生物技术通报
·特约综述· BIOTECHNOLOGY BULLETIN 2015, 31(11):89-101

肿瘤生物信息学数据库

杨健 蔡浩洋
（四川大学生命科学学院 生长代谢衰老研究中心，成都 610065）

摘 要： 恶性肿瘤已经成为严重危害人类健康的主要疾病之一。近年来，高通量检测技术迅速发展，成为肿瘤研究的重要手段之一，使得与肿瘤相关的组学数据迅速积累。这些数据对于研究肿瘤的发生发展机制具有重要意义。对海量生物学数据的管理和挖掘已经成为癌症研究的基础与重要方向。主要介绍人类肿瘤研究中经常用到的生物信息学数据库，包括综合性数据库、基因组、转录组、蛋白组、表观遗传组数据库等。在总结国内外肿瘤数据库发展现状的基础上，讨论了目前数据库开发存在的问题，旨为现有的研究提供帮助。

关键词： 生物信息学；数据库；生物学大数据；肿瘤；组学数据
DOI：10.13560/j.cnki.biotech.bull.1985.2015.11.010

The Cancer-related Bioinformatics Databases

Yang Jian Cai Haoyang
（*Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu* 610065）

Abstract: Malignant tumor has become one of the major diseases that takes seriously risks to human health. In recent years, the rapid development of high-throughput detection technology has become an important means in cancer research. In this way, the cancer genomics data accumulated rapidly. These data is important for the research of mechanisms of tumor initiation and development. Massive biological data management and mining have become the foundation and an important field in cancer research. This article describes the frequently used bioinformatics databases of human tumors, including comprehensive databases, databases of genomics, transcriptomics, proteomics, epigenetics, etc. Here we sum up the status quo of database development in China and abroad, and discuss the existing problems to assist current research.

Key words: bioinformatics；database；biological big data；tumor；OMICs data

随着人们生活方式和环境的改变，恶性肿瘤已经成为疾病三大死亡病因之一，占全球每年死亡人数的15%以上[1]。我国在人口老龄化进程不断加快的情况下，恶性肿瘤的发病率及死亡率一直呈上升趋势，估计每年新发肿瘤病例超过300万例，严重威胁人们的健康和生命。当前已知的肿瘤类型接近200种，已发现的与肿瘤相关的原癌基因或抑癌基因超过500个。世界各国都投入了大量的人力与经费用于癌症的基础研究与诊断治疗，美国国立癌症研究所（National Cancer Institute，NCI）2015年度的研究预算达到49.5亿美元。近年来，随着高通量检测和分析技术的发展与普及，与肿瘤相关的生物学数据呈指数级增长，利用数据挖掘的方法从海量数据中找出驱动基因与突变有助于阐明肿瘤发生的分子机制。然而，这些数据的管理和分析成为研究人员面临的一大挑战。不同的检测技术产生了复杂的、不同结构的生物学数据，这些原始数据必须经过标准化、结构化、添加注释及统计分析才能成

收稿日期：2015-10-13
作者简介：杨健，男，博士研究生，研究方向：生物信息学；E-mail：stardustcx@163.com
通讯作者：蔡浩洋，男，副研究员，研究方向：生物信息学；E-mail：haoyang.cai@scu.edu.cn

为有价值的信息。同时，在高通量测序技术的价格不断降低的情况下，当前的研究对肿瘤样本的测序深度越来越高，对单个肿瘤样本的测序可以产生超过 150 GB 的数据，这对海量数据的存储与利用也提出了新的挑战。因此，生物信息学数据库的构建成为肿瘤研究的一个重要方向，也是信息处理的基础，通过大量肿瘤样本的数据分析可以得到单个实验难以获得的规律性结论。利用生物信息技术收集、存储、分析并共享与肿瘤相关的生物学数据正逐渐成为癌症研究中必不可少的技术手段，高质量的肿瘤数据库将为研究人员提供便捷的数据分析服务与数据共享平台，为揭示癌症的发生发展机制奠定基础，如图 1 所示。

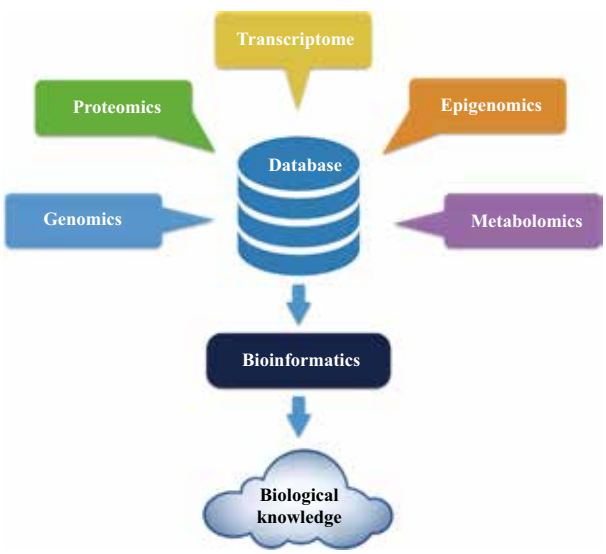

图 1 利用生物信息学整合多种类型的数据并构建数据库

现今生物学数据库因其重要作用已获得了广泛的关注与研究。伴随着生物信息学这一交叉学科的快速发展，目前已经产生了大量的数据库，并在生物学的各个研究领域产生了广泛的影响。例如，国际上三大核酸与蛋白质数据库，包括美国国家生物信息中心的 GenBank[2]、欧洲生物信息研究所的 EMBL[3] 以及日本生物信息学中心的 DDBJ[4]，是目前最具有影响力的生物全领域数据库，为研究人员提供了获取与共享数据的平台，极大地促进了包括肿瘤在内的相关领域研究。从 1994 年开始，Nucleic Acids Research 杂志每年都出版一期数据库专辑，收录重要的生物学数据库。Bioinformatics 杂志也设立了数据库专栏，介绍各种生物信息学数据库。另外，“精准医疗计划”的提出为生物信息学数据库的发展带来了新的契机。“精准医疗”旨在根据个体的差异为每一个病人制定个性化的预防和治疗方案，达到精确用药的目的。该计划的短期目标主要与恶性肿瘤相关，根据肿瘤的基因变异研发靶向药物，然后对病人进行临床基因诊断，按个体基因变异的情况使用不同的靶向药物。该项目一方面需要构建大量人群的肿瘤变异数据库，另一方面需要开发新的数据分析算法进行海量数据的挖掘与整合。表 1 列出了主要的肿瘤生物信息学数据库网站。本文将这些数据库按照研究领域或者数据类型进行分类并作介绍。

1 综合性肿瘤数据库

早期的基因芯片和近年来广泛应用的二代测序技术产生了大量的生物学数据，包括 DNA 拷贝数变化（Copy number aberration，CNA）、基因突变、表达谱以及全基因组测序数据。这些数据中蕴含着潜在的有价值的生物学信息，可能帮助人们更加深入地理解癌症，因此对海量数据的存储和分析也具有重要意义。目前已有多个机构致力于这些数据的收集、存储以及分析。在这一部分中，我们将对几个重要的综合数据库作简要介绍。

