

# 常用肿瘤基因分析方法及基于 TCGA 数据库的分析应用

李鑫, 李梦玮, 张依楠, 徐寒梅

中国药科大学多肽药物创制工程中心, 南京 211198

**摘要:** 随着二代测序技术的快速发展, 数据量不断累积, 肿瘤学家的目光逐渐由多物种测序转移至高通量测序数据的分析和比对。基因数据分析方法层出不穷, 高通量的组学分析手段不断优化和创新, 基因数据的挖掘和分析工作正处于飞速发展的时期。以肿瘤病人样本为核心的数据库 The Cancer Genome Atlas (TCGA) 由此应运而生, 该数据库全方位记录了从临床肿瘤病人样本得到的基因数据如 DNA 序列、转录本信息、表观遗传学修饰等。本文主要从数据分析方法、TCGA 数据库及其应用实例等 3 个方面详细介绍了肿瘤相关基因数据的深入挖掘和生物信息学分析方法的最新研究进展, 以期研究人员利用大数据发现肿瘤防治相关的新靶点提供借鉴和参考。

**关键词:** 基因数据; TCGA 数据库; 肿瘤

## Common cancer genetic analysis methods and application study based on TCGA database

Xin Li, Mengwei Li, Yinan Zhang, Hanmei Xu

*Engineering Research Center of Peptide Drug Discovery and Development, China Pharmaceutical University,  
Nanjing 211198, China*

**Abstract:** The development of second-generation sequencing (NGS) technology is providing numerous data which shifts the focus of cancer research from the sequencing of multi-species to the analysis and comparison of select data via high-throughput sequencing. The NGS also facilitates the diversity of available genetic data analysis methods, the constant optimization and innovation of analytical approaches for high-throughput genomics as well as the rapid development of genetic data mining and analysis models. The Cancer Genome Atlas (TCGA) database is a direct result of this work. The TCGA database provides a comprehensive record of genetic data collected from a tumor patient's sample, including its

收稿日期: 2018-11-20; 修回日期: 2019-01-27

基金项目: 国家“重大新药创制”科技重大专项(编号: 2018ZX09301053-001, 2018ZX09301039-002, 2018ZX09201001-004-001)和江苏高校优势学科建设工程项目资助[Supported by the National Science and Technology Major Projects of New Drugs (Nos. 2018ZX09301053-001, 2018ZX09301039-002, 2018ZX09201001-004-001) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD)]

作者简介: 李鑫, 硕士研究生, 专业方向: 海洋药学。E-mail: cpu\_lixin@163.com

通讯作者: 徐寒梅, 博士, 教授, 研究方向: 多肽类药物研究与开发。E-mail: 13913925346@126.com

DOI: 10.16288/j.yczs.18-279

网络出版时间: 2019/2/25 15:23:47

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20190225.1523.006.html>

DNA sequence, transcriptional information, epigenetic modification and related. This review elaborates the latest progress in both the mining algorithm and analysis methods for tumor genomics. Specially, we introduce and review the TCGA database and data analysis approaches while demonstrating its applicability using representative cases. This review may shed light on new tumor-related targets discovery for researchers by means of big data.

**Keywords:** gene data analysis; TCGA database; cancer

近年来,随着高性能计算机集群技术支持的新一代测序机和自动化分析的高通量测序平台不断问世、基因组测序分析成本大幅降低、基因组数据共享平台层出不穷,以及大量的基因组数据被上传至互联网,为研究人员开展大规模的基因组学研究创造了便利条件,同时肿瘤基因组学的研究也越来越深入。由此,整合多种癌症基因组数据的 The Cancer Genome Atlas (TCGA)数据库应运而生,为研究人员快速、准确地获取肿瘤基因组数据提供了很好的途径。

数据挖掘是一门随着计算机科学发展而快速发展的学科,其在生命科学领域的作用随着大量测序数据的累计而逐渐显现。现阶段,国内大部分实验室对基因组数据挖掘和处理还处于起步阶段,不仅缺乏相应的数据处理平台,更缺乏具有相应知识背景的科研人员,而在国际上基因组数据研究已经是一个迅猛发展的领域。本文重点介绍了常见基因数据分析方法、TCGA 数据库以及近年来围绕 TCGA 数据库所得到的研究成果,期望为相关科研人员提供一些利用数据库资源研究肿瘤基因组学的新思路。

## 1 常见基因数据分析方法

### 1.1 生存分析

生存分析是一类用于计算在一个集合内对于给定的时间段中影响因素与给定结果或时间事件之间关联的统计学方法,该方法的特点是可以对时间事件进行分析,其中 Kaplan-Meier 生存分析和 Cox 回归分析是两种最常用的时间事件标准化统计学方法。Kaplan-Meier 生存分析可以基于一个影响因素对事件进行分析,每个独立个体的时间范围由记录点开始一直延续至事件发生点。Cox 回归分析是一种多参数回归模型,该模型以生存结局和生存时间为因

