·综 述·

朱云平 北京蛋白质组研究中心研究员，硕士生导师，生物信息学研究室 PI。中国人类蛋白质组组织理事，北京放射医学研究所学术委员会委员。研究方向为：人类蛋白质组的生物信息学、系统生物学。参与 863 计划、973 计划及国家自然科学基金课题的研究。负责国家 973 课题 1 项。建立了世界上最大的人类蛋白质组数据及数据管理平台，对人类肝脏蛋白质组进行了系统的生物信息研究，包括蛋白质鉴定、修饰、定位、相互作用网络、代谢通路等，正在进行肿瘤标志物发现的研究。发表 SCI 论文 20 余篇，主编、主审专著两部，参编多部。获国家科技进步成果三等奖 1 项，北京市科学技术进步一等奖 1 项，军队科技进步成果二等奖两项。

从蛋白质基因组学视角出发的精准肿瘤学

黄雨柔[1,2,3]，吴松锋[2,3]，舒坤贤[1]，朱云平[2,3]

1 重庆邮电大学 生物信息学院，重庆 400065
2 军事科学院军事医学研究院生命组学研究所 国家蛋白质科学中心 (北京)，北京 102206
3 北京蛋白质组研究中心，北京 102206

黄雨柔, 吴松锋, 舒坤贤, 朱云平. 从蛋白质基因组学视角出发的精准肿瘤学. 生物工程学报, 2022, 38(10): 3616-3627.
HUANG YR, WU SF, SHU KX, ZHU YP. Precision oncology from a proteogenomics perspective. Chin J Biotech, 2022, 38(10): 3616-3627.

摘 要：癌症是一种机制复杂的异质性疾病，需要有针对性的精准医疗策略。精准医疗的发展离不开基因组学的飞速发展，但基因组学在分子表型分析中具有一定的局限性，蛋白质基因组学应运而生。蛋白质基因组学是蛋白质组学和基因组学的融合学科。文中描述了基因组学分析的局限性，强调了蛋白质基因组学的重要性，旨在从蛋白质基因组的视角重新了解精准肿瘤学。此外，还简要介绍了蛋白质基因组学在精准肿瘤学中的应用，对相关的公共数据项目进行了描述，最后，提出了现阶段需要克服的困难。

关键词：癌症；蛋白质基因组学；精准肿瘤学；公共数据库

Received: July 7, 2022; **Accepted:** October 9, 2022; **Published online:** October 18, 2022
Supported by: National Key Research and Development Program of China (2021YFA1301603); Open Project Program of State Key Laboratory of Proteomics (SKLP-O2020005)
Corresponding author: ZHU Yunping. Tel: +86-10-61777058; E-mail: zhuyunping@ncpsb.org.cn
基金项目：国家重点研发计划 (2021YFA1301603)；蛋白质组学国家重点实验室开放课题 (SKLP-O2020005)

Precision oncology from a proteogenomics perspective

HUANG Yurou[1,2,3], WU Songfeng[2,3], SHU Kunxian[1], ZHU Yunping[2,3]

1 School of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2 National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 102206, China
3 Proteome Research Center, Beijing 102206, China

Abstract: Cancer is a heterogeneous disease with complex mechanisms that requires targeted precision medicine strategies. The growth of precision medicine is indispensable from the rapid development of genomics. However, genomics has certain limitations in molecular phenotype analysis, proteogenomics thus arose at the right time. Proteogenomics is the merging of proteomics and genomics. This review describes the limitations of genomic analysis and highlights the importance of proteogenomics to re-understand precision oncology from a proteogenomic perspective. In addition, the application of proteogenomics in precision oncology is briefly introduced, the related public data projects are described, and finally, the challenges that need to be addressed at this stage are proposed.

Keywords: cancer; proteogenomics; precision oncology; public database

全球范围内，癌症一直是难以攻坚的医学难题之一。2020 年，癌症导致了近 1 000 万人的死亡[1]，成为全球第二大常见死因。虽然这些年通过各个领域学者的共同努力，癌症的死亡率得到了一定的控制[2]，但国际癌症研究机构 (International Agency for Research on Cancer, IARC) 发布的 Globocan 2020 报告表明，全球癌症负担已上升至 1 930 万例，到 2040 年预计将上升至 3 030 万例。总体而言，全球癌症发病率和死亡率正在迅速增加，抗癌刻不容缓。

癌症的发病机制非常复杂，它涉及年龄、性别、遗传等多方面的因素。面对这一棘手的全球健康问题，传统的手术、化疗、放疗等方式取得一定的效果，但是复杂的癌症发病机制阻碍着传统治疗方法的发展。人们逐渐意识到没有两例癌症是完全相同的，寻找新的治疗方法即精准医疗成为发展的趋势。“精准医疗”一词首先由美国国家研究委员会提出，旨在试图通过知识网络构建疾病分类，并提出建立新的数据网络，将治疗过程中患者的临床数据和生物医学研究结合起来[3]。近年来，精准医疗飞速发展，特别是在肿瘤领域大放异彩。众所周知，肿瘤治疗效果差主要是因为肿瘤异质性[4]，而精准肿瘤学旨在解决肿瘤间和肿瘤内的异质性。精准肿瘤学的发展离不开飞速发展的基于下一代测序技术的基因组学和基于质谱的蛋白质组学。在精准癌症医学的愿景中，使用整合多组学研究癌症已成为发展趋势。如下详述，仅靠基因组学 (为了简化，基因组学将包括基因组学、表观遗传组学和转录组学) 的数据来分析癌症是很局限的，只有将蛋白质组学结合起来才能突破现有瓶颈。

