癌症基因组图谱计划数据及分析*

邓祯祥 综述 李金明 审校

摘要 人类癌症细胞中通常藏匿着多种引起恶性病变的染色体变异，核酸替换和表观遗传修饰。癌症基因组图谱（the cancer genome atlas，TCGA）计划的目标是获取、刻画并分析人类癌症中大规模、多种变异的分子特征，并且为癌症研究者迅速地提供数据。本文对TCGA数据的四个产生流程以及包含的癌症种类、数据类型、数据水平、分析流程和常用的几种分析工具等进行阐述，同时以卵巢癌（ovarian cancer）为例详细介绍了TCGA数据在突变分析、拷贝数分析、表达分析和通路分析等方面的应用，并对TCGA研究团队近几年有关胶质母细胞瘤（glioblastoma，GBM）的研究方法和结果以及已经完成分析的癌症类型进行综述。

关键词 癌症基因组图谱 数据类型 数据分析 数据挖掘

doi:10.3969/j.issn.1000-8179.20130649

Data and analysis of the cancer genome atlas

Zhenxiang DENG, Jinming LI

Correspondence to: Jinming LI; E-mail: jmli@smu.edu.cn

Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China.

This work was supported by the Startup Grant for Talent Recruitment of High Education in Guangdong Province (No. YCJ 2011-430).

Abstract Multiple chromosomal aberrations, nucleotide substitutions, and epigenetic modifications may occur in human cancer cells, which drive malignant transformation. The Cancer Genome Atlas (TCGA) project aims to promote large-scale multi-dimensional analysis of these molecular characteristics in human cancer and rapidly provide data to researchers. In this study, we introduce four flow paths of the production of TCGA data, the collections of various cancer types, the data category and level, and the standardized pipeline of data analysis, as well as several existing data analytical tools. We used ovarian cancer as an example to introduce the application of the TCGA data in the analyses of mutation, copy number, analysis, and expression. We summarized the important findings of glioblastoma by TCGA teams.

Keywords: cancer genome atlas, data category, data analysis, data mining

癌症基因组图谱计划是由美国国立卫生研究院（national institutes of health，NIH）组织美国国立癌症研究所（nationalcancerinstitute，NCI）和国立人类基因组研究所（national human genome researchinstitute，NHGRI）于2006年启动的大型研究。TCGA计划目的是希望能够全面的、系统性的了解恶性肿瘤的形成、生长、转移等过程的生物学基础，以及与病理机制相关的基因组变化，促进癌症的早期诊断和加速癌症治疗的发展步伐，并且能进一步的预防癌症的发生。目前在美国，7所医院的临床癌症中心和3所临床基因测序中心已经加入到这项研究计划，并新建了样本资源中心和临床遗传信息处理中心。TCGA计划是通过利用这些研究中心的组织样本、仪器设备以及研究团队完成。大规模的收集了数百例特定癌症患者的临床信息，肿瘤组织及其相对应的正常组织样本或血液样本，并进行全面的基因组数据分析和整合分析，以便能够进一步加深对癌症分子生物机制的了解。

1 TCGA数据

1.1 数据简介

TCGA数据主要是通过组织处理（BCR）、整合研究、数据分享和团队研究四个方面来获得的。组织处理的主要工作包括收集癌症患者捐赠的肿瘤组织和正常组织，以及对样本组织进行标准化处理并获取基因组数据和临床数据。整合研究的工作主要由癌症基因组中心（CGCCs）、基因组测序中心（GSC）和基因组数据分析中心（GDAC）三个部门共同完成。其中CGCCs收集了几百个肿瘤和正常组织样本的基

作者单位：南方医科大学基础医学院生物信息教研室（广州市 510515）

*本文课题受广东省高校引进人才专项资金（编号：YCJ2011-430）资助

通信作者：李金明 jmli@smu.edu.cn

网络出版日期：2013-12-11 网络出版地址：http://www.cnki.net/kcms/detail/12.1099.R.20131211.0910.002.html

因组数据并进行大规模的统计学分析，识别其中包含的差异表达基因和DNA拷贝数变异基因；GSC通过对癌症关联特征鉴定的候选基因和基因组区域进行大规模的高通量测序；GDAC完成数据处理、统计分析，并为所有研究团队提供图表报告。数据分享主要通过数据协调中心（data coordination center，DCC）实现，DCC对TCGA所产生的数据建立数据库进行分析，并定期在网上公布以便全球的临床科研机构可以迅速、准确的获取测序信息和基因组分析结果等。团队研究是TCGA计划希望所有癌症研究组织都能有效的利用其数据进行研究分析，以改进现有的诊断、治疗方法，降低癌症患者死亡率，造福全人类。

