

• 生物信息专栏 •



口腔鳞状细胞癌遗传变异数据库

施雯靖¹, 潘贤润², 吕哲宇³, 詹超英¹, 沈百荣^{1*}

(1. 四川大学华西医院疾病分子网络前沿科学中心 成都 610212; 2. 成都中医药大学医学技术学院 成都 610075;

3. 四川化工职业技术学院数字经济学院 四川 泸州 646300)

【摘要】口腔鳞状细胞癌 (OSCC) 作为一种特发在口腔的恶性肿瘤, 属于头颈部最恶性、危害最大的肿瘤。口腔鳞状细胞癌的发生发展具有很强的异质性, 是遗传和环境包括生活习惯共同作用的结果。为了在系统遗传学层次上了解口腔鳞状细胞癌发生发展的机制, 迫切需要一个口腔鳞状细胞癌遗传变异的数据库。通过自然语言处理结合人工收集、整理了 PubMed 文献中口腔鳞状细胞癌相关的遗传变异数据, 构建了口腔鳞状细胞癌遗传变异数据库 (GVDoscc)。该数据库包含从 334 篇原始文章中提取的 1 020 种遗传变异和 608 条临床样本信息, 为口腔鳞状细胞癌的变异规律寻找和分子机制研究提供了一个可靠的开源平台。

关键词 数据库; 遗传变异; 口腔鳞状细胞癌; 系统遗传

中图分类号 TP392; Q319 文献标志码 A doi:10.12178/1001-0548.2022044

A Genetic Variation Database for Oral Squamous Cell Carcinoma

SHI Wenjing¹, PAN Xianrun², LYU Zheyu³, ZHAN Chaoying¹, and SHEN Bairong^{1*}

(1. Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University Chengdu 610212;

2. College of Medical Technology, Chengdu University of Traditional Chinese Medicine Chengdu 610075;

3. College of Digital Economy, Sichuan Polytechnic of Chemical Industry Luzhou Sichuan 646300)

Abstract Oral squamous cell carcinoma (OSCC) is a malignant tumor idiopathic in the oral cavity, which is the most malignant and most harmful tumor of head and neck. The occurrence and development of oral squamous cell carcinoma is very heterogeneous, caused by the interaction between heredity and environment, including personalized lifestyles. To understand the complex mechanisms of OSCC at the systems biological level, there is an urgent need for a OSCC genetic variation database. In this research, the OSCC genetic variation database (GVDoscc) was established by both natural language processing and manually curating of data from PubMed. GVDoscc contains 1 020 genetic variations and 608 clinical sample records extracted from 334 original articles, providing a reliable and accurate open-source for OSCC associated genetic variation pattern identification and molecular mechanism investigation.

Key words database; genetic variation; oral squamous cell carcinoma; systems genetics

口腔鳞状细胞癌 (oral squamous cell carcinoma, OSCC) 是一种特发在口腔的恶性肿瘤, 癌变的鳞状细胞可发生在牙龈、硬腭、舌、颊粘膜、唇等器官, 属于头颈部最恶性的肿瘤, 超过 50% 的头颈部癌症为口腔鳞状细胞癌^[1]。2020 年, 全球范围内口腔癌死亡病例 177 757 例 (占全部癌症的 1.8%), 新增病例 377 713 例 (占全部癌症的 2%)^[2], 而超过 90% 的口腔癌为口腔鳞状细胞癌^[3]。尽管发病率

和死亡率从 1975 年起有所下降^[4], 但目前没有针对口腔癌的常规筛查测试或计划, 也没有血液测试可以进行诊断, 几乎一半口腔癌或口咽癌在首次诊断时已扩散到淋巴结。2006 年, 以 EGRF 为靶点的西妥昔单抗被 FDA 批准用于治疗口腔鳞状细胞癌, 是目前唯一可用于口腔鳞状细胞癌分子靶向疗法^[5], 西妥昔单抗的有效性与病人的遗传和基因表达谱有关^[6]。

收稿日期: 2022-02-06; 修回日期: 2022-04-06

基金项目: 国家自然科学基金 (32070672)