变量,可同时分析多种因素对生存期的影响<sup>[1]</sup>。在随机对照临床试验中,Kaplan-Meier 生存分析是首选的数据分析方法<sup>[2]</sup>。对于多影响因素事件,可选用 Cox 回归分析。基于这两种分析方法的特点,在基因数据分析中,Kaplan-Meier 多用于分析基因表达与生存周期的关系,而 Cox 回归多用于分析预后影响因素与生存周期的关系<sup>[3]</sup>。

### 1.2 差异表达分析和聚类分析

差异表达是指同一基因在两个条件中的检测结果在排除系统误差、人为误差等因素后具有较为明显的差异,通常用  $P$  值来表示。这种差异可以通过外显子测序、芯片筛选等方法检测。比较同一基因在不同条件下的表达量差异是筛选潜在功能基因的第一步,通常由统计学工具辅助完成。常用的算法包括倍数法、 $t$  检验法、方差分析、SAM 法、贝叶斯法和信息熵法等<sup>[4]</sup>,这些统计学方法各有其优势和不足(表 1)。

聚类分析在基因表达数据研究中被大量应用且在不断优化,它可以在模式分类数不确定的情况下对基因数据进行分组,其数学意义是将研究对象分为相对同质的群组。从生物学的角度,这种方法就是将具有潜在相同作用的基因分为同一组,如对于一组肿瘤组织高表达基因可以假定其存在促肿瘤生长活性,对于一组低表达基因则可假定其存在抗肿瘤活性,或认为同一组基因可能受同一转录因子的调控等。

两个影响聚类分析结果的重要指标是评价研究对象相似性程度的距离尺度和将研究对象分组的聚类算法,其中距离尺度可以根据不同的筛选目的分为几何距离、线性相关系数和非线性相关系数 3 种,分别对应的是衡量样本间的相似性、衡量样本间是否具有相同变化趋势和衡量样本间在同一时间节点

表 1 基因差异表达分析方法优缺点

Table 1 Advantages and disadvantages of gene differential expression analysis methods

分析方法	优点	缺点
倍数法	计算量小, 一般用于大规模初筛	具体阈值较难确定
$t$ 检验法	能充分利用样本信息, 检验效率高	在数据量较小时, 对总体方差的估计不准确
方差分析	不受比较组数的限制, 且可以同时分析多个因素的作用	多重假设检验可能带来放大的假阳性率
SAM 法	假阳性率低	诊断能力较差, ROC 指数相对偏低
贝叶斯法	样本量小时也可得到较好的分析结果	对卡方分布和指数分布的数据不敏感
信息熵法	无需样本的类别信息即可进行筛选	不能得到差异表达的基因

的波动趋势是否相似。而常用的聚类算法主要包括简单聚类、层次聚类、模糊聚类、 $k$  均值聚类、双向聚类和自组织映射神经网络聚类等。对于聚类结果, 一般选择对其进行可视化处理, 使其更易于接受和直观的分析, 常用的有热图(heatmap)、点线图和冰柱图等<sup>[5]</sup>。

### 1.3 受试者工作特征曲线分析

受试者工作特征曲线分析(receiver operating characteristic, ROC)最早起源于第二次世界大战时期, 最初用来降低雷达兵们的误报率和漏报率, 现多用于临床疾病诊断临界点寻找、不同检测方法对同一疾病的识别能力的比较、单一生物标志物对疾病的诊断准确度和筛选对疾病发生发展有显著影响的潜在基因。ROC 曲线是一条通过二分类方式拟合的非线性曲线, 其纵坐标为敏感度, 横坐标为(1-特异性), 评价指标为曲线下面积(area under the curve, AUC)。与生存分析最大的不同点在于 ROC 曲线分析不考虑时间因素, 且不需要将试验结果分为两类, 因此一般不用于分析预后等时间相关事件。ROC 曲线分析的优点是直观、简单, 可用肉眼看出结果。而缺点是对临界点的寻找没有明确的限定, 可能一定程度上影响数据分析结果。在许多生物信息学分类分析时, ROC 分析经常出现正相关显著低于负相关的现象, 因此研究人员对其进行了改进, 加入了精确率与反馈率曲线 (precision-recall, PR), 这一优化使正负分类结果相对平衡, 已经在 R 语言中实现了应用。对于不同条件间 ROC 比较, 则需要分别对其 AUC 进行处理, 消除抽样误差带来的影响, 常用的处理方法有 Delong 法和 Hanley 法<sup>[6,7]</sup>。