TCGA试点计划已经完成，并且成功在胶质母细胞瘤[1]、卵巢癌[2]和结直肠癌[3]等癌症中证实了特定癌症的基因组变化图谱是可以绘制出来的，并且证明资源集中策略能有效地加速癌症研究。近年来TCGA的研究对象已经覆盖超过了20多种癌症（表1）。到目前为止，TCGA研究团队及其他利用TCGA数据的研究小组已在国际知名期刊上发表了多篇论文[1-10]。

TCGA计划的大部分数据和研究结果都公开在TCGA数据门户网站上，并且可以免费下载。世界各地的研究者在TCGA研究团队第1次使用数据发表文章后都可以将这些数据应用到自己的相关研究领域中。至今已有7种癌症的相关论文已发表或接近发表，因此这些癌症数据可以供研究者使用。另有肺腺癌、甲状腺癌、头颈癌、黑色素瘤、胃腺癌、低程度胶质瘤、宫颈鳞状细胞癌、膀胱癌8种癌症的数据资料已公开，但是发表基于这些数据的文章需解禁后才能使用，暂时受到限制。其他癌症数据因样本量尚未达到预期目标，解禁日期未定。

1.2 TCGA数据分类

TCGA数据门户网站上的数据有独特的数据分类体系，其将数据分成不同的数据类型和不同的数据水平。

表1 TCGA包含的25种癌症以及收集到的样本和可供下载的样本数

Table 1 The 25 types of cancer in TCGA, cancer samples, and number of downloadable cancer samples collected

Available Cancer Types	Cases Shipped by BCR	Cases with Data	Date Last Updated
Acute Myeloid Leukemia [LAML]	202	200	13-2-15
Bladder Urothelial Carcinoma [BLCA]	171	153	13-4-12
Brain Lower Grade Glioma [LGG]	247	222	13-4-13
Breast invasive carcinoma [BRCA]	972	940	13-4-12
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	144	134	13-4-11
Colon adenocarcinoma [COAD]	424	424	13-4-12
Esophageal carcinoma [ESCA]	50	0	03-3-25
Glioblastoma multiforme [GBM]	600	597	13-3-01
Head and Neck squamous cell carcinoma [HNSC]	374	358	13-4-12
Kidney Chromophobe [KICH]	66	66	13-4-11
Kidney renal clear cell carcinoma [KIRC]	513	501	13-4-12
Kidney renal papillary cell carcinoma [KIRP]	142	127	03-3-27
Liver hepatocellular carcinoma [LIHC]	126	99	13-4-03
Lung adenocarcinoma [LUAD]	563	538	13-4-12
Lung squamous cell carcinoma [LUSC]	494	399	13-4-12
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]	28	18	13-4-08
Ovarian serous cystadenocarcinoma [OV]	597	601	13-4-11
Pancreatic adenocarcinoma [PAAD]	73	57	13-4-12
Prostate adenocarcinoma [PRAD]	199	188	13-4-09
Rectum adenocarcinoma [READ]	169	165	13-3-25
Sarcoma [SARC]	73	54	13-4-12
Skin Cutaneous Melanoma [SKCM]	336	318	13-4-12
Stomach adenocarcinoma [STAD]	325	325	13-4-12
Thyroid carcinoma [THCA]	500	508	13-4-14
Uterine Corpus Endometrioid Carcinoma [UCEC]	525	500	13-4-11

1.2.1　数据类型　在TCGA计划中，各个研究团队通过Agilent、Illumina、RNAseq等平台获得mRNA表达数据、microRNA表达数据、拷贝数数据、蛋白质数据、基因突变数据以及甲基化数据，同时收集患者基本资料、治疗进程、临床分期和生存状况等临床数据。每一个平台都可以测得上述多种类的数据（数据类型），包括样本基因组中的基因突变（插入/缺失）、DNA拷贝数、mRNA表达、microRNA表达、蛋白质表达和DNA甲基化数据。TCGA研究团队同时还获得了一些与癌症组织配对（matched）和不配对（unmatched）的正常组织样本的mRNA/microRNA表达数据、DNA拷贝数数据或者甲基化数据。