作者简介: 施雯靖 (1997-), 女, 主要从事生物信息学方面的研究。

*通信作者: 沈百荣, E-mail: bairong.shen@scu.edu.cn

烟草和酒精是口腔鳞状细胞癌最强的两个生活习惯相关的风险因素。文献 [7] 指出, 口腔癌的发病风险会随着吸烟时间和每日吸烟频率的增加而显著增加。烟草中发现的许多化学物质会直接破坏 DNA, 造成生长失控。目前的研究并不确定酒精是否会直接损害 DNA, 但已经有研究证明, 酒精有助于许多破坏 DNA 的化学物质更容易进入细胞^[8], 并且酗酒和大量吸烟的人患口腔癌的风险比不吸烟不喝酒的人高 30 倍^[9]。遗传突变决定了部分人对某些类型的致癌化学物质进行分解的能力较差, 因此他们对烟草、酒精和某些工业化学品的致癌作用更为敏感。文献 [6] 提出与非 OSCC 的头颈癌和其他鳞状细胞癌相比, HRAS 和 PIK3CA 是唯一显著突变的致癌基因, 并且此类突变的分布在具有不同风险习惯 (吸烟、过度饮酒或嚼槟榔) 的人群中也是不同的。

2011 年, 文献 [10] 发布了口腔癌相关基因数据库 OCDB v.2, 包括了 374 个基因。但其相关性的可信度不足, 数据库中并不能提供明确的报道证实某一基因与口腔癌相关。该数据库的初始版本距今已过十年, 数据在此期间没有进行更新。文献 [11] 建立了哥本哈根口腔鳞状细胞癌数据库, 但该研究重点关注患者的临床信息, 没有收录关于基因或突变方面的数据, 且出于隐私保护的考虑, 该数据库并未公开发表。文献 [12] 于 2021 年发表了口腔癌基因组变异数据库 dbGENVOC, 但其数据仅包括 325 名印度口腔癌患者和手动整理出版文献中的 118 名患者数据, 同时只简单纳入了变异的描述信息, 对患者没有任何描述。遗传变异与疾病的关系并不是单一的, 在不同的患者间也存在异质性的问题。因此, 上述 3 个数据库对临床的参考与应用价值有限。对于目前已有的遗传变异数据库, 如 Clinvar 和 OMIM, 尽管它们包含的疾病种类非常全面, 但并未收录研究对象的临床信息, 也就无法直接在临床上利用其数据。

精准医疗和数据驱动的第四科学研究范式的发展^[13], 需要基因组学和临床组学融合的数据。本文通过把碎片性的口腔鳞状细胞癌相关遗传变异与对应的临床信息融合起来, 构建了口腔鳞状细胞癌遗传变异数据库 (the OSCC genetic variation database, GVDoscc) (<http://sysbio.org.cn/GVDoscc>)。为口腔领域的相关工作者提供了高质量的开源数据集。

1 材料方法

1.1 数据来源

从美国国家生物信息技术中心 (the national center for biotechnology information, NCBI) 建立的 Pubmed (<https://pubmed.ncbi.nlm.nih.gov/>) 数据库使用检索词 “(Oral Squamous Cell Carcinoma[ti] OR OSCC[ti]) NOT review[ptyp] AND English[LA] AND (1991/01/01[DP]: 2021/7/1[DP])” 获取英文非综述类文献 6 137 篇, 作为本文的数据源。

1.2 技术路线

数据库的收集、搭建流程和功能简介如图 1 所示, 从 Pubmed 数据库获取到口腔鳞状细胞癌的相关文献后, 利用 PubTator 对全文中的基因或突变进行注释。PubTator 是一种网络服务工具, 为基因与蛋白质、遗传变异、疾病、化学物质、物种和细胞系提供来自文本挖掘系统的自动注释^[14]。若整篇文章中都未曾出现一次基因或突变的名称, 则删去此文献; 然后对筛查得到的 3 884 篇文献进行人工注释, 并对 Pubtator 的注释进行二次核查。数据库采用 B/S 结构 (Browser/Server, 浏览器/服务器模式), 用户可以通过浏览器直接进行访问, 前端网页由 html 和 JavaScript 完成, 后端服务器采用 Apache 搭建而成, PHP 用于连接后台数据库, 数据存储存储在 MySQL 数据库中。用户可以通过在线访问数据库来进行分类浏览或检索自己需要的信息, 统计页面也提供了对所有收录数据的整体统计, 用户还可以在提交页面上传新数据, 经管理员审核通过后, 将会被添加进数据库中。

