

肿瘤基因组数据挖掘及其应用*

郑 思 侯 丽 李 姣

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 简要介绍肿瘤基因组数据的积累和基因组数据挖掘情况; 结合现有的高通量技术与数据挖掘策略, 以白血病为例, 提出一套基于基因组数据研究的肿瘤精准诊断治疗和风险预测策略; 最后总结肿瘤基因组数据挖掘研究面临的机遇与挑战。

〔关键词〕 基因组; 肿瘤; 肿瘤基因组数据; 数据挖掘

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.01.015

Data Mining and Application of Cancer Genome ZHENG Si, HOU Li, LI Jiao, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China

〔Abstract〕 The paper briefly introduces the situations relevant to data accumulation of cancer genome and mining of genome data, puts forward a set of accurate genome data research based cancer diagnosis and treatment and risk prediction strategies by combining the current high-throughput technology and data mining strategy taking leukemia as an example, and finally summarizes the opportunities and challenges faced by mining and research of cancer genome data.

〔Keywords〕 Genome; Cancer; Cancer genomic data; Data mining

1 引言

生命体遗传信息的获得对于生命科学领域的研究具有重要意义。1953 年沃森和克里克发现了 DNA 双螺旋结构以来, 人们对 DNA 有了进一步的了解

并开始进行更深入的探索。20 世纪 70 年代 DNA 测序技术的诞生, 为生命科学研究提供了一种新方法, 彻底改变了人们对人类生物学的认知方式^[1-2]。21 世纪初人类基因组计划 (Human Genome project, HGP) 破译了人类基因信息, 是人类第 1 次在分子水平上全面认识自我, 巩固了基因组信息在生命科学领域研究中的地位。后来, 随着高通量测序技术的产生和发展, 人们能够低成本、高效率地对几万到几百万条 DNA 分子序列进行一次性测定, 对未来的临床医学的研究产生了重大影响^[3-7]。高通量数据的不断积累, 产生了一系列数据收集、存储与管理平台, 为生命科学研究提供了宝贵的资源。例如, 基因组信息可用于分析并制订与人类健康管理、疾病风险预测相关的标准, 进行人群的患病风险预测^[8]。

在不同种类的疾病中, 肿瘤是一种死亡率较高的亚类, 现有的大概 200 种不同类型的肿瘤以及更

〔修回日期〕 2016-10-14

〔作者简介〕 郑思, 实习研究员, 发表论文 1 篇; 通讯作者: 李姣, 副研究员。

〔基金项目〕 中央级公益性科研院所基本科研业务费资助课题“医学信息学理论方法在转化医学研究中的应用”(项目编号: 13R0101); 中国工程科技知识中心建设项目“医药卫生专业知识服务系统”(项目编号: CKCEST-2015-2-3)。

多的子类,都存在遗传信息的变异,这些变异会导致细胞不受控制的发展。识别肿瘤完整的遗传信息变化,以及理解这些变化是如何交互驱动疾病过程的,可为癌症诊断治疗提供新的思路与解决方案^[9]。目前国际上已经开展了很多基于癌症基因组数据挖掘进行疾病内部遗传机制研究的项目。例如,美国开展了肿瘤基因组图谱计划(The Cancer Genome Atlas, TCGA),其目标就是创建一个与肿瘤发生发展相关的基因目录,促进肿瘤的诊断与治疗^[10]。肿瘤基因组数据挖掘不仅能揭示肿瘤发生发展的理论基础,而且也将逐渐应用于肿瘤的临床诊断和治疗当中^[11]。然而,如何将这些理论研究成果转化为临床实践是一个需要探索的问题,依赖于更加精准的数据挖掘预测模型和高通量技术手段。

2 肿瘤基因组数据积累

2.1 概述

全球范围内的肿瘤基因组计划加快了肿瘤基因组数据积累,其中,TCGA 由美国国立卫生研究院(National Institute of Health, NIH)于 2006 年开始建立,目标是获得一个综合不同癌症类型及其亚类的基因组变异图谱。TCGA 的合作研究网络汇集了来自不同学科和不同机构的科研工作者,进而产生有价值的参考数据供全球研究社区使用。NIH 聘请了来自美国国家癌症研究所(National Cancer Institute, NCI)和国际人类基因组研究中心(National Human Genome Research Institute, NHGRI)的科学家及工作人员来开展肿瘤基因组图谱的研究工作。此外,还有其他一些来自美国和欧洲的科学家参与其中,通过协同合作开展大规模肿瘤基因组学数据挖掘,目前,TCGA 已经揭示了近 1 000 万个与癌症相关的基因突变,并且产生了可开放获取的多类型组学数据,这些数据的积累为更深入地开展肿瘤疾病的研究提供了宝贵的资源。

2.2 数据采集和处理流程

TCGA 拥有一套完善的样本采集和数据处理流程,包括:(1)具有广泛的临床基地,能够获得肿

瘤样本及其匹配的正常组织样本,从组织样本中有效地提取 DNA、RNA 等分子。(2)迅速地扫描成千上万的 DNA 和 RNA 序列样本,获得序列信息。(3)协同分析和集成不同类型的数据。(4)高效地将数据提供给研究机构,从而避免病人的隐私泄露。目前,该计划初步落下帷幕,通过大规模基因组测序技术及整合性的多维分析手段,TCGA 的研究人员已经成功地对 30 多种人类常见肿瘤进行了研究^[12]。

2.3 数据积累与数据共享

目前,TCGA 已经取得了一系列研究成果^[13],为临床实践研究提供了宝贵的信息。比如,通过项目的开展,研究人员能够发现某些癌症亚型的特异性通路,有助于发现肿瘤的靶向疗法或组合疗法,进而大大缩短药物开发的时间和降低药物开发的成本。同时,TCGA 存储的肿瘤组学数据是一个宝贵的资源,包含基因表达、蛋白表达、DNA 甲基化、体细胞突变等各种不同种类的数据,可供开展大规模的肿瘤基因组研究^[14]。作为 TCGA 研究项目的一部分,Kandoth 等^[15]对来自 12 种肿瘤的 3 000 多个样本的基因组数据进行挖掘,找到了 127 个在已知的癌症细胞进程中显著突变的基因。此外,通过对比分析,找到了不同类型肿瘤中共同存在的突变,以及每种肿瘤特异的突变。Brodie 等^[16]通过分析 TCGA 的肺癌组学测序数据,发现了与 CHFR 表达相关的疾病亚类。这些理论研究积累了很多疾病相关的遗传分子信息,为发现致癌物质并设计新的癌症诊断治疗工具创造了基础。通过将研究获得的遗传学变异同病人健康状况、治疗历史和对治疗的反应等详细的临床信息整合在一起,可研究组学突变如何影响病人的预后以及对治疗的反应,进而采取有针对性的治疗措施。

3 肿瘤基因组数据挖掘

3.1 肿瘤基因组异质性挖掘分析

肿瘤的发生发展过程是异常复杂的,而且具有很强的异质性(Heterogeneity)^[17]。即使对于同一种疾病,不同个体的基因组之间也存在数以百万的差异位点,影响生命活动调节功能。例如,基因组内

核苷酸发生细微的变化,可能会导致最后行使生物学功能的蛋白质发生改变,引起机体的代谢功能紊乱,进而使机体表现出不同的表征(比如耐药性、缓解率、复发率等)。此外,肿瘤内部细胞群体会存在一个动态演化过程^[18],而不同阶段的肿瘤细胞会存在不同的基因组突变,这种内部的遗传异质性可以为肿瘤的靶向治疗提供关键信息^[19-20]。因此,在实际科学研究中,需要结合信息学手段,通过对肿瘤患者的基因组数据进行挖掘分析,发现肿瘤内部异质性,为肿瘤的分子诊断提供依据,同时也可以对疾病开展特异性的治疗策略^[20-22]。

3.2 数据挖掘到治疗方案的转换

在临床实践中,可以尝试对肿瘤患者采取精准治疗^[23],将个体异质性考虑在内的疾病预防和治疗策略,将会成为医学研究的新方向^[11, 24]。具体而言,通过对个体的基因组进行测试分析,同时结合大规模肿瘤基因组数据挖掘获取的研究成果,医生不仅可以推断出个体患有肿瘤的概率,还可以预测现有的药物对患者是否有效,帮助选取合适的个体化预防和治疗策略。精准医疗是根据个体基因特征、环境以及生活习惯进行疾病干预以及治疗的最佳方法。美国精准医疗的实施将推动肿瘤个体化医疗的发展,该计划包括采集与分析百万人群基因组数据与临床信息、构建高效存储的数据库、制定相关监管制度、制定数据标准、保护患者隐私和信息安全。其中,癌症被选为该计划的短期目标,通过分析疾病内部遗传分子的变异来进行疾病的风险预测、分类评估和治疗策略选取。因此,肿瘤精准医疗需要一个大的可开放获取的知识库,能够覆盖研究获得的如疾病分型、致病机制、肿瘤相关的分子变异、组学分析模型等肿瘤基因组学信息。

4 肿瘤基因组数据挖掘与应用案例——以白血病为例

4.1 概述

基于肿瘤基因组信息及现有的肿瘤基因组学研究成果,可以对患病人群进行疾病的精准诊断、治

疗和预后预测,同时也可以对正常人群进行疾病的风险预测。本文以白血病为例,设计基于肿瘤基因组数据挖掘开展肿瘤精准预防和诊断治疗策略的方案。在不同类型的肿瘤中,白血病是一种恶性程度极高的血液肿瘤。造血干细胞恶性克隆增殖、分化受阻、不正常凋亡并在骨髓和造血组织中大量累积,导致正常造血功能受阻而形成疾病。白血病的异质性强、亚类多,目前还没有精确有效的预防和治疗手段。早期研究中,科学家已经开始基于基因组数据对白血病开展相关研究^[25-26]。此后,科学家们结合基因测序技术和信息分析手段,开展了大量关于白血病致病机制以及预后影响因素的研究,逐渐加深了人们对这种疾病的认识。结合现有的血液肿瘤研究成果、高通量基因检测技术及数据挖掘方法,可以对白血病进行精准的诊断和治疗,见图1,同时也可以对正常人群进行疾病的患病风险预测,见图2。

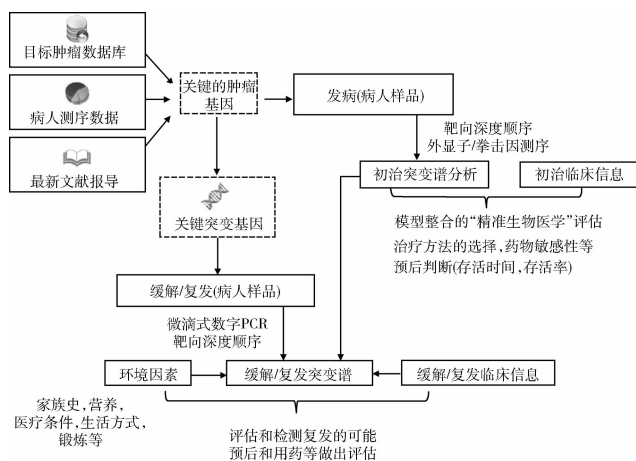


图1 基因组数据挖掘与应用——精准诊疗方案

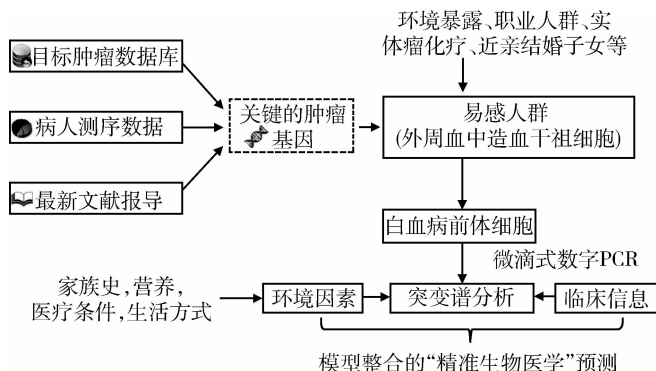


图2 基因组数据挖掘与应用——风险预测

4.2 基于数据挖掘的诊断和治疗

首先,通过挖掘国际大规模肿瘤测序数据库(例如 TCGA、ICGC^[27]、COSMIC^[28])和中国人群的测序数据并及时跟踪最新的血液肿瘤研究报道(例如 PubMed、BIREME^[29]),筛选出关键的白血病相关基因集合并进行分类,这部分的信息将为后续的研究提供重要的数据基础。其次,对于初治的白血病患者,取样进行靶向深度测序、外显子测序或者全基因组测序,结合文献和数据库筛查的白血病基因信息,分析获得患者的关键血病基因突变信息,构建初治突变图谱。此处,靶向深度测序是指针对特定的位点进行测序分析。可以选取一些常见的白血病易感位点,或者根据疾病的表型特征选取特定功能区域进行研究。初治突变图谱只是对肿瘤患者的遗传变异情况进行一个初步的描绘,需要进一步结合患者的临床表型信息及环境因素,采用精准数学模型对患者的发病机理及预后进行预测,对患者有一个初步的诊断。这个过程将患者基因型和表型信息转化为临床实践,精准生物学评估具有一定的挑战,而依据该评估结果可以制定合适的治疗方案。此外,假如患者得到缓解,可以用微滴式数字 PCR(该检测技术具有较高的灵敏度)检测其外周血中关键基因突变及频率变化,以评估和检测复发的可能。如果病人出现复发,通过获取复发时的白血病基因突变图谱,与初治突变图谱进行比较,对病人预后和用药等做出评估。

4.3 基于数据挖掘的风险预测

环境暴露、职业因素、近亲结婚等因素会增加正常人群的患病概率,这些人群也被称为白血病易感人群,可以通过基因组学方法进行疾病的患病风险预测。首先,通过挖掘国际大规模肿瘤测序数据库和中国人群的测序数据,及时跟踪最新的血液肿瘤研究报道,筛选出关键的白血病相关基因集合并进行分类,这部分的信息将为后续的研究提供重要的数据基础。其次,易感个体的外周血中存在造血干细胞,可能是潜在的白血病前体细胞,可以采用微滴式数字 PCR 获得突变谱并进行分析。结合文

献和数据库筛查的白血病基因信息,分析获得易感个体是否存在关键白血病基因的突变(包括致病的和易感的位点)以及突变频率。最后,进一步结合个体临床表型信息及环境因素,采用精准数学模型对易感个体的患病风险进行预测。

5 结语

近几年来,基因组测序技术的出现为肿瘤基因组研究及其应用带来了空前的景象,数以百计的肿瘤基因组案例及其比较研究已经产生,而且这个进程在不断地被推进^[30]。随着组学测序技术的普及和费用的降低,在疾病研究、临床诊断和药物研发领域,对个体基因组进行测序研究并应用于个人的健康管理将变得越来越普遍。每个人都拥有一张自己的基因组图谱,可以根据自己的基因组图谱采取个体化的肿瘤防治策略。而对于大样本数据的研究,可以更好地挖掘肿瘤的发生机制,找到具有统计学证据的癌症生物分子标记。未来仍然还有很多可以挖掘和分析的空间。例如过去人们发现了许多位于编码区的驱动突变,而对于占人类基因组序列约 10%~20% 的非编码区,可能也存在许多的肿瘤相关信息^[31]。

然而,在肿瘤基因组数据挖掘及其应用中,仍然面临着管理、技术、人才和法律方面的挑战。从管理方面,如何将不同项目的研究结果进行汇总,建成类似于国际核酸序列数据库的组织系统,为世界各地的研究者提供一个统一的资源共享平台,仍然比较困难。从技术方面,对于精准医疗而言,由于肿瘤的多样性与致病机制的复杂性,创建精准的数学模型来评估基因组突变与疾病之间的关系,并给出详细的疾病治疗策略,仍然需要提升。对于一些罕见疾病,由于样本数量的限制,很难给出一个合理的生物学模型或者有统计学意义的生物标记来对疾病进行评估。从人才方面,将理论研究成果转化到实际的临床应用上也面临一些困难。将基因检测技术应用到临床上肿瘤的诊断与治疗的时候,如何让医生与患者读懂一份检测报告并不容易。针对这个问题,美国有相应的遗传咨询行业^[32],通过美

国遗传咨询师认证行会 (American Board of Genetic Counselor, ABGC) 认证的注册遗传咨询师, 在学术与临床方面均有能力为个人或家庭提供关于出现某种遗传疾病或先天缺陷或其发生风险方面的基因健康的咨询服务。从法律方面, 我国目前的基因检测市场的行业标准不完善, 相应的法律法规还需要进一步完善。基因检测会存在一些伦理、法律和社会相关问题。此外, 如何保护病人的数据隐私也是一个亟需解决的问题。尽管面临的困难还比较多, 但肿瘤基因组研究改变了对于肿瘤的认知与研究方式, 未来将会不断推进医学的发展。

参考文献

- Sanegr F, Coulson AR. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase [J]. *Journal of Molecular Biology*, 1975, 94 (3): 441–448.
- Sanger F, Nicklen S, Coulson AR. DNA Sequencing with Chain-terminating Inhibitors [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 1977, 74 (12): 5463–5467.
- Margulies M, Egholm M, Altman WE, et al. Genome Sequencing in Microfabricated High-density Picolitre Reactors [J]. *Nature*, 2005, 437 (7057): 376–380.
- Fedurco M, Romieu A, Williams S, et al. BTA, A Novel Reagent for DNA Attachment on Glass and Efficient Generation of Solid-phase Amplified DNA Colonies [J]. *Nucleic Acids Research*, 2006, 34 (3): e22.
- Shendure J, Porreca GJ, Reppas NB, et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome [J]. *Science*, 2005, 309 (5741): 1728–1732.
- Gullapalli RR, Desai KV, Santana-santos L, et al. Next Generation Sequencing in Clinical Medicine: challenges and lessons for pathology and biomedical informatics [J]. *Journal of Pathology Informatics*, 2012, (3): 40.
- Shapiro E, Biezuner T, Linnarsson S. Single-cell Sequencing-based Technologies will Revolutionize Whole-Organism Science [J]. *Nature Reviews Genetics*, 2013, 14 (9): 618–630.
- Worthe EA, Mayer AN, Syverson GD, et al. Making a Definitive Diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease [J]. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 2011, 13 (3): 255–262.
- Hanahan D, Weinberg RA. Hallmarks of Cancer: the next generation [J]. *Cell*, 2011, 144 (5): 646–674.
- Tomczak K, Czerwinski P, Wizeowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge [J]. *Contemporary Oncology*, 2015, 19 (1A): A68–A77.
- Collins FS, Varmus H. A New Initiative on Precision Medicine [J]. *The New England Journal of Medicine*, 2015, 372 (9): 793–795.
- Ledford H. End of Cancer – Genome Project Prompts Re-think [J]. *Nature*, 2015, 517 (7533): 128–129.
- Devarakonda S, Morgensztern D, Govindan R. Clinical Applications of The Cancer Genome Atlas Project (TCGA) for Squamous Cell Lung Carcinoma [J]. *Oncology*, 2013, 27 (9): 899–906.
- Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project [J]. *Nature Genetics*, 2013, 45 (10): 1113–1120.
- Kandoth C, McLellan MD, Vandin F, et al. Mutational Landscape and Significance Across 12 Major Cancer Types [J]. *Nature*, 2013, 502 (7471): 333–339.
- Brodie SA, Li G, Brandes JC. Molecular Characteristics of Non-small Cell Lung Cancer with Reduced CHFR Expression in The Cancer Genome Atlas (TCGA) Project [J]. *Respiratory Medicine*, 2015, 109 (1): 131–136.
- Lawrence MS, Stojanov P, Polak P, et al. Mutational Heterogeneity in Cancer and the Search for New Cancer-associated Genes [J]. *Nature*, 2013, 499 (7457): 214–218.
- Jan M, Majeti R. Clonal Evolution of Acute Leukemia Genomes [J]. *Oncogene*, 2013, 32 (2): 135–140.
- Shlush LI, Zandi S, Mitchell A, et al. Identification of Pre-leukaemic Haematopoietic Stem Cells in Acute Leukaemia [J]. *Nature*, 2014, 506 (7488): 328–333.
- Gerlinger M, Rowan A J, Horswell S, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing [J]. *The New England Journal of Medicine*, 2012, 366 (10): 883–892.
- Ding L, Wendl MC, McMichael JF, et al. Expanding the Computational Toolbox for Mining cancer Genomes [J].

- Nature Reviews Genetics, 2014, 15 (8): 556 – 570.
- 22 Fietze S, Ogeen H, Littlepage LE, et al. Global Analysis of ZNF217 Chromatin occupancy in the Breast Cancer Cell Genome Reveals an Association with ERalpha [J]. BMC Genomics, 2014, (15): 520.
 - 23 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 [M]. National Academies Press, 2012: 34 – 37.
 - 24 Katsnelson A. Momentum Grows to Make Personalized Medicine More Precise [J]. Nat Med, 2013, 19 (3): 249.
 - 25 Rowley J D. Letter: A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia Identified by Quinacrine Fluorescence and Giemsa Staining [J]. Nature, 1973, 243 (5405): 290 – 293.
 - 26 Ley TJ, Mardis ER, Ding L, et al. DNA Sequencing of a Cytogenetically Normal Acute Myeloid Leukaemia Genome [J]. Nature, 2008, 456 (7218): 66 – 72.
 - 27 International Cancer Genome C, Hudson TJ, Anderson W, et al. International Network of Cancer Genome Projects [J]. Nature, 2010, 464 (7291): 993 – 998.
 - 28 Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer [J]. Nucleic Acids Research, 2015, 43 (Database issue): D805 – D811.
 - 29 Bonham M D. BIREME: latin american and caribbean health sciences information center [J]. Bulletin of the Medical Library Association, 1990, 78 (2): 119 – 123.
 - 30 Ashworth A, Hudson TJ. Genomics: comparisons across cancers [J]. Nature, 2013, 502 (7471): 306 – 307.
 - 31 Huang T, Alvarez A, Hu B, et al. Noncoding RNAs in Cancer and Cancer Stem Cells [J]. Chinese Journal of Cancer, 2013, 32 (11): 582 – 593.
 - 32 Trepanier A, Ahrens M, McKinnon W, et al. Genetic Cancer Risk Assessment and Counseling: recommendations of the national society of genetic counselors [J]. Journal of Genetic Counseling, 2004, 13 (2): 83 – 114.

(上接第 63 页)

5 结语

大数据给人的第一感觉是技术层面的事情，但大数据平台“鼓励从更多、更广、更深的角度看待、分析数据”的工作理念对工作的影响更为深远。大数据平台的应用促使人们从更广、更深的角度看待和分析数据，更加宏观地思考监督数据与监督之外的数据如何结合，如何通过更广泛的数据联系来发现可能的关系和数据价值，例如与国家宏观经济数据、卫生统计数据等之间进行关联分析，发现可能帮助开展卫生计生监督工作的有意义的结论。大数

据平台在大数据管理和利用方面提供了一个扩展性、可管理性、集成性方面都比较好的技术手段。大数据平台可以让业务人员更方便去思考和实现数据的多角度利用，便于从概念验证转向生产系统。

参考文献

- 1 樊春, 徐庆华, 胡庆生, 等. 安徽省 3 市黑诊所周边居民对非法行医认知及其影响因素研究 [J]. 中国卫生监督杂志, 2015, 22 (3): 262 – 266.
- 2 涂惠敏. 浙江省非法行医现状调查及长效治理策略分析研究 [D]. 杭州: 浙江大学, 2008.