International Journal of Advanced Biotechnology and Research
ISSN 0976-2612, Online ISSN 2278–599X,
Vol 3, Issue 4, 2012, pp 824 -834
http://www.bipublication.com

ANALYSIS OF LUNG CANCER MICROARRAY DATA IDENTIFIES NEW POTENTIAL GENES TARGETS FOR INHIBITOR DESIGN.

Bhagavathi S.[1], Gulshan Wadhwa[2] and Anil Prakash [3]

[1,2]Dept. of Biotechnology, Apex Bioinformatics Centre, Department of Biotechnology, Ministry of Science and Technology, Government of India, Block-2, 7th Floor, C.G.O. Complex Lodhi Road, New Delhi-110003
[3]Barkatullah University, Bhopal
Corresponding Author: S.BHAGAVATHI
Correspondence:bhagavathikanagaraj@gmail.com

[Received-16/10/2012, Accepted-02/12/2012]

ABSTRACT

Lung cancer is the leading cause of cancer associated deaths Worldwide and has one of the poorest prognoses among all cancer types. It is a disease which causes deaths all over the world and is rapidly increasing in number. Lung cancer is the most common cause of cancer-related mortality worldwide, with >160 000 deaths in the United States in 2008 and a poor 5-year survival rate, which remains stable at 15% .Analysis of the genetic alterations occurring in lung cancer has shown that histopathological differences are in line with genetic heterogeneity of the disease .Consistent with the genetic heterogeneity, expression studies have not only unveiled profound differences between SCLC and NSCLC but also within different NSCLC subtypes , However, until recently all NSCLC lung cancer forms have been treated with similar approaches regardless of their biological differences. Microarray studies in cancer compare expression levels between two or more sample groups on thousands of genes. Data analysis follows the approach comparison of sample mean to identify differentially expressed genes. This leads to the discovery of 'population-level' markers, i.e., genes with the expression patterns A > B and B > A. These tests and the gene expression pattern grid may be useful for the identification of therapeutic targets and diagnostic or prognostic markers that are present only in subsets of cancer patients, and provide a more complete portrait of differential expression in cancer. Various tests and experimental parameters were set up and from this we characterized about 8 differentially expressed genes and finally tested/validated using statistical tests.

Key Words: Lung cancer, Microarray data analysis, gene expression.

1. INTRODUCTION

Microarray data analysis technologies are a powerful approach for genomic research. The multistep, data intensive nature of this technology has created an unprecedented informatics and analytical challenge. It is important to understand the crucial steps that can affect the outcome of the analysis. These technologies have provided a contemporary trend on various analytical steps in the micro array data analysis which includes experimental design, data standardization, image acquisition and analysis, normalization, statistical significance inference, exploratory data analysis, class prediction and pathway analysis as well as various considerations relevant to their implementation.[30]. Micro array technologies permit the simultaneous monitoring of gene expression values for tens of thousands of genes in one experiment. [49,32,14] Studies of differential expression of individual genes often find genes that are up-regulated in some tumors, and down-regulated in others. Microarray studies typically seek to identify differentially expressed genes using use fold-change [14] t-tests [3] and models. [27,54,6,21]

Micro array are used to assay gene expression within a single sample to compare gene expression in two different cell types or tissue samples. Because a microarray can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way scientists examine gene expression.

The methods of analysis for identifying differentially expressed genes in data from microarray experiments vary widely, but all are focused on the question of whether genes are over-expressed or under-expressed in samples in group A (e.g., tumor, or treatment, or metastatic, or responder) compared to samples in group B (e.g., normal, or control, or quiescent, or nonresponder). These patterns can efficiently be referred to as AB (over expressed in A) and BA (under expressed in A) patterns. Typically, researchers use study designs that favor biological replication to maximize the ability to detect reproducibly genes that are differentially expressed in a patient population, at a sacrifice of the ability to detect individual-specific patterns of differential expression with technical replication. Most cancers are diseases with heterogeneous etiologies; moreover, the development of every primary tumor in different individuals is a unique biological event. Thus, the expression levels of genes in the individual patient are also important; some important gene dysregulation may be highly specific to each individual

Nevertheless, the study of gene expression represents an important and necessary first step under in our understanding and cataloguing of the human genome. As more information accumulates, Scientists will be able to use micro arrays to ask increasingly complex questions and perform more intricate experiments. With new advances, researchers will be able to infer probable functions of new genes based on similarities in expression patterns with those of known genes. Ultimately these studies promises to expand the size of existing gene families reveal new patterns of coordinated gene expression across gene families and uncover entirely new categories of genes.

Furthermore, because the product of any one gene usually interacts with those of many others, our understanding of how these genes coordinate will become clearer through such analyses and precise knowledge of these inter relationships will emerge. The use of micro arrays may also speed the identification of gene involved in the development of various diseases by enabling the scientists to examine a much larger number of genes.