1.2.1 筛选标准

本研究在文献初筛后, 基于人工阅读方式, 采用以下纳入和排除标准对文献进行筛选。

纳入标准。1) 在 Pubmed 数据库中公开发表的有关口腔鳞状细胞癌遗传变异的流行病学研究; 2) 具有临床上确诊为口腔鳞状细胞癌的患者及其对照样本信息。

排除标准。1) 评论、社论等不予收录; 2) 病例报告所包含的样本量太少, 不具有普适性, 不予收录; 3) 综述及 meta 分析文章, 追溯文中数据来源原文按纳入排除标准处理。

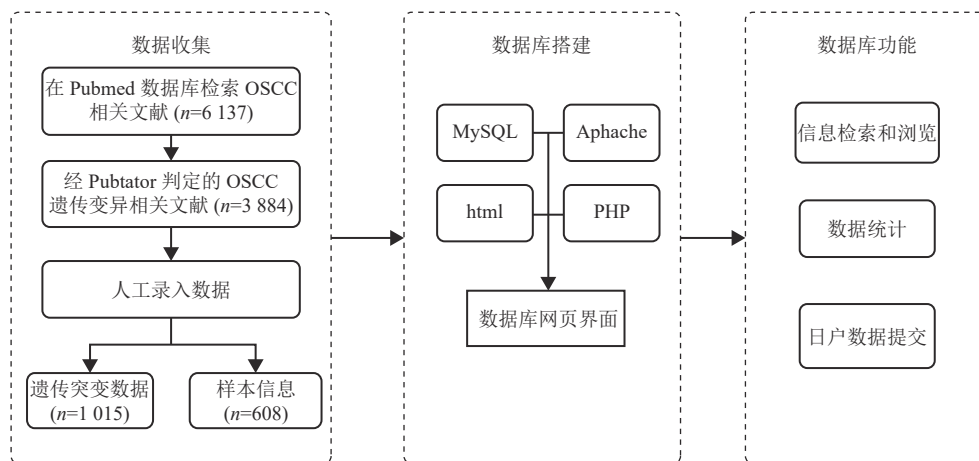


图1 GVDoscc 数据库搭建流程

1.2.2 数据库结构

数据库主要由3张实体表和1张关系主表(Main)组成, 实体表包括样临床样本表(Sample)、参考文献表(Reference)以及变异表(Mutation)。具体描述如下。

样本表: 样本编号(Sample ID)、试验对象年龄(Age)、试验对象国籍(Country)、试验人数(Size)、性别(Gender)、样本来源(Source);

参考文献表: PMID(该文献在Pubmed中的编号)、发表年份(Year)、文献题目(Title)、作者(Author)和期刊名(Journal);

变异表: 变异编号(Mutation ID)、变异分类(Classification)、变异名称(Name)、变异位点(Position)、碱基变化(Base variation)、变异结果(Consequence)、基因(Symbol)、基因所在位点(Gene location)、氨基酸变化(Amino Acid variation);

主表: 数据唯一的主编号(Main ID)、变异所导致的病理作用(Pathological Function)、临床特征(Clinical Significance)、临床应用(Application)、患者的变异频率(Patient-frequency)、对照的变异频率(Control-frequency)、结论(Summary)。

1.3 生物信息学分析

1.3.1 PPI 网络分析

String 数据库旨在收集、评分和整合所有公开可用的蛋白质-蛋白质相互作用信息来源, 并通过计算预测来补充这些信息。String 目前版本为 v11.5, 已收录人类(Homo sapiens)的 19 303 个蛋白^[15]。将收录的基因映射到 String 数据库中(置信度>0.9), 构建了蛋白质互作(protein protein interaction, PPI)网络, 并通过 Cytoscape 计算节点的度

中心性, 以降序排列, 取前 5% 为关键基因。Cytoscape 是一个专注于网络可视化和分析的开源软件项目, 用于将生物分子交互网络、高通量基因表达数据和其他的分子状态信息整合在一起^[16]。

