早期和晚期肺腺癌的差异基因和信号通路富集分析

王桂平[1],韦　敏[1],廖洪映[2],周　琼[1],王　妍[1],林竹贞[1],张彦焘[1]
(1.广州医学院卫生职业技术学院,广东 广州 510180;2.中山大学附属第三医院,广东 广州 510630)

摘　要:[目的]分析早期和晚期肺腺癌的差异基因和信号通路。[方法]从美国国立生物信息中心(NCBI)的GEO数据库下载GSE10072数据集,去除临床指标缺失的样本,按照TNM分期将肺腺癌样本分为早期(Ⅰ期,共16例)和晚期(Ⅲ~Ⅳ期,共15例)两组。原始数据经dChip进行质量检验、标准化,然后进行差异基因分析。从MsigDB数据库获得344个生物信号基因集,通过GSEA进行信号通路富集分析。[结果]获得SEMA3、PLAU、CDKN2A等14个明显差异基因,获取的差异基因主要与细胞凋亡、细胞黏附等过程密切相关。选取MsigDB中来源于Bicarta、KEGG、GenMAPP三大数据库的344个基因集进行富集分析,结果发现Death pathway、Leukocyte transendothelial migration和Focaladhesion等22条信号通路在晚期肺腺癌中明显富集,富集通路主要涉及细胞凋亡、细胞黏附和迁移等过程。[结论]早期与晚期肺癌中存在一些明显的差异表达基因,有部分信号通路在晚期肺腺癌中明显富集。
关键词:肺腺癌;表达谱;基因通路;基因富集
中图分类号:R734.2　　**文献标识码:**A　　**文章编号:**1004-0242(2013)01-0054-05

An Analysis of Differential Expression Genes and Gene Sets Associated with Lung Adenocarcinoma in Early and Advanced Lung Adenocarcinoma

WANG Gui-ping[1], WEI Min[1], LIAO Hong-ying[2], et al.
(1.Health College, Guangzhou Medical University, Guangzhou 510180, China;
2.The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou 510630, China)

Abstract: [Purpose] To analyze the differential expression genes and gene sets associated with lung adenocarcinoma in early and advanced lung adenocarcinoma. [Methods] Lung adenocarcinoma gene expression profile data GSE10072 were obtained from Gene Expression Omnibus (GEO) database of National Center for Biotechnology Information. Samples without clinical data were excluded. The samples were divided into early stage group (stage Ⅰ, 16 samples) and advanced stage group (stage Ⅲ and Ⅳ, 15 samples). Raw data were normalized, quality control and analyses of differentially expressed genes by dChip software. Three hundred and forty-four gene sets of cell signal pathways from MsigDB bank.The enrichment of gene sets was analyzed by GSEA software. [Results] Fourteen differentially expressed genes including SEMA3, PLAU, CDKN2A were obtained by dChip analysis. Data mining showed that differentially expressed genes obtained were related with developmental programmed cell death, cell adhesion. Analysis of gene sets enrichment against 344 pathways from Bicarta, KEGG, GenMAPP, showed that three pathways including Death pathway, Leukocyte transendothelial migration and Focal adhesion were enriched in advanced lung adenocarcinoma. [Conclusion] There are some differential expression genes between early and advanced lung adenocarcinoma, and some pathways are enriching in advanced lung adenocarcinoma.
Key words: lung adenocarcinoma; gene expression profile; signal pathway; GSEA

肺癌已成为严重威胁人类健康的一种重要疾病。根据我国居民死因调查结果,肺癌死亡率从20世纪70年代中期至90年代初期的20年增加近1.5倍,是增长最快的恶性肿瘤[1,2]。肺腺癌(adenocarcinoma of lung)是具有腺样分化或黏液分泌的恶性上皮肿瘤,发病率约占原发性肺癌的20%~30%。早期肺腺癌手术切除后5年生存率可达30%以上,病变越早,预后越好。然而,80%以上非小细胞肺癌(non-small cell lung cancer, NSCLC)发现时已为晚期,因

收稿日期:2012-05-02;修回日期:2012-09-04
基金项目:广东省自然科学基金项目(S2011010004147);广州医学院2011年科研基金博士启动基金(2011C06)
通讯作者:王桂平,E-mail:docgpwang@yahoo.com.cn

而无法手术治疗,5 年生存率约 8%~15%。因此,寻找和发现晚期肺腺癌恶性改变或者预后特征基因或信号通路,对肺腺癌的早期诊断、预后判断和分子治疗等均具有重要意义[3]。

目前认为肺癌患者的预后受多方面因素的影响,主要有病理类型、分化程度、p-TNM 分期等。上述因素之间往往存在着错踪复杂的交互关系,应用常规方法很难筛选出预后特征分子[4]。基因芯片凭借其高通量、快速检测基因的优点,在肿瘤发生机制、早期诊断、肿瘤基因分型、指导治疗、评估预后等研究方面有着广泛的应用前景,是当前肿瘤研究领域的热点之一[5-7]。本研究基于基因表达谱分析手段,对肺腺癌表达谱数据进行差异特殊基因和生物通路富集分析,以期进一步从基因表达谱的角度阐明晚期肺腺癌的分子机制。

1 材料与方法

1.1 数据集的获取和数据预处理

首先,我们从美国国立生物中心(NCBI)的 GEO 数据库 (http:www.ncbi.nlm.nih.gov/geo) 中下载 GSE10072 数据集。GSE10072 来源于美国国家卫生研究院(NIH)遗传流行病学部(genetic epidemiology branch),采用 GPL96 芯片平台,疾病组织类型为肺腺癌,包括 58 个腺癌和 49 个正常肺组织样本,共 107 个样本。原始文件下载后解压缩为.cel 格式文件。导入 dChip 软件还原原始芯片扫描图像,以总荧光强度为中位数的芯片 N2 为基准,对所有芯片进行标准化。根据芯片分析报告判断芯片中探针交叉杂交和芯片污染的情况,去除探针交叉杂交和污染大于 5%的芯片样本,共获得 31 个样本,其中Ⅰ期肺腺癌样本 16 个,Ⅲ和Ⅳ期肺腺癌样本 15 个。

1.2 肺腺癌差异表达基因分析[8]

基因差异表达分析采用 dChip 软件分析包进行。dChip 由哈佛大学生物统计系 Li 等联合开发,是综合性芯片分析软件。该软件运行在于 Windows 平台,主要分析 Affymetrix 基因表达谱及 SNP 芯片数据,dChip 可进行差异基因识别、方差分析、主成分分析、时间序列分析、层次聚类、连锁分析及 SNP 的拷贝数分析等。我们对 GSE10072 数据集中质量合格芯片样本分别采用 dChip 进行差异基因分析,具体操作方法按 dChip 操作指南进行(http://www.dchip.org),1.2-fold change(差异倍数)的基因被选择为差异表达基因。

1.3 差异基因功能注释与生物学意义分析

采用 Toppgene(http://toppgene.cchmc.org/)进线分析工具,对获取的差异基因进行生物过程、亚细胞分布和分子功能等分析。通过基因组疾病相关数据库(genetic association DB disease)分析差异基因与疾病的相关性。

1.4 基因集富集分析

参照文献[9,10]方法,对从 GEO 数据库获得的肺腺癌数据集进行基因集富集分析 (gene set enrichment analysis,GSEA)。首先,我们从 GSEA 网站 MsigDB 数据库中获得 344 个与生物信号传导相关的基因集(gene sets) 作为参照基因集。按 default weighted enrichment statistic 方法,每次分析重复 1 000 次,标准化的 $P<0.05$ 的基因集定义为阳性基因集。GSEA 软件及说明见 http://www.broadinstitute.org/gsea/ index.jsp。

2 结 果

2.1 芯片数据的标化和归一化

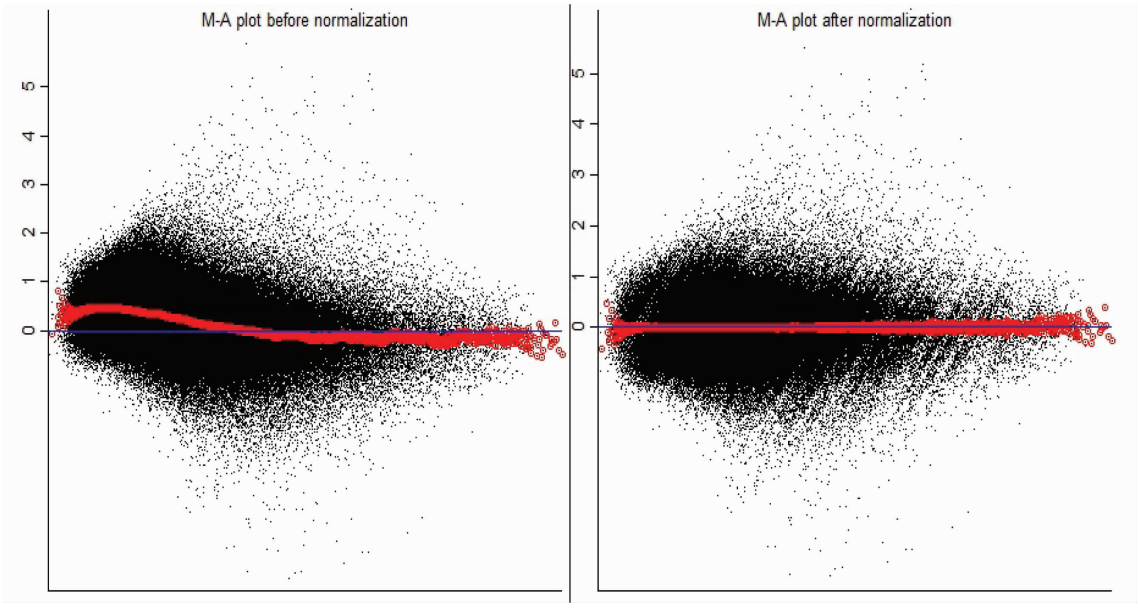
进行数据分析之前,首先对芯片的质量进行检测,质量合格的芯片再经归一化后才进行后续的数据分析,归一化结果(Figure 1)。芯片样本的散点 MA 图(M-A plot after normalization)可以直观观察系统偏移的形式以及是否存在强度一寸的系统偏移。图中的曲线是归一化产生的回归曲线,反映了M 与 A 之间的函数关系。从图中可以看出,经过归一化后,M 值与 A 值已经没有相关性了,这说明基因的表达差异与基因的测量值之间没有相关性,即归一化后系统性的偏差消除了。

2.2 差异表达基因分析

采用 dChip 分析软件包对 GSE10072 数据集中 31 个(Ⅰ期 *vs* Ⅲ和Ⅳ期)合格芯片样本进行差异基因分析,获得 14 个明显差异表达基因(Table 1)。

2.3 差异表达基因生物学意义

为揭示所获取的差异表达基因生物学意义,我们对所获取的差异表达基因进行生物通路过程分析。结果发现,我们获取的差异表达基因主要与细胞



a. M-A plot of microarrays before normalization　　b. M-A plot of microarrays after normalization

Figure 1　M-Aplot of microarrays red/green ratio

Table 1　Differential expression genes of lung adenocarcinoma from gene sets GSE10072

