多组学大数据与医学发展

刘斯洋[1],林星辰[1],程丝[2],王超龙[3],李昊[2*]

1. 中山大学公共卫生学院(深圳),深圳 518107
2. 首都医科大学附属北京天坛医院,国家神经系统疾病临床医学研究中心,北京 100070
3. 华中科技大学公共卫生学院,武汉 430030

摘要　多组学技术、队列研究设计、数据科学和机器学习的进步已经开始改变循证医学,为下一代"深度"医学的未来提供了诱人的前景。总结了基因组与基因组修饰测序、转录组与单细胞转录组、蛋白组、代谢组、微生物组、影像组与生物传感器等多组学实验技术和全基因组关联分析、全基因组关联信号解读、多基因风险评分、孟德尔随机化与人工智能算法等大数据分析技术的发展趋势,探讨了这些技术在疾病分型、诊断与预测、药物研发和临床试验设计等方面的临床应用。针对多组学大数据与医学发展面临的挑战,展望了未来队列设计、数据管理与共享、国际合作等发展方向。

关键词　多组学;大数据;医学研究;临床应用

过去几十年,医学科学研究取得了惊人且前所未有的进步。从对基础疾病过程的病理生理学更好的理解,到揭示细胞机械的原子分辨率,再到开发改变医学领域疾病进程和结果的治疗方法,都取得了重大突破。这些进展得益于基于基因组学、转录组学、蛋白质组学和代谢组学等多组学高通量测序及质谱检测技术的发展,为医学研究提供了大量高质量的生物数据。同时,大数据科学、生物信息学和人工智能等领域的发展也为医学研究带来了新的机遇。此外,基因编辑技术,如CRISPR(clustered regularly interspaced short palindromic repeats)-Cas9(CRISPR-associated protein 9)的出现,也为个性化医疗开辟了新的方向。

尽管基础科学和技术取得了重要的进步,但主要医学领域从实验室到临床应用的快速转化却相对滞后。药物开发和临床试验成本高昂且失败率高,这与长期困扰医疗系统的低效和缺陷等问题共同导致了临床研究的危机。应对这一困境的关键在于研究设计及研究方法的改进。

收稿日期:2024-04-28;修回日期:2024-06-11
基金项目:国家重点研发计划项目(2022YFC2502400)
作者简介:刘斯洋,副教授,研究方向为医学遗传学方法研发与应用,电子信箱:liusy99@mail.sysu.edu.cn;李昊(通信作者),教授,研究方向为临床流行病方法学,电子信箱:lihao@ncrcnd.org.cn
引用格式:刘斯洋, 林星辰, 程丝, 等. 多组学大数据与医学发展[J]. 科技导报, 2024, 42(12): 51-74;
doi:10.3981/j.issn.1000-7857.2024.05.00538

1 多组学技术

以中心法则的脱氧核糖核酸(deoxyribonucleic acid,DNA)复制、转录、蛋白质生物合成为基础,疾病是基因与环境相互作用的结果,其中会经历一系列生理生化过程的变化。基因组学、转录组学、蛋白质组学、代谢组学、微生物组学和影像组学等多组学实验技术的研发和改进,可以刻画疾病发生发展中的生物过程与分子相互作用(图1)。

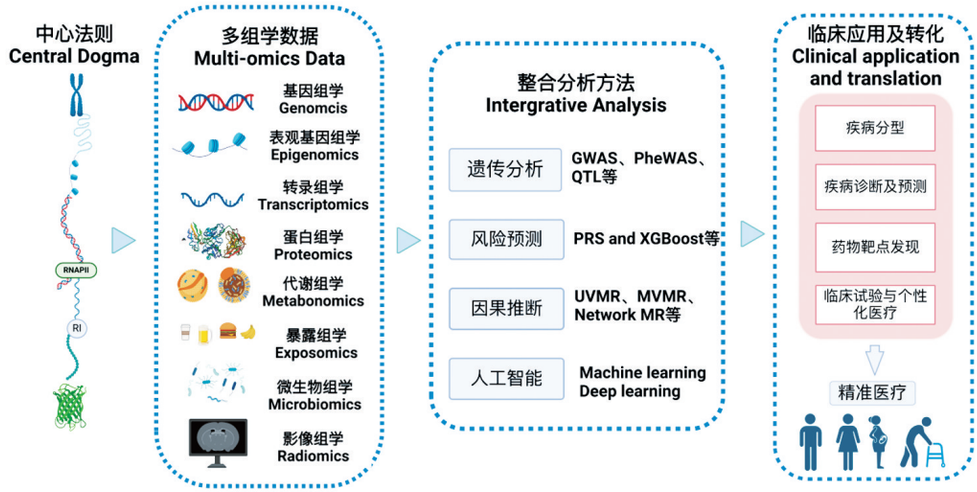

图1 主要组学实验技术发展趋势

1.1 基因组与基因组修饰

基因组学是研究基因组的结构、功能、演变及其相互作用的学科,旨在收集和量化一个生物体的所有基因,揭示它们的相互关系及其对生物体的影响。主要相关技术包括基因分型阵列[1-3]、全基因组测序[4-5]和外显子测序[6]。随着测序技术的进步,一个人类基因组测序的成本,已经从1999年近30亿美元降至100美元[7],极大地促进了对遗传变异如何导致疾病易感性和治疗反应差异的理解。其中,长读长测序技术(如单分子实时测序和纳米孔测序)更昂贵,也能够更好地捕捉到结构性重排,且同时测量表观遗传修饰[8]。此外,还有其他技术,如单色多重定量聚合酶链反应(polymerase chain reaction,PCR)的方法可用来测量染色体端粒长度[8]。通过将表型与基因组联系起来,可以识别疾病与表型的关键基因与通路,构建遗传风险评分模型[8]。根据计算的遗传风险因素对患者进行亚群分类,离个性化治疗方案更近了一步[9]。

尽管基因组序列本身基本是静态的,但它在各种组织中不断经历化学修饰(例如,在某些胞嘧啶核苷酸上附加甲基基团)[10],从而影响其表达。通过可逆的化学修饰和结构变化调控基因表达的所有表观遗传标记的集合又被称为表观基因组(epigenome)。胞嘧啶的甲基化修饰是最常见的一种化学修饰,在人的基因组中,大约有70%~90%的胞嘧啶处于甲基化状态[11]。常用的检测甲基化修饰的技术包括2种:一种使用重亚硫酸盐处理化学修饰DNA后进行测序的技术用于检测这些甲基化的胞嘧啶[12],另外则可以通过3代测序技术中的电信号直接检测甲基化修饰[13]。表观基因组对疾病扰动和外部环境暴露做出反应[14],可以捕捉与衰老[15]、癌症[10]、怀孕[16]等相关的组织特异性和时间特异性信息,以及饮食、压力、体育活动和吸烟等生活方式因素[17]。然而,解释这些数据需要谨慎考虑,因为表观遗传变化与健康结果之间的因果关系可能十分复杂,并且可能因个体和群体而异[18]。

1.2　转录组

转录组学研究特定细胞或组织中所有核糖核酸(ribonucleic acid,RNA)分子(即转录组)的表达,提供分子动态变化的整体视角。RNA转录组的检测包括编码蛋白的RNA(messenger RNA,mRNA)、长非编码RNA、短非编码RNA(如microRNA、小干扰RNA、小核RNA、piwi互作RNA和增强子RNA)及环状RNA。RNA测序(RNA sequencing,RNA-seq)是常用的转录组学技术,可从少量RNA样本中定量和定性RNA转录[19-21]。无论是mRNA,还是非编码RNA,都被发现与疾病密切相关,包括但不限于肿瘤[22]、发育疾病[23]和神经退行性疾病[24-25]等。

首次报告于2009年的单细胞转录组(single-cell RNA sequencing,scRNA-seq)[26]及后续不断快速发展的实验室与计算方法[27],推动了大规模细胞图谱项目,例如,分别旨在测序人体和大脑中的所有细胞类型的人类细胞图谱[28]和美国国立卫生研究院(National Institutes of Health,NIH)的脑计划[29]。这些项目的成果包括发现新的细胞类型[30]和癌症的不同发病机制[31]。

scRNA-seq正迅速成为生物学家工具包中的标准组成部分,未来10年可能会像今天的批量RNA测序一样广泛使用。scRNA-seq也可以进一步结合空间组学方法,通过"空间编码"[32]和"原位转录组学"[32]定位空间转录信息。空间编码方法在RNA测序文库准备期间记录空间信息,如激光捕获显微切割(laser capture microdissection,LCM)和直接从组织切片捕获mRNA。原位转录组学方法则通过测序或成像在组织切片中生成数据。空间转录组学方法通过在组织中定位分子定义的细胞类型,同时检测其形态、活动或连接性,可以将分子细胞类型与形态、生理和行为相关性联系起来,这对于理解大脑功能、发育和疾病十分重要[33]。目前,所有空间组学方法在深度转录组数据、细胞分辨率和高成本方面都存在限制,但正在快速改进,并已应用于临床样本[34]。如果能够克服技术限制,未来空间组学的应用将更加广泛。

1.3　蛋白质组

蛋白质组学能够最大限度地识别和量化细胞或组织中的所有蛋白质。常规临床实验室,如酶联免疫吸附分析(enzyme-linked immunosorbent assay,ELISA)通常只测量少量血液蛋白,但血浆中包含来自各个器官和组织的数千种蛋白质。蛋白质的大规模研究主要通过质谱法、基于亲和力的蛋白质组检测、液相蛋白质芯片检测[35]。其中,质谱法被视为精确蛋白质检测和发现的金标准,而基于亲和力的蛋白质组检测可能是更具成本效益和可扩展性的定量选择。基于亲和力蛋白质组检测方法主要分为2类:基于抗体亲和的方法,以Olink公司为代表;基于适体亲和的方法[36],以SomaLogic公司为代表。这2种方法都可以实现高通量、多重免疫测定,允许同时检测血清、血浆及其他体液或组织提取物中的数千种蛋白质,从而实现系统性蛋白质分析的方法。

尽管蛋白质组学技术能够识别大量蛋白质,但这些测量的敏感性、特异性和可重复性可能会有所不同,尤其是对于低丰度蛋白质。解释蛋白质组学数据需要理解潜在的混杂因素,包括样品处理、储存条件和个体间的固有差异。此外,蛋白质和蛋白质的相互作用也可以通过免疫沉淀和质谱的结合来识别,例如,通过抗体纯化目标蛋白,然后通过质谱检测相互作用的未知蛋白,最终获得相互作用的蛋白[37]。值得注意的是,蛋白质翻译后的修饰,如磷酸化、糖基化、泛素化、乙酰化和硝基化,在蛋白质翻译后广泛存在,这些修饰对于细胞内信号转导、蛋白质运输和酶活性至关重要[38-39]。

1.4　代谢组

"代谢组"指的是与细胞代谢相关的所有小分子,包括氨基酸、脂类、碳水化合物和微量元素等。临床上经常会在血浆、尿液、唾液或其他体液中测量几种特定的代谢物和激素,其中还包括呼气中的内源性和外源性挥发性有机化合物。一般而言,代谢物分析可以立即反映细胞生理的动态变化,代谢物水平或比例异常可能导致疾病[40]。值得注意的是,代谢物化学性质多样,丰度变化大,这给高通量代谢组学检测带来了挑战。尽管可以通过有针对性的方法进行特定假设的测试和定量精确测量,但系统级分析通常需要无针对性的方法以检测尽可

能多的代谢物。与代谢组学相关的数据库和技术包括核磁共振(nuclear magnetic resonance,NMR)[41]和基于质谱(mass spectrometry,MS)的方法(气相色谱-质谱(gas chromatography MS,GC-MS)、液相色谱串联质谱(liquid chromatograph MS,LC-MS)和毛细管电泳-质谱(capillary electrophoresis MS,CE-MS)。质谱法[42-43]是当前主要的商业代谢组学应用基础,能够定量分析1000~2000种代谢物,这提供了关于异质性的更多重要信息。家用技术也在兴起,例如,通过安装在马桶上的设备捕获并分析尿液生物标志物[44]。然而,区分真正相关的代谢变化和无害波动仍然是一个限制,特别是要考虑由饮食到昼夜节律等多种因素会影响代谢物水平。

人类代谢组数据库(Human Metabolome Database)是一个包含人体内小分子代谢物详细信息的免费数据库[45],可用于代谢组学研究,相应的代谢物分析可以通过MetaboAnalyst 5.0平台进行[45]。Karsten对目前已发表的、关于利用非靶向代谢组学进行代谢组学全基因组关联研究(mGWAS)分析的文章进行了总结,并提供了相关链接(http://www.metabolomix. com /list-of-all-published-gwas-with-metabolomics),可以为研究者提供参考。