The Cancer Genome Atlas（TCGA，https://tcga-data.nci.nih.gov/tcga）[5] 由美国国立癌症研究所和国家人类基因组研究所（National Human Genome Research Institute，NHGRI）资助，关注与癌症的发生和发展相关的分子突变图谱。根据癌症的发病率，TCGA 选取了 34 种癌症及其对应的正常组织样本进行比较研究，每种肿瘤都有大量的样本重复以进行癌症变异数据的深度挖掘。TCGA 拥有基因组测序中心（Genome Sequencing Centers，GSCs）、基因组数据分析中心（Genome Data Analysis Centers，GDACs）以及基因组描述中心（Genome Characterization Centers，GCCs）等，能够对样本进行外显子组和基因组测序及分析，提供包括基因组拷贝数变化、表观遗传、基因表达谱、miRNA 等数据。TCGA 的数据访问权限分为两种，公开的数据

表 1 主要的肿瘤生物信息学数据库

Database	Description	Link	Refs.
ArrayExpress	Microarray gene expression data	https://www.ebi.ac.uk/arrayexpress/	[31]
arrayMap	Reference resource for genomic copy number imbalances	http://www.arraymap.org	[18]
BioMuta	Integrated sequence feature database	https://hive.biochemistry.gwu.edu/tools/biomuta/	[22]
Cancer3D	Cancer mutations and protein structures	http://cancer3d.org/	[41]
CancerDR	Cancer drug resistance database	http://crdd.osdd.net/raghava/cancerdr	[49]
CancerPPD	Anticancer peptides and proteins	http://crdd.osdd.net/raghava/cancerppd/	[40]
CancerResource	Cancer-relevant proteins and compound interactions	http://bioinformatics.charite.de/cancerresource	[48]
canEvolve	Web portal for integrative oncogenomics	http://www.canevolve.org	[17]
CanGEM	Cancer GEnome Mine	http://www.cangem.org/	[20]
CanProVar	Cancer Proteome Variation Database	http://bioinfo.vanderbilt.edu/canprovar/	[39]
canSAR	Cancer research and drug discovery knowledgebase	http://cansar.icr.ac.uk	[47]
CaSNP	Copy number alterations of cancer genome from SNP array data	http://cistrome.dfci.harvard.edu/CaSNP/	[19]
cBioPortal	cBioPortal for Cancer Genomics	http://www.cbioportal.org	[14]
CGAP	Cancer Genome Anatomy Project	http://cgap.nci.nih.gov	[12]
CGHub	Cancer Genomics Hub	https://cghub.ucsc.edu	[10]
CGP	Cancer Genome Project	http://www.sanger.ac.uk/research/projects/cancergenome	[21]
CGWB	Cancer Genome Work Bench	https://cgwb.nci.nih.gov/	[16]
ChiTaRS	Chimeric transcripts and RNA-sequencing data	http://chitars.bioinfo.cnio.es/	[36]
COSMIC	Catalogue Of Somatic Mutations In Cancer	http://cancer.sanger.ac.uk/cosmic	[13]
CPTAC	Clinical Proteomic Tumor Analysis Consortium	http://proteomics.cancer.gov/programs/cptacnetwork	[37]
dbDEPC	Differentially expressed proteins in human cancers	http://lifecenter.sgst.cn/dbdepc/	[38]
DiseaseMeth	Human disease methylation database	http://bioinfo.hrbmu.edu.cn/diseasemeth	[28]
DriverDB	Exome sequencing database for cancer driver gene	http://driverdb.ym.edu.tw/DriverDB/	[42]
EGA	European Genome-phenome Archiv	https://ega.crg.eu	[9]
GDSC	Genomics of Drug Sensitivity in Cancer	http://www.cancerRxgene.org	[46]
GEO	Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/	[27]
ICGC	International Cancer Genome Consortium	https://icgc.org/	[11]
MENT	Methylation and expression database of normal and tumor tissues	http://mgrc.kribb.re.kr:8080/MENT	[26]
MethDB	Common resource for epigenetic phenomenon	http://www.methdb.de/	[29]
MethHC	DNA methylation and gene expression in human cancer	http://MethHC.mbc.nctu.edu.tw	[25]
MethyCancer	Human DNA Methylation and Cancer	http://methycancer.psych.ac.cn/	[24]
miRCancer	MicroRNA-cancer association database	http://mircancer.ecu.edu/	[34]
NCG	Network of Cancer Genes	http://bio.ifom-ieo-campus.it/ncg	[43]
NGSmethDB	Next-generation sequencing single-cytosine-resolution DNA methylation	http://bioinfo2.ugr.es/NGSmethDB	[30]
Oncomine	Cancer microarray database	https://www.oncomine.org/	[32]
OncomiRDB	Experimentally verified oncogenic and tumor-suppressive microRNAs	http://bioinfo.au.tsinghua.edu.cn/member/jgu/oncomirdb/	[33]
Platinum	Mutations on structurally defined protein-ligand complexes	http://structure.bioc.cam.ac.uk/platinum	[50]
SomamiR	Somatic mutations impacting microRNA function in cancer	http://compbio.uthsc.edu/SomamiR/	[35]
TCGA	The Cancer Genome Atlas	https://cancergenome.nih.gov/	[5]
TP53MULTLoad	TP53 mutation database	http://p53.fr	[44]
UCSC Genome Browser	UCSC Cancer Genomics Browser	https://genome-cancer.ucsc.edu/	[15]
UMD TP53	TP53 database	http://www.umd.be:2072/	[45]

包括了临床和人口数据、基因表达数据、CNA 数据、表观数据等，而需要授权的数据主要是一些个人特有数据，如原始的测序数据、单核苷酸多态性（Single nucleotide polymorphism，SNP）数据以及 VCF 文件等。现在来源于 TCGA 的测序原始数据存储在癌症基因组中心（Cancer Genomics Hub，CGHub），而序列分析数据则可在 TCGA 的数据中心（TCGA Data Portal）下载。随着 TCGA 的数据的增长，目前有许多基于 TCGA 的研究，包括对癌症分类的探索[6]、癌症的突变标志物研究[7]、药物靶点研究[8]等。

European Genome-phenome Archive（EGA，https://ega.crg.eu）[9]收集了多种测序以及分型数据，如基因组关联分析、分子诊断以及各种目的的测序数据。目前，该数据库已收集了超过 800 项研究的数据，数据量也达到了 1.7 PB 之巨，其中约 60% 都与肿瘤相关。这些数据的访问受到严格的控制，用户可通过浏览或搜索找到需要的数据项，但是下载则需要向指定的数据访问控制机构申请。为了方便用户下载数据，EGA 还开发了基于 java 的下载工具。