No	Gene name	EntrezGene	Fold change	Difference of means
1	EPHX1	2052	−2.1	−622.23
2	IMPA2	3613	−2.11	−254.98
3	CX3CL1	6376	2.03	192.13
4	SEMA3	10512	1.75	183.29
5	KIT	3815	−2.12	−258.88
6	PLAU	5328	2.22	555.95
7	ASS1	445	−1.81	−220.88
8	HMGA2	8091	2.52	122.12
9	CDKN2A	1029	−2.02	−454.42
10	CYR61	3491	1.67	232.35
11	FER1L3	26509	1.63	165.73
12	C14orf147	171546	−1.54	−107.68
13	SLC38A1	81539	−1.76	−181.31
14	PLAC8	51316	2.07	229.66

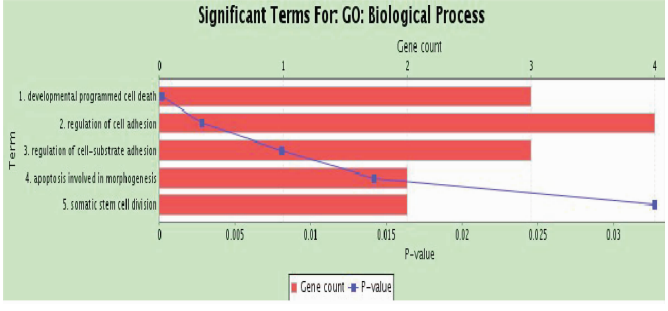


Figure 2　Biological process of lung adenocarcinoma differential genes

凋亡、细胞黏附等过程密切相关(Figure 2)。同时，我们也通过基因组疾病相关数据库进行分析，发现所获得的差异表达基因中有 7 个基因已证实与人类肿瘤密切相关，并且 SEMA3、PLAU、CDKN2A 均已证实与肺癌发生发展相关(Table 2)。

2.4　GSEA 信号通路富集分析

传统的单基因差异分析对生物学过程的研究作用是非常有限的，不能有效揭示基因表达谱数据的生物学意义。为了从生物学意义水平上发现一些与肺腺癌恶性改变密切相关的信号通路，我们采用 GSEA 对所获取的表达谱数据进行基因集富集分析。富集得分(enrichment score，ES)是 GSEA 分析的原始结果，它反映的是将全部杂交数据排序后，在此序列的前部或后部一个功能基因集富集的程度。归一化富集得分 (normalized enrichment score，NES)是标准化结果，正值表示该基因集与第一个表型相关，负值表示与第二个表型相关。一般而言，NES 的绝对值越大，错误发现率的值就越小，说明富集度越高的功能基因集，分析结果的可信度就越高。我们对 MsigDB 中 C2 收集的来源于 Bicarta、KEGG、GenMAPP 三大数据

Table 2　Data mining results of lung adenocarcinoma differential genes

Gene symble	Gene name	Related cancer
EPHX1	epoxide hydrolase 1	Many cancers
IMPA2	inositol(myo)-1(or 4)-monophosphatase 2	Stroke, bipolar disorder
CX3CL1	chemokine (C-X3-C motif) ligand 1	Colorectal cancer
SEMA3	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	Lung cancer, prostate cancer
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	Ameloblastoma, acute T cell leukemia, gastrointestinal stromal tumor
PLAU	plasminogen activator, urokinase	Lung cancer and many other cancers
ASS1	argininosuccinate synthetase 1	No
HMGA2	high mobility group AT-hook 2	No
CDKN2A	cyclin-dependent kinase inhibitor 2A	Lung cancer and many other cancers
CYR61	cysteine-rich, angiogenic inducer, 61	No
FER1L3	fer-1-like 3, myoferlin (C. elegans)	No
C14orf147	chromosome 14 open reading frame 147	No
SLC38A1	solute carrier family 38, member 1	No
PLAC8	placenta-specific 8	No

库的 344 个基因集进行富集分析，结果发现：有 206 条通路在肺腺癌Ⅲ~Ⅳ期患者中上调表达，但显著性上调的基因集数仅有 22 个，错误发现率< 25%的基因集 3 个，主要包括细胞凋亡、细胞黏附和迁移等信号通路过程(Table 3)。