1.5 微生物组

开创性工作研究迅速揭示了人体内寄居着一个由微生物组成的生态系统,主要包括4个区域:肠道、皮肤、阴道和口腔微生物群[46]。常用的调查微生物组成的方法包括测序扩增标记基因(如编码16S rRNA的基因)[47],而宏基因组DNA测序则能够研究样本中所有基因序列的全貌[48]。微生物组学与代谢组学相辅相成,人体血液中已知有超过200种化合物由微生物代谢或独立产生[49],并且可以通过算法在一定程度上从微生物组数据推算个体的代谢组状态[50]。

微生物组在影响人体生理方面的重要性日益凸显,包括对大脑[51]、神经发育[52]、肝脏[53]、免疫系统[53]和消化健康[54]的影响,并在药物疗效[55]和代谢[56]中起关键作用。随着时间推移,跟踪微生物组信号有助于丰富现有的表型数据,并理解个人健康轨迹的趋势。由于微生物组是一个复杂且动态的生态系统,其数据受个体间巨大差异的限制,这些差异源于遗传因素、饮食、年龄和环境,使得建立"正常"或"健康"微生物组基准变得具有挑战性[57]。此外,微生物组分析的准确性往往依赖于参考数据库的质量,目前的数据库中许多微生物种类(特别是非西方人群中的种类)缺乏或不存在[58]。

1.6 影像组

在医学领域,影像组学指的是从放射影像(计算机断层扫描(computed tomography,CT)、核磁共振成像(nuclear magnetic resonance,NMR)、断层扫描(positron emission tomography,PET)等)中提取感兴趣区域的高通量图像特征,并通过机器学习(machine learning,ML)方法对病变区域和关键信息(如生物标志物)进行精确量化,最终帮助诊断、分类或分级疾病[59]。2014年,英国生物库启动了世界上最大的多模态成像研究,邀请10万名参与者接受脑部、心脏和腹部磁共振成像、双能X射线吸收测量和颈动脉超声检查[60]。大规模多模态成像与丰富的表型和遗传数据相结合,为科学家开展与健康相关的研究提供了前所未有的资源。研究人员能够研究成像表型与广泛的生活方式、环境和遗传因素之间的关系,并探讨这些前因如何通过组织结构和/或功能的变化影响疾病风险。迄今为止,已发表的研究成果主要集中在探索生活方式因素与影像生成表型(imagine derived phenotypes,IDPs)之间的横断面关联。例如,较高的体重指数(body mass index,BMI)和腰臀比与大脑不同区域的体积减小相关[61],而高血压和其他血管风险因素与异常的白质微结构异常和其他不良脑部指标相关[62]。这些早期发现有助于理解血管风险因素与神经退行性疾病(如阿尔茨海默病)之间的关系机制[62]。心血管风险因素也与心脏结构和功能相关[63],此外,还发现了其他一些不太明显的关联,如与空气污染、绝经激素治疗和肺功能等的关联。

新的研究成果也不断涌现。例如,糖尿病与心脏4个腔室的异常形态和功能相关,而以前通常认为只有左室受到糖尿病的影响[64]。在新冠肺炎疫情期间,一项研究使用磁共振成像对英国生物银行的785名参与者进行了2次磁共振成像扫描,其中

401名参与者在2次扫描之间的平均间隔141 d内被检测出SARS-CoV-2感染阳性,并与384名未感染的对照组进行了比较。研究结果表明,感染SARS-CoV-2的参与者在2次扫描之间显示出明显的长期效应,与对照组相比,感染SARS-CoV-2的参与者在颞叶前额叶皮质和海马回的灰质厚度和组织对比度方面减少更多,与主要嗅觉皮层功能连接的区域在组织损伤标志物方面发生更大变化。此外,SARS-CoV-2感染组的全脑尺寸减小更多,感染SARS-CoV-2的参与者在2次时间点之间的认知能力下降也更明显[65]。这证明了新冠感染对脑健康的长期影响。另外,一项基于IDPs与精神障碍的孟德尔随机研究揭示了9种IDPs存在对精神分裂症、厌食症和双相情感障碍风险具有因果影响的证据[66]。

在影像组学中,类别不平衡和过拟合是常见问题,例如,在队列中某些疾病的低发病率导致无法区分PET图像中受影响和未受影响的病变区域[67]。

1.7　其他表型组

数字健康相关技术的发展,将有助于实现个性化健康和大规模远程医疗。最初,手表式设备的设计是为了计算每日步数以鼓励锻炼,但近年来,心脏超声等新设备,提供了深入测量和跟踪大量生活方式和生理因素的新方法,包括心率、心率变异性、睡眠模式和血氧水平,甚至可以进行双点心电图[68]。在数字健康领域,“可穿戴设备”的作用在过去10年中不断增强。各种设备可以进行非侵入性信号测量(如光学测量)和生物流体(如汗液或间质液)测量[69]。连续葡萄糖监测器可供糖尿病患者使用,并开始在一些肥胖治疗计划中使用。此外,过去10年中,智能手机的数字健康应用程序数量也在不断增长,这一趋势可能会继续下去。将这些信号与分子数据相关联,可以开发候选的数字生物标志物,从而揭示个体的健康状态,甚至针对个体特定器官或系统进行测量。

使用数字健康工具对于评估和监测脑健康特别有益,因为直接收集用于产生组学数据的生物样本比较困难。越来越多的可穿戴设备作为交感神经和副交感神经活动的指标可以测量心率变异性。基于应用程序的可扩展评估,能够应用于已知的具有潜在临床意义的功能领域[70]。数字健康工具通过提供每日体力活动输出数据,为个体的表型数据增添了丰富性,并与健康变量的网络相互关联。然而,需要注意的是,可穿戴设备的准确性、持久性和用户合规性可能会影响所收集数据的一致性和可靠性。

2　组学大数据的分析方法

与多组学技术发展并驾齐驱的是组学大数据分析方法。过去几十年,科学家研发了多种具有里程碑意义的分析方法,并不断对其进行完善,以揭示疾病发生发展的潜在分子机制。

2.1　全基因组关联分析

全基因组关联研究(genome-wide association studies,GWAS)是在人类全基因组范围内识别存在的序列变异。其中最常见的变异形式是单核苷酸多态性变异(single nucleotide polymorphism,SNP)。自2005年第1项GWAS在老年黄斑性病变研究[71]中取得成功以来,GWAS在随后的15年间彻底改变了复杂疾病遗传学领域,为人类复杂的性状和疾病提供了许多有力的关联证据[71]。截至2022年7月,约6000项研究报道了有400000余个在遗传变异和常见疾病性状之间具有全基因组显著性的遗传关联($P<5\times10^{-8}$)[72-73]。目前,在国际上,如卒中[74]、站立身高[75]、吸烟[76]、受教育水平[77]和血压[78]等复杂疾病和性状的全基因组关联研究的样本量都超过了100万,成功定位了多个重要基因。

尽管GWAS获得了巨大的成功,但也面临一些重要挑战。首先是统计功效和准确性的问题。复杂疾病与性状往往由多种基因和环境因素共同作用决定。如果基因效应较小,则需要大量样本才能识别这些基因位点,耗费大量时间和资源。此外,数据的异质性和多重比较也会影响统计的功效和准确性。最后,基于SNP芯片技术的GWAS依赖于预先存在的遗传变异参考组合,难以检测超罕见突变(次要等位基因频率(minor allele frequency,MAF)<0.01%)以及较为复杂的插入缺失和结构

性变异。

应对这些问题的主要策略有2种。首先是尽可能增大样本量。随着测序技术的推广,结合了特定人群参考面板先验信息的低深度测序方法将会是最优的基因组关联研究策略[79-81]。此外,通过算法研发可以在一定程度上提高关联分析检验的功效。例如,使用分层荟萃分析方法(stratified meta-analysis,SMA)和联合混合模型方法(joint mixed model,JMM)[82]可减少群体结构带来的问题;通过加入多基因评分作为协变量[83],可以提高祖先内部和跨祖先的全基因组关联研究的功效;通过负荷检测等方法可以增加罕见疾病突变的检出功效。随着机器学习和长读长测序技术的发展,越来越多包含了结构性变异的高质量人群单体型参考面板得以公布。

2.2 全基因组关联信号的解读

完成全基因组关联分析后的一个重要问题涉及解读与验证关联信号。连锁不平衡(linkage disequilibrium,LD)现象[84]、广泛存在的非编码区域[85]和基因多效性[86]为GWAS确定因果变异和基因带来相当大的挑战。迄今为止,能够将复杂性状的致病变异与介导这种易感性的分子和细胞功能联系起来的研究相对较少。一个早期的成功案例描述了*SORT1*非编码变异如何影响低密度脂蛋白胆固醇和心肌梗死风险[87]。最近的一些研究则集中于肥胖相关的*FTO*基因内含子变异、*IRX3*和*IRX5*表达的改变,以及脂肪细胞[88]和下丘脑功能[89]之间的关系。类似的功能描述也出现在与精神分裂症[90]和心血管疾病[91]相关的单个位点的研究中。过去10年里,功能基因组学努力将这种“一次一个位点”的工作流程转变为系统的、多维的、整合的方法,希望能够提供与全基因组变异发现相匹配的全基因组功能分析。

在分子层面上,一个基石是生成全基因组范围的功能活动目录。例如,DNA元素百科全书(Encyclopedia of DNA Elements,ENCODE)和表观基因组学路线图(Roadmap Epigenomics)项目已经在数百种细胞类型和组织中生成了组蛋白修饰、转录因子结合、染色质可及性、三维基因组结构和其他调控注释的图谱[85, 92]。这些数据与GWAS结果之间的基因组重叠模式使得风险变异的功能推断成为可能,提供了驱动疾病发病机制的特定细胞类型的线索[93-94],并加速了特定位点的机制见解。

与此同时,业界也在努力拓展连接性状相关的调控变异与其在与疾病相关的细胞类型中调控的基因和过程数据。例如,GTEx(Genotype-Tissue Expression)联盟在数百名个体和几十种组织中绘制了成千上万的表达数量性状位点(QTL)[95]。DNA接近性检测(如Hi-C)和单细胞数据可以进一步提供调控变异与其效应基因之间关系的线索。人类生物分子图谱计划(HubMAP)[96]和人类细胞图谱计划(Human Cell Atlas)[28]等项目将提供各种发育阶段中单个人类细胞类型的全面、高分辨率参考图谱,提供新的机会来理解调控遗传变异如何导致细胞和有机体表型。

GWAS结果的比较分析已成为探索不同性状之间病因学联系的有用工具。连锁不平衡评分回归(linkage disequilibrium score regression,LDSC)[97]是常用的遗传相关性分析方法,通过估计每个SNP的LD评分(一种衡量SNP与周围SNP之间LD关联程度的统计量)来量化2个性状的遗传变异的平均比例,既表型间的遗传关联度。然而,LDSC通常需要使用2个性状的大样本量GWAS分析结果才能保证准确性。另一种流行的共定位分析(colocalization analysis)[98]通过贝叶斯方法鉴定2个性状是否由同一个基因组区域的同一个因果变异驱动,从而提供2个性状之间的遗传关联证据。共定位方法假设每个性状在该基因组区域内最多存在一个因果变异,列举了2个性状之间的4种因果变异假设模型及无因果变异假设模型,并通过贝叶斯方法得到不同模型的后验概率。关于基因组区域关联和遗传共享的5个相互排斥的假设如下。H_0:无关联;H_1:仅与性状一关联;H_2:仅与性状二关联;H_3:2个性状通过不同因果变异相关联;H_4:2个性状通过相同因果变异相关联。当$PP.H_4$>0.8时,判断H_4假设成立,既2个性状由同一个因果变异驱动。

快速发展的技术,如大规模平行报告基因测定[99]和CRISPR基因编辑,支持了大规模靶向序列

扰动的功能表征。这些方法的变体使得基因(通过敲除筛选)[100]、调控元件(使用CRISPR干扰和CRISPR激活筛选[100-101])和遗传变异(碱基编辑器[102])的功能评估在更大规模和更高分辨率上成为可能[103]。结合复杂的读取方法,包括高内涵成像[104]和单细胞转录组学及表观基因组学[105-106],这些方法可以生成经验"真相"数据,支持开发预测致病变异、效应转录本和细胞效应的计算模型。最终,这些模型可减少单个研究对所有细胞类型中所有变异功能进行详尽实验表征的需求,提高机制解释的效率。

2.3　多基因风险评分

多基因风险评分(polygenic risk score,PRS)是依据个体全基因组基因型,根据GWAS汇总统计数据中得出的相应基因型效应大小估计值加权,从而量化个体对疾病易感程度的一种评估工具[107]。作为精准医疗的有力工具,PRS可以应用于疾病风险预测和诊断改进,预测疾病的进展和复发,部署精准治疗,提高人群筛查效率[108]。