### 1.4 Meta 分析

Meta 分析是一种对同类研究结果进行整合定量分析的统计学方法, 其目的是通过整合多个已有的研究数据来增大样本含量, 从而减少由随机误差所导致的数据差异, 进而增大检验学效能。在临床研究中常用于病因学、诊断性试验、发病机制、病人费用和效益、流行病学、干预措施评价、随访和预后测评等方面的分析。一般的分析流程为提出问题、文献与资料收集、数据构建、Meta 分析和实验验证。其中文献与资料收集是影响 Meta 分析结果的关键步骤, 涉及到文献搜索策略和数据纳入排除标准的建立<sup>[8]</sup>。

一般来说, 同一领域不同研究组之间的操作和研究方法会存在一定区别, 进而带来一些人为误差。这种差异被称为异质性, 一般分为方法异质性、临床异质性和统计学异质性。异质性检验是验证所构建标准是否良好的常用方法。对于基因表达常用的芯片 Meta 分析, 一般选用同一测序平台来源的数据以避免测序方法对分析结果的干扰。Meta 分析根据实际要求不同可以分为多种类型包括单组率 Meta 分析、网状 Meta 分析和诊断性 Meta 分析等, 其具体分类依据在许多文章中都有报道过, 因此不再叙述<sup>[9]</sup>。

## 2 TCGA 数据库

### 2.1 数据库简介

肿瘤被认为是人类最复杂疾病之一, 目前为止人类已经发现了超过 200 种肿瘤亚型。肿瘤病人

基因中发生的变化如体细胞突变、拷贝数变异、基因表达量差异和表观修饰变化与其特定的肿瘤亚型是相对应的。因此,为了更好地发现、诊断和治疗肿瘤,对其基因变化进行深入研究和建立相应数据库是目前所急需的<sup>[10]</sup>。2006年,美国国立癌症研究院(National Cancer Institute, NCI)和美国国立人类基因组研究院(National Human Genome Research Institute, NHGRI)合作开展了 The Cancer Genome Atlas (TCGA)数据库计划,该计划旨在通过大规模基因测序和综合性、多维度的分析手段来寻找由肿瘤发生发展造成的基因变化,构建肿瘤基因相关的全方位“地图集”<sup>[11]</sup>。

TCGA 计划分为两个部分:第一部分从 2006~2008 年选择了具有严重不良预后且危害公共健康的 3 种常见肿瘤(脑癌、肺癌和卵巢癌)进行数据采集和分析,从而对其数据库整体框架的构建进行基本测试;从 2009 年开始进入第二阶段,扩大肿瘤类型至 33 种并扩大样本量进行 6 种数据类型的记录和分析(图 1 A 和 B)这一过程虽然耗资巨大但成果显著。近年来科研人员已经依据 TCGA 数据库在多种肿瘤中发现了潜在的临床标志物和治疗靶点<sup>[12~15]</sup>。

## 2.2 TCGA 数据类型

TCGA 使用基于芯片技术的高通量测序方法和二代测序技术来精确记录肿瘤基因组的全方位信息,除此之外,TCGA 还记录并追踪了病人的临床信息包括性别、年龄、肿瘤分期、复发和预后情况等,从而有利于对其开展多因素综合性的分析。以下为 TCGA 数据库中较为常见的数据类型。

### 2.2.1 RNA 测序数据

RNA 测序(RNA-seq)是一种针对转录组进行测序的高通量技术,其特点是在大量样本中快速识别和量化不同表达水平的转录组,检测异构体变化、找到新的转录组、筛选融合基因和非编码 RNA(ncRNA)。TCGA 数据库中提供了 RNA 序列、基因表达量、外显子序列和突变点等信息的记录,这一数据库为肿瘤转录组研究人员提供了大量数据和样本信息支持<sup>[16,17]</sup>。

### 2.2.2 MicroRNA 测序数据

MicroRNA 是一种长度约 20nt 的非编码小 RNA 分子,通过与 mRNA 相互作用影响目标 mRNA 的稳定性及转录翻译等过程,最终调控基因表达、诱导靶基因沉默、影响细胞生长、发育等生物过程<sup>[18]</sup>,近年来也有研究以 miRNA 作为靶点的抗肿瘤药物<sup>[19]</sup>。TCGA 数据库提供了肿瘤样本的 miRNA 表达、异构体情况,可以用于分析肿瘤相关基因的互作网络关系和探索未被发现的 miRNA<sup>[20,21]</sup>。

### 2.2.3 DNA 测序数据

DNA 测序(DNA-seq)是一种高通量手段来测定 DNA 序列从而找到 DNA 的变化如插入、缺失、点突变、多态性、拷贝数改变、突变频率和病毒基因组侵入。TCGA 数据库以 Sanger 测序技术为基础构建了 DNA 测序数据集,构建该数据集是为了探究在不同肿瘤类型中基因组的多样性,从而进一步找到具有诊断和治疗意义的新靶点<sup>[22,23]</sup>。