Taking all pros and cons into consideration, this study has been initiated with the objective to identify genes which may be responsible for Lung cancer using microarray data in-silico approach, so that this information could be used to identify the genes which are differentially expressed during lung cancer conditions and to understand the behavior of genes and the same could be utilized for drug designing , [25] therapy and advance diagnosis of cancer.

In the field of cancer research, the set of differentially expressed genes represents a potential goldmine of markers for molecular medicine. Specially, studies have demonstrated the vast potential of differentially expressed genes as biomarkers for molecular diagnosis [13] including stage and grade.[37,42,1,23,45,34], prognosis [4], therapy outcome prediction [28,39,36], malignancy and metastasis [31,51,17,41] and tumour aggressiveness [9]. With the ability to detect drug-resistant tumour using a drug response index that summarize signature profiles (positive and negative), physician will be able to prescribe drugs that are less likely to fail and thus generate more individualized prescriptions that have higher positive response rates. Eventual uses of this information will extend beyond profiling.

Tests for identifying differentially expressed genes should be as powerful (sensitive) and yet as specific as possible. There is a growing demand for easy- to –use tools for the analysis and interpretation of global patterns of gene expression from microarray experiments. While commercial platforms are available, most are specially designed for data from one platform. [24].

Genome wide micro array technologies which are widely used to monitor global gene expression in cancer have identified numerous differentially expressed genes. [10,19].

These studies shows a strong association of differentially expressed genes which is significantly more expressed in tumors as opposed to normal tissues from lung cancer patients and have demonstrated that it can also be a strong Prognostic indicator. [2,20,26]

2. MATERIALS AND METHODS

The gene expression datasets on lung cancer gene expression by Garber et. al were done using Affymetrix chips and the expression levels measured. In the present study the number of samples used were 72, the number of genes were 916 and sample classification was down by giving group (1) abnormal (2) normal genes.

The expression of microarray data from pilot studies, basic research and large scale clinical trials require the development of integrative computational tools that can not only analyze gene expression but that can also evaluate methods of analysis. The microanalysis steps include the experimental design and implementation, data collection and archival image acquisition, image analysis, data pre-processing, data normalization and identifying differentially expressed genes with exploratory data analysis. To tackle this

problem, c*aGEDA* format 1 was used which facilitates post analysis translation interpretation.
Global correction of sample mean, mean verses variance graph, MA plot, score histogram of genes, mean histogram expression pattern grid, test for differential expressed genes, permutation test, K-mean clustering etc were used to normalize the data, their correction and finding out significantly expressed genes to ultimately marking the genes which may be possible potential candidates for causing the Lung cancer. Various tests and experimental parameters were set up. The output was generated which has graphical output, class comparison output, expression pattern output and various tests output, these outputs were analyzed and a detailed analyses of each of the graphical outputs was done in order to get gene expression analysis. Gene expression results were studied for the differentially expressed genes and the putative cancer causing genes were identified.

2.1. Data formats

Currently, caGEDA accepts data in two formats, caGEDA format 1 and caGEDA format2. They are identical with one difference: format2 includes the accession number to facilitate post-analysis translational interpretation we are using caGEDA format1.

2.2. Missing Values

Missing values are encoded as 'NA' or'?' and can be estimated using k-nearest neighbours. A total of three neighbours in Euclidean space are used; if one of the neighbours also has a missing value, the next neighbour is found.

2.3. Finding differentially expressed genes

Differentially expressed genes (DEGs) are genes whose expression levels are significantly different between two groups of experiments. The genes are relevant for discovering potential pharmaceutical targets and diagnostic or prognostic markers **[50].** Identification of differential gene expression is the first task of an in depth microarray analysis**. [44]**
These include a number of new tests J5, D1, simple separability, weighed separability, the permutation percentile separability test (PPST) and the ABA test.

2.4. Statistical data

2.4.1. Simple separability test

We define N_1 as the number of samples in group A for which the expression value for the *ith* gene is less than the minimum value of that gene of group B (the subscript *i* in N_{1i} has been omitted for ease of reading) ; N_2 as the number of samples in group A for which the expression value for the *ith g*ene is greater than the maximum value of that gene of group B, N_3 as the number of samples in group B for which the expression value for the $_{ith}$ gene is less than the minimum value of that gene of group A; and N_4 as the number of samples in group B for which the expression value for the *ith* gene is greater than the maximum value of that gene of group A . This is equivalent to the proportion of non-overlap between the two distributions.

The counts N_1, N_2, and N_3, N_4 are paired such that in the perfectly separable case when $N_1=N_A$, then $N_2=0$, or vice versa; and when $N_3=N_B$, then $N_4=0$, or vice versa. Our score Si, which is:

$S_i = (N_1+N_4)/(N_A+N_B)$ or $(N_2+N_3)(N_A+N_B)$ or $(N_2+N_3)(N_A+N_B)$

(Whichever is greater), has an upper limit of 1.0 the threshold, T, ranges from 0 to 1. Genes with S scores greater than or equal to T are retained (1= perfect separability, no overlap in the distribution).

