

Analysis of Gene Function Enrichment and Drug Treatment of Oral Mucositis after Radiotherapy and Chemotherapy for Head and Neck Tumor Based on Computer Text Mining

Ting Rong¹, Si Rong², Yifei Zeng³, Chaobin Zhou¹, Jiamei Gu¹, Zhen Li¹, Naidong Wang⁴

¹Qingdao University Medical College, Qingdao Shandong

²Hunan Xiangxi Vocational and Technical College, Jishou Hunan

³Yongzhou Third Hospital, Yongzhou Hunan

⁴Affiliated Hospital of Qingdao University, Qingdao Shandong

Email: wangnaidong163@163.com

Received: Jul. 1st, 2020; accepted: Jul. 14th, 2020; published: Jul. 21st, 2020

Abstract

Objective: Use computer software to mine and analyze the published databases to determine the genes, proteins and signal pathways related to Oral mucositis (OM) and wound healing, and to predict and explore potential drugs that may be effective in the treatment of OM. **Method:** Through the text mining software pubmed2ensembl, the genes related to OM and wound healing were identified. After de duplication, their gene intersections were screened with Venny 2.0. David enriched and analyzed the gene functions of the genes and the path analysis of Kyoto Encyclopedia of genes and genomes (KEGG). String was used to analyze the network of protein-protein interaction and cytcap was used to analyze the module of protein-protein interaction. Gepia further screened the results of module analysis. Finally, dgidb was used to analyze the drug gene interaction. **Result:** After text mining and de duplication, 372 intersecting genes were obtained. The biological process of these genes was analyzed by function enrichment, and KEGG pathway was analyzed. Through protein interaction analysis and module analysis, 37 related genes were further screened. Three genes with statistically significant difference between box plot and survival rate were screened out. Finally, 38 corresponding drugs were screened out. **Conclusion:** Using bioinformatics tools, such as text mining, gene function enrichment analysis, molecular pathway and so on, we can find potential drugs for the treatment of OM.

Keywords

Text Mining, Head and Neck Tumor, Oral Mucositis, Wound Healing

基于计算机文本挖掘的头颈部肿瘤放化疗后口腔粘膜炎症基因功能富集及药物治疗分析

荣婷¹, 荣思², 曾奕斐³, 周超斌¹, 谷家美¹, 李臻¹, 王乃东⁴

¹青岛大学医学院, 山东 青岛

²湖南省湘西职业技术学院, 湖南 吉首

³永州市三医院, 湖南 永州

⁴青岛大学附属医院, 山东 青岛

Email: wangnaidong163@163.com

收稿日期: 2020年7月1日; 录用日期: 2020年7月14日; 发布日期: 2020年7月21日

摘要

目的: 利用计算机软件对已经公开的数据库进行挖掘和分析, 确定口腔粘膜炎症(Oral mucositis, OM)和伤口愈合相关的基因、蛋白、信号通路, 对治疗OM可能有效的潜在药物进行预测和探索。**方法:** 通过文本挖掘软件pubmed2ensembl确定与OM、伤口愈合相关的基因, 去重后, 用Venny 2.0筛选他们基因的交集, DAVID对交集基因进行基因功能的富集分析和京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes, KEGG)通路分析。使用STRING对富集结果进行蛋白质-蛋白质相互作用的网络分析, Cytoscape进行蛋白质-蛋白质相互作用的模块分析。GEPIA进一步筛选模块分析的结果。最后, 使用DGIdb进行分析, 得到药物-基因相互作用的结果。**结果:** 通过文本挖掘去重后, 得到372个交集基因, 对这些基因的生物过程进行功能富集分析, 并分析KEGG通路, 通过蛋白质-蛋白质相互作用分析和模块分析, 进一步筛选出37个相关基因。筛选出箱式图与生存率同时存在统计学差异的3个基因。最后, 筛选出与之对应的38种药物。**结论:** 使用文本挖掘、基因功能富集分析、分子途径等生物信息学工具, 可以发现潜在的治疗OM的药物, 为头颈部肿瘤放化疗后的放射性黏膜炎提供了新的线索。

