RESEARCH ARTICLE

Finding Genes Discriminating Smokers from Non-smokers by Applying a Growing Self-organizing Clustering Method to Large Airway Epithelium Cell Microarray Data

Maryam Shahdoust[1]*, Ebrahim Hajizadeh[2], Hossein Mozdarani[3], Ali Chehrei[3]

Abstract

Background: Cigarette smoking is the major risk factor for development of lung cancer. Identification of effects of tobacco on airway gene expression may provide insight into the causes. This research aimed to compare gene expression of large airway epithelium cells in normal smokers (n=13) and non-smokers (n=9) in order to find genes which discriminate the two groups and assess cigarette smoking effects on large airway epithelium cells. Materials and Methods: Genes discriminating smokers from non-smokers were identified by applying a neural network clustering method, growing self-organizing maps (GSOM), to microarray data according to class discrimination scores. An index was computed based on differentiation between each mean of gene expression in the two groups. This clustering approach provided the possibility of comparing thousands of genes simultaneously. Results: The applied approach compared the mean of 7,129 genes in smokers and non-smokers simultaneously and classified the genes of large airway epithelium cells which had differently expressed in smokers comparing with non-smokers. Seven genes were identified which had the highest different expression in smokers compared with the non-smokers group: NQO1, H19, ALDH3A1, AKR1C1, ABHD2, GPX2 and ADH7. Most (NQO1, ALDH3A1, AKR1C1, H19 and GPX2) are known to be clinically notable in lung cancer studies. Furthermore, statistical discriminate analysis showed that these genes could classify samples in smokers and non-smokers correctly with 100% accuracy. With the performed GSOM map, other nodes with high average discriminate scores included genes with alterations strongly related to the lung cancer such as AKR1C3, CYP1B1, UCHL1 and AKR1B10. Conclusions: This clustering by comparing expression of thousands of genes at the same time revealed alteration in normal smokers. Most of the identified genes were strongly relevant to lung cancer in the existing literature. The genes may be utilized to identify smokers with increased risk for lung cancer. A large sample study is now recommended to determine relations between the genes ABHD2 and ADH7 and smoking.

Keywords: Lung cancer - cigarette smoking - gene expression - microarray - growing self-organizing maps

Asian Pacific J Cancer Prev, **14 (1)**, 111-116

Introduction

Lung cancer is one of the most frequent human cancers and the leading cause of cancer- related death in males and the second leading cause of cancer death among females (Jemal et al., 2011). Smoking, particularly smoking cigarette, is one of the main contributor to lung cancer (Spira et al., 2004; Jemal et al., 2011). Cigarette smoking injures airway epithelium cells exposed to it. A number of studies show that noncancerous large-airway epithelium cells of current and former smokers with and without lung cancer exhibit allelic loss (Wistuba et al.,1997; Powell et al., 1999), P53 mutation (Franklin et al., 1997), changes in DNA methylation in the promoter regions of several genes (Guo et al., 2004) and also increased telomerase activity (Miyazu et al., 2005). Some microarray studies show that cigarette smoking up-regulates the expression of some lung cancer marker genes such as UCHL1 (Spira et al., 2004; Brendan et al., 2006; Beane et al.,2007; Spira et al., 2007; Cote et al., 2009; Pickett et al., 2009). The identification of effects of smoking on airway gene expression may provide an insight to study the cause of this elevated risk and to diagnosis and prognosis of the lung cancer. Therefore to asses these alterations, finding the genes which have the different expression and distinguish smokers from non-smokers could be useful. In 2003 Hsu et al., have introduced an approach to cancer class discovery and marker genes identification based on GSOMs. The approach has three phases; cancer class discovery, marker gene identification and refinements. The applied approach in this article is part of Hsu approach to compare smokers and non-smokers large airway epithelium cells gene expression in order to find genes which expressed differently in smokers group.

*[1]Biostatistician, [2]Department of Biostatistics, [3]Department of Medical Genetics, Tarbiat Modares University, Pathologist, Thyroid Disorder Research Center, Arak University of Medical Sciences, Iran *For correspondence: m.shahdoost@gmail.com*

Date set is microarray gene expression data of large airway epithelium cells (Brendan et al., 2006). The clustering variable was class discrimination score which was calculated based on differentiation between each mean of gene expression in smokers and non-smokers groups. This paper was aimed to identify the genes which discriminate the smokers from non-smokers in order to assess the effects of cigarette smoking on large airway epithelium cells by applying a neural network clustering method, growing self-organizing maps (GSOM) (Alahokoon et al., 2000; Hsu et al., 2003), to compare the gene expression of large airway epithelium cells in the normal smokers and the non-smokers. By applying the approach, we were able to compare the expression of genes at the same time in order to find differentiations and also to identify the effects of cigarette smoking.

Materials and Methods

Data set

Data set included large airway epithelium cells microarray information from 9 normal non-smokers and 13 normal smokers of their left lung. Each sample composed of 7129 genes expression levels. The data was a part of the Brendan et al. (2006) study, the up-regulation of expression of the ubiqutin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers, prepared in Dr Crystal lab. The data has been deposited in the Gene Expression Omnibus site, which is curated by the national Center for bioinformatics. The dataset is available in www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2489 .

Data standardization before applying algorithm is necessary. Equation (1) could be used for standardizing data (Hsu et al., 2003).

$$X'=(x-x_{min})/(x_{max}-x_{min}) \quad (1)$$

Identifying discriminating genes

To identify genes that were differently expressed between the smokers and the non-smokers in large-epithelium cells, the hexagonal Growing self-organizing maps (GSOMs) clustering method (Hsu et al.,2003) had been trained from "*class discrimination scores*" in smokers group (Galub et al., 2000; Hsu et al., 2003). This score was calculated by equation (2) for all genes.

$$P(g,1,2)=[\mu(g)_1-\mu(g)_2]/[\sigma(g)_1+\sigma(g)_2] \quad (2)$$

Where $\mu(g)_1$ and $\mu(g)_2$ are the mean of expression gene (g) in smokers (group 1) and non-smokers (group 2) and $\sigma(g)_1$ and $\sigma(g)_2$ are the standard deviations of the expression level of gene (g) for all samples belonging to smokers and non-smokers respectively. The high value of absolute $P(g,1,2)$ shows that the gene (g) is strongly suitable for discriminating smokers from non-smokers. In the map of nodes provided by the applied GSOMs, the node with highest average of class discrimination scores included the genes which discriminated smokers from non-smokers. We call these genes discriminating genes at the follow.

To evaluate the prediction strength of identified discriminating genes, a weighted vote for each gene was calculated by equation (3) for each sample (Galub et al., 2000; Hsu et al., 2003). A weighted vote of gene g for a sample shows the vote of gene g to asset the sample to the smokers group. It was supposed that the means of identified discriminating genes votes in the smoker group would be higher than the means of identified discriminating genes votes in the non-smokers.

$$v(g)=P(g,1,2)\{[x(g)]-[[\mu(g)_1+\mu(g)_2]/2]\} \quad (3)$$

Where $P(g,1,2)$ is the discrimination score for gene g, $x(g)$ is the gene expression for each sample is tested and $\mu(g)_1$ and $\mu(g)_2$ are the means of gene g for non-smokers and smokers groups respectively.

Man-Whitney U test was applied to compare the means of genes weighted votes of non-smokers and smokers. Also, discriminate analysis had been applied to see how much these marker genes are able to discriminate smokers from non-smokers.

In respect to the other experiences, marker genes are either in the vicinity of the highest average node on the trained GSOM or are included in identified predictor genes. So in the article, all the nodes were in the vicinity of the introduced node had been studied and also all the identified genes had been matched by clinical studies too.