3　讨　论

肿瘤发生是多种遗传事件共同作用的结果，高通量表达谱分析提供了从整体水平了解肿瘤发生机制的途径。在生物学过程中，某条通路中大多数基因

Table 3　Enrichment of genesets related to biological pathways

No	Pathway	Size	NES	*P*	FDR
1	Death pathway	32	−1.893	<0.001	0.102
2	Leukocyte transendothelial migration	101	−1.794	0.002	0.223
3	Focal adhesion	187	−1.752	0.004	0.251
4	ECM-receptor interaction	81	−1.726	0.004	0.261
5	Apoptosis KEGG	49	−1.669	0.026	0.387
6	Mitochondria pathway	20	−1.664	0.006	0.338
7	Caspase pathway	22	−1.638	0.020	0.380
8	Apoptosis GenMAPP	43	−1.616	0.024	0.411
9	Apoptosis	79	−1.593	0.012	0.455
10	Proteasome	17	−1.592	0.008	0.414
11	Cell communication	109	−1.590	0.019	0.382
12	LairPathway	15	−1.548	0.021	0.500
13	Epithelial cell signaling in Helicobacter pylori infection	61	−1.546	0.016	0.468
14	Cytokine-cytokine receptor interaction	226	−1.544	0.034	0.443
15	Statin pathway PharmGKB	17	−1.536	0.040	0.440
16	Regulation of Actin Cytoskeleton	181	−1.516	0.018	0.484
17	Hypertrophy model	20	−1.508	0.037	0.484
18	Cell adhesion molecules	115	−1.508	0.054	0.458
19	Akt pathway	17	−1.499	0.040	0.439
20	TGF beta signaling pathway	81	−1.464	0.050	0.465
21	GH pathway	27	−1.440	0.031	0.438
22	Tight junction	120	−1.432	0.014	0.446

Note：data in the table were lined by normalized enrichment score.

表达都发生轻微改变，从而使得整条通路的功能发生变化，这样的生物学意义显然远远高于表达改变倍数高的单个基因。基因集富集分析（gene set enrichment analysis，GSEA）关注的是整个杂交数据在特定功能基因集中的表达一致性，而不是某几个表达发生显著改变的基因，可以从全局水平分析不同表型的基因表达变化[11]。本研究从差异表达基因和生物信号通路两个角度，对晚期肺腺癌分子机制进行初步研究，以期揭示晚期肺腺癌预后、恶性改变等事件的分子机制。

基因表达差异的比较分析是在转录水平上鉴别组织或细胞间基因表达与否和基因表达量差异的技术，是揭示生物体发育和分化机制最有效的途径，在疾病相关基因分离等研究领域有极广泛的应用，是基因组学研究的核心领域之一。本研究中，我们通过基因差异表达分析方法，发现 SEMA3、PLAU、CDKN2A 等 14 个基因在晚期（Ⅲ~Ⅵ期）与早期（Ⅰ期）肺腺癌组织中明显异常表达，经基因注释表明 14 个差异表达基因主要与细胞凋亡、细胞黏附等过程密切相关，经文献挖掘发现，有 7 个基因已证实与人类肿瘤发生有关，并且 SEMA3、PLAU、CDKN2A、EPHX1 等基因已被证实与肺癌发生发展密切相关[12,13]。

Death pathway(死亡通路)和 Leukocyte transendothelial migration(白细胞经内皮移行)信号通路与肿瘤发生发展关系密切[14,15]。本研究通过 GSEA 通路富集分析发现：Death pathway 和 Leukocyte transendothelial migration 两条信号通路在晚期肺腺癌中表达明显上调，提示这两条信号通路的异常改变可能是晚期肺腺癌重要的分子事件。综上研究，我们发现 GSEA 与差异基因分析结果是一致的，即细胞黏附和细胞凋亡异常可能是晚期肺腺癌重要的分子事件。

基于基因表达谱分析，本研究发现了一些与晚期肺腺癌相关的分子，特别是首次发现 Death pathway 和 Leukocyte transendothelial migration 两条信号通路可能是晚期肺腺癌重要的分子事件。本研究对进一步深入研究晚期肺腺癌预后标志、分子治疗等均具有重要意义。然而，本研究仅从生物信息学角度获得了晚期肺腺癌的分子机制，仍需通过进一步的实验验证。

参考文献：

[1] Parkin DM，Bray F，Ferlay J，et al. Global cancer statistics，2002[J]. CA Cancer J Clin，2005，55(2): 74–108.

