Available Online through
www.ijpbs.com (or) www.ijpbsonline.com

IJPBS |Volume 3| Issue 3 |JUL-SEP|2013|07-26

Research Article
Biological Sciences

SIGNAL TRANSDUCTION PATHWAY ANALYSIS OF LUNG CANCER THAT MAY PROVIDE POTENTIAL DRUG TARGET IDENTIFICATION

Suchithra S, Shambhu M G & Kusum Paul

Department of Biotechnology, The Oxford College of Engineering, Hosur Road, Bangalore

***Corresponding Author Email:** *suchi.maran0224@gmail.com*

ABSTRACT

Signaling pathways belong to a complex system of communication that governs cellular processes. They represent signal transduction from an extracellular stimulus through a receptor to intracellular mediators, as well as intracellular interactions. Any alteration in signaling cascade often leads to changes in cell function and cause many diseases, including cancer. Lung cancer is common complication of cancer for both smokers and non-smokers. While multiple pathways are implicated in the pathophysiology of lung cancer, there are no specific treatments and no means to predict lung cancer onset or progression. Here, we identify gene expression signatures related to lung cancer for both smokers and non smokers, Microarray experiment performed to identify the genes which involved for lung cancer in both smokers and non smokers conditions. A bioinformatics analysis identified differentially expressed genes and their networks and biological pathways potentially responsible for the progression of lung cancer. Using differential gene expression analysis; we identified unknown gene fault caused by two separate genes like BMK1/ERK1 that causing lung cancer.

KEY WORDS

Lung Cancer (LC); Mitogen activator protein kinase (MAPK); Adenocarcinomas (AC); Big mitogen activated protein kinase(BMK); Extracellular signal regulated protein kinase(ERK). Non tumor (NT).

INTRODUCTION

Lung cancer is most leading cause of cancer deaths in both men and women in most developed countries. Researchers estimated 1,618,800 new cases and 1,388,400 deaths in 2008 [1]. Adenocarcinoma (AC) is a common type of lung cancer and has increased relative to other histological types of lung cancer. Based on histological studies, there are 2 types of lung cancer (a) Non small cell lung cancer (NSCLC): it is common in 80% of cancers. (b) Small cell lung cancer (SCLC): It is present in 20% of people. Small cell lung cancer is a rapidly growing, quick spreading tumor caused primarily by smoking [2]. Other rare forms of cancer, as well as cancer that spread from other regions of the body are found in the lungs.

Based on histological studies, there are 80-90% of patients are affected from tobacco smoking and 10% of patients are non-tobacco smokers [3]. There are two major classes of carcinogens including N-nitrosamines and polycyclic aromatic hydrocarbons present in both patient samples. There is no much information on disease prevention, screening or therapeutic regimen on lung cancer. Researchers were studies on different patient samples based environmental and genetic factors [4]. In molecular studies on lung cancer there are many genes influencing in the pathogenesis of lung cancer produce proteins involved in cell growth and differentiation, apoptosis, angiogenesis, tumor progression, cell cycle processes and immune regulation.

The tobacco smoke contains various toxic chemicals whose metabolites bind to DNA and

induce activating point mutation in the p53 tumor suppressor and K-RAS gene [5]. In a comprehensive study of genetics, the gene mutations in the p53 tumor suppressor and K-RAS proto-oncogene is frequent alterations in many human tumors [6, 7]. In fact, the oncogenic mutations of the K-RAS gene can alone cause cancer and this mutation is found in 30% of human lung adenocarcinoma [8, 9]. There are large gene expression Data set that can provide novel insights into lung cancer [10]. There are specific differences between lung cancer in smokers and none have compared nonmalignant to tumor tissue in the same subject [11].

Researchers were identified mutant genes involved in lung cancer, there are various geometrical changes associated with lung adenocarcinoma. There are 26 genes that are frequently mutated in lung cancer, further increasing opportunities for individualized diagnosis and treatment of the country's leading cause of cancer deaths. Mutations in the EGFR gene of smokers and non smokers of male, as well in women and people under 45 [12]. Mutations on ROS1 gene encode a protein that is important for cell growth, cell survival, and deregulation of ROS1 through chromosomal rearrangement drives the growth of tumors [13]. In clinical studies on lung cancer mutations in ROS1 genes shows 23% causing KRAS, mutations of which account for about 25 % of cases; EGFR, accounting for 10-15%; and ALK, rearranged in about 4 % EPHA3 is 5- 10%. Altogether, known cancer causing genetic changes.

In genetic abnormalities in lung cancer it interlinks with many signaling pathways having their main functions altered, focusing on individual factors. In significant growth promoting pathways (EGFR/k-Ras/PI3K), growth inhibitory pathways (p53/Rb/P14ARF, STK11), apoptotic pathways (Bcl-2/Bax/Fas/FasL), DNA repair and immortalisation genes are mainly targeted as drug targets in lung cancer [14]. The mitogen-activated protein kinase (MAPK) gene is commonly present in lung cancer of both smokers and nonsmokers including male and female. There are significant cascaded of BMK1 pathway is the most recently discovered and least-studied mammalian mitogen-activated protein (MAP) kinase cascade [15]. In order to study the function of this MAP kinase in lung cancer patients, in order to study the BMK1/ERK1. Big MAP kinase 1 (BMK1 or ERK5) is a key mediator of endothelial cell (EC) function as shown by impaired embryonic angiogenesis and vascular collapse in BMK1 [16].

The BMK1 is a new member of mitogen-activated protein kinase (MAPK) family [17]. Mutations in MAPK signaling pathways have been shown to play a significant role in many types of cancer. There are four different MAPKs that have been identified in cancer cells, ERK1/2 and BMK1 exhibit significant structural similarity [18]. The ERK1/2 and BMK1 cascades are both activated by mitogens and by oncogenic signals and, thus, are strongly implicated in tumorogenesis [19]. In fact, recent research has shown that some pharmacological compounds which have been considered to be specific inhibitors of ERK1/2 also interfere with the lesser known BMK1 [20].

2. MATERIALS AND METHODS

SELECTION OF RAW DATA:

Data searching: We have evaluated all published case control studies and the diseased datasets were selected using various databases. The gene expression databases include GEO (Gene expression Omnibus), Array express (EBI database), SMD (Stanford microarray database) and PUMAdb (Princeton University Microarray database). In order to classify the different genes present in both smokers and non-smokers with lung cancer. We have determined the expression profile of different tumor stage adenocarcinoma and paired normal lung tissues of current, former and never smokers. The study comprised of Gene expression profiling with Affymetrix chip data were selected from a GEO database (accession number- GDS3257).

The GDS3257 contains 107 datasets of adenocarcinoma and the non-tumor tissue samples were processed for microarray analysis using Affymetrix Human Genome U133A Array

GeneChip. The samples include 107 datasets and were finally examined based on expression from 58 tumors and 49 non-tumor tissues from 20 never smokers, 28 current smokers and 26 former smokers. The Cigarette smoking effect on lung adenocarcinoma from which these probe sets were derived and was selected from dbEST, RefSeq and GenBank. The sequence clusters were created from the UniGene database and then refined by analysis and comparison with a number of other publicly available databases.

EXPERIMENTAL DESIGN

These 107 data sets were chosen because they provide comparative data on adenocarcinoma vs. normal lung tissue. The microarray data sets were transformed in such a way that all the elements of the microarray profiles represented the signal measurements.

AC/ NLT= Adenocarcinoma (AC) /Normal Lung tissue (NLT).

In the primary lung adenocarcinoma and histological non-malignant lung tissue, the dataset contains 107 sets of samples for both tobacco smokers and non smokers selected by matching the following criteria in a descending order based on cell type, histologic stage of differentiation, clinical stage and patient age. Each sample was accompanied by an adjacent section for histologic confirmation.

The 107 chosen datasets were studied based on two types.

The first type included experiments in which disease on adenocarcinoma tissue was used as the reference sample in hybridization. In this case, the signal AC/NLT was calculated as the fluorescent intensity ratio Cy5/Cy3.

The second type included experiments in which the reference sample was the Normal Lung tissue of control data sets.

