European Review for Medical and Pharmacological Sciences 2014; 18: 117-125

Screening of genes related to lung cancer caused by smoking with RNA-Seq

C. ZHOU, H. CHEN, L. HAN, F. XUE, A. WANG, Y.-J. LIANG[1]

Department of Respiratory Medicine, Zhou Pu Hospital, Shanghai, China
[1]Department of Respiratory Medicine, East Hospital, Tongji University, Shanghai, China

Abstract. – AIM: To study the lung cancer caused by smoking from RNA-seq data and its mechanism at molecular level.

MATERIALS AND METHODS: We downloaded gene expression profile SRA (Sequence Read Archive) data from Gene Expression Omnibus database that included two samples: one was lung cancer tissue samples from smoker (GSM718710) and the other was from non-smoker (GSM718709). We analyzed differential expression of genes with packages software TopHat and Cufflinks, and did Gene Ontology (GO) function clustering of the differentially expressed genes by BLASTX. Then we utilized KEGG Orthology Based Annotation System (KOBAS) to make pathway annotation and do enrichment analysis of KEGG pathway. After that, we searched for probable alternative splicing of the selected differentially expressed genes and found closely-linked genes.

RESULTS: we screened 1603 differentially expressed genes, most of which were involved in cellular processes. We also identified that the possible alternative splicing of gene FCGBP might have an important impact on lung cancer.

CONCLUSIONS: These findings in this study may help better understand the relationship between smoking and lung cancer pathogenesis.

Key Words:
Lung cancer, Smoking, Differentially expressed genes, FCGBP.

Introduction

Lung cancer is one of the leading malignant tumors and the common cause of cancer-related deaths[1]. The 5 year survival for lung cancer still remains relatively poor, possibly because lung cancer is often diagnosed at advanced stage and treatment options are limited. The predominant causal factor for lung cancer is tobacco smoking, which can alter the activity of chemo-preventive drugs[2,3], stimulate the clearance of selected targeted anticancer therapies[4], reduce the efficacy of cancer treatment[5,6] and increase the risk of second primary tumors. Therefore, it is urgently needed to improve the understanding of the molecular and cellular mechanisms of lung cancer caused by smoking in order to develop new more effective strategies in preventing and treating lung cancer.

Tobacco smoking is a major risk factor in the etiology of lung cancer which has been reported in many researches[7-9]. There is evidence that smoking affects the mechanism of lung carcinogenesis. The common genetic changes found in lung cancers include loss of heterozygosity (LOH) at alleles on chromosome 3p encompassing fragile histidine triade gene (FHIT)[8], mutations in TP53, defects in the p16INK4/RB pathway, Semaphorin 3B (SEMA3B) and RASSF1A, aberrant promoter methylation in O6-methylguanine-DNA methyltransferase (O6MGMT), p16INK4, death-associated protein kinase (DAPK), tissue inhibitor of and metalloproteinase-3 (TIMP-3)[9]. However, the mechanism of lung carcinogenesis especially the roles that lung cancer related genes play are still not very clear.

In this study, we analyzed RNA-seq data from lung cancer samples caused by smoking to screen the genes closely related with lung cancer in order to study and better understand its molecular mechanisms and explore new therapy strategies.

Materials and Methods

Source of RNA-seq Sequencing Data

SRA (Sequence Read Archive) sequencing data was downloaded from the gene expression database of GEO (Gene Expression Omnibus), which included two samples: one was from a smoker with lung cancer (GSM718710), and the other was from a smoker without lung cancer (GSM718709). Each sample was sequenced by pair-ended of 100 bp (Table I).

Corresponding Author: Chao Zhou, MM; Yongjie Liang, MD; e-mail: chaozhouz@hotmail.com

Table I. The data samples for analysis.

Sample name	Type	Library	Data size	ENA ID
GSM718709	Smoker without Lung cancer	Paired-end	4.45G base	SRR192335
GSM718710	Smoker with Lung cancer	Paired-end	4.18G base	SRR192336

Processing of the Download Raw Data

The downloaded data were the sequencing results of mRNA in different samples by pair-ended using next generation sequencing technologies. SRA Toolkit[10] software provided by the NCBI (National Center for Biotechnology Information) was used to transform the original data in SRA format, so as to obtain the standard sequencing data in FASTQ[11] format.

Positioning of the Original Reads

We took the human reference genome hg19 provided by University of California Santa Cruz (UCSC)[12] website as analysis reference in our study, and then aligned reads to hg19 by TopHat software. There were three steps in Tophat mapping, at first the reads with ability to match up with hg19 were aligned directly, then connected the exons of the reference genome as a reference and aligned the rest of reads according to the corresponding annotation files of hg19, finally the rest reads were cut and mapped to the reference genome, moreover we did statistics on the results by aligning and positioning.

Calculation of Gene Expression Values

We used cufflinks to calculate the gene expression values (RPKM value). RPKM[13] was the result with the number of reads mapped to gene divided by the number of all reads mapped to genome (in million) and the length of RNA (in KB).