[2] Yang L，Li LD，Chen YD，et al.Mortality time trends and the incidence and mortality estimation and projection for lung cancer in China[J]. Chin J Lung Cancer，2005，8(4):274–278.[杨玲，李连弟，陈育德，等. 中国肺癌死亡趋势分析及发病、死亡的估计与预测 [J]. 中国肺癌杂志，2005，8(4):274–278.]

[3] Soon YY，Stockler MR，Askie LM，et al. Duration of chemotherapy for advanced non-small-cell lung cancer: a systematic review and meta-analysis of randomized trials [J]. J Clin Oncol，2009，27(20):3277–3283.

[4] Krepela E，Dankova P，Moravcikova E，et al. Increased expression of inhibitor of apoptosis proteins，survivin and XIAP，in non-small cell lung carcinoma [J]. Int J Oncol，2009，35(6):1449–1462.

[5] Talbot SG，Estilo C，Maghami E，et al. Gene expression profiling allows distinction between primary and metastatic squamous cell carcinomas in the lung [J]. Cancer Res，2005，65(8):3063–3071.

[6] Glinsky GV，Glinskii AB，Stephenson AJ，et al. Gene expression profiling predicts clinical outcome of prostate cancer [J]. J Clin Invest，2004，113(6):913–923.

[7] Korrat A，Greiner T，Maurer M，et al. Gene signature-based prediction of tumor response to cyclophosphamide[J]. Cancer Genomics Proteomics，2007，4(3):187–195.

[8] Amin SB，Shah PK，Yan A，et al.The dChip survival analysis module for microarray data[J]. BMC Bioinformatics，2011，12:72.

[9] Wang GP，Ye Y，Yang XQ，et al.Gene expression profiles-based approach identifies candidate therapeutic compounds for lung adenocarcinoma[J].Chinese Journal of Clinical Pharmacology and Therapeutics，2010，15(3) :266–272.[王桂平，叶云，杨晓勤，等.基于基因表达谱的途径筛选肺腺癌治疗药物[J].中国临床药理学与治疗学，2010，15(3) :266–272.]

[10] Irizarry RA，Wang C，Zhou Y，et al. Gene set enrichment analysis made simple [J]. Stat Methods Med Res，2009，18(6): 565–575.

[11] Subramanian A，Kuehn H，Gould J，et al. GSEA-P:a desktop application for gene set enrichment analysis[J]. Bioinformatics，2007，23(23): 3251–3253.

[12] Andujar P，Wang J，Descatha A，et al.P16INK4A inactivation mechanisms in non-small-cell lung cancer patients occupationally exposed to asbestos[J].Lung Cancer，2010，67(1):23–30.

[13] Tomizawa Y，Sekido Y，Kondo M，et al. Inhibition of lung cancer cell growth and induction of apoptosis after reexpression of 3p21.3 candidate tumor suppressor gene SEMA3B[J]. PNAS，2001，98(24):13954–13959.

[14] Zhang X，Miao X，Sun T，et al.Functional polymorphisms in cell death pathway genes FAS and FASL contribute to risk of lung cancer [J]. J Med Genet，2005，42(6):479–484.

[15] Jöhrer K，Janke K，Krugmann J，et al. Transendothelial migration of myeloma mells is increased by tumor necrosis factor (TNF)-α via TNF receptor 2 and autocrine up-regulation of MCP-1[J]. Clin Cancer Res，2004，10(6):1901–1910.