Cancer Genomics Hub（CGHub，https：//cghub.ucsc.edu）[10]收集来自 3 个国家癌症协会项目的基因组信息：包括癌症基因图谱项目（TCGA）、癌症细胞系百科全书（Cancer Cell Line Encyclopedia，CCLE）以及为有效治疗进行的治疗方案研究项目（Therapeutically Applicable Research to Generate Effective Treatments，TARGET）。CGHub 收集了来自 25 种不同类型癌症的测序数据，以 BAM 文件形式存储，目前的数据量已经超过 2 PB，并且以每周约 50 TB 的速率增长。CGHub 支持对癌症测序数据的浏览和受控制的访问，对于来自于 CCLE 的数据是完全公开，而另外两个项目的数据则是需要具有授权才可以下载。CGHub 还提供了一款数据下载软件 GeneTorrent，并可在多个平台上使用。GCHub 提供的原始数据对于整合和共享癌症相关数据具有重要作用，对癌症基础研究具有极大的促进作用。

International Cancer Genome Consortium（ICGC，https://icgc.org）[11]是由多个国家多个研究机构组成的癌症研究团体，包含来自亚洲、澳大利亚、欧洲、北美和南美的 88 个研究团队。其目标是获取包括胆道癌、膀胱癌、血癌等多达 50 种肿瘤及其亚型的基因组、转录组和表观遗传的全部信息，并以最快的速度和最少的限制将这些数据提供给整个科研团体，促进癌症的机理和治疗研究。到目前为止，ICGC（release 19）提供了 12 979 个癌症基因组的数据，包含了 16 459 160 个简单的体细胞突变，涉及到 57 543 个基因。用户可通过 ICGC 的数据中心搜索感兴趣的数据，并利用网站提供的工具下载。与 TCGA 一样，ICGC 的测序原始数据和涉及个体信息的数据如生殖细胞突变需要得到 ICGC 的授权。

Cancer Genome Anatomy Project（CGAP，http://cgap.nci.nih.gov）[12]是 NCI 的一个研究项目，主要收集了正常组织、前癌组织以及癌细胞的基因表达水平，以期改善癌症的检测、诊断以及病患治疗。CGAP 网站主要提供了 cDNA 克隆、文库、基因表达、SNP 以及基因组变异信息，并且提供了一系列的分析工具，可实现对一个或多个基因、文库的搜索，发掘基因组和基因中的 SNP，获取文库中差异表达的基因，比较两个文库的差异表达基因，分析基因参与的通路，并且将这些信息可视化。

Catalogue Of Somatic Mutations In Cancer（COSMIC，http://cancer.sanger.ac.uk/cosmic）[13]是世界上最大最全面的有关肿瘤的体细胞突变以及其影响的资源。主要提供多种肿瘤细胞基因组中的 CNA、甲基化、基因融合、SNP 及基因表达信息等。最新的版本（v74，Aug 2015）中描述了超过 100 万个肿瘤样本中的 2 002 811 个点突变，涉及到大部分的人类基因。除此之外，COSMIC 中还提供了超过 6×10^6 个非编码点突变、10 534 个基因融合、61 299 个基因组重排、695 504 个拷贝数异常、60 119 787 个表达异常的详细信息，并且这些信息在基因组和编码基因中都进行了注释，进而与疾病和变异类型关联起来。COSMIC 给癌症用户提供了十分重要而全面的肿瘤基因组变异信息。

cBioPortal for Cancer Genomics（cBioPortal，http：//www.cbioportal.org）[14]是一个癌症基因组数据探索、可视化及分析平台，提供 CNA、基因突变信息，并根据数据完整程度提供包括 mRNA 丰度、蛋白丰度以及 DNA 甲基化水平等信息。目前，该平台收集了 105 个肿瘤研究中的 10 473 个样本数据。用户可选取特定的样本，形成数据集，并定义一系

列感兴趣的基因，分析这些基因在样本中的 CNA 的出现频率和基因突变频率。在结果中除了汇总信息外，还会针对每个基因给出 CNA 和突变在样本中的分布、突变位点和频率、共表达基因以及生存曲线等；而对于用户提供的基因列表，还可生成互作网络并提供已知的相互作用的药物。cBioPortal 在发现肿瘤相关突变、分析基因的生物学功能以及药物选择等方面的研究中具有重要推进作用。

UCSC Cancer Genomics Browser（UCSC 癌症基因组浏览器，https://genome-cancer.ucsc.edu）[15] 保存癌症基因组及临床数据，并提供了数据可视化和分析的工具。该平台中收集了样本的多种信息，包括基因表达水平、CNA、通路信息等。在 UCSC 的癌症基因组浏览器中，研究人员可以对一个或几个实验中的样本及其关联的临床信息进行研究，可实现不同样本以及癌症类型之间的比较，分析基因组变异与表型之间的相关性。目前，该平台收集了来自 TCGA、CLLE、Connectivity Map 以及 TARGET 的 575 个数据集，包含了超过 22 700 个样本的数据。

除了 UCSC 癌症基因组浏览器外，还有多个综合性分析平台基于 TCGA 等数据库的基因组信息综合性地分析基因组的变异与临床数据和基因表达谱的关联性。例如，癌症基因组工作平台（Cancer Genome Work Bench，CGWB，https://cgwb.nci.nih.gov）[16] 提供了一系列工具来挖掘、整合以及可视化 TCGA 等数据库中的基因组和临床数据，用户可快速地比较患者临床信息与基因组的变异及甲基化等；而 canEvolve[17] 数据库收集了来自 90 个研究的超过 10 000 位病人的数据，为用户提供两种水平的分析数据：其一是 mRNA、miRNA、蛋白的表达水平、基因组变异、蛋白相互作用数据；其二是综合分析数据，如基因表达与 miRNA 表达、基因表达与 CNA 之间的关联、基因富集、网络分析及生存分析等。

2　肿瘤基因组数据库

一般而言，肿瘤细胞的基因组中都存在着大量的变异，主要包括染色体结构的变异、CNA、基因融合以及 SNP 等。对肿瘤的基因组变异信息的收集和整理可促进研究者对肿瘤发生发展的认识。以下介绍一些收集和整理这类信息的数据库。

arrayMap（http://www.arraymap.org）[18] 是由苏黎世大学分子生命科学研究所构建的，提供预处理过的肿瘤基因组芯片数据以及 CNA 图谱。目前，最新版的 arrayMAP（Jan 2015）包含了约 250 种癌症中获得的 64 000 多个基因组芯片数据集。用户可通过关键字搜索自己感兴趣的样本或者搜索特定文献中的样本，并在此基础上分析感兴趣的基因或基因组片段上的 CNA；用户还可以选择两个样本来比较二者的 CNA 的差异。

CaSNP（http://cistrome.dfci.harvard.edu/CaSNP/）[19] 数据库收集了来源于 SNP 芯片的 CNA 数据，并提供查询服务。CaSNP 从 34 种肿瘤的 104 项研究中获取了约 11 500 张 SNP 芯片，基于这些芯片整理出了肿瘤基因组中的 CNA。用户可以搜索基因或者感兴趣的基因组区域，CaSNP 将返回各项研究中的染色体区域得失频率及平均的拷贝数，并提供下载链接或在 UCSC 基因组浏览器中可视化。

CanGEM（http://www.cangem.org/）[20] 是一个公开的存储肿瘤样本的临床和芯片数据的数据库。它主要利用 arrayCGH 芯片来发掘基因的拷贝数变异。用户可以通过关键字搜索特定类型的肿瘤样本或者发掘特定基因发生拷贝数变化的样本构建个性化的数据集情况，然后基于这些样本计算变异发生的频率。CanGEM 还提供原始数据下载服务，用户可以对感兴趣的数据集进行深入的分析。