Screening of the Differentially Expressed Genes

We calculated the differences in gene expression among different samples with cuffdiff module in cufflinks and *p*-value of significant difference. *p* value less than 0.05 was thought as significant difference.

Functional Annotation of the Differentially Expressed Genes

We compared the sequence of differential genes treated by 3 nmol/L and 10 nmol/L of genistein with COG[14] (clusters of orthologous groups of proteins) (http://www.ncbi.nlm.nih.gov/COG) by BLASTX[15] (with similar threshold: E-value <1e-05). The functional annotation of the differentially expressed genes and COG function classification were obtained at the same time. Besides the function regulated by genes expressed specifically through COG classifications could be understood intuitively and emotionally.

Analysis of KEGG Pathway Annotation and Enrichment of the Differentially Expressed Genes[16]

KOBAS[17] was used to analyze the pathway annotation and enrichment of the differentially expressed genes and chose FDR (false discovery rate) less than 0.05 as the threshold. The statistical method used was cumulative hypergeometric distribution.

Clustering Analysis of the Differentially Expressed Genes

The clustering map of the differentially expressed genes was drawn according to the expression values of differential genes in different samples by gplots package in R language.

Analysis of Alternative Splicing of the Differentially Expressed Genes

We concluded all the possible splicing situations of the genes to our interest based on the situation of reads alignment and visualized them by R language.

Results

Results of the original Reads Positioning and Gene Expression Values

We aligned the reads from sequencing to hg19 by using TopHat software and hg19 from UCSC as reference. All results of the reads positioning are shown in Table II.

The gene expression values (RPKM: reads per kilobase per million reads value) were calculated by cufflinks and the distribution of which was also counted (as shown in Table III and Figure 1),

Table II. Position results of the sequencing data.

Clean Rds Num	Mapping Rds	Mapping rate	Total gene num	Mapping gene num	Mapping gene rate
44511672	35564170	79.89%	25245	17961	71.14%
41861340	33912253	81.01%	25245	17874	70.80%

Table III. Distribution of expression values of the differentially expressed genes.

RPKM range	SRR192335 (frequency)	SRR192336 (frequency)
(0, 10)	12086	11815
[10, 20)	2509	2544
[20, 30)	1071	1104
[30, 40)	564	585
[40, 50)	360	375
[50, 60)	221	245
[60, 70)	167	171
[70, 80)	126	138
[80, 90)	124	100
[90, 100)	76	96
[100, INF)	657	701

from which we were able to draw that most of the expression values were distributed between 1-10.

Screening Results of the Differentially Expressed Genes

Firstly we did *t*-test[18] on the gene expression values of the two samples to gain p-value and then corrected it by FDR (false discovery rate) to obtain adj-value. After that, we removed the genes greater than the default threshold (| log FC | > 1, and FDR < 0.01). Finally, we found a total of 1603 differentially expressed genes. The similarities and differences of gene expression in two samples are shown in Figure 2.