关键词

文本挖掘, 头颈部肿瘤, 口腔脑膜炎, 伤口愈合

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

放射性粘膜反应是头颈部肿瘤放化疗后主要不良反应之一, 包括口腔黏膜炎、鼻腔黏膜炎以及咽喉炎等, 口腔粘膜炎(Oral mucositis, OM)出现的频率最高[1]。OM表现为在口腔粘膜表面形成片状薄层白色粘膜, 红斑, 弥漫性糜烂, 脱皮, 充血, 出血等, 不仅影响患者正常进食、吞咽、语言等, 还可能导致脱水, 酸中毒, 甚至影响头颈部肿瘤的治疗或者延长其疗程。由于其成因复杂, 目前, 尚无特效药物可以治疗[2]。生物信息学的工具, 可以在众多研究的数据中, 挖掘出相应的、有针对性的基因、蛋白或者信号通路, 有可能提供新的药物研发。本研究利用生物信息学工具, 筛选出可有效治疗OM的基因、分

子途径以及蛋白质 - 蛋白质相互作用, 进一步找到针对性的药物, 提供治疗 OM 的新线索。

2. 资料与方法

2.1. 文本挖掘

使用 pubmed2ensembl (<http://pubmed2ensembl.ls.manchester.ac.uk/>)进行文本挖掘, 关键词为“Oral mucositis”和“wound healing”, 检索出相关基因后, 使用 Venny (<https://bioinfo.cnb.csic.es/tools/venny/>), 取两者交集的基因集, 保存此基因集, 进行下一步数据分析。

2.2. 基因富集分析

DAVID 是一个整合了多个生物学数据、分析工具的生物信息数据库, 提供了大规模基因、蛋白系统综合的生物功能注释信息。应用 DAVID (<https://david.ncifcrf.gov/home.jsp>)进行基因富集分析, 包括生物学过程(Biological process, BP), 细胞成分(Cellular Component, CC), 分子功能(Molecular Function, MF)以及 KEGG 信号通路(KEGG Pathway)分析。我们对保存的基因集进行了 BP、CC、MF、KEGG Pathway 分析。

2.3. 蛋白质 - 蛋白质相互作用网络的分析

将上一步基因富集分析得到的结果应用 STRING (<http://string-db.org>)进行蛋白质-蛋白质相互作用网络分析。STRING 提供了一个基于 PubMed 文本挖掘数据、实验数据以及整合多个数据库资源的平台, 可用来分析蛋白质之间的关系以及相互作用。选择“多种蛋白质” - 种属“Homo” - 检索 - 设置置信区间水平为最高(≥ 0.9)、隐藏单个的蛋白质。

下载 Cytoscape (<https://cytoscape.org/>), 导入基因集, 下载并运行 MCODE、STRING 模块, 所有设置均为默认设置, 即 K-Core 为 2, Node Score Cutoff 为 0.2。Cytoscape 是一个基于 Java 的开放源代码的网络可视化软件, 可绘制基因表达调控网络、蛋白质 - 蛋白质作用网络等与网络、层级有关的数据。选择 Nodes 分别为 37、33 的蛋白质相互作用模块, 构建可视化网络。选择与 OM 密切相关的 Nodes 为 37 的模块 Shared Name, 保存, 并进行下一步分析。

2.4. 药物 - 基因相互作用分析

将上一步保存的基因集作为潜在靶点, 打开 GEPIA (<http://gepia.cancer-pku.cn/detail.php>), 在“expression DIY”中检索基因集中的每一个基因, “[Log2FC] Cutoff”为 1, “P-value Cutoff”取 0.01, 数据集组织选择头颈部鳞状细胞癌“HNSC”, 数据来源选择 TCGA 和 GTEx 数据库。在“Survival”里面选择“Survival Plots”, “Cutoff-High”、“Cutoff-Low”均选择 50%, 数据集组织选择头颈部鳞状细胞癌“HNSC”, 绘制生存曲线图。