### 2.2.4 单核苷酸多态性检测数据

单核苷酸多态性检测(single nucleotide polymorphisms, SNPs)是指由单一核苷酸的改变所引起的序列多态性,TCGA 选择了 Illumina 平台的分子量阵列技术来检测多种肿瘤基因组中 SNP 水平的变化,此外还能记录拷贝数变异(copy number variation, CNV)和杂合性缺失(loss of heterozygosity, LOH)<sup>[24]</sup>。

### 2.2.5 DNA 甲基化测序数据

DNA 甲基化测序可以检测全基因组的表观遗传学改变,在 CpG 位点上的甲基化和去甲基化修饰是最早和最常见的肿瘤相关表观遗传变异,这些表观遗传变异具有成为特异性肿瘤标志物的可能。TCGA 数据库中的甲基化数据是基于 Illumina 测序平台获得的,保证了单碱基对的分辨率,高测量精度和低样品 DNA 需要量,不仅记录了信号强度、探查可信度还收录了用于进一步确定 DNA 甲基化水平的计算  $\beta$  值等<sup>[25~27]</sup>。

### 2.2.6 反向蛋白质阵列表达数据

反向蛋白质阵列(reverse-phase protein array, RPPA)



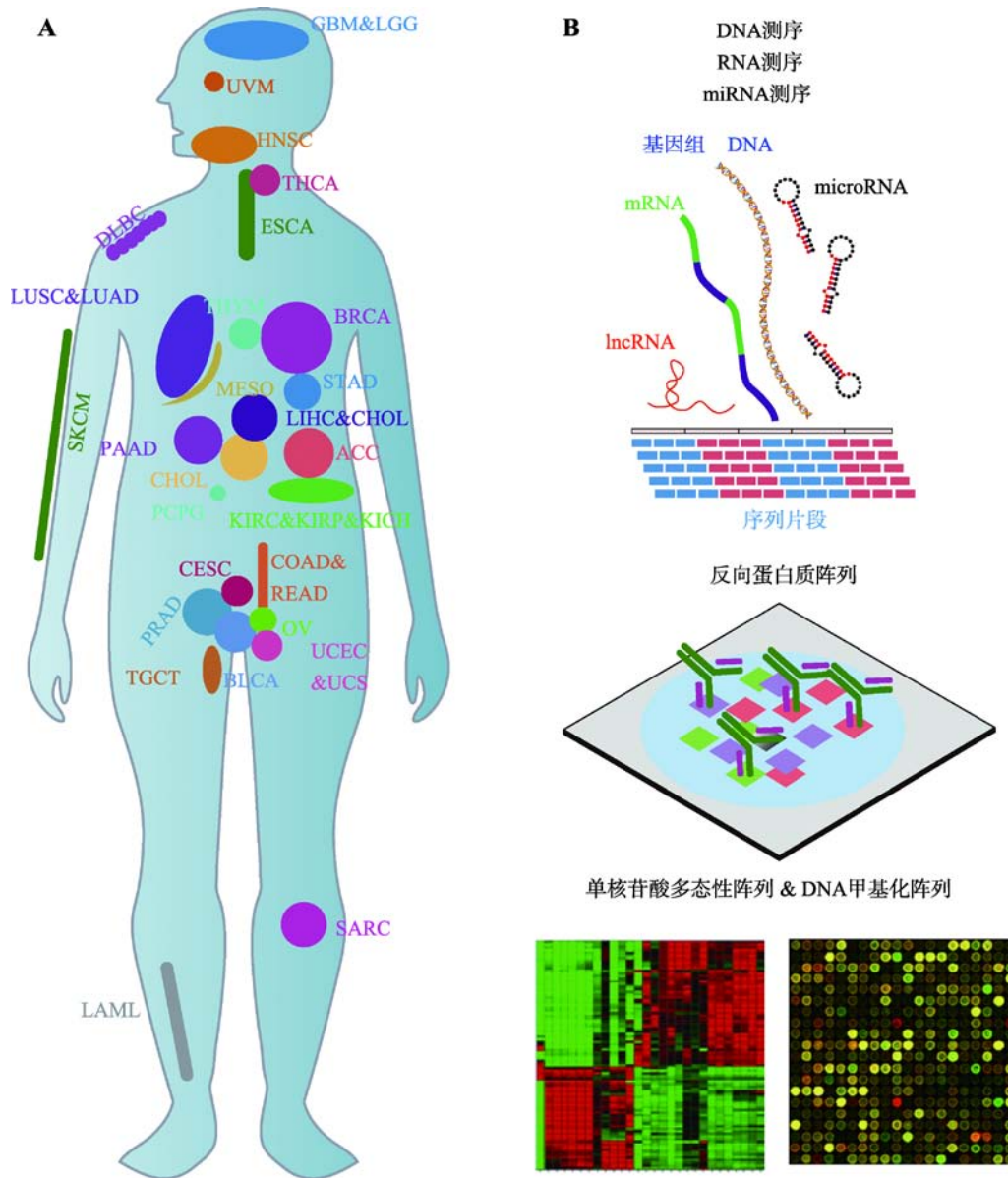


图 1 TCGA 数据库收录的肿瘤类型和数据类型

Fig. 1 Tumor types and numeric types of TCGA database

A: TCGA 收录的 33 种肿瘤类型的体内分布示意图。ACC: 肾上腺皮质癌; BLCA: 膀胱癌; BRCA: 乳腺癌; CESC: 宫颈鳞状细胞癌; CHOL: 胆癌; COAD: 结肠腺癌; DLBC: 弥漫性大 B 细胞淋巴瘤; ESCA: 食管癌; GBM: 多形性胶质母细胞瘤; HNSC: 头颈部鳞癌; KICH: 肾嫌色细胞癌; KIRC: 肾透明细胞癌; KIRP: 乳头状肾细胞癌; LAML: 骨髓瘤; LBB: 低分化脑胶质细胞瘤; LIHC: 肝癌; LUAD: 肺腺癌; LUSC: 肺鳞状细胞癌; MESO: 间皮瘤; OV: 卵巢癌; PAAD: 胰腺癌; PCPG: 肾上腺癌; PRAD: 前列腺癌; READ: 直肠癌; SARC: 肉瘤; SKCM: 皮肤黑色素瘤; STAD: 胃癌; TGCT: 睾丸癌; THCA: 甲状腺癌; THYM: 胸腺癌; UCEC: 子宫内膜癌; UCS: 子宫癌; UVM: 葡萄膜黑色素瘤。B: TCGA 记录的 6 种测序数据类型。