2.4.2. Weighted Separability

Clearly, simple separately is not as informative as it might be, especially in the comparison of two sample distributions that are perfectly separated, one by a large magnitude and the other by a smaller magnitude. Under weighted separability, the score for each gene is weighted using information on the magnitude of the difference between the group means. Specifically, the weight of each of m given genes is:

W_i-($mean_{Ai}$-$mean_{Bi}$) / (1/m)sum(-$mean_{Ai}$-$mean_{Bi}$)

Under weighted separability, the score S_i become w_iS_i and has a lower limit of 0 and an upper limit of N_A+N_B

2.4.3. J5 test

The J5 test is a gene-specific ratio between the mean difference in expression intensity between two groups A and B, to the average mean group difference of all m genes.

$$J5_i = \frac{\overline{A}_i - \overline{B}_i}{\frac{1}{m}\sum_{j=1}^{m} |\overline{A}_j - \overline{B}_j|}$$

The J5 test is likely to be useful in pilot studies where, due to high variance, t-test are likely to exhibit unacceptably low specificity (high false discovery rates).

2.4.4. Permutation percentile separability test

PPST examines whether a gene is over expressed (or under expressed) in the test group compared with the control group and vice versa. This is achieved by first counting the number of individuals in a group that exhibit an expression intensity that is greater (or less than) than the upper (or lower) 95th percentile of the alternate group. This number is compared with a distribution of counts that results when an arbitrarily large number of permutations of sample class labels are performed. PPST identifies genes with AB (i.e., where A>B) and BA (i.e., where B>A) patterns of differential expression.

2.4.5. ABA test

The ABA test is a special form of the PPST (Permutation percentile separability test) and reports genes that show both a surprisingly large number genes that are over expressed compared with the alternate group and a surprisingly large number of samples that are under expressed compared with the alternate group. The test is called ABA to reflect the pattern in which genes from the A group are found in the tails of the B group, and it reports genes that exhibit either significant ABA or BAB patterns. caGEDA provides output for both, and an expression pattern grid for the ABA pattern genes (tumour-normal –tumour).

3. RESULTS AND DISCUSSION

The goal of our paper was analysis of microarray data on Lung cancer. The main task of our analysis was the identification of up- and down-regulated genes. With the solution of the main task we also resolved others:

- Comparison of different subtypes of lung cancer
- Estimation of the number of genes differentially expressed in different subtypes of lung cancer
- Identification of up- and down-regulated biological processes according to Gene Ontology
- Detection of putative targets for some known chemical compounds.

The differentially expressed genes can be identified using statistical testing. We find an abundance of AB and BA pattern genes, with roughly the same number of genes called significant under the parametric t-test. We also find large numbers of genes with significant ABA test scores, and some with 'BAB' pattern genes. There is a marked tendency in most data sets for more ABA (cancer-normal-cancer) type genes than BAB pattern genes. These patterns are also reflected in 'expression pattern grids' of gene with significant *s3* (ABA) statistics (Fig.1) [24]

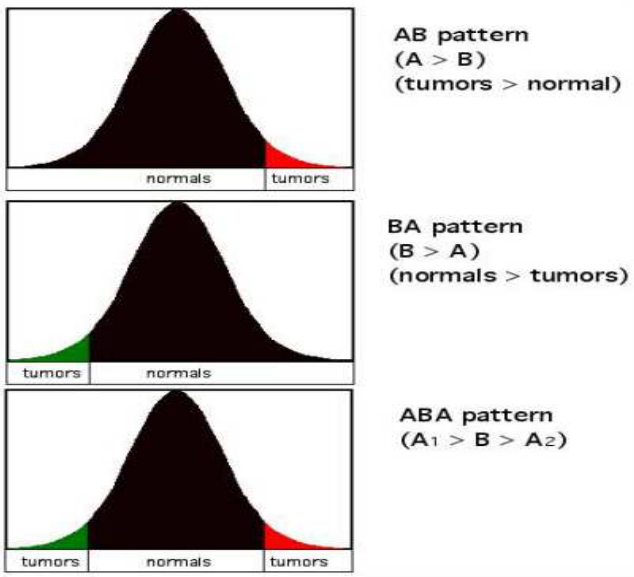

Fig.1: Conceptual representation of AB, BA, and ABA patterns of differential expression. The colored tails represent the placement of expression values of a given gene in tumours when compared to the distribution of expression values in normal samples. Standard AB and BA patterns are represented by red and black, respectively.

The result of user defined group classification revealed group (1) as abnormal and group (2) as normal. Performance evaluation was conducted for number of times bipartition recovered at 0.0. The genes retained were 64. The correlation between micro array after normalization was 0.172 (Fig. 2). The red portion in Fig. 2 indicates the differentially expressed genes. To evaluate the relationship between means of both the groups and their respective variances, the means and variances were presented graphically where means ranges from-51 to 54 on X-axis and variances from 0 to 12325 on Y-axis (Figure.3).