利用 DGIdb 工具(http://www.dgidb.org/search_interactions)进行药物 - 基因相互作用分析。DGIdb 包含了 NCBI Entrez、DrugBank、PharmGKB 等临床试验数据库以及 PubMed 已经发表的文献数据等 27 个数据库的信息。选择表达量、生存曲线图均有统计学差异($P < 0.05$)的基因, 点击“Find Drug-Gene Interactions”, 即可得到与此基因相互作用的药物。

3. 结果

3.1. 文本挖掘

通过文本挖掘, 去掉重复基因后, 与 OM 相关的基因为 803 个, 与伤口愈合相关的为 1409 个基因,

Venny2.0 取交集后, 筛选出 372 个与这两个关键词均相关的基因(图 1)。

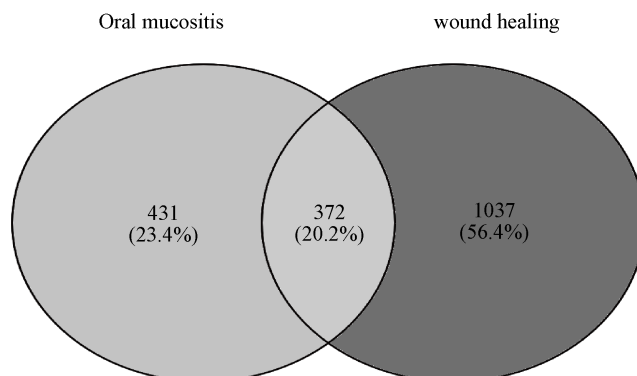


Figure 1. Venny diagram
图 1. Venny 图

3.2. 基因功能富集及 KEGG 信号通路富集分析

通过 DAVID 的基因功能富集分析以及信号通路分析, 按照统计学显著性 P 值大小从小到大的顺序排序, 排在前五的 BP、CC、MF 以及 KEGG Pathway 显示: 对有机物质刺激的反应很敏感, 在细胞因子 - 细胞因子受体相互作用的调节中最丰富(图 2)。

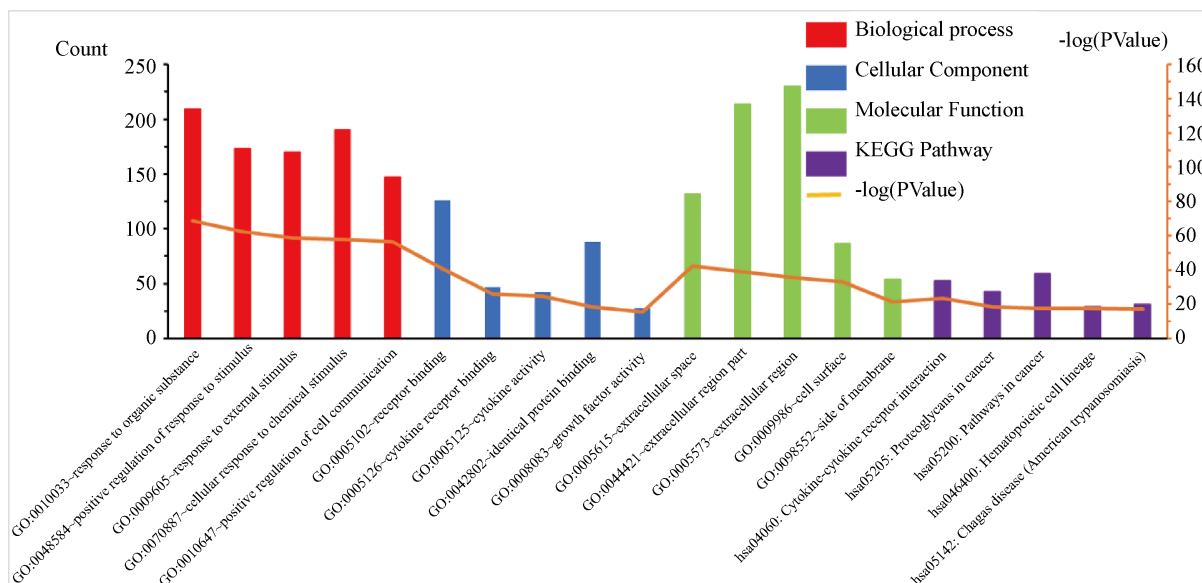


Figure 2. GO diagram
图 2. GO 图

3.3. 蛋白质相互作用分析

利用 STRING, 将交集基因集进行蛋白质 - 蛋白质相互作用分析, 得到蛋白质相互作用网络图。图 3 总共包含 370 个基因, 有 2 个基因没有形成蛋白质 - 蛋白质相互作用网络。利用 Cytoscape 软件的 MCODE 进行模块分析, 排名第一的是节点为 37 的蛋白质相互作用网络图。该分析结果表明主要与细胞因子 - 细胞因子受体相互作用的信号通路有关(图 4)。

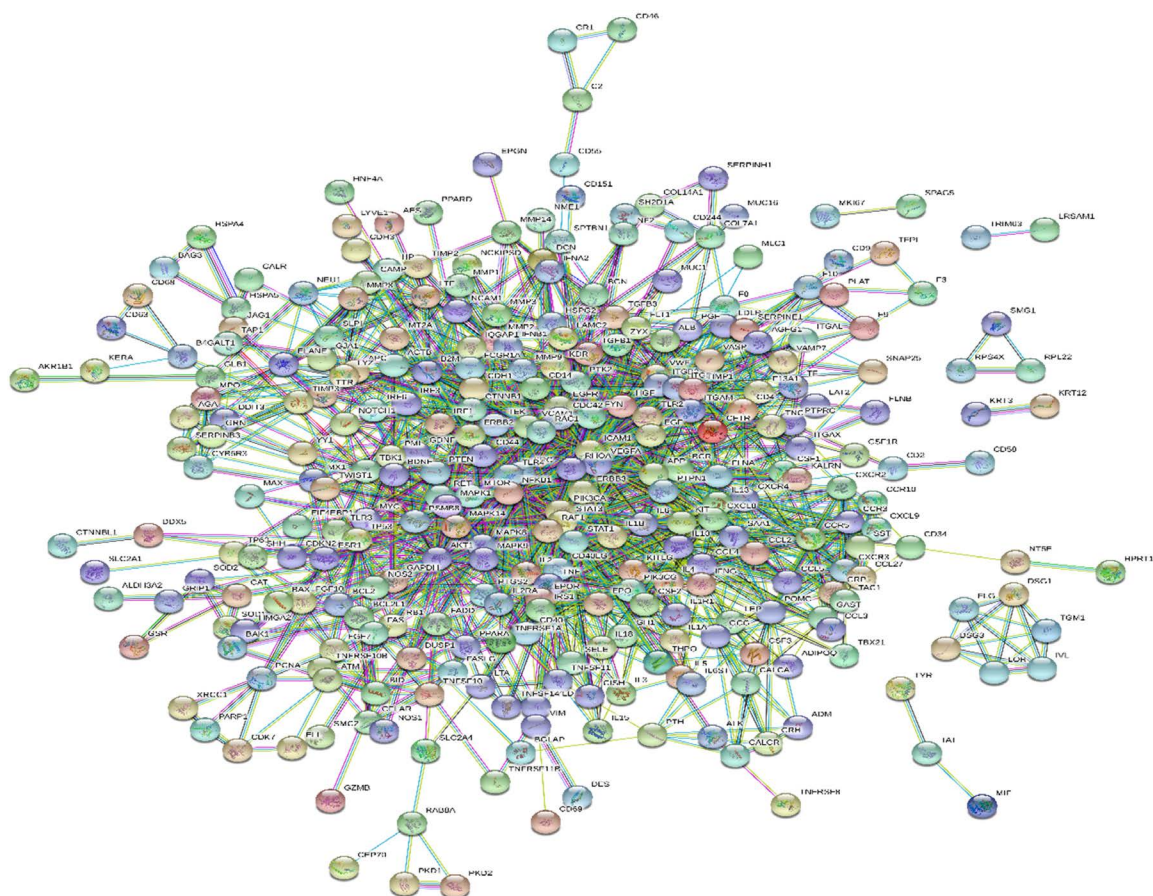


Figure 3. STRING protein interaction network diagram
图 3. STRING 蛋白质相互作用网络图

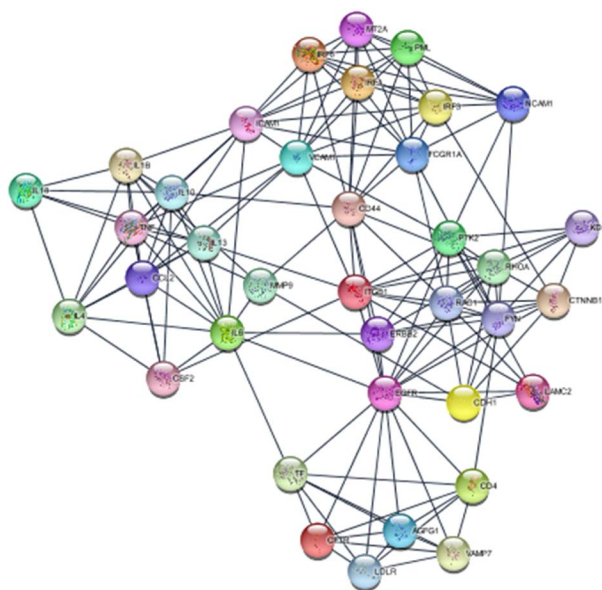


Figure 4. The most important modules in the protein interaction network
图 4. 蛋白质相互作用网络中最重要模块图

3.4. 基因 - 药物相互作用筛选