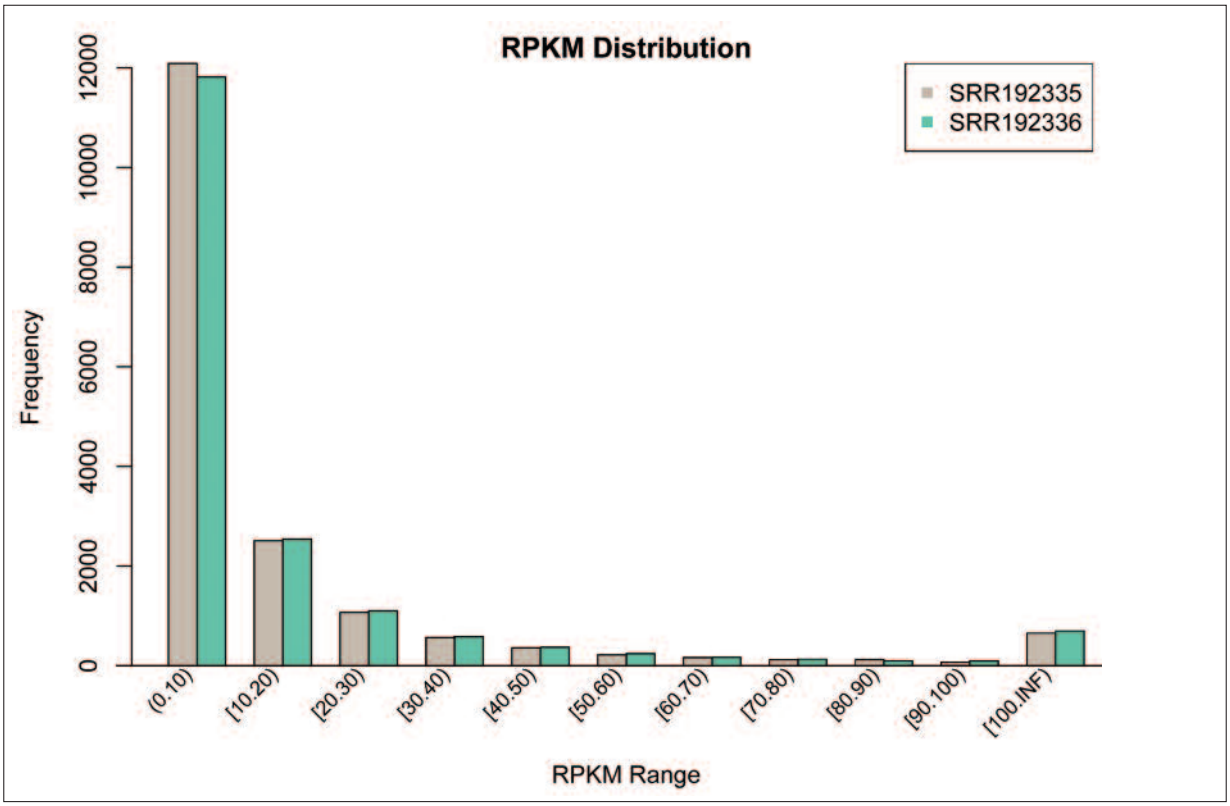

Figure 1. Distribution of gene expression values.

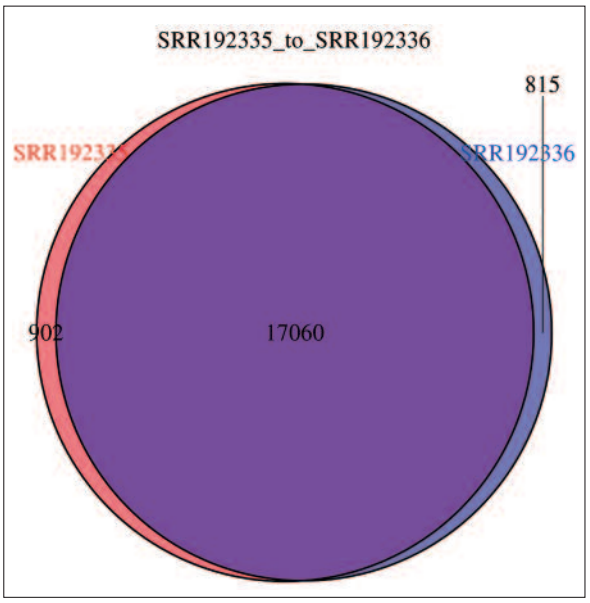

Figure 2. Gene expression in two different samples.

Functional Classification and Pathway of the Differentially Expressed Genes

The sequences of the differentially expressed genes screened from two groups separately were aligned with COG[19] orthologous clusters database by BLASTX (similar threshold: E-value < 1e-05). The GO function nodes of the differentially expressed genes were classified based on similarity degree between the sequence and gene sequences on each GO[20] node recorded in Ontology (as shown in Figure 3). In addition, we did statistics on the differentially expressed genes participated in every GO function node. Ultimately, we got GO function described from three directions: cell component (such as organelle, membrane-enclosed lumen) biological processes (such as nitrogen utilization, carbohydrate utilization) and molecular function (such as chemo-repellent activity, translation regulator activity). In view of the differentially expressed genes involved in cellular processes, the number of genes participated in the cell biological process was the maxim.

As pathway analysis played an important role in explaining the pathogenesis of the disease researched in this study, so we made KEGG annotation and did pathway enrichment analysis of all the differentially expressed genes by using KOBAS software. The detailed results are exhibited in Table IV. There were 42 pathways significantly enriched including Neurotrophin signaling pathway, p53 signaling pathway, ErbB signaling pathway and so on. The hsa04722 pathway is shown in Figure 4.

Clustering and Alternative Splicing of the Differentially Expressed Genes

In order to illustrate the distance of expression values among genes, an expression clustering map was drawn based on values of the differential genes expressed in different samples with R gplots package (As shown in Figure 5).

We discovered that only one gene IgG Fc binding protein (FCGBP) in all the differentially expressed genes had alternative splicing[21] (as shown in Figure 6).

Discussion

In this study, the next generation sequencing technology RNA-seq was chosen to sequence the mRNA in normal tissues and cancer cells and constructed a cDNA library. Then, we found genes with significant expression difference by RNA-seq analysis method so as to explore the pathogenesis of lung cancer at genetic level. The results showed, the differentially expressed genes screened were mostly involved in cellular processes which indicated that smoking could influence the normal cell growth in lung to induce carcinomatous changes. As shown in Figure 4, a number of differentially expressed genes were identified including interleukin 6 (IL-6), interleukin 8 (IL-8), matrix metalloproteinase 10 (MMP10) and so on. Among which, IL-6 and IL-8 are of particular interest because they are expressed in pre-malignant epithelial cells, and their expression is associated with a poor prognosis in lung cancer patients[22,23]. Evidence shows that inflammatory mediators contribute to the pathogenesis of many human cancers, including lung cancer[24,25,26], as under inflammatory stress, IL-6 and IL-8 participate in tumorigenesis by acting directly on lung epithelial cells via signaling through the nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (NFkB1) pathway[27]. In addition, IL-6 and IL-8 are expressed by lung cancer cells and act in an autocrine and/or paracrine fashion to stimulate cancer cell proliferation[28,29], migration, and invasion[30]. What's more, IL-6 can regulate MMP10 expression via JAK2/STAT3 signaling pathway[31]. MMP10, also known as stromelysin-2, is one of the well characterized members of the MMP family. MMP10 has been demonstrated to be over-expressed in several human tumors of epithelial origin, including gastric cancer[32], skin carcinoma[33] and non-small cell lung cancer