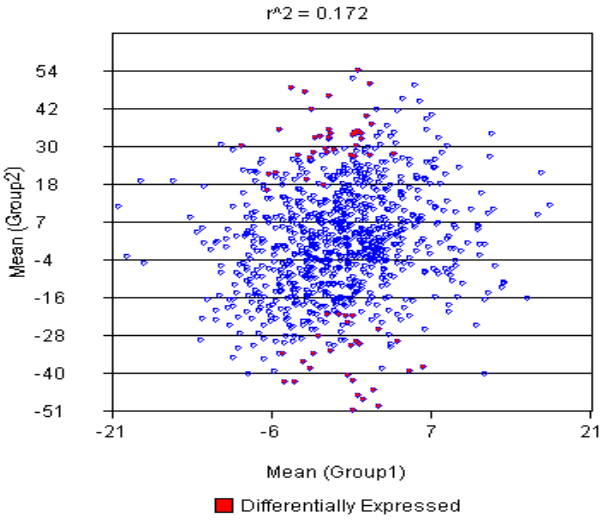

Fig.2: Global correction graph of means of differential expressed genes of lung cancer

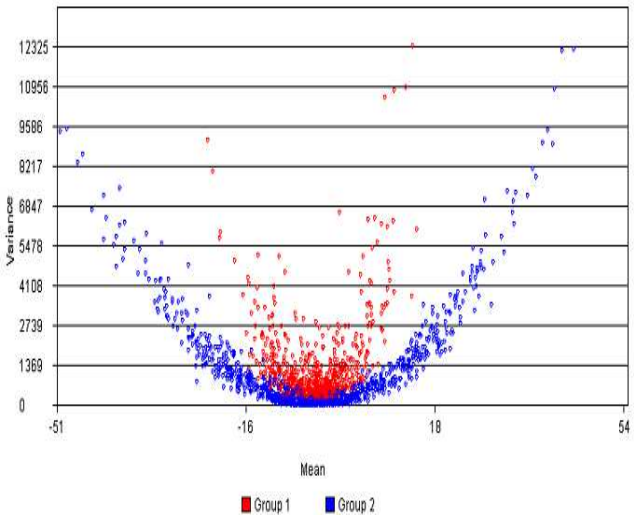

Fig. 3: Relationship between means of both the groups of differentially expressed genes and genes not expressed & their respective variances

The coefficient of variation between arrays before normalization was 17.577 and after normalization was 0.172. This is a very important graphical output as this helps in calculating the differentially expressed genes.

MA plot was constructed by depicting M= log (mean1/mean2) on Y-axis and A= log (mean1 x mean 2) on X- axis (Figure.4). The values of M range from -7 to -5 and the values of a range from -1 to -3. The correlation between these two variables was 0.38. The red region depicts the differentially expressed genes.

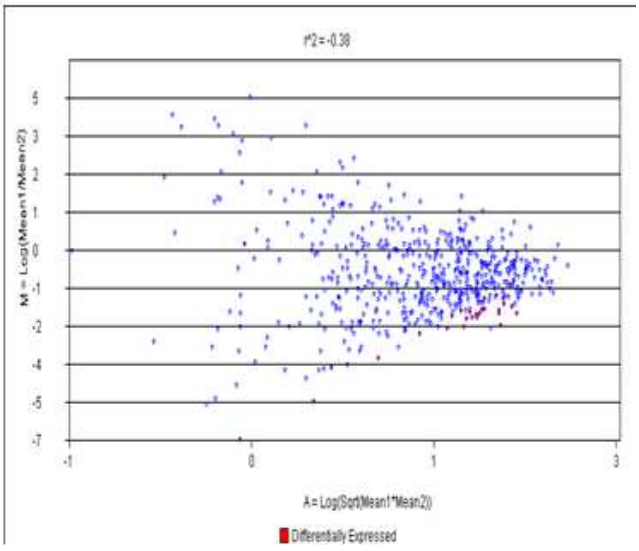

Fig. 4: MA values of differentially expressed genes

The score histogram is a bar chart graphic representation of the score distribution report (Figure.5). The scores are converted to a percentage on the horizontal axis and the frequency is plotted on the vertical axis. It represents the J5 test score which varies from -3.94 to 3.94. The score histogram of genes on the basis of scores depicts the under expressed and over expressed genes.

The confounding indexes of 64 retained genes before normalization and after normalization were -9.029 and -2.027 respectively.

The next output of the result was picture showing the mean histogram based on the mean of the data (Figure 6). The mean values were obtained based on the gene values. The means of group 1 and group 2 were grouped into two groups, i.e. under expressed and over expressed genes.

The most important of all outputs is the expression pattern grid (Figure 7) which depicts the genes which are differentially expressed in a manner which covers the entire region of genes.