The total dataset contains 44928 genes with 107 samples of both adenocarcinoma and normal lung tissue.

In order to identify the expression of carcinoma of lung tissue based on the AC/NLT signal was calculated as the ratio of the intensity ratios (Cy5/Cy3) ngn3/ (Cy5/Cy3) ESC, where (Cy5/Cy3) adenocarcinoma and (Cy5/Cy3) normal lung tissue are the intensity ratios for cancer and normal pancreatic neuronal cell development, respectively. The aim is to test expression ratio-based analysis to differentiating between normal and lung cancer.

MICROARRAY DATA PREPROCESSING:

The Affymetrix CEL files were used to calculate the absolute values of raw data and then they were normalized by natural logarithm transformation. This preprocessing was performed by using R statistical software version 2.80. The analysis of microarray quality was assessed using the probe-level modeling and quality metrics provided by the Affy package of BioConductor.

IDENTIFICATION OF BMK1/ERK1 ASSOCIATED GENE EXPRESSION:

BioConductor packages used to calculate the processed data. GCRMA (Gene-Chip Robust Multiarray Average) was used for signal normalization. Adenocarcinoma data were defined as the baseline to generate differential expression values for all hybridizations. A patient sample pair was excluded from further analysis from one of the samples did not meet the quality control. Micro-array data was subsequently normalized using the Robust Microarray Analysis (RMA) algorithm. A meticulous analysis of variance was performed to generate p values of the probe set. To evaluated transcriptome similarities among treatment groups, different expressed genes with expression of at least 2 and not greater than -2 and p values of not greater than 0.001 were selected from each group and combined for hierarchical clustering using an average clustering method, correlation similarity metric and by clustering both rows and columns.

Since the expression data is approximately log normally distributed, we used the log-transformed data as produced by the RMA algorithm for all subsequent statistical tests. For visualization purposes, we centered the log-transformed expression data by subtracting the

average probeset log-expression values. Probe sets with relatively low expression (average expression values below 100 Affymetrix units) or with nearly constant expression values (standard deviation below 50) were excluded from further consideration of the 44928 probe sets on the Affymetrix Human Genome U133A GeneChip.

There are 39,000 transcript variants, which represent greater than 33,000 of the good characterized human genes. An unpaired t-test was used to determine the probe sets (genes) that are differentially expressed between the normal and adenocarcinoma tissue samples. The robust multi-array average approach averages normalized expression levels across all probes for the gene (probe set level analysis) whereas Genomatix Chip Inspector calculates the change in each probe (probe level analysis). Genes were deemed as 'differentially expressed genes' using Cyber-T, based on a Bayesian regularized t-test, P50.05 in the robust multi-array average approach and a false discovery rate (FDR) 50.1% using Chip Inspector with a minimum of four probes per transcript. The first 800 probe sets with the lowest t-test p-values (corresponding to a p-value cutoff of 9x10-12) were selected for further analysis. We also used a more stringent fold-change constraint that excluded the probe sets with log-fold change < 1 (roughly corresponding to a fold change < 2), where the log-fold change of gene g between classes N ('normal') and T ('tumor') is defined as

Log-fc (*g*, N-T) =log2 *g* (T) – log2 *g* (N),
Log2 *g* (C) = ∑si€clog2g (Si) /|C|...Equation (1)

The equation represents the average log- value of gene 'g' in the samples Si of class C.

Also, probe set lists of BMK1/ERK1-induced differentially expressed genes were produced for each different context. The intersecting probe sets representing "context-independent" transcriptional expressed genes were systematically evaluated for significant enrichment of canonical signaling pathways.

FUNCTIONAL ENRICHMENT ANALYSES:

The Database for Annotation, Visualization and Integrated Discovery (DAVID).

(http://david.abcc.ncifcrf.gov) and Concept Gen (http://conceptgen.ncibi.org) were used to identify over-represented biological functions and pathways among the differentially expressed genes.

NETWORK ANALYSIS:

A gene co-citation network of the differentially expressed genes was generated by using a sentence level co-citation filter. This network analysis allows visualization of the differentially expressed genes and their potential associations with each other identified in the literature. The topology of the network was analyzed using the Fast Greedy community structure identification algorithm, implemented in the Cytoscape plug-in G Lay (http://brainarray.mbni.med.umich.edu/surgeon/clay) to identify coherent sub networks. Identified sub networks were subjected to functional enrichment analyses by Gene ontology tool (GO rilla) to reveal over-represented biological functions within each subnetwork.

Protein-protein interaction (PPI) networks are the assembly of the protein signal cascades that transfer the biological function and information through the pathways. Current public PPI databases provide rich information and they mostly differ in the way they acquire or validate their data. For example: MINT, BIND, HPRD, and MIPS are manually curated; this means a team of biologists Checked the literature to find new interactions and once an interaction is confirmed it is added to the database. On other hand, IntAct and DIP are based on literature mining and they achieve these using computational methods that retrieve the interaction knowledge automatically from published papers.

We have observed limited intersection and overlap between the six major databases (BioGRID, BIND, MINT, HPRD, IntAct, and DIP). The information contained in these databases is partly complementary and the knowledge of the protein interactions can be increased and improved by combining multiple databases and integrate PPI data warehouse with 6 major

databases and erase the duplicated interaction pairs using the synonym of the protein name. We map protein name and then we erase the duplicated interaction pairs and successfully gather 44928 available and non-redundant PPI pairs among 33000 proteins.

Due to the limitation of the knowledge of the directed interactions between transcription factors and genes under specific condition and we gather the associated interactions under pancreatic cancer neurological development. Who used statistical assessments to construct directly regulatory associations between transcription factors and genes based on the gene expression data and transcription factor binding site prediction toolkit. We have constructed the relationships between neurological transcription factors (MAP3K8, BRAF, PARK2, PPP2R1B, SLC22A18, KRAS, CYP2A6, EGFR, ERBB2, DLEC1, RASSF1, NEK2, TTK, and PRC1) and their regulatory networks and validated by public literature and databases. Finally, we collect 3600 directed edges and 1960 bi-direction edges as our network.

Neighborhood Scoring is a local method for prioritizing candidates based on the distribution of differentially expressed genes in the network. We adopted the method such that every network object is assigned a score, which is mainly based partly on its expression fold change and partly on the expression fold changes of its neighbors. The differential expression levels of the genes are mapped to the corresponding network objects. Next adjusted differential expression level and the score can be calculated for each network object as follows:

$$Score(i) = \frac{1}{2} \cdot FC(i) + \frac{1}{2} \cdot \frac{\sum_{n \in N(i)} FC(n)}{|N(i)|}$$ ..Equation (2)

The score of network object i equally depend on its fold change (FC) and on the fold changes of its neighbors' n, where N (i) includes all neighboring network objects of i. The Network objects are not differentially expressed and that do not have any differentially expressed genes in their direct neighborhood are assigned a score of 0.

INTERCONNECTIVITY:

Interconnectivity is a local method that prioritizes candidates based on their overall connectivity to the differentially expressed genes. First, an interconnectivity score is calculated for each pair of interacting network objects. The interconnectivity score is mainly based on both the direct interaction between a pair and the indirect interactions with a path length of 2 which define as the shared neighborhood of two network objects. We adopted the method to score interactions of differentially expressed genes based on their direct interaction and on their shared neighborhood as follows:

$$ICN(i,j) = e(i,j) \cdot \left(\frac{2 + |N(i) \cap N(j)|}{\sqrt{\deg(i) \cdot \deg(j)}} \right)$$ ..Equation (3)

e (i, j) describes an edge between the two network objects i and j and it is set to 1 if edge exists and 0 else. Besides direct interaction between i and j and size of their shared neighborhood N is taken into account and normalized by the overall degrees of the two network objects.

Next, each network object receives its final score based on the interconnectivity to all differentially expressed genes:

$$Score(i) = \frac{1}{|DEG|} \cdot \sum_{d \in DEG} ICN(i,d)$$ .....Equation (4)

Where, d represents a differently expressed gene and DEG the set of all.

IDENTIFICATION OF DRUG TARGETS:

We used R and BioConductor to identify the level of gene expression signatures and at the level of predicted drug targets. At the level of gene expression we calculated the distance matrix for the disease pairs based on the overlap between sets of differentially expressed genes using the correlation coefficient as a measure of the overlap. The same approach was used for calculating the distance matrix at the level of predicted drug targets, where the top 100 predicted drug targets for each disease were used for the calculation.

Next, the diseases were clustered using hierarchical clustering with complete linkage and we used Mantel test to assess the similarity

between the genetic signature-based distance matrices and the predicted drug target-based distance matrices. The Mantel test calculates the correlation between 2 matrices where, the p-value is a departure from 0 correlation over 1000 permutations of the rows and columns, and these results is potentially used as a drug targets.

RESULT & DISCUSSION

GENE SIGNATURES IDENTIFICATION BASED ON DIFFERENTIAL GENE EXPRESSION OF BMK1/ERK1:

The normal and tumor developed patient samples were used for Affymetrix Human Genome U133A array data set contains 44928 gene entries and more than 39000 is transcript variants, which in turn represents greater than 33000 of the best characterized human genes. The Affymetrix Data contains 58 samples of tumor [16 samples were never smokers, 18 samples were former smokers and 24 samples are current smokers] and 49 samples are normal [16 samples are Never a smoker, 18 are former smoker and 16 is current smokers] met the RNA quality criteria and were hybridized to Affymetrix gene expression microarrays. Excluding two outlets and one mislabeled sample identified during the quality assessment process were used in our analyses (Tab-1).

Based on histology of adenocarcinoma and normal lung tissue samples is collected. The datasets of normal and diseased datasets of both male and female in different stages are (i) nonsmokers (Normal and diseased: 4 stages of early stage and late stage of the tumor) (ii) former smokers (Normal and Diseased: 3 stages of early and late stage tumor tissue) (iii) current smokers (Normal and diseased: 4 stages of early and late stage tumor tissue).

QUALITY ANALYSIS OF PRE-PROCESSED DATA:

The experimental analysis of GDS3257 datasets was performed with the R statistical analysis software using customized versions of the simplified and affyQCReport packages (and depending packages) from BioConductor. We are predicting our dataset of QC to identify thresholds that can be used to qualify arrays. The unprocessed GDS3257 raw data is subject to unprocessed variation and removed unwanted sources of variability of unweighted samples. The unprocessed data being assayed should be processed using box plot. The unprocessed probe intensities across all arrays with the overall higher level of probe signal intensity. Frequently, this measure is not very sensitive and problematic arrays may look like their better quality counterparts. **(Fig: 1, Fig: 2, Fig: 3, Fig:4)**

RNA DEGRADATION PLOT:

The quality analysis of preprocessed data on RNA oligonucleotide probesets were predicted using RNA degradation plot **(Fig: 5, Fig: 6, Fig: 7, Fig: 8)**

QUALITY ANALYSIS OF PRE-PROCESSING OF RAW DATA AFTER NORMALIZATION:

We have predicted normalization of data sets with a large majority of genes will not have their relative expression levels changed from one treatment group to the next, and assumption that departures of the response from linearity are small and slowly varying and local regression is used to estimate the normalized expression levels as well as the expression level-dependent error variance **(Fig: 9, Fig: 10, Fig: 11).**

MA PLOTS:

An MA plot is analysis of variables in the analysis of dual dye arrays and understanding their meaning of the crucial concept of analysis. A is defines as

$$A=\log_2 \sqrt{Cy5.Cy3} = 1/2[log2(Cy5) + log2(Cy3)]$$ .... Equation (5)

Cy5 and Cy3 represent respectively the red and green dye intensities of a particular spot. So A is the green and red intensities of a spot multiplied together, square rooted and log transformed. It is essentially a measure of the total log transformed intensity of a spot. Essentially, if combined green and red intensities are high for a particular spot, then A will also be high.

M is defined as

$M = \log_2 \frac{Cy5}{Cy3} = \log_2 \frac{Cy5}{Cy3} = log2(Cy5) - log2(Cy3)$ ..... Equation (6)

So M is the log transformed red dye intensity divided by the green dye intensity. It gives an indication of whether the red or green dye binds more to the array at a given spot. The purpose of an MA-plot is to investigate intensity bias. If a disproportionate amount of spots on the plot are above or below the x-axis then it indicates a problem with an array. These kinds of problems can sometimes be addressed with normalization. MA-plots can be viewed for a whole array, or for the individual print tip groups on an array. This diagram gives us a good indication of whether normalization within an array is needed for the complete view of median and IQR values of all the 107 arrays before and after normalization.

MA plot is done to show how genes are sequentially expressed in an array. In normalization process of MA plot median and IQR values of all the arrays with unknown function and null value will be removed and median value will be reduced.

After normalization the median values for some arrays are assigned as zero this is because the genes are not expressed i.e. lesser the median value the expression level will be less, greater the median value expression level is more.

IDENTIFICATION OF DIFFERENTIALLY EXPRESSION GENES ON BMK1/ERK1 SIGNALING CASCADES:

After normalization of gene expression data has recognized 21001 genes out of 22283 gene terms entered by the user. 20986 genes were recognized by gene symbol and 15 genes by other gene IDs. 7996 duplicate genes were removed (keeping the highest ranking instance of each gene) leaving a total of 13005 genes. Only 12073 of these genes are associated with a GO term. As expected, we observed distinct hierarchical clustering by tissue type within each smoking group. Overall, there were 2436 probes (1920 genes) that significantly ($p<0.001$) distinguished adenocarcinoma from normal lung tissue. Among the 1066 genes with a fold change >1.5 and $p<0.001$, 116 were similarly altered in each smoking group, while 364, 261, and 60 genes differentiated adenocarcinoma/ normal lung tissue only in former smokers, current smokers, and never smokers, respectively. The remaining 265 genes were similarly altered in some but not all smoking groups. **(Fig: 12, Fig: 13, Table 1)**

PRINCIPLE COMPONENT ANALYSIS:

PCA is used to reduce multidimensional datasets to lower dimensions for analysis; it is a technique that can determine the key features of high-dimensional datasets. In PCA essential cluster arrays are formed by groups that are of most significantly deregulated probes. Clustering first the most significant group, and then by progressively less significant groups. Given the experimental design of the dataset that we are attempting to analyze here, adenocarcinoma and normal lung tissue groups with most of the variance in expression level should have been introduced because of the conditions under scrutiny. **(Fig: 14)**

SCATTER PLOT:

It is also known as scatter graph is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. The data is displayed as a collection of points, each having the value of the variable determining the position on the other variable determining the position on the vertical axis. **(Fig: 15, Fig: 16, Fig: 17, Fig: 18)**

FUNCTIONAL ENRICHMENT ANALYSIS:

Functional enrichment analyses of the 1066 differentially expressed genes were performed to identify over-represented biological functions using Gene Ontology terms and pathways. Functional enrichment analysis based on DAVID classification shows 116 were similarly altered in each smoking group, while 261, 364, and 60 genes differentiated adenocarcinoma/ normal lung tissue only in current, former, and never smokers, respectively. The remaining 265 genes

were similarly altered in some but not all smoking groups. The over expressed biological functions among the up and down regulated differentially expressed genes in the parental and induced BMK1/ERK1 group, respectively (DAVID P<50.05). (Tab-7) overlapping genes with p-value &q-value)

We also used more robust methods designed to show relations between subject groups. These methods do not allow comparisons between tumor and nontumor sample from the same subject. with False distribution rate with p value, Functional enrichment analysis on cluster of gene expression in many cell signaling pathways, and List of Gene Id's, gene symbols with gene names.

NETWORK ANALYSIS

Identified sub networks were subjected to functional enrichment analyses by 'GO rilla a tool for identifying the GO term' to reveal over-represented biological functions within each subnetwork.

MAPKs are protein Kinases which on activation phosphorlate their specific nuclear or cytosolic substrates at serine or threonine residues or both. Such phosphorylation events can either positively or negatively regulate substrate, and thus entire signaling cascade activity.

The major cytosolic target of activated ERKs is RSKs (Ribosomal protein S6 Kinase). Active RSKs translocates to the nucleus and phosphorylates such factors as c-Fos on Ser362, SRF (Serum Response Factor) at Ser103, and CREB (Cyclic AMP Response Element-Binding protein) at Ser133.

In the nucleus activated ERKs phosphorylate many other targets such as MSKs (Mitogen- and Stress-activated protein kinases), MNK (MAP interacting kinase) and Elk1on Serine383 and Serine389. ERK can directly phosphorylate CREB and also AP-1 components c-Jun and c-Fos. Another important target of ERK is NF-KappaB. This study reveals that nuclear pore proteins are direct substrates for ERK, Other ERK nuclear targets include c-Myc, HSF1 (Heat-Shock Factor-1), STAT1/3 (Signal Transducer and Activator of Transcription-1/3), and many more transcription factors. **(Fig: 11, Fig: 12)**

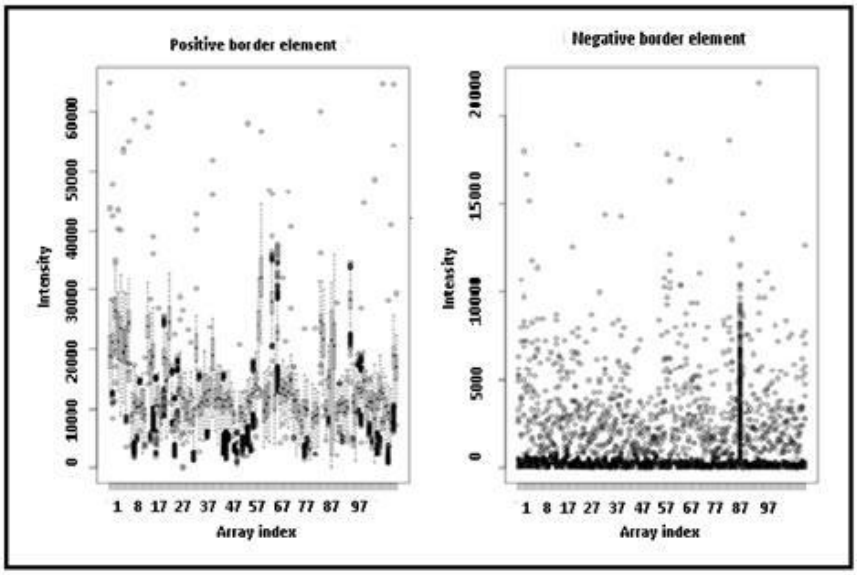
Table 1: Genes common in current, never and former smokers

Probe	Gene Symbol	Pvalue	HR	Lower .95	Upper .95
215616_s_at	KDM4B	2.26E-05	1.7967	1.3701	2.3559
208942_s_at	SEC62	4.06E-05	0.45	0.3074	0.6589
216627_s_at	B4GALT1	0.000132	1.8344	1.3439	2.5038
221044_s_at	TRIM34	0.000198	1.4886	1.2073	1.8356
206448_at	ZNF365	0.000208	1.5436	1.2272	1.9416
215000_s_at	FEZ2	0.000287	1.6395	1.2551	2.1415
210620_s_at	GTF3C2	0.000376	1.635	1.247	2.1437
217877_s_at	GPBP1L1	0.000447	1.7256	1.2725	2.3401
221115_s_at	LENEP	0.000634	1.5902	1.2187	2.075
213850_s_at	SFRS2IP	0.000784	0.5237	0.359	0.7639
212442_s_at	LASS6	0.000785	1.6998	1.2472	2.3167
222242_s_at	KLK5	0.000786	1.4736	1.1751	1.8478
212115_at	HN1L	0.000802	1.5381	1.1958	1.9784
218751_s_at	FBXW7	0.000814	0.4169	0.2498	0.6958
218672_at	SCNM1	0.000833	1.4891	1.1789	1.8808
216920_s_at	TARP	0.000862	0.329	0.1711	0.6327
205769_at	SLC27A2	0.000929	1.4456	1.1623	1.798
41660_at	CELSR1	0.000942	1.5144	1.1842	1.9367
204718_at	EPHB6	0.001049	1.4146	1.1496	1.7407
209744_x_at	ITCH	0.001052	1.6405	1.22	2.2058
204019_s_at	SH3YL1	0.001213	1.5522	1.1893	2.0258

204415_at	IFI6	0.001213	1.5374	1.1848	1.9949
221563_at	DUSP10	0.001238	1.5711	1.1944	2.0666
212778_at	PACS2	0.001403	1.3684	1.1288	1.6587
209376_x_at	SFRS2IP	0.001406	0.5324	0.3616	0.7839
212517_at	ATRN	0.001463	0.4986	0.3248	0.7655
209797_at	CNPY2	0.001467	0.4752	0.3005	0.7515
205155_s_at	SPTBN2	0.001483	1.652	1.2121	2.2514
213622_at	COL9A2	0.001532	1.369	1.1273	1.6625
34406_at	PACS2	0.001536	1.313	1.1094	1.5539
208835_s_at	LUC7L3	0.001585	0.4828	0.3072	0.7586
221709_s_at	ZNF839	0.001656	1.6082	1.1962	2.1622
212446_s_at	LASS6	0.001657	1.5315	1.1743	1.9974
207493_x_at	SSX2	0.001691	1.323	1.1109	1.5756
215243_s_at	GJB3	0.001712	1.4293	1.1434	1.7868
203153_at	IFIT1	0.001712	1.4114	1.138	1.7506
206921_at	GLE1	0.001722	1.492	1.1617	1.916
209778_at	TRIP11	0.001768	1.4917	1.1609	1.9168
214754_at	TET3	0.001808	1.6322	1.1998	2.2204
210017_at	MALT1	0.001937	0.5484	0.3751	0.8018
217404_s_at	COL2A1	0.001944	1.5701	1.1803	2.0886
204211_x_at	EIF2AK2	0.001975	1.5796	1.1824	2.1101
205131_x_at	CLEC11A	0.002	1.6334	1.1966	2.2296
207194_s_at	ICAM4	0.002134	1.3999	1.1294	1.7352
214635_at	CLDN9	0.002145	1.5983	1.1847	2.1562
217949_s_at	VKORC1	0.002153	1.5498	1.1715	2.0503
220779_at	PADI3	0.002235	1.547	1.1695	2.0464
205157_s_at	KRT17	0.002269	1.3386	1.11	1.6143
65521_at	UBE2D4	0.00227	1.4669	1.147	1.8762
204914_s_at	SOX11	0.0023	1.5018	1.1563	1.9505
213361_at	TDRD7	0.002306	1.4633	1.1455	1.8692
217502_at	IFIT2	0.002325	1.4162	1.132	1.7718
221794_at	DOCK6	0.002357	1.6545	1.196	2.2887
205387_s_at	CGB	0.002466	1.3312	1.1061	1.6021
208952_s_at	LARP4B	0.002589	1.4513	1.139	1.8493
209973_at	NFKBIL1	0.002634	1.5875	1.1747	2.1454
213026_at	ATG12	0.002647	0.544	0.3658	0.8091
222148_s_at	RHOT1	0.002668	1.5308	1.1595	2.0212
219461_at	PAK6	0.002809	1.4643	1.1402	1.8806
218927_s_at	CHST12	0.002864	1.4697	1.1411	1.8929
218144_s_at	INF2	0.002907	1.4114	1.125	1.7707
207282_s_at	MYOG	0.002916	1.6048	1.1753	2.1912
206498_at	OCA2	0.002942	1.4181	1.1265	1.7852
201952_at	ALCAM	0.003017	1.5275	1.1545	2.021

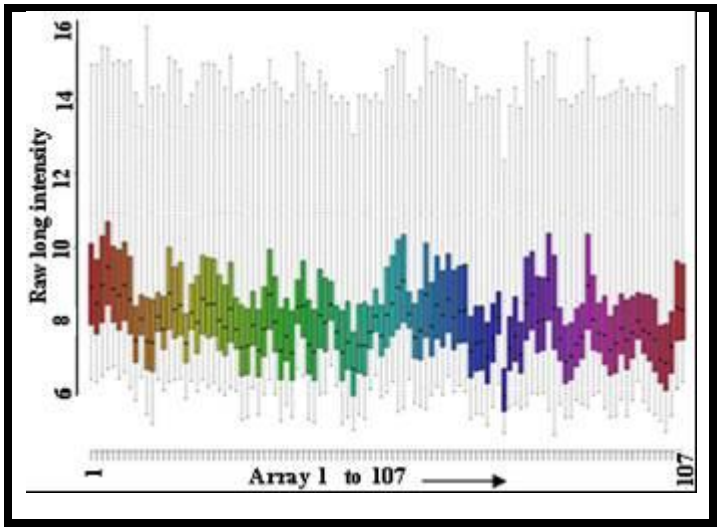
200975_at	PPT1	0.003041	1.3952	1.1194	1.7389
215505_s_at	STRN3	0.003176	1.3969	1.1187	1.7441
55692_at	ELMO2	0.003282	0.589	0.4139	0.8383
218387_s_at	PGLS	0.003292	1.5932	1.1678	2.1735
204104_at	SNAPC2	0.003319	1.5725	1.1625	2.1272
213711_at	KRT81	0.003349	1.3167	1.0956	1.5825
211295_x_at	CYP2A6	0.003379	1.2751	1.0838	1.5001
215392_at	AU148154	0.00351	0.5792	0.4014	0.8357
220909_at	TRIM46	0.00352	1.4355	1.126	1.8299
213505_s_at	SFRS14	0.003598	0.5826	0.405	0.8382
206457_s_at	DIO1	0.003731	1.5142	1.1439	2.0042
221690_s_at	NLRP2	0.003769	1.408	1.117	1.7748
212006_at	UBXN4	0.003878	1.4742	1.1328	1.9184
205853_at	ZBTB7B	0.003945	1.5783	1.1573	2.1525
211398_at	FGFR2	0.003995	1.5516	1.1505	2.0925
201107_s_at	THBS1	0.004126	1.3397	1.097	1.636
206277_at	P2RY2	0.004136	1.5315	1.1444	2.0495
220340_at	GREB1L	0.004177	1.4917	1.1346	1.9611
204156_at	SIK3	0.004184	0.5875	0.4082	0.8454
212599_at	AUTS2	0.004305	1.4925	1.1338	1.9648
220879_at	---	0.004305	1.5353	1.1438	2.0606
201166_s_at	PUM1	0.004305	1.4539	1.1245	1.8798
219763_at	DENND1A	0.004358	1.4545	1.1242	1.8817
207639_at	FZD9	0.004428	1.281	1.0801	1.5193
219370_at	RPRM	0.0045	1.4748	1.128	1.9281
213771_at	IRF2BP1	0.004541	1.3682	1.1019	1.699
214303_x_at	MUC5AC	0.004569	1.4135	1.1128	1.7955
208415_x_at	ING1	0.004661	0.56	0.3748	0.8368
219389_at	SUSD4	0.00468	1.4464	1.12	1.868
205738_s_at	FABP3	0.004702	1.3949	1.1074	1.757
205289_at	BMP2	0.004712	0.4119	0.2226	0.762
220123_at	SLC35F5	0.004714	1.4836	1.1285	1.9505
204915_s_at	SOX11	0.004773	1.4138	1.1116	1.7982
205290_s_at	BMP2	0.004795	0.4061	0.2171	0.7596
204913_s_at	SOX11	0.004819	1.4842	1.1279	1.9531
206862_at	ZNF254	0.004903	1.3846	1.1038	1.737
212051_at	WIPF2	0.004903	1.4452	1.1182	1.8678
201428_at	CLDN4	0.004905	1.5241	1.1363	2.0442
212177_at	SFRS18	0.004928	0.5404	0.3518	0.8299
216647_at	TCF3	0.005	1.525	1.1358	2.0476

Fig-1: Box plot of raw probe intensity values



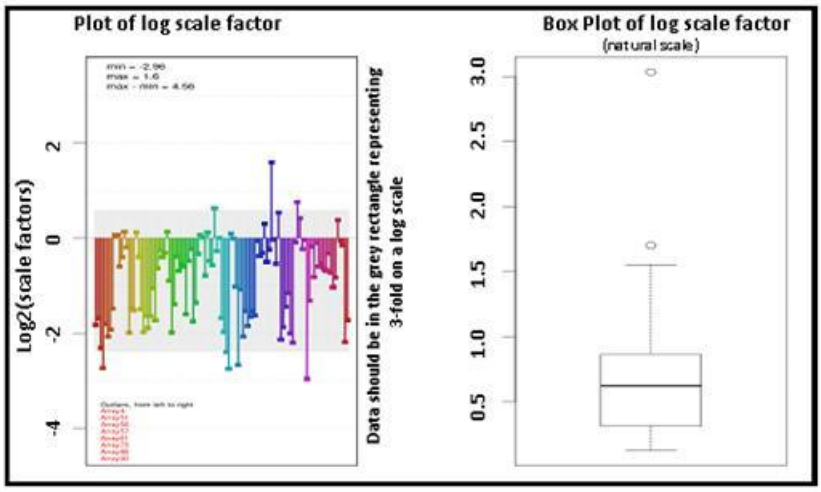
A boxplot is represented to compare the probe intensity levels between the arrays of a data set. Either end of the box represents the upper and lower quartile; the middle of the box represents the median. Horizontal lines, connected to the box by "whiskers", indicate the largest and smallest values not considered outliers. Outliers are values that lay more than 1.5 times the interquartile range from the first of the third quartile (the edges of the box); they are represented by a small circle.

Fig-2: Box plot of $\log_2$ transformed probe intensity values: microarray data were preprocessed based on the intensity of means vales. The X-axis represents the number of datasets based on statistical data and the y-axis represents log intensities ($p<0.005$).



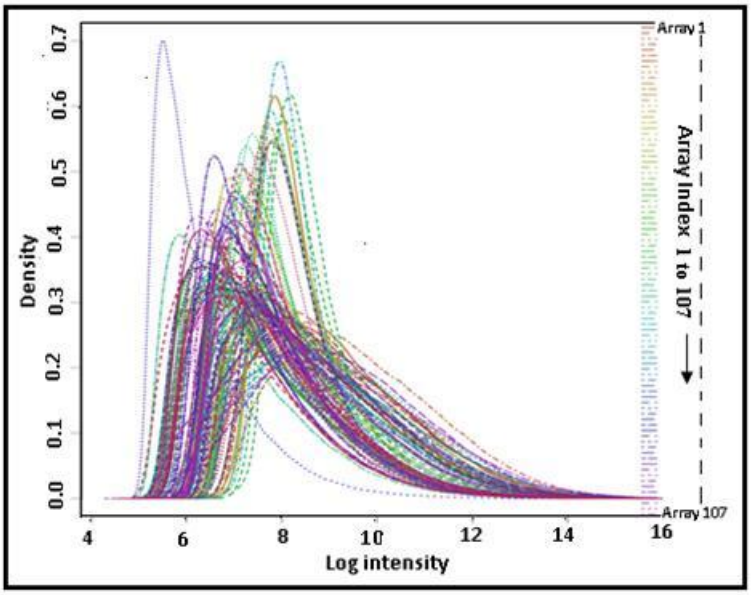
There are more arrays have different intensity levels which are drastically different from the rest of the array were corrected by normalization. For microarray data, these graphs are always constructed using log_2 transformed probe intensity values, as the graph would be virtually unreadable using raw values, as you can see below, where raw values are juxtaposed with log_2 transformed values.

Fig-3: Log2 intensity of more differential arrays with an outlier of 3 fold intensity calculations using GCRMA.



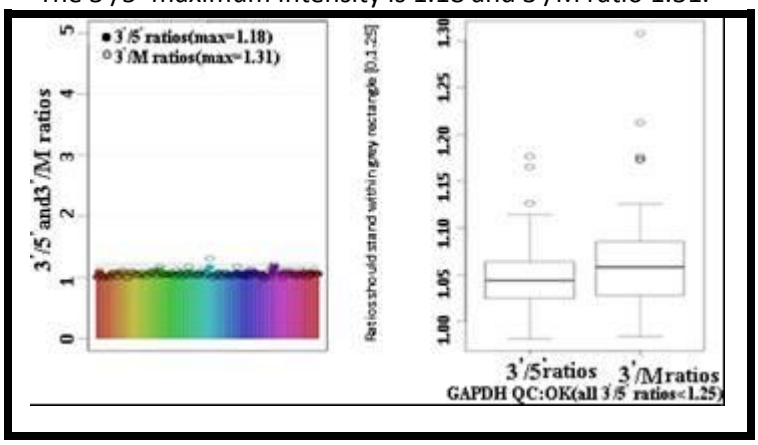
The arrays4, array51, array56, array57, array61, array75, array86 and array90 are significantly different on 3 folds of log2 score. The significance includes min:-2.96 and max: 1.6 with high fold intensities. The significant results include 3 are never smoker (GSM254629, GSM254728 and GSM254726 in early stage tumor), former smoker (GSM254633 in early stage tumor) and current smoker (GSM254641, GSM254663 and GSM254678 in early stage tumor) is compared with Normal tissue significance of early stage tumor tissue (GSM254640).

Fig-4: Raw probe intensity values plotted on a histogram



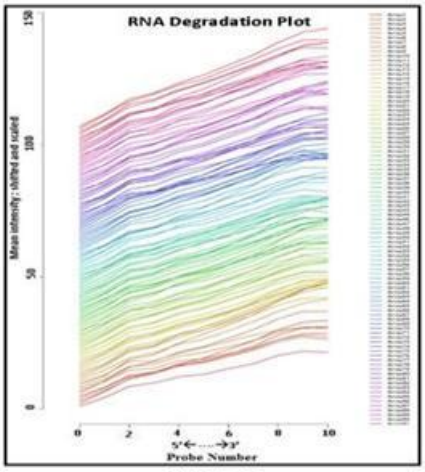
A histogram of array intensity levels is visualizing the spread of data and compare and contrast probe intensity between the arrays of the dataset. We predicted the intensity of the dyes used to absorb on hybridization, the data were predicted using an intensity map. The intensity of color was predicted according to distributions of signal intensity with log2 values (Fig: 4). The x-axis represents probe density level and the y-axis indicates probe intensity. As our data arrays all arrays are more deviated to the right, which as stated above could indicate high levels of background noise. We used all arrays as an intensity of a variable calculation.

Fig-5: RNA degradation map on GAPDH ratio of all 3'/5' ratios <1.25.
The 3'/5' maximum intensity is 1.18 and 3'/M ratio 1.31.



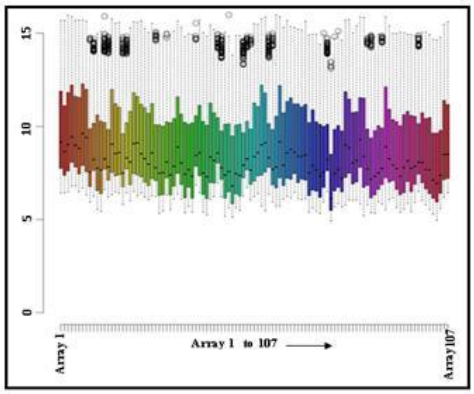
The above diagram gives a good indication of the quality of the sample that has been hybridized to the array, mRNA degradation occurs when molecule begins to break down and is therefore ineffective in determining gene expression. Because this kind of degradation starts at the 5' end of the molecule and progresses to the 3' end it can be easily measured using oligonucleotide arrays, where each PM probe is numbered sequentially from the 5' end of the targeted mRNA transcript. When RNA degradation is advanced, PM probe intensity at the 3' end of a probe set should be elevated when compared with the 5' end.

Fig-6: RNA degradation plot



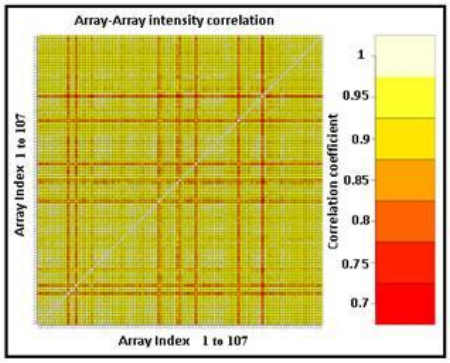
When dealing with high quality RNA a slope of between 0.5 and 1.7 is typical, depending on the type of array; slopes that exceed these values by a factor of 2 or higher could indicate excessive degradation, the actual value is however less important than agreement between the chips, because if all the arrays have similar slopes then comparisons within genes across the arrays may still be valid. Fig: 6 is an RNA degradation plot for the dataset which we are assaying. The slope falls within the recommended range, which indicates that all of the samples were of good quality and very strong correlation between the various arrays in the dataset.

Fig-7: Array of control HGU-133 Affymetrix array



We analyzed the expression of Affymetrix arrays which contain control probe sets, and annotated probe sets. The samples were hybridized according to log-intensity profiles of all probes with different intensities of arrays. The results were predicted in Fig 7. The image contains average curve, control probe sets with adenocarcinomas and normal tissue of lung induced stem cells data of too many probes so the profile plot cannot be read properly and used for QC.

Fig-8: Affymetrix array on preprocessed datasets arranged based on probability of intensity mean (p<0.005).



The processed probe intensities across all arrays with the overall higher level of probe signal intensity. Frequently measure is not very sensitive and problematic arrays may look like their better quality counterparts. We are predicting our dataset of QC of normalized datasets with the threshold intensities were predicted using quality arrays. Box plots of the log - intensity distribution are plotted for between-array comparison. The distributions of raw PM (perfect match probes) log-intensities are not expected to be identical but still not totally different while the distributions of normalized probe-set log-intensities are expected to be more comparable if not identical. Drawing these box plots before and after normalization allows also checking the normalization step.

Fig-9: Boxplot representation after normalization.

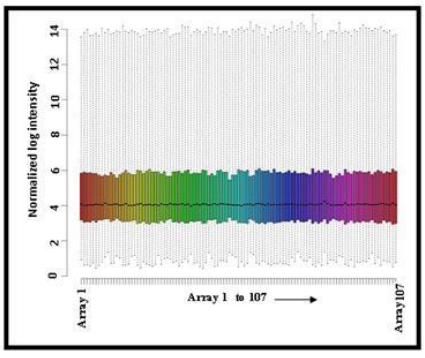


Fig-10: Intensity map of log_2 transformation

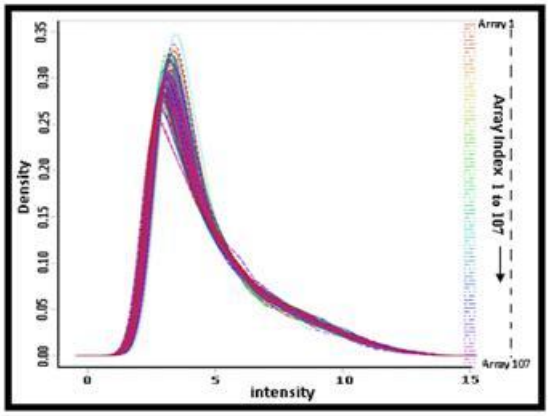


Fig-11: Affymetrix map after normalization to correct the intensity values based on correlation coefficient.

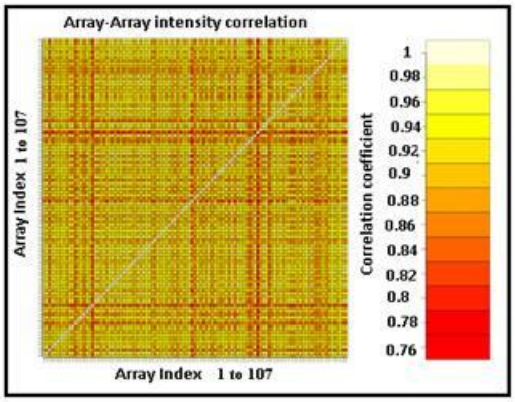
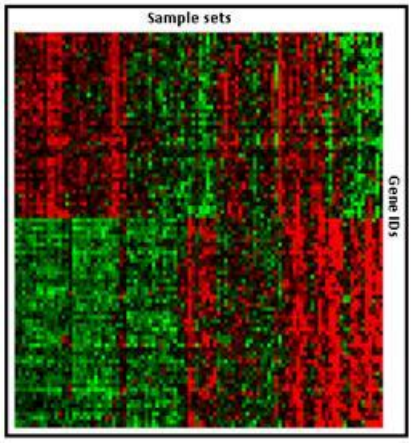
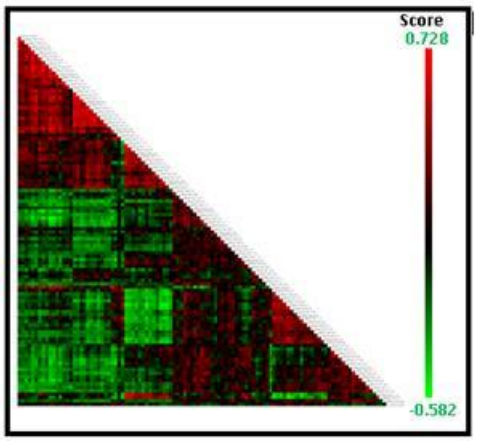


Fig-12: Hierarchical clustering of differentially expressed genes based on P-values (<0.001)



Using Gene functional enrichment analysis of <1.5 fold and the probability of 0.001 class difference of adenocarcinoma and normal lung tissue genes was compared. The significance alterations of gene expressions are altered in stage I and Stage II in the early stage of smokers. We specified differently expressed 64 up- and 98 down-regulated probesets, with 54 up- and 81 down-regulated genes.

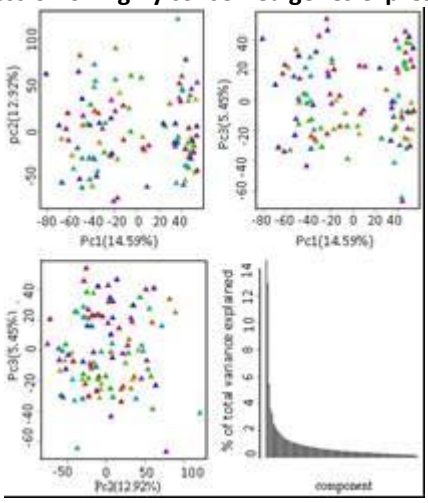
Fig- 13: Pearson Correlation coefficient of differential gene expression data analysis.



To verify whether the current and never smoking signature in the tumor was present also in former smokers, we compared the current with never and former and never smoking signatures in tumor and found 26 probes (22 down- and 4 unregulated, representing 21 genes) that differentiated both current with never and former and never smoking using stringent selection criteria.

PRINCIPLE COMPONENT ANALYSIS:

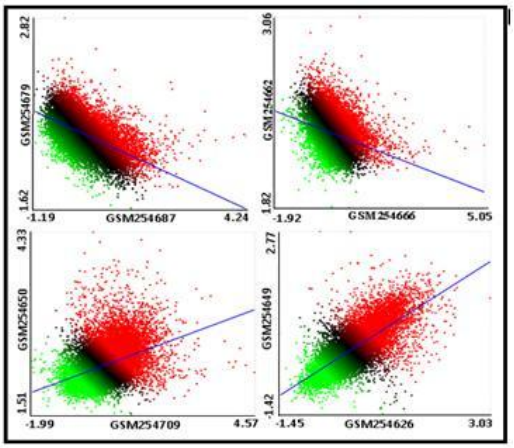
Fig-14: Principle component analysis after GCRMA normalization of gene expression of highly conserved genes expressed Gene.



These plots provide information on high levels of background noise or otherwise compromised data cannot be corrected by normalization, as becomes clear from its non-clustering with its fellow group members before or after preprocessing. Given this and previous information we can be quite confident that this array shouldn't be included in the differential expression analysis.

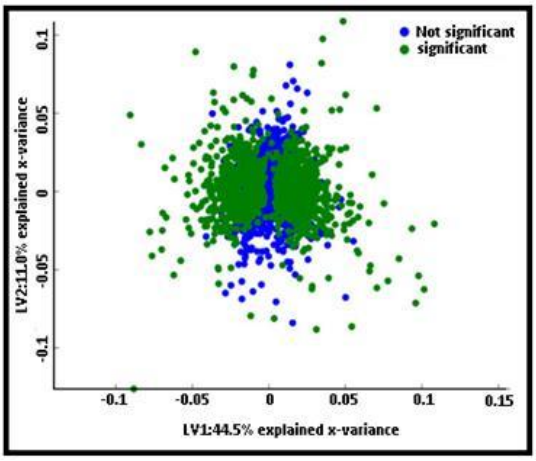
SIGNIFICANCE OF GENE EXPRESSION

Fig-15: Scatter plot of significance up regulated and down regulated current smoker and never smoker gene datasets.



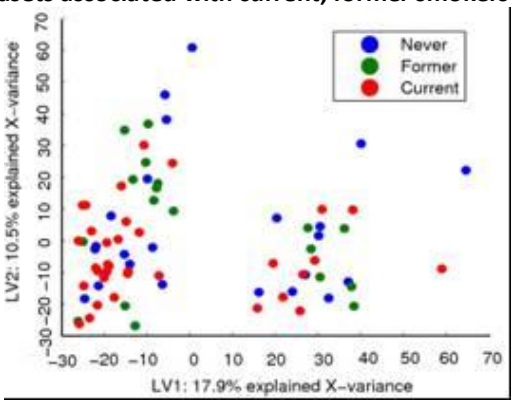
The above diagram shows the distribution of fold change between the upregulated and down regulated current smoker and never smoker genes data set, the upregulated and down regulated Prob sets are compared and the expression level is identified.

Fig-16: Scatter plot on significance of never smoker datasets using Pearson correlation coefficient with significance of <0.005.



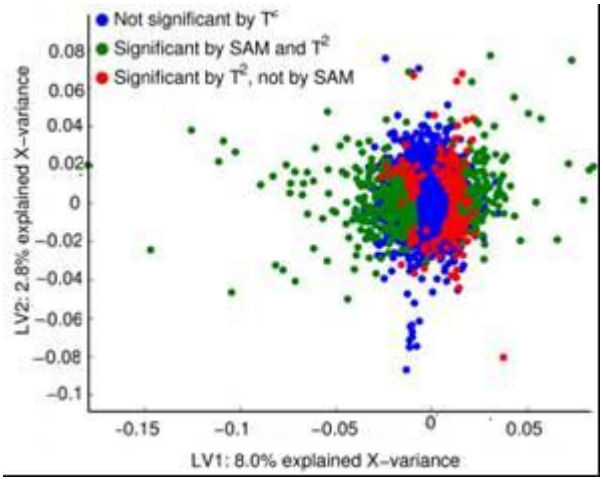
The above graph **Fig-16** represents the significant and not significant genes which is plotted across level1 and level2 which was taken at different intensities at different time interval. Green color indicates the highly expressed genes at and the blue color indicates less expressed genes.

Fig-17: Datasets associated with current, former smokers and never.



The above graph **Fig-17** represents the never, former, and current smokers data set, which is plotted across x and y axis as level1 and level2 with different intensity values, which was taken at different time interval. Blue color indicates the gene expression of never smoker, green color indicates gene expression of former smoker and red color indicates gene expression of current smoker, higher the intensity value higher will be the level of expression.

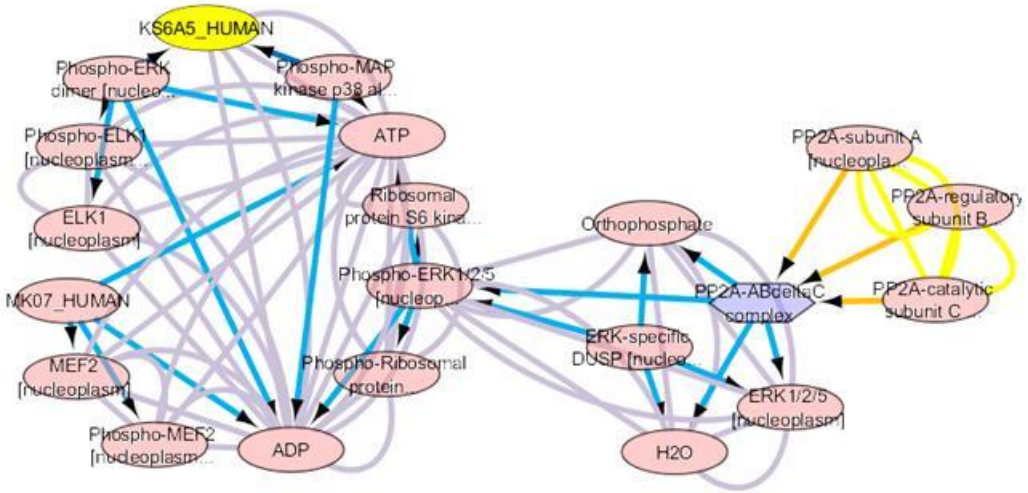
Fig-18: Common number of genes in both current, never and former smoker datasets.



The above graph Fig-4.11c represents the not significant by T^2, significant by SAM and T^2 and significant by T^2 and not by SAM data sets, which is plotted across x and y axis as level1 and level2 with different intensity values, which was taken at different time interval. Blue color indicates the expression of not significant genes by T^2, green color indicates expression of significant genes by SAM and T^2 and red color indicates expression of significant genes by T^2 and not by SAM, higher the intensity value higher will be the level of expression.

NETWORK ANALYSIS:

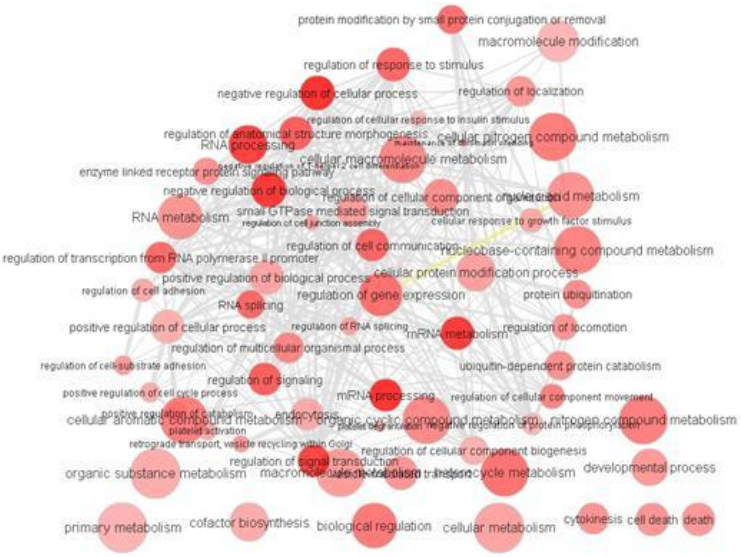
Fig- 19: Phosphorylation of ERK1/2/5 protein in adenocarcinomas



MAPKs are protein Kinases which on activation phosphorlate their specific nuclear or cytosolic substrates at serine or threonine residues or both. Such phosphorylation events can either positively or negatively regulate substrate, and thus entire signaling cascade activity.

The major cytosolic target of activated ERKs is RSKs (Ribosomal protein S6 Kinase). Active RSKs translocates to the nucleus and phosphorylates such factors as c-Fos on Ser362, SRF (Serum Response Factor) at Ser103, and CREB (Cyclic AMP Response Element-Binding protein) at Ser133.

Fig-20: The regulatory genes present in lung cancer that contains final activities



3. CONCLUSION

This study, although preliminary, suggests that lung cancer of smokers and nonsmokers have different etiologies, and involve different pathways of cell transformation. This implies that optimal approaches to treatment might differ in lung cancer of smokers and nonsmokers. Our study also suggests that a number of premalignant changes occur in the noninvolved lungs of smokers; these changes in gene expression might represent targets for preventative therapy. The possibilities raised in this article suggest that a larger study comparing both noncancerous and tumor tissue in smokers and nonsmokers, and examining gene expression samples in tissue from smokers with and without cancer, would be of considerable clinical importance.

4. ACKNOWLEDGMENTS

We extend our sincere thanks to the management of The Oxford College of Engineering and Scientific biominds for their valuable support and suggestions for preparing the manuscript and also to the Department of Biotechnology for providing necessary resources.

5. REFERENCE

[1] Jemal, A.; Bray, F.; Center, M.M.; Ferlay, J.; Word, E.; Forman, D.; J. Clin Global cancer statistics. CA Cancer. 2011.

[2] Cho, W.C.; Yip, T.T.; Cheng, W.W.; Au, J.S. Serum amyloid A is elevated in the serum of lung cancer patients with poor prognosis. Cancer 2010.

[3] Samet JM, Avila-Tang E, Boffetta P, et al. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Cancer 2009.

[4] Enstrom JE, Kabat GC. Environmental tobacco smoke and tobacco related mortality in a prospective study of Californians, 1960-98. BMJ. 2003.

[5] Mumford, J. L., He, X. Z., Chapman, R. S., Cao, S. R., Harris, D. B., Li, X. M., Xian, Y. L., Jiang, W. Z., Xu, C. W., Chuang, J. C., Wilson, W. E., and Cooke, M. Lung cancer and indoor air pollution in Xuan Wei, China. Science (Washington DC), 1987.

[6] Rodenhuis, S., and Slebos, R. J. C. Clinical significance of ras oncogene activation in human lung cancer. 1992.

[7] Greenblatt, M. S., Bennett, W. P., Hollstein, M., and Harris, C. C. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis.1994.

[8] Keohavong, P., DeMichelle, M. A. A., Melacrinos, A. C., Landreneau, R. J., Weyant, R. J., and Siegfried, J. M. Detection of K-ras mutations in lung carcinomas: relationship to prognosis. 1996.

[9] Siegfried, J. M., Gillespie, A. T., Mera, R., Casey, T. J., Keohavong, P., Testa, J. R., and Hunt, J. D. Prognostic value of specific K-ras mutations in lung adenocarcinomas. Cancer Epidemiol. Biomark. Prev, 1997.

[10] Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson. 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas subclasses. Proc. Natl. Acad. Sci. USA 98.

[11] Anbazhagan, R., T. Tihan, D. M. Bornman, J. C. Johnston, J. H. Saltz, A. Weigering, S. Piantadosi, and E. Gabriel son.. Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles 1999.

[12] Dr Alan E Guttmacher; Genome Study Finds 26 Lung Cancer Genes; 23 Oct 2008.

[13] Alice Shaw, MD, PhD, New Genetic Subtype of Lung Cancer Defined; Journal of Clinical Oncology; 31 Jan 2012;

[14] E. Brambilla and A. Gazda, Pathogenesis of lung cancer signaling pathways: roadmap for therapies; Eur Respir J. 2009 June

[15] Lee, J.-D., Ulevitch, R.J. and Han, J.H. Primary structure of BMK-1; A new mammalian MAP kinase. Biochem. Biophys. Res. Comm., 1995.

[16] Kato,Y., V.V.Kravchenko, R.I.Tapping, J.Han, R.J.Ulevitch, and J.-D.Lee. BMK1/ERK5 regulates serum-induced early gene expression through transcription factor MEF2C. EMBO J.

[17] Hayashi Y, Iwashita T, Murakamai H, Kato Y, Kawai K, Kurokawa K, Tohnai I, Ueda M, Takahashi M. Activation of BMK1 via tyrosine 1062 in RET by GDNF and MEN2A mutation. Biochem Biophys Res, 2001 Mar 2

[18] Lee JD, Ulevitch RJ, Han J. Primary structure of BMK1: a new mammalian map kinase. Biochem Biophys Res. 1995

[19] Chang L, Karin M. Mammalian MAP kinase signaling cascades. Nature 2001.

[20] Weldon CB, Scandurro AB, Rolfe KW, Clayton JL, Elliott S, Butler NN, Melnik LI, Alam J, McLachlan JA, Jaffe BM, et al. Identification of mitogen-activated protein kinase as a chemo resistant pathway in MCF-7 cells by using gene expression microarray. Surgery. 2002

***Corresponding Author:**
Suchithra S*
Department of Biotechnology
The Oxford College of Engineering
Hosur Road, Bangalore