是一种高通量、高灵敏度、可重复的蛋白检测技术，可同时用 500 个抗体对超过 1000 个样本进行检测，可以用于分子标志物筛选、分子靶标识别、肿瘤细胞亚型分析和药效学评价。TCGA 数据库收录了 RPPA 分析的原始图片，原始信号强度，相对蛋白表

达量以及标准化后的蛋白信号<sup>[28]</sup>。

### 2.3 TCGA 数据库资源获取方法

TCGA 数据库提供的数据量较大，一般需要专业的工具下载和处理，研究人员可以直接访问 TCGA

数据库网站(<https://portal.gdc.cancer.gov/>), 使用其自带的 GDC-Client 进行下载。也可以利用编程语言 R 中的多种包如 TCGA2STAT、RTCGA 等进行下载。此外, 还可以使用一些研究人员制作的第三方工具如 TCGA-Assemble 等进行数据下载和初始化处理。

### 3 基于 TCGA 数据库分析的应用实例

#### 3.1 针对单一类型数据的研究

三阴性乳腺癌(triple negative breast cancer, TNBC)是一种高异质性和侵略性的疾病, 且目前为止没有明确有效的治疗靶点, 在依据肿瘤亚型为基准的个体化医疗时代, TNBC 相比于其他类型的乳腺癌有更高的死亡率。但在临床中发现, 约有 1/3 的病人通过常规化疗手段使病情得到完全缓解。因此, Jiang 等<sup>[29]</sup>以对化疗敏感为条件在 TCGA、METAVRIC 等数据库中选择了约 400 例样本的肿瘤组织和正常组织外显子序列进行研究。在分析中他们发现以 BRCA1 分子为核心的 AR-和 FOXA-调节网络的突变与化疗敏感性有较高的相关度。进一步分析发现以 BRCA1/2 低表达为表型的 BRCA 基因缺陷型 TNBC 病人有更高的化疗敏感性和更长的化疗后生存周期。除此之外, 通过体外实验发现 BRCA 缺陷型 TNBC 病人体内不仅有相对更高的突变率且体内表达了一种可以增强免疫细胞活性的新抗原。因此, BRCA 缺陷可以作为一个潜在的三阴性乳腺癌分类标签。

IsomiRNA 是一类序列或长度发生变化的异构体 miRNA, 这类 RNA 的靶点和功能会较原有的标准 miRNA 有所变化。在肿瘤发生过程中, 这类 miRNA 被认为对其有潜在的调控作用。Omar 等<sup>[30]</sup>通过对 TCGA-miRNA 数据集中乳腺癌的数据进行分析, 发现 has-miR-140-3p 和 5'isomiR-140-3p 在乳腺癌中均高表达。他们对这两种 miRNA 进行功能分析发现, 两者均能通过作用于增殖和迁移相关的基因从而对肿瘤细胞的生存和转移有显著的调控作用, 且二者之间存在协同作用关系。