然而,PRS的准确性往往会受到遗传因素和社会环境因素的影响。不同人群的致病等位基因频率、等位基因效应大小和LD模式均存在差异。即使在相对同质的人群中,PRS的预测准确性也可能因年龄、性别、社会经济地位和环境暴露等社会环境风险因素而异。如果社会环境风险因素与遗传血统相关,可以通过主成分分析(principal component analysis,PCA)矫正PRS模型;但当社会环境因素仅与共同的人口统计学历史有关时,PCA的校正能力有限。

适用于多血统和混合队列的GWAS分析方法的快速发展,有助于提高PRSs在不同人群之间的可转移性。最近的方法允许对2种以上祖先血统的GWAS汇总统计数据进行PRS建模。例如,CT-SLEB[109]使用C+T算法选择目标人群的PRS中要包含的SNP,并使用经验贝叶斯算法有效地估计这些SNP的效应大小。PRS-CSx[110]通过贝叶斯回归和连续收缩先验进行多基因预测,通过整合多个族裔的GWAS汇总统计数据来提升跨群体PRS的预测能力。此外,Amariuta等[111]在707种细胞中针对转录因子介导的细胞特异性调控位点进行表型相关活性转录的推断和建模(inference and modeling of phenotype-related active transcription,IMPACT),并在传统PRS的基础上使用IMPACT注释来降低不同人群间的LD偏差,从而提高PRS的跨祖先可转移性。

2.4　孟德尔随机化研究

孟德尔随机化(Mendelian randomization,MR)是一种重要且常用的基于观察性数据进行因果推断的遗传学分析方法[112-113],在解析疾病发生发展过程中的风险因素和生物机制方面中发挥了重要作用,成功预测多项随机临床试验(randomized clinical trial,RCT)的结果[114-116],并推动了安塞曲匹、托莱西单抗、卡那单抗等药物的发现与应用[117-118]。在等位基因随机分裂过程中,个体根据是否携带改变暴露的遗传变异被自然分成"干预组"和"对照组"。MR使用与长时间暴露关联的遗传变异作为工具变量来分析暴露因素对疾病易感性的影响,有效避免了反向因果和混杂因素干扰,从而更为准确地推测出暴露和结局之间的因果关系。运用MR进行因果推断时,需满足3个重要的假设前提[119]。(1) 相关性(relevance):该工具变量与暴露因素显著关联。(2) 可交换性(exchangeability):工具变量和结果的混杂因素独立。(3) 排除限制性(exclusion restriction):工具变量除了通过暴露因素影响结果外,不直接对结果产生影响。

多变量孟德尔随机化(multivariate Mendelian randomization,MVMR)研究是MR的扩展,允许估计2种或多种暴露对结果的因果影响。在挑选应用于MVMR的遗传工具时,需要注意以下问题:检验遗传工具与某个暴露变量之间是否满足相关性假设时,必须将其他的暴露变量纳入估计[120];无论使用单样本数据或双样本数据,遗传工具个数应该大于或等于暴露变量个数[121];影响1个以上暴露变量但对结果不产生直接影响(除非通过包括的变量)的遗传工具可以被MVMR分析纳入,前提是遗传工具相关的表型都作为暴露变量纳入同一个MVMR模型中[121],从而避免多效性偏差和根据强度选择工具变量而产生的潜在偏差[122];需要检验

SNP结果关联的异质性,可以使用基于个人水平数据的Sargan统计量或基于汇总数据的Cochran Q统计量[121];使用MR-Egger截距检验多效性[123]。

一般来说,单变量MR估计暴露对结果的"总体"影响,而MVMR估计每次暴露对结果的"直接"影响。这些影响是否相同取决于模型中暴露变量之间的关系,以及其他暴露变量与结果之间的关系。假设一个考虑2种暴露的模型:X_1,是感兴趣的主要暴露;X_2,是感兴趣的次要暴露。当X_2是X_1的混杂因素或对撞机因子时,单变量MR和MVMR估计相同的因果效应。当X_2是X_1的中介因素时,单变量MR和MVMR估计的因果效应不相同,其中单变量MR估计X_1对结果的总效应;MVMR估计X_1不通过X_2对结果的直接效应[121]。

然而,随着多组学实验技术的发展,基因—暴露和暴露—结果因果关联的研究越来越复杂,MR分析也面临新的挑战。目前方法学研究主要集中在解决以下MR悖论:MR假设遗传变异仅通过单一途径影响暴露和结局变量,而广泛存在的水平基因多效性现象会导致无效的MR因果估计;如果遗传变异破坏了暴露目标的正常功能(如暴露与靶受体的结合),则有可能会产生矛盾关联(靶受体失效,作为保护因素的生物标志物水平增加,同时疾病风险增高);当遗传变异与同一途径上效应相反的多个依赖性状相关,也会影响因果推断;如果暴露是时间依赖性的,那么尽管MR结果表明存在因果效应,但在生活中改变暴露不一定会改变疾病风险;如果暴露在多年累积后才能导致疾病,MR分析会产生比RCT或观察性流行病学研究更大的因果估计值;将多个性状和遗传变异组合到同一个模型的多变量MR方法正在被广泛应用,但这可能导致重叠性状等问题。

2.5 人工智能算法与分子医学

近年来,机器学习、深度学习和大语言模型等人工智能(artificial intelligence,AI)算法迅猛发展。随着生物数据规模的不断扩大,AI算法逐渐被应用于多组学大数据的分析,并取得了良好成效[124]。

在基因组层面,最重要的进展是将AI应用于变异检测的目标。当个体读取被映射到参考基因组中的相应位置时,它们可以被可视化为一个"堆叠",其中与参考不同的碱基被突出显示。这种视觉呈现有助于在基因组中复杂区域进行快速手动审核,这一借鉴了计算机视觉和图像识别方面进展的见解推动了用于变异检测的深度学习方法的发展[125]。此外,其他变异检测方法则在更窄的应用领域中使用机器学习,例如,用于特定变异或基因组区域的技术校准错误的模式[126]。机器学习在罕见疾病变异的优先级排序方面也表现出极高的效用。例如,一种基于逻辑回归的机器学习利用大量文献衍生数据集将表型与候选基因匹配,以帮助鉴定遗传病的潜在致病基因[127]。另一种方法则应用最大似然估计和贝叶斯网络来达到同样的目的[128]。这些方法在识别罕见遗传疾病方面特别成功,多项研究显示未诊断的遗传疾病的解决率达到30%~50%[129-130]。在一项研究中,来自英国和爱尔兰的13449名先证者,诊断率为41%[131]。此外,与先前的方法相比,纳米孔长读取测序的临床应用不仅提高了准确性[132],而且还有可能在不到8 h的时间内进行罕见疾病背景下的临床诊断[133]。

在转录组层面,最初的转录组分析方法通过将每个基因的表达谱与参考范围进行比较,识别疾病关联的差异表达基因[134]。引入贝叶斯模型后,转录组分析可以进一步预测罕见变异的调控效应[135]。在一大批未诊断的罕见疾病患者中,血液转录组测序在8%的患者中识别出了致病变异[136]。后来,一个包含基因表达、等位基因特异性表达和替代剪接数据的分层贝叶斯模型则被用于识别遗传驱动的转录组异常[137]。AI还被用于解决剪接点预测这一难题。一个使用32层深度神经网络的深度学习模型显示出改善罕见疾病诊断的潜力[138]。自编码器(神经网络的一种,它能够高效地学习如何将输入数据编码为压缩表示,然后再将其解码回原始输入的表示)也展现出改善RNA测序数据中异常剪接预测的能力[139]。这些方法成功应用于一个出现了发育退化、震颤和癫痫的12岁女孩的罕见病案例诊断中。短读长基因组测序确定了96个候选基因变异,但没有一个看起来能够解释患者的病情。通过添加基于患者血液的RNA测序的剪接异常检测

算法,识别出了KCTD7中的一个剪接增益变异,从而确立了进行性肌阵挛性癫痫的诊断[140]。

在蛋白质组层面,深度学习则在几乎整个工作流程都取得了重大进展[141]。例如,通过对已知化学实体的光谱图案进行训练,深度学习的方法改进了候选肽段光谱的预测[142]。利用卷积神经网络工具,还可以准确预测肽段滞留时间,即肽段从液相色谱柱中洗脱的时间点[143]。采用了卷积神经网络和长短期记忆方法,还可以实现全新的肽段测序和蛋白质鉴定[144]。此外,还有研究者将大型语言模型应用于蛋白质功能预测,旨在加速药物发现[145]。另外,深度学习还为临床一个重要焦点——生物标志物的预测提供了好的解决策略。在一项研究中,基于来自5个独立队列约17000名无重大疾病的个体蛋白质定量数据,训练了一系列机器学习模型(包括基于逻辑回归的模型和随机森林)对预防医学领域通常用于预测健康的11个不同指标(如5年内发生心血管主要事件的风险)进行预测。在验证队列中,94种蛋白质的定量预测肝脏脂肪的C统计量为0.83,表明可以用于非侵入性检测非酒精性脂肪肝[146]。辅助机器学习的蛋白质组学方法还发现了酒精性肝病、阿尔茨海默病和帕金森病的循环生物标志物[147]。

在代谢组学层面,深度学习主要用于增强先天代谢错误的诊断。Liu等[148]将非靶向代谢组学筛查方法应用于先天性代谢错误诊断,通过临床表型、分子检测数据等其他临床数据复验诊断结果,并与临床指南推荐的传统靶向代谢组学筛查方法(使用血浆氨基酸、血浆酰基肉碱和尿有机酸作为筛选标志物)对比,发现非靶向代谢组学将诊断率增加了6倍,并被证明是针对非氧化戊糖磷酸途径缺陷的有效策略。在最近的一项研究中,外显子测序结合代谢组学改进了变异分类。例如,代谢指纹法通过支持向量机建立了丙酮酸激酶缺乏症的诊断,支持向量机通过在n维空间中找到一个超平面来识别亚组[149-150]。在另一个实例中,金属蛋白基因的变异提供了多通道卷积神经网络的训练数据,该网络显示与代谢性疾病更密切相关的是金属蛋白的铁结合位点的突变,而不是其他位置的突变[151]。

3　多组学大数据的临床实践

尽管基础科学和技术取得了显著进步,但多组学大数据在主要医学领域的临床应用转化却相对滞后,是需要重点发展的环节。以下是多组学大数据在疾病分型、疾病诊断与预测、药物靶点选择以及个性化医疗等方向的一些较为成功的案例。

3.1　疾病分型

随着临床诊断技术的不断发展,许多疾病的定义正在逐渐完善,但也伴随着疾病内部异质性的加剧和治疗措施与患者情况不匹配的现象。这在复杂疾病中尤为明显,如心血管疾病、精神类疾病,单靠指南推荐的宏观人体测量值和实验室检测指标难以准确分型[74, 152],由此带来的科研开发和临床实践的不良后果不容忽视[153]。通过多组学大数据分析将特定疾病的患者细分为同质亚型,是改善诊断、预测、治疗、预防和预后的一种有效的策略[154]。

随着越来越多的全基因组疾病关联信号的发现,遗传亚分类揭示了对心血管疾病[155]、过敏性疾病[156]、肥胖[157]及糖尿病[158]等复杂疾病异质性的生物学机制。然而,GWAS关联信号往往只能解释复杂疾病变异的一小部分,不同的基因表达模式[159]在复杂疾病内部异质性转变[160]中起到重要作用。Reichart等[161]使用单细胞核RNA测序技术(single nucleus RNA sequencing,snRNAseq),比较不同心肌病基因型之间的转录组特征差异和细胞谱系差异,揭示了基因型与病理性心脏重构之间的内在关联,改变了“心衰是由共同机制引起”的固有推论,为心脏靶向治疗和个性化医疗提供了线索。需要注意的是,表观遗传模式不仅与细胞特异性基因表达和转录因子结合有关[162],还与许多环境因素有关,如吸烟[163]。蛋白质是生物学功能的载体,描绘清晰的蛋白分子网络对于解析疾病异质性和制定后续治疗策略具有重要意义,尤其在肿瘤学方面。Li等[164]通过对结直肠癌患者的基因组学、蛋白质组学和磷酸化蛋白质组学数据进行共识聚类分析,发现原发肿瘤与转移病灶的遗传特征几乎没有差异,而蛋白质组学和磷酸化蛋白质组学联合分析能够很好地区分转移和非转移性患者。相比于蛋白质