1.2.2　数据水平分类　数据水平是TCGA计划使用的数据分类方法，促进研究者交流和定位感兴趣的数据。每一种数据类型，实验平台和实验中心都会获得4个水平的数据。其中水平1是单个样本的低水平且未经过标准化的原始数据；水平2是经过标准化后的单样本数据或者是对存在或不存在特定分子异常的解释数据；水平3是对单个样本经过处理的数据汇集或者是已探测的基因位点集合形成的较大的连续区域；水平4是感兴趣的区域或概要，主要包括量化各类样本之间的关联、基于两个或多个数据的关联以及存在分子异常、样本特征和临床变量的数据。

1.2.3　数据类型和数据水平分类之间的关系　由于每一种平台都能产生多种数据类型，所以为了理解数据的分类有必要弄清数据类型和数据水平之间的关系。对于每一种数据类型，数据水平进一步对数据的分析程度及其结果进行分类。每个中心、平台可能会有略微不同的数据水平，这主要取决于数据类型的分析平台和计算方法。

2　TCGA数据分析流程及工具

2.1　分析流程

为了帮助临床研究者和癌症生物学者更有效的利用癌症基因组图谱计划产生的数据，TCGA中多个团队组建了GDAC，建立了TCGA数据分析的标准化流程并定期将这些结果公开。主要包括MutSig（mutation significance）分析[11]：整合所有患者的基因突变，识别差异具有统计学意义的突变基因；GISTIC（genome identification of significant targets in cancer）分析[4]：整合所有患者的基因组拷贝数数据，识别染色体上差异具有统计学意义的扩增或缺失区域，并列出这些区域中所包含的基因；CNMF聚类：通过对各种数据完成非负矩阵因子非监督聚类分析，寻找肿瘤中可能的亚型；各种临床相关和生存分析：寻找与癌症分期或生存等预后最相关的基因表达、突变和肿瘤亚型以及Pathway分析等。这些分析不仅提供各式图表利于生物学者建立或验证研究假说，而且由于分析方法的标准化、统一化大幅提高了各种数据间的可比性。

2.2　分析工具

目前已经存在多个数据分析工具能够对TCGA数据分析起作用。这些工具主要包括：GenePattern：能够对基因表达、蛋白质组和单核苷酸多态性（single nucleotide polymorphism，SNP）数据提供超过90种计算和可视化工具；Genboree：一个用于基因组研究的软件系统，能够研究阵列比较基因组杂交技术（aCGH）、基于PCR的重测序技术、比较序列拼接的基因组重测序技术、使用配对末端标签和序列的基因组重映射技术、基因组注释、多基因组比较和通过基因组自身比较的模型发现技术等产生的数据中基因组变异；Cancer Genome Workbench：一个用于观察和分析肿瘤样本中体细胞突变的计算平台，提高识别体细胞突变的精度；CaIntegrator：一个数据集成平台，研究人员通过其能够研究与临床相关的参数，如临床效果和基因组之间的关系。

3　TCGA在数据挖掘中的应用

目前，TCGA已经得到越来越多的关注，在Pubmed中搜索关键字TCGA，发现收录的文章中有887篇同TCGA相关。2011年发表的1篇有关卵巢癌[2]的文章中，TCGA研究团队利用卵巢癌的临床数据、外显子数据、拷贝数数据、mRNA和microRNA表达数据以及甲基化数据来挖掘癌症中潜在的基因组变异和表观遗传变异信息并找出可能的驱动基因。

3.1　突变分析

通过使用MUSIC[12]（mutation significance in cancer）和MutSig两种不同的方法对316个卵巢癌样本的癌症组织和正常组织的外显子进行数据分析。在303个样本中发现了与之前报道过的TP53突变相一致的结果，并在样本中发现了BRCA1（8%）和BRCA2（9%）生殖细胞突变，识别存在RNA剪接调控过程中常见的RB1、NF1、FAT3、CSMD3、GABRA6和CDK12显著突变基因。将该研究中发现的突变基因与体细胞突变目录数据库和人类孟德尔在线遗传数据库相比较，产生了477和211个匹配，包括BRAF、PIK3CA、KRAS和NRAS突变。这些突变有很强的转换活性，因此研究人员相信这些突变在卵巢癌中是罕见的但却是重要的驱动突变。

3.2　拷贝数分析

有研究显示，为了识别卵巢癌中差异具有统计学意义的拷贝数畸变区域及识别位于这些区域内的基因，使用GISTIC统计学方法对489个卵巢癌样本

的拷贝数数据进行分析。结果在卵巢癌中识别出了63个局部扩增区域和50个局部缺失区域，通过对扩增区域的进一步分析以及与其他数据资源的比较，在至少10%的病例中发现了22个靶向治疗基因，包括MECOM、MAPK1、CCNE1和KRAS[2]。