#### 3.2 针对多组学数据的研究

由于胰腺腺癌病患的异质性高导致现阶段的治

疗效果不理想, Gibori 等<sup>[31]</sup>尝试利用 RNAi 技术进行多靶点给药, 从而解决这一问题。他们首先通过对 TCGA 数据库中胰管腺癌的蛋白质阵列数据和 microRNA 测序数据进行分析, 结合病人的生存情况找出与生存时间显著正相关的 microRNA 和显著负相关的蛋白质, 分别为 miR-34a 和 PLK1。他们还利用两亲性谷氨酰胺聚合物作为纳米载体, 将 miR-34a 的类似物(miR-34a mimic)和抑制 PLK1 蛋白表达的 siRNA 共同偶联至载体表面进行体内外给药实验。小鼠移植瘤模型研究发现这种双靶点纳米制剂可以有效靶向至胰管腺癌的发病部位并抑制肿瘤生长, 这为胰管腺癌的治疗提供了新思路。

TCGA 数据库提供了 30 余种肿瘤类型的相关数据, 这使得泛肿瘤研究的进展大大提升, Thorsson 等<sup>[32]</sup>对 TCGA 中 33 种肿瘤类型的超过 10000 例样本的全部 6 种数据进行免疫基因组分析, 使用 160 个免疫表达特征进行打分, 通过聚类分析将这 10000 余个样本进行分类, 最终基于不同的免疫表达特征分为 6 类, 包括 IFN- $\gamma$  主导型、炎症型、淋巴细胞耗尽型、免疫沉默型和 TGF- $\beta$  主导型等。基于这 6 种分类, 研究人员对不同类别中的肿瘤免疫浸润构成、免疫反应与体细胞多样性的相关性、免疫反应与预后的相关性、不同免疫亚型与预后的相关性、免疫原性的变化、免疫调节剂的表达差异等进行了进一步的关联分析, 从而证明了这种分类的准确性。这一分类几乎包括了人类所有的恶性肿瘤类型, 这为从免疫基因组学角度预测疾病走向和病人预后提供了帮助。

Berger 等<sup>[33]</sup>通过对 TCGA 数据库中包括乳腺癌在内的 5 种妇科肿瘤类型的 2579 例样本进行综合的多平台分析并与其余肿瘤类型样本数据进行对比, 发现了这 5 种肿瘤病人样本中特有的基因组和表观基因组特征, 包括 3 个体细胞拷贝数变异、46 个显著突变基因以及与之报道相同的多种 miRNA 和 lncRNA 异常表达, 研究人员通过多种聚类分析将这 5 种具有共性的妇科肿瘤类型基于 16 个特异性分子指标分为了 5 个亚型, 进一步验证发现这 5 种亚型病人的生存时间存在显著差异, 最终研究人员在保证分类精确度的基础上, 使用二分决策树将 16 个特异性分子优化至 6 个, 这为未来妇科肿瘤的分类

和诊断提供了帮助。

精准肿瘤学是一门分析个体差异从而指导肿瘤治疗的学科。近年来研究发现,多组学特征可以用来预测肿瘤患者的临床特征,但多组学数据计算量大,分析难度高且大部分医生没有学习过相关的生物信息学知识,因此 Yu 等<sup>[34]</sup>建立了 Omics Analysis System for PRrecision Oncology (OASISPRO)系统,用于挖掘和量化 TCGA 数据库中的多组学数据。该系统可以将临床样本数据可视化,并基于机器学习相关算法找出与临床分期相关的基因,以及预测患者生存时间,这对精准治疗和个体化用药提供了指导。

Omics Pipe 是一个模块化的云计算平台,该平台可以根据用户要求自动获取 TCGA 数据库中的相关数据集,并进行多组学整合分析,此外还可以自定义组学分析和在平台框架基础上加入自己的计算模块,自由度更高。该平台是用 python 代码构建而来,所有的计算与分析工作都是依托亚马逊云服务器完成,平台构建的目的是为广大生物学家提供一个模块化的高通量数据分析框架,使数据分析变得更简单和高效<sup>[35]</sup>。

## 4 结语与展望

二代测序技术作为 21 世纪的重大科学技术进步之一,为肿瘤基因组学研究提供了极大的帮助,随着肿瘤基因组数据库和患者样本信息的不断丰富,科研人员对肿瘤基因的分析日趋深入,而对分析方法和工具的选择要求也不断提高。目前对肿瘤基因组的分析仍然处于起步阶段,虽然 TCGA 构建了立体化的多元素肿瘤基因组数据库,但多组学的基因数据很少作为一个整体进行立体化的分析,大多数研究都只局限于某一特定的数据类型如 SNP、miRNA 和表观修饰等。这也从侧面体现了现阶段统计学算法的局限性。