Cancer Genome Project（CGP，http://www.sanger.ac.uk/research/projects/cancergenome）[21] 是 The Wellcome Trust Sanger Institute 下属的一个项目，主要目标是利用人类基因组序列和高通量的突变检测技术识别体细胞突变，进而发现人类肿瘤发生过程中重要的基因。该项目提供了肿瘤中的 CNA 及基因型信息，同时也提供了一些识别突变、CNA 的软件，如 BioView、GRAFT 等。

BioMuta（https://hive.biochemistry.gwu.edu/tools/biomuta/）[22] 数据库存储了癌症细胞中基因的非同义单核苷酸变异，这些突变会影响基因的正常功能。BioMuta 中的数据来源于 COSMIC、ClinVar、UniProtKB 以及一些文献中，最新版本（v2.0）中包含了 26 种癌症中的 322 922 个 SNP。用户可搜索感

兴趣的基因，获得该基因在癌细胞中的突变位点及其分布频率。

3 肿瘤 DNA 甲基化数据库

DNA 甲基化修饰是表观遗传的一个重要形式，可调控基因的转录水平，对于维持细胞正常功能具有重要作用。DNA 甲基化模式改变可能导致癌症的发生，一些抑癌基因的高甲基化导致基因表达量降低引起癌症发生，也可能导致一些抑癌的 miRNA 转录水平下降同样会引发癌症[23]。目前也有部分数据库收集和整理肿瘤中的甲基化模式，并可与基因的表达水平比较。以下对这些数据库作简要介绍。

MethyCancer（http://methycancer.psych.ac.cn/）[24] 数据库收集了肿瘤中的 DNA 甲基化、重复序列、癌症相关基因、突变、CpG 岛以及肿瘤相关信息。用户可搜索感兴趣的基因或基因组区域，获得相关的甲基化、重复序列、基因以及 CpG 岛等信息。另外，网站还提供了一个可视化工具 MethyView，可在一个窗口中查看一个基因组区域内上述元素的相互关系。MethyCancer 可作为分析人类基因组中 IpG 岛的分布、启动子区 DNA 甲基化形式的平台，能帮助研究人员识别肿瘤中受 DNA 甲基化影响的基因，发掘潜在的表观遗传靶点。

MethHC（http://MethHC.mbc.nctu.edu.tw）[25] 系统性地整理了来自 TCGA 的肿瘤基因组甲基化、基因表达、miRNA 甲基化、miRNA 表达以及甲基化和基因表达水平的关联关系。目前，数据库收集了 18 种人类肿瘤的超过 6 000 个样本、6 548 张芯片以及 12 567 个 RNA 测序数据。MethHC 提供了基因及其上下游的多个区域的甲基化水平、甲基化和基因表达关系、基于甲基化位点的癌症分层聚类以及每种癌症中高甲基化和低甲基化的前 250 个基因列表。

MENT（http://mgrc.kribb.re.kr：8080/MENT）[26] 数据库收集和整合了来自 Gene Expression Omnibus（GEO）[27] 和 TCGA 的 DNA 甲基化、基因表达水平数据，同时将 DNA 甲基化和基因表达水平关联起来。MENT 提供了友好的界面，用户可通过基因搜索或数据集搜索来发掘差异甲基化。基因搜索返回目标基因在哪些条件下发生差异甲基化，而数据集搜索则返回一定条件下所有差异甲基化的基因。两种搜索都可以通过设定方向、差异甲基化值和 p 值对结果进行筛选。

DiseaseMeth（http://bioinfo.hrbmu.edu.cn/diseasemeth）[28] 收集和整理了多种人类疾病中的甲基化数据，包括癌症、神经发育和退行性疾病、自身免疫疾病等。目前，DiseaseMeth 整合了 175 个高通量数据集的数据，用户可以多种方式搜索自己感兴趣的内容，如 gene ID、疾病名称等，还可以比较疾病与疾病之间、基因与基因之间以及疾病与基因之间的甲基化关系。除此之外，该数据库还支持甲基化数据下载，研究者可将数据整合到自己的研究中。

除了上述针对癌症基因组甲基化的数据库外，还有一些数据库搜集和整理更为广泛的甲基化数据，如 MethDB 和 NGSmethDB。MethDB（http://www.methdb.de/）[29] 是较早的 DNA 甲基化数据库，主要集中于环境因子对甲基化的影响；而 NGSmethDB（http://bioinfo2.ugr.es/NGSmethDB）[30] 基于高通量测序数据，最近更新中还包含了 SNP 信息，以便后续分析。

4 肿瘤转录组数据库

肿瘤细胞具有较强的生长和繁殖能力，生命活动旺盛，因此与正常细胞相比，基因的转录水平和模式也存在较大的差异。转录组是特定条件下细胞内全部转录物的总和，包括多种类型的 RNA，而通常我们更关心的是编码基因的产物 mRNA 以及近年来比较热门的非编码 RNA，如小 RNA（miRNA）及长非编码 RNA（lncRNA）。我们将针对一些与肿瘤相关的转录组数据库作介绍。

Gene Expression Omnibus（GEO，http://www.ncbi.nlm.nih.gov/geo/）[27] 是美国国家生物技术中心（NCBI）的一个子数据库，是一个免费且公开的生物数据存储平台，主要存储包括基因芯片、第二代测序以及其它高通量的功能基因组学数据。GEO 将提交的原始数据分为 3 个层次：平台、系列和样本。这些原始数据又进一步组成不同的数据集，并在 GEO 生成基因表达谱。用户可通过搜索获得感兴趣的数据集，利用 GEO 提供的 *t* 检验或聚类发掘感兴趣的基因及其表达谱，还可进一步搜索与之表达谱相似的基因。GEO 的原始数据符合 MIAME（Mini-mum

information about a microarray experiment）数据标准（http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html），提供了包括原始数据、处理后数据、样本信息、实验设计方案、芯片注释信息以及实验和数据处理流程等信息。GEO 还支持数据下载，用户可将感兴趣的样本或数据集下载下来，用于自己的研究。

ArrayExpress（https://www.ebi.ac.uk/arrayexpress/）[31]是欧洲生物信息协会（EMBL-EBI）下属的功能基因组数据库，收集整理基于芯片和测序的基因组数据。其数据一部分是直接提交到 ArrayExpress，另一部分是从 GEO 导入的，目前收集了 7 000 个测序研究以及 42 000 个基于芯片的研究中的超过 1.5×10^6 个样本数据。ArrayExpress 的数据格式符合 MIAME 和 Minimum Information about Sequencing Experiment（MINSEQE，http://www.fged.org/projects/minseqe/）标准，包含了详细的样本和实验信息，用户可通过关键字搜索感兴趣的样本。网站还提供了统一的数据提交工具 Annotare，方便用户提交数据。