1 蛋白质基因组学的发展

基因是生命体的源头，蛋白质是生命活动的承担者。从基因组学到蛋白质组学是从基因型到表型的复杂过程，且每个步骤都受到不同程度的调控 (图 1)。肿瘤基因组的复杂性正在迅速被揭示，但它们是如何被调节成功能性蛋白质组的仍不清楚[5]。

1.1 蛋白质基因组的概念

蛋白质基因组学是蛋白质组学和基因组学的融合学科，一般指利用蛋白质组数据结合基因组数据开展研究，补充和完善基因组注释。在肿瘤相关的蛋白质基因组研究中，以特定肿瘤基因组为指导对蛋白质丰度和修饰程度进行解释，并整合基因组和蛋白质组结果，以便更深入理解和预测癌症表型[6]。蛋白质基因组学于 2004 年首次提出[7]，最初用于描述蛋白质组数据改进基因组注释和蛋白质编码潜力表征的研究。随后的研究大部分集中在基因组注释或揭示蛋白质组的复杂性[8]。随着蛋白质组技术的进步以及研究的深入，研究人员发现蛋白质组具有和基因组不一样的特征，能更直接揭示肿瘤发病机制[9]。蛋白质基因组学在癌症应用研究中的基本流程如图 2 所示，通过整合多组学分析可以更好地阐述癌症的发病机制以及相应的生物学过程，从而进一步提高癌症治疗和诊断的能力。

1.2 基因组学的局限性

过往的研究已经确定肿瘤的突变谱可以用于癌症患者群体治疗[10]，但也有相关研究证明基于突变谱的治疗策略具有一定的局限

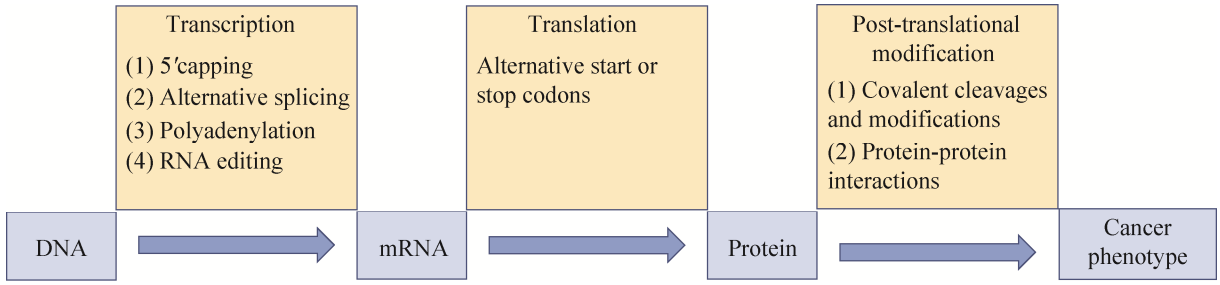

图 1 基因型到表型的复杂过程

Figure 1 The complex process from genotype to phenotype.

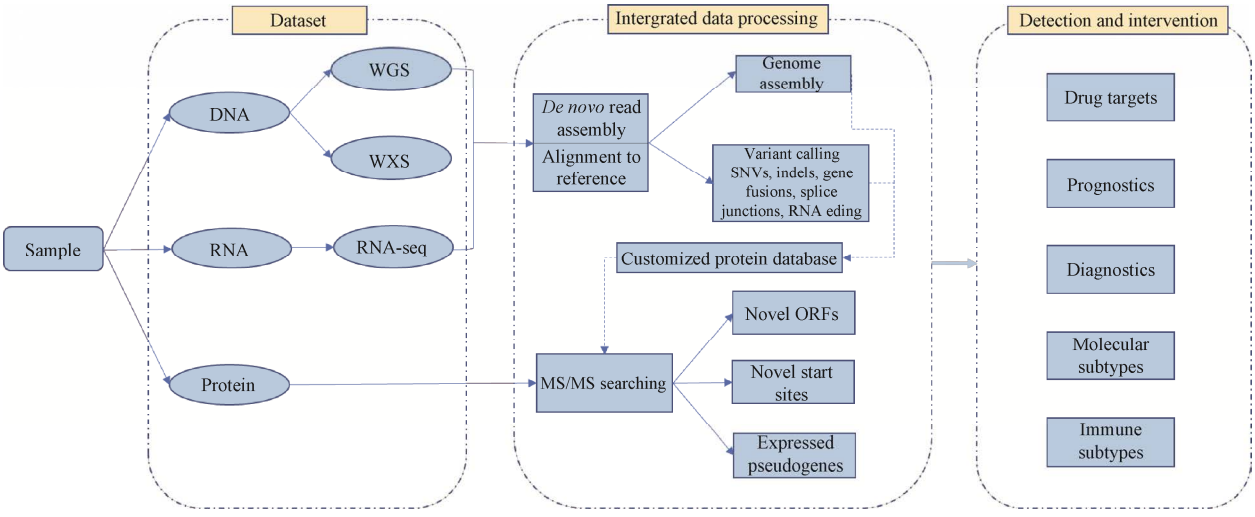

图 2 蛋白质基因组学在癌症中应用的基本流程

Figure 2 Workflow for the application of proteogenomics in cancer research. WGS: whole genome sequencing; WXS: whole-exome sequencing; SNVs: single nucleotide variants.

性[11-12]。一般可采用联合转录组的方法来选择特定的癌症治疗方法[13]，但这并不能可靠地预测蛋白质水平的变化，对于疾病诊断及治疗往往不够全面。此外，靶向药物的失效也是难题之一，基于基因检测的靶向治疗药物的受益人群低于 10%[14]。甚至，具有相同基因突变的患者在面对相同的靶向药物处理也会产生不同的效果。例如，在黑色素瘤或结直肠癌患者中，两种肿瘤类型具有相同的鼠类肉瘤滤过性毒菌致癌同源体 B (v-raf murine sarcoma viral oncogene homolog B, BRAF) 基因突变。然而，对黑色素瘤有效的靶向治疗可能无法帮助结直肠癌患者[15-16]。此外，蛋白质与转录本之间的相关性较弱[17]。整合蛋白质基因组分析显示，对于许多基因而言，肿瘤中 mRNAs 丰度的变化与其匹配蛋白丰度的变化相关性很差。尽管一些基因具有高水平的 mRNA-蛋白质相关性，但 30%–70%的基因在统计上无显著的正相关性[18-19]。大部分情况下，转录水平不足以预测蛋白质水平，尤其是蛋白质翻译后修饰，如蛋白质磷酸化[18,20]。例如，在结直肠研究中[18]，RNA 丰度不能准确地预测蛋白质丰度，大多数 DNA 局部扩增并没有导致蛋白质水平的相应升高。

综上所述，尽管 RNA-seq 数据的深度仍然大于蛋白质组数据的深度[21]，但仅依靠基因组学进行的诊断和治疗不足以应付癌症的复杂性，癌症在蛋白质水平的研究可以大大缩小癌症基因型和癌症表型之间的差距。在癌症的诊断和治疗方面，蛋白质组学的研究有可能提供新的思路和视角。但需要注意的是蛋白质组学也并不能独立解释癌症，只有将蛋白质组学与基因组学相结合时，效用才能最大化。

1.3 蛋白质基因组的过去、现在和未来

早期研究中，蛋白质组学技术整体灵敏度较低，在一定程度上阻碍了蛋白质基因组学工作的开展。随着蛋白质组技术的不断成熟这一困难逐渐被克服，质谱仪器和分析方法的进步极大地提高了蛋白质组的覆盖率，减少了偏差，提高了精确度[22]。

近年来，多个大型蛋白质组研究项目的蓬勃发展极大加快了蛋白质基因组学研究的步伐。2003 年 10 月，“人类肝脏蛋白质组计划”[23]由我国启动，先后 11 个国家参与其中。2014 年，贺福初院士领导启动“中国人蛋白质组计划”(Chinese Human Proteome Project, CNHPP)，对多种肿瘤进行了蛋白质层面的分析。国际上，2015 年初，奥巴马提出精准医学计划[24]。2016 年初，癌症登月计划[25]使精准医学计划开始走进大众视野，精准医学很快成为医学研究热点。癌症基因组图谱 (the cancer genome atlas, TCGA) 以及临床蛋白质组肿瘤分析协会 (Clinical Proteomic Tumor Analysis Consortium, CPTAC) 的推进无疑也加速了精准肿瘤学的发展，发布了多种人类肿瘤的蛋白基因组学研究，利用癌症特征肽段来进行诊断或治疗[26-32]。综上所述，质谱技术的成熟和蛋白质数据的日益丰富为蛋白质基因组的发展奠定坚实的基础。

2 蛋白质基因组学在精准肿瘤学中的应用

蛋白质基因组学弥补了单一基因组的局限性，促进了精准肿瘤学的快速发展，为癌症治疗带来了新的突破。蛋白质基因组学可以将基因组、转录组、蛋白质组、修饰蛋白质组的数据整合，从多个层面重新定义癌症的分型、挖掘潜在的治疗靶点，以进行更精准的治疗方法选择和用药指导。蛋白质基因组学已广泛应用于人类肿瘤的分析，通过回顾整理近几年前人的研究工作，总结蛋白质基因组学在精准肿瘤学领域中有如下应用。

2.1 阐释肿瘤生物学过程

在乳腺癌的研究中[33]，该研究利用蛋白质基因组学方法，通过对 122 例原发乳腺癌治疗前的肿瘤标本进行多组学分析，特别是对蛋白质磷酸化和乙酰化等蛋白质翻译后化学修饰进行了全面分析。磷酸蛋白质组学分析揭示了肿瘤抑制因子丢失和靶向激酶之间的新关联。乙酰蛋白质组分析强调了参与 DNA 损伤反应的关键核蛋白的乙酰化，并揭示了细胞质、线粒体乙酰化和代谢之间的相互作用。这发现蛋白质基因组学能够更精准地分析乳腺癌的可靶向蛋白质信号通路和生物学特征，突显了其对乳腺癌临床研究的巨大潜力。

在结肠癌的研究中[18]，通过对 95 个经基因组注释的癌症基因组图谱 (TCGA) 结直肠癌样本进行蛋白质组学表征，对候选驱动基因进行了优先排序，筛选出了潜在的位于 20 号染色体长臂的候选基因，包括肝细胞核因子 (hepatocyte nuclear factor 4 alpha, HNF4A) 基因、线粒体外膜转位酶 (translocase of outer mitochondrial membrane 34, TOMM34) 基因和酪氨酸蛋白激酶 (SRC proto-oncogene, non-receptor tyrosine kinase, SRC) 基因。综合蛋白质组学分析可以对人类癌症中的基因组异常进行功能性解释，并为理解癌症生物学提供了新的角度。

2.2 检测生物标志物

癌症标志物是指特异存在于恶性肿瘤细胞，或因机体对肿瘤刺激而产生的物质，它能反应肿瘤的发生、发展，并对治疗疗效起到监视作用。在结肠癌的研究中[34]，研究人员采用多组学的研究分析方法，全面剖析了结直肠癌相关生物标志物，发现了基于基因组层面不能检测到的新的生物标志物。他们的研究结果可能为结肠癌相关研究提供新的研究思路。

此外，在头颈部鳞状细胞癌 (head and neck squamous cell carcinomas, HNSCCs) 的研究中[26]，通过对 108 例人乳头瘤病毒 (human papilloma virus, HPV) 阴性 HNSCCs 患者和 66 个匹配的癌旁组织样本的蛋白质基因组谱分析表明，符合美国食品药品监督管理局 (Food and Drug Administration, FDA) 标准的用于 HNSCC 治疗的研究药物可从候选的生物标志物中进行选择。

2.3 寻找治疗靶点

目前，除了传统的外科化疗和放射治疗外，还有另一种有效且方便的癌症治疗方法，即靶向治疗。肿瘤基因组的表征能让人们了解到癌症中的体细胞突变，以便选择合适的治疗方法。但是，如前所述，癌症基因组与蛋白质组的相关性较弱，仅凭基因组层面的分析不足以筛选出有效的靶点。在胰腺导管腺癌的研究中[32]，研究人员通过对胰腺癌及其正常癌旁组织进行多组学分析研究，确定了潜在的胰腺导管腺癌治疗靶点和早期诊断标志物。结果显示，多组学分析能够更精确地筛选出癌症靶点。

在子宫内膜癌 (uterine corpus endometrial carcinoma, UCEC) 的研究中[30]，对浆液性肿瘤和子宫内膜样肿瘤的比较显示，在浆液性肿瘤中肿瘤蛋白 p53 结合蛋白 1 (tumor protein p53 binding protein 1, TP53BP1-S1763) 和检查点激酶 2 (checkpoint kinase 2, CHEK2-S163) 磷酸化水平升高，与更高的 DNA 损伤反应相关，这是癌症治疗中一个有吸引力的靶点。

2.4 助力癌症分型

癌症是病理复杂的疾病。每一种亚型的特征、最优治疗方式、预后等都不同。基于蛋白质基因组学的疾病分子分型研究是目前肿瘤精准医疗领域的热点。在乙型肝炎 (hepatitis B virus, HBV) 相关性肝癌的研究中[35]，研究人员

通过对 159 例乙肝病毒阳性的肝癌和癌旁样本进行多组学的研究分析，绘制出了目前全球最大规模的全景式肝癌队列的蛋白质基因组学图谱。研究发现，肝癌患者的蛋白质组数据整体上可以分为 3 类亚型，它们分别是代谢驱动型、微环境失调型和增殖驱动型。这 3 类亚型患者的预后具有极其显著的差别，这将为肝癌的临床预后判别以及肝癌的个性化治疗起到重要的指导作用。

在前列腺癌的研究中，Sinha 等[36]对 76 例前列腺癌患者的肿瘤组织样本进行了基因组学、表观基因组学、转录组学和蛋白质组学分析。蛋白质组分析确定了 5 种蛋白质组亚型，它们在很大程度上独立于先前报道的基因组亚型。

2.5 助力免疫治疗

癌症的治疗方法从化疗、靶向治疗逐渐转向免疫疗法发展，癌症疫苗是免疫治疗的一大热门方向。在结肠癌的研究中[34]，与结肠癌相关的蛋白编码基因与已知的癌症基因几乎没有重叠，这有助于开发个性化疫苗以及其他治疗方法。

上述 UCEC 研究中[30]免疫景观的蛋白质基因组学研究确定了 8 种复发性过度表达的癌-睾丸抗原，并在 49%的肿瘤中发现了推定的新抗原。具有高肿瘤突变负荷 (tumor mutation burden, TMB) 的 UCEC 细胞已经发展出几种机制来抑制抗原处理和呈递，这可能导致免疫逃避。这些机制包括人类酪氨酸激酶蛋白 (janus kinase 1, JAK1) 和/或信号传导转录激活因子 1 (signal transducer and activator of transcription 1, STAT1) 突变，以及抗原肽转运蛋白 1 (transporter 1, TAP1)、转运蛋白 2 (transporter 2, TAP2) 和 TAP2 结合蛋白(TAP2 binding protein, TAP2BP) 的蛋白水平降低。

免疫蛋白质基因组学分析显示 HNSCC 肿瘤中存在广泛的免疫细胞浸润水平[26]。这些肿瘤中多个免疫检查点蛋白的一致上调可能解释了抗程序性死亡受体 1 (programmed death 1, PD1) 单一疗法的中等应答率[37]，并为研究高水平免疫细胞浸润肿瘤中的联合检查点阻断提供了理论基础。

除以上提及的癌症，在其他癌症中蛋白质基因组视角的精准肿瘤学也发挥着重要作用(表 1)。在实践中，多组学的意义在于它可以从不同维度更全面地探索复杂的癌症机制，为每一位癌症病人提供更全面的分子层面的信息，从而为癌症的诊断及治疗提供新思路、新方法。综上所述，蛋白质基因组学在精准肿瘤学领域能够在生物标志物、治疗靶点、癌症分型、疫苗研发、临床试验、药物开发等方面促进发展。

3 蛋白质基因组学相关项目

此外，蛋白质基因组学的发展离不开数据的积累，如今已经有许多储存数据的项目 (表 2)。在癌症方面，美国国家癌症研究院 (National Cancer Institute, NCI) 支持的项目，如 TCGA 和 CPTAC，已经生成了大量数据集，积累了 PB 量级的数据。此外，在蛋白质组数据收集方面，国际上成立了 ProteomeXchange 联盟[46]，目前共有 6 个成员 (iProX[47]为国内成员)。接下来将简单介绍几个相关的数据项目。

3.1 TCGA

TCGA 成立于 2006 年，并于 2010 年扩展[48]，它对 33 种癌症的肿瘤样本及其匹配的正常样本进行了分子表征[49]，产生了可供癌症研究领域继续挖掘的大量数据[50]。收集的数据类型包括 DNA 拷贝数阵列、DNA 甲基化、外显子组和全基因组测序、mRNA 阵列、microRNA 测序和反相蛋白质阵列，共计约 2.5 PB 的数据[51]。TCGA 对分子层面的表征为理解复杂的癌症机制铺平了新的道路[49]。近年来，许多癌症相关的研究相

表 1 蛋白质基因组学在肿瘤中的运用相关部分论文

Table 1 Papers related to the application of proteomics in tumors

Cancer type	Conceptual insights	Year	References
ALL	Identification of potential driver genes and/or pathway	2019	[38]
Brain cancer	Identification of potential driver genes and/or pathway	2021	[28]
	Identification of putative targets and biomarkers	2020	[39]
	Insights into mechanisms of drug resistance	2020	[40]
Breast cancer	Identification of potential driver genes and/or pathways	2016	[20]
	Identification of potential driver genes and/or pathway	2020	[33]
	Refinement and extended characterization of cancer subtypes		
ccRCC	Refinement and extended characterization of cancer subtypes	2019	[27]
CRC	Identification of putative targets and biomarkers	2019	[34]
	Identification of potential driver genes and/or pathways	2014	[18]
EnC	Insights into mechanisms of immune evasion	2020	[30]
Gastric cancer	Identification of putative targets and biomarker	2019	[41]
	Refinement and extended characterization of cancer subtypes		
HNSCC	Insights into mechanisms of immune evasion and drug resistance	2021	[26]
HCC	Identification of potential biomarkers for patient selection	2019	[35]
Lung cancer	Refinement and extended characterization of cancer subtypes	2022	[42]
	Insights into mechanisms of immune evasion		
	Identification of putative targets and biomarkers		
	Refinement and extended characterization of cancer subtypes	2020	[43]
	Identification of putative targets and biomarkers		
	Refinement and extended characterization of cancer subtypes	2021	[31]
	Insights into mechanisms of immune evasion		
	Identification of putative targets and biomarkers		
	Identification of putative targets and biomarkers	2020	[44]
OvC	Identification of putative targets and biomarkers	2020	[45]
	Identification of potential driver genes and/or pathways	2016	[19]
PDA	Identification of putative targets and biomarker	2021	[32]
	Refinement and extended characterization of cancer subtypes		

ALL: acute lymphoblastic leukaemia; ccRCC: clear cell renal cell carcinoma; CRC: colorectal cancer; EnC: endometrial carcinoma; HCC: hepatocellular carcinoma; HNSCC: head and neck squamous cell carcinoma; OvC: ovarian cancer; PDA: pancreatic ductal adenocarcinoma.

表 2 与蛋白质基因组相关的数据项目信息

Table 2 Data item information related to proteogenomics

Projects	Responsible agency	URL
The cancer genome atlas (TCGA)	National Cancer Institute National Human Genome Research Institute	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
Therapeutically applicable research to generate effective treatments (TARGET)	NCI Office of Cancer Genomics	https://ocg.cancer.gov/programs/target
Clinical proteomic tumor analysis consortium (CPTAC)	NCI Office of Cancer Clinical Proteomics Research	https://proteomics.cancer.gov/programs/cptac
Genomic data commons (GDC)	NCI Center for Cancer Genomics	https://portal.gdc.cancer.gov
Proteomic data commons (PDC)	NCI Center for Cancer Genomics	https://pdc.cancer.gov
iProX	National Center for Protein Sciences (Beijing)	https://www.iprox.org
NODE (national omics data encyclopedia)	Shanghai Institute of Nutrition and Health	https://biosino.org/

继利用 TCGA 数据库进行数据再挖掘以发现新的癌症分子机制以及生物标志物[52-53]。此外，它还通过分子将肿瘤进行分型，从而改变了诊断思路，使患者的治疗比以前更加精确。这些变化可能为开发成功概率更高的临床试验提供新思路。

3.2 CPTAC

CPTAC 于 2006 年由美国食品药品监督管理局 (Food and Drug Administration, FDA) 和美国临床化学协会 (American Association for Clinical Chemistry, AACC) 合作启动[18-20]，它的目标是在蛋白质水平上分析癌症，将基因型与蛋白质型联系起来，目的是了解癌症表型的基础。CPTAC 第一阶段包括技术质量保证研究[54]。作为对 TCGA 研究的补充，CPTAC 第二阶段包括对 TCGA 乳腺、卵巢和结直肠样本进行基于质谱的蛋白质组分析[18-20]。CPTAC 第三阶段是对从其他癌症类型中收集的组织进行蛋白质基因组分析[27,55]。此外，为了支持精准肿瘤学，CPTAC 第三阶段已经建立了蛋白质基因组翻译研究中心，该中心将研究癌症治疗对单个肿瘤样本的疗效，以生成预测模型。CPTAC 目前总共有约 26 TB 的数据，在 CPTACIII 完成后，这一数字预计将增加 4 倍，达到约 66 TB 的数据。CPTAC 关于结直肠癌、乳腺癌和卵巢癌的 3 份报告为癌症研究界提供了宝贵资源，通过指出这些癌症的哪些基因组和转录组特征在蛋白质水平上被检测到，以及通过提供翻译后修饰 (特别是磷酸化和乙酰化) 与 DNA 修复相关的蛋白质功能活性的实质性影响的新见解，为癌症的发病机制探索带来了新的突破[56]。CPTAC 开创性地利用蛋白质基因组学对肿瘤进行临床表征，结合多种组学方法可以更全面地理解癌症并可能带来新的治疗方法。

4 面对的困难与展望

蛋白质基因组研究在精准医学计划[57]中发挥了重要的作用，已经逐步加大了识别和理解癌症的能力。蛋白质基因组学弥补了临床采用单一组学的传统方法的局限，但它也面临着一些挑战。为了更好地诠释和理解癌症机制，有必要克服这些挑战。

4.1 数据质量参差不齐

蛋白质基因组学的一个主要挑战是在公共领域缺乏足够数量的优质蛋白质组数据，部分原因是早期蛋白质组学界普遍存在的“数据囤积”心态。此外，质谱仪的更新换代以及实验流程的不同，导致产出数据的深度以及精度都不同。不同年代、不同团队或者不同技术平台产出质谱数据的质量可能相差甚远。目前，蛋白质组学领域的数据共享已明显转向更加开放[58]。因此，越来越多的蛋白质组数据集可以在公共数据库[59]中获得。这意味着蛋白质基因组组学的瓶颈已经从数据积累转向海量数据的分析处理。

4.2 数据分析存在挑战

数据的井喷式发展带来了数据分析的挑战。低丰度肽的难以检测是蛋白质组数据分析瓶颈的一部分。低丰度肽的难以检测主要有以下两个原因：(1) 蛋白质丰度的大动态范围和缺乏对所有母体离子的选择[60]。(2) 通常存在一个明确的 mRNA 丰度阈值，低于该阈值，无法识别肽[61]。但随着数据独立采集 (data independent acquisition, DIA) 技术[62]的发展，使低丰度肽段的检测成为可能。

将蛋白质组学数据与其他类型的基因组学或代谢组学数据整合起来也是数据分析瓶颈的一部分。这种整合对于理解癌细胞的表型以及开发多模式生物标记物具有重大意义。数据分析困难就将挑战转移到对生物信息学工具的需求上。

4.3 生物信息学工具需求巨大

随着基于下一代测序技术和质谱技术的蛋白质基因组学数据集的规模和复杂性的增加，迫切需要高效且易于使用的工具来进行生物信息学分析。在蛋白质基因组工作流程中，原始核苷酸数据必须与现有的基因注释进行比对，并翻译成多肽序列。蛋白质组学数据集还需要复杂的工作流程，包括数据库搜索和统计结果的正确解释。如今已经开发了一些工具来帮助蛋白质基因组数据库的构建[63-64]和基因组上肽的可视化或与基因模型的对比[65-66]。目前需要不断努力来开发用于分析基于蛋白质基因组学方法的生物信息学工具。常见的联合分析以及可视化工具如表 3 所示。在未来，笔者认为，基因组测序项目将从一开始就包括蛋白质组学分析，从而基于蛋白质组学数据对蛋白质编码基因进行注释。

5 总结

本文中，我们从蛋白质基因组学的全新视角围绕精准肿瘤学进行了详细阐述。通过对基因组学以及蛋白质组学的回顾，我们讨论总结了单一基因组学的局限性并对蛋白质基因组学进行了展望。笔者认为，想要进一步了解癌症复杂的发病机制就必须将基因组学和蛋白质组学联合起来分析，这样才能够更好地与临床结合起来，帮助癌症的诊断以及治疗。随着公共数据库的不断完善，蛋白质基因组发展迅猛。毫无疑问，基因组学和蛋白质组学之间的协同关系将继续发展，并将成为未来几十年人类癌症完整表征的关键。总之，各种迹象表明，蛋白质基因组学将会大大推动精准肿瘤领域的发展。

表 3 常见的联合分析以及可视化工具

Table 3 Common tools for association discovery and visualization

Projects	Introduction	References
LinkedOmics	Interactive visualization and analysis of multi-omics data from TCGA and CPTAC within and across cancer types	[67]
multiOmicsviz	Visualizing the effect of one omics data type on other omics data along the chromosome	[18]
iProFun	Characterization of functional consequences of copy number alterations (CNAs) and methylation alterations in tumors	[68]
NetsaM	Network seriation and modularization	[69]
tsNet	A statistical model for cell type-specific inference based on bulk tumor profiling data	[70]
Perseus	Bioinformatics platform for integrative analysis of proteomic data	[71-72]

REFERENCES

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2021, 71(3): 209-249.

[2] Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2021. CA Cancer J Clin, 2021, 71(1): 7-33.

[3] National Research Council (US) Committee on a Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington (DC): National Academies Press (US), 2011: 4-6.

[4] Network CGAR, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet, 2013, 45(10): 1113-1120.

[5] Alfaro JA, Sinha A, Kislinger T, et al. Onco-proteogenomics: cancer proteomics joins forces with genomics. Nat Methods, 2014, 11(11): 1107-1113.

[6] Renuse S, Chaerkady R, Pandey A. Proteogenomics. Proteomics, 2011, 11(4): 620-630.

[7] Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform

genome annotation. Proteomics, 2004, 4(1): 59-77.
[8] Menschaert G, Fenyö D. Proteogenomics from a bioinformatics angle: a growing field. Mass Spectrom Rev, 2017, 36(5): 584-599.
[9] Wilhelm M, Schlegl J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature, 2014, 509(7502): 582-7.
[10] Flaherty KT, Gray R, Chen A, et al. The molecular analysis for therapy choice (NCI-MATCH) trial: lessons for genomic trial design. J Natl Cancer Inst, 2020, 112(10): 1021-1029.
[11] Le Tourneau C, Delord JP, Gonçalves A, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol, 2015, 16(13): 1324-1334.
[12] Saad E D, Paoletti X, Burzykowski T, et al. Precision medicine needs randomized clinical trials. Nat Rev Clin Oncol, 2017, 14(5): 317-323.
[13] Rodon J, Soria JC, Berger R, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med, 2019, 25(5):
[14] Marquart J, Chen EY, Prasad V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. JAMA Oncol, 2018, 4(8): 1093-1098.
[15] Hyman DM, Puzanov I, Subbiah V, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. N Engl J Med, 2015, 373(8): 726-736.
[16] Sclafani F, Gullo G, Sheahan K, et al. BRAF mutations in melanoma and colorectal cancer: a single oncogenic mutation with different tumour phenotypes and clinical implications. Crit Rev Oncol Hematol, 2013, 87(1): 55-68.
[17] Liu YS, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell, 2016, 165(3): 535-550.
[18] Zhang B, Wang J, Wang XJ, et al. Proteogenomic characterization of human colon and rectal cancer. Nature, 2014, 513(7518): 382-387.
[19] Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell, 2016, 166(3): 755-765.
[20] Mertins P, Mani D R, Ruggles K V, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature, 2016, 534(7605): 55-62.
[21] Mudge JM, Frankish A, Harrow J. Functional transcriptomics in the post-ENCODE era. Genome Res, 2013, 23(12): 1961-1973.
[22] Mann M, Kulak NA, Nagaraj N, et al. The coming age of complete, accurate, and ubiquitous proteomes. Mol Cell, 2013, 49(4): 583-590.
[23] 贺福初. 国家人类肝脏蛋白质组计划. 医学研究通讯, 2004(7): 2.
He FC. National human liver proteome project. Bull Med Res, 2004(7): 2 (in Chinese).
[24] Bahcall O. Precision medicine. Nature, 2015, 526(7573): 335.
[25] Obama's cancer moonshot. Nat Biotechnol, 2016, 34(2): 119.
[26] Huang C, Chen LJ, Savage SR, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell, 2021, 39(3): 361-379.e16.
[27] Clark DJ, Dhanasekaran SM, Petralia F, et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell, 2019, 179(4): 964-983.e31.
[28] Wang LB, Karpova A, Gritsenko MA, et al. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell, 2021, 39(4): 509-528.e20.
[29] Gillette MA, Satpathy S, Cao S, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell, 2020, 182(1): 200-225.e35.
[30] Dou YC, Kawaler EA, Zhou DC, et al. Proteogenomic characterization of endometrial carcinoma. Cell, 2020, 180(4): 729-748.e26.
[31] Satpathy S, Krug K, Jean Beltran PM, et al. A proteogenomic portrait of lung squamous cell carcinoma. Cell, 2021, 184(16): 4348-4371.e40.
[32] Cao LW, Huang C, Zhou DC, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell, 2021, 184(19): 5031-5052.e26.
[33] Krug K, Jaehnig EJ, Satpathy S, et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell, 2020, 183(5): 1436-1456.e31.
[34] Vasaikar S, Huang C, Wang XJ, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell, 2019, 177(4): 1035-1049.e19.
[35] Gao Q, Zhu HW, Dong LQ, et al. Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell, 2019, 179(2): 561-

577.e22.
[36] Sinha A, Huang V, Livingstone J, et al. The proteogenomic landscape of curable prostate cancer. Cancer Cell, 2019, 35(3): 414-427.e6.
[37] Seiwert TY, Burtness B, Mehra R, et al. Safety and clinical activity of pembrolizumab for treatment of recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-012): an open-label, multicentre, phase 1b trial. Lancet Oncol, 2016, 17(7): 956-965.
[38] Yang MJ, Vesterlund M, Siavelis I, et al. Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. Nat Commun, 2019, 10(1): 1519.
[39] Petralia F, Tignor N, Reva B, et al. Integrated proteogenomic characterization across major histological types of pediatric brain cancer. Cell, 2020, 183(7): 1962-1985.e31.
[40] Satpathy S, Jaehnig EJ, Krug K, et al. Microscaled proteogenomic methods for precision oncology. Nat Commun, 2020, 11(1): 532.
[41] Mun DG, Bhin J, Kim S, et al. Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell, 2019, 35(1): 111-124.e10.
[42] Dong LQ, Lu DY, Chen R, et al. Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma. Cancer Cell, 2022, 40(1): 70-87.e15.
[43] Xu JY, Zhang CC, Wang X, et al. Integrative proteomic characterization of human lung adenocarcinoma. Cell, 2020, 182(1): 245-261.e17.
[44] Chen YJ, Roumeliotis TI, Chang YH, et al. Proteogenomics of non-smoking lung cancer in east Asia delineates molecular signatures of pathogenesis and progression. Cell, 2020, 182(1): 226-244.e17.
[45] McDermott JE, Arshad OA, Petyuk VA, et al. Proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability. Cell Rep Med, 2020, 1(1): 100004.
[46] Deutsch EW, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. Nucleic Acids Res, 2020, 48(D1): D1145-D1152.
[47] Ma J, Chen T, Wu SF, et al. iProX: an integrated proteome resource. Nucleic Acids Res, 2018, 47(D1): D1211-D1217.
[48] Hanauer D, Rhodes D, Sinha-Kumar C, et al. Bioinformatics approaches in the study of cancer. Curr Mol Med, 2007, 7(1): 133-141.
[49] Blum A, Wang P, Zenklusen JC. SnapShot: TCGA-analyzed tumors. Cell, 2018, 173(2): 530.
[50] Network CGAR. Comprehensive molecular characterization of gastric adenocarcinoma. Nature, 2014, 513(7517): 202-209.
[51] Hinkson IV, Davidsen TM, Klemm JD, et al. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. Front Cell Dev Biol, 2017, 5: 83.
[52] Donehower LA, Soussi T, Korkut A, et al. Integrated analysis of TP53 gene and pathway alterations in the cancer genome atlas. Cell Rep, 2019, 28(5): 1370-1384.e5.
[53] Danaher P, Warren S, Lu RZ, et al. Pan-cancer adaptive immune resistance as defined by the tumor inflammation signature (TIS): results from the cancer genome atlas (TCGA). J Immunother Cancer, 2018, 6(1): 63.
[54] Paulovich AG, Billheimer D, Ham AJL, et al. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. Mol Cell Proteom, 2010, 9(2): 242-254.
[55] Clark DJ, Hu YW, Bocik W, et al. Evaluation of NCI-7 cell line panel as a reference material for clinical proteomics. J Proteome Res, 2018, 17(6): 2205-2215.
[56] Zhang B, Whiteaker J R, Hoofnagle A N, et al. Clinical potential of mass spectrometry-based proteogenomics. Nat Rev Clin Oncol, 2019, 16(4): 256-268.
[57] Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med, 2015, 372(9): 793-795.
[58] Rodriguez H, Snyder M, Uhlén M, et al. Recommendations from the 2008 international summit on proteomics data release and sharing policy: the Amsterdam principles. J Proteome Res, 2009, 8(7): 3689-3692.
[59] Vizcaíno JA, Deutsch EW, Wang R, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol, 2014, 32(3): 223-226.
[60] Liu HB, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem, 2004, 76(14):

4193-4201.

[61] Wang XJ, Slebos RJC, Wang D, et al. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res, 2012, 11(2): 1009-1017.

[62] Ludwig C, Gillet L, Rosenberger G, et al. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol, 2018, 14(8): e8126.

[63] Sheynkman GM, Johnson JE, Jagtap PD, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics, 2014, 15(1): 703.

[64] Wen B, Xu SH, Sheynkman GM, et al. sapFinder: an R/bioconductor package for detection of variant peptides in shotgun proteomics experiments. Bioinformatics, 2014, 30(21): 3136-3138.

[65] Pang CNI, Tay AP, Aya C, et al. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. J Proteome Res, 2014, 13(1): 84-98.

[66] Nagaraj SH, Waddell N, Madugundu AK, et al. PGTools: a software suite for proteogenomic data analysis and visualization. J Proteome Res, 2015, 14(5): 2255-2266.

[67] Vasaikar SV, Straub P, Wang J, et al. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res, 2018, 46(D1): D956-D963.

[68] Song XY, Ji JY, Gleason KJ, et al. Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. Mol Cell Proteomics, 2019, 18(8 suppl 1): S52-S65.

[69] Shi ZA, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. Nat Methods, 2013, 10(7): 597-598.

[70] Petralia F, Wang L, Peng J, et al. A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. Bioinformatics, 2018, 34(13): i528-i536.

[71] Tyanova S, Temu T, Sinitcyn P, et al. The perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods, 2016, 13(9): 731-40.

[72] Tyanova S, Cox J. Perseus: a bioinformatics platform for integrative analysis of proteomics data in cancer research. Methods Mol Biol, 2018, 1711: 133-148.

(本文责编　郝丽芳)