3.3 mRNA和microRNA表达与DNA甲基化分析

TCGA研究团队结合Agilent、Affymetrixhuex和Affymetrixu133a独立平台测得水平3的表达数据，采用CNMF聚类来识别亚型，并且结合临床数据预测每类亚型的预后效果。CNMF聚类分析mRNA表达数据识别了4个亚型，相同的分析方法用到另一个公共数据集[13]也产生了4个亚型，因此可以认为在卵巢癌中至少存在4个稳定的表达亚型。同样采用CNMF聚类分析microRNA表达数据识别了3个亚型，这3个亚型的生存时间存在显著性差异，即microRNA亚型-1的肿瘤患者的生存时间显著长于另外2个亚型。最后对所有样本的DNA甲基化数据进行一致聚类识别了4个亚型，与不同的年龄、BRCA失活和生存之间有显著相关性[2]。

3.4 信号通路分析

通过分析常见的包含一个或多个基因突变、拷贝数变化、基因表达变化的癌症通路，发现在患者中存在的RBI(67%)和PI3K/RAS(45%)通路均发生了失调。在一个PPI网络中使用HOTNET搜索畸变的子网络识别了多个已知通路，其中包括在卵巢癌样本中发生畸变的NOTCH信号通路(22%)[2]。

在上述分析卵巢癌TCGA数据挖掘的例子中，通过对TCGA提供的各类数据进行分析可以得到相应癌症可能潜在的致癌驱动基因或者抑癌基因，同时也可以用来验证实验结果和行相应的生存分析。

4 TCGA已发表的研究报道

TCGA的第1篇研究GBM报道发表于2008年Nature，拷贝数变异分析检测出在GBM中未报道的多个显著变异，如NF1和PARK2同源缺失、AKT3扩增；整合基因表达和拷贝数变异发现拷贝数变异区域内76%基因的表达模型与拷贝数相关。整合基因突变和DNA拷贝数扩增或缺失的结果发现致癌机制主要影响3个通路：RTK/RAS/PI3K signaling(88%)、TP53 signaling(87%)、RB1/CDK4 pathway(78%)[1]。后续研究发现GBM可分为4个亚型：经典型、原神经细胞型、神经元型、间质型[6]。2011年应用TCGA数据发表的另1篇论文，发现41个基因的突变与其表达发生变化[14]。为了研究拷贝数变异如何影响基因表达，Jörnsten等[15]在2011年开发了一个框架模型并在GBM的TCGA数据中获得验证。

TCGA研究团队2011年开始陆续报道了卵巢癌[16]、结直肠癌[17-18]、肺上皮细胞癌[19]及乳腺癌[20]的研究。这些报道提供了各种癌症特征基因的突变，染色体扩增和缺失以及受影响的信号通路。

近两年来已有超过几十篇应用TCGA数据发表的论文，涵盖了生物信息、统计学以及癌症分子生物研究，如microRNA分析[21-22]、甲基化分析[7,23]和拷贝数分析[24-26]。

5 小结

TCGA是一个以促进研究者对癌症的分子生物机制进一步了解为目标的宝贵资源。通过收集并整合分析临床数据和各种类型的基因组数据，使病理学科在以主观形态为诊断标准的应用技术和职能任务方面发生革命性改变，为敏感与耐药不同的癌症患者定制个性化医疗，为临床肿瘤研究者提供大量有价值的信息，为新的临床检测提供靶基因，为肿瘤预防和治疗提供明确的分子生物标记物，在可治愈期尽早发现肿瘤。

TCGA数据对结合分子生物标记物和临床数据，在生物统计或生物信息分析的研究初期或者实验结果的验证方面有很重要的作用。现在TCGA样本的主要来源是欧美人种，虽导致癌症的致病机制可能会存在地域差异，但在国内大规模癌症数据库建立之前，TCG数据提供重要的信息，并为未来研究打下重要基础。

参考文献

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

2 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma[J]. Nature, 2011, 474(7353):609-615.

3 Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer[J]. Nature, 2012, 487(7407):330-337.

4 Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma[J]. Proc Natl Acad Sci U S A, 2007, 104(50):20007-20012.

5 Cope L, Wu RC, Shih IeM, et al. High level of chromosomal aberration in ovarian cancer genome correlates with poor clinical outcome[J]. Gynecol Oncol, 2013, 128(3):500-505.

6 Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1[J]. Cancer Cell, 2010, 17(1):98-110.

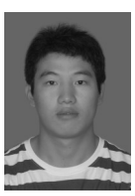
7 Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma[J]. Cancer Cell, 2010, 17(5):510-522.

8 Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers[J]. Nature, 2012, 489(7417):519-525.

9 Cancer Genome Atlas Research Network. Comprehensive molecular

portraits of human breast tumours[J]. Nature, 2012, 490(7418):61–70.
10 Bolton KL, Chenevix–Trench G, Goh C, et al. Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer[J]. JAMA, 2012, 307(4):382–390.
11 Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma[J]. Nature, 2011, 471(7339):467–472.
12 Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes[J]. Genome Res, 2012, 22(8):1589–1598.
13 Tothill RW, Tinker AV, George J, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome[J]. Clin Cancer Res, 2008, 14(16):5198–5208.
14 Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival [J]. Cancer Res, 2011, 71(13):4550–4561.
15 Jörnsten R, Abenius T, Kling T, et al. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma[J]. Mol Syst Biol, 2011, 7:486.
16 Dutta P, Bui T, Bauckman KA, et al. EVI1 splice variants modulate functional responses in ovarian cancer cells[J]. Mol Oncol, 2013, 7(3): 647–668.
17 Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data[J]. Proc Natl Acad Sci U S A, 2013, 110(11):4245–4250.
18 Li Y, Zhang L, Ball RL, et al. Comparative analysis of somatic copy–number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots[J]. Hum Mol Genet, 2012, 21(22):4957–4965.
19 Sproul D, Kitchen RR, Nestor CE, et al. Tissue of origin determines cancer–associated CpG island promoter hypermethylation patterns [J]. Genome Biol, 2012, 13(10):R84.
20 Wang C, Pécot T, Zynger DL, et al. Identifying survival associated morphological features of triple negative breast cancer using multiple datasets[J]. J Am Med Inform Assoc, 2013, 20(4):680–687.
21 Creighton CJ, Hernandez–Herrera A, Jacobsen A, et al. Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high–grade serous ovarian carcinoma[J]. PLoS One, 2012, 7(3):e34546.
22 Genovese G, Ergun A, Shukla SA, et al. microRNA regulatory network inference identifies miR–34a as a novel regulator of TGF–β signaling in glioblastoma[J]. Cancer Discov, 2012, 2(8):736–749.
23 Andreopoulos B, Anastassiou D. Integrated analysis reveals hsa–miR–142 as a representative of a lymphocyte–specific gene expression and methylation signature[J]. Cancer Inform, 2012, 11:61–75.
24 Chen H, Xing H, Zhang NR. Estimation of parent specific DNA copy number in tumors using high–density genotyping arrays[J]. PLoS Comput Biol, 2011, 7(1):e1001060.
25 Standfuss C, Pospisil H, Klein A. SNP microarray analyses reveal copy number alterations and progressive genome reorganization during tumor development in SVT/t driven mice breast cancer[J]. BMC Cancer, 2012, 12:380.
26 Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing[J]. Genome Res, 2012, 22(3):568–576.

（2013–04–29 收稿）
（2013–11–05 修回）
（本文编辑：张侃）



作者简介
邓祯祥　在读硕士研究生。研究方向为生物信息学。
E–mail：ghost_fx135@163.com

·读者·作者·编者·

科技论文材料与方法的撰写要求

材料（资料或对象）与方法让读者在同等条件下可重复出该结果，是科技论文的重要特征。其撰写应注意以下几个方面：1）紧扣主题；2）科学真实；3）典型而新颖；4）符合伦理学原则。

按重要性程度、时间排序或实验过程书写。包括：1）实验动物：品系名称，级别，雌雄，月龄或周龄，体重，合格证号，提供单位和生产许可证号，饲养及环境；2）药品及试剂：主要试剂的名称，级别，来源，批号和配置方法；3）主要仪器：主要仪器名称，生产厂家；4）方法：随机分组方法，给药剂量与途径，动物处理，指标测定方法，数据处理方法。

信息不全，主次不分，顺序颠倒，写入一般常规试剂是撰写禁忌。常见的问题：1）组别名称随心所欲，全文不统一，没有直呼其名，用A、B组代号，读者阅读时要反复前后对应，文字表述不简明，术语使用不当；小标题太多冗长；2）过繁：避免教科书式的撰写方法；3）过简：介绍的内容无法使读者重复该实验；4）不规范：应选用规范的单位和书写符号。