·信息研究·

特发性肺纤维化相关基因的筛选和生物信息学分析

邢 静，黄鑫炎，郭禹标
（中山大学附属第一医院呼吸内科，广东 广州 510080）

摘 要：【目的】通过生物信息学的方法发现特发性肺纤维化（IPF）的致病基因并为进一步研究提供靶点。【方法】从GEO数据库中下载基因芯片数据集GSE53845、GSE24206、GSE10667，并使用GEO2R分析工具筛选出正常组织与IPF的差异表达基因。在DAVID数据库中对差异表达基因进行GO分析和KEGG通路富集分析，以便找到IPF发病过程中差异表达基因主要参与的生物功能及其集中的信号通路。为了研究差异表达基因与蛋白之间的作用关系，使用STRING和CYTOSCAPE软件来构建蛋白相互作用网络，使用MCODE软件来提取蛋白相互作用网络中的子网络模块。【结果】发现了110个差异表达基因，其中有92个在IPF中高表达，18个低表达。GO富集分析表明IPF中上调的差异表达基因主要影响细胞粘附、生物粘附、胶原蛋白代谢等相关的生物过程，富集的分子功能主要参与细胞外基质结构的构成、钙离子的结合；IPF中下调的蛋白则主要涉及感觉调节的生物过程。KEGG通路分析表明IPF中上调的差异表达基因主要参与受体相互作用、细胞粘附等信号通路。【结论】利用生物信息学筛选出差异表达基因，其中部分基因已被证实参与IPF，部分基因尚未有研究，提示其可能是IPF发病机制研究新的研究靶点。

关键词：特发性肺纤维化；差异表达基因；生物信息学
中图分类号：R563.9 **文献标志码：**A **文章编号：**1672-3554(2017)06-0926-05

Screening and Bioinformatics Analysis of Idiopathic Pulmonary Fibrosis Related Genes

XING Jing, HUANG Xin-yan, GUO Yu-biao
(Department of Respiratory, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China)
Corresponding to: GUO Yu-biao; E-mail:guoyubiao@hotmail.com

Abstract:【Objective】We aimed to explore the pathogenic genes of Idiopathic pulmonary fibrosis (IPF) by bioinformatics analysis and provide a target for further research.【Methods】Gene data sets GSE53845, GSE24206, GSE10667 were downloaded from the Gene Expression Omnibus database and the differential expression genes of normal tissue and IPF were screened with GEO2R analysis tool. GO analysis and KEGG pathway enrichment analysis of differentially expressed genes were performed in DAVID database in order to find out the biological function and its focused signal pathway in differentially expressed genes during IPF development. In order to study the relationship between differential genes and proteins, STRING and CYTOSCAPE software were used to construct the protein interaction network and MCODE software was used to extract the sub-network modules in the protein-interacting network.【Results】This study found 110 differentially expressed genes, of which 92 were high expression in IPF and 18 were low expression. GO enrichment analysis showed that the up-regulated genes in IPF mainly affected the biological processes such as cell adhesion, bio-adhesion and collagen metabolism. The enriched molecular function was mainly involved in the composition of extracellular matrix structure and the binding of calcium ions. The down-regulated proteins are mainly involved in the sensory regulation of the biological process in IPF. KEGG pathway analysis showed that the up-regulated genes in IPF were mainly involved in receptor interactions, cell adhesion and other signaling pathways.【Conclusions】This study uses bioinformatics to screen out the differential genes, some of which have been shown to be involved in IPF, and some genes have not been studied, suggesting that it may be a new research target for IPF pathogenesis.

Key word: idiopathic pulmonary fibrosis; differentially expressed gene; bioinformatics analysis
[J SUN Yat-sen Univ(Med Sci), 2017, 38(6): 926-930; 937]

收稿日期：2017-06-19

作者简介：邢静，硕士研究生，研究方向：呼吸病学，E-mail：474053670@qq.com；郭禹标，通信作者，教授，E-mail：guoyubiao@hotmail.com

特发性肺纤维化（idiopathic pulmonary fibrosis，IPF）是以成纤维细胞增殖和细胞外基质沉积为特征的肺部疾病[1]。其临床特征表现为不明原因的持续性进行性呼吸困难，并常伴有咳嗽，双肺基底部吸气末 Velcro 啰音，IPF 病程进展较慢，会引起肺部弥漫性纤维化，最终导致呼吸功能衰竭[2]。IPF 的病因至今尚不明确。危险因素包括吸烟、环境暴露等，存在一定的遗传易感性。以往的研究认为，持续的炎症反应导致肺脏损伤和纤维化形成是 IPF 发病的主要机制。现在的研究更多地认为是肺泡上皮细胞损伤后的异常修复是 IPF 发病的主要机制[3]。IPF 的中位生存时间为 3~5 年，目前尚无有效的治疗手段[4]。近年来发现 IPF 多呈现家族聚集，提示其可能是一种多基因共同作用的疾病。但至今对 IPF 的致病基因尚未有太多的研究。本研究希望通过生物信息学的方法发现 IPF 的致病基因并揭示进一步发病机制研究的方向。

1　材料与方法

1.1　数据来源

从 GEO 数据库（GEO，http：// www. ncbi. nlm. nih. gov/geo/）[5]。下载基因芯片数据集 GSE53845[6]、GSE24206[7]、GSE10667[8]。GSE53845 包含 40 个 IPF 标本和 8 个正常组织标本，GSE24206 包含 17 个 IPF 标本和 6 个正常组织标本，GSE10667 包含 31 个 IPF 标本和 15 个正常组织标本。

1.2　确定差异表达基因

使用 GEO2R（https：//www.ncbi.nlm.nih.gov/geo/geo2r/）筛选差异表达基因（differentially expressed gene，DEG），DEG 的筛选条件为 adj.P.Val < 0.05 和 |logFC| > 1。

1.3　差异表达基因的 GO 富集分析和 KEGG 通路分析

GO 分析依据基因产物的相关分子功能、生物过程和细胞组成进行简单注释。使用 DAVID 在线软件（https：//david-d.ncifcrf.gov/）[9]。进行 GO 分析和 KEGG 分析。$P < 0.05$ 被认为是具有统计学差异。

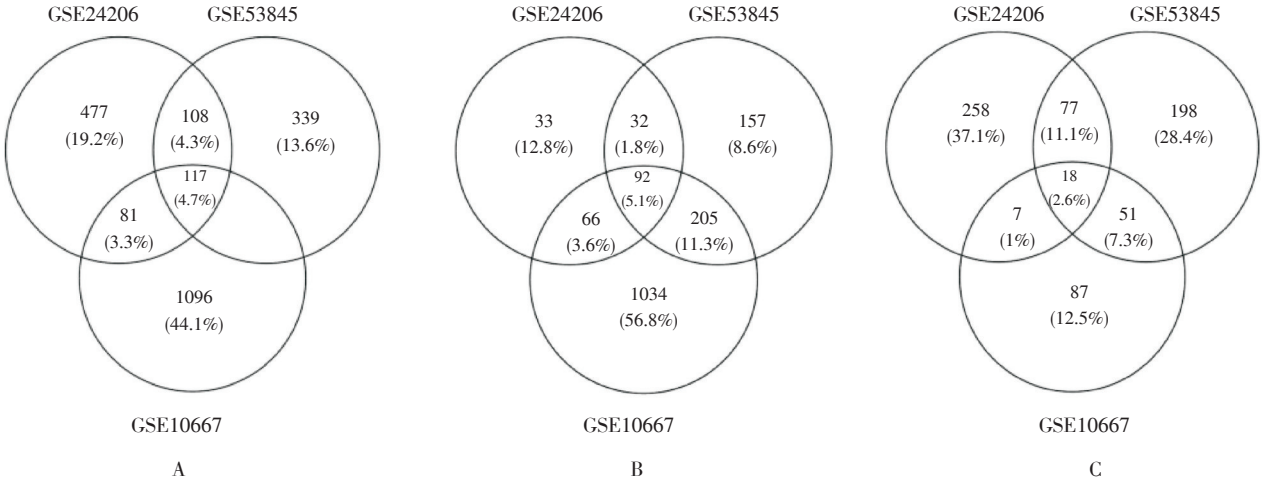
1.4　差异表达基因的蛋白互作网络图及互作网络图构建基因的分析

通过 STRING（http：// www. string-db.org/）[10] 可以预测蛋白之间的相互关系，选择 confidence score 为 0.4。CYTOSCAPE 中 MCODE 用于对这个网络中关系密切的模块进行分析，参数为：Degree Cutoff：2，Node Score Cutoff：0.2，K-Core：2，Max. Depth：100。