3.1. Expression Pattern Grid: The Gene Expression Pattern Grid, which is generated for any set of differentially expressed genes with the GEDA web application, summarizes the types of differential expression in a way that is related to the PPST test. Black signifies that the expression intensity of a given gene in a given sample is within the upper and lower 95th percentile boundaries of the alternate group. On screen green signifies that the expression intensity is below the lower 95th percentile of the alternate group, and red signifies that the expression value is above the upper 95th percentile of the alternate group (red = overexpression; green = underexpression).

This representation includes information on both the population-level informativeness as well as which individuals appear to exhibit uniquely differentially expressed profiles. Samples within sample group are arranged according to their relative position in a hierarchical agglomerative clustering with pair wise distance = 1-Pearson's correlation coefficient.' Not expressed' is a hypothesis generated in these graphs when the expression intensity value of that gene for that individual falls in the lower 95th %tile of the entire data set. Expression pattern grids were produced online with the Gene Expression Data Analysis web application http://bioinformatics.upmc.edu/.

3.2. Differentially expressed genes: After using the different steps discussed above, here we name some genes which are differentially expressed and is the putative cancer causing genes:

3.2.1. EPAS1: Endothelial pas domain protein 1 which is also designated as HIF-1 alpha like factor (HLF) is a recently identified Bhlh-PAS protein that shares 48% sequence identity with Hlf-1 ALPHA. The EPAS1 provides novel insight into the regulatory mechanisms of EPAS1 gene expression that may contribute to the adaptation of tumor cells under the

hypoxic environment. [15] EPAS1 activates its own promoter and seems to be regulates. EPAS1 has been identified as a member of the basic helix -loop -helix (bHLH) PAS protein family and plays a critical role in the regulation of hypoxia inducible genes.

3.2.2. Polo like Kinase (PLK 1): PLK is involved in targeting cyclin B1, to the nucleus (Toyoshima-morimoto et.al.) PLK 1 is over expressed in human tumors and has prognostic potential in cancer, indicating its involvement in carcinogenesis & its potential as a therapeutic target. The first identified PLK 1 family member was Polo, which is a serine/threonine kinase, in Drosophila melanogaster. PLK 1 is associated with tumorigenesis and belongs to a family of disease relevant protein kinases that can be targeted by different drugs. It represents a promising approach for the development of novel anticancer therapies. PLK1 activity is suppressed and cell cycle arrests to repair the damaged DNA which is considered as novel prognostic marker for several tumors. NSLC patients whose tumour showed that high PLK1 expression had poorly curative effect, and were resistant to chemotherapy, which implied that PLK 1 might be associated with drug resistance.

3.2.3. Metallo proteinase domain 12: These are a family of proteolytic enzymes with over 20 members that break down proteins in the extracellular matrix. MMPs are regulated by specific inhibitors known as the tissue inhibitors of metallo proteinases (TIMPs). These are highly expressed in many different cancer types including lung cancer. The importance of MMP12 in human lung has however been shown by its role in the development of emphysema caused by cigarette smoke The MMP12 polymorphism is located in the coding region of the hemopexin domain that is responsible for MMP12 activity

3.2.4. Trophinin (TRo-tastin): Trophinin was identified as one of the best correlated genes. The trophinin expression in lung cancer specimens were examined by immunohistochemical staining. The role of trophinin in cancer metastasis was further investigated by approaches of over expression and knock down with small interfering RNA (siRNA). Over expression of trophinin increases cell invasion ability and knock down with siRNA inhibits cell invasion trophinin could enhance cell invasion as a novel prognostic factor for early stage lung cancer. Trophinin expression is high in human embryonic tissue as well as lung cancer and is undetectable in normal adult tissue.

3.2.5. Metallothionien: Metallothioniens are a group of low molecular weight (6000-7000 Da), cysteine rich (30%) intracellular proteins with high affinity for certain metals but no known enzymatic activity. This protein is rich in sulphydryi groups and thus is an excellent candidate for attacks by electrophiles, such as the platinum drugs, resulting in drug resistance. The protein MT plays an important role in Cd detoxification and it has been suggested that differential inducibility of pulmonary may lead to interspecies susceptibility. Recent *in vitro* studies have suggested that modulation of MT synthesis in certain tumors may provide a promising protocol to enhance the efficacy of drugs in patients undergoing chemotherapy for malignant diseases.[46,43] MTs also have an important role in carcinogenesis because of their over expression in variety of tumors. A high cancer of MT occurs during early development in variety of tissues, including liver, lung tissues, with levels declining to very low concentrations adult live MT expression in lung tumors has not been evaluated adequately and there is only one published study in the literature. [18]

3.2.6. Pleoitrophin (PTN): Pleoitrophin is a heparin binding growth factor involved in the differentiation & proliferation of neuronal tissue during embryogenesis, and also secreted by melanoma & carcinoma cells. Pleoitropin is very basic protein of an apparent mass of 1&kda, which is differentially expressed during pre-and post natal development .During embryogenesis PTN is strongly expressed in brain, liver, Spleen , lung, bone & tongue is strongly expressed in human lung cancer lines, particularly in those are lines derived from small lung cancer . PTN might be a prognostic factor for lung cancer and larger prospective further studies do required to confirm this hypothesis.

