

胶质瘤生物信息学研究资源与进展

吴从严, 李镭, 杨金成, 赵耀东

(南京医科大学附属上海市第一人民医院神经外科, 上海 200080)

[摘要] 胶质瘤是一种常见的原发性恶性脑瘤, 尽管治疗方法不断更新, 但患者预后仍然很差, 其中最具侵袭性的胶质瘤生存期仅有几个月。生物信息学起源于 20 世纪 70 年代, 随着分子生物学和计算机科学技术的进步, 生物信息学发展很快, 带动基因组、转录组、蛋白组、代谢组等组学概念应运而生, 从而推动了胶质瘤等疾病诊疗技术的进步。例如胶质瘤分类得到细化, 已将分子亚型融入胶质瘤分类体系。现常用于胶质瘤研究的数据库有: GEO 数据库、TCGA 数据库、CGGA 数据库、REMBRANDT 数据库等, 利用这些数据库中上传的样本测序信息, 能够从基因组、转录组、蛋白组、代谢组、甲基化组等多组学的角度阐明胶质瘤的发生发展机制, 为胶质瘤诊疗技术的进步带来新的希望。

[关键词] 神经胶质瘤; 计算生物学; 综述

DOI: 10.3969/J.issn.1672-6790.2020.06.032

Research progress in the bioinformatics of glioma Wu Congyan, Li Lei, Yang Jincheng, Zhao Yaodong (Department of Neurosurgery, Shanghai First People's Hospital of Nanjing Medical University, Shanghai 200080, China)

Corresponding author: Zhao Yaodong, Email: zhaoyd@aliyun.com

[Abstract] Glioma is a common primary malignant brain tumor. Despite continuous improvement in treatment, the prognosis is still poor, and the survival time is only a few months for the most aggressive glioma patient. Bioinformatics, which originated in the 1970s, has been booming with the development of new molecular biology technologies and the progress of computer science. It leads to the emergence of such omics concepts as genome, transcriptome, proteome and metabolome, and promotes the progress of glioma diagnosis and treatment, such as the refinement of glioma classification and the integration of molecular subtypes into the classification of glioma. Currently, it is commonly used in glioma research databases such as GEO, TCGA, CGGA and REMBRANDT. By using the sample sequencing information uploaded in these databases, the occurrence and development mechanism of glioma can be studied from the perspective of multi-omics, including genome, transcriptome, proteome, metabolome, methylome, and bring new hope for the improvement of diagnosis and treatment of glioma.

[Keywords] Glioma; Computational biology; Review

胶质瘤是最常见的原发恶性脑瘤, 也是最致命的恶性脑瘤之一。据统计每年胶质瘤新增确诊病例高达 17 000 例。对比世卫组织 2007 年和 2016 年版中枢神经系统肿瘤组织病理学分类标准^[1-2]可以发现, 2016 版增加了分子遗传学特征作为分型的指标, 如 II 级少突胶质细胞瘤[异柠檬酸脱氢酶 (IDH) 突变型和 1p/19q 共缺失型]和弥漫性星形细胞瘤 (IDH 突变型), III 级间变性少突胶质细胞瘤 (IDH 突变型和 1p/19q 共缺失型) 和间变性星形细胞瘤

(IDH 突变型), IV 级的胶质母细胞瘤和弥漫性中线胶质瘤 (h3k27 突变型)^[3]。将分子亚型融入胶质瘤的分类, 有助于加深对胶质瘤本质的理解。而这一巨大进步是与生物信息学的发展密不可分的。尽管胶质瘤的诊断治疗方法不断进步, 但患者预后仍然很差, 其中最具侵袭性的胶质瘤, 如胶质母细胞瘤和弥漫性中线胶质瘤, 生存期仅有几个月^[4]。因此, 迫切需要深入了解胶质瘤的发病机制, 寻找新的分子靶标, 开发新的治疗技术, 而蓬勃发展的生物信息学

基金项目: 国家自然科学基金 (81172400, 81101909)

作者简介: 吴从严, 硕士研究生, Email: wcy1995qt@163.com

通信作者: 赵耀东, 副主任医师, 硕士生导师, Email: zhaoyd@aliyun.com

为此提供了可能。

1 生物信息学发展概述

“生物信息学”名词由荷兰理论生物学家 Paulien Hogeweg 在 1970 年提出,他当时意思是指“生物系统中的信息过程研究”^[5]。1995 年美国在人类基因组计划的第一个五年总结报告中重新定义了生物信息学:作为一门交叉学科,通过综合运用数学、计算机科学、生命技术理论和工具,实现了对生物科学领域中信息的综合获取、存储、加工、分析、解释,阐明高通量生物数据中所包含的生物学意义^[6]。生物信息学作为一门建立在数据积累,尤其是原始数据积累上的学科,它的主要特点是收集数据的规模化、数据处理的程序化与数据分析的专门化。生物信息学研究目前有 3 个层面:第 1 个层面是根据现有的生物信息学资源(如生物信息学数据库),利用相应的工具(如专业网站、软件等)解决问题;第 2 个层面是利用数值计算方法、数据统计学方法以及相关工具对原始数据进行处理研究;第 3 个层面是提出有价值的生物信息学问题,研究新型方法,开发新型工具,引领生物信息学研究。

20 世纪 60 年代初,随着计算方法在蛋白质序列分析中的应用,奠定了生物信息学的基础;然后随着分子生物学新技术的开发和计算机科学的进步,生物信息学获得突破性进展,由官方设立的大型医学数据管理机构先后设立并快速壮大,如美国国立生物技术信息中心(NCBI)、欧洲生物信息学研究所(EBI)等;随着 90 年代人类基因组计划实施,测序技术的进一步发展,基因组、转录组、蛋白组与代谢组等组学概念应运而生。进入 21 世纪,微阵列技术和高通量测序具有更高的分辨率和更广的检测范围,以及技术进步带来检测成本的下降,使得生物信息学得到更广泛的应用。2009 年建立的单细胞 RNA 测序(scRNA-seq)技术,能够检测单细胞水平上的基因表达,具有比高通量测序更高的细胞差异分辨率,将极大地促进对细胞功能、疾病发展和治疗反应的认识^[6-7]。但由于其测序成本昂贵,目前开展得还不够广泛。

2 胶质瘤的生物信息学研究资源

开发数据库是生物信息学研究的关键,现有的生物信息学数据库几乎覆盖了生物医学的所有领域。下面就胶质瘤研究领域的常用生物信息学数据库进行简述。

2.1 GEO(Gene Expression Omnibus)数据库 2000

年 NCBI 启动了 GEO 数据库项目,主要存储高通量测序获得的基因表达数据^[8]。虽然其大多数(90%)数据是基因表达数据,但其数据范围也扩展到基因组甲基化、蛋白质分析、染色体构象研究和基因组变异/拷贝数的研究^[9]。由于基因组研究数据的复杂性,分析工作具有很强的专业性。为了方便尽可能多的研究人员使用数据库,GEO 开发了在线的数据查询、可视化和分析工具 GEO-2R,使得研究人员不需要下载或操作数据文件,可以直接在 GEO 网站上进行简单的数据分析^[10]。

2.2 癌症基因组图谱(The Cancer Genome Atlas, TCGA) 2005 年 NCBI 启动 TCGA 项目,目的是为了构建反映恶性肿瘤关键基因组变化的全面多维图谱,目前存储的数据包括 DNA 拷贝数变异、甲基化、mRNA 和 microRNA 表达、蛋白质表达和临床数据。目前已有基于 TCGA 数据集的探索性分析工具和数据库。例如癌症基因组学门户网站(<http://www.cbioportal.org>)具有开放式获取资源以及资源开源的特点,已经开发了交互式探索癌症基因组的数据集和直观呈现数据的方法,包括快速查看与癌症相关的基因或通路,生存预后情况以及进行相关的生物网络分析^[11-12]。此外,这些 TCGA 数据集也可以通过定期更新的 FireBrowse 网站直接下载。这些工具提供了快速、直观的癌症基因组图谱和匹配的临床数据,并允许将这些有价值的数据进行生物学解读和临床应用。此外,为了研究 TCGA 数据库中长链非编码 RNA(lncRNA)的生物学特性,Cerami 等^[11]和 Gao 等^[12]分别开发了 TANRIC 和 Co-LncRNA 数据库,促进了 lncRNA 生物学功能的研究和临床应用。

2.3 中国胶质瘤基因组图谱(Chinese Glioma Genome Atlas, CGGA) CGGA 项目(<http://www.cgga.org.cn>)由北京神经外科研究所和中国胶质瘤协作组主办,旨在对导致胶质瘤进展的主要基因组变异进行编目和识别,并为中国胶质瘤样本提供详细的基因组特征描述。CGGA 门户网站是一个开放访问平台,提供多维胶质瘤基因组的数据集。GLIOMASdb(<http://www.cgga.org.cn:9091/gliomasdb>)作为一个免费提供在线访问的分析平台,能够对数据库信息进行简单分析^[13]。

2.4 其他数据集 如:分子脑瘤数据库(REMBRANDT)是一个癌症临床基因组数据库和一个在线挖掘和分析平台,方便人们通过有效地结合临床

信息和基因组特征来提高对胶质瘤的认识^[14]。

3 生物信息学促进胶质瘤研究快速发展

目前在胶质瘤的研究中,通过对样本测序获得的基因组、转录组、蛋白组、代谢组与甲基化组等数据进行生物信息学分析,使研究人员能够了解胶质瘤的致病机制,明确相关分子标志物,为胶质瘤的诊断、治疗、预后评估提供新的思路。这方面进展简述如下:

临床上发现低级别胶质瘤几乎无一例外地进展为继发性胶质母细胞瘤(sGBM),治疗选择有限,机制也不清楚。有学者^[15]通过对 118 例 sGBM 测序获取了突变图谱,发现 sGBM 的 MET 第 14 号外显子跳跃(METex14),以及 PTPRZ1-MET(ZM)融合基因和 METex14 等 MET 基因相关突变可促进胶质瘤恶性进展;并由此开展了 MET 抑制剂 PLB-1001(伯瑞替尼)治疗高级别胶质瘤的 I 期临床试验,确定了 II 期临床试验的推荐剂量。

在转录组学层面,长链非编码 RNA(Long non-coding RNA, lncRNA)是一组长度超过 200 个核苷酸的非编码核苷酸序列,lncRNA 可以作为竞争性的内源性 RNA(ceRNA)来调节 miRNA 的表达,进而调节基因的表达^[16]。Han 等^[17]通过对 TCGA 和 CGGA 中低级别和高级别胶质瘤的 RNA-seq 数据库中的 lncRNA 进行分析,筛选出表达差异最明显的 lncRNA-MIR22HG,通过分子实验证明了沉默胶质母细胞瘤的 MIR22HG 会导致 miR-22-3p 和-5p 表达下调,进而抑制 Wnt/ntcatenin 信号通路,从而导致细胞增殖、侵袭能力减弱并且抑制体内肿瘤生长,最后还基于 miR-22 的前体结构研制了小分子抑制剂 AC1L6JTK,阻止 miR-22 成熟体的形成,从而抑制体内肿瘤的生长。

虽然胶质瘤基因扩增或丢失等遗传学机制已经得到充分研究,但关于不同级别胶质瘤之间蛋白组学的变化信息却很少。Buser 等^[18]通过质谱和生物信息学分析相结合的方法,对人胶质瘤组织切片进行无偏倚定量蛋白质组学研究,发现介导细胞内吞作用的网格蛋白、AP-2 接头蛋白和吞蛋白水平下调,导致细胞表面的相应受体增多,表明内吞机制成分的缺失是不同胶质瘤分级和亚类的共同特征。

异柠檬酸脱氢酶 1(IDH1)的突变是低度胶质瘤和继发性胶质母细胞瘤的高频事件,而 IDH1 突变型胶质瘤易受干预措施影响。新陈代谢的重编程是癌症的标志之一,而 IDH1 突变相较 IDH1 野生型

的胶质瘤新陈代谢的改变尚不清楚。Zhou 等^[19]在 IDH1 突变型胶质瘤组织中发现了从糖酵解到脂质代谢方式的改变,IDH1 突变型胶质瘤中低脂酰基的降低可能是由长链酰基辅酶 a 合成酶 1(ACSL1)、ACSL4 和极长链酰基辅酶 a 合成酶 3(ACSVL3)所介导的;通过在 RNA-seq 数据库中验证,发现较低的 ACSL1 表达与 IDH1 突变型胶质瘤患者生存预后更好有关。

胶质瘤的进展与异常的 DNA 甲基化密切相关,MGMT 启动子甲基化状态被认为是高级别胶质瘤的独立危险因素,决定了烷基化剂的化疗效果^[20]。大多数研究集中在蛋白质编码基因的甲基化或 lncRNA 如何调节蛋白质编码基因的甲基化,而 lncRNA 基因的甲基化概况及其对肿瘤发生和发展的相关后果却知之甚少。Li 等^[21]通过对公共数据库中 II 级、III 级、IV 级胶质瘤的 lncRNA、甲基化及预后数据进行比对分析,发现 60 种 lncRNA 有显著差异表达,其中包括已知的 PVT1、HOTAIR、H19 和了解很少的 CARD8-AS、MIR4435-2HG;通过整合 HM450K 甲基化微阵列数据,证明了其中一些 lncRNA 受甲基化的表观遗传学调控;并发现 MIR4435-2HG 在高级别胶质瘤中高表达,可能影响 EMT 和 TNF 信号通路,并且能通过降低 miR-125-5-p 和 miR-125b-5-p 来增加 CD44 的表达。

4 展望

胶质瘤作为一种高致死性的中枢神经系统恶性肿瘤,有着复杂的致病机制,至今尚不清楚。随着微阵列技术、高通量测序等测序方法的推广应用,通过使用生物信息学方法对组学数据进行加工分析,使研究人员能够从大样本大数据层面阐明胶质瘤的致病机制,为胶质瘤的临床病理分子分型、预后评估、分子靶向治疗及改善胶质瘤对放化疗的耐受性提供了可能。随着生物信息学在胶质瘤研究应用的不断深入,将为胶质瘤的分型、诊断与治疗提供更多依据。

参考文献

- [1] LOUIS D N, PERRY A, REIFENBERGER G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary[J]. Acta Neuropathol, 2016, 131(6): 803-820.
- [2] LOUIS D N, OHGAKI H, WIESTLER O D, et al. The 2007 WHO classification of tumours of the central nervous system[J]. Acta Neuropathol, 2007, 114(2): 97-109.
- [3] ZHAO Z, ZHANG K, WANG Z, et al. A comprehensive

- review of available omics data resources and molecular profiling for precision glioma studies [J]. *Biomed Rep*, 2019, 10(1):3-9.
- [4] ZHAO S, CAI J, LI J, et al. Bioinformatic profiling identifies a glucose-related risk signature for the malignancy of glioma and the survival of patients [J]. *Mol Neurobiol*, 2017, 54(10):8203-8210.
- [5] VINCENT A T, CHARETTE S J. Who qualifies to be a bioinformatician? [J]. *Front Genet*, 2015, 6:164.
- [6] ZHANG X, LI T, LIU F, et al. Comparative analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems [J]. *Mol Cell*, 2019, 73(1):130-142.
- [7] GAO S. Data Analysis in single-cell transcriptome sequencing [J]. *Methods Mol Biol*, 2018, 1754:311-326.
- [8] EDGAR R, DOMRACHEV M, LASH A E. Gene expression omnibus:NCBI gene expression and hybridization array data repository [J]. *Nucleic Acids Res*, 2002, 30(1):207-210.
- [9] BARRETT T, WILHITE S E, LEDOUX P, et al. NCBI GEO:archive for functional genomics data sets--update [J]. *Nucleic Acids Res*, 2013, 41:D991-D995.
- [10] CLOUGH E, BARRETT T. The gene expression omnibus database [J]. *Methods Mol Biol*, 2016, 1418:93-110.
- [11] CERAMI E, GAO J, DOGRUSOZ U, et al. The cBio cancer genomics portal:an open platform for exploring multi-dimensional cancer genomics data [J]. *Cancer Discov*, 2012, 2(5):401-404.
- [12] GAO J, AKSOY B A, DOGRUSOZ U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal [J]. *Sci Signal*, 2013, 6(269):p11.
- [13] ZHAO Z, MENG F, WANG W, et al. Comprehensive RNA-seqtranscriptomic profiling in the malignant progression of gliomas [J]. *Sci Data*, 2017, 4:170024.
- [14] MADHAVAN S, ZENKLUSEN J C, KOTLIAROV Y, et al. Rembrandt: helping personalized medicine become a reality through integrative translational research [J]. *Mol Cancer Res*, 2009, 7(2):157-167.
- [15] HU H, MU Q, BAO Z, et al. Mutational landscape of secondary glioblastoma guides MET-Targeted trial in brain tumor [J]. *Cell*, 2018, 175(6):1665-1678.
- [16] XIONG Z, WANG L, WANG Q, et al. LncRNA MALAT1/miR-129 axis promotes gliomatumorigenesis by targeting SOX2 [J]. *J Cell Mol Med*, 2018, 22:3929-3940.
- [17] HAN M, WANG S, FRITAH S, et al. Interfering with long non-coding RNA MIR22HG processing inhibits glioblastoma progression through suppression of Wnt/ β -catenin signalling [J]. *Brain*, 2020, 143(2):512-530.
- [18] BUSER D P, RITZ M F, MOES S, et al. Quantitative proteomics reveals reduction of endocytic machinery components in gliomas [J]. *EBio Medicine*, 2019, 46:32-41.
- [19] ZHOU L, WANG Z, HU C, et al. Integrated metabolomics and lipidomics analyses reveal metabolic reprogramming in human glioma with IDH1 mutation [J]. *J Proteome Res*, 2019, 18(3):960-969.
- [20] WICK W, WELLER M, VAN DEN BENT M, et al. MGMT testing--the challenges for biomarker-based glioma treatment [J]. *Nat Rev Neurol*, 2014, 10(7):372-385.
- [21] LI Z, TAN H, ZHAO W, et al. Integrative analysis of DNA methylation and gene expression profiles identifies MIR4435-2HG as an oncogenic lncRNA for glioma progression [J]. *Gene*, 2019, 715:144012.

(收稿日期:2020-08-18)

严正声明！ 请通过《中国临床保健杂志》官网核实信息

近期, 本刊陆续接到作者投诉, 反映有些网站假冒本刊官网、有些个人或中介机构以本刊名义诱骗作者投稿, 骗取钱财, 致使很多作者上当受骗, 也严重损害了本刊的形象和声誉。

因此, 本刊严正声明: 本刊未委托任何个人或中介机构进行征稿。请广大读者和作者务必访问《中国临床保健杂志》官方网站 (www.zglcbjzz.com) 核实信息 (在"百度"搜索引擎中输入"中国临床保健杂志", 可查询到"官网"认证的链接)。本刊正在收集相关证据, 将依法对侵权者追责!

编辑部电话: 0551-62608457

Email: J. C. Healthcare@163.com