2　结　果

2.1　确定差异表达基因

在基因芯片数据集 GSE53845、GSE24206、GSE10667 分别筛选出 831、783、1560 个差异表达



A: gene; B: high expression; C: low-middle expression

图 1　GSE53845、GSE24206、GSE10667 中差异表达基因的表达

Fig.1　Identifications of differentially expressed gene in mRNA expression profile of GSE53845, GSE24206, and GSE10667

基因。在三个基因芯片数据集中117个基因被筛选出来,其中110个有相同的表达趋势。在110个差异表达基因中有92个在IPF中高表达,18个低表达(图1)。

2.2 差异表达基因的GO富集分析和KEGG通路分析结果

GO富集分析表明IPF中上调的差异表达基因参与了细胞粘附、生物粘附、胶原蛋白代谢等相关的生物过程,分子功能主要参与细胞外基质结构的构成、钙离子的结合;IPF中下调的蛋白则主要涉及感觉调节的生物过程(表1)。KEGG通路分析表明IPF中上调的差异表达基因主要参与受体相互作用、细胞粘附等信号通路(表2)。

表2 KEGG通路分析

Table 2 KEGG Pathway of DEG

Up-regulated KEGG	Count	%	*P*
04512:ECM-receptor interaction	6	69.36	<0.001
04510:Focal adhesion	7	80.92	0.001
04360:Axon guidance	4	46.24	0.046

2.3 差异表达基因的蛋白互作网络图

生物体内,蛋白质的功能行使不是独立的,还需借助其他蛋白质的调节和介导等相互作用而实现。蛋白质通过形成互作网络而发挥功能[11]。差异表达基因的蛋白互作网络是由51个节点和84

表1 GO富集分析结果

Table 1 GO terms enriched by DEGs

Term	Count	%	*P*
Up-regulated			
BP(Biological Process)			
GO:0007155~cell adhesion	17	1.97	< 0.001
GO:0022610~biological adhesion	17	1.97	< 0.001
GO:0032963~collagen metabolic process	6	0.69	< 0.001
GO:0044259~multicellular organismal macromolecule metabolic process	6	0.69	< 0.001
GO:0044236~multicellular organismal metabolic process	6	0.69	< 0.001
MF (Molecular Function)			
GO:0005201~extracellular matrix structural constituent	8	0.92	< 0.001
GO:0005509~calcium ion binding	14	1.62	< 0.001
GO:0048407~platelet-derived growth factor binding	3	0.35	0.001
GO:0008201~heparin binding	5	0.58	0.001
GO:0019838~growth factor binding	5	0.58	0.002
CC (Cell Component)			
GO:0044421~extracellular region part	30	3.47	< 0.001
GO:0031012~extracellular matrix	20	2.31	< 0.001
GO:0005576~extracellular region	40	4.62	< 0.001
GO:0005578~proteinaceous extracellular matrix	19	2.20	< 0.001
GO:0044420~extracellular matrix part	10	1.16	< 0.001
Down-regulated			
BP			
GO:0051930~regulation of sensory perception of pain	2	11.11	0.008
GO:0051931~regulation of sensory perception	2	11.11	0.008
GO:0051345~positive regulation of hydrolase activity	3	16.67	0.009
GO:0016477~cell migration	3	16.67	0.020
GO:0051674~localization of cell	3	16.67	0.025

条边组成。每个节点就代表一个蛋白质，每条边则表示蛋白质之间的相互作用[12]。其中度值比较高的基因是*COL1A1*、*COL1A2*、*COL3A1*、*COL14A1*、*MMP7*、*POSTN*、*COL8A2*、*THBS2*。用MCODE插件分析后只有一个子模块满足参数要求，其由7个节点和21条边组成（图2）。