1.3.2 功能富集分析

基因本体(gene ontology, GO)是一个在生物信息学领域中广泛使用的本体, 包含了生物学领域知识体系本质的表示形式, 从基因的分子功能、生物过程和细胞组分这3方面对目标进行注释^[17]。GO 可用于系统分析和标注基因产物的功能; KEGG 数据库则是用于分子水平信息, 尤其是基因组测序等高通量生成的大规模数据集, 来了解细胞、生物和生态等生物系统的高级功能和效用^[18]; KEGG 的通路数据库可用于系统层次的疾病机制的分析。在数据库构建的基础上, 本文对数据库中的基因进行了 GO 富集分析(P 值<0.05)和 KEGG 通路富集分析(P 值<0.05, 富集倍数>2)。

2 结果

2.1 数据统计

在经过上述的筛选和整理后, 本文共收录了 1 020 种口腔鳞状细胞癌相关遗传变异, 涉及到 436 个蛋白编码基因和 18 个 miRNA 基因, 相关临床样本总数为 82 863 个。其中, 明确标明为男性和女性的分别为 44 475 个和 15 169 个, 具体数据统计见表 1。其中, 部分样本数据没有注释性别, 导致男性样本和女性样本人数之和小于样本总人数。

根据在原始试验中, 该遗传变异参与的正常生物学过程、病理过程及治疗干预药理学反应, 将其按生物标志物类型分为诊断、治疗和预后 3 类

(8%、29%、30%)，其中，非单一类型的生物标志物(占诊断、治疗和预后的 2 种及以上)共 182 个，占 33%。

表 1 数据库基本数据统计

| 遗传变异种类 | 计数/种 |
|---------|------|
| 基因拷贝数变异 | 278 |
| 多态性 | 78 |
| 杂合性丢失 | 24 |
| 甲基化 | 278 |
| 单核苷酸多态性 | 198 |
| 点突变 | 50 |
| 其他 | 114 |

2.2 网页设计

按照数据库的结构和应用需求，本文设计了 7 个在线使用的功能模块：网页基本介绍、分类浏览、高级检索、统计结果展示、数据提交、用户帮助和管理维护。

基本介绍：对数据库的内容进行基本介绍，展示引用相关资源、相关数据库和研究机构网址等信息。

分类浏览：用户可根据 3 种分类方式(变异类型、临床样本来源、突变的相关基因分类等)对数据进行浏览。

高级检索：用户根据突变名称、数据来源(地点)、基因名称等条件检索需要的信息。

统计结果：主要由遗传变异的分类、频率排名前 10 位的基因、突变在染色体上分布和临床样本来源的地理位置分布热图组成。

数据提交：用户只需填写 4 个条目，包括突变名称、参考文献、联系邮箱和细节，即可提交新数据，其中突变名称与邮箱为必填项。在审核成功后，系统会将新的数据添加进本数据库中。

用户帮助：详细介绍了数据库的操作方式，确保用户能够正确查找到自己需要的信息。

管理维护：对数据库进行更新与维护，更新间隔保持在 3~6 个月，在数据更新时，相应的统计也会进行更新。

2.3 生物信息学分析结果

2.3.1 PPI 分析结果

首先将收录的 436 个编码基因通过 String 映射到 PPI 网络中，成功映射了 256 个节点与 1 085 条关系。通过度中心性，共筛选出了 12 个关键基因，如图 2 与表 2 所示，包括 PIK3R1、TP53、PIK3CA、MAPK1、HRAS、AKT1、KRAS、RHOA、MYC、EGFR、RELA 以及 NRAS。