计算机性能的不断提升使数据量不再是限制科研人员的主要因素,而如何将多组学数据整合到一起才更为关键。现阶段的多组学分析还比较简单,大多数研究都围绕聚类分析展开,将多组学数据依照临床样本信息进行分类,筛选出潜在的肿瘤标志物。而这种分析对肿瘤的多组学发病机理研究帮助

较小,无法系统的阐明不同组学水平之间的关联性。但由于机器学习等人工智能算法的出现,科研人员将从更宏观的角度来分析肿瘤基因组数据,TCGA 数据库也已经与多个高校及科研机构合作,尝试进行高通量多组学的肿瘤基因数据分析,但其分析结果的准确性还有待进一步的验证,同时,分析结果的具体临床应用也有待开发。随着算法的不断发展,多组学分析将为肿瘤学研究提供强有力的支持,并从宏观的角度阐述不同分子水平对肿瘤的调控作用以及之间的联系。相信未来会出现基于多组学基因数据的整合分析方法,更全面的阐述肿瘤的发生和发展过程,为肿瘤诊断和治疗提供帮助。

此外,现有数据库主要针对白种人构建,而亚洲人种数据库还尚处于起步阶段,存在数据量少、数据类型单一、临床信息不全面等缺陷,但近年来也有一些成果出现,如中国科学院的生命与健康大数据中心等<sup>[36]</sup>,相信随着政府部门的重视和国内测序产业的发展,黄种人多组学数据库也将逐步完善,成为肿瘤基因组学研究的新支柱。

## 参考文献(References):

- [1] George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol*, 2014, 21: 686–694. [DOI]
- [2] Rasmussen L, Pratt N, Hansen MR, Hallas J, Pottgard A. Using the "proportion of patients covered" and the Kaplan-Meier survival analysis to describe treatment persistence. *Pharmacoepidemiol Drug Saf*, 2018, 27: 867–871. [DOI]
- [3] Hsu CH, Yu M. Cox regression analysis with missing covariates via nonparametric multiple imputation. *Stat Methods Med Res*, 2018, 962280218772592. [DOI]
- [4] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 2015, 43: e47. [DOI]
- [5] Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*, 1999, 9: 1106–1115. [DOI]
- [6] Bunker R, Mallet RT. Metabolomics and receiver operating characteristic analysis: a promising approach for sepsis diagnosis. *Crit Care Med*, 2016, 44: 1784–1785. [DOI]
- [7] Grau J, Grosse I, Keilwagen J. PRROC: computing and



- visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 2015, 31: 2595–2597. [DOI]
- [8] Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, Raitakari OT, Jarvelin MR, Salomaa V, Ala-Korpela M, Ripatti S, Pirinen M. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*, 2016, 32: 1981–1989. [DOI]
- [9] Dimou NL, Tsigirgos KD, Elofsson A, Bagos PG. GVAR: robust analysis and meta-analysis of genome-wide association studies. *Bioinformatics*, 2017, 33: 1521–1527. [DOI]
- [10] Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, 2011, 17: 297–303. [DOI]
- [11] Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, 2015, 19: A68–77. [DOI]
- [12] Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*, 2000, 100: 57–70. [DOI]
- [13] Sirintrapun SJ, Zehir A, Syed A, Gao J, Schultz N, Cheng DT. Translational bioinformatics and clinical research (biomedical) informatics. *Clin Lab Med*, 2016, 36: 153–181. [DOI]
- [14] Li QK, Pavlovich CP, Zhang H, Kinsinger CR, Chan DW. Challenges and opportunities in the proteomic characterization of clear cell renal cell carcinoma (ccRCC): a critical step towards the personalized care of renal cancers. *Semin Cancer Biol*, 2018, DOI:10.1016/j.semcancer.2018.06.004. [DOI]
- [15] Smith CC, Beckermann KE, Bortone DS, de Cubas AA, Bixby LM, Lee SJ, Panda A, Ganesan S, Bhanot G, Wallen EM, Milowsky MI, Kim WY, Rathmell WK, Swannstrom R, Parker JS, Serody JS, Selitsky SR, Vincent BG. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. *J Clin Invest*, 2018, 128(11): 4804–4820. [DOI]
- [16] Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet*, 2016, 17: 257–271. [DOI]
- [17] Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Cancer Genome Atlas Research N, Liang H. A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell*, 2018, 173: 386–399 e312. [DOI]
- [18] Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. *Nat Rev Mol Cell Biol*, 2018, 20(1): 21–37. [DOI]
- [19] Rupaimoole R, Slack FJ. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat Rev Drug Discov*, 2017, 16: 203–222. [DOI]
- [20] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*, 2009, 136: 215–233. [DOI]
- [21] Cortez MA, Ivan C, Valdecanas D, Wang X, Peltier HJ, Ye Y, Araujo L, Carbone DP, Shilo K, Giri DK, Kelnar K, Martin D, Komaki R, Gomez DR, Krishnan S, Calin GA, Bader AG, Welsh JW. PDL1 Regulation by p53 via miR-34. *J Natl Cancer Inst*, 2016, 108. [DOI]
- [22] Boyd SD. Diagnostic applications of high-throughput DNA sequencing. *Annu Rev Pathol*, 2013, 8: 381–410. [DOI]
- [23] Lasken RS, McLean JS. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet*, 2014, 15: 577–584. [DOI]
- [24] McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shaperro MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 2008, 40: 1166–1174. [DOI]
- [25] Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. *Science*, 2017, 357(6348): pii: eaal2380. [DOI]
- [26] Okugawa Y, Grady WM, Goel A. Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers. *Gastroenterology*, 2015, 149: 1204–1225 e1212. [DOI]
- [27] Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. *Lancet*, 2018, 392(10149): 777–786. [DOI]
- [28] Lu Y, Ling S, Hegde AM, Byers LA, Coombes K, Mills GB, Akbani R. Using reverse-phase protein arrays as pharmacodynamic assays for functional proteomics, biomarker discovery, and drug development in cancer. *Semin Oncol*, 2016, 43: 476–483. [DOI]
- [29] Jiang T, Shi W, Wali VB, Pongor LS, Li C, Lau R, Gyorffy B, Lifton RP, Symmans WF, Pusztai L, Hatzis C. Predictors of chemosensitivity in triple negative breast cancer: an integrated genomic analysis. *PLoS Med*, 2016, 13: e1002193. [DOI]
- [30] Salem O, Erdem N, Jung J, Munstermann E, Worner A, Wilhelm H, Wiemann S, Korner C. The highly expressed



- 5'isomiR of hsa-miR-140-3p contributes to the tumor-suppressive effects of miR-140 by reducing breast cancer proliferation and migration. *BMC Genomics*, 2016, 17: 566. [DOI]
- [31] Gibori H, Eliyahu S, Krivitsky A, Ben-Shushan D, Epshtein Y, Tiram G, Blau R, Ofek P, Lee JS, Ruppin E, Landsman L, Barshack I, Golan T, Merquiol E, Blum G, Satchi-Fainaro R. Amphiphilic nanocarrier-induced modulation of PLK1 and miR-34a leads to improved therapeutic response in pancreatic cancer. *Nat Commun*, 2018, 9: 16. [DOI]
- [32] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedomallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CE, Cancer Genome Atlas Research N, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich L. The immune landscape of cancer. *Immunity*, 2018, 48: 812–830 e814. [DOI]
- [33] Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, Tiezzi DG, Wang C, Cherniack AD, Godwin AK, Kuderer NM, Rader JS, Zuna RE, Sood AK, Lazar AJ, Ojesina AI, Adebamowo C, Adebamowo SN, Baggerly KA, Chen TW, Chiu HS, Lefever S, Liu L, MacKenzie K, Orsulic S, Roszik J, Shelley CS, Song Q, Vellano CP, Wentzensen N, Cancer Genome Atlas Research N, Weinstein JN, Mills GB, Levine DA, Akbani R. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, 2018, 33: 690–705 e699. [DOI]
- [34] Yu KH, Fitzpatrick MR, Pappas L, Chan W, Kung J, Snyder M. Omics analysis system for precision oncology (OASISPRO): a web-based omics analysis tool for clinical phenotype prediction. *Bioinformatics*, 2018, 34(2): 319–320. [DOI]
- [35] Fisch KM, Meissner T, Gioia L, Ducom JC, Carland TM, Loguercio S, Su AI. Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics*, 2015, 31: 1724–1728. [DOI]
- [36] Zhang YS, Xia L, Sang J, Li M, Liu L, Li MG, Niu GY, Cao JB, Teng XF, Zhou Q, Zhang Z. The BIG Data Center's database resources. *Hereditas(Beijing)*, 2018, 40(11): 1039–1043.
- 张源笙, 夏琳, 桑健, 李漫, 刘琳, 李萌伟, 牛广艺, 曹佳宝, 滕徐菲, 周晴, 章张. 生命与健康大数据中心资源. *遗传*, 2018, 40(11): 1039–1043. [DOI]

(责任编辑: 方向东)