组学,代谢组学通常被定义为对小分子化合物(化合物<1500 Da)的综合分析。因为代谢组小分子对生理通量变化具有更高的敏感性,特别适用于心血管[165]、糖尿病[166]等复杂疾病的机制和分型研究,尤其是考虑到遗传和环境因素及其相互作用等因素的作用下。

虽然不同组学数据从各个维度揭示了疾病异质性的生物学机制,但是由于不同学科和疾病领域的研究人员都在提出新的亚型,疾病分类学的不断扩充也伴随着"多样性"的缺点,既缺乏统一定义、方法或统计标准。通过机器学习方法建立多个组学数据之间的关联网络可能是解决这一问题的关键途径。例如,Ding等[167]通过整合7695例卒中患者的临床表型、生物标志物和全基因组测序数据探究非心源性栓塞性缺血性卒中(noncardioembolic ischemic stroke,NCIS)生物异质性和亚群的综合景观。在60个生物标志物中通过分层聚类确定了30个预后相关分子簇,并使用降维分析揭示了与特定生物标志物相关的精细亚群结构。此外,多组学数据整合在癌细胞亚群分析的应用还有助于指导肿瘤靶向治疗,有助于推进癌症精准医学。Liu等[168]通过对食管鳞状细胞癌(esophageal squamous cell carcinomas,ESCCs)患者的多组学数据聚类分析,将ESCCs分为4种亚型,其中免疫调节型(immune modulation,IM)对免疫检查点阻断疗法反应最好。研究者们进一步开发了识别ESCCs患者IM亚型的分类器,并成功以85.7%的敏感性和90%的特异性预测抗PD-1治疗反应。

综上所述,通过基于患者个体的多组学数据,医疗团队可实现精准诊断和制定个性化治疗方案的制定[169]。这种个性化医疗模式可根据患者的遗传背景、表观遗传学特征和生物标志物等信息,为患者提供更有效的治疗方案,提高治疗成功率[170]。

3.2 疾病诊断与预测

受限于医学成像分辨率不足、缺乏生物标志物或生物标志物的灵敏性和特异性不足,以及对疾病生物学机制及影响因素了解不充分,许多疾病临床诊断上容易出现漏诊、误诊、早诊困难、过度诊断及治疗方案不适当导致的药物耐药性和副作用等问题。多组学大数据分析可以使用基因组、分子或成像数据来开发准确的诊断工具,可以帮助医生在疾病早期,甚至在临床症状出现之前进行诊断疾病,进行疾病预后和预测,以及指导治疗和预防策略,是迈向精准医疗的重要一步[171]。肿瘤学[172]和罕见病学是多组学大数据分析疾病诊断与预测应用发展较好的领域。

罕见病患者,尤其是危重婴儿和儿童,通常需要通过快速准确的诊断以尽早得到临床治疗和管理。随着罕见病基因的大量发现、基因检测成本的大幅降低,以及政府在基因检测方面的不断投入,基于快速基因组检测显著改善了罕见病诊断的及时性和公平性[173]。英国10万基因组计划在一项涵盖2183个家庭的4660名参与者和161种罕见疾病的试点研究中,探究了基因组测序对常规护理后未确诊的罕见疾病患者中的作用,发现基因组测序将罕见病诊断率提高到31%~33%,其中25%的患者具有临床可操作性[174]。

基因检测也已经成为癌症非侵入性辅助诊断的重要工具,可以评估种系遗传癌症风险,识别特定类型癌症的体细胞变异,以用于癌症诊断、预测或指导药物治疗[175-176]。2013年,美国医学遗传学学会(ACMG)[177]发布了关于对基因组测序患者的继发性发现进行负责任管理的建议,现已包括82个基因。大量的临床证据证明这些基因导致严重疾病的可能性很高[178-179]。截至2023年7月,美国FDA已经批准了97个经过验证可在临床上用于预后和治疗的肿瘤生物标志物(https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-list)。对于有癌症家族史的人,种系变异的风险评估是疾病预测、筛查和早期检测的关键。在癌症起始的双打击模型中[180],肿瘤抑制基因的一个等位基因被种系变异破坏,第2个等位基因通过体细胞突变被破坏,导致肿瘤发生。虽然癌症相关基因的数据库,如Sanger Cancer Gene Census(https://cancer. sanger.ac.uk/cosmic),包含数百个条目,但只有十几个基因是家族性癌症综合征的主要驱动因素,这些基因的种系变异会增加癌症的风险。肿瘤-正常种系样本配对分析是探究肿瘤种

系变异和获得性体细胞变异复杂相互作用的重要方法[180]。*BRCA*1和其他高外显率癌症易感基因(包括*BRCA*2、*CHEK*2、*PALB*2、*ATM*、*VHL*、*BAP*1和*MSH*2等)的致病性种系变异的临床试验现已成为越来越多癌症类型标准管理的一个组成部分[180]。

循环肿瘤DNA(circulating tumor DNA,ctDNA)指患者体液中脱落于新形成肿瘤并正在发生坏死的体细胞中的DNA。液体活检技术可以鉴定患者血浆中的游离DNA(cell-free DNA,cfDNA)中的ctDNA,并应用于肿瘤治疗反应和耐药性检测、残留病灶监测、治疗指导和早期疾病检测等方面[181]。其中,液体活检最悠久、广泛的应用是提供有关癌症患者观察到的治疗反应和疾病进展差异的潜在遗传原因的信息。例如,cfDNA检测识别EGFR变异型非小细胞肺癌患者接受EGFR抑制剂治疗后出现的表皮生长因子受体(EGFR)T790M突变[182]。此外,液体活检还广泛用于监测治疗耐药性并了解耐药机制。获得性耐药通常以个体患者中多个耐药亚克隆的克隆生长为特征,相比于单病灶肿瘤活检,cfDNA的一个关键优势是能够捕获与耐药性相关的分子异质性。例如,一项比较疾病进展时匹配的肿瘤活检和cfDNA检测的研究表明,在多达2/3的病例中,cfDNA检测可能会揭示单次肿瘤活检无法识别的其他改变[183]。但需要注意的是,由于癌症患者整体cfDNA中ctDNA的比例在不同肿瘤类型、不同个体间差异较大[184],且肿瘤治疗会降低患者ctDNA水平[185],液体活检在低ctDNA水平患者中的应用存在较大困难。

除了基因变异,表观遗传重编程也是肿瘤可塑性[186]和适应性[187]的主要贡献者。近几十年来,大型项目扩大了影响表观遗传因素的癌症相关基因突变的已知范围[188-189],包括调控组蛋白标记的染色质重塑剂和修饰剂[190]、DNA甲基化[191]、microRNA[192]和3D基因组折叠[193],证实了表观遗传畸变在血液和实体恶性肿瘤病因中的作用[194]。

3.3　药物靶点发现

药物研发是一个漫长而昂贵的过程[195],药物基因组学等多组学技术在临床实践中的应用对于药物研发和药物安全至关重要,基因支持的靶点更有可能在Ⅱ期和Ⅲ期试验中取得成功[196-197]。欧美的多家制药巨头都大力投资英国样本生物样本库中50万人的基因组[198]、蛋白组[199]等测序研究,并获得了显著回报[200-202]。实践证明,多组学数据分析在药物靶点的发现和预测、药物重利用及药物相应预测等方面具有极高的应用价值[203]。

基因组学方法是药物靶点发现和预测的新途径。基于GWAS关联信号的生物信息学解读和实验随访可以为药物开发提供信息,有可能通过识别新的药物靶点和使用遗传工具间接测量药物效应来加速药物发现,并为药物重新定位提供证据[204]。例如,Okada等[205]通过GWAS评估了来自超过10万名欧洲和亚洲受试者的约100万个SNP,识别了101个类风湿性关节炎风险位点(其中42个为新发位点),并通过基于功能注释、顺势作用元件数量性状位点、通路分析的生物信息学分析方法,以及基于与人类原发性免疫缺陷、血液癌体细胞突变和基因敲除小鼠表型的遗传共享分析的新方法,在101个遗传位点中鉴定了98个生物学候选基因。在肿瘤学中,测序技术能够识别体细胞癌基因组中的驱动突变,促进了针对这些突变的药物或药物组合的开发,显著改善患者预后[206]。例如,靶向慢性粒细胞白血病(CML)中*BCR-ABL1*融合基因的抗癌药物伊马替尼(酪氨酸激酶抑制剂)的问世,就对CML患者的预后产生了革命性影响,使其预期寿命与普通人群相似[202]。

当药物靶点和疾病结果之间的准确中介尚不清楚时,使用多组学数据分析可以提供新的见解。例如,秋水仙碱[207]在大规模的Ⅲ期试验中显示出心血管疾病(cardiovascular disorder,CVD)风险的有效性,但其确切作用机制未知,对接受该药物治疗的个体进行蛋白质组学探索可能有助于鉴定这些药物修饰的蛋白质[208]。他汀类降脂药物则是另一个例子,研究表明他汀类药物对上皮性卵巢癌具有潜在的化学预防作用,为了进一步确定作用机制,研究者通过MR检查他汀类药物靶点(HMG-CoA还原酶)与卵巢癌的关联,并在2次分析中检查其他降脂药物靶点与卵巢癌的关联,最终确定了遗传代理的HMG-CoA还原酶与上皮性卵巢癌的发生

概率降低显著相关[209]。

转录组学分析是通过药物特征匹配(signature matching)实现药物再利用(drug reperposing)的重要方法,可用于将药物、靶点或疾病的信息与蛋白网络、蛋白质-化合物或蛋白质-代谢物网络的拓扑信息相结合[210-211],以评估药物-疾病相似性和药物-药物相似性。2006年,第1份"连接图"(connectivity map,CMap)[212]参考序列集被公布,该序列集通过共同的基因表达特征将基因、药物和疾病状态联系起来。2017年,随着高效基因表达谱分析方法L1000的出现,下一代CMap[213]的规模较之前扩大了1000多倍,成为囊括了19811种化合物的庞大基因表达数据库。药物基因表达图谱的出现极大地促进了药物重利用的进展。Wei等[214]比较了包含164种药物化合物的基因表达特征谱和对糖皮质激素治疗敏感或耐药的急性淋巴细胞白血病基因表达特征谱,计算两者间的负相关程度并排序,发现免疫抑制剂雷帕霉素对糖皮质激素耐药型急性淋巴细胞白血病的预测疗效最强,这一结论得到了临床前实验的支持[215-216]。

虽然药物基因表达图谱对于药物靶点的发现具有重要意义,但是在应对新冠肺炎这种突发新型传染病时,传统的图谱构建速度仍然不满足需要。因此,基于深度学习的药物基因表达图谱预测方法成为新的研究方向。例如,Pham等[217]开发了一种多头注意力机制驱动的图神经网络方法DeepCE,该方法先利用图卷积网络从数据中自动提取化学子结构特征,再通过注意力机制用于捕获化学亚结构和基因之间,以及细胞系中基因之间的关联,最后,在多输出多层前馈神经网络中预测受新化学物质扰动的差异基因表达谱,并比较基因表达谱与COVID19患者基因表达谱之间的Spearman等级相关分数来筛选药物库中的药物,最终确定的10种药物中,有5种药物被既往研究认为具有临床应用前景。

除了药物靶点选择外,药物疗效[218]和安全性[219]也是药物研发的重点,然而普遍存在的药物响应个体化差异对患者预后产生不利影响,并增加了资源有限的卫生保健系统成本[203],这在肿瘤等复杂疾病的治疗中尤为明显。为了解决这一问题,研究者构建了GDSC(genomics of drug sensitivity in cancer)[220]、CCLE(cancer cell line encyclopedia)[221]等开创性的大型癌症药物响应筛查数据集,这些数据集囊括了癌细胞基因组学、转录组学、表观遗传组学、影像组学、蛋白组学数据,并将这些数据与药物敏感性、基因敲低或敲除数据等功能表征相结合,使科学家能够通过更深入地了解药物基因组学,以实现药物响应预测。此外,为了解决不同数据集之间获得敏感性评分的实验设置和预处理方法的差异问题,Yingtaweesittikul等[222]还开发了用于癌症药物反应预测的综合概率数据库CREAMMIST,该数据库基于贝叶斯框架系统地整合了不同数据集中所有可用的剂量反应值以获得综合剂量反应曲线,使得跨数据集的药物响应预测成为可能。

3.4 临床试验和个体化医疗

临床试验的设计是临床研究中最重要的步骤之一。更好的方案设计能带来更高效的临床试验实施和更快的"继续/停止"决策。此外,设计不良的失败试验带来的损失不仅是财务上的损失,还有社会成本。结合了基因组学、免疫学和精准医学,以及传统临床试验设计的新范式在解决缺血性脑卒中患者的精准治疗中发挥了重要作用。

在20世纪末和21世纪初,生活方式的改变(如戒烟和规律锻炼)、高血压和高脂血症治疗的进步,以及糖尿病血管并发症的更好控制,降低了卒中风险[223]。抗血小板药物也发挥了重要作用,抗血栓试验协作组的荟萃分析显示,抗血小板疗法在2年内将卒中相对风险降低约25%,但增加了出血风险[224]。最初,使用不同机制的双药方案未能进一步降低卒中风险。MATCH[225]和CHARISMA[226]试验未显示双重治疗的益处,MATCH试验还显示出出血风险增加。然而,鉴于组合疗法显示出更强的血小板抑制效果,研究者继续探索短期双重抗血小板治疗(DAPT),以在复发卒中的高风险期内最大化益处并最小化出血风险。CHANCE[226]和POINT[227]试验均显示,在高风险TIA或小卒中后的早期DAPT

有益,POINT试验中较长的治疗时间导致出血率增加。基于这些试验,指南认可并广泛实施TIA或小卒中后的短期DAPT。通过短期使用抗血小板药物的方案进行微调,研究人员展示了一种在抗血栓益处和出血风险之间成功平衡的方法。

Wang等[228]进行了另一项DAPT试验,即CHANCE-2(涉及急性非致残性脑血管事件的高风险患者使用替格瑞洛或氯吡格雷加阿司匹林),该试验考虑了基于药物代谢的基因变异对氯吡格雷疗效的抵抗性。氯吡格雷是一种前体药物,必须通过肝酶*CYP2C19*代谢成活性药物。替格瑞洛与氯吡格雷类似,通过血小板P2Y12腺苷二磷酸受体起作用,但不需要代谢修饰。携带1种或多种*CYP2C19*失活等位基因的患者对氯吡格雷的反应预计会降低。在支持氯吡格雷失活等位基因携带者对氯吡格雷效果降低的假设中,作者发现,在选定人群中,替格瑞洛加阿司匹林比氯吡格雷加阿司匹林更有益。这一著名的CHANCE-2临床试验是首个在临床试验中考虑基因型变异从而取得成功的研究之一,也说明了基于个人基因谱的针对性治疗的重要前景。

个体化医疗致力于将"一刀切"的治疗方式转变成为每一位患者提供最好的照护。随着基因检测可及性越来越高以及医学数据数字化管理不断推进,正处于个性化医疗的加速发展之中[171]。在未来10年中,复杂多组学信息与常规临床护理相整合可能会改变人们今天所了解的医学领域。例如,CRISPR-Cas9工具的可编程性和强大的活性促进了它们快速整合到各种适应证的治疗中,用于过继细胞治疗前的细胞离体工程和体内基因校正[229]。通过小干扰RNA(small interfering RNA,siRNA)靶向抑制目标基因表达的RNA干扰(RNA interference,RNAi)技术也具有极高的应用潜力。Nissen等[230]进行的一起临床试验表明,Zerlasiran(旨在降低机体Lp(a)生产的siRNA药物)单次给药后,Lp(a)浓度较基线最大下降>96%,150 d后疗效开始逐渐减弱,210 d时Lp(a)浓度仍相较于基线下降50%和45%,直到365 d时Lp(a)浓度相较于基线仍下降30%和29%,具有较强降血脂效果。

4　医学伦理与数据共享

尽管科学界越来越重视基因组研究的多样性和包容性,但对于基因组数据供科学研究者访问的呼吁尚未完全解决维护公众信任等问题。

4.1　当前已有的伦理规范和数据保护措施

人类遗传资源的共享和保护一直是各国关注的重点,许多国家通过建立公共人类遗传数据存储库,并对访问者实施限制或约束,以保障数据安全。例如,要访问来自美国国立卫生研究院国家生物技术信息中心(National Center for Biotechnology Information,NCBI)基因型和表型数据库(Database of Genotypes and Phenotypes,dbGaP)和英国生物样本库的数据,研究人员必须向数据访问审查委员会提出申请,该委员会将确保该项目符合知情同意条款。中国也采用了相同的策略。自《人类遗传资源管理条例》和《中华人民共和国生物安全法》颁布实施以来,中国开始从行政法规层面对人类遗传资源信息共享进行规制,并建立了科研领域数据共享制度,即人类遗传资源数据共享平台(Genome Sequence Archive for Human,GSA-human)[231]。根据规定,访问者向GSA-human提交申请并审核通过后,可以在限制期限内下载感兴趣的数据,数据分析的结果属于数据使用者。

在临床实践中,医疗数据的收集也引发了许多法律和道德隐私问题。医疗数据来源不同,如来自医疗电子病历、医疗保险、互联网设备和社交媒体等。美国和欧洲在医疗数据的管理上存在不同[232]。美国隐私法对医疗数据的处理方式取决于数据的创建方式和处理数据的人(即保管人)。相比之下,欧盟《通用数据保护条例》为医疗数据(以及其他数据)规定了一个广义的单一制度,即无论其格式、收集方式或保管人。中国自2016年以来相继发布了《中华人民共和国网络安全法》《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》,建立了个人健康医疗数据保护、数据分类分级与安全合规评估等制度,构建了国家人口健康科学数据中心、国家基因组科学数据中心等国家级科研数据共享平台;但中国目前尚未建立统一的临床医疗健

康数据管理平台,不同的医疗机构、部门和系统之间缺乏有效的互联互通机制。

4.2 未来数据共享的可能发展方向

2016年,国务院发布的《关于促进和规范健康医疗大数据应用发展的指导意见》提出了3项健康医疗数据应用应当坚持的原则:以人为本、驱动创新;规范有序、安全可控;开放融合、共建共享。应当建立一个统一的健康医疗数据共享平台,以提供标准化的数据格式和接口,从而实现医疗机构、部门和系统之间数据的互联互通。同时,该平台应具备安全可靠的数据存储和传输能力,以确保在数据共享过程中的保密性和完整性。

中国未来的健康医疗数据共享平台将促进不同生物样本库之间的数据交换和交互,而人类遗传数据流动受到数据交换和共享法规的约束。因此,研发新的多组学分析技术以应对这样的挑战十分有必要[233]。在基因组学领域,全基因组关联荟萃分析(genome-wide association meta-analysis, GWAMA)是一种共享汇总统计数据的安全方法,但是汇总统计量经过预处理和调整,限制了GWAMA的参数化空间,未来应当加强这一方面的研究。此外,基因型插补(genotype imputation)对于GWAMA最大化统计功效至关重要,但需要高质量的参考面板和更好的种族匹配。在人工智能领域,随着聊天生成预训练转换器(chat generative pre-trained transformer, ChatGPT)在美国医疗执照考试中崭露头角,研究者对基于自然语言处理的个性化医疗大语言模型(large language model, LLM)的研发越来越感兴趣。随着ICD-11编码、医疗数据电子化管理的普及和医疗数据共享的不断推进,LLM在未来的医疗人工智能系统中将会发挥重要作用。但是目前LLM在处理高质量的庞大医疗数据(尤其是影像学数据)方面仍存在较大困难。

5 结论

正如过去20年来对人类基因组的理解程度不断提高一样,人们对于准确描述疾病发生发展过程中的关键生物学过程的能力也在不断提高。可以看到多组学和大数据技术的发展推动了循证医学的发展。随着这些技术的改进,工作重点应该在于基础研究的深入与临床应用的转化。这些技术的持续发展将包括对样本处理的优化:结果需要在几小时或几天内返回给参与者和患者,而不是需要几个月的时间。大量数据的涌入结合强大的人工智能方法将带来可操作的建议和可解释的结果,使临床医生能够前所未有地实施医学实践,专注于保持健康人的健康,防止疾病在其出现之前就发生,并逆转和治愈已发生的疾病。

面向上述愿景,国内外都启动了个性化人口健康与疾病研究计划(表1)。

这些自然人群或专病的队列,在规模和测量深度方面各不相同,范围从遗传学到广泛的纵向表型测量。如果能够把基于这些队列的研究成果转化为临床应用,那么对于理解和促进人类健康将产生非凡的影响。为了实现这一目标,需要创新策略来

表1 部分代表性的组学研究计划

计划	位置	目标	队列规模	数据范围	网址
STROMICS	中国	脑健康	10000+	基因组、多组学、临床表型	http://www.stromics.org.cn
中国慢性病前瞻性队列	中国	人群健康	500000+	基因组、多组学、临床表型	https://www.ckbiobank.org
中国表型组计划	中国	人群基线	10000+	基因组、多组学、人体测量指标	https://hupi.fudan.edu.cn
江苏出生队列	中国	发育健康	10000+	基因组、多组学、人体测量指标	https://jbc-cnbc.njmu.edu.cn
广州出生队列	中国	发育健康	10000+	基因组、多组学、人体测量指标	http://bigcs.com.cn
All of US	美国	人群健康	1000000+	基因组、临床	https://allofus.nih.gov
MVP	美国	人群健康	1000000+	基因组、临床	https://www.mvp.va.gov/pwa
UK Biobank	中国	人群健康	500000+	基因组、多组学、人体测量指标	https://www.ukbiobank.ac.uk
FinnGen	芬兰	人群健康	500000+	基因组、多组学、人体测量指标	https://www.finngen.fi/en

吸引患者并生成必要的证据,以推动新进展进入临床试验,从而改善公共健康。

综上所述,提出以下原则,或许有助于改善后续中国医学研究项目的设计,产生社会、经济效应。

1)促进转化应用。项目伊始,将转化医学的思维应用于基于多组学大数据的研究设计和应用。把基础实验的前提假设建立在人群多组学大数据研究发现的基础上。

2)构建基因解读功能平台。测量代表中国人群的多组学大数据对于将基因组和表型变异的生物学意义置于上下文中至关重要。需要建立GTEx和Human Cell Atlas等数据库。

3)结合人工智能。利用人工智能方法,将大量数据转化为可操作的建议和可解释的结果。

4)促进数据共享。参考欧美国家大型生物样本库的建设经验,从国家层面促进数据共享,降低科研人员采集数据的难度,使其专注于智力创新性工作,可能是有效的提高效率的方法之一。

5)加强合作。在全球范围内推动个性化健康和疾病研究计划,加强各国间及各学科间的合作与数据共享。

参考文献(References)

[1] Vissers L E L M, de Vries B B A, Osoegawa K, et al. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities[J]. American Journal of Human Genetics, 2003, 73(6): 1261-1270.

[2] Consortium I H. The international HapMap Project[J]. Nature, 2003, 426(6968): 789-796.

[3] Frazer K A, Ballinger D G, Cox D R, et al. A second generation human haplotype map of over 3.1 million SNPs [J]. Nature, 2007, 449(7164): 851-861.

[4] Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome[J]. Nature, 2001, 409(6822): 860-921.

[5] Wheeler D A, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing[J]. Nature, 2008, 452(7189): 872-876.

[6] Ng S B, Turner E H, Robertson P D, et al. Targeted capture and massively parallel sequencing of 12 human exomes[J]. Nature, 2009, 461(7261): 272-276.

[7] Pennisi E. A $100 genome? New DNA sequencers could be a 'game changer' for biology, medicine[EB/OL]. [2022-03-01]. https://www. science. org/content/article/100-genome-new-dna-sequencers-could-be-game-changer-bi - ology-medicine.

[8] Wang Y H, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics and applications[J]. Nature Biotechnology, 2021, 39(11): 1348-1365.

[9] Patel A P, Wang M X, Ruan Y F, et al. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease[J]. Nature Medicine, 2023, 29(7): 1793-1803.

[10] Cedar H, Bergman Y. Linking DNA methylation and histone modification: Patterns and paradigms[J]. Nature Reviews Genetics, 2009, 10(5): 295-304.

[11] Loyfer N, Magenheim J, Peretz A, et al. A DNA methylation atlas of normal human cell types[J]. Nature, 2023, 613(7943): 355-364.

[12] Weber M, Davies J J, Wittig D, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells[J]. Nature Genetics, 2005, 37(8): 853-862.

[13] Simpson J T, Workman R E, Zuzarte P C, et al. Detecting DNA cytosine methylation using nanopore sequencing[J]. Nature Methods, 2017, 14(4): 407-410.

[14] Yousefi P D, Suderman M, Langdon R, et al. DNA methylation-based predictors of health: Applications and statistical considerations[J]. Nature Reviews Genetics, 2022, 23(6): 369-383.

[15] Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing[J]. Nature Reviews Genetics, 2018, 19(6): 371-384.

[16] Apicella C, Ruano C S M, Méhats C, et al. The role of epigenetics in placental development and the etiology of preeclampsia[J]. International Journal of Molecular Sciences, 2019, 20(11): 2837.

[17] Lim U, Song M A. Dietary and lifestyle factors of DNA methylation[J]. Methods in Molecular Biology, 2012, 863: 359-376.

[18] Rozek L S, Dolinoy D C, Sartor M A, et al. Epigenetics: Relevance and implications for public health[J]. Annual Review of Public Health, 2014, 35: 105-122.

[19] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics[J]. Nature Reviews Genetics, 2009, 10(1): 57-63.

[20] Byron S A, Van Keuren-Jensen K R, Engelthaler D M,

et al. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges[J]. Nature Reviews Genetics, 2016, 17(5): 257-271.

[21] Stark R, Grzelak M, Hadfield J. RNA sequencing: The teenage years[J]. Nature Reviews Genetics, 2019, 20 (11): 631-656.

[22] Sparano J A, Gray R J, Makower D F, et al. Prospective validation of a 21-gene expression assay in breast cancer[J]. The New England Journal of Medicine, 2015, 373 (21): 2005-2014.

[23] Dear R, Wagstyl K, Seidlitz J, et al. Cortical gene expression architecture links healthy neurodevelopment to the imaging, transcriptomics and genetics of autism and schizophrenia[J]. Nature Neuroscience, 2024, 27(6): 1075-1086.

[24] Crist A M, Hinkle K M, Wang X, et al. Transcriptomic analysis to identify genes associated with selective hippocampal vulnerability in Alzheimer's disease[J]. Nature Communications, 2021, 12(1): 2311.

[25] de Goede O M, Nachun D C, Ferraro N M, et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease[J]. Cell, 2021, 184(10): 2633-2648.e19.

[26] Tang F C, Barbacioru C, Wang Y Z, et al. mRNA-Seq whole-transcriptome analysis of a single cell[J]. Nature Methods, 2009, 6(5): 377-382.

[27] Stegle O, Teichmann S A, Marioni J C. Computational and analytical challenges in single-cell transcriptomics [J]. Nature Reviews Genetics, 2015, 16(3): 133-145.

[28] Regev A, Teichmann S A, Lander E S, et al. The human cell atlas[J]. Elife, 2017, 6: e27041.

[29] Insel T R, Landis S C, Collins F S. Research priorities. The NIH BRAIN initiative[J]. Science, 2013, 340(6133): 687-688.

[30] Montoro D T, Haber A L, Biton M, et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes[J]. Nature, 2018, 560(7718): 319-324.

[31] Young M D, Mitchell T J, Vieira Braga F A, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors[J]. Science, 2018, 361 (6402): 594-599.

[32] Chen J, Suo S B, Tam P P, et al. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq [J]. Nature Protocols, 2017, 12(3): 566-580.

[33] Lein E, Borm L E, Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing[J]. Science, 2017, 358(6359): 64-69.

[34] Karras P, Bordeu I, Pozniak J, et al. A cellular hierarchy in melanoma uncouples growth and metastasis[J]. Nature, 2022, 610(7930): 190-198.

[35] Suhre K, McCarthy M I, Schwenk J M. Genetics meets proteomics: Perspectives for large population-based studies[J]. Nature Reviews Genetics, 2021, 22(1): 19-37.

[36] Assarsson E, Lundberg M, Holmquist G, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability[J]. PLoS One, 2014, 9(4): e95192.

[37] Jia J L, Jin J P, Chen Q, et al. Eukaryotic expression, Co-IP and MS identify BMPR-1B protein-protein interaction network[J]. Biological Research, 2020, 53(1): 24.

[38] Mann M, Jensen O N. Proteomic analysis of post-translational modifications[J]. Nature Biotechnology, 2003, 21 (3): 255-261.

[39] Rosenberger G, Liu Y S, Röst H L, et al. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS[J]. Nature Biotechnology, 2017, 35(8): 781-788.

[40] Newgard C B. Metabolomics and metabolic diseases: Where do we stand? [J]. Cell Metabolism, 2017, 25(1): 43-56.

[41] Viant M R, Rosenblum E S, Tieerdema R S. NMR-based metabolomics: A powerful approach for characterizing the effects of environmental stressors on organism health[J]. Environmental Science & Technology, 2003, 37(21): 4982-4989.

[42] Kennedy A D, Wittmann B M, Evans A M, et al. Metabolomics in the clinic: A review of the shared and unique features of untargeted metabolomics for clinical research and clinical testing[J]. Journal of Mass Spectrometry, 2018, 53(11): 1143-1154.

[43] Würtz P, Kangas A J, Soininen P, et al. Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: A primer on-omic technologies[J]. American Journal of Epidemiology, 2017, 186(9): 1084-1096.

[44] Tasoglu S. Toilet-based continuous health monitoring using urine[J]. Nature Reviews Urology, 2022, 19(4): 219-230.

[45] Wishart D S, Guo A C, Oler E, et al. HMDB 5.0: The human metabolome database for 2022[J]. Nucleic Acids Research, 2022, 50(D1): D622-D631.

[46] Pace N R. A molecular view of microbial diversity and the biosphere[J]. Science, 1997, 276(5313): 734-740.

[47] Claesson M J, Wang Q, O'Sullivan O, et al. Comparison

of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions[J]. Nucleic Acids Research, 2010, 38(22): e200.

[48] Qin J J, Li R Q, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing [J]. Nature, 2010, 464(7285): 59-65.

[49] Heinken A, Hertel J, Acharya G, et al. Genome-scale metabolic reconstruction of 7302 human microorganisms for personalized medicine[J]. Nature Biotechnology, 2023, 41(9): 1320-1331.

[50] Garza D R, van Verk M C, Huynen M A, et al. Towards predicting the environmental metabolome from metagenomics with a mechanistic model[J]. Nature Microbiology, 2018, 3(4): 456-460.

[51] Hattori N, Yamashiro Y. The gut-brain axis[J]. Annals of Nutrition and Metabolism, 2021, 77(Suppl. 2): 1-3.

[52] Yap C X, Henders A K, Alvares G A, et al. Autism-related dietary preferences mediate autism-gut microbiome associations[J]. Cell, 2021, 184(24): 5916-5931. e17.

[53] Tilg H, Adolph T E, Trauner M. Gut-liver axis: Pathophysiological concepts and clinical implications[J]. Cell Metabolism, 2022, 34(11): 1700-1718.

[54] Andoh A, Nishida A. Alteration of the gut microbiome in inflammatory bowel disease[J]. Digestion, 2023, 104 (1): 16-23.

[55] Nichols R G, Peters J M, Patterson A D. Interplay between the host, the human microbiome, and drug metabolism[J]. Human Genomics, 2019, 13(1): 27.

[56] Wilmanski T, Kornilov S A, Diener C, et al. Heterogeneity in statin responses explained by variation in the human gut microbiome[J]. Med, 2022, 3(6): 388-405.e6.

[57] Wilmanski T, Rappaport N, Diener C, et al. From taxonomy to metabolic output: What factors define gut microbiome health?[J]. Gut Microbes, 2021, 13(1): 1-20.

[58] Abdill R J, Adamowicz E M, Blekhman R. Public human microbiome data are dominated by highly developed countries[J]. PLoS Biology, 2022, 20(2): e3001536.

[59] van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: The path to the clinic[J]. Nature Medicine, 2021, 27(5): 775-784.

[60] Littlejohns T J, Holliday J, Gibson L M, et al. The UK Biobank imaging enhancement of 100, 000 participants: Rationale, data collection, management and future directions[J]. Nature Communications, 2020, 11: 2624.

[61] Hamer M, Batty G D. Association of body mass index and waist-to-hip ratio with brain structure: UK Biobank study[J]. Neurology, 2019, 92(6): e594-e600.

[62] Cox S R, Lyall D M, Ritchie S J, et al. Associations between vascular risk factors and brain MRI indices in UK Biobank[J]. European Heart Journal, 2019, 40(28): 2290-2300.

[63] Woodbridge S P, Aung N, Paiva J M, et al. Physical activity and left ventricular trabeculation in the UK Biobank community-based cohort study[J]. Heart, 2019, 105 (13): 990-998.

[64] Jensen M T, Fung K, Aung N, et al. Changes in cardiac morphology and function in individuals with diabetes mellitus: The UK biobank cardiovascular magnetic resonance substudy[J]. Circulation Cardiovascular Imaging, 2019, 12(9): e009476.

[65] Douaud G, Lee S, Alfaro-Almagro F, et al. SARS-CoV-2 is associated with changes in brain structure in UK Biobank[J]. Nature, 2022, 604(7907): 697-707.

[66] Guo J, Yu K, Dong S S, et al. Mendelian randomization analyses support causal relationships between brain imaging-derived phenotypes and risk of psychiatric disorders[J]. Nature Neuroscience, 2022, 25(11): 1519-1527.

[67] Mayerhoefer M E, Materka A, Langs G, et al. Introduction to radiomics[J]. Journal of Nuclear Medicine, 2020, 61(4): 488-495.

[68] Ip J E. Wearable devices for cardiac rhythm diagnosis and management[J]. JAMA, 2019, 321(4): 337-338.

[69] Ates H C, Nguyen P Q, Gonzalez-Macia L, et al. End-to-end design of wearable sensors[J]. Nature Reviews Materials, 2022, 7(11): 887-907.

[70] Öhman F, Hassenstab J, Berron D, et al. Current advances in digital cognitive assessment for preclinical Alzheimer's disease[J]. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2021, 13(1): e12217-e12217.

[71] Klein R J, Zeiss C, Chew E Y, et al. Complement factor H polymorphism in age-related macular degeneration[J]. Science, 2005, 308(5720): 385-389.

[72] Abdellaoui A, Yengo L, Verweij K J H, et al. 15 years of GWAS discovery: Realizing the promise[J]. American Journal of Human Genetics, 2023, 110(2): 179-194.

[73] Sollis E, Mosaku A, Abid A, et al. The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource [J]. Nucleic Acids Research, 2023, 51(D1): D977-D985.

[74] Mishra A, Malik R, Hachiya T, et al. Stroke genetics informs drug discovery and risk prediction across ancestries[J]. Nature, 2022, 611(7934): 115-123.

[75] Yengo L, Vedantam S, Marouli E, et al. A saturated map of common genetic variants associated with human height [J]. Nature, 2022, 610(7933): 704-712.

[76] Liu M Z, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use[J]. Nature Genetics, 2019, 51(2): 237-244.

[77] Lee J J, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals [J]. Nature Genetics, 2018, 50(8): 1112-1121.

[78] Evangelou E, Warren H R, Mosen-Ansorena D, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits[J]. Nature Genetics, 2018, 50(10): 1412-1425.

[79] Li Y, Sidore C, Kang H M, et al. Low-coverage sequencing: Implications for design of complex trait association studies[J]. Genome Research, 2011, 21(6): 940-951.

[80] Huang S J, Liu S Y, Huang M X, et al. The Born in Guangzhou Cohort Study enables generational genetic discoveries[J]. Nature, 2024, 626(7999): 565-573.

[81] Liu S Y, Huang S J, Chen F, et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history[J]. Cell, 2018, 175(2): 347-359.

[82] Peterson R E, Kuchenbaecker K, Walters R K, et al. Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations[J]. Cell, 2019, 179(3): 589-603.

[83] Campos A I, Namba S, Lin S C, et al. Boosting the power of genome-wide association studies within and across ancestries by using polygenic scores[J]. Nature Genetics, 2023, 55(10): 1769-1776.

[84] Slatkin M. Linkage disequilibrium: Understanding the evolutionary past and mapping the medical future[J]. Nature Reviews Genetics, 2008, 9(6): 477-485.

[85] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome[J]. Nature, 2012, 489(7414): 57-74.

[86] Bulik-Sullivan B, Finucane H K, Anttila V, et al. An atlas of genetic correlations across human diseases and traits[J]. Nature Genetics, 2015, 47(11): 1236-1241.

[87] Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus[J]. Nature, 2010, 466(7307): 714-719.

[88] Claussnitzer M, Dankel S N, Kim K H, et al. FTO obesity variant circuitry and adipocyte browning in humans [J]. The New England Journal of Medicine, 2015, 373 (10): 895-907.

[89] Smemo S, Tena J J, Kim K H, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3[J]. Nature, 2014, 507(7492): 371-375.

[90] Sekar A, Bialas A R, de Rivera H, et al. Schizophrenia risk from complex variation of complement component 4 [J]. Nature, 2016, 530(7589): 177-183.

[91] Gupta R M, Hadaya J, Trehan A, et al. A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression[J]. Cell, 2017, 170(3): 522-533.e15.

[92] Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes[J]. Nature, 2015, 518(7539): 317-330.

[93] Farh K K H, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants[J]. Nature, 2015, 518(7539): 337-343.

[94] Tansey K E, Cameron D, Hill M J. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks[J]. Genome Medicine, 2018, 10(1): 14.

[95] Battle A, Brown C D, Engelhardt B E, et al. Genetic effects on gene expression across human tissues[J]. Nature, 2017, 550(7675): 204-213.

[96] HuBMAP Consortium. The human body at cellular resolution: The NIH Human Biomolecular Atlas Program[J]. Nature, 2019, 574(7777): 187-192.

[97] Bulik-Sullivan B K, Loh P R, Finucane H K, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies[J]. Nature Genetics, 2015, 47(3): 291-295.

[98] Giambartolomei C, Vukcevic D, Schadt E E, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics[J]. PLoS Genetics, 2014, 10(5): e1004383.

[99] Patwardhan R P, Lee C, Litvin O, et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis[J]. Nature Biotechnology, 2009, 27(12): 1173-1175.

[100] Potting C, Crochemore C, Moretti F, et al. Genome-wide CRISPR screen for PARKIN regulators reveals transcriptional repression as a determinant of mitophagy [J]. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(2): E180-E189.

[101] Fulco C P, Nasser J, Jones T R, et al. Activity-by-con-

tact model of enhancer-promoter regulation from thousands of CRISPR perturbations[J]. Nature Genetics, 2019, 51(12): 1664-1669.

[102] Komor A C, Kim Y B, Packer M S, et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage[J]. Nature, 2016, 533(7603): 420-424.

[103] Findlay G M, Daza R M, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing[J]. Nature, 2018, 562(7726): 217-222.

[104] Feldman D, Singh A, Schmid-Burgk J L, et al. Optical pooled screens in human cells[J]. Cell, 2019, 179(3): 787-799.

[105] Dixit A, Parnas O, Li B Y, et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens[J]. Cell, 2016, 167(7): 1853-1866.

[106] Datlinger P, Rendeiro A F, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout[J]. Nature Methods, 2017, 14(3): 297-301.

[107] Choi S W, Mak T S H, O'Reilly P F. Tutorial: A guide to performing polygenic risk score analyses[J]. Nature Protocols, 2020, 15(9): 2759-2772.

[108] Torkamani A, Wineinger N E, Topol E J. The personal and clinical utility of polygenic risk scores[J]. Nature Reviews Genetics, 2018, 19(9): 581-590.

[109] Zhang H Y, Zhan J N, Jin J, et al. A new method for multiancestry polygenic prediction improves performance across diverse populations[J]. Nature Genetics, 2023, 55(10): 1757-1768.

[110] Ruan Y F, Lin Y F, Feng Y C A, et al. Improving polygenic prediction in ancestrally diverse populations[J]. Nature Genetics, 2022, 54(5): 573-580.

[111] Amariuta T, Ishigaki K, Sugishita H, et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements[J]. Nature Genetics, 2020, 52(12): 1346-1354.

[112] Smith G D, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease?[J]. International Journal of Epidemiology, 2003, 32(1): 1-22.

[113] Pingault J B, O'Reilly P F, Schoeler T, et al. Using genetic data to strengthen causal inference in observational research[J]. Nature Reviews Genetics, 2018, 19: 566-580.

[114] Ference B A, Kastelein J J P, Ginsberg H N, et al. Association of genetic variants related to CETP inhibitors and statins with lipoprotein levels and cardiovascular risk[J]. JAMA, 2017, 318(10): 947-956.

[115] Ference B A, Ray K K, Catapano A L, et al. Mendelian randomization study of ACLY and cardiovascular disease[J]. New England Journal of Medicine, 2019, 380(11): 1033-1042.

[116] Ference B A, Yoo W, Alesh I, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: A Mendelian randomization analysis[J]. Journal of the American College of Cardiology, 2012, 60(25): 2631-2639.

[117] Holmes M V, Richardson T G, Ference B A, et al. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development[J]. Nature Reviews Cardiology, 2021, 18(6): 435-453.

[118] Ridker P M, Everett B M, Thuren T, et al. Antiinflammatory therapy with canakinumab for atherosclerotic disease[J]. The New England Journal of Medicine, 2017, 377(12): 1119-1131.

[119] Sanderson E, Glymour M M, Holmes M V, et al. Mendelian randomization[J]. Nature Reviews Methods Primers, 2022, 2: 6.

[120] Sanderson E. Multivariable Mendelian randomization and mediation[J]. Cold Spring Harbor Perspectives in Medicine, 2021, 11(2): a038984.

[121] Sanderson E, Davey Smith G, Windmeijer F, et al. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings[J]. International Journal of Epidemiology, 2019, 48(3): 713-727.

[122] Burgess S, Thompson S G. Avoiding bias from weak instruments in Mendelian randomization studies[J]. International Journal of Epidemiology, 2011, 40(3): 755-764.

[123] Rees J M B, Wood A M, Burgess S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy[J]. Statistics in Medicine, 2017, 36(29): 4705-4718.

[124] Gomes B, Ashley E A. Artificial intelligence in molecular medicine[J]. The New England Journal of Medicine, 2023, 388(26): 2456-2465.

[125] Poplin R, Chang P C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks[J]. Nature Biotechnology, 2018, 36(10): 983-

987.

[126] DePristo M A, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data[J]. Nature Genetics, 2011, 43(5): 491-498.

[127] Birgmeier J, Haeussler M, Deisseroth C A, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature[J]. Science Translational Medicine, 2020, 12(544): eaau9113.

[128] De La Vega F M, Chowdhury S, Moore B, et al. Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases[J]. Genome Medicine, 2021, 13(1): 153.

[129] Splinter K, Adams D R, Bacino C A, et al. Effect of genetic diagnosis on patients with previously undiagnosed disease[J]. The New England Journal of Medicine, 2018, 379(22): 2131-2139.

[130] Lee H N, Deignan J L, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders[J]. JAMA, 2014, 312(18): 1880.

[131] Wright C F, Campbell P, Eberhardt R Y, et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland[J]. The New England Journal of Medicine, 2023, 388(17): 1559-1571.

[132] Dewey F E, Grove M E, Pan C P, et al. Clinical interpretation and implications of whole-genome sequencing [J]. JAMA, 2014, 311(10): 1035.

[133] Gorzynski J E, Goenka S D, Shafin K, et al. Ultrarapid nanopore genome sequencing in a critical care setting [J]. The New England Journal of Medicine, 2022, 386(7): 700-702.

[134] Li X, Battle A, Karczewski K J, et al. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants[J]. American Journal of Human Genetics, 2014, 95(3): 245-256.

[135] Li X, Kim Y, Tsang E K, et al. The impact of rare variation on gene expression across tissues[J]. Nature, 2017, 550(7675): 239-243.

[136] Frésard L, Smail C, Ferraro N M, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts[J]. Nature Medicine, 2019, 25(6): 911-919.

[137] Ferraro N M, Strober B J, Einson J, et al. Transcriptomic signatures across human tissues identify functional rare genetic variation[J]. Science, 2020, 369(6509): eaaz5900.

[138] Jaganathan K, Panagiotopoulou K S, McRae J F, et al. Predicting splicing from primary sequence with deep learning[J]. Cell, 2019, 176(3): 535-548.e24.

[139] Mertes C, Scheller I F, Yépez V A, et al. Detection of aberrant splicing events in RNA-seq data using FRASER[J]. Nature Communications, 2021, 12(1): 529.

[140] Park C Y, Zhou J, Wong A K, et al. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk[J]. Nature Genetics, 2021, 53(2): 166-173.

[141] Wen B, Zeng W F, Liao Y X, et al. Deep learning in proteomics[J]. Proteomics, 2020, 20(21/22): e1900335.

[142] Zhou X X, Zeng W F, Chi H, et al. pDeep: Predicting MS/MS spectra of peptides with deep learning[J]. Analytical Chemistry, 2017, 89(23): 12690-12697.

[143] Bouwmeester R, Gabriels R, Hulstaert N, et al. DeepLC can predict retention times for peptides that carry as-yet unseen modifications[J]. Nature Methods, 2021, 18(11): 1363-1369.

[144] Sinitcyn P, Richards A L, Weatheritt R J, et al. Global detection of human variants and isoforms by deep proteome sequencing[J]. Nature Biotechnology, 2023, 41(12): 1776-1786.

[145] Brandes N, Goldman G, Wang C H, et al. Genome-wide prediction of disease variant effects with a deep protein language model[J]. Nature Genetics, 2023, 55(9): 1512-1522.

[146] Williams S A, Kivimaki M, Langenberg C, et al. Plasma protein patterns as comprehensive indicators of health[J]. Nature Medicine, 2019, 25(12): 1851-1857.

[147] Mann M, Kumar C, Zeng W F, et al. Artificial intelligence for proteomics and biomarker discovery[J]. Cell Systems, 2021, 12(8): 759-770.

[148] Liu N, Xiao J, Gijavanekar C, et al. Comparison of untargeted metabolomic profiling vs traditional metabolic screening to identify inborn errors of metabolism[J]. JAMA Network Open, 2021, 4(7): e2114155.

[149] Shayota B J, Donti T R, Xiao J, et al. Untargeted metabolomics as an unbiased approach to the diagnosis of inborn errors of metabolism of the non-oxidative branch of the pentose phosphate pathway[J]. Molecular Genetics and Metabolism, 2020, 131(1/2): 147-154.

[150] Van Dooijeweert B, Broeks M H, Verhoeven-Duif N M, et al. Untargeted metabolic profiling in dried blood spots identifies disease fingerprint for pyruvate kinase deficiency[J]. Haematologica, 2021, 106(10): 2720-2725.

[151] Koohi-Moghadam M, Wang H B, Wang Y C, et al. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach[J]. Nature Machine Intelligence, 2019, 1(12): 561-567.

[152] Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates?[J]. Nature Reviews Drug Discovery, 2004, 3(8): 711-715.

[153] Kola I, Bell J. A call to reform the taxonomy of human disease[J]. Nature Reviews Drug Discovery, 2011, 10 (9): 641-642.

[154] Johansson Å, Andreassen O A, Brunak S, et al. Precision medicine in complex diseases-Molecular subgrouping for improved prediction and treatment stratification[J]. Journal of Internal Medicine, 2023, 294(4): 378-396.

[155] Antman E M, Loscalzo J. Precision medicine in cardiology[J]. Nature Reviews Cardiology, 2016, 13(10): 591-602.

[156] Melén E, Koppelman G H, Vicedo-Cabrera A M, et al. Allergies to food and airborne allergens in children and adolescents: Role of epigenetics in a changing environment[J]. The Lancet Child & Adolescent Health, 2022, 6(11): 810-819.

[157] Loos R J F, Yeo G S H. The genetics of obesity: From discovery to biology[J]. Nature Reviews Genetics, 2022, 23(2): 120-133.

[158] Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables[J]. The Lancet Diabetes & Endocrinology, 2018, 6(5): 361-369.

[159] Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals[J]. Nature Genetics, 2003, 33(Suppl): 245-254.

[160] Parreno V, Loubiere V, Schuettengruber B, et al. Transient loss of Polycomb components induces an epigenetic cancer fate[J]. Nature, 2024, 629(8012): 688-696.

[161] Reichart D, Lindberg E L, Maatz H, et al. Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies[J]. Science, 2022, 377 (6606): eabo1984.

[162] Schübeler D. Function and information content of DNA methylation[J]. Nature, 2015, 517(7534): 321-326.

[163] Besingi W, Johansson A. Smoke-related DNA methylation changes in the etiology of human disease[J]. Human Molecular Genetics, 2014, 23(9): 2290-2297.

[164] Li C, Sun Y D, Yu G Y, et al. Integrated omics of metastatic colorectal cancer[J]. Cancer Cell, 2020, 38(5): 734-747.

[165] Laaksonen R, Ekroos K, Sysi-Aho M, et al. Plasma ceramides predict cardiovascular death in patients with stable coronary artery disease and acute coronary syndromes beyond LDL-cholesterol[J]. European Heart Journal, 2016, 37(25): 1967-1976.

[166] Wigger L, Cruciani-Guglielmacci C, Nicolas A, et al. Plasma dihydroceramides are diabetes susceptibility biomarker candidates in mice and humans[J]. Cell Reports, 2017, 18(9): 2269-2279.

[167] Ding L L, Liu Y, Meng X, et al. Biomarker and genomic analyses reveal molecular signatures of non-cardioembolic ischemic stroke[J]. Signal Transduction and Targeted Therapy, 2023, 8(1): 222.

[168] Liu Z H, Zhao Y H, Kong P Z, et al. Integrated multi-omics profiling yields a clinically relevant molecular classification for esophageal squamous cell carcinoma [J]. Cancer Cell, 2023, 41(1): 181-195.

[169] Abul-Husn N S, Kenny E E. Personalized medicine and the power of electronic health records[J]. Cell, 2019, 177(1): 58-69.

[170] Shendure J, Findlay G M, Snyder M W. Genomic medicine-progress, pitfalls, and promise[J]. Cell, 2019, 177 (1): 45-57.

[171] Yurkovich J T, Evans S J, Rappaport N, et al. The transition from genomics to phenomics in personalized population health[J]. Nature Reviews Genetics, 2024, 25 (4): 286-302.

[172] Vockley J G, Niederhuber J E. Diagnosis and treatment of cancer using genomics[J]. BMJ, 2015, 350: h1832.

[173] Stark Z, Dolman L, Manolio T A, et al. Integrating genomics into healthcare: A global responsibility[J]. American Journal of Human Genetics, 2019, 104(1): 13-20.

[174] 100,000 Genomes Project Pilot Investigators. 100,000 genomes pilot on rare-disease diagnosis in health care-preliminary report[J]. New England Journal of Medicine, 2021, 385(20): 1868-1880.

[175] Thavaneswaran S, Rath E, Tucker K, et al. Therapeutic implications of germline genetic findings in cancer[J]. Nature Reviews Clinical Oncology, 2019, 16(6): 386-396.

[176] Chakravarty D, Solit D B. Clinical cancer genomic profiling[J]. Nature Reviews Genetics, 2021, 22(8): 483-501.

[177] ACMG Board of Directors. ACMG policy statement: Updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing[J]. Genetics in Medicine, 2015, 17(1): 68-69.

[178] Miyazaki N, Kobayashi T, Komiya T, et al. Postoperative malignant hyperthermia confirmed by calcium-induced calcium release rate after breast cancer surgery, in which prompt recognition and immediate dantrolene administration were life-saving: A case report[J]. Journal of Medical Case Reports, 2021, 15(1): 201.

[179] Narod S A, Foulkes W D. BRCA1 and BRCA2: 1994 and beyond[J]. Nature Reviews Cancer, 2004, 4(9): 665-676.

[180] Knudson A G Jr. Mutation and cancer: Statistical study of retinoblastoma[J]. Proceedings of the National Academy of Sciences of the United States of America, 1971, 68(4): 820-823.

[181] Xie W, Suryaprakash S, Wu C, et al. Trends in the use of liquid biopsy in oncology[J]. Nature Reviews Drug Discovery, 2023, 22(8): 612-613.

[182] Oxnard G R, Thress K S, Alden R S, et al. Association between plasma genotyping and outcomes of treatment with osimertinib (AZD9291) in advanced non-small-cell lung cancer[J]. Journal of Clinical Oncology, 2016, 34(28): 3375-3382.

[183] Strickler J H, Loree J M, Ahronian L G, et al. Genomic landscape of cell-free DNA in patients with colorectal cancer[J]. Cancer Discovery, 2018, 8(2): 164-173.

[184] Diehl F, Schmidt K, Choti M A, et al. Circulating mutant DNA to assess tumor dynamics[J]. Nature Medicine, 2008, 14(9): 985-990.

[185] Bettegowda C, Sausen M, Leary R, et al. Abstract 5606: Detection of circulating tumor DNA in early and late stage human malignancies[J]. Cancer Research, 2014, 74(19_Supplement): 5606.

[186] Nam A S, Chaligne R, Landau D A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics[J]. Nature Reviews Genetics, 2021, 22(1): 3-18.

[187] Shaffer S M, Dunagin M C, Torborg S R, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance[J]. Nature, 2017, 546(7658): 431-435.

[188] Stunnenberg H G, Hirst M. The international human epigenome consortium: A blueprint for scientific collaboration and discovery[J]. Cell, 2016, 167(5): 1145-1149.

[189] Hutter C, Zenklusen J C. The cancer genome atlas: Creating lasting value beyond its data[J]. Cell, 2018, 173(2): 283-285.

[190] Piunti A, Shilatifard A. Epigenetic balance of gene expression by Polycomb and COMPASS families[J]. Science, 2016, 352(6290): aad9780.

[191] Müller D, Gyrffy B. DNA methylation-based diagnostic, prognostic, and predictive biomarkers in colorectal cancer[J]. Biochimica et Biophysica Acta Reviews on Cancer, 2022, 1877(3): 188722.

[192] Pon J R, Marra M A. Driver and passenger mutations in cancer[J]. Annual Review of Pathology, 2015, 10: 25-50.

[193] Kloetgen A, Thandapani P, Tsirigos A, et al. 3D chromosomal landscapes in hematopoiesis and immunity[J]. Trends in Immunology, 2019, 40(9): 809-824.

[194] Feinberg A P, Koldobskiy M A, Göndör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression[J]. Nature Reviews Genetics, 2016, 17(5): 284-299.

[195] Wouters O J, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018[J]. JAMA, 2020, 323(9): 844-853.

[196] Ochoa D, Karim M, Ghoussaini M, et al. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs[J]. Nature Reviews Drug Discovery, 2022, 21(8): 551.

[197] Nelson M R, Tipney H, Painter J L, et al. The support of human genetic evidence for approved drug indications[J]. Nature Genetics, 2015, 47(8): 856-860.

[198] Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data [J]. Nature, 2018, 562(7726): 203-209.

[199] Sun B B, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and health in the UK Biobank[J]. Nature, 2023, 622(7982): 329-338.

[200] Dewey F E, Gusarova V, Dunbar R L, et al. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease[J]. The New England Journal of Medicine, 2017, 377(3): 211-221.

[201] Ozen A, Comrie W A, Ardy R C, et al. CD55 deficiency, early-onset protein-losing enteropathy, and thrombosis[J]. The New England Journal of Medicine, 2017, 377(1): 52-61.

[202] O'Brien S G, Guilhot F, Larson R A, et al. Imatinib compared with interferon and low-dose cytarabine for

newly diagnosed chronic-phase chronic myeloid leukemia[J]. The New England Journal of Medicine, 2003, 348(11): 994-1004.

[203] 武士华. 药物基因组学研究发展和前景分析[J]. 军事医学科学院院刊, 2002, 26(3): 218-221.

[204] Sanseau P, Agarwal P, Barnes M R, et al. Use of genome-wide association studies for drug repositioning [J]. Nature Biotechnology, 2012, 30(4): 317-320.

[205] Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery [J]. Nature, 2014, 506(7488): 376-381.

[206] Abida W, Armenia J, Gopalan A, et al. Prospective genomic profiling of prostate cancer across disease states reveals germline and somatic alterations that may affect clinical decision making[J]. JCO Precision Oncology, 2017, 2017: PO.17.00029.

[207] Tardif J C, Kouz S, Waters D D, et al. Efficacy and safety of low-dose colchicine after myocardial infarction [J]. The New England Journal of Medicine, 2019, 381 (26): 2497-2505.

[208] Opstal T S J, Hoogeveen R M, Fiolet A T L, et al. Colchicine attenuates inflammation beyond the inflammasome in chronic coronary artery disease: A LoDoCo2 proteomic substudy[J]. Circulation, 2020, 142(20): 1996-1998.

[209] Yarmolinsky J, Bull C J, Vincent E E, et al. Association between genetically proxied inhibition of HMG-CoA reductase and epithelial ovarian cancer[J]. JAMA, 2020, 323(7): 646-655.

[210] Pushpakom S, Iorio F, Eyers P A, et al. Drug repurposing: Progress, challenges and recommendations[J]. Nature Reviews Drug Discovery, 2019, 18(1): 41-58.

[211] Iorio F, Rittman T, Ge H, et al. Transcriptional data: A new gateway to drug repositioning?[J]. Drug Discovery Today, 2013, 18(7/8): 350-357.

[212] Lamb J, Crawford E D, Peck D, et al. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease[J]. Science, 2006, 313(5795): 1929-1935.

[213] Subramanian A, Narayan R, Corsello S M, et al. A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles[J]. Cell, 2017, 171(6): 1437-1452.e17.

[214] Wei G, Twomey D, Lamb J, et al. Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance[J]. Cancer Cell, 2006, 10(4): 331-342.

[215] Laukkanen S, Veloso A, Yan C, et al. Therapeutic targeting of LCK tyrosine kinase and mTOR signaling in T-cell acute lymphoblastic leukemia[J]. Blood, 2022, 140(17): 1891-1906.

[216] Blackburn J S, Liu S L, Wilder J L, et al. Clonal evolution enhances leukemia-propagating cell frequency in T cell acute lymphoblastic leukemia through Akt/mTORC1 pathway activation[J]. Cancer Cell, 2014, 25 (3): 366-378.

[217] Pham T H, Qiu Y, Zeng J C, et al. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing[J]. Nature Machine Intelligence, 2021, 3(3): 247-257.

[218] Schork N J. Personalized medicine: Time for one-person trials[J]. Nature, 2015, 520(7549): 609-611.

[219] Pirmohamed M, James S, Meakin S, et al. Adverse drug reactions as cause of admission to hospital: Prospective analysis of 18 820 patients[J]. BMJ, 2004, 329 (7456): 15-19.

[220] Iorio F, Knijnenburg T A, Vis D J, et al. A landscape of pharmacogenomic interactions in cancer[J]. Cell, 2016, 166(3): 740-754.

[221] Ghandi M, Huang F W, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia[J]. Nature, 2019, 569(7757): 503-508.

[222] Yingtaweesittikul H, Wu J X, Mongia A, et al. CREAM-MIST: An integrative probabilistic database for cancer drug response prediction[J]. Nucleic Acids Research, 2023, 51(D1): D1242-D1248.

[223] Virani S S, Alonso A, Aparicio H J, et al. Heart disease and stroke statistics-2021 update: A report from the American heart association[J]. Circulation, 2021, 143(8): e254-e743.

[224] Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients[J]. BMJ-British Medical Journal, 2002, 324(7329): 71-86.

[225] Diener H C, Bogousslavsky J, Brass L M, et al. Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischaemic attack in high-risk patients (MATCH): Randomised, double-blind, placebo-controlled trial[J]. Lancet, 2004, 364 (9431): 331-337.

[226] Bhatt D L, Fox K A A, Hacke W, et al. Clopidogrel and aspirin versus aspirin alone for the prevention of

atherothrombotic events[J]. The New England Journal of Medicine, 2006, 354(16): 1706-1717.
[227] Johnston S C, Easton J D, Farrant M, et al. Clopidogrel and aspirin in acute ischemic stroke and high-risk TIA [J]. The New England Journal of Medicine, 2018, 379 (3): 215-225.
[228] Wang Y J, Meng X, Wang A X, et al. Ticagrelor versus clopidogrel in CYP2C19 loss-of-function carriers with stroke or TIA[J]. The New England Journal of Medicine, 2021, 385(27): 2520-2530.
[229] Villiger L, Joung J, Koblan L, et al. CRISPR technologies for genome, epigenome and transcriptome editing [J]. Nature Reviews Molecular Cell Biology, 2024, 25 (6): 464-487.
[230] Nissen S E, Wolski K, Watts G F, et al. Single ascending and multiple-dose trial of zerlasiran, a short interfering RNA targeting lipoprotein(a): A randomized clinical trial[J]. JAMA, 2024, 331(18): 1534-1543.
[231] 张思思, 陈旭, 陈婷婷, 等. GSA-Human: 人类遗传资源数据管理的公共系统[J]. 遗传, 2021, 43(10): 988-993.
[232] Price W N, Cohen I G. Privacy in the age of medical big data[J]. Nature Medicine, 2019, 25(1): 37-43.
[233] Chen G B, Liu S Y, Zhang L, et al. Building and sharing medical cohorts for research[J]. Innovation (Cambridge (Mass)), 2024, 5(3): 100623.

Multi-omics big data and medical advancements

LIU Siyang[1], LIN Xingchen[1], CHENG Si[2], WANG Chaolong[3], LI Hao[2*]

1. School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen 518107, China
2. China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China
3. School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Abstract Advances in multi-omics technologies, cohort study design, data science, and machine learning are transforming evidence-based medicine, offering a promising outlook for the future of next-generation "deep" medicine. We hereby summarized the development trends in multi-omics experimental techniques, including genomics and epigenomics sequencing, transcriptomics and single-cell transcriptomics, proteomics, metabolomics, microbiomics, imaging, and biosensors. Furthermore, we introduced progress in big data analysis methods such as genome-wide association studies, interpretation of genome-wide association signals, polygenic risk scoring, Mendelian randomization, and artificial intelligence algorithms. Additionally, we discussed the clinical applications of these technologies in disease subtyping, diagnosis and prediction, drug development, and clinical trial design. Finally, we discussed the challenges faced and explored future directions in cohort study design, data management and sharing, and the enhancement of international collaboration.

Keywords multi-omics; big data; medical research; clinical applications

（责任编辑　王微）