在 GEPIA 中, 筛选出 CSF2、MT2A、LAMC2 共 3 个基因, 箱式图、生存曲线均有统计学差异($P < 0.05$)。即 CSF2 (图 5)、MT2A、LAMC2 在头颈部肿瘤放疗化疗后 OM 中表达水平明显高于正常组织, 生存率明显低于正常组织。

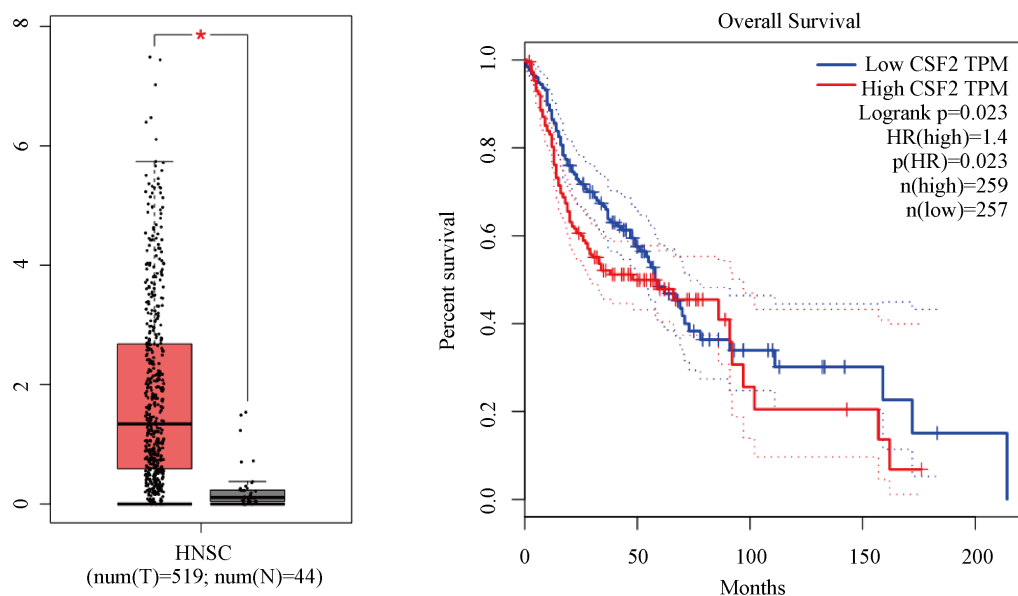


Figure 5. CSF2 box diagram (left), survival curve (right)
图 5. CSF2 箱式图(左)、生存曲线图(右)

利用 DGIdb 工具在线检索了针对 CSF2、MT2A、LAMC2 三个基因已经研发的或者正在研究中的药物 48 种。这 48 种包括经典的抗肿瘤药物、激素、抗病毒药物、免疫抑制剂等(表 1)。