3.2.7. Thrombomodulin: Thrombomodulin which is a receptor for thrombin on the surface of the vascular endothelial cells neutralizes thrombin and the formed thrombin- TM complex activates protein C. TM is not only a thrombin receptor but also an onco developmental antigen, which is found in lung cancers.[22]**.**TM plays an important role in the regulation of intravascular coagulation by exerting the inhibitory activity on thrombin induced platelets aggregation. TM expression in the lung cancer cells appears to vary depending on the cellular conditions. It can be roughly speculated that functionally active TM on lung cancer cells may modulate the biological behaviors of these cells, such as invasiveness and metastatic potential.

3.2.8. GPRC5A: GPRC5A is a member of G-coupled receptors, which was originally identified as an all trans-retinoic acid-induced gene. Although recent studies reported that this gene was highly expressed in the cancer cell lines and that GPRC5A might positively regulate cell proliferation, its

mechanism remains unknown, GPRCSA gene, which is under expressed in human lung cancer, suppresses lung tumors in mouse models & could provide a key to attacking lung cancer in humans. The researchers compared 186 lung tumors to 17 normal tissues using gene expression profiling with a micro array. The four tumor types all had a fraction of GPRC5A gene expression shown in the normal cells: adeno carcinoma 46.2. In the end they inserted the GPRC5A gene back into lung cancer lines in a laboratory experiment & suppressed colony formation of human lung cancer by 91% into cell lines. [52,29]**.**Further study substaining the role of GPRC5A gene in human lung cancer could lead to the development of novel approaches for lung cancer prevention diagnosis & treatment.

3.3. PUTATIVE CANCER CAUSING GENES:

In the previous section we named only a few genes which may be responsible for development of the disease. A large series of evidence has conclusively demonstrated that the development and progression of cancer are due to the accumulation of a number of genetic alterations which finally result in a final malignancy states one of the main point emerging from the investigations is that most tumor types show genetic aberrations modified in the protein engine which allows cells to divide correctly.

One of the basic goals in the analysis of microarray gene expression data is the identification of differentially expressed genes in the comparison of different types of cell or tissue samples. In order to control the biological and experimental variability of the measurements, statistical inference has to be based on an adequate number of replicate experiments.

For the following, we assume that the data are given either as absolute intensities or as relative values with respect to a common reference sample, and have been calibrated. To identify differentially expressed genes with respect to a certain biological question, a suitable statistical test may be performed for each gene [12]**.**The choice of the test statistic depends on the biological question and on the nature of the available experimental data. Microarrays have an important role in finding novel drug targets; the thinking that guides the design and interpretation of such experiments has been expressed by (Lonnstedt and Speed, 2002.) [33]

The number of genes selected would depend on the size, aim, background and follow-up plans of the experiment." Often, interest is restricted to so-called 'druggable' target classes, thus thinning out the set of eligible genes considerably. One of the basic goals in the analysis of microarray gene expression data is the identification of differentially expressed genes in the comparison of different types of cell or tissue samples. In order to control the biological and experimental variability of the measurements, statistical inference has to be based on an adequate number of replicate experiments. Microarray based experiments are frequently seen as the stronghold of hypothesis-free genome research. While debatable in itself, this assertion simply shifts the responsibility to the computational scientist analyzing the data. In the absence of a clear hypothesis much of the analysis will be of an exploratory nature. Once this leads to a hypothesis further independent verification is needed. This embeds microarray experiments and statistical analysis into a feedback cycle producing new experiments.

Gene expression profiling provides insight into the functions of genes at a molecular level. Microarray technology measures the relative activity of previously identified target genes. For understanding the disease network fundamentals of lung cancer, analysis of gene expression profiling data derived from micro-array technology was done. DNA microarray technology has been widely hailed as a powerful tool to study the global gene expression in organisms or tissues. Microarray can be applied to a wide range of studies including gene regulation, disease diagnosis and prognosis, cancer classification, bio-marker discovery and drug development. The microarray's capacity to compare gene expression patterns in different tissues or conditions threatens to change the way biology is practiced and understood.

Also we know that cancer is not the effect of single gene translocations. It involves the fusion of the genes which is a result of translocations of the chromosomes. These include finding out the analysis of the genes in consideration and find out the basis of these translocations. If we get to the core of these translocations and chromosomes aberrations we will be able to generate a putative method for finding out the targets for the drug discovery.

With the help of our procedure using the microarray method we are able to identify genes that may be the possible cause of the disease and thus would establish the basis of drug designing in our next step of analysis.