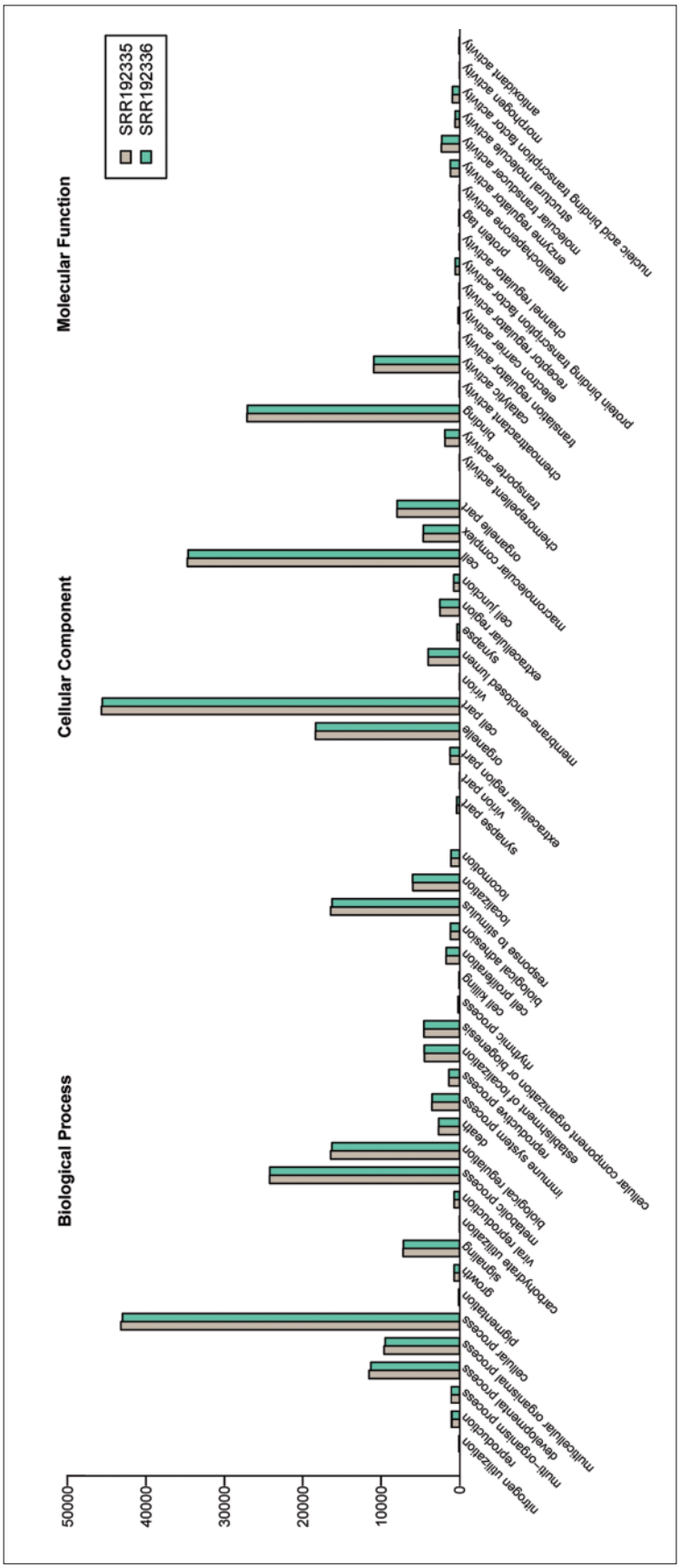

Figure 3. COG classification of the differentially expressed genes.

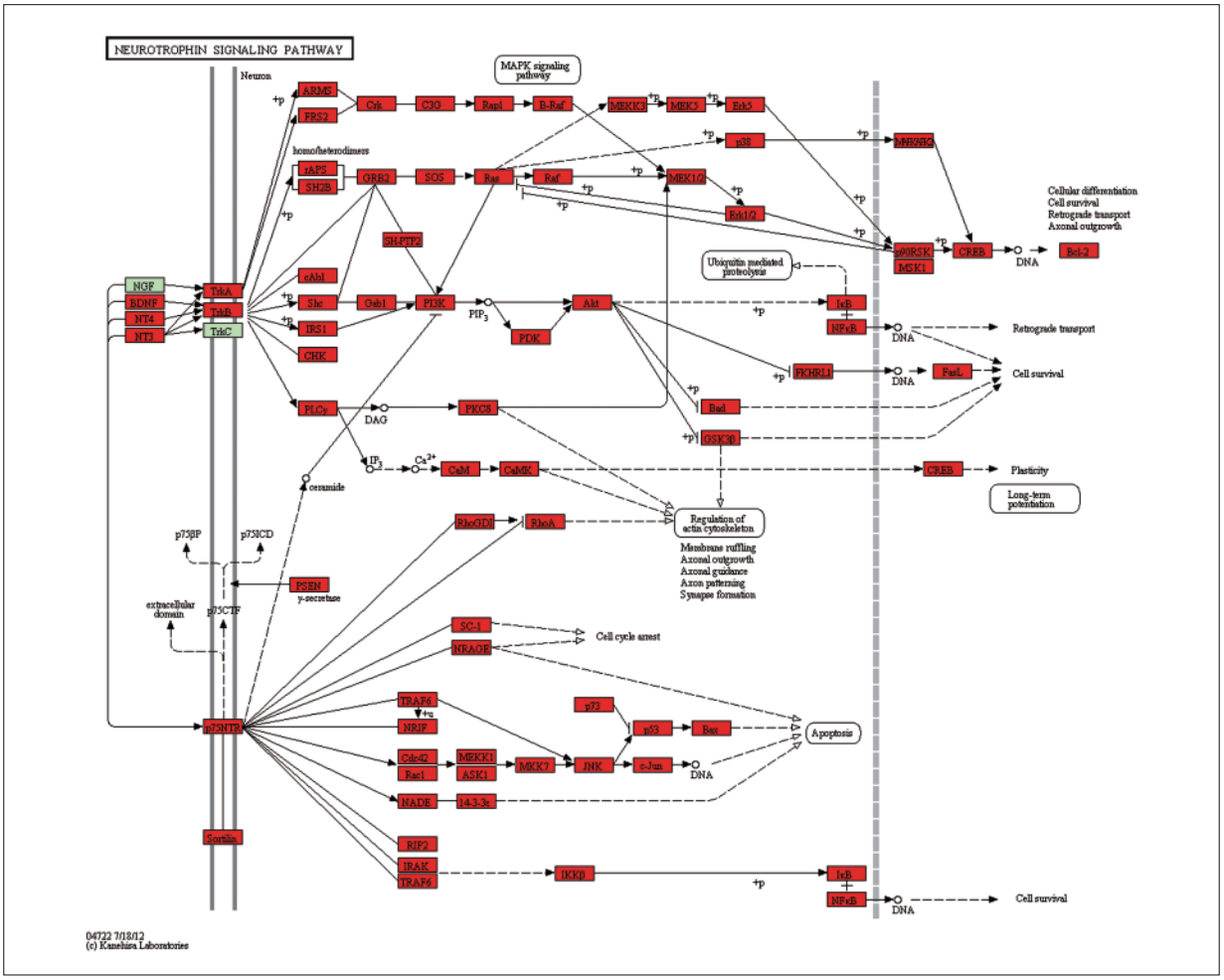

Figure 4. The pathway enriched significantly (hsa04722). The position marked red was the differential genes annotated.

(NSCLC)[34] and so on, which indicates MMP10 may play an important role in the development and progression of lung cancer.

It is worth to mention that cytochrome P450 (CYP) related enzymes can convert the carcinogens present in tobacco and tobacco smoke into DNA reactive metabolites. For example, CYP2A6 and CYP4X1, both were genetic variants of CYP, were also found in the clustering Figure of the differentially expressed genes. It is reported that some CYP variants are associated with increased risks for cancers of the lung, esophagus, and head and neck. For example, CYP2A6 can mediate 7-hydroxylation of coumarin, a component of cigarette smoke, and activates several nitrosamines in tobacco smoke, including NNK (nicotine-derived nitrosamine ketone)[35,36], moreover, individuals who lack functional *CYP2A6* have impaired nicotine metabolism and may thus be protected against tobacco dependence. Judged from the pathway results, the most relevant one was for signal transduction in cell process[37].

From the Neurotrophin signaling pathway in Figure 4, we could see brain derived neurotrophic factor (BDNF), a member of the neurotrophin family of growth factors, and its high-affinity receptor, tropomyosin receptor kinase A, B (TrkA, TrkB) were differentially expressed. Evidence shows that TrKB and BDNF promote tumor cells survival[38] and angiogenesis, and contribute to resistance to cytotoxic drugs and anoikis, enabling cells to acquire many of the characteristic features required for tumorigenesis[39]. A number of studies have focused on the BDNF/TrkB signal transduction pathway to explore their expression in different tumor types, correlating expression with prognosis and tumor stage[40,41].

In addition, we also found IgG Fc-binding protein (FCGBP) was the only gene with alternative splicing. FCGBP was first identified as an

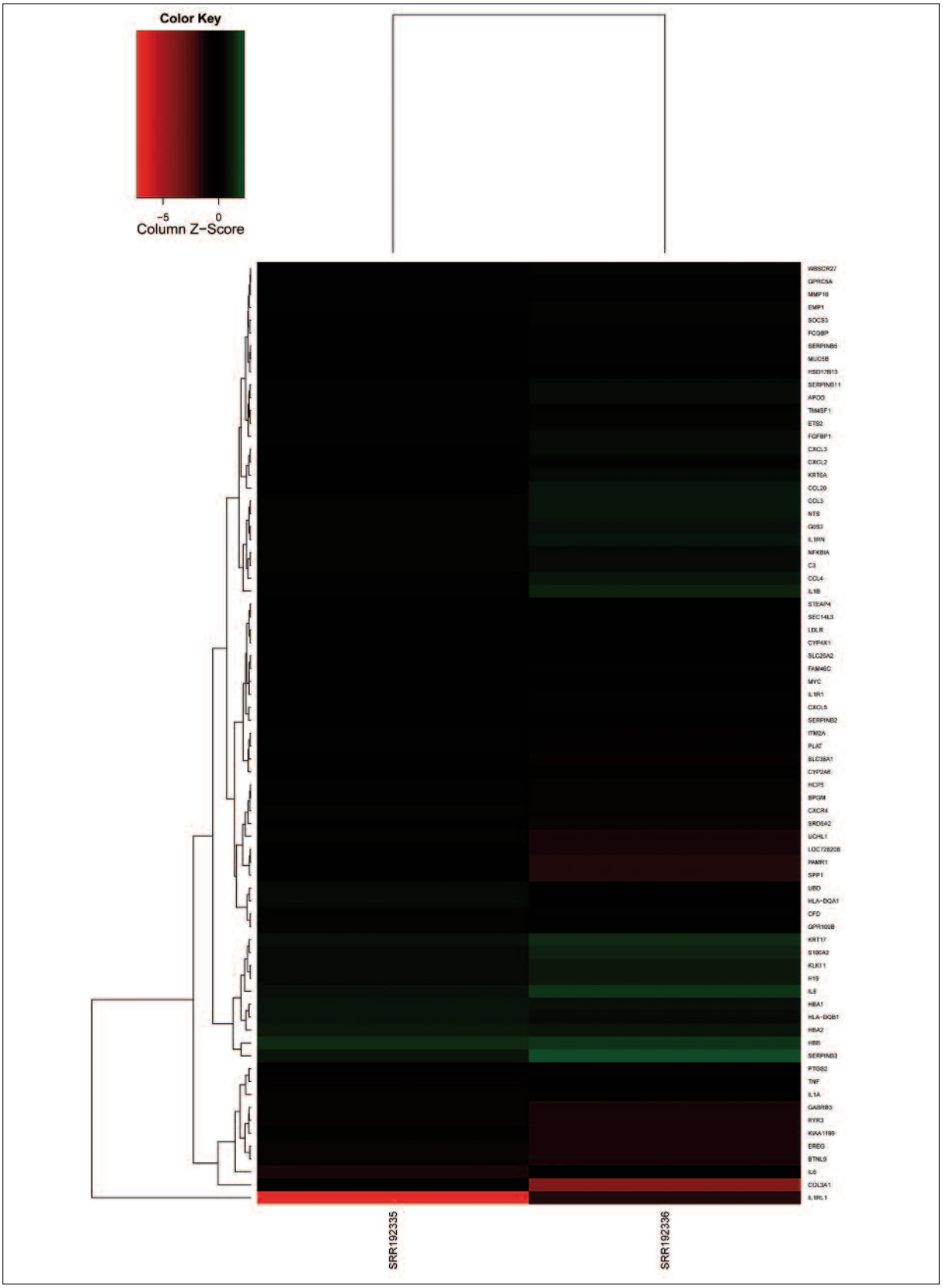

Figure 5. Clustering of the differentially expressed genes.

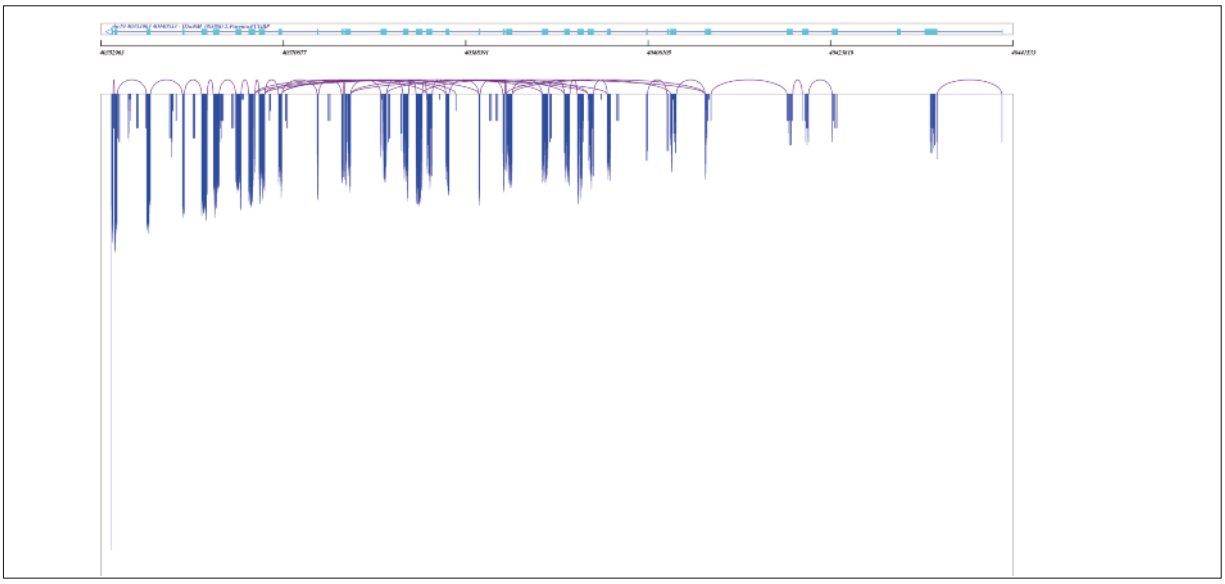

Figure 6. The possible alternative splicing of FCGBP. The arrow indicated the direction of the gene in the chromosome: 3 '-> 5'. Blue area represented the depth of base covering in this region, one blue area indicated an expressed exons, and two exons connected by a purple line mean the reads with one side in a exon while the other in another exon, that was the junction read.

IgG Fc binding site in intestinal and colonic epithelia[42] and it was produced by goblet cells in the colon secreted into the bowel lumen with mucus, which suggested that it might possibly contribute to immune protection in lung.

Conclusions

We have found several differentially expressed genes related to lung cancer caused by smoking such as IL6, IL8, and FCGBP in this study, which may help better understand the pathogenesis of lung cancer, especially FCGBP may be used as a new biomarker for diagnosis, prevention or treatment for lung cancer.

Conflict of Interest

The Authors declare that there are no conflicts of interest.

References

1) Youlden DR, Cramb SM, Baade PD. The international epidemiology of lung cancer: geographical distribution and secular trends. J Thor Oncol 2008; 3: 819-831.
2) Mayne ST, Lippman SM. Cigarettes: a smoking gun in cancer chemoprevention. J Natl Cancer Inst 2005; 97: 1319-1321.
3) [No authors listed] The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. N Engl J Med 1994; 330: 1029-1035.
4) Hamilton M, Wolf JL, Rusk J, Beard SE, Clark GM, Witt K, Cagnoni PJ. Effects of smoking on the pharmacokinetics of erlotinib. Clin Cancer Res 2006; 12: 2166-2171.
5) Fox JL, Rosenzweig KE, Ostroff JS. The effect of smoking status on survival following radiation therapy for non-small cell lung cancer. Lung Cancer 2004; 44: 287-293.
6) Pantarotto J, Malone S, Dahrouge S, Gallant V, Eapen L. Smoking is associated with worse outcomes in patients with prostate cancer treated by radical radiotherapy. BJU Int 2007; 99: 564-569.
7) Park SK, Cho LY, Yang JJ, Park B, Chang SH, Lee K-S, Kim H, Yoo K-Y, Lee C-T. Lung cancer risk and cigarette smoking, lung tuberculosis according to histologic type and gender in a population based case–control study. Lung Cancer 2010; 68: 20-26.
8) Holschneider CH, Baldwin RL, Tumber K, Aoyama C, Karlan BY. The fragile histidine triad gene: a molecular link between cigarette smoking and cervical cancer. Clin Cancer Res 2005; 11: 5756-5763.
9) Yokota J, Kohno T. Molecular footprints of human lung cancer progression. Cancer Sci 2004; 95: 197-204.
10) Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y. DDBJ launches a new archive database with analytical tools for next-generation sequence data. Nucleic Acids Res 2010; 38: D33-D38.
11) Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009; 25: 1966-1967.
12) Simone D, Calabrese FM, Lang M, Gasparre G, Attimonelli M. The reference human nuclear mitochondrial sequences compilation validated and

implemented on the UCSC genome browser. BMC Genomics 2011; 12: 517.

13) Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012; 131: 281-285.

14) Natale D, Galperin M, Tatusov R, Koonin E. Using the COG database to improve gene recognition in complete genomes. Genetica 2000; 108: 9-17.

15) Mount DW. Using the basic local alignment search tool (BLAST). Cold Spring Harbor Protocols 2007; 2007: pdb. top17.

16) Aoki-Kinoshita KF, Kanehisa M. Gene annotation and pathway mapping in KEGG, in Comparative Genomics. Springer, 2007; pp. 71-91.

17) Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C-Y, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011; 39: W316-W322.

18) Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3: 3.

19) Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001; 29: 22-28.

20) Blake J, Chan J, Kishore R, Sternberg P, Van Auken K. Gene Ontology annotations and resources. Nucleic Acids Res 2013; 41: D530-D535.

21) Pervouchine DD, Knowles DG, Guigó R. Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics 2013; 29: 273-274.

22) Davalos AR, Coppe J-P, Campisi J, Desprez P-Y. Senescent cells as a source of inflammatory factors for tumor progression. Cancer Metastasis Rev 2010; 29: 273-283.

23) Seike M, Yanaihara N, Bowman ED, Zanetti KA, Budhu A, Kumamoto K, Mechanic LE, Matsumoto S, Yokota J, Shibata T. Use of a cytokine gene expression signature in lung adenocarcinoma and the surrounding tissue as a prognostic classifier. J Natl Cancer Inst 2007; 99: 1257-1269.

24) Coussens LM, Werb Z. Inflammation and cancer. Nature 2002; 420: 860-867.

25) Engels EA. Inflammation in the development of lung cancer: epidemiological evidence. Exp Rev Anticancer Ther 2008; 8: 605-615.

26) Perwez Hussain S, Harris CC. Inflammation and cancer: an ancient link with novel potentials. Int J Cancer 2007; 121: 2373-2380.

27) Landi S, Moreno V, Patricola LG, Guino E, Navarro M, De Oca J, Capella G, Canziani F, and; for the Bellvitage Colorectal Cancer Study Group. Association of common polymorphisms in inflammatory genes interleukin (IL) 6, IL8, Tumor Necrosis Factor α, NFK B1, and peroxisome proliferator-activated receptor δ with colorectal cancer. Cancer Res 2003; 63: 3560-3566.

28) Kamohara H, Ogawa M, Ishiko T, Sakamoto K, Baba H. Leukemia inhibitory factor functions as a growth factor in pancreas carcinoma cells: Involvement of regulation of LIF and its receptor expression. Int J Oncol 2007; 30: 977-983.

29) Takamori H, Oades ZG, Hoch RC, Burger M, Schraufstatter IU. Autocrine Growth Effect of IL-8 and GRO [alpha] on a Human Pancreatic Cancer Cell Line, Capan-1. Pancreas 2000; 21: 52-56.

30) Lang K, Niggemann B, Zanker KS, Entschladen F. Signal processing in migrating T24 human bladder carcinoma cells: Role of the autocrine interleukin 8 loop. Int J Cancer 2002; 99: 673-680.

31) Zhang X, Yin P, Di D, Luo G, Zheng L, Wei J, Zhang J, Shi Y, Zhang J, Xu N. IL-6 regulates MMP-10 expression via JAK2/STAT3 signaling pathway in a human lung adenocarcinoma cell line. Anticancer Res 2009; 29: 4497-4501.

32) Aung P, Oue N, Mitani Y, Nakayama H, Yoshida K, Noguchi T, Bosserhoff A, Yasui W. Systematic search for gastric cancer-specific genes based on SAGE data: melanoma inhibitory activity and matrix metalloproteinase-10 are novel prognostic factors in patients with gastric cancer. Oncogene 2005; 25: 2546-2557.

33) Kerkelä E, Ala-Aho R, Jeskanen L, Lohi J, Grénman R, M-Kähäri V, Saarialho-Kere U. Differential patterns of stromelysin-2 (MMP-10) and MT1-MMP (MMP-14) expression in epithelial skin cancers. Br J Cancer 2001; 84: 659.

34) Zhang X, Zhu S, Luo G, Zheng L, Wei J, Zhu J, Mu Q, Xu N. Expression of MMP-10 in lung cancer. Anticancer Res 2007; 27: 2791-2795.

35) Hecht SS. DNA adduct formation from tobacco-specific N-nitrosamines. Mutat Res 1999; 424: 127.

36) Su T, Bao Z, Zhang Q-Y, Smith TJ, Hong J-Y, Ding X. Human cytochrome P450 CYP2A13: predominant expression in the respiratory tract and its high efficiency metabolic activation of a tobacco-specific carcinogen, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone. Cancer Res 2000; 60: 5074-5079.

37) Capaccione KM, Pine SR. The notch signaling pathway as a mediator of tumor survival. Carcinogenesis 2013;

38) Pearse RN, Swendeman SL, Li Y, Rafii D, Hempstead BL. A neurotrophin axis in myeloma: TrkB and BDNF promote tumor-cell survival. Blood 2005; 105: 4429-4436.

39) Thiele CJ, Li Z, McKee AE. On Trk–the TrkB signal transduction pathway is an increasingly important target in cancer biology. Clin Cancer Res 2009; 15: 5962-5967.

40) Asgharzadeh S, Pique-Regi R, Sposto R, Wang H, Yang Y, Shimada H, Matthay K, Buckley J, Ortega A, Seeger RC. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. J Natl Cancer Ins 2006; 98: 1193-1203.

41) Nakagawara A, Azar CG, Scavarda NJ, Brodeur GM. Expression and function of TRK-B and BDNF in human neuroblastomas. Mol Cell Biol 1994; 14: 759-767.

42) Kobayashi K, Blaser M, Brown W. Identification of a unique IgG Fc binding site in human intestinal epithelium. J Immunol 1989; 143: 2567-2574.