Oncomine（https://www.oncomine.org/）[32]致力于收集、标准化并分析肿瘤样本的基因表达谱芯片数据，为生物医药领域的研究者提供肿瘤转录组数据。目前，Oncomine 已经收集了来自 715 个数据集的 86 733 个样本，用于识别肿瘤基因组中失调的基因、通路和调控网络。Oncomine 可提供基因在肿瘤样本和正常样本间、肿瘤样本和肿瘤样本间、正常样本和正常样本间的差异表达、基因表达谱、共表达基因等信息。用户可选择一组样本，如肿瘤类型、赖药性、组织类型等，获得显著高表达和低表达的基因，同时可联合不同样本，分析共同显著差异的基因，帮助用户从大量的差异表达基因中挑选在多样本中都显著差异的基因。对于获得的一系列感兴趣基因，用户还可进行筛选，作富集分析，并可视化受影响的通路等。需要注意的是，Oncomine 是一个面向非盈利团体的受密码保护的数据分析平台，因此用户需要注册才可使用其服务。

OncomiRDB（http://bioinfo.au.tsinghua.edu.cn/member/jgu/oncomirdb/）[33]的目标是收集和注释通过实验验证的对癌症具有促进或抑制作用的 miRNA。该数据库中的 miRNA 至少符合以下一条：调控至少一种与肿瘤相关的表型或细胞过程，如增殖、凋亡、迁移、侵袭、衰老和细胞周期调节；或者有实验证据证明直接调控至少一个原癌基因或抑癌基因。该数据库的所有数据是通过人工收集和整理，目前包含 2 259 条调控关系，涉及到 328 个 miRNA 以及 829 个靶基因。用户可直接搜索某种 miRNA，也可以通过模糊搜索得到 miRNA 以及靶基因列表，还可以限定组织、肿瘤类型、以及 miRNA 功能分类，获得特定细胞类型中的特定类型的 miRNA 及其靶基因，结果可以以列表和图形方式给出。用户还可直接下载 OncomiRDB 中提供的全部调控关系，这些高可信度的 miRNA-靶基因关系是 miRNA 功能研究的重要资源。

miRCancer（http://mircancer.ecu.edu/）[34]提供了较为全面的 miRNA 集合以及它们在多种肿瘤中的表达情况。miRCancer 中的数据获取过程如下：首先利用文本挖掘方式从 PubMed 中搜索与 miRNA 相关的文章，并获取 miRNA 的表达情况，再人工验证，以提高数据的准确度。目前，数据库中已经搜集了 44 000 余种 miRNA，包括 176 种肿瘤中的 3 700 多个肿瘤相关的 miRNA。用户可直接搜索某种 miRNA，结果页面将给出其在不同肿瘤样本中的表达情况，以及相关文献；也可以限定 miRNA 和肿瘤类型，结果页只列出该肿瘤中的相关研究。另外，数据库还提供了两种分析工具，可对不同物种或肿瘤中的 miRNA 进行聚类分析或卡方检验。

SomamiR（http://compbio.uthsc.edu/SomamiR/）[35]数据库主要收集 miRNA 及其靶序列上的突变，miRNA 上的突变会改变其识别的靶序列，而靶序列上的突变则可能导致 miRNA 结合能力减弱甚至不能结合。SomamiR 数据库提供了 miRNA 序列上的体细胞突变、利用 CLASH、PAR-CLIP、HITS-CLIP 实验获得的靶序列中的体细胞突变、预测的靶序列中的体细胞突变等。另外，数据库还提供了存在 miRNA 靶序列体细胞突变且肿瘤相关的基因及其参与的通路，受影响的通路可在 KEGG 通路中展示。数据库中的所有内容都可以免费下载。

ChiTaRS（http://chitars.bioinfo.cnio.es/）[36]数据库记录了来自人类、小鼠、果蝇等 8 个物种中的嵌合转录本，同时收集了 1 400 个人类癌症基因组序列断点以及与之对应的嵌合转录本的表达水平数据。

用户可搜索某种疾病中染色体上的断点及涉及的基因，也可提供一段 DNA 序列检查是否存在断点，还可以比较不同物种中的断点。这些断点信息以及在各物种中的比较可帮助我们理解嵌合转录本的进化以及其在肿瘤发展中的作用。

5 肿瘤蛋白组数据库

蛋白是生命活动的主要承担者，细胞的各项生命活动都与蛋白有着密切的联系，因此细胞内蛋白的种类、数目和形式对细胞功能起着重要的作用。蛋白结构变异、蛋白修饰的改变以及蛋白含量的变化等导致细胞的生长和代谢变化是肿瘤发生的重要因素。对于肿瘤细胞中蛋白的种类、含量以及修饰的记录对于解析肿瘤的表型具有重要的价值。我们将介绍一些与肿瘤细胞中蛋白组相关的数据库。

Clinical Proteomic Tumor Analysis Consortium（CPTAC，http://proteomics.cancer.gov/programs/cptacnetwork）[37]是由 NCI 启动的一项旨在识别和描述肿瘤组织和正常组织中的全部蛋白，整合基因组和蛋白组的数据，发掘可作为肿瘤生物标记的候选蛋白并排序，最终在一组相关样本中验证。CPTAC 由蛋白组特征研究中心（PCCs）、数据整合中心以及资源中心组成。PCCs 通过质谱测定肿瘤组织中的蛋白类型、含量、蛋白修饰等，数据整合中心负责将 PCCs 的数据整理并公开，资源中心负责整理和发放样品以及实验的参考材料等。目前，CPTAC 已发表近 20 项蛋白组研究，主要是直结肠癌、卵巢癌和乳癌中的蛋白组研究以及一些对于实验条件和技术的测试性研究。目前 CPTAC 提供的数据还较少，还处于起步阶段，但是与 TCGA 这类大型的基因组研究项目类似，未来 CPTAC 可能成为蛋白组研究的重要资源，可提供高质量的癌症蛋白组数据。

dbDEPC（http://lifecenter.sgst.cn/dbdepc/）[38] 是一个专门收集肿瘤样本中出现的差异表达蛋白的数据库。最新版本（v2.0）收集了来自 241 篇文献的 331 项质谱数据，在 20 种肿瘤中发现了 4 029 个差异表达蛋白。用户可通过关键字或蛋白序列搜索特定蛋白，获得该蛋白发生差异表达的样本及其表达谱；也可以浏览特定样本或特定质谱数据中差异表达的蛋白。

Cancer Proteome Variation Database（CanProVar，http://bioinfo.vanderbilt.edu/canprovar/）[39] 是一个存储人类蛋白组中的体细胞和生殖细胞发生的单个氨基酸突变，特别是那些与肿瘤发生和发展有关系的氨基酸突变。CanProVar 中的数据主要来源于 TCGA、COSMIC、OMIM、HPI 等数据库以及一些研究文献。目前，该数据库包含了 11 445 个与肿瘤相关的蛋白突变位点以及超过 40 000 个与癌症无关的蛋白突变位点。用户可在网站中搜索特定蛋白或者某种肿瘤，获取蛋白的突变情况，在结果页面会给出蛋白的基本信息、GO 注释以及相关的研究文献。

CancerPPD（http://crdd.osdd.net/raghava/cancerppd/）[40]收集了通过实验验证的具有抗肿瘤作用的肽段（ACP）和蛋白，这些数据来源于公开发表的文献、专利和其他的数据库。目前，CancerPPD 包含了 3 491 个 ACP 以及 121 个抗肿瘤的蛋白。对于每一个条目，该数据库都提供了全面的注释信息，包括来源、肽段的特性、抗癌活性、羧基端和氨基端修饰、构象以及肽段的四级结构等。用户可搜索和浏览蛋白、ACP，查看与之相关的注释信息。除此之外，网站提供了多种比对工具，用户可通过比对来搜索序列或结构相似的肽段。

Cancer3D（http://cancer3d.org/）[41] 数据库整合了来自 TCGA 和 CCLE 的体细胞错义突变信息，在蛋白结构水平上分析其对蛋白功能的影响。该数据库为每个蛋白提供了两个不同的分析工具：e-Driver 和 e-Drug。E-Driver 可展示突变在蛋白中的位置、存在的结构域、与之相互作用的蛋白，并提供蛋白结构的 3D 视图，帮助用户判断突变对蛋白功能的潜在影响；e-Drug 可提供蛋白突变对药物活性的影响，可查看不同结构域上的突变对药物活性的影响，可帮助用户发掘出蛋白中可能的药物靶点结构域。Cancer3D 提供的这两项服务可帮助研究者评估突变对蛋白功能以及药物效果的影响，理解肿瘤突变和赖药性的关系，具有重要的应用价值。

6 癌基因数据库

肿瘤相关基因包括原癌基因和抑癌基因，大部分都与细胞的生长、增殖、迁移、侵袭、衰老、凋亡以及细胞周期有关。随着研究的深入，已经发现

许多与肿瘤相关的基因。对于已知肿瘤相关基因的收集、整理并共享可帮助研究者快速获得大量的肿瘤相关基因的信息，减少研究者的时间成本。以下对一些肿瘤相关基因的数据库作简单介绍。

DriverDB（http://driverdb.ym.edu.tw/DriverDB/）[42]收集了来自TCGA、ICGC、TARGET等数据库的总共6 000多个外显子组测序数据，并利用dbSNP、COSMIC等注释信息和生物信息学方法识别肿瘤驱动基因。用户可通过“Cancer”页面选取特定的肿瘤类型，得到该肿瘤中的驱动基因列表，并可获得它们的基因本体信息（GO）、参与的通路以及基因间的互作关系等。而通过“Gene”页面可搜索感兴趣的基因，查看该基因不同区域在不同肿瘤中的突变频度。另外，网站还提供元分析（Meta-Analysis），用户可选取一组样本做个性化分析。

Network of Cancer Genes（NCG，http://bio.ifom-ieo-campus.it/ncg）[43]收集和整理了多种肿瘤中的已知和候选的肿瘤相关基因。候选基因数据主要来源于基因组测序、外显子测序以及基因筛选实验（gene panel screening）。最新版（v5.0）中包含了518个已知的肿瘤相关基因及1 053个候选基因，覆盖了49种肿瘤，同时提供了miRNA与基因之间的调控关系。用户可浏览或搜索一个或多个基因，获得与该基因相关的功能和疾病注释信息、突变信息、表达谱、miRNA及蛋白互作关系等，还可以可视化miRNA调控关系和蛋白互作网络，用户可保存获得的结果。除此之外，用户还可下载全部肿瘤相关基因。

TP53MULTLoad（http：//p53.fr）[44]是一个人工收集的有关TP53基因突变的网站，包含了UMD TP53（http://www.umd.be：2072/）[45]数据库以及与TP53有关的信息。用户可利用该网站获取到p53蛋白的所有点突变的相关信息，如生化活性等。同时，该网站还提供有关TP53的分析工具，如TP53 Mut Assessor，允许用户在个人电脑上获取p53各种突变多方面的信息。

7　肿瘤与药物数据库

肿瘤细胞的耐药性是临床肿瘤治疗失败的主要原因之一，因此寻找耐药靶点成为肿瘤药物开发领域的热点之一。除了耐药性，肿瘤细胞对药物的敏感性、药物的副作用、肿瘤细胞的潜在药物靶点开发等也是肿瘤药物的重要研究方向，且依赖于医疗临床大数据的采集和分析。数据库的构建使得结构化的数据便于进行统计分析，从而研究治疗方案及疗效评价、药物副反应情况、肿瘤病人的治疗现状等，有助于深度挖掘肿瘤细胞与药物之间的关联，为精准医疗提供参考依据，促进肿瘤新药研发。

Genomics of Drug Sensitivity in Cancer（GDSC，www.cancerRxgene.org）[46]由英国桑格研究院（Sanger Institute）开发，收集肿瘤细胞对药物的敏感度和反应。癌基因组的变异会影响临床治疗的效果，不同的靶点对药物的反应也有很大不同。因此这类数据对于发现潜在的肿瘤治疗靶点十分重要。GDSC的数据来自75 000个实验，描述了约200个抗癌药物在1 000多种肿瘤细胞中的反应。该数据库中的癌基因组突变信息来自COSMIC数据库，包括癌基因点突变、基因扩增与丢失、组织类型以及表达谱等。用户可以从化合物、癌基因和细胞系3个层面对数据库进行检索，癌基因或细胞系对不同药物的反应会被详细列出，并且结果会以图形化的界面加以展示，包括统计分析，火山图及相关文献等。检索结果以及整个数据库都可由用户下载以进行后续分析。

canSAR（http://cansar.icr.ac.uk）[47]是由英国癌症研究院（The Institute of Cancer Research）开发，致力于帮助药物开发与肿瘤转化医学研究。该数据库包含了多种类型的数据，包括生物学、药理学、化学、结构生物学和蛋白质相互作用网络。这些不同类型的数据被整合起来以解决复杂的生物学问题，例如某个蛋白在不同肿瘤类型中的表达情况或突变情况，哪些化合物可以影响某类肿瘤细胞系的生长，某类药物会结合哪些蛋白并影响其生物学活性等。用户可以通过基因、蛋白、蛋白家族、蛋白3D结构、细胞系及药物来浏览或查询整个数据库，结果以详细列表的形式展示，并链接到相关信息资源。目前canSAR包含2万多个蛋白，约1.2万种细胞系，100万个化合物结构，整合了ArrayExpress、UniProt、COSMIC等11种数据源的数据。

CancerResource（http://bioinformatics.charite.de/cancerresource）[48]致力于收集与肿瘤相关的化合

物与靶标之间的联系，由柏林夏洛特医科大学开发。众多的生物学和医学实验发现了多种化合物可以用于激活或者抑制与肿瘤相关的癌基因，这些化合物可能成为潜在的药物靶点。然而这些信息都存在于大量的文献中，需要用文献挖掘的方法提取有用的信息。CancerResource 通过文献挖掘以及整合多种数据源的方式收集并发现了大量化合物及其靶点的信息。用户可以选择多种检索数据库的方式，包括搜索化合物、靶标、细胞系、突变、信号通路等。结果页面包含化合物与靶标的详细信息、表达图谱及相关数据来源链接等。该数据库收录了近 50 000 个化合物，3 000 多个与肿瘤相关的蛋白，2 000 多个细胞系以及约 9×10^5 条突变信息。由于整合了多种数据源，CancerResource 提供的数据资源非常全面，将有助于精准医药的开发与研究。

CancerDR（http://crdd.osdd.net/raghava/cancerdr）[49] 是另一个有助于精准医疗的数据库，由印度 CSIR 微生物技术研究所开发维护。耐药性是肿瘤治疗的一大障碍，药物靶点的突变是肿瘤产生耐药性的重要原因之一。CancerDR 收集了 148 种抗癌药物以及它们在 952 种细胞系中的药理状况，对于每一个药物靶点提供了序列的天然变体、突变体、三维结构和序列突变信息。其界面允许用户通过药物靶点、细胞系、药物名称和三维结构来检索或者浏览数据库，检索结果将以列表的形式展现。同时，作者还开发了一些在线分析工具，例如突变序列比对和聚类分析等。该数据库有助于发现新的药物靶点突变，并识别能杀死多种癌细胞的药物分子，从而促进肿瘤耐药性的治疗。

另一个更广泛收集耐药性信息的数据库是由剑桥大学开发的 Platinum（http://structure.bioc.cam.ac.uk/platinum）[50]。该数据库不局限于肿瘤数据，包含超过 1 000 种蛋白配体复合物的三维结构突变，以及这些突变对其亲和力的影响。这些数据由人工从 180 多篇相关文献中提取得到，共有 200 多个复合物。Platinum 的用户搜索界面包括多种限制条件，能使用户快速精确地从数据库中检索出需要的信息。该数据库将蛋白质结构突变与配体的亲和力关联起来，有助于研究由突变引起的疾病耐药性。

8 问题与展望

肿瘤生物信息学数据库发展迅速，但同时也存在一些问题与挑战。例如，与肿瘤相关的数据积累越来越快，单个研究课题就可能产生 10 TB 以上的原始数据，分析处理这些数据将耗费巨大的计算资源，如果要进行大规模数据分析所需要的资源将是难以承受的，如何将这些海量数据有效地存储起来，并以适当的格式提供给研究人员成为急需解决的问题。在数据迅速积累的情况下保持数据库的及时更新与升级也是非常重要的问题。另外，由于组学数据格式并不统一，现有的数据库大多只针对某一种组学数据或某一类特定的数据类型，整合多种数据类型可以促进寻找肿瘤驱动基因及治疗靶点，如何将独立的、分散的数据库中的信息整合到一起并开发新的数据整合算法，形成标准化、全方面的肿瘤信息数据库是目前该研究领域的新挑战。最后，目前广泛应用的肿瘤数据库主要集中在欧美等国，而我国有一些高发肿瘤类型在西方国家并不高发，如鼻咽癌和食管癌，因此这两类肿瘤的相关数据相对较少，研究也不多；反之，在西方国家高发的黑色素瘤在我国发病率极低。此外，由于人种的差异，同一种肿瘤在不同人种中的易感位点和基因突变频率也不尽相同。因此需要开发一些针对我国特有高发肿瘤类型或者针对亚洲人群的数据库，为我国的肿瘤研究提供高质量的数据服务与对比分析，同时完善全球肿瘤研究的数据资源。

目前，国内癌症研究相关数据库主要涉及到癌症病例的收集和整理的肿瘤登记数据库以及针对 miRNA、甲基化等热门领域的数据库。前者根据癌症病例数据的特点设计适宜的数据库结构，提高病例信息的管理水平，是循证医学十分重要的资源。目前，已有针对乳腺癌[51]、原发骨肿瘤[52]、脑肿瘤[53]等癌症的数据库，收集和整理了不同癌症患者的病例信息。而 miRNA、甲基化是目前生物学研究的前沿和热门领域，也是国内癌症研究的重要方向。除了前述的 MethyCancer[24]、DiseaseMeth[28] 及 OncomiRDB[33] 等数据库外，还有多个数据库也是针对这些热门领域的。dbDEMC[54] 和 nc2Cancer[55] 都是人类癌症相关的非编码 RNA 数据库，分别记录

了非编码 RNA 的表达谱及其与肿瘤的关系；而李孟娇等[56]构建的有关喉癌的数据库则整理与喉癌相关的基因、蛋白以及 miRNA 甲基化和表达数据。这些数据库的构建为国内的癌症研究积累了重要的具有地域特色的癌症基础数据，为针对本国的肿瘤研究奠定了一定的基础。值得注意的是，我国肿瘤病例登记目前还处在初级阶段，信息分散且数据量比较有限，需要更多的努力来整合并扩大覆盖面；而针对热门领域的数据库要注重数据库的维持和更新，保持数据库的时效性，进一步提高数据的科研和应用价值。

虽然存在问题与挑战，肿瘤生物信息学数据库已经为肿瘤研究做出了巨大的贡献。癌症研究领域丰富的实验数据促进了一大批肿瘤生物信息学数据库的出现，这些数据库所提供的在线数据分析功能与下载平台又大大地促进了我们对肿瘤发生发展机制的认识。随着生物学大数据时代的到来，利用生物信息学进行数据分析与诠释已经成为实验研究不可或缺的手段与资源。随着日新月异的技术革新与精准医疗项目的开展，必定会出现更多的肿瘤数据库，并最终从根本上改变癌症的诊断和治疗方式。

参 考 文 献

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

[2] Benson DA, Clark K, Karsch-Mizrachi I, et al. GenBank [J]. Nucleic Acids Res, 2015, 43 (Database issue): D30-35.

[3] Li W, Cowley A, Uludag M, et al. The EMBL-EBI bioinformatics web and programmatic tools framework [J]. Nucleic Acids Res, 2015, 43 (W1): W580-584.

[4] Kodama Y, Mashima J, Kosuge T, et al. The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data [J]. Nucleic Acids Res, 2015, 43 (Database issue): D18-22.

[5] Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project [J]. Nat Genet, 2013, 45: 1113-1120.

[6] Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin [J]. Cell, 2014, 158 (4): 929-944.

[7] Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer [J]. Nature, 2013, 500 (7463): 415-421.

[8] Grieb BC, Chen X, Eischen CM. MTBP is overexpressed in triple-negative breast cancer and contributes to its growth and survival [J]. Mol Cancer Res, 2014, 12 (9): 1216-1224.

[9] Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research [J]. Nat Genet, 2015, 47 (7): 692-695.

[10] Wilks C, Cline MS, Weiler E, et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data [J]. Database (Oxford), 2014, 2014. pii: bau093.

[11] International Cancer Genome Consortium. International network of cancer genome projects. [J]. Nature, 2010, 464 (7291): 993-998.

[12] Strausberg RL, Buetow KH, Emmert-Buck MR, et al. The cancer genome anatomy project: building an annotated gene index [J]. Trends Genet, 2000, 16 (3): 103-106.

[13] Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer [J]. Nucleic Acids Res, 2015, 43 (Database issue): D805-811.

[14] Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal [J]. Sci Signal, 2013, 6 (269): pl1.

[15] Goldman M, Craft B, Swatloski T, et al. The UCSC Cancer Genomics Browser: update 2015 [J]. Nucleic Acids Res, 2015, 43 (Database issue): D812-817.

[16] Zhang J, Finney RP, Rowe W, et al. Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB) [J]. Genome Res, 2007, 17 (7): 1111-1117.

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

[18] Cai H, Gupta S, Rath P, et al. arrayMap 2014: an updated cancer genome resource. [J]. Nucleic Acids Res, 2015, 43 (Database issue): D825-830.

[19] Cao Q, Zhou M, Wang X, et al. CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data [J]. Nucleic Acids Res, 2011, 39 (Database issue): D968-974.

[20] Scheinin I, Myllykangas S, Borze I, et al. CanGEM: mining gene

copy number changes in cancer.[J]. Nucleic Acids Res, 2008, 36 (Database issue): D830-835.

[21] Timms B. Cancer genome project to start [J]. Eur J Cancer, 2000, 36 (6): 687.

[22] Wu TJ, Shamsaddini A, Pan Y, et al. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE) [J]. Database (Oxford), 2014, 2014: bau022.

[23] Formosa A, Lena AM, Markert EK, et al. DNA methylation silences miR-132 in prostate cancer [J]. Oncogene, 2013, 32 (1): 127-134.

[24] He X, Chang S, Zhang J, et al. MethyCancer: the database of human DNA methylation and cancer [J]. Nucleic Acids Res, 2008, 36 (Database issue): D836-841.

[25] Huang WY, Hsu SD, Huang HY, et al. MethHC: a database of DNA methylation and gene expression in human cancer [J]. Nucleic Acids Res, 2015, 43 (Database issue): D856-861.

[26] Baek SJ, Yang S, Kang TW, et al. MENT: methylation and expression database of normal and tumor tissues [J]. Gene, 2013, 518 (1): 194-200.

[27] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update [J]. Nucleic Acids Res, 2013, 41 (Database issue): D991-995.

[28] Lv J, Liu H, Su J, et al. DiseaseMeth: a human disease methylation database [J]. Nucleic Acids Res, 2012, 40 (Database issue): D1030-1035.

[29] Negre V, Grunau C. The MethDB DAS server: adding an epigenetic information layer to the human genome [J]. Epigenetics, 2006, 1 (2): 101-105.

[30] Geisen S, Barturen G, Alganza AM, et al. NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes[J]. Nucleic Acids Res, 2014, 42(Database issue): D53-59.

[31] Kolesnikov N, Hastings E, Keays M, et al. ArrayExpress update--simplifying data submissions [J]. Nucleic Acids Res, 2015, 43 (Database issue): D1113-1116.

[32] Rhodes DR, Kalyana-Sundaram S, Mahavisno V, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18, 000 cancer gene expression profiles [J]. Neoplasia, 2007, 9 (2): 166-180.

[33] Wang D, Gu J, Wang T, et al. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs [J]. Bioinformatics, 2014, 30 (15): 2237-2238.

[34] Xie B, Ding Q, Han H, et al. miRCancer: a microRNA-cancer association database constructed by text mining on literature [J]. Bioinformatics, 2013, 29 (5): 638-644.

[35] Bhattacharya A, Ziebarth JD, Cui Y. SomamiR: a database for somatic mutations impacting microRNA function in cancer [J]. Nucleic Acids Res, 2013, 41 (Database issue): D977-982.

[36] Frenkel-Morgenstern M, Gorohovski A, Vucenovic D, et al. ChiTaRS 2.1--an improved database of the chimeric transcripts and RNA-seq data with novel sense-antisense chimeric RNA transcripts [J]. Nucleic Acids Res, 2015, 43 (Database issue): D68-75.

[37] Ellis MJ, Gillette M, Carr SA, et al. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium [J]. Cancer Discov, 2013, 3 (10): 1108-1112.

[38] He Y, Zhang M, Ju Y, et al. dbDEPC 2.0: updated database of differentially expressed proteins in human cancers [J]. Nucleic Acids Res, 2012, 40 (Database issue): D964-971.

[39] Li J, Duncan DT, Zhang B. CanProVar: a human cancer proteome variation database [J]. Hum Mutat, 2010, 31 (3): 219-228.

[40] Tyagi A, Tuknait A, Anand P, et al. CancerPPD: a database of anticancer peptides and proteins [J]. Nucleic Acids Res, 2015, 43 (Database issue): D837-843.

[41] Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures [J]. Nucleic Acids Res, 2015, 43 (Database issue): D968-973.

[42] Cheng WC, Chung IF, Chen CY, et al. DriverDB: an exome sequencing database for cancer driver gene identification [J]. Nucleic Acids Res, 2014, 42 (Database issue): D1048-1054.

[43] An O, Pendino V, D'Antonio M, et al. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes [J]. Database (Oxford), 2014, 2014: bau015.

[44] Leroy B, Fournier JL, Ishioka C, et al. The TP53 website: an integrative resource centre for the TP53 mutation database and TP53 mutant analysis[J]. Nucleic Acids Res, 2013, 41(Database issue): D962-969.

[45] Hamroun D, Kato S, Ishioka C, et al. The UMD TP53 database and website: update and revisions [J]. Hum Mutat, 2006, 27 (1):

14-20.

[46] Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells [J]. Nucleic Acids Res, 2013, 41 (Database issue): D955-961.

[47] Bulusu KC, Tym JE, Coker EA, et al. canSAR: updated cancer research and drug discovery knowledgebase [J]. Nucleic Acids Res, 2014, 42 (Database issue): D1040-1047.

[48] Ahmed J, Meinel T, Dunkel M, et al. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge [J]. Nucleic Acids Res, 2011, 39 (Database issue): D960-967.

[49] Kumar R, Chaudhary K, Gupta S, et al. CancerDR: cancer drug resistance database [J]. Sci Rep, 2013, 3: 1445.

[50] Pires DE, Blundell TL, Ascher DB. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes [J]. Nucleic Acids Res, 2015, 43 (Database issue): D387-391.

[51] 李金平, 李宏, 廉斌, 等. 乳腺癌电子数据库的建立及临床应用 [J]. 海南医学, 2013, 24 (9): 1371-1372.

[52] 单华超, 徐海荣, 李远, 等. 原发骨肿瘤流行病学数据库的建立与使用 [J]. 中国骨与关节杂志, 2015 (9): 693-696.

[53] 郑虎, 张红波, 孙彦辉, 等. 脑肿瘤患者认知障碍数据库的初步建立及临床意义 [J]. 数理医药学杂志, 2013 (4): 410-412.

[54] 崔洪亮, 张阳德, 任菲. dbDEMC2.0: 人类癌症相关 miRNA 数据库 2.0 [J]. 中国现代医学杂志, 2014, 24 (3): 77-79.

[55] 程卓, 刘珂, 严章明, 等. nc2Cancer: 一个研究与癌症相关人类非编码 RNA 的数据库 [J]. 生物信息学, 2015, 13 (2): 77-81.

[56] 李孟娇, 鄂琪敏, 刘加林, 等. 喉癌相关基因和 miRNA 综合数据库的构建 [J]. 中华耳鼻咽喉头颈外科杂志, 2015, 50 (9): 765-768.

(责任编辑　马鑫)