Recently two extensive studies on different histological types of lung cancer using high-density microarrays were published. [5,19]**.** In both of these studies the different types of lung cancer could clearly be separated according to gene expression profiles by hierarchical clustering. Different survival for patients in different clusters was also

demonstrated. Besides, a number of smaller array studies have also been published which are conducted mostly on cell culture level and investigate different aspects of lung cancer, including metastatic potential and classification [11].

A better framework of significance inference includes calculation of a statistic based on replicate array data for ranking genes according to their possibilities of differential expression and selection of a cut- off value for rejecting the null –hypothesis that the gene is not differentially expressed [38,40].

Micro array experiments are the most abundant form of gene expression data. In summary, we propose some of the differential expressed genes like EPAS1, Polo like Kinase (PLK 1) , Metallo proteinase domain 12 , Trophinin (TRo-tastin), Pleoitrophin (PTN),GPRC5A. [8] .The EGFR pathway is one of the most extensively studied signalling pathways relevant to cancer and the EGFR interactome is one of the most well-described interactomes, both biochemically and theoretically . Thus, it is likely that the poor lung cancer response rates may be due, at least in part, to a too homogeneous treatment approach used in the past for a highly heterogeneous disease [7]. Only recently, lung cancer heterogeneity has started to gain therapeutic relevance, as documented by the attempt to propose preferential chemotherapeutic options for each tumor histotype[53,47].

By harnessing the power of genomic research, this pioneering work has painted the clearest and most complete portrait yet of lung cancer's molecular complexities. "This big picture perspective will help to focus our research vision and speed our efforts to develop new strategies for disarming this common and devastating disease. Microarray data projects include data from pilot studies, basic research (in vitro and model animal), pre-clinical studies and large-scale clinical trials. Cancer microarray data normally contains a small number of samples which have a large number of gene expression levels as features. To select relevant genes involved in different types of cancer remains a challenge. [55].Each new study reporting results from microarray experiments is accompanied by new methods of analysis. These creative efforts have led to a wealth of methods but a dearth of comprehension on the comparative performance of these methods.

Tests for identifying differentially expressed genes should be as powerful (sensitive) and yet as specific as possible. There is a growing demand for easy- to use tools for the analysis and interpretation of global patterns of gene expression from microarray experiments. While commercial platforms are available, most are specially designed for data from one platform.

This study had been a stepping towards the analysis of data by using micro array technology which is a powerful approach for genomics research. It is important to understand the crucial steps that can affect the outcome of the analysis of the contemporary trends on various main analysis steps in the microarray data analysis process which includes experimental design, data standardization, image acquisition and analysis, normalization, statistical significance inference, exploratory data analysis class prediction and pathway analysis, as well as various considerations relevant to their implementation. By carrying out this analysis we are able to find out the putative drug targets for the disease and this approach will be increasingly useful and this work can be taken further for purpose of analysis.

4. CONCLUSION

This study shows that a total of 916 genes are expressed in lung cancer samples studied. The 64 genes are retained for showing under and over expression under disease conditions. These genes were segregated using J5 Test, T-test, and calculation of P values. An interesting potential application of the expression pattern grid and the study of modes of differential expression is the search of potential gene deletions (samples in group A down regulated compared with samples in group B) . Crossed-out boxes in the expression pattern grid provide us a visual aid in formulating hypothesis about deletion events. Furthermore the gene expression pattern grid is useful in identifying false positives that occur due to outliers. Various features of the analysis carried out by caGEDA are described, and more information on its implementation is available on (http://bioinformatics.upmc.edu/Help/GEDADescription.html). Using the generalized micro-analysis, we detected a number of putative targets for some known chemical compounds. Lung Cancer is a widespread disease and finding out the genes which may be responsible for its effect is one of the preliminary steps of investigation. It is possible to go about with the process of drug designing and thus establishing remarkable step for the treatment of cancer.

REFERENCES:

1. Allander, S.V., Illei, P.B., Chen, Y., et al 2002. Expression profiling of synovial sarcoma by cDNA microarrays: association of ERBB2, IGFBP2, and

ELF3 with epithelial differentiation. Am. J Pathol, 161, 1587-95.

2. Alves, F., Vogel, W., Mossie, K., Millauer, B., Hofler, H., Ullrich, A., 1995. Distinct structural characteristics of discoidin subfamily receptor tyrosine kinases and complementary expression in human cancer. Oncogene, 10,609-618

3. Baldi, P., Long, A.D., 2001.A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 17, 509-519.

4. Beer, D.G., Kardia, S.L., Huang, C.C., et al., 2002. Gene expression profiles predict survival of patients with lung adenocarcinoma. Nat Med, 8, 816-24.

5. Bhattacharjee, A., Richards, W.G., Staunton, J., Li. C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker, D.J, Meyerson, M.2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.Proc Natl Acad Sci U.S.A.20; 98(24):13790-5.