Table 1. Summary table of potential drugs for OM treatment against target genes

表 1. 针对靶基因的治疗 OM 的潜在药物汇总表

编号	靶基因	药物	来源	pmids
1	CSF2	PERFOSFAMIDE	NCI	1283360
2	CSF2	ACLARUBICIN	NCI	7689127
3	CSF2	AMIKACIN	NCI	7540696 8158029
4	CSF2	CYTARABINE	NCI	8819077 8450676
5	CSF2	DIAZQUONE	NCI	8617583
6	CSF2	BCG VACCINE	NCI	12126899
7	CSF2	BECLOMETHASONE DIPROPIONATE	NCI	9230235
8	CSF2	CETIRIZINE	NCI	16783450
9	CSF2	CILOSTAZOL	NCI	15744080
10	CSF2	STAVUDINE	NCI	2538549
11	CSF2	DITIOCARB	NCI	1381236

Continued

12	CSF2	EPIGALOCATECHIN GALLATE	NCI	9000119 15637567
13	CSF2	ETANERCEPT	NCI	12906870
14	CSF2	ALCOHOL	NCI	16272341
15	CSF2	FAMCICLOVIR	NCI	10522033
16	CSF2	FLUTICASONE PROPIONATE	NCI	10433509
17	CSF2	LACTULOSE HYDRATE	NCI	11355146
18	CSF2	MECHLORETHAMINE	NCI	10640980
19	CSF2	METHIMAZOLE	NCI	8836317 7991239
20	CSF2	MYCOPHENOLATE MOFETIL	NCI	16029632 12974756
21	CSF2	MYCOPHENOLIC ACID	NCI	9822358
22	CSF2	VINORELBIN DITARTRATE	NCI	15934496
23	CSF2	NORDIHYDROGUAIARETIC ACID	NCI	2453577
24	CSF2	OMEPRAZOLE	NCI	11100337
25	CSF2	PROCARBAZINE	NCI	10640980
26	CSF2	ROQUINIMEX	NCI	9811314
27	CSF2	SODIUM SALICYLATE	NCI	10574634
28	CSF2	STAUROSPORINE	NCI	9359403
29	CSF2	STREPTOZOTOCIN	NCI	16342200
30	CSF2	TEMOZOLOMIDE	NCI	12610499 16100942
31	CSF2	TERFENADINE	NCI	8836342
32	CSF2	TETANUS TOXOID	NCI	8704175
33	CSF2	TICARCILLIN	NCI	7875147
34	CSF2	TRIFLUOPERAZINE	NCI	1690578
35	CSF2	VINBLASTINE	NCI	10640980
36	CSF2	KB-003		ChemblInteractions
37	CSF2	MOR-103		TdgClinicalTrial ChemblInteractions
38	CSF2	REGRAMOSTIM		TdgClinicalTrial
39	CSF2	MILTEFOSINE	NCI	1555249
40	CSF2	ISOPROTERENOL	NCI	9730867
41	CSF2	NAMILUMAB		ChemblInteractions
42	CSF2	IDARUBICIN	NCI	8915668
43	CSF2	INDINAVIR	NCI	10479132
44	CSF2	INTERFERON ALFA-2B	NCI	10522033
45	CSF2	MT-203		TdgClinicalTrial
46	MT2A	ALCOHOL	NCI	1733046
47	MT2A	ACETAMINOPHEN	NCI	9570328
48	LAMC2	OCRIPLASMIN		ChemblInteractions

4. 讨论

头颈癌是全球第六大常见的恶性肿瘤，每年约有 600,000 例新增病例[3]。大多数为头颈部鳞状细胞癌(head and neck squamous cell carcinoma, HNSC)，晚期 HNSC 的治疗常常采用化疗、放疗等结合的方法治疗。放化疗后，有众多的不良反应，以 OM 最为常见。对于口腔黏膜炎的发病机制、药物等研究，已经足够深入，本研究则是在这些研究成果的基础上，进行文本数据挖掘，利用生物信息学软件、网站等工具，对有可能治疗 OM 的药物进行预测。尽管预测结果并不一定会成为临床用药依据，但，可作为一个探索方向，为以后的研究工作提供新思路。

本研究是基于高通量癌症基因组学数据，进行文本挖掘产生的研究假设，揭示 OM 与伤口愈合可能的关系[4]。利用 pubmed2ensembl 对已经证实了的，与 OM、伤口愈合相关的基因进行筛选，用 Venny 取两基因集的交集，共 372 个基因。

我们对第一步高通量文本挖掘得到的 372 个基因集进行基因富集分析，发现与对有机物质的应答、上调刺激产生的应答等生物学过程有关；与激活细胞因子、生长因子，与细胞因子受体以及生长因子受体结合等分子功能相关；与胞外区域、胞膜表面等有关。KEGG 通路富集分析结果提示可能与细胞因子 - 细胞因子受体信号通路高度相关，同时与癌症的蛋白多糖、癌症的分子途径等有关。从而，我们认为，HNSC 放化疗后的口腔黏膜炎发病机制复杂，涉及众多的基因、蛋白、信号通路。

我们希望通过 STRING、Cytoscape 分析蛋白质相互作用网络分析，进一步筛选出与 OM 作用最为紧密的基因，进一步揭露发病机制。经过 GEPIA 的再次筛选，最后，靶基因确定为 CSF2、MT2A、LAMC2。CSF2 为巨噬细胞集落刺激因子的一员，参与了多种癌症的发生发展，如黑色素瘤，结直肠癌等[5]，在脑转移性甲状腺癌中 CSF2 是最重要的上调基因之一[6] [7] [8]。CSF2 还可诱导小胶质细胞瘤形成，刺激神经胶质细胞瘤的生长与侵袭[5]。MT2A 是金属硫蛋白的一类，与星形胶质细胞氧化应激有关，过表达的 MT2A 可促进多种肿瘤的发生[8] [9]，可能与金属硫蛋白(MT)的促进细胞增殖、抗凋亡作用有关[10]。在 Martina Raudenska 等人的研究中，证实 MT2A 与 HNSC 分级呈正相关，HNSC 中 MT2A 的量远远高于瘤旁组织，DutschWicherek 等人的研究表明当存在淋巴结转移时，瘤旁组织的 MT2A 浓度要高于瘤内组织[11]。LAMC2 即层粘连蛋白 5 γ -2，有研究表明：LAMC2 在间变性甲状腺癌中高表达，与间变性甲状腺癌的增殖、细胞周期、迁移能力有关，通过激活细胞外调节蛋白激酶 1 和 2 增强癌细胞的侵袭能力[12]。LAMC2 的存在可能增加儿童神经母细胞瘤患病率[13]，LAMC2 可能参与口腔鳞状细胞癌的表面遗传学变化，抑制 LAMC2 的表达，可降低口腔鳞状细胞癌中肿瘤干细胞的侵袭、迁移能力。

基因 - 药物分析，针对靶基因的药物，为抑制剂和未知作用的药物，这些药物，大部分已经应用于临床多年，极少数还在试验阶段。大体上，安全性是有保障的，只是，部分药物未应用于治疗头颈部肿瘤放化疗后的 OM。如果条件允许，可以先进行这部分的药物研发，尝试局部治疗，这个过程，可能需要探索多个基因、蛋白、信号通路[14] [15]。

综上所述，本研究应用文本挖掘工具，通过对之前的研究进行探索，利用生物信息学工具，进行基因富集分析以及信号通路分析，建立蛋白质相互作用网络，并最终筛选出最密切的基因，找到潜在的可能有效治疗头颈部肿瘤放化疗后的口腔黏膜炎药物。最终，这些药物是否有效，则需要基础实验以及临床试验验证。

参考文献

- [1] 郭静, 李晓珊, 张甫婷. 如何应对头颈肿瘤 5 大化疗常见不良反应[N]. 广东科技报, A04.
- [2] 张新良, 王晓萍. 急性放射性口腔黏膜炎的诊断与治疗[J]. 临床肿瘤学杂志, 2006, 1(4): 312-313.
- [3] Solomon, B., Young, R.J. and Rischin, D. (2018) Head and Neck Squamous Cell Carcinoma: Genomics and Emerging

Biomarkers for Immunomodulatory Cancer Treatments. *Seminars in Cancer Biology*, **52**, 228-240.

<https://doi.org/10.1016/j.semcancer.2018.01.008>

- [4] 刘燕, 等. 基于文本挖掘的高通量癌症基因组数据注释[J]. 中华医学图书情报杂志, 2016, 25(12): 34-39.
- [5] Sielska, M., *et al.* (2020) Tumour-Derived CSF2/Granulocyte Macrophage Colony Stimulating Factor Controls Myeloid Cell Accumulation and Progression of Gliomas. *British Journal of Cancer*.
<https://doi.org/10.1038/s41416-020-0862-2>
- [6] Chitu, V., *et al.* (2020) Microglial Homeostasis Requires Balanced CSF-1/CSF-2 Receptor Signaling. *Cell Reports*, **30**, 3004-3019.e5. <https://doi.org/10.1016/j.celrep.2020.02.028>
- [7] Schulten, H.J., *et al.* (2016) Microarray Expression Profiling Identifies Genes, Including Cytokines, and Biofunctions, as Diapedesis, Associated with a Brain Metastasis from a Papillary Thyroid Carcinoma. *American Journal of Cancer Research*, **6**, 2140-2161.
- [8] Xu, Z., *et al.* (2019) Demethylation and Overexpression of CSF2 Are Involved in Immune Response, Chemotherapy Resistance, and Poor Prognosis in Colorectal Cancer. *OncoTargets and Therapy*, **12**, 11255-11269.
<https://doi.org/10.2147/OTT.S216829>
- [9] Cui, Y., *et al.* (2003) ECRG2, a Novel Candidate of Tumor Suppressor Gene in the Esophageal Carcinoma, Interacts Directly with Metallothionein 2A and Links to Apoptosis. *Biochemical and Biophysical Research Communications*, **302**, 904-915. [https://doi.org/10.1016/S0006-291X\(03\)00122-0](https://doi.org/10.1016/S0006-291X(03)00122-0)
- [10] Dutsch-Wicherek, M., *et al.* (2013) Analysis of Metallothionein and Vimentin Immunoreactivity in Pharyngeal Squamous Cell Carcinoma and Its Microenvironment. *Cell and Tissue Research*, **352**, 341-349.
<https://doi.org/10.1007/s00441-013-1566-1>
- [11] Raudenska, M., *et al.* (2015) Prognostic Significance of the Tumour-Adjacent Tissue in Head and Neck Cancers. *Tumor Biology*, **36**, 9929-9939. <https://doi.org/10.1007/s13277-015-3755-x>
- [12] Garg, M., *et al.* (2014) Laminin-5 γ -2 (LAMC2) Is Highly Expressed in Anaplastic Thyroid Carcinoma and Is Associated with Tumor Progression, Migration, and Invasion by Modulating Signaling of EGFR. *The Journal of Clinical Endocrinology & Metabolism*, **99**, E62-E72. <https://doi.org/10.1210/jc.2013-2994>
- [13] Yang, T., *et al.* (2018) The rs2147578 C > G Polymorphism in the Inc-LAMC2-1:1 Gene Is Associated with Increased Neuroblastoma Risk in the Henan Children. *BMC Cancer*, **18**, 948. <https://doi.org/10.1186/s12885-018-4847-y>
- [14] Zhou, Y.M., *et al.* (2020) MicroRNA-134 Inhibits Tumor Stem Cell Migration and Invasion in Oral Squamous Cell Carcinomas via Downregulation of PI3K-Akt Signaling Pathway by Inhibiting LAMC2 Expression. *Cancer Biomarkers*. <https://doi.org/10.3233/CBM-191362>
- [15] Zhao, C., *et al.* (2018) An Integrated Methylation and Gene Expression Microarray Analysis Reveals Significant Prognostic Biomarkers in Oral Squamous Cell Carcinoma. *Oncology Reports*, **40**, 2637-2647.
<https://doi.org/10.3892/or.2018.6702>