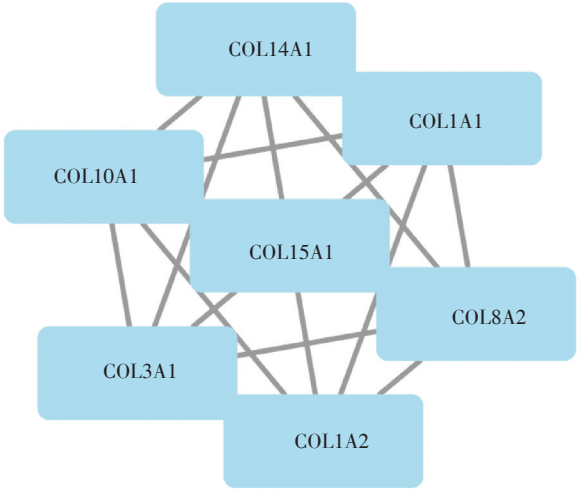


图2　差异表达蛋白的子模块相互作用图

Fig.2　Significant modules from PPI network

3　讨　论

生物信息学是一门新兴的交叉学科，其利用数学、计算机科学和生物学等工具对生物学数据进行分析和研究，目的在于了解数据中所蕴含的生物学意义。近年来，生物信息学被广泛应用于发病机制的研究，它能够找出疾病发展过程中起主要作用的基因，从而发现疾病新的致病靶点，对于了解疾病的发展机制和开发新的治疗方法具有重大意义。IPF是一种严重威胁生命的慢性疾病，其特征为异常的肺重塑，细胞外基质沉积，其发病机制目前仍不十分清楚。近年来虽然有多种药物进行了基础研究及临床试验，但其治疗效果并不令人满意。因此深入该病的发病机制才能进一步发现新的治疗方法从而改善病人的预后。新近观点认为，环境和基因等因素诱导上皮细胞损伤，激活了炎症反应、细胞因子途径、损伤修复等多重途径，导致了肺内促纤维化及抗纤维化失衡；这些活化的介质反过来激活多种细胞组分，又进一步引起细胞功能改变和细胞间相互作用，最终导致肺纤维化的进展。本研究中，我们在三个基因芯片数据集，共包含88个IPF样本和29个正常对照样本中，发现与正常对照组相比，差异表达基因中有92个在IPF中高表达，18个低表达。GO分析说明IPF中上调的差异表达基因主要参与了细胞粘附、生物粘附、胶原蛋白代谢等相关的生物过程，分子功能主要与细胞外基质结构的构成、钙离子的结合。IPF中下调的蛋白主要参与感觉调节的生物过程。KEGG通路分析表明IPF中上调的差异表达基因主要参与受体相互作用、细胞粘附等信号通路。通过构建的蛋白相互作用网络图，我们确定了8个度值比较高表达基因:*COL1A1*、*COL1A2*、*COL3A1*、*COL14A1*、*MMP7*、*POSTN*、*COL8A2*、*THBS2*。

研究证实IPF的疾病进展过程中伴随着严重胶原代谢紊乱[13]。本研究所确定的基因*COL1A1*、*COL1A2*、*COL3A1*、*COL14A1*是胶原蛋白家族的成员。在肺纤维化发病过程中，大量的细胞外基质沉积在肺中，在IPF肺部组织中高度表达。Ⅰ型胶原蛋白是主要的肺胶原蛋白，包含两个α1链（COL1A1）和一个α2链（COL1A2）。IPF病理表现为早期肺泡壁增厚和上皮细胞存在死亡的肺泡炎，这些改变最终导致肺纤维化形成，其特征在于Ⅰ型胶原蛋白的过表达[14]。Ⅰ型和Ⅲ型胶原蛋白在纤维化肺的发展中起主要作用。在IPF中，Ⅰ型和Ⅲ型胶原蛋白产量都有所增加[15]。Ⅷ型胶原细胞及细胞基质间的相互作用间起重要作用，参与多种细胞活动，比如增殖、粘附、迁移及金属蛋白酶的生成，而且在血管生成和血管重构中也发挥重要作用。我们尚未发现COL8A参与肺纤维化中的相关研究，但在动脉粥样硬化的纤维帽形成中发挥作用[16]。COL14A是一种不规则三螺旋结构的纤维胶原，在调节纤维直径的宽度中发挥重要作用。COL14及产物内皮抑素参与发性肺动脉高压IPAH的发展，并且作为疾病进展的标志物[17]。

基质金属蛋白酶（matrix metalloproteinases，MMP）含有一组能够降低细胞外基质的具有蛋白酶活性的含锌蛋白酶类，参与细胞外基质的降解和重建[18]。IPF与正常对照样品相比，IPF中MMP7血清水平升高[19]，且IPF患者的死亡率与血浆中MMP7有相关性[20]。骨膜蛋白（POSTN）是成骨或成纤维细胞的一种间质特异性分泌蛋白，与