6. Black, M.A., Doerge, R.W., 2002. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics, 18, 1609-1616.

7. Borczuk, A.C., Toonkel ,R.L., Powell, C.A., 2009.Genomics of lung cancer.Proc AmThorac, Soc 6,152-158.

8. Broderick, Stephen ,Yoshizawa , Akihiko F.,Riegel A., T., Wellstein A., 1994.Ribozyme-targeting elucidates a direct role of pleiotrophin in tumor growth. J. Biol. Chem., 269, 21358-21363.

9. Chan,W.C., Huang, J.Z., 2001. Gene expression analysis in aggressive NHL. Ann Hematol, 80(Suppl 3):B38-41.

10. Chang- Tze Ricky YU, Jung Mao HSU, Yuan-Chll Gladys Lee , Ann Ping Tsou, Chen-Kung Chou and Chi Ying F., etc 2005, phosphorylation & stabilization of HURP by Aurora–A: implication of HURP as a transformation target of Aurora. A. Mol Cell Biol Bid, 25, 5789-800).

11. Churchill, G.A., 2002. Fundamentals of experimental design for cDNA micro arrays. Nat. Genet. 32 (suppl.2) 490-495.

12. Claverie, J.M., 1999.Computational methods for the identification of differential and coordinated gene expression. Human Molecular Genetics., 8, 1821–1832.

13. DeRisi et al 1996., Welford et al 1998, Ding, Li; Getz, Gad; Wheeler modulation of somatic mutations affect key pathways in lung adenocarcinoma

14. DeRisi, J.L, Iyer V.R, Brown P.O. 1997. Exploring metabolic and generic control of gene expression on a genomic scale. Science, 24, 680-686.

15. Folkman, J., 1975. Tumor angiogenesis: a possible control point in tumor growth. Ann. Intern Med. 82(1):96-100.

16. Garber, M.E., Troyanskaya O.G., Schluens K, et al.2001. Diversity of gene expression in Adenocarcinoma of the lung .Proc Natl Acad Sci USA. 98, 13784-9.

17. Gildea, J.J, Seraj M.J., Oxford, G., et al 2002. RhoGDI2 is an invasion and metastasis supressor gene in human cancer. Cancer Res, 62,6418-23.

18. Hart, B.A., Voss, G.W., Vacek P.M., 1933. Metallothionine in human lung carcinoma .Cancer Letters Volume 75,121-128

19. Hong, T.M., Yang, P.C, Peck, K., Chen, J.J., Yang ,S.C., Chen, Y.C., Wu C.W., Am J Respir, 2000. profiling the downstream gene of tumour suppressor PTEN in lung cancer cells by complementary DNA microarray. Cell Mol. Biol. vol. 23 no. 3 , 355-363.

20. Husgafvel-Pursiainen ,K., Hackman, P., Ridanpaa, M., Anttila, S., Karjalainen ,A.,Partanen, T., Taikina-Aho,O., Heikkila, L., and Vaino, H.,1993. K-ras mutations in human adenocarcinoma of the lung: association with smoking and exposure to asbestos. Int J Cancer 53, 250–256

21. Ideker T, Thorsson V, Siegel AF, Hood LE, 2000. Testing for differentially- expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol, 7:805-817.

22. Imada, S., Yamaguchi, H., Nagumo, M., Katayanagi, S., Iwasaki, H., Imada, M.,1990. Identification of fetomodulin, a surface marker protein of fetal development, as thrombomodulin by gene cloning and functional assays. Dev Biol. 1990 Jul; 140(1):113-22.

23. Inoue, H., Matsuyama, A., Mimori, K ., et al 2002. Prognastic score of gastric cancer determined by cDNA microarray. Clin Cancer Res, 8,3475-9.

24. James Lyons-Weiler, Satish Patel, Michael, J., Becich ,Tony, E., Godfrey. 2004. Tests for finding complex patterns of differential expression in cancers: towards individualized medicine: BMC Bioinformatics 2004, 5:110.

25. Jayakanthan, M., Gulshan Wadhwa, Madhan Mohan, T., Arul,L., Balasubramanian P., and Sundar,D., 2009. "Computer-Aided Drug Design for Cancer-Causing H-Ras p21 Mutant Protein" Letters in Drug Design and Discovery, 6, 14-20.

26. Keohavong, P., DeMichele, M.A.A., Melacrinos, A.C., Landreneau, R.J., Weyant, R.J., and Siegfried, J.M., 1996. Detection of K-ras mutations in lung carcinomas: Relationship to prognosis. Clin Cancer Res 2, 411–418.

27. Kerr, M.K., Martin, M., Churchill, G.A., 2000. Analysis of variance for gene expression microarray data. J Comput Biol, 7,819-837.

28. Kihara, C., Tsunoda, T., Tanaka, T., et al 2001. Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by c DNA microarray analysis of gene expression profiles. Cancer Res, