BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(5)

An *in silico* analytical study of lung cancer and smokers datasets from gene expression omnibus (GEO) for prediction of differentially expressed genes

Atif Noorul Hasan[1, 2], Mohammad Wakil Ahmad[3], Inamul Hasan Madar[4], B Leena Grace[5] & Tarique Noorul Hasan[2, 6 *]

[1]Dept. of Bioinformatics, Jamia Millia Islamia, New Delhi, India; [2]Division of Bioinformatics, Noor-Amna Foundation for Research and Education, Bettiah, Bihar, India; [3]Dept. of Software Engg, College of Computer Science, King Saud University, Riyadh, Saudi Arabia; [4]Dept. of Biotechnology and Bioinformatics, Bishop Heber College, Tiruchirappalli, TN, India; [5]Dept of Biotechnology, Vinayaka Missions University, Salem, TN, India; [6]R & D Center, Bharathiar University, Coimbatore-641046, TN, India; Tarique Noorul Hasan - Email: tariquenh@gmail.com; Phone: +91 9472060067 ; *Corresponding author

Received February 26, 2015; Revised March 31, 2015; Accepted April 15, 2015; Published May 28, 2015

Abstract:
Smoking is the leading cause of lung cancer development and several genes have been identified as potential biomarker for lungs cancer. Contributing to the present scientific knowledge of biomarkers for lung cancer two different data sets, i.e. GDS3257 and GDS3054 were downloaded from NCBI's GEO database and normalized by RMA and GRMA packages (Bioconductor). Diffrentially expressed genes were extracted by using and were R (3.1.2); DAVID online tool was used for gene annotation and GENE MANIA tool was used for construction of gene regulatory network. Nine smoking independent gene were found whereas average expressions of those genes were almost similar in both the datasets. Five genes among them were found to be associated with cancer subtypes. Thirty smoking specific genes were identified; among those genes eight were associated with cancer sub types. *GPR110*, *IL1RN* and *HSP90AA1* were found directly associated with lung cancer. *SEMA6A* differentially expresses in only non-smoking lung cancer samples. *FLG* is differentially expressed smoking specific gene and is related to onset of various cancer subtypes. Functional annotation and network analysis revealed that *FLG* participates in various epidermal tissue developmental processes and is co-expressed with other genes. Lung tissues are epidermal tissues and thus it suggests that alteration in *FLG* may cause lung cancer. We conclude that smoking alters expression of several genes and associated biological pathways during development of lung cancers.

Background:

Lung cancer is one of the leading causes of cancer deaths worldwide. It accounts for more than 1.3 million deaths per year **[1].** Tobacco smoking causes about 90% of lungs cancer (Landi *et al.*, 2008). Non-smoker's lungs cancer accounts for 10%-15%. The etiology behind the non-smokers lungs cancer are genetic factors **[2]**, chemical physical agents and radiation, clinical history of lung disease **[3]**, air pollution which may have second hand smoke. Lung cancer is the uncontrolled growth of altered cells which may begin in one or both lungs. Most often the cells that line the air passage are affected. small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) are the primary type of lungs cancer. The secondary lung cancer originates somewhere else and metastasizes to lung. Reported genes which are associated with lung cancer are *TP53* **[4]**, *EGFR* **[5]**, *KRAS* **[6]**, *PIK3CA* **[7]**, *EML4-ALK* **[8]**, *PTEN*, *FHIT*, *MYO18B* **[9, 10, 11]**, *SEMA5A* and *SEMA6A* **[12].**

Several researches have been done in order to understand the mechanism behind uncontrolled cell growth in lung tissue. A high mutation rate was observed in factors such as p53, p16, Rb, and EGFR **[13].** p21 has significant role in controlling G_1/S transition and if altered may cause lung cancer **[14].** Many genes such as *myc*, *her2* and *neu* are involved in the cell growth and differentiation, are found mutated in lung cancer cells **[15].** The involvement of cyclin D1 in the development and progression of NSCLCs was well demonstrated **[16].** ARF proteins encoded by INK4A oncogene undergoes alteration due to gene hyper methylation may cause NSCLC **[17].** Higher expression of endothelial growth factor VEGF is inversely related to lung cancer survivals **[18].**

It is clear from the scientific studies that any cancer is caused by genetic alteration **[19]** and researchers have identified several marker genes and potential drug targets for different types of cancer **[20].** Cancer is considered complex and heterogeneous in nature and about 5% - 10% of our total pool of genes play significant role in oncogenesis **[21].**

Bioinformatics provides important tool and platforms for performing *in-silico* studies. Microarray is one such *in silico* approach which examines tens of thousands of genes together at a time and measures their expression levels. The best usage of microarray is to compare the expression level of genes from cell maintained in a particular condition to a same set of genes from a reference cell maintained under normal condition.

High mortality rate caused due to lung cancer can be reduced by early stage diagnosis and effective therapeutic methods. Methods for lung cancer diagnosis vary from person to person as medical team suggests. Different kind of imaging tests such as CT, PET and bone scan and direct tests such as bronchoscopy, endobronchial ultrasound, fine needle biopsy, sputum cytology, endoscopic esophageal ultrasound and thoracentesis are prescribed.

Apart from the conventional diagnostic methods, biomarkers may serve as an easy diagnostic method for cancer diagnosis and may add to the development of new molecular targets for drug development. Microarray data analysis has a great role in disease biomarker identification **[22, 23].** As data analysis is cheaper and less time consuming than wet lab procedures, this sets our goal to identify new biomarkers for lungs cancer.

The earlier researchers has successfully identified single or two three genes involved in lungs cancer **[24, 25, 26].** Contributing additionally to the present scientific knowledge this study was undertaken with two objectives. The first was to find out, if there is any differentially expressed gene (DEG) in smoker and non-smokers normal tissue samples which is common in smoker and non-smoker tumor tissue sample. The common DEG could be regarded as the biomarker for potential development of lung tumor in smokers. The second objective was to find out the relation of smoking with lung cancer development. To achieve the goal, genome wide analysis of three different types of expression data was conducted. In order to identify biomarkers for lung cancer caused due to smoking specifically, only those data sets were selected which were related to smoking.

Methodology:

Data and Data Source

All the gene expression datasets were downloaded from NCBI's GEO Database (http:/ /www.ncbi.nlm.nih.gov /sites/GDSbrowser). The first data set GDS3837 has samples taken from nonsmoking females. It has a total of 120 samples out of which 60 were normal and 60 were tumor. The second data set GDS3257 has samples taken from former, current and never smokers. It has a total of 107 samples out of which 58 were tumor and 49 were normal. The third data set GDS3054 has the samples taken from the buccal epithelia of smokers and non-smokers. Out of 10 samples 5 samples each were of smokers and non-smokers. The data sets were downloaded in .CEL format and were analyzed on R (3.1.2).

Data Normalization

Normalization of expression data helps in adjusting individual hybridization intensities and balances data appropriately in such a way that a meaningful biological comparisons could be made. Moreover it brings data closer and makes data less scattered. Data normalization was done using bioconductor's packages such as RMA and GCRMA.

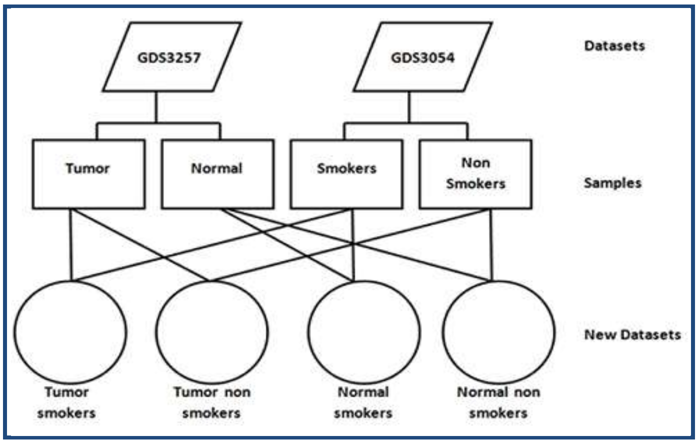

Figure 1: Schema for the generation of new data sets from GDS3257 and GDS3054. Sample subtypes from both the data sets were separated and shuffled and four new data sets were made, i.e. tumor smokers, tumor non-smokers, normal smokers and normal non-smokers.

Differentially expressed genes (DEGs) extraction

The procedure was done twice by fixing the p-value to 0.05. The top 50 DEGs were used for further analysis. Firstly DEGs of datasets GDS3837 and GDS3257 were extracted and common DEGs among both dataset were selected out. For the second DEG extraction datasets GDS3257 and GDS3054 were used. Their samples were shuffled to make four new datasets and then DEGs from each datasets were extracted. The new datasets were *tumor smokers, tumor non-smokers, normal smokers* and *normal non-smokers* **(Figure 1).**

Extraction of smoking specific DEGs

The newly formed four datasets were further analyzed comparatively to extract DEGs specific to smoking. Top 50 DEGs were selected from each dataset and p-value was 0.05. DEGs which were present in *tumor smokers* and were not present in *tumor non-smokers* (set 1) were selected out. In the same way the DEGs which were present in *normal smokers* and were not present in *normal non-smokers* (set 2) were selected out.

Gene annotation and GRN construction

David online gene annotation tool was used for DEGs annotation in order to check their relation with different types of diseases and the pathways in which they are enriched. GENE MANIA a Cytoscape plugin tool was used for GRN (Gene Regulatory Network) construction. GRN was constructed for analyzing the co-expression association of smoking specific genes.

Results:

Differentially expressed genes (DEGs) extraction and Gene Annotation

Nine DEGs were found common in both types of datasets, i.e. tumor and normal samples from non-smokers (GDS3837) and tumor and normal samples from smokers and non-smokers (GDS3257). Those DEGs are *AGER*, *CA4*, *EDNRB*, *FAM107A*, *GPM6A*, *NPR1*, *PECAM1*, *RASIP1*, and *TGFBR3*. These DEGs are independent of smoking, as these are common in both datasets (or in any tumor samples, i.e. smokers or non-smokers). Average expressions of these DEGs in both the datasets are very close.

These DEGs were then subjected to annotation in DAVID online tool. Gene *AGER*, *CA4*, *EDNRB*, *PECAM1* and *TGFBR3* were found to be associated with NSCLC, Colon, Cancer, Cancer and Prostate cancer respectively. Table1 shows average expression of these DEGs in both the data sets and their cancer sub type association.

A total of sixteen and fourteen smoking specific DEGs were found in set1 and set2 respectively, out of which three genes, *ANXA9*, *CLCA4* and *GPR110* were found common in both datasets **Table 2 (see supplementary material).** Gene annotation and literature survey of all the DEGs revealed that CLCA4, *ANXA9, GPR110, IL1RN, KLK12, PPARD, FLG* and *HSP90AA1* including the above mentioned three genes were associated with cancer sub types **Table 3 (see supplementary material) [27-45].** Among These eight genes, *GPR110, IL1RN* and *HSP90AA1*were associated with lung cancer and six genes were found to be enriched in several growth promoting pathways such as Akt signaling pathway, Ubiquitin proteasome pathway, Wnt signaling pathway, etc.

GRN construction of all the genes from set1 and set2 revealed that only *FLG* is associated with functions like epidermal cell differentiation, epidermis development, epithelial cell differentiation, skin development and keratinocyte differentiation **(Figure 2) Table 4 (see supplementary material).** False Discovery Rate corresponding to each function varies from 1.44E-07 to 5.71E-05. Moreover *FLG* was found to be co-expressed with query genes such as *KLK13*, *GYS2*, *KLK12*, *IL1RN*, *TMPRSS11D*, *ZNF185*, *CLCA4* and *HTN3*.

Discussion:

Biomarker identification of any disease helps in the development of better diagnostics and improves the clinical treatment efficacy and microarray data analysis has been widely used for disease biomarker discovery **[46, 47, 48].** With the aim to discover biomarkers for lung cancer independent of smoking, we identify nine Differentially Expressed Genes (DEGs), *AGER, CA4, EDNRB, FAM107A, GPM6A, NPR1, PECAM1, RASIP1* and *TGFBR3*. These genes were found common in both lung cancer datasets and their average expression was very similar in either type of dataset **Table 1 (see supplementary material)**, that is, lung cancer dataset of non-smokers (GDS3837) and lung cancer dataset of smokers (GDS3257). Gene annotation of these DEGs revealed that genes like *AGER*, *CA4* and *TGFBR3* were involved in *NSCLC*, colon cancer and prostate cancer respectively Whereas *EDNRB* was found to be associated with several human cancers such as prostate, colorectal and oral. *PECAM1* was found to be associated with lung cancer, breast cancer, colorectal cancer ovary and prostate cancer. These genes were not reported in the researches from where both dataset were taken for study. Dataset GDS3837 corresponds to the research of Lu et al (2010) **[12].** They identified several semaphorin gene family members such as *SEMA5A*, SEMA6A, *SEMA3B and SEMA3F* to be potentially associated with lung cancer among non-smokers. Our analysis to the same dataset correlates with previous findings that semaphorin family gene are associated with lung cancer among non smoker. *SEMA6A* was found to be up regulated with an average expression of 7.061 and a log fold change of 2.280.

Dataset GDS3257 corresponds to the research of Landi *et al* (2008) **[49].** They successfully identified candidate target genes for lungs cancer among smokers. Even they did not found any semaphorin family genes associated with lung cancer among smokers. This clearly suggests that semaphorin family genes are only associated with lung cancer independent of smoking.

Our further research was aimed to explore the correlation of smoking with gene alteration. In the way to achieve our goal we selected out DEGs specific to smoking, with the help of new datasets which was formed by sample shuffling of datasets GDS3257 and GDS3054 **(Figure 1).** A total of sixteen DEGs were extracted and categorized in two sets, i.e. set1 and set2 Gene annotation revealed that among those sixteen DEGs, *CLCA4*, *ANXA9, GPR110, IL1RN, KLK12, PPARD, FLG* and *HSP90AA1* were involved in cancer subtypes such as lungs, breast, colorectal, prostate, vulva, urothelial, oral, gastric, vaginal and different epithelial cancers **Table 3 (see supplementary material).** *GPR110*, *IL1RN* and *HSP90AA1*, were involved in lungs cancer.

Lung tissues are epithelial tissue and cancer in lung tissue are the result of uncontrolled cellular growth in it. Lung cancer develops in multistep process involving several genetic and epigenetic alterations. Alteration includes activation of growth promoting pathways and inhibition of tumour suppressor pathways. These pathways regulate different cellular cycle events such cell proliferation arrest, DNA unwinding, phosphorylation, DNA replication, transcription factor expression, protein ubiquitination, mitotic spindle assembly, nuclear envelop break down, etc. Pathway analysis of tumor specific DEGs revealed that six genes, *CLCA4*, *GYS2*, *HSP90AA1*, *ITCH*, *PPARD* and *IL1RN*, were enriched in several growth promoting pathways. *HSP90AA1* is enriched in Akt signaling pathway, Ahr sinaling transduction and NOD-like receptor signaling pathway. *ITCH* is enriched in ubiquitin proteasome pathway. *PPARD* is enriched in wnt signaling pathway.

The GRN analysis for all smoking specific genes was carried to investigate gene function association and gene co-expression. Only *FLG* was found to be involved in fuctions such as epidermal cell differentiation, epidermis development, epithelial cell differentiation, skin development and keratinocyte differentiation **Table 4 (see supplementary material).** Several nonquery genes (non smoking specific) such as *SPRR2B*, *SCEL*, *SPRR3*, *SPRR1A*, *TGM3*, *TGM1*, *SPINK5*, *KRT16* and *RHCG* were co-expressed with *FLG* **(Figure 2)**. Earlier studies have suggested the fact that, a single gene is not responsible for the tumor development, rather group of genes expresses together for the development of such clinical condition **[50, 51].** In the favor of this fact we found that alone *FLG* among the query genes is involved in epidermal cell regulation, moreover *FLG* was found to be co-expressed with several query genes such as *KLK13*, *GYS2*, *KLK12*, *IL1RN*, *TMPRSS11D*, *ZNF185*, *CLCA4* and *HTN3*.

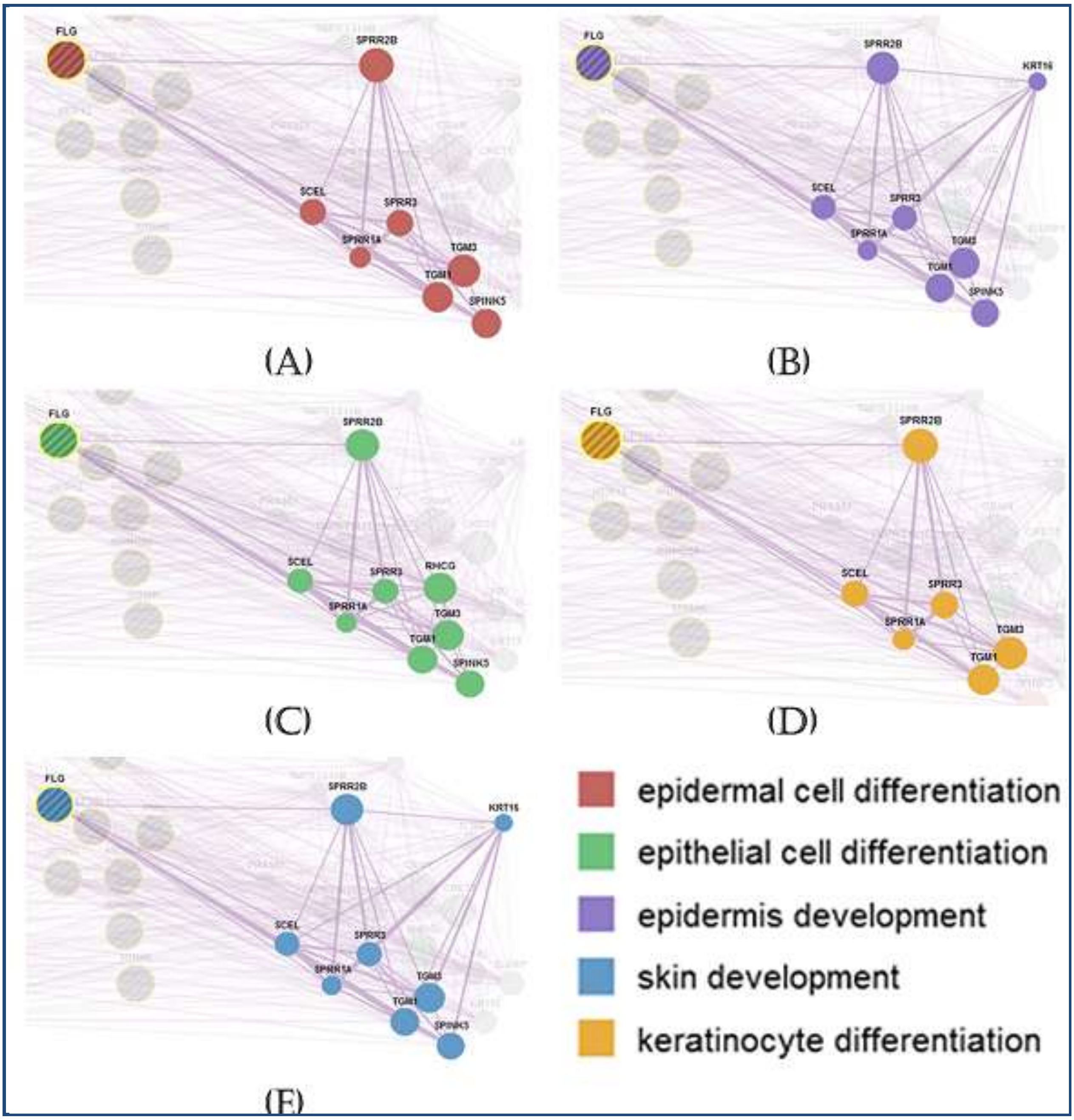

Figure 2: Gene FLG gene regulatory networks (GRN). FLG GRN for **(A)** epidermal cell differentiation, **(B)** epidermis development, **(C)** epithelial cell differentiation, **(D)** keratinocyte differentiation, **(E)** skin development.

Conclusion:

AGER, *CA4*, *EDNRB*, *FAM107A*, *GPM6A*, *NPR1*, *PECAM1*, *RASIP1*, *TGFBR3* are the DEGs which are smoking independent whereas *AGER*, *CA4*, *EDNRB*, *PECAM1* and *TGFBR3* are associated with various cancer subtypes. Semaphoring family gene *SEMA6A* is differentially expressed only in non-smoking lung cancer samples. *CLCA4, ANXA9, GPR110, IL1RN, KLK12, PPARD, FLG* and *HSP90AAl* are smoking dependent DEGs whereas *GPR110, IL1RN* and *HSP90AAl* are associated with lung cancer *FLG* is co-expressed with several query genes which are enriched in different cellular growth promoting pathways and *FLG* together with other genes plays key role in epidermal cell regulation. These query genes were found to be smoking specific. This suggests that smoking alters several genes and associated pathways. This study needs to get validated *in vitro* so that effective biomarkers could be identified and effective targeting of affected biological pathway could be achieved for treatment and prognosis of lung cancer.

References:

[1] Ferlay J *et al. International Journal of Cancer* 2010 **127:** 2893 [PMID: 21351269]
[2] Bryant A & Cerfolio RJ, *Chest* 2007 **132:** 185 [PMID: 17573517]
[3] Wu H *Cancer research* 1988, **48**: 7279 [PMID: 3191498]
[4] Hainaut P & Pfeifer GP *Carcinogenesis* 2001 **22:**367 [PMID: 11238174]
[5] Martin P *et al. Cancer Control* 2006 **13:**129 [PMID: 16735987]
[6] Herbst RS *et al. Journal of clinical oncology* 2005 **23:** 5900 [PMID: 16043828]
[7] Yamamoto H *et al. Cancer research* 2008 **68:** 6913 [PMID: 18757405]
[8] Wong DW *et al. Cancer* 2009 **115:** 1723 [PMID: 19170230]
[9] Minna JD *et al. Cancer cell* 2002 **1:** 49 [PMID: 12086887]
[10] Sekido Y *et al. Annual Review of Medicine* 2003 **54:** 73 [PMID: 12471176]
[11] Yokota J & Kohno T *Cancer Science* 2004 **95:**197 [PMID: 15016317]
[12] Lu TP *et al. Cancer Epidemiology Biomarkers & Prevention* 2010 **19:** 2590 [PMID: 20802022]
[13] Yanaihara N *et al. Cancer cell* 2006 **9:** 189 [PMID: 16530703]
[14] Niklinski J *et al. Lung Cancer* 2001 **34:** S53 [PMID: 11720742]
[15] Jacobson DR *et al. Annals of oncology* 1995 **6(suppl 3):** S3 [PMID: 8616111]
[16] Mishina T *et al. British journal of cancer* 1999 **80:** 1289 [PMID: 10376986]
[17] Zhao Z *et al. Chinese medical journal* 2008 **121:** 445 [PMID: 18364119]
[18] Volm M & Koom ÃR *British journal of cancer* 1997 **75:** 1774 [PMID: 9192980]
[19] Vogelstein B & Kinzler KW *Nature medicine* 2004 **10:** 789 [PMID: 15286780]
[20] Futreal PA *et al. Nature Reviews Cancer* 2004, **4:**177 [PMID: 14993899]
[21] Huang SM & Harari PM *Investigational new drugs* 1999 **17:**259 [PMID: 10665478]
[22] Miller JC *et al. Proteomics* 2003 **3:** 56 [PMID: 12548634]
[23] Nagata M *et al. International journal of cancer* 2003 **106:** 683 [PMID: 12866027]
[24] Lu TP *et al. Cancer Epidemiology Biomarkers & Prevention* 2010 **19:** 2590 [PMID: 20802022]
[25] Belinsky SA *et al. Proceedings of the National Academy of Sciences* 1998 **95:** 11891 [PMID: 9751761]
[26] Lynch TJ *et al. New England Journal of Medicine* 2004 **350:** 2129 [PMID: 15118073]
[27] Yu Y *et al. PloS one* 2013 **8:** e83943 [PMID: 24386311]
[28] Bundela S *et al. PloS one* 2014 **9:** e102610 [PMID: 25029526]
[29] Yang B *et al. PloS one* 2013 **8:** e60861 [PMID: 23593331].
[30] Hansel DE *et al. BMC medical genomics* 2013 **6:** 42 [PMID: 24134934]
[31] Miyoshi N *et al. Oncology letters* 2014 **8:** 2313 [PMID: 25289111]
[32] Hu Z *et al. U.S. Patent* #8,198,254 B2 (2012)
[33] Lum AM *et al. BMC cancer* 2010 **10:** 40 [PMID: 20149256]
[34] Trivedi MV *et al. Cancer Research* 2013, **73(24 Supplement):** P6–04-05 [PMID: NA]
[35] Hu Z *et al. Cancer letters* 2006 **236:** 269 [PMID: 16019127]
[36] Konwar R *et al. Oncology Research* 2009 **17:** 367 [PMID: 19544973]
[37] Cheng I *et al. Cancer Epidemiology Biomarkers & Prevention* 2007 **16:**1309 [PMID: 17548705]
[38] Grimm C *et al. Gynecologic oncology* 2004 **92:** 936 [PMID: 14984963]
[39] Xue H *et al. Journal of gastroenterology and hepatology* 2010 **25:**1604 [PMID: 20880168]
[40] Burada F *et al. Molecular biology reports* 2013 **40:** 2851 [PMID: 23192617]
[41] Zhao EH *et al. World Journal of Gastroenterology* 2012 **18:** 6597 [PMID: 23236234]
[42] Ticha I *et al. PloS one* 2013 **8:** e83952 [PMID: 24391853]
[43] Skaaby T *et al. PloS one* 2014 **9:** e99437 [PMID: 24905740]
[44] Adnane J *et al. Oncogene* 1991 **6:** 659 [PMID: 1851551]
[45] Coskunpinar E *et al. Anticancer research* 2014 **34:** 753 [PMID: 24511009]
[46] Cooper CS *et al. Nature Clinical Practice Urology* 2007 **4:** 677 [PMID: 18059348]
[47] Scherzer CR *et al. Proceedings of the National Academy of Sciences* 2007 **104:** 955 [PMID: 17215369]
[48] Landi MT *et al. PloS one* 2008 **3:** e1651 [PMID: 18297132]
[49] Allam A & Gumpeny RS *International journal of Alzheimer's disease* 2012 **2012:** 649456 [PMID: 22482075]
[50] Alon U *et al. Proceedings of the National Academy of Sciences* 1999 **96:** 6745 [PMID: 10359783]
[51] Eisen MB *et al. Proceedings of the National Academy of Sciences* 1998 **95:** 14863 [PMID: 9843981]

Edited by P Kangueane
Citation: **Hasan *et al.*** Bioinformation 11(5): 229-235 (2015)

Supplementary material:

Table 1: Smoking independent DEGs: Common smoking independent DEGs from datasets GDS3837 and GDS3257 and their average expression (Avg Exp) in both datasets. Associated cancer subtypes was obtained with the help of David online functional annotation tool

Common DEGs	Avg Exp GDS3837	Avg Exp GDS3257	Associated Cancer sub type
AGER	10.13694	9.35774	NSCLC
CA4	8.18887	8.021088	Colon
EDNRB	9.329481	8.724439	Cancer
FAM107A	9.263178	9.013393	NA
GPM6A	7.713325	7.252757	NA
NPR1	7.739213	7.480749	NA
PECAM1	7.966476	10.12169	Cancer
RASIP1	8.783135	8.942669	NA
TGFBR3	9.573514	8.323888	Prostate

Table 2: Smoking dependent DEGs: Smoking dependent DEGs and their corresponding log fold change (Log FC)

SET 1		SET 2	
Gene	Log FC	Gene	Log FC
CLCA4	-5.625	MXD1	-4.813
S100A12	-4.787	FLG	-6.004
ANXA9	-4.039	KLK12	-4.303
IL1RN	-3.179	ANXA9	-4.384
C18orf25	-2.851	TMPRSS11D	-4.833
HTN3	-2.767	CLCA4	-5.678
HSP90AA1	4.744	TTC9	-3.847
GYS2	-2.893	ARNTL2	-4.437
RMND5A	-3.852	GPR110	-3.087
EPCAM	5.410	AIM1L	-3.843
PPARD	-1.294	EPS8L1	-3.859
ZNF365	-3.465	PAX9	-2.668
KLK13	-4.565	ZNF185	-3.103
AHNAK	-3.405	ITCH	-2.869
IL36RN	-3.584		
SEC62	3.075		
GPR110	-2.998		

Table 3: Common DEGs of set1 and set2: Common DEGs among set1 and set2 and associated cancer subtypes. Other associated diseases were obtained with David functional annotation online tool. These DEGs are smoking dependent.

Gene	Associated Cancer	Associated other disease
CLCA4	Breast [27], Oral [28], Colorectal [29], Urothelial [30]	cyctic fibrosis,
ANXA9	Colorectal [31], Different epithelial cancers [32]	pemphigus vulgaris
GPR110	Lung & Prostate [33] Breast [34]	NA
IL1RN	Lung [35], breast [36] Prostate [37], Vulvar [38], Gastric [39], Colorectal [40],	Alzheimer's Diseases, asthama, arthritis, diabetes

KLK12	Gastric [41]	eczema herpeticum
PPARD	Colorectal [42]	Alzheimer's Disease, atherosclerosis, diabetes
FLG	Cervix,Vagina, Vulva[43], Breast [44]	ichthyosis vulgaris
HSP90AA1	NSCLC [45]	hand-foot-genital syndrome, lobular neoplasia

Table 4: *FLG* associated functions: *FLG* and its associated functions and corresponding false discovery rate (FDR), as predicted by Gene Mania online tool. Among the common smoking specific DEGs, only *FLG* was associated with various epidermal tissue developmental processes.

Function	Associated query genes	FDR
Epidermal cell differentiation	FLG	1.44E-07
Epidermis development	FLG	2.71E-07
Epithelial cell differentiation	FLG	5.71E-05
Skin development	FLG	2.71E-07
Keratinocyte differentiation	FLG	2.71E-07