纤维化程度密切相关。特发性肺纤维化细胞合成的POSTN的mRNA含量比对照组高,且骨膜素蛋白表达增加,并主要集中在成纤维活动活跃的部位[21]。POSTN还能促进Ⅰ型胶原的形成,它的水平升高,将会导致IPF患者的肺功能恶化[22,23]。血小板反应蛋白2(THBS2)属于血小板反应蛋白家族,是成纤维细胞等细胞分泌的细胞外基质糖蛋白,参与细胞间基质蛋白结合等功能。有研究在增殖性糖尿病视网膜病变中观察到升高的THBS2[24]。但据我们所知,该基因从未在特发性肺纤维化研究中提及。

综上所述,本研究表明 *COL1A1*、*COL1A2*、*COL3A1*、*COL14A1*、*MMP7*、*POSTN*、*COL8A2*、*THBS2* 等基因参与的分子功能、生物过程、细胞组成及信号通路可能与IPF的发病密切相关。其中 *COL8A* 和 *THBS2* 两个基因尚未在IPF中有研究。这些研究结果为特发性肺纤维化发病机制研究提供了新的思路。

本研究利用生物信息学筛选出差异表达基因,其中部分基因已被证实参与IPF,部分基因尚未有研究,提示其可能是IPF发病机制研究新的研究靶点,可能需要体内外实验及临床资料验证。

参考文献

[1] Selman M, King TE, Pardo A. Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy[J]. Ann Intern Med, 2001, 134(2): 136-151.

[2] 中华医学会呼吸病学分会间质性肺疾病学组. 特发性肺纤维化诊断和治疗中国专家共识[J]. 中华结核和呼吸杂志, 2016, 39(6): 427-432.
Group Interstitial Lung Diseases of Chinese Thoracic Society. Chinese expert consensus statement on clinical diagnosis and treatment of idiopathic pulmonary fibrosisInterstitial lung disease study group of chinese thoracic society[J]. Chin J Tubercul RespirDis, 2016, 39(6): 427-432.

[3] 孙锦涛,徐兴祥. 特发性肺纤维化发病机制及药物治疗的研究进展[J]. 中华肺部疾病杂志(电子版), 2015, 8(2): 82-87.
Sun JT, Xu XX. Research progress for pathogenesis of idiopathic pulmonary fibrosisand its drug treatment[J]. Chin J Lung Dis(E Ed)2015, 8(2): 82-87.

[4] Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management[J]. Am J Respir Crit Care Med, 2011, 183(6): 788-824.

[5] 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-D995.

[6] Depianto DJ, Chandriani S, Abbas AR, et al. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis[J]. Thorax, 2015, 70(1): 48-56.

[7] Meltzer EB, Barry WT, D'Amico TA, et al. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: proof-of-principle[J]. BMC Med Genomics, 2011, 4(1): 70-77.

[8] Konishi K, Gibson KF, Lindell KO, et al. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis[J]. Am J Respir Crit Care Med, 2009, 180(2): 167-175.

[9] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nat Protoc, 2009, 4(1): 44-57.

[10] 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.

[11] Snider J, Kotlyar M, Saraon P, et al. Fundamentals of protein interaction network mapping[J]. Mol Syst Biol, 2015, 11(12): 848-856.

[12] 李敏,孟祥茂. 动态蛋白质网络的构建、分析及应用研究进展[J]. 计算机研究与发展, 2017, 54(6): 1281-1299.
Li M, Meng XM. The construction, analysis, and applications of dynamic protein-protein interaction networks[J]. J Comp Res Devel, 2017, 54(6): 1281-1299.

[13] Van Hoozen BE, Grimmer KL, Marelich GP, et al. Early phase collagen synthesis in lungs of rats exposed to bleomycin[J]. Toxicology, 2000, 147(1): 1-13.

[14] Vuga LJ, Milosevic J, Pandit K, et al. Cartilage oligomeric matrix protein in idiopathic pulmonary fibrosis[J]. PLoS One, 2013, 8(12): e83120.

[15] Lepparanta O, Tikkanen JM, Bespalov MM, et al.

(下转第937页 to page 937)