Results

The nodes with highest average score of class discrimination (first ranked node) included seven genes (Table 1, Figure 1). Other nodes with high averages are in the neighborhood of the first ranked node, as expected (Table 2).

Table 3 shows the discriminating genes weighted votes for each sample. The smokers genes weighted votes were higher than non-smokers genes weighted votes. Also, Man-Whitney U tests results showed that the means of each genes weighted votes in two groups were significantly different ($p<0.05$).

The discriminate analysis results showed that seven identified genes of first ranked node could classify 100%

Table 1. Identified Marker Genes: Genes of First Node with High Average Score of Class Discrimination Scores (average score=1.45)

Number	P(g,1,2)	Gene name
1	1.38	NQO1
2	1.58	H19
3	1.43	ALDH3A1
4	1.56	AKR1C1
5	1.34	ABHD2
6	1.34	GPX2
7	1.59	ADH7

Figure 1. Part of the GSOM Map Trained from Class Discrimination Scores for Normal Smokers Group Showing Location of Marker Genes. The node number 183 is the node with highest discriminate score

Table 2. Genes in Other Nodes with High Discrimination Score Averages

Node rank	Discrimination score average	Genes names
2	1.22	GCLC
3	1.13	AKR1C3,CYP1B1,CLDN10
4	1.1	TALDO1,FTH1
5	1.06	PGD,TXNRD1
6	1.005	ME1,UCHL1,AKR1B10,PIR

2003). The clustering variable was calculated based on the differentiations between each mean of gene expression in two groups. Clustering of genes performed a viewpoint in gene expression alteration in the smokers and identified the effects of cigarette smoking on the airway genes expression. In the provided GSOM map the node with the highest average of discriminating score included seven genes (discriminating genes): NQO1, ALDH3A1, H19, AKR1C1, ABHD2, GPX2, ADH7.

Table 3. Genes Weighted Votes for Each Sample and Man-whitney Test Result of Comparison of Each Weighted Votes in the Smoker Group and the Non-smoker Group

Group		Genes weighted votes						
	Sample number	NQO1	H19	ALDH3A1	AKR1C1	ABHD2	GPX2	ADH7
Non-smokers	1	0.001725	0.009342	0.062896	0.014509	0.001155	0.006143	0.006636
	2	0.004996	0.002495	0.041414	0.030475	0.000244	0.004265	0.008748
	3	0.003759	0.008468	0.052937	0.024697	0.000896	0.001558	0.007351
	4	0.005199	0.006869	0.055985	0.031604	0.00034	0.003767	0.013298
	5	0.003898	0.005932	0.042807	0.023688	0.000628	0.004968	0.00588
	6	0.006025	0.008509	0.0066749	0.022703	0.000284	0.007944	0.011765
	7	0.00316	0.005166	0.060754	0.028277	0.000135	0.006901	0.003317
	8	0.005988	0.007146	0.059739	0.025897	0.000721	0.006806	0.012127
	9	0.005358	0.004795	0.057966	0.029374	0.000621	0.007478	0.012461
Smokers	10	-0.00741	-0.01504	0.004688	-0.00722	-0.00112	-0.00811	-0.01104
	11	-0.0029	-0.00486	-0.0391	0.001066	0.0000172	0.00376	-0.00855
	12	-0.00247	-0.00723	-0.15148	-0.02107	-0.00076	-0.00581	0.001895
	13	-0.00579	-0.00235	-0.22414	-0.05396	-0.00058	-0.015	0.002704
	14	-0.00246	-0.00415	-0.08474	-0.07544	0.000291	-0.00648	-0.01632
	15	-0.01845	-0.00624	-0.06568	-0.06299	-0.00101	-0.01597	-0.02396
	16	-0.00207	-0.01065	-0.05895	-0.01875	-0.00032	-0.0044	-0.00973
	17	-0.00481	-0.0106	-0.01573	-0.00229	-0.00103	-0.00457	-0.00913
	18	0.001081	-0.00148	-0.03298	-0.01292	-0.00048	-0.00354	0.002519
	19	-0.00235	-0.01375	-0.06812	-0.05943	-0.00015	-0.01041	-0.01161
	20	-0.00276	-0.00053	0.024058	-0.01513	-0.00068	0.001032	-0.01674
	21	-0.00833	-0.01346	-0.02372	-0.01759	-0.00131	-0.00634	-0.01098
	22	0.000801	0.005528	0.01188	0.01173	-0.00028	0.003857	-0.0069
Man-whitney test	p-value	0	0	0	0	0	0	0

Table 4. Results of Discriminate Analysis

Original group	Non-smokers	Smokers	Total
Non-smokers	9 (100%)	0 (0%)	9
Smokers	0 (100%)	13 (100%)	13

of the samples correctly, in order to their actual groups (Table 4).

Discussion

There are lots of evidences proving that smokers with a particular mutation have a dramatically higher risk to develop lung cancer. Therefore the comparison of the smokers' genes expression with the non-smokers' could be helpful in order to discover the effect of smoking on airway gene expression. In this study we used microarray data of large epithelium lung cells to compare normal non-smokers with normal smokers. The aim of the study was finding the genes which could discriminate the smoker samples from the non-smoker samples. Finding these genes could be useful to study the effect of cigarette smoking, on large airway epithelium cells. We applied a neural network clustering approach; GSOM (Hsu et al.,

NQO1, NAD(P)H: quinine oxidoreductase, is a detoxification enzyme that protects against the regeneration of reactive oxygen species chemically induced by oxidative stress, cytotoxicity, mutagenicity, and carcinogenicity (Joseph et al., 1998; Kiyohara et al., 2005; Saldivar et al., 2005; Kolesar et al., 2011). There are evidences suggest that tobacco smoking demonstrates a strong increase in expression of NQO1 (Cote et al., 2009; Pickett et al., 2009; Boyle et al., 2010; Timofeeva et al., 2010)

AlDH3A1 which is from Aldehye dehydrogenases was among the first seven discriminating genes. Aldehye dehydrogenases activity is a functional marker for lung cancer (Ucar et al., 2008; Sullivan et al., 2010; Muzio et al., 2011). ALDH3A1 are up-regulated by smoking (Beane et al., 2007; Petal et al., 2007). In fact Aldehye dehydrogenases, such as AlDH3A1, are involved in the oxidation of toxic aldehydes produced from oxidative stress and exposure to tobacco smoke (Vasiliou et al., 2005; Muzio et al., 2011).

The other identified discriminating gene was AKR1C1. This gene comes from the aldo-keto reductase (AKRs) superfamily. The AKR1 family contains many of the human isoforms, which include AKR1A, AKR1B,

AKR1C and AKR1D (Penning, 2005; Jin et al., 2007). In our study, AKR1C1, AKR1C3 and AKR1B10 were identified in either the first ranked node or in its vicinities. AKR1C1, AKR1C3 and AKR1B10 are up-regulated by cigarette smoking (Spira et al., 2004; Woenckhaus et al., 2006; Beane et al., 2007). AkR1B10 is a diagnostic marker of non-small cell lung carcinoma in smokers (Penning, 2005; Miller et al., 2012).

GPX2, glutathione peroxidase 2, is another discriminating gene. This gene is involved in the xenobiotics metabolism and there are some evidences that confirm its inducement by exposing to cigarette smoke (Spira et al., 2004; Woenckhaus et al., 2006; Brigelius-Flohe et al., 2012).

The other identified discriminating genes was H19, imprinted maternally expressed transcript (Matouk et al., 2007; 2010). There are some studies that show the up-regulation of H19 in respiratory epithelia exposed to cigarette smoking (Kaplan et al., 2003; Liu et al., 2010). Some studies suggest that overexpression and eventual loss of imprinting of H19 may represent early markers in the progression of airway epithelium toward lung cancer (Kaplan et al., 2003).

The two last identified discriminating gene were ABHD2 and ADH7. ABHD2, Abhydrolase domain-containing protein 2, encodes a protein containing an alpha/beta hydrolase fold, which is a catalytic domain found in a very wide range of enzymes. The function of this protein has not been determined. Alternative splicing of this gene results in two transcript variants encoding the same protein (Entrez gene, available in: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=11057). ADH7, alcohol dehydrogenose. This gene encodes class IV alcohol dehydrogenase 7 mu or sigma subunit, a member of the alcohol dehydrogenase family. This family metabolize a variety of substrates, including ethanol, retinol, other aliphatic alcohols, hydroxysteroids, and lipid peroxidation products (Seitz et al., 2007). Up to the time of writing this article, there were not enough studies which confirmed the relationship between these two genes alterations and smoking, so further studies are recommended to find the effect of cigarette smoking on the expression these gene.

Also, in the vicinity of identified discriminating genes there were two genes which their relevancy with lung cancer is strongly confirmed by biological studies; CYP1B3 and UCHL1.

CYP1B1 comes from cytochrome P450 family 1. CYP1B1 have a significant role in the oxidation of a veriety of carcinogens. The gene is expressed in the lungs and is up-regulated in response to cigarette smoke (Spira et al., 2004; Nagaraj et al., 2005; Wenzlaff et al., 2005; Beane et al., 2007,Xu et al., 2012).

UCHL1, ubiquitin carboxyl-terminal hydrolase L1, is used as a marker of the lung cancer (Hibi et al., 1999) and it is up-regulated in the large and the small airway epithelium of cigarette smokers, including normal smokers with early chronic obstructive lung disease (Brendan et al., 2006; Orr et al., 2011; Hurst-Kennedy et al., 2012).

Most of the identified genes agree with other similar recent studies results. For example the study of Boyle et al. (2010) comparing the oral mucosa and airway epithelium transcriptome of smokers versus non-smokers showed the overexpression of CYP1A1, CYP1B1, AKRs,ALDH2A1,NQO1 and UGTs. Pickett et al. (2009) investigated the effects of cigarette smoking condensated on airway epithelium cells. Their findings demonstrated a strong increase in expression of genes that coded for xenobiotic and detoxifying functions such as CYP1A1 and CYP1B1 and antioxidants such as GPX2 and NQO1. The results of Beane s, et al.,study indicated that many of the rapidly reversible genes such as CYP1A1, CYP1B1,AKR1B10, AKR1C1 and ALDH3A1 are up-regulated by smoking and involved in a protective or adaptive response to tobacco exposure and the detoxification of tobacco smoke components.

Most of these articles had applied common multivariate clustering methods such as hierarchical clustering which has several drawbacks such as being time-consuming and lack of robustness when there is strong presence of noise in data. But by applying GSOMs clustering according to the difference of means expression of genes, we were able to compare 7129 genes of smokers with nonsmokers in just a few minutes. Also, the map of GSOM provided a visual viewpoint to find genes discriminating the smokers from the non-smokers and could suggest further studies about co- expression of genes which were placed in the same node. In addition, it was possible to evaluate the strength of identified discriminating genes which were supposed to distinguish two groups in a systematic way by calculating weighted votes.

In our study the majority of genes in the first ranked node and its vicinity have been always interesting in lung cancer studies such as gene NQO1 (Eom et al., 2009; Timofeeva et al., 2010; Guo et al., 2012; Liu et al., 2012) and even some of them such as ALDH3A1, AKR1B10,UCHL1 are known as marker for lung cancer (Penning et al., 2005; Petal et al., 2007; Ucer et al., 2008). The identified genes except ADH7 and ABHD2 had strong relevancy to lung cancer and were supported by existing literatures but we did not do any laboratory study to investigate the correlation between ADH7 and ABHD2 or other genes which were placed in other high average nodes with smoking and also the lung cancer. Therefore a large sample experimental study is needed to study the altered expression of the genes in smokers comparing non-smokers.

Acknowledgements

We thank Arthur L Hsu of University of Melbourne, for valuable help in providing the GSOM Pak and details of preprocessing used.

References

Alahakoon D, Halgamuge SK, Srinivasan B (2000). Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE*, **11**, 601-14.

Amit D, Hochberg A (2012). Development of targeted therapy for a broad spectrum of cancers (pancreatic cancer, ovarian cancer, glioblastoma and HCC) mediated by a double

promoter plasmid expressing diphtheria toxin under the control of H19 and IGF2-P4 regulatory sequences. *Int J Clin Exp Med*, **5**, 296-305.

Beane J, Liu G, Sebastiani P, et al (2007). Reversible and permanent effect of tobacco smoke exposure on airway epithelium gene expression. *Genome Bioogy*, **8**, 201.

Brendan JC, Harvey B, Heguy A, et al (2006). Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human Airway epithelium of cigarette smokers. *Cancer Res*, **66**, 10729-39.

Brigelius-flohe R, Müller M, Lippmann D, Kipp AP (2012). The yin and yang of Nrf2-regulated selenoproteins in carcinogenesis. *IJCB*, **10**, 1155.

Boyle J, Gumus Z, Kacker A, et al (2010).Transcriptome effects of cigarette smoke on the human oral mucosal. *Cancer Prev Res*, **3**, 264-78.

Chao CH, Berthiller J, Zhang Z, et al (2006). polymorphism and the risk of Lung, bladder, and colorectal NAD(P)H:quinone oxidoreductase 1 (NQO1) pro187Ser polymorphism and the risk of lung, bladder, and colorectalcancers: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*, **15**, 978-87.

Cote M, Yoo W, Wenzlaff A, et al (2009). Tobacco and estrogen metabolic polymorphisms and risk of non-small cell lung cancer in women. *Carcinogenesis*, **30**, 626-35.

Eom S, Kim S, Zhang Y, et al (2009). Influence of NQO1,ALDH2, and CYP2E1 genetic polymorphisms, smoking, and alcohol drinking on the risk of lung cancer in Koreans. *Cancer Causes Control*, **20**, 137-45.

Franklin WA, Gazdar AF, Haney J, et al (1997). Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J Clin Invest*, **100**, 2133-7.

Golub T, Slonim D, Tamayo P, et al (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-7.

Getz G, Domany E, Levine E (2000). Coupled two-way clustering analysis of gene microarray data. *PNAS*, **97**, 12079-84.

Gebel S, Bosio A, Gerstmayer B, et al (2004). Gene expression profiling in respiratory tissues from rats exposed to mainstream cigarette smoke. *Carcinogenesis*, **25**, 169-78.

Guo M, Hooker C, House M, et al (2004). Promoter hypermethylation of resected bronchial margins:a field defect of changes? *Clin Cancer Res*, **10**, 5131-6.

Guan P, He M, Huang D, et al (2009). Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. *J Exp Clin Cancer Res*, **28**, 103.

Guo S, Gao M, Li X, et al (2012). Lack of association between NADPH quinone oxidoreductase 1 (NQO1)gene C609T polymorphism and lung cancer: a case-control study and a meta-analysis. *PLoS One*, **7**, 1371.

Hallberg J, Dominicus A, Eriksson U, et al (2008). Interaction between smoking and genetic factors in the development of chronic bronchitis. *Am J Respir Crit Care Med*, **77**, 486-90.

Hibi K, Borges M, Westra WH, et al (1999). PGP9.5 as a candidate tumor marker for non-small-cell lung cancer. *Am J Pathol*, **155**, 711-5.

Hsu A, Tang S, Halgamuge S (2003). An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data. *Bioinformatics*, **19**, 2131-40.

Hsu A (2003). Notes and usage of JAVA software: GSOMPak. Available at http://www.mame.mu.oz.au.

Hsu A, Halgamuge S (2003). Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualization. *Int J Approximate Reasoning*, **23**, 259-79.

Hurst-Keneddy J, Chin L, Li L (2012). Ubiquitin C-Terminal hydrolase L1 in tumorigenesis. *Biochem Res Int*, **10**, 1155.

Joseph P, Jaiswal AK (1998). NAD(P)H: quinine oxidoreductase 1 reduces the mutagenecity of DNA caused by NADPH: P450 reductase activated metabolities of benzo[a]pyrene quinines. *Br J Cancer*, **77**, 709-19.

Kim B, Lee H, Choi H, et al (2007). Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Res*, **67**, 7431-8.

Jin Y, Penning TM (2007). Aldo-keto redutases and bioactivation/ detoxification. *Annu Rev Phamacol Toxicol*, **47**, 263-92.

Jemal A, Bray F, Cetnter M, et al (2011). Global cancer statistics. *Ca Cancer J Clin*, **6**, 69-90.

Kaplan R, Heguy A, Luettich K, et al (2003). Monoallelic up-regulation of the imprinted H19 gene in airway epithelium of phenotypically normal cigarette smokers. *Cancer Res*, **63**, 1475-83.

Kiyohara Ch, Takayama K, Yoshimasu K, et al (2005). NQO1, MPO, and the risk of lung cancer: a huge review. *Genetic In Med*, **7**, 463-78.

Koggalage M (2004). Unsupervised class discovery and feature selection using an improved hierarchical dynamic self-organizing map. NIPS, 3.

Kolesar JM, Dahlberg SE, Marsh S, et al (2011). The NQO1*2/*2 polymorphism is associated with poor overall survival in patients following resection of stages II and IIIa non-small cell lung cancer. *Oncol Rep*, **25**, 1765-72.

Liu F, Killian JK,Yang M, et al (2010). Epigenomic alterations and gene expression profiles in respiratory epithelia exposed to cigarette smoke condensate. *Oncogene*, **29**, 3650-64.

Liu F, Yu G, Wang G, et al (2012). An NQO1-initiated and p53-independent apoptotic pathway determines the snti-tumoreffect of tanshinone IIA against non-small cell lung cancer. *PLoS ONE*, **7**, 42138.

Matouk IJ, DeGroot N, Mezan S, et al (2007). The H19 non-coding RNA is essential for human tumor growth. *PLOS ONE*, **2**, 845.

Matouk IJ, Mezan S, Mizrahi A, et al (2010). The oncofetal H19 RNA connection: Hypoxia, p53 and cancer. *Biochim Biophsy Acta*, **1803**, 443-51.

Miller V, Lin H, Murugan P, et al (2012). Aldo-keto reductase family 1 member C3 (AKR1C3) is expressed in adenocarcinoma and squamous cell carcinoma but not small cell carcinoma. *Int J Clin Exp Pathol*, **5**, 278-89

Miyata K, Nakayam M, Mizuta S, et al (2007). Elevated mature macrophage expression of human ABHD2 gene in vulnerable plaque. *Biochem Biophys Res Commun*, **365**, 207-13

Miyazu YM, Miyazawa K, Kurimoto N (2005). Telomerase expression in noncancerous bronchial epithelia is possible market of early development of lung cancer. *Cancer Res*, **65**, 9623-27.

Muzio G, Maggiora M, Paiuzzi E, et al (2011). Aldehyde dehydrogenases and cell proliferation. *Free Radic Biol Med*, **52**, 735-46.

Nacht M, Dracheva T, Gao Y, et al (2001). Molecular characteristics of non-small cell lung cancer. *Proc Natl Acad Sci USA*, **98**, 15203-8.

Nagaraj NS, Beckers S, Mansah JK, et al (2006). Cigarette smoke condensate induces cytochromes P450 and aldo-keto reductases in oral cancer cells. *Toxicol Lett*, **165**, 182-94.

Orr K, Shi Zh, Brown W, et al (2011). Potential prognostic marker ubiquitin carboxyl-terminal hydrolase –L1 does not predict patient survival in non-small cell lung carcinoma. *J Exp Clin Cancer Res*, **30**, 79.

Powel CA, Klares S, O Cannor G, et al (1999). Loss of heterozygosity in epithelium cells obtained by bronchial brushing:clinical utility in lung cancer. *Clin Cancer Res*,

5, 2025-34.
Penning T (2005). AKR1B10: a new diagnostic marker of non-small cell lung carcinoma in smokers. *Clinical Canser Res*, **11**, 1687-90.
Patel M, Lu L, Zander D, et al (2007). ALDH1A1 and ALDH3A1 expression in lung cancers: correlation with histologic type and potential precursors. *Lung Cancer*, **59**, 340-9.
Pickett G, Seagrave J, Boggs S, et al (2010). Effects of 10 cigarette smoke condensates on primary human airway epithelium cells by comparative gene and cytokine expression studies. *Toxicol Sci*, **114**, 79-89.
Saldivar S, Wang Y, Zhao H, et al (2005). An association between a Nqo1 genetic polymorphism and risk of lung cancer. *Mutation Res*, **582**, 71-8.
Seitz HK, Becker P (2007). Alcohol metabolism and cancer risk. *Alcohol Res Health*, **30**, 38-41.
Sullivan J, Spinola M, Dodge M, et al (2010). Aldehyde dehydrogenase activity selects for lung adenocarcinoma stem cells dependent on notch signaling. *Cancer Res*, **70**, 9937-48.
Spira A, Beane J, Shah V, et al (2004). Effect of cigarette smoke on the human airway epithelium cell transcriptome. PNAS, **101**, 10143-8.
Spira A, Beane J, Shah V, et al (2007). Airway epitheilial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, **13**, 361-5.
Timofeeva M, Kropp S, Sauter W, et al (2010). Genetic polymorphisms of MPO,GSTT1,GSTM1,GSTP1,EPHX1 and NQO1 as risk factors of early-oneset lung cancer. *Int J Cancer*, **127**, 1547-61.
Ucar D, Cogle Ch, Zucali J, et al (2008). Aldehyde dehydrogenase activity as a functional marker for lung cancer. *Chemico-Biological Interactions*, **178**, 48-55.
Vasiliou V, Nebert DW (2005). Analysis and update of the human aldehyde dehydrogenase (ALDH) gene family. *Hum Genomics*, **2**, 138-43.
Wistuba I, Gazdar A, Wirmani A, et al (1997). Molecular damage in the bronchial epithelium of current and former smokers. *J Natl Cancer Inst*, **89**, 1366-73.
Wenzlaff AS, Bock CH, Cote M, et al (2005). CYP1A1 and CYP1B1 polymorphisms and risk of lung cancer among never smokers: a population-based study. *Carcinogenesis*, **26**, 2207-12.
Woenckhaus M, Klein-Hitpass L, Grepmeier U, et al (2006). Smoking and cancer –related gene expression in bronchial epithelium and non-small-cell lung cancer. *J Pathol*, **210**, 192-204.
Wang W, Chiu J, Lin Ch, et al (2007). Reversal of inflammation-associated dihydrodiol dehydrogenases (AKR1C1 and AKR1C2) overexpression and drug resistance in nonsmall cell lung cancer cells by wogonin and chrysin. *Int J Cancer*, **120**, 2019-27.
Wei Sh, Liu Z, Zhao H, et al (2010). Single nucleotide polymorphism ADH7 A92G is associated with risk of squamous cell carcinoma of the head and neck. *Cancer*, **116**, 2984-92.
Xu W, Zhou Y, Hang X, et al (2012). Current evidence on the relationship between CYP1B1 polymorphisms and lung cancer risk: a meta-analysis. *Mol Biol Rep*, **39**, 2821-9.