张　卓,江　帅,李睿江,等. 面向肿瘤精准医学的综合数据资源 TCGA 及其相关在线分析工具推荐[J]. 中华医学图书情报杂志,2018,27(3):5-9.

DOI:10.3969/j.issn.1671-3982.2018.03.002

·专题·

面向肿瘤精准医学的综合数据资源 TCGA 及其相关在线分析工具推荐

张　卓,江　帅,李睿江,李宛莹,李　昊,陈河兵,伯晓晨

[摘要]癌症基因组图谱(TCGA)是一个公共资助的项目,旨在编目和发现引起癌变的主要人类基因组变化,目标是创建癌症基因组的全面"图谱"。TCGA 数据库收录了多种癌症组学数据,包括转录组数据、表观遗传组学数据、基因突变数据和疾病样本临床数据等,为认识肿瘤发生的相关知识提供了丰富的资源,可以帮助科研人员更好地学习和认识癌症相关领域知识并促进肿瘤精准医学的实现。调查整理了 TCGA 数据在线分析工具并对其进行筛选推荐,可以帮助研究人员方便地进行 TCGA 数据分析。

[关键词]癌症;多组学数据;精准医学;分析工具;肿瘤;基因组

[中图分类号]R730.2　　[文献标志码]A　　[文章编号]1671-3982(2018)03-0005-05

Comprehensive data resource TCGA for oncology precision medicine and its online analysis tools

ZHANG Zhuo, JIANG Shuai, LI Rui-jiang, LI Wan-ying, LI Hao, CHEN He-bing, BO Xiao-chen

(Institute of Radiation Medicine, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 100850, China)

Corresponding author: BO Xiao-chen

[**Abstract**] TCGA, a public-funded project, is aimed at cataloguing and discovering the major human genome variations that induce canceration. Its goal is to establish a comprehensive cancer genome atlas. A variety of cancer genomics data are covered in TCGA database, including transcriptomics data, epigenomics data, gene mutation data, and disease sample clinical data, which provide a wealth of resources for understanding the knowledge in relation with tumorgenesis, and can thus help scientific researchers to effectively learn and understand cancer-related knowledge, and speed up the realization of oncology precision medicine. The online analysis tools of TCGA data were assessed in order to help scientific researchers to analyze the TCGA data.

[**Key words**] Cancer; Multiomics data; Precision medicine; Analysis tools; Tumor; Genome

癌症是一种极为复杂的人类疾病,涉及基因组的多种动态变化[1]。每种类型的癌症,发生的遗传畸变都是独特的,包括体细胞突变、拷贝数变异、基因表达谱差异和表观遗传改变。因此需要更好地理解肿瘤的各种遗传变化,才能更好地对其进行诊断、治疗和预防。全基因组测序和生物信息技术的发展为癌症基因组研究提供了新的线索[2]。典型的综合数据资源是癌症基因组图谱(The Cancer Genome Atlas,TCGA)项目,它收集整理了大量癌症基因组数据,并利用新的基因组分析技术以加速对癌症的全面了解。

TCGA 数据库的目标是完成一套完整的与所有

[基金项目]"精准医学研究"重点专项"精准医学大数据管理和共享技术平台"(2016YFC0901600)

[作者单位]军事科学院军事医学研究院辐射医学研究所,北京 100850

[作者简介]张　卓(1987-),男,山西临汾人,博士,工程师,主要从事生物信息学研究。

[通讯作者]伯晓晨(1973-),男,天津市人,博士,研究员,主要从事生物信息学研究。E-mail:boxc@bmi.ac.cn

癌症基因组改变相关的“图谱”,旨在获得癌症生物学的新见解,从而有助于癌症的治疗。该项目是2006 年由美国国立卫生研究院牵头的一项大型癌症基因组计划,自 2008 年开始有阶段性成果发表[3],2009 年继续投资 2.75 亿美元,增加了多种类型的癌症数据,到 2014 年已收集了 36 类癌症数据,包括临床数据、DNA、RNA、蛋白质等多层次的数据。在数据生成方面,该项目取得了无可争议的成功。随着样品采集、测序和分析技术的快速发展,TCGA 收录的肿瘤相关数据呈指数增长。目前,新成立的 NCI Genomics Data Commons 将 TCGA 的数据整合在该门户网站中,并且为基因组数据用户提供了交互式支持和更清晰友好的界面。

我们可以用前所未有的微观视角来看待癌症,但是还没有达到能够解释这种疾病的全貌的程度,对其发病机制亦不完全清楚。而 TCGA 数据已被用于发现新的突变,确定内在的肿瘤类型,确定泛癌相似性和差异性,同时收集肿瘤演变的证据。目前已经开发了大量针对 TCGA 数据的生物信息学工具,反映出 TCGA 数据资源的重要性。

1 TCGA 数据介绍

为了全面分析癌症基因组图谱,TCGA 应用基于微阵列和下一代测序方法的高通量技术,产生了癌症的多种数据类型信息。

TCGA 中的癌症数据通过各种标识符(ID)进行识别和编目(表 1),每种癌症类型都包括体细胞突变、拷贝数、基因表达、miRNA 表达、DNA 甲基化、逆转蛋白相位阵列(RPPA)和临床信息。除原始排序文件外(表 2),每种数据类型都包括可供公开下载的原始数据和已处理的数据。

表 1 TCGA 数据库中的 ID 号

ID 类型	描述	示例
File UUID	TCGA 中数据文件 ID	00a2364d-7385-4fa8-8562-b4f19548505a
File Submitted ID	上传至 TCGA 的文件 ID	147f470-7440-42b8-8e3a-4e28b654916e-beta-value
Case UUID	TCGA 中的样本 ID	942c0088-c9a0-428c-a879-e16f8c5bfdb8
Case Submitted ID	上传至 TCGA 的样本 ID,一般用来代表样本	TCGA-CJ-4642
Project ID	样本属于的项目 ID	TCGA-BRCA

表 2 数据类型和可获取水平

数据类型	描述	可访问级别
Aligned Reads	原始测序数据	受限
Raw Simple Somatic Mutation	原始突变信息数据	受限
Annotated Somatic Mutation	注释突变信息数据	受限
Aggregated Somatic Mutation	聚合的突变信息数据	受限
Masked Somatic Mutation	转换后的突变信息数据	开放
Gene Expression Quantification	基因表达数据	开放
Copy Number Segment	拷贝数信息数据	开放
Masked Copy Number Segment	转换后的拷贝数信息数据	开放
Methylation Beta Value	甲基化数据	开放
Isoform Expression Quantification	城市 microRNA 表达数据	开放
miRNA Expression Quantification	microRNA 表达数据	开放
Biospecimen Supplement	生物样本信息	开放
Clinical Supplement	临床信息	开放

2　TCGA 数据在线分析工具

目前 TCGA 数据分析很复杂，涉及多个步骤，为获得有意义的生物学结果，需要仔细考虑分析每个步骤，并将特定工具应用于某些实验模型。为现有数据开发相关的探索工具，需要实验科学家和计算科学家之间的协调。然而，实验科学家很难使用计算科学家开发的计算工具，因为这些计算工具需要数据准备以及安装和使用打包软件，而且某些软件往往只适用于某些特定平台或操作系统。一些更高级的计算工具往往难以理解或使用，从而限制了其应用。不过有基于网络的工具可以提供方便的计算解决方案，帮助实验科学家使用和分析复杂的癌症基因组数据。这些工具帮助无生物信息学背景的生物学家和医学家获得更多的生物学和医学见解，但是选择适当的工具并不是一项简单的任务，对于没有经验的用户来说尤其如此。

本文整理了一个基于网络的可用于分析 TCGA 数据的公开工具列表，并将这些工具进行分类以便更好地进行查询和使用。

表 3 显示了基于网络工具的 32 个在线分析资源，它们代表了当前可用于分析 TCGA 数据的主要资源。为了进一步区分和指导这些工具的选择，本文将所有资源工具分为全局分析工具、目标分析工具和辅助分析工具三大类。

表 3　针对 TCGA 数据的在线分析资源

分类	工具名称	可视化类型	下载
全局分析工具(Ⅰ类)	Broad GDAC Firehose	矩阵、直方图	是
	Cancer Landscapes[4]	矩阵、网络图	否
	canEvolve[5]	网络图、热图	是
	Regulome Explorer[6]	矩阵、Circos、基因组坐标图、网络图	是
	TCGA Mbatch	矩阵、PCA 图、分层聚类图	是
	TCGA NG-CHM	热图	是
	TCPA[7]	热图、网络图	是
全局分析工具(Ⅱ类)	MethHC[8]	矩阵、热图	是
	OASISPRO[9]	直方图、线图/箱线图	是
	OncoScape[10]	矩阵、热图、通路图、散点图	是
	TCGA Clinial Explorer[11]	矩阵、直方图	是
	TCGA SpliceSeq[12]	矩阵	是
目标分析工具	Cancer3D[13]	基因组坐标图、网络图、散点图/箱线图、三维结构图	是
	Cbioportal[14]	矩阵、热图、网络图	是
	GEPIA[15]	矩阵、条状图/箱线图/点线图	是
	IntOGen[16]	矩阵、热图、直方图	是
	KMplotter	线图	是
	MEXPRESS[17]	基因组坐标图	是
	PROGgeneV2[18]	线图	是
	TANRIC[19]	热图	是
	TCGA4U[20]	矩阵、热图、直方图	是
	UALCAN[21]	热图、箱线图、线图	是
	UCSC Xena	热图、散点图、直方图	是
	Wanderer[22]	基因组坐标图、散点图	是
	Zodiac[23]	矩阵、网络图	否
辅助分析工具	BCMD[24]	图像	否
	CDSA[25]	图像	否
	CELLX[26]	矩阵、热图	是
	GDISC[27]	矩阵、箱线图	是
	PathwayMapper[28]	通路图	是
	TCIA[29]	图像	是
	Vanno[30]	矩阵、Circos、三维结构图、热图	是

全局分析工具能够检查癌症基因组的整体特征,可以成为刚刚开始研究癌症基因组数据研究人员的宝贵资源。全局分析工具有两种类型即Ⅰ型和Ⅱ型,前者仅提供全局分析,后者则提供除全局分析之外的选定目标分析。

目标分析工具是研究人员最常使用的基于网络的公共工具。这些工具可以令研究人员深入分析具体的基因或者基因集,甚至 miRNA 等研究对象,方便使用者调查癌症数据中自己感兴趣的目标。

基于公共网络的辅助分析工具可以将 TCGA 数据转换为易于访问、浏览和下载的在线资源。这些数据可以帮助用户补充实验结果或者提供额外的证据和解释,帮助研究人员更全面地分析自己的研究和促进生物学发现。

3 TCGA 数据在线分析工具推荐

首先可以由本文的分类区分不同工具的使用类型,缩小选择范围;然后根据实际需要结合具体研究(如数据来源、数据类型、分析方法、研究目的),选择具体的工具进行进一步的分析。以下是对 TCGA 数据进行不同分析时建议选择的一些工具,但这些工具都不能完全取代先进的计算和统计方法,只是为研究人员提供一些使用帮助,扩展他们癌症组学、癌症复杂性和癌症网络等方面的相关知识。

3.1 突变分析

有 10 种在线工具(Broad GDAC Firehose, Cancer3D, cbioportal, CELLX, IntOGen, TANRIC, TCGA Clinical Explorer, TCGA4U, UCSC Xena 和 Vanno)可以进行突变分析。一般来说,推荐使用 cbioportal,因其包含多种癌症类型和多种可视化分析功能,功能强且易于使用。

3.2 相关性分析

有 17 种在线工具(Broad GDAC Firehose, Cancer Landscapes, canEvolve, cbioportal, CELLX, GDISC, GEPIA, MethHC, MEXPRESS, OASISPRO, Regulome Explorer, TANRIC, TCGA Clinical Explorer, TCGA NG-CHM, TCPA, Wanderer 和 Zodiac)可以进行相关性分析。总的来说,推荐使用麻省理工学院和哈佛大学 Broad 研究所研发的 Broad GDAC Firehose,因其有多种分析算法供用户使用,功能全面,且包含多种分析工具。

3.3 差异分析

有 12 种在线工具(Broad GDAC Firehose, canEvolve, cbioportal, CELLX, GEPIA, MEXPRESS, OncoScape, TANRIC, TCGA4U, TCPA, UALCAN 和 Wanderer)可以进行差异分析,一般推荐使用分析基因表达谱的工具 GEPIA。差异分析是该工具的主要分析功能,其在线分析界面简单易懂,非常易于理解和使用。

3.4 通路分析

有 8 种在线工具(Broad GDAC Firehose, Cancer Landscapes, canEvolve, MethHC, OncoScape, PathwayMapper, Regulome Explorer 和 TCGA NG-CHM)可以进行通路分析。推荐使用 Broad GDAC Firehose 和 OncoScape,前者分析方法丰富,后者简单直观。

3.5 生存分析

有 16 种在线工具(Broad GDAC Firehose, Cancer Landscapes, canEvolve, cbioportal, CELLX, GDISC, GEPIA, KMplotter, OASISPRO, PROGgeneV2, TANRIC, TCGA Clinical Explorer, TCGA4U, TCPA, UALCAN 和 UCSC Xena)可以进行生存分析。如果仅想进行单一的生存分析,推荐使用 PROGgeneV2,因其具有广泛的数据来源和多种可选参数设置。

3.6 泛癌分析

有 8 种在线工具(Broad GDAC Firehose, Cancer Landscapes, cbioportal, IntOGen, Regulome Explorer, TCGA NG-CHM, UCSC Xena 和 Zodiac)可以进行泛癌症分析(pan-cancer analysis)。一般来说,推荐使用 cbioportal 和 Cancer Landscapes,前者收集了来自泛癌研究的大量样本且拥有强大的分析能力;后者的癌症图谱模型中包含了泛癌模型,可以直接用于分析。

4 总结

科学家们开发出多种生物信息学工具进行数据挖掘和分析,以便寻找新发现。不久的将来,新发现将有助于诊断、治疗和预防癌症。TCGA 提供的癌症基因组学数据可以系统地揭示癌症分子生物学的新图景。这些大量公开可用的数据,为世界各地的研究人员提供了癌症遗传学的知识来源,结合多种分析有助于开发个性化癌症药物。本文全面整理了基于网络的公共可用的在线分析资源和工具,可以

帮助研究人员方便地查找和使用合适的工具，增进他们对癌症基因组学的理解。

【参考文献】

[1] Hanahan D, Weinberg RA. The hallmarks of cancer[J]. Cell, 2000, 100(1): 57-70.

[2] Stratton MR, Campbell PJ, Futreal PA. The cancer genome[J]. Nature, 2009, 458(7239): 719-724.

[3] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways[J]. Nature, 2008, 455(7216): 1061-1068.

[4] Kling T, Johansson P, Sanchez J, *et al*. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content[J]. Nucleic Acids Research, 2015, 43(15): e98.

[5] Samur M K, Yan Z, Wang X, *et al*. canEvolve: a web portal for integrative oncogenomics[J]. PLoS One, 2013, 8(2): e56228.

[6] Madhavan S, Gusev Y, Natarajan TG, *et al*. Genome-wide multi-omics profiling of colorectal cancer identifies immune determinants strongly associated with relapse[J]. Frontiers Genetics, 2013, 4: 236.

[7] Li J, Lu Y, Akbani R, *et al*. TCPA: a resource for cancer functional proteomics data[J]. Nature Methods, 2013, 10(11): 1046-1047.

[8] Huang WY, Hsu SD, Huang HY, *et al*. MethHC: a database of DNA methylation and gene expression in human cancer[J]. Nucleic Acids Research, 2015, 43 (Database issue): 856-861.

[9] Yu KH, Fitzpatrick MR, Pappas L, *et al*. Omics AnalySIs System for PRecision Oncology (OASISPRO): a web-based omics analysis tool for clinical phenotype prediction[J]. Bioinformatics, 2017, 34(2): 319-320.

[10] Andreas S, Magali M, Rubayte R, *et al*. OncoScape: exploring the cancer aberration landscape by genomic data fusion[J]. Scientific Reports, 2016, 6(1): 28103.

[11] Lee HJ, Palm J, Grimes SM, *et al*. The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations[J]. Genome Medicine, 2015, 7(1): 1-14.

[12] Ryan M, Wong WC, Brown R, *et al*. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer[J]. Nucleic Acids Research, 2016, 44(1): 1018-1022.

[13] Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures[J]. Nucleic Acids Research, 2015, 43(1): 968-973.

[14] Cerami E, Gao J, Dogrusoz U, *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data[J]. Cancer Discovery, 2012, 2(5): 401-404.

[15] Tang Z, Li C, Kang B, *et al*. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses[J]. Nucleic Acids Research, 2017, 45(1): 98-102.

[16] Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, *et al*. IntOGen-mutations identifies cancer drivers across tumor types[J]. Nature Methods, 2013, 10(11): 1081-1082.

[17] Koch A, De TM, Jeschke J, *et al*. MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data[J]. BMC Genomics, 2015, 16(1): 636.

[18] Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database[J]. BMC Cancer, 2014, 14: 970.

[19] Li J, Han L, Roebuck P, *et al*. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer[J]. Cancer Research, 2015, 75(18): 3728-3737.

[20] Huang ZZ, Duan HL, Li HM. Identification of gene expression pattern related to breast cancer survival using integrated TCGA datasets and genomic tools[J]. Biomed Research International, 2015(6): 878546.

[21] Chandrashekar DS, Bashel B, Balasubramanya SAH, *et al*. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses[J]. Neoplasia, 2017, 19(8): 649-658.

[22] Diez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer[J]. Epigenetics Chromatin, 2015, 8(1): 1-8.

[23] Zhu Y, Xu Y, Jr HD, *et al*. Zodiac: a comprehensive depiction of genetic interactions in cancer by integrating TCGA data[J]. Journal of the National Cancer Insttitute, 2015, 107(8).

[24] Chang H, Han J, Borowsky A, *et al*. Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association[J]. IEEE Transactions Medical Imaging, 2013, 32(4): 670-682.

[25] Gutman DA, Cobb J, Somanna D, *et al*. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data[J]. Journal of American Medical Informatics Association Jamia, 2013, 20(6): 1091-1098.

[26] Ching KA, Wang K, Kan Z, *et al*. Cell Index Database (CELLX): a web tool for cancer precision medicine[J]. Pacific Symposium on Biocomputing, 2015: 10-19.

[27] Spainhour J C G, Lim J, Qiu P. GDISC: a web portal for integrative analysis of gene-drug interaction for survival in cancer[J]. Bioinformatics, 2017, 33(9): 1426-1428.

[28] Bahceci I, Dogrusoz U, La KC, *et al*. PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data[J]. Bioinformatics, 2017, 33(14): 2238-2240.

[29] Clark K, Vendt B, Smith K, *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository[J]. Journal of Digittal Imaging, 2013, 26(6): 1045-1057.

[30] Huang PJ, Lee CC, Tan BC, *et al*. Vanno: a visualizationaided variant annotation tool[J]. Human Mutation, 2015, 36(2): 167-174.

[收稿日期：2018-02-01]

[本文编辑：黄思敏]