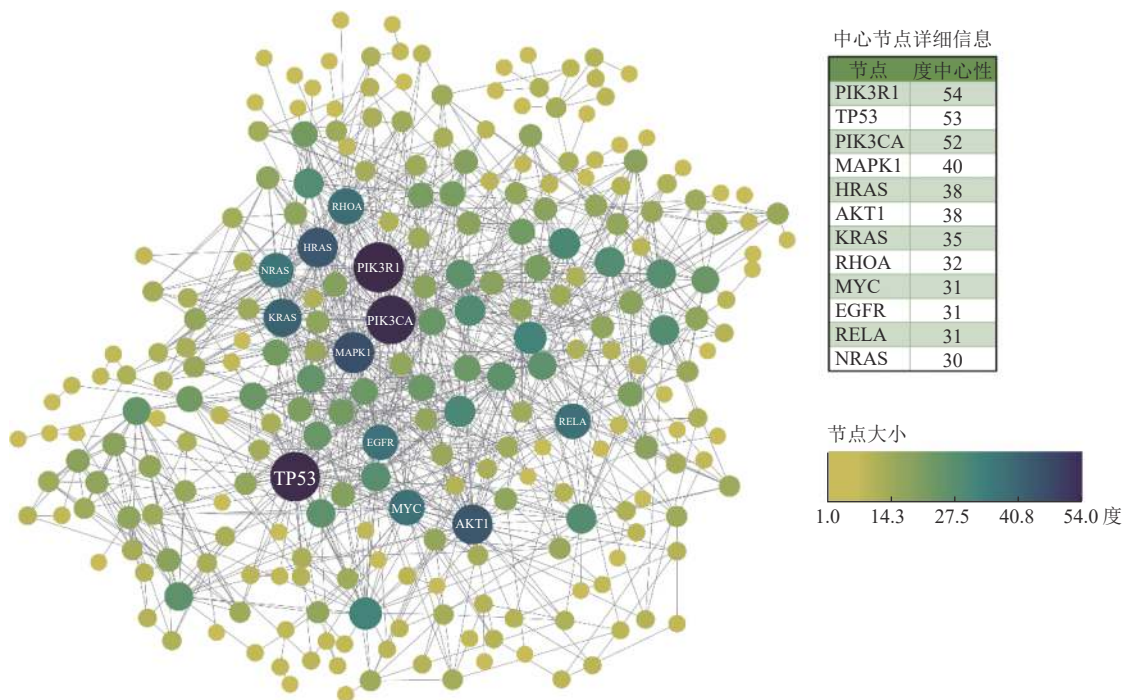


图 2 口腔鳞状细胞癌相关基因所构成的蛋白质相互作用网络

表 2 关键基因在 OSCC 中的功能

| 基因 | 在OSCC中的功能 | PMID编号 |
|--------|-----------------------------|----------|
| PIK3R1 | 细胞粘附和发育 | 32115621 |
| TP53 | 抑癌基因, 调控细胞凋亡或细胞周期 | 31949709 |
| PIK3CA | 肿瘤生长 | 31516747 |
| MAPK1 | 与患癌风险高度相关 | 25265753 |
| HRAS | 转导信号, 指示细胞生长和分裂 | 31845386 |
| AKT1 | 调控细胞增殖和生长, 参与细胞凋亡和葡萄糖代谢 | 27889930 |
| KRAS | 调控细胞生长和活化 | 20813562 |
| RHOA | 促进肌动蛋白细胞骨架的重组并调节细胞的形状、附着和运动 | 21334929 |
| MYC | 促进细胞增殖, 永生化, 去分化和转化 | 26708050 |
| EGFR | 参与细胞的增殖、血管生成、肿瘤侵袭、转移及细胞凋亡 | 33680380 |
| RELA | 肿瘤的发生 | 33710605 |
| NRAS | 转导信号, 指示细胞生长和分裂 | 28938622 |

2.3.2 功能富集分析结果

将收录的基因进行 GO 富集分析, 得到了 549 个生物过程、63 个细胞组分和 106 个分子功能。在经过 KEGG 通路分析后, 得到了 118 个相关通路, 对上述 4 种分析方式的前 10 条结果按基因数目降序排列, 见表 3~表 6。

表 3 根据基因数目排列的前十位生物过程

| GO编号 | 生物过程 | 基因数目/个 | P值 |
|---------|-------------------|--------|------------------------|
| 0045944 | RNA聚合酶II启动子转录的正调控 | 79 | 7.71×10 ⁻²¹ |
| 0006351 | 转录, DNA模板 | 65 | 9.62×10 ⁻³ |
| 0006355 | 转录调控, DNA模板 | 50 | 2.47×10 ⁻² |
| 0043066 | 凋亡过程的负调控 | 48 | 4.51×10 ⁻¹⁷ |
| 0045893 | 转录的正调控, DNA模板 | 48 | 6.82×10 ⁻¹⁵ |
| 0008285 | 细胞增殖的负调控 | 46 | 5.74×10 ⁻¹⁸ |
| 0000122 | RNA聚合酶I转录的负调控 | 45 | 2.41×10 ⁻⁸ |
| 0007165 | 信号转导 | 43 | 7.50×10 ⁻³ |
| 0006915 | 凋亡过程 | 42 | 6.10×10 ⁻¹⁰ |
| 0016049 | 细胞生长 | 40 | 7.17×10 ⁻¹³ |

由表 3 可见, 富集最多在 RNA 聚合酶 II 启动子转录的正调控 (17.59%) 上, RNA 聚合酶 II 的作用是催化 DNA 转录, 合成 mRNA 及大多数 hnRNA 和 miRNA 的前体; 细胞组分: 超过 30% 的基因富集在细胞核 (43.43%)、细胞质 (37.42%) 和细胞膜 (32.74%) 上; 分子功能: 富集结果都与结合密切相

关, 其中蛋白结合 (64.14%) 上在此次分析中富集的基因最多, 牵涉了 288 个基因。在经过 KEGG 通路分析后发现基因显著富集在癌症途径 (pathways in cancer) 上, 并且参与了 PI3K-Akt 信号通路和黏着斑 (focal adhesion)。

表 4 根据基因数目排列的前十位细胞成分

| GO编号 | 细胞成分 | 基因数目/个 | P值 |
|---------|---------|--------|------------------------|
| 0005634 | 细胞核 | 195 | 5.85×10 ⁻¹³ |
| 0005737 | 细胞质 | 168 | 5.93×10 ⁻⁷ |
| 0005886 | 细胞质膜 | 147 | 7.90×10 ⁻⁹ |
| 0005829 | 胞质溶胶 | 128 | 1.08×10 ⁻⁹ |
| 0005654 | 细胞核质 | 105 | 2.36×10 ⁻⁷ |
| 0070062 | 外泌体 | 91 | 6.39×10 ⁻⁴ |
| 0005576 | 胞外区 | 75 | 6.93×10 ⁻⁹ |
| 0016020 | 植物细胞膜 | 66 | 2.18×10 ⁻² |
| 0005615 | 胞外空间 | 62 | 3.02×10 ⁻⁷ |
| 0005887 | 质膜的组成部分 | 51 | 1.74×10 ⁻³ |

表 5 根据基因数目排列的前十位分子功能

| GO编号 | 分子功能 | 基因数目/个 | P值 |
|---------|--------------------|--------|------------------------|
| 0005515 | 蛋白质结合 | 288 | 1.56×10 ⁻¹⁵ |
| 0046872 | 金属离子结合 | 70 | 2.41×10 ⁻³ |
| 0005524 | ATP结合 | 67 | 7.05×10 ⁻⁷ |
| 0003677 | DNA结合 | 59 | 2.41×10 ⁻³ |
| 0042802 | 相同蛋白质结合 | 50 | 1.41×10 ⁻¹⁰ |
| 0003700 | 转录因子活性, 序列特异性DNA结合 | 47 | 5.41×10 ⁻⁶ |
| 0008270 | 锌离子结合 | 43 | 5.24×10 ⁻³ |
| 0042803 | 蛋白质同源二聚化活性 | 42 | 3.44×10 ⁻⁷ |
| 0019899 | 酶结合 | 34 | 4.65×10 ⁻¹² |
| 0005509 | 钙离子结合 | 34 | 2.51×10 ⁻⁴ |

表 6 根据基因数目排列的前十位 KEGG 通路

| KEGG编号 | KEGG通路 | 基因数目/个 | P值 |
|--------|--------------|--------|------------------------|
| 05200 | 癌症途径 | 71 | 7.15×10 ⁻²⁹ |
| 05218 | 黑色素瘤 | 28 | 1.77×10 ⁻²⁰ |
| 05220 | 慢性髓性白血病 | 28 | 2.73×10 ⁻²⁰ |
| 05219 | 膀胱癌 | 22 | 1.76×10 ⁻¹⁹ |
| 04151 | PI3K-Akt信号通路 | 55 | 2.04×10 ⁻¹⁹ |
| 05205 | 癌症中的蛋白多糖 | 40 | 1.70×10 ⁻¹⁷ |
| 05210 | 结直肠癌 | 24 | 2.56×10 ⁻¹⁷ |
| 05206 | 癌症中的miRNA | 47 | 4.65×10 ⁻¹⁷ |
| 05230 | 癌症的中枢代谢 | 24 | 5.84×10 ⁻¹⁷ |
| 05213 | 子宫内膜癌 | 22 | 8.35×10 ⁻¹⁷ |

3 结束语

基于基因表达模式的分类表明, OSCC 可以分

为不同的亚型,文献[19]通过免疫组化发现在不同 OSCC 亚型中, EGFR、MTDH 和 ERCC1 的表达水平存在显著的差别。文献[20]发现不同的肿瘤亚型对不同类型的疗法存在优先反应^[20]。因此,需要高质量的数据库来对患者的免疫状态进行精确判断,本文将 Pubmed 中碎片化的结论收集起来,基于 1991–2021 年间的 334 篇文献中的 1 020 条遗传变异数据,构建了一个全面、专业的口腔鳞状细胞癌数据库,对变异进行分类和对样本信息进行归纳。在生物信息学分析后发现, PPI 的关键基因包括 PIK3、RAS 家族和基因 TP53 等。PI3K 参与的 PI3K-Akt 通路,调节了细胞的增殖分化、凋亡等多种功能,在 KEGG 同样也被富集,其过度活化与人体肿瘤的发展密切相关^[21],早在 2006 年就有第一个新型 PI3K 抑制剂开始了临床试验^[22]; RAS 基因的突变存在于 30% 的人类恶性肿瘤中,其功能是调节细胞的分化增殖,被称为细胞信号网络传递中的“分子开关”^[23]; p53 作为 TP53 基因表达的蛋白产物,在人类恶性肿瘤中普遍存在^[24],被认为是目前最重要的抑癌基因。GO 分析结果主要富集在细胞的转录和调控、基本组成结构(细胞核、细胞膜、细胞质)上,说明此类基因与肿瘤的发生、发展或转移密切相关。

癌症是复杂的、异质的和动态的疾病,将肿瘤的基因分析与个体生理状态的深度纵向分析(深度表型)相结合是预防、诊断和治疗癌症的关键,而挖掘深度表型首先需要提高诊断的精度^[25]。从建模的角度,可以通过对不同的生物标志物进行组合和建模,提高预测口腔鳞状细胞癌的发生和进展的准确性。未来将进一步更新数据库的内容和结构,同时基于数据库建立知识图谱和知识引导的模型,通过结合临床研究、对口腔鳞状细胞癌的基因变异异质性进行深度分析,为临床的精准诊疗提供数据和信息支撑。

参 考 文 献

- [1] EL-NAGGAR A K, CHAN J K, GRANDIS J R. WHO classification of head and neck tumours[M]. [S.l.]: WHO/IARC Classification of Tumours, 2017: 200-210.
- [2] SUNG H, FERLAY J, SIEGEL R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. *CA Cancer J Clin*, 2021, 71(3): 209-249.
- [3] 张陈平. 口腔癌治疗规范的思考[J]. *中国肿瘤临床*, 2010, 37(16): 905-907.
ZHANG C P. Consideration of standardized treatment protocols for oral cancer[J]. *Chinese Journal of Clinical Oncology*, 2010, 37(16): 905-907.
- [4] National Cancer Institute. SEER cancer stat facts: Oral cavity and pharynx cancer[EB/OL]. [2022-01-07]. <https://seer.cancer.gov/statfacts/html/oralcav.html>.
- [5] BONNER J A, HARARI P M, GIRALT J, et al. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck[J]. *N Engl J Med*, 2006, 354(6): 567-578.
- [6] CHAI A W Y, LIM K P, CHEONG S C. Translational genomics and recent advances in oral squamous cell carcinoma[J]. *Semin Cancer Biol*, 2020, 61: 71-83.
- [7] 刘宇, 罗治彬, 王彝, 等. 212 例口腔癌患者的流行病学危险因素分析[J]. *实用癌症杂志*, 2014, 29(2): 160-161, 167.
LIU Y, LUO Z B, WANG Y, et al. Epidemiological analysis of risk factors of 212 cases of oral cancer[J]. *The Practical Journal of Cancer*, 2014, 29(2): 160-161, 167.
- [8] REIDY J T, MCHUGH E E, STASSEN L F. A review of the role of alcohol in the pathogenesis of oral cancer and the link between alcohol-containing mouthrinses and oral cancer[J]. *J Ir Dent Assoc*, 2011, 57(4): 200-202.
- [9] TURATI F, GARAVELLO W, TRAMACERE I, et al. A meta-analysis of alcohol drinking and oral and pharyngeal cancers. Part 2: Results by subsites[J]. *Oral Oncol*, 2010, 46(10): 720-726.
- [10] GADEWAL N S, ZINGDE S M. Database and interaction network of genes involved in oral cancer: Version II[J]. *Bioinformatics*, 2011, 6(4): 169-170.
- [11] SCHMIDT JENSEN J, JAKOBSEN K K, MIRIAN C, et al. The Copenhagen oral cavity squamous cell carcinoma database: Protocol and report on establishing a comprehensive oral cavity cancer database[J]. *Clin Epidemiol*, 2019, 11: 733-741.
- [12] PRADHAN S, DAS S, SINGH A K, et al. dbGENVOC: Database of genomic variants of oral cancer, with special reference to India[J]. *Database*, 2021, 2021: baab034.
- [13] SHEN L, BAI J W, WANG J, et al. The fourth scientific discovery paradigm for precision medicine and healthcare: Challenges ahead[J]. *Precis Clin Med*, 2021, 4(2): 80-84.
- [14] WEI C H, ALLOT A, LEAMAN R, et al. PubTator central: Automated concept annotation for biomedical full text articles[J]. *Nucleic Acids Research*, 2019, 47(1): 587-593.
- [15] SZKLARCZYK D, GABLE A L, LYON D, et al. STRING v11: Protein-Protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets[J]. *Nucleic Acids Res*, 2019, 47(1): 607-613.
- [16] SHANNON P, MARKIEL A, OZIER O, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks[J]. *Genome Res*, 2003, 13(11): 2498-2504.
- [17] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: Tool for the unification of biology. The gene ontology consortium[J]. *Nat Genet*, 2000, 25(1): 25-29.
- [18] KANEHISA M, GOTO S. KEGG: Kyoto encyclopedia of genes and genomes[J]. *Nucleic Acids Res*, 2000, 28(1):

- 27-30.
- [19] 陈蔚华, 裴婧, 贾云香. EGFR、MTDH、ERCC1 在口腔特殊亚型的鳞状细胞癌中的表达及临床意义[J]. 口腔医学研究, 2018, 34(6): 623-626.
CHEN W H, PEI J, JIA Y X. Expression and clinical significance of EGFR, MTDH, and ERCC1 in special subtypes of oral squamous carcinoma[J]. Journal of Oral Science Research, 2018, 34(6): 623-626.
- [20] CHUNG C H, PARKER J S, KARACA G, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression[J]. *Cancer Cell*, 2004, 5(5): 489-500.
- [21] 韦柳婷, 冯洁, 莫书荣. PI3K-Akt 信号通路与肿瘤相关性的研究进展[J]. *肿瘤学杂志*, 2014, 20(4): 331-336.
WEI L T, FENG J, MO S R. Progress in correlation of PI3K-Akt signal pathway with tumor[J]. *Journal of Chinese Oncology*, 2014, 20(4): 331-336.
- [22] 赵文楠, 金美花, 孔德新. 新型分子靶向抗癌药物-PI3K 抑制剂[J]. *食品与药品*, 2013, 15(1): 54-59.
ZHAO W N, JIN M H, KONG D X. Novel molecular-targeted antitumor drugs-PI3K inhibitor[J]. *Food and Drug*, 2013, 15(1): 54-59.
- [23] 刘广伟, 张再兴, 戴建军. ras 基因在肿瘤发生、诊断及诊断中的作用研究进展[J]. *山东医药*, 2014, 54(27): 93-95, 98.
LIU G W, ZHANG Z X, DAI J J. Research progress on the role of ras Gene in tumorigenesis, development and diagnosis[J]. *Shandong Medical Journal*, 2014, 54(27): 93-95, 98.
- [24] 李海丽, 邵驰浩, 周若宇, 等. 抑癌基因与癌基因的最新研究进展[J]. *基础医学与临床*, 2018, 38(7): 1029-1033.
LI H L, SHAO C H, ZHUO R Y, et al. Latest progress of the tumor suppressor gene and oncogene[J]. *Basic and Clinical Medicine*, 2018, 38(7): 1029-1033.
- [25] ALAM S, SULTANA A, 王姣, 等. 癌症的异质性表征与深度表型[J]. *中华医学图书情报杂志*, 2019, 28(10): 1-6.
ALAM S, SULTANA A, WANG J, et al. Cancer heterogeneity characterization and deep phenotyping[J]. *Chinese Journal of Medical Library and Information Science*, 2019, 28(10): 1-6.

编辑 刘飞阳