将多组学数据转换为二维网络后利用卷积神经网络来预测疾病状态; Yang等人[105]提出的Subtype-GAN是结合对抗网络和共识聚类来识别癌症亚型; Moon等人[106]提出的双模态数据整合算法(MOMA)是利用注意力机制来提取两个组学之间的重要模块从而实现信息融合; Zuo等人[107]提出了深度跨组学循环注意力(DCCA)模型来实现单细胞转录组和表观基因组的联合分析.

2.2.3 后期整合方法(late integration method)

后期整合方法(late integration method)的核心思想是对不同组学分别进行建模, 然后利用投票法、加权平均等方法将模型输出的分类结果或预测概率进行合并. 对于有监督问题, 最常见的后期整合方法是针对每个组学构建一个分类器, 然后利用投票法进行合并. 针对无监督问题的后期整合方法则多采用加权平均的方法, 例如基于集群分配聚类的COCA[108]、基于多核学习聚类的KLIC算法[109]、基于加权最近邻算法整合单细胞多组学的Seurat V4[110]等. 此外, 还有些研究者基于深度学习进行后期整合, 例如基于视图相关性发现网络(VCDN)开发的MOGONET[111]和MOGAT[112]. 后期整合方法的优势在于其灵活性, 它允许研究者将针对单组学开发的方法进行重新组装. 但是后期整合方法无法捕捉多组学间相互作用, 利用组学间的互补信息.

3 多组学整合在复杂疾病精准诊疗中的应用

在临床应用中, 复杂疾病的患者往往存在极高的异质性, 亟需个性化的诊疗方案. 但是仅依靠医生积累的临床经验很难达到精准医疗的目标, 因此越来越多的研究者开始尝试将多组学数据和临床问题相结合, 建立复杂疾病的诊疗模型, 挖掘相应的生物标志物, 为实现精准医疗提供支持. 在上述内容中, 我们介绍了不同类型的多组学整合方法, 接下来我们按这些方法在复杂疾病中的应用场景进行了归纳(图4), 阐述了多组学整合方法在疾病筛查、分型、预后和药物反应预测中的应用.

疾病筛查旨在及时、准确地发现疾病病变, 以实现早期治疗和干预, 从而提高生存率. 相比传统的医学影像检查和组织活检, 近年来新兴的液体活检技术是一种更微创、更易于推广的疾病筛查手段. 但是基于循环肿瘤细胞(circulating tumor cells, CTC)、循环肿瘤DNA(circulating tumor DNA, ctDNA)和无细胞RNA(cell free RNA, cfRNA)等技术的单组学液体活检往往不能满足早期筛查的精度需求, 因此研究者开始探索体液数据的多组学整合策略. 例如, 约翰霍普金斯大学提出的CancerSEEK, 通过整合61种ctDNA和8种蛋白质实现8种癌症筛查[113]; Zhu等人[114]整合不同类型的RNA以实现肝癌无创诊断; Liu等人[66]提出的Pathformer整合了7种转录后调控事件, 在血浆和血小板中实现泛癌筛查.

疾病分型是指对患者进行聚类或分类, 以研究不同亚型之间的分子差异, 为个性化治疗提供依据. 早期, 基于贝叶斯、相似网络、联合降维和自编码器等方法的无监督的多组学整合方法被广泛应用于疾病分型. 近来, 随着疾病亚型金标准数据的不断完善, 也出现了如P-NET[65]、MOGONET[111]和MOGAT[112]等有监督的多组学整合方法.

疾病预后预测是指根据分子特征和临床特征评估患者的生存期, 辅助医生制定合适治疗方案. 由于患者预后受到治疗手段、遗传因素和分子特征等多个因素影响, 因此非常适合利用多组学数据进行建模. 针对疾病预后预测的多组学整合方法主要以机器学习或深度学习结合Cox比例风险模型为主(例如CoxPath[60]、GDP[62]和DL-Cox[115]等), 还有些研究者将高风险患者和低风险患者分为两类进行建模(例如P-NET[65]和Pathformer[66]等).

药物反应预测的意义在于通过分析患者的分子特征, 可以更好地了解药物对不同患者的疗效和副作用, 从而为医生提供更精准的用药建议. 多组学整合方法通常通过有监督方法实现药物反应预测, 将患者分为应答者(包括完全应答和部分应答)和非应答者(包括疾病稳定和疾病进展), 如P-NET[65]、Pathformer[66]和MOLI[95]. 此外, 有些研究者还通过整合单细胞多组学数据, 来挖掘新的治疗靶点[107].

4 挑战和未来方向

在本篇综述中, 我们探讨了复杂疾病相关的多组学类型和获取途径, 对多组学整合方法进行了详细地分类, 并探讨了其在复杂疾病不同场景中的应用潜力. 在此过程中, 我们发现多组学整合在复杂疾病中的应用也面临着很多挑战和机遇, 可以将其归纳为以下三类:

(1) 样本层面. 在相同的病人和样本中相互匹配的多组学数据缺失是目前主要面临的挑战. 尽管已有众多多组学数据库(见表1), 但它们通常针对基因组学或转录组学, 且不同组学间的样本交集有限. 例如, TCGA数据库的1218个乳腺癌样本中仅752个样本同时具有mRNA表达、DNA甲基化和拷贝数变异数据. 这导致大部分现有的多组学整合方法没有足够的训练数据, 无法发挥最佳效果. 建立统一的数据质控和处理标准流程, 鼓励跨机构数据共享协作, 是扩大样本规模和弥补数据缺失最直接的方法. 但是数据共享带来的异

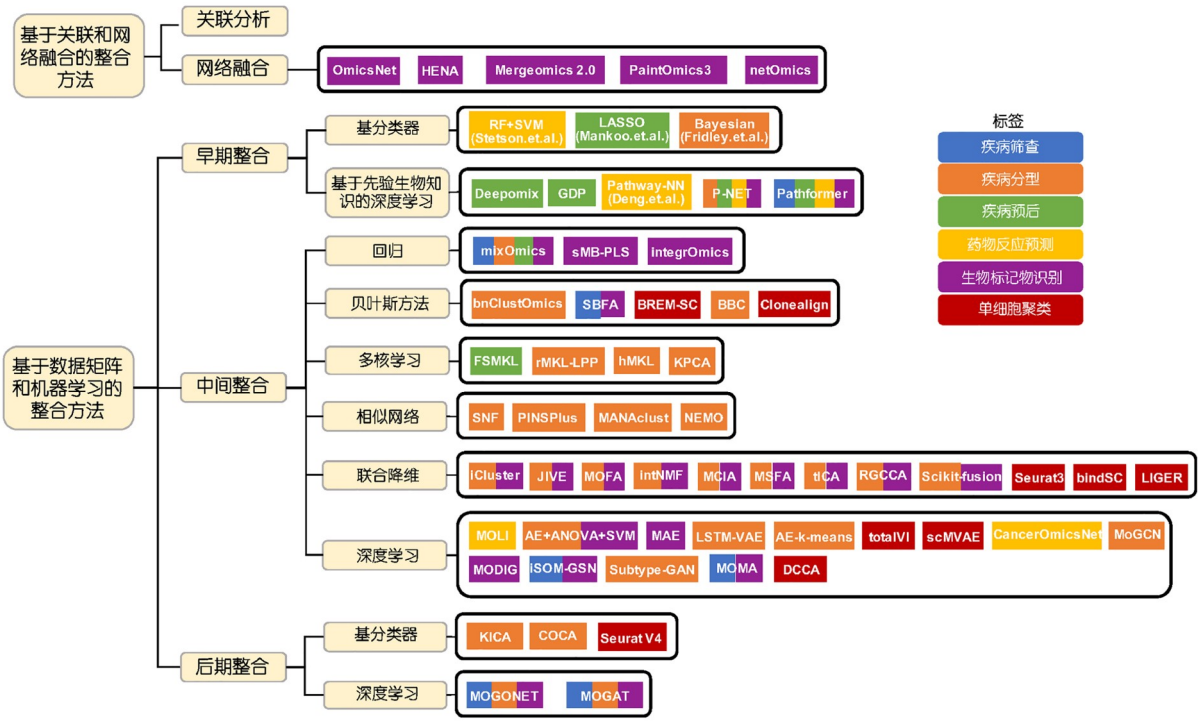

图 4 多组学整合方法概述. 这些方法根据其模型类型进行分组, 并根据其在复杂疾病中的应用场景进行颜色编码

Figure 4 Overview of multi-omics integration methods. These methods are grouped according to their model types and color-coded based on their application scenarios in complex diseases

质性和隐私泄露问题也不可忽略, 因此开发针对不匹配数据的整合算法也是未来的主要方向之一. 目前有些研究者提出了基于网络融合的整合方法(例如NEMO[81]), 将不匹配数据整合转化为不同样本网络融合, 解决数据对齐问题. 还有些研究者尝试利用迁移学习、零样本学习等新兴技术, 通过迁移其他相似领域的知识来规避数据缺失的问题, 已成功应用于增强癌症诊断以及治疗反应预测[116,117].

(2) 数据层面. 疾病多组学数据的高维、高噪音、高异质性和批次化效应的特性是目前主要面临的挑战. 不同测序技术产生的数据差异和复杂疾病的复杂性导致了多组学数据的高噪音和高异质性, 大量的致病突变和基因异常表达也给多组学数据带来了高维低样本的问题. 批次化效应是指相同和类似的样本在不同场所(例如不同医院)和不同的时间点获得的数据会有很大的不同, 同一批次(例如同一医院)样本的数据往往更相似, 这造成了在构建预测模型时的困难, 非常容易造成过拟合、泛化能力差的结果. 一个突出体现就是很多预测方法在文章发表时准确度很高(甚至过高), 但是在真实世界表现效果不佳. 采用更高效的、更前沿的数学模型和深度学习模型是解决以上这些挑战的主要方向. 例如, Amodio等人[118]提出的SAUCIE模型, 利用自编器对单细胞数据进行插补、去噪和去除批次效应, 以实现低维可视化和无监督聚类; Wang等人[119]提出的对抗配对转移网络是利用生成对抗网络和自编码器消除不同scRNA-seq数据集之间的批次效应; Yu等人[120]开发了基于深度度量学习的模型, 以初始聚类中最近邻信息为指导去除样本的批次效应, 整合不同来源的scRNA-seq数据. 研究者也成功地将自编码器、卷积神经网络和Transformer等结构引入多组学整合(如MAE[97]、MoGCN[103]和Pathformer [66]), 利用特征提取和潜层空间整合来弥补数据质量问题. 另外, 利用大语言模型进行预训练(如scGPT[121]), 解决单细胞多组学整合、细胞注释、疾病标志物识别等问题也成为多组学模型未来发展的趋势.

(3) 模型和计算层面. 计算模型缺乏可解释性、计算效率低和存在隐私泄露风险是目前主要面临的挑战. 深度学习模型虽然取得了较好的效果但可解释性不足, 通常无法直观地表示生物标志物和表型之间的关系, 这极大地阻碍了多组学整合从实验室到临床转化的发

展进程. 因此, 可解释的深度学习模型是未来重要的发展方向之一. 有些研究者通过可视化工具来解释深度学习的决策过程, 例如通过模型的权重分布来理解哪些生物标志物在分类或预测中起重要作用; 还有些研究者将生物学先验知识融入神经网络(DeepOmix[63]和P-NET[65]), 通过引入基于路径或网络的先验信息来指导学习过程, 使模型学习到的特征与生物学过程紧密相关. 在提高计算效率方面, 研究者也尝试通过迁移学习、模型蒸馏技术、GPU加速和云计算等新技术, 在不显著降低准确度的前提下减少模型的复杂度、时间和资源消耗. 另外, 计算过程中的患者隐私保护也是一个日益重要的话题. 差分隐私和联邦学习等在共享原始数据下进行模型训练的深度学习技术也受到越来越多研究者的关注. 此外, 未来的发展还可能包括对模型泛化能力的改善, 以确保模型在不同的人群、不同的环境条件下都能保持高效和准确, 从而推动多组学整合在临床应用中的发展.

总之, 多组学整合在复杂疾病的诊疗中展现出巨大的潜力, 但同时也面临着诸多挑战. 样本层面的数据缺失问题, 数据层面的高维、高噪音、高异质性和批次化效应, 以及模型层面的可解释性不足, 这些都是当前多组学整合亟待解决的问题. 然而, 随着技术的进步和研究方法的创新, 我们有理由相信, 这些问题将逐步得到解决. 而且随着跨学科合作的加强, 多组学整合将在复杂疾病的诊疗中发挥越来越重要的作用, 为人类的健康和福祉做出更大的贡献.

参考文献

1 Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. Nat Comput Sci, 2021, 1: 395–402
2 Cano-Gamez E, Trynka G. From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. Front Genet, 2020, 11: 505357
3 Lim J, Chin V, Fairfax K, et al. Transitioning single-cell genomics into the clinic. Nat Rev Genet, 2023, 24: 573–584
4 van der Harst P, de Windt L J, Chambers J C. Translational perspective on epigenetics in cardiovascular disease. J Am Coll Cardiol, 2017, 70: 590–606
5 Dawson M A, Kouzarides T. Cancer epigenetics: From mechanism to therapy. Cell, 2012, 150: 12–27
6 Zheng Y, Luo Y, Chen X, et al. The role of mRNA in the development, diagnosis, treatment and prognosis of neural tumors. Mol Cancer, 2021, 20: 49
7 Miao L, Zhang Y, Huang L. mRNA vaccine for cancer immunotherapy. Mol Cancer, 2021, 20: 41
8 Gupta R A, Shah N, Wang K C, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, 2010, 464: 1071–1076
9 Matsuo M, Masumura T, Nishio H, et al. Exon skipping during splicing of dystrophin mRNA precursor due to an intraexon deletion in the dystrophin gene of Duchenne muscular dystrophy kobe. J Clin Invest, 1991, 87: 2127–2131
10 Xia Z, Donehower L A, Cooper T A, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun, 2014, 5: 5274
11 De Zuani M, Xue H, Park J S, et al. Single-cell and spatial transcriptomics analysis of non-small cell lung cancer. Nat Commun, 2024, 15: 4388
12 Yarden Y. Biology of HER2 and its importance in breast cancer. Oncology, 2001, 61: 1–13
13 Morishima-Kawashima M, Ihara Y. Alzheimer's disease: β-amyloid protein and tau. J Neurosci Res, 2002, 70: 392–401
14 Batista T M, Jayavelu A K, Wewer Albrechtsen N J, et al. A cell-autonomous signature of dysregulated protein phosphorylation underlies muscle insulin resistance in type 2 diabetes. Cell Metab, 2020, 32: 844–859.e5
15 Xu J Y, Zhang C, Wang X, et al. Integrative proteomic characterization of human lung adenocarcinoma. Cell, 2020, 182: 245–261.e17
16 Yang Q, Vijayakumar A, Kahn B B. Metabolites as regulators of insulin sensitivity and metabolism. Nat Rev Mol Cell Biol, 2018, 19: 654–672
17 Goldberg I J, Trent C M, Schulze P C. Lipid metabolism and toxicity in the heart. Cell Metab, 2012, 15: 805–812
18 Schmidt D R, Patel R, Kirsch D G, et al. Metabolomics in cancer research and emerging applications in clinical oncology. CA Cancer J Clin, 2021, 71: 333–358
19 Graham S F, Chevallier O P, Roberts D, et al. Investigation of the human brain metabolome to identify potential markers for early diagnosis and therapeutic targets of Alzheimer's disease. Anal Chem, 2013, 85: 1803–1811
20 Ahn J, Sinha R, Pei Z, et al. Human gut microbiome and risk for colorectal cancer. JNCI-J Natl Cancer Institute, 2013, 105: 1907–1911
21 Dunne J L, Triplett E W, Gevers D, et al. The intestinal microbiome in type 1 diabetes. Clin Exp Immunol, 2014, 177: 30–37
22 Ghaisas S, Maher J, Kanthasamy A. Gut microbiome in health and disease: Linking the microbiome–gut–brain axis and environmental factors in

the pathogenesis of systemic and neurodegenerative diseases. Pharmacol Ther, 2016, 158: 52–62
23 Chattopadhyay I, Verma M, Panda M. Role of oral microbiome signatures in diagnosis and prognosis of oral cancer. Technol Cancer Res Treat, 2019, 18: 153303381986735
24 Tonelli A, Lumngwena E N, Ntusi N A B. The oral microbiome in the pathophysiology of cardiovascular disease. Nat Rev Cardiol, 2023, 20: 386–403
25 Farwell M D, Pryma D A, Mankoff D A. PET/CT imaging in cancer: Current applications and future directions. Cancer, 2014, 120: 3433–3445
26 Friedrich M G. Tissue characterization of acute myocardial infarction and myocarditis by cardiac magnetic resonance. JACC-Cardiovasc Imag, 2008, 1: 652–662
27 Liu C H, Lai Y L, Shen P C, et al. DriverDBv4: A multi-omics integration database for cancer driver gene research. Nucleic Acids Res, 2024, 52: D1246–D1252
28 Li M, Zhou T, Han M, et al. cfOmics: A cell-free multi-Omics database for diseases. Nucleic Acids Res, 2024, 52: D607–D621
29 Kumar A, Kumar K V, Kundal K, et al. MyeloDB: A multi-omics resource for multiple myeloma. Funct Integr Genomics, 2024, 24: 17
30 Wang D, Kumar V, Burnham K L, et al. COMBATdb: A database for the COVID-19 multi-omics blood ATlas. Nucleic Acids Res, 2023, 51: D896–D905
31 Zhang W, Suo J, Yan Y, et al. iSMOD: An integrative browser for image-based single-cell multi-omics data. Nucleic Acids Res, 2023, 51: 8348–8366
32 Huang L, Zhu H, Luo Z, et al. FPIA: A database for gene fusion profiling and interactive analyses. Intl J Cancer, 2022, 150: 1504–1511
33 Sun Y Z, Hu Y F, Zhang Y, et al. FibROAD: A manually curated resource for multi-omics level evidence integration of fibrosis research. Database, 2022, 2022: baac015
34 Arif M, Zhang C, Li X, et al. iNetModels 2.0: An interactive visualization and database of multi-omics data. Nucleic Acids Res, 2021, 49: W271–W276
35 Zhou Y, Fang J, Bekris L M, et al. AlzGPS: A genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery. AlzheimersRes Ther, 2021, 13: 24
36 Tang Z, Fan W, Li Q, et al. MVIP: Multi-omics portal of viral infection. Nucleic Acids Res, 2022, 50: D817–D827
37 Fanidis D, Moulos P, Aidinis V. Fibromine is a multi-omics database and mining tool for target discovery in pulmonary fibrosis. Sci Rep, 2021, 11: 21712
38 Yuan J, Chen F, Fan D, et al. EyeDiseases: An integrated resource for dedicating to genetic variants, gene expression and epigenetic factors of human eye diseases. NAR Genomics BioInf, 2021, 3: lqab050
39 Lloyd-Price J, Arze C, Ananthakrishnan A N, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature, 2019, 569: 655–662
40 Zhang J, Bajari R, Andric D, et al. The international cancer genome consortium data portal. Nat Biotechnol, 2019, 37: 367–369
41 Ghandi M, Huang F W, Jané-Valbuena J, et al. Next-generation characterization of the cancer cell line encyclopedia. Nature, 2019, 569: 503–508
42 Fernandes M, Patel A, Husi H, et al. C/VDdb: A multi-omics expression profiling database for a knowledge-driven approach in cardiovascular disease (CVD). PLoS ONE, 2018, 13: e0207371
43 Xie B, Yuan Z, Yang Y, et al. MOBCdb: A comprehensive database integrating multi-omics data on breast cancer for precision medicine. Breast Cancer Res Treat, 2018, 169: 625–632
44 Vasaikar S V, Straub P, Wang J, et al. LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res, 2018, 46: D956–D963
45 Fernandes M, Husi H. Establishment of a integrative multi-omics expression database CKDdb in the context of chronic kidney disease (CKD). Sci Rep, 2017, 7: 40367
46 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: 406–409
47 V A, Nayar P G, Murugesan R, et al. CardioGenBase: A literature based multi-omics database for major cardiovascular diseases. PLoS ONE, 2015, 10: e0143188
48 Biobank U. About us. UK Biobank, 2014. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us
49 Weinstein J N, Collisson E A, Mills G B, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet, 2013, 45: 1113–1120
50 Curtis C, Shah S P, Chin S F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 2012, 486: 346–352
51 Huang Y T, VanderWeele T J, Lin X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. Ann Appl Stat, 2014, 8
52 Mangleburg C G, Wu T, Yalamanchili H K, et al. Integrated analysis of the aging brain transcriptome and proteome in tauopathy. Mol

Neurodegeneration, 2020, 15: 1–7
53 Zhou G, Xia J. OmicsNet: A web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Res, 2018, 46: W514–W522
54 Sügis E, Dauvillier J, Leontjeva A, et al. HENA, heterogeneous network-based data set for Alzheimer's disease. Sci Data, 2019, 6: 151
55 Ding J, Blencowe M, Nghiem T, et al. Mergeomics 2.0: A web server for multi-omics data integration to elucidate disease networks and predict therapeutics. Nucleic Acids Res, 2021, 49: W375–W387
56 Hernández-de-Diego R, Tarazona S, Martínez-Mira C, et al. PaintOmics 3: A web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res, 2018, 46: W503–W509
57 Bodein A, Scott-Boyer M P, Perin O, et al. Interpretation of network-based integration from multi-omics longitudinal data. Nucleic Acids Res, 2022, 50: e27
58 Picard M, Scott-Boyer M P, Bodein A, et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J, 2021, 19: 3735–3746
59 Stetson L C, Pearl T, Chen Y, et al. Computational identification of multi-omic correlates of anticancer therapeutic response. BMC Genomics, 2014, 15: 1–8
60 Mankoo P K, Shen R, Schultz N, et al. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS ONE, 2011, 6: e24709
61 Fridley B L, Lund S, Jenkins G D, et al. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet Epidemiol, 2012, 36: 352–359
62 Xie G, Dong C, Kong Y, et al. Group Lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. Genes, 2019, 10: 240
63 Zhao L, Dong Q, Luo C, et al. DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. Comput Struct Biotechnol J, 2021, 19: 2719–2725
64 Deng L, Cai Y, Zhang W, et al. Pathway-guided deep neural network toward interpretable and predictive modeling of drug sensitivity. J Chem Inf Model, 2020, 60: 4497–4505
65 Elmarakeby H A, Hwang J, Arafeh R, et al. Biologically informed deep neural network for prostate cancer discovery. Nature, 2021, 598: 348–352
66 Liu X, Tao Y, Cai Z, et al. Pathformer: A biological pathway informed transformer integrating multi-omics data for disease diagnosis and prognosis. bioRxiv, 2023: 2023.05. 23.541554
67 Rohart F, Gautier B, Singh A, et al. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol, 2017, 13: e1005752
68 Li W, Zhang S, Liu C C, et al. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics, 2012, 28: 2458–2466
69 Lê Cao K A, González I, Déjean S. integrOmics: An R package to unravel relationships between two omics datasets. Bioinformatics, 2009, 25: 2855–2856
70 Suter P, Dazert E, Kuipers J, et al. Multi-omics subtyping of hepatocellular carcinoma patients using a Bayesian network mixture model. PLoS Comput Biol, 2022, 18: e1009767
71 Bao J, Chang C, Zhang Q, et al. Integrative analysis of multi-omics and imaging data with incorporation of biological information via structural Bayesian factor analysis. Brief BioInf, 2023, 24: bbad073
72 Wang X, Sun Z, Zhang Y, et al. BREM-SC: A bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res, 2020, 48: 5814–5824
73 Lock E F, Dunson D B. Bayesian consensus clustering. Bioinformatics, 2013, 29: 2610–2616
74 Campbell K R, Steif A, Laks E, et al. clonealign: Statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol, 2019, 20: 1–2
75 Speicher N K, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics, 2015, 31: i268–i275
76 Wei Y, Li L, Zhao X, et al. Cancer subtyping with heterogeneous multi-omics data via hierarchical multi-kernel learning. Brief BioInf, 2023, 24: bbac488
77 Mariette J, Villa-Vialaneix N, Wren J. Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics, 2018, 34: 1009–1015
78 Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods, 2014, 11: 333–337
79 Nguyen H, Shrestha S, Draghici S, et al. PINSPlus: A tool for tumor subtype discovery in integrated genomic data. Bioinformatics, 2019, 35: 2843–2846

80 Tyler S R, Chun Y, Ribeiro V M, et al. Merged Affinity Network Association Clustering: Joint multi-omic/clinical clustering to identify disease endotypes. Cell Rep, 2021, 35: 108975
81 Rappoport N, Shamir R, Schwartz R. NEMO: Cancer subtyping by integration of partial multi-omic data. Bioinformatics, 2019, 35: 3348–3356
82 Shen R, Olshen A B, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics, 2009, 25: 2906–2912
83 Lock E F, Hoadley K A, Marron J S, et al. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat, 2013, 7: 523–542
84 Chalise P, Fridley B L, Peddada S D. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS ONE, 2017, 12: e0176278
85 Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis—A framework for unsupervised integration of multi-omics data sets. Mol Syst Biol, 2018, 14: e8124
86 Bady P, Dolédec S, Dumont B, et al. Multiple co-inertia analysis: A tool for assessing synchrony in the temporal variability of aquatic communities. Comptes Rendus Biologies, 2004, 327: 29–36
87 De Vito R, Bellio R, Trippa L, et al. Multi-study factor analysis. Biometrics, 2019, 75: 337–346
88 Tenenhaus M, Tenenhaus A, Groenen P J F. Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. Psychometrika, 2017, 82: 737–777
89 Teschendorff A E, Jing H, Paul D S, et al. Tensorial blind source separation for improved analysis of multi-omic data. Genome Biol, 2018, 19: 76
90 Zitnik M, Zupan B. Data fusion by matrix factorization. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 41–53
91 Cantini L, Zakeri P, Hernandez C, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun, 2021, 12: 124
92 Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell, 2019, 177: 1888–1902.e21
93 Dou J, Liang S, Mohanty V, et al. Unbiased integration of single cell multi-omics data. bioRxiv, 2020. Doi: 10.1101/2020.12.11.422014
94 Welch J, Kozareva V, Ferreira A, et al. Integrative inference of brain cell similarities and differences from single-cell genomics. bioRxiv, 2018: 459891
95 Sharifi-Noghabi H, Zolotareva O, Collins C C, et al. MOLI: Multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics, 2019, 35: i501–i509
96 Chaudhary K, Poirion O B, Lu L, et al. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res, 2018, 24: 1248–1259
97 Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). BMC Genomics, 2019, 20: 944
98 Chung N C, Mirza B, Choi H, et al. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. Methods, 2019, 166: 66–73
99 Miao Z, Humphreys B D, McMahon A P, et al. Multi-omics integration in the age of million single-cell data. Nat Rev Nephrol, 2021, 17: 710–724
100 Gayoso A, Steier Z, Lopez R, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods, 2021, 18: 272–282
101 Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. Brief BioInf, 2021, 22: bbaa287
102 Pu L, Singha M, Ramanujam J, et al. CancerOmicsNet: A multi-omics network-based approach to anti-cancer drug profiling. Oncotarget, 2022, 13: 695–706
103 Li X, Ma J, Leng L, et al. MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. Front Genet, 2022, 13: 806842
104 Fatima N, Rueda L, Jonathan W. iSOM-GSN: An integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps. Bioinformatics, 2020, 36: 4248–4254
105 Yang H, Chen R, Li D, et al. Subtype-GAN: A deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics, 2021, 37: 2231–2237
106 Moon S, Lee H, Lu Z. MOMA: A multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics, 2022, 38: 2287–2296
107 Zuo C, Dai H, Chen L, et al. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. Bioinformatics, 2021, 37: 4091–4099
108 Hoadley K A, Yau C, Wolf D M, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 2014, 158: 929–944
109 Cabassi A, Kirk P D W, Xu J. Multiple kernel learning for integrative consensus clustering of omic datasets. Bioinformatics, 2020, 36: 4789–4796

110 Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell, 2021, 184: 3573–3587.e29
111 Wang T, Shao W, Huang Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun, 2021, 12: 3445
112 Tanvir R B, Islam M M, Sobhan M, et al. MOGAT: A multi-omics integration framework using graph attention networks for cancer subtype prediction. Int J Mol Sci, 2024, 25: 2788
113 Cohen J D, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, 2018, 359: 926–930
114 Zhu Y, Wang S, Xi X, et al. Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer. Theranostics, 2021, 11: 181–193
115 Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics, 2019, 35: i446–i454
116 Chougrad H, Zouaki H, Alheyane O. Multi-label transfer learning for the early diagnosis of breast cancer. Neurocomputing, 2020, 392: 168–180
117 Zhu Y, Brettin T, Evrard Y A, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. Sci Rep, 2020, 10: 18040
118 Amodio M, van Dijk D, Srinivasan K, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods, 2019, 16: 1139–1145
119 Wang D, Hou S, Zhang L, et al. iMAP: Integration of multiple single-cell datasets by adversarial paired transfer networks. Genome Biol, 2021, 22: 63
120 Yu X, Xu X, Zhang J, et al. Batch alignment of single-cell transcriptomics data using deep metric learning. Nat Commun, 2023, 14: 960
121 Cui H, Wang C, Maan H, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods, 2024, 21: 1470–1480

Summary for “复杂疾病中多组学多模态数据的生物信息学研究进展”

Progress of bioinformatics studies for multi-omics and multi-modal data in complex diseases

Xiaofan Liu[1,2] & Zhi John Lu[1,2*]

[1] *MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China;*
[2] *Institute for Precision Medicine, Tsinghua University, Beijing 100084, China*
* Corresponding author, E-mail: zhilu@tsinghua.edu.cn

With the advancement of high-throughput sequencing technologies, the integration of multi-omics and multi-modal data has become an important trend in the study of complex diseases. Multi-omics/multi-modal data provide new perspectives for a deeper understanding of the pathogenesis and development of diseases, offering crucial support for the precision diagnosis and treatment of complex diseases. This review first introduces various types of omics data and their contributions to complex disease research. Genomics reveals the genetic background and mutations associated with diseases by analyzing gene sequences; transcriptomics uncovers gene regulatory relationships related to diseases by studying expression patterns; proteomics focuses on the expression, modification, and interactions of proteins; metabolomics reflects adjustments in metabolic pathways before and after illness through changes in metabolites; radiomics shows disease-induced alterations via medical imaging. Integrating and analyzing these omics data can compensate for the information gaps of single omics data, enabling a more comprehensive understanding of the molecular mechanisms of complex diseases. Furthermore, this review introduces multi-omics databases related to complex diseases, covering diseases such as cancer, cardiovascular and cerebrovascular diseases, organ fibrosis, chronic kidney disease, Alzheimer’s disease, and inflammatory bowel disease. These databases facilitate researchers in obtaining and analyzing multi-omics data.

Next, this review systematically categorizes the existing multi-omics integration methods into two types: correlation and network-based methods and data matrix and machine learning-based methods. These two approaches use different means to reveal the potential connections between data, thereby providing deeper insights into complex biological systems. Correlation and network-based methods involve using association analysis or complex network analysis to identify the intrinsic connections between different omics, thereby discovering biomarkers related to phenotypes. Data matrix and machine learning-based methods refer to utilizing statistical analysis, machine learning, and deep learning models to achieve data fusion for clustering or classification tasks, while revealing the inherent relationships between multi-omics data and identifying disease-related biomarkers. Data matrix and machine learning-based methods are further divided into early integration, intermediate integration, and late integration. Early integration method involves merging multi-omics data into a joint matrix and then applying machine learning or deep learning models for classification. Intermediate integration method involves modeling each omics data separately, followed by the integration of the transformed matrices or models. Late integration method independently models each omics data and then combines the model output results. Building on this, the review also discusses the applications of multi-omics integration models in complex diseases, such as disease screening, subtyping, prognosis, and drug response prediction.

Finally, this review summarizes the current challenges in multi-omics/multi-modal data integration, which are divided into three levels: sample, data, and model. At the sample level, the absence of matching data limits the efficacy of integration methods, and researchers are addressing this issue through the development of data sharing and new algorithms. At the data level, the characteristics of high dimensionality, noise, and heterogeneity necessitate the use of more efficient deep learning methods for data integration. At the model level, the key challenges include lack of interpretability, low computational efficiency, and privacy concerns. Researchers are enhancing model interpretability through visualization tools and incorporating biological prior knowledge into deep learning models, while also exploring new technologies such as model acceleration and privacy-preserving computation to improve model efficiency and security. Despite the challenges, multi-omics integration has demonstrated significant potential in the diagnosis and treatment of complex diseases.

multi-omics, multi-modal data, bioinformatics, complex diseases, data integration

doi: 10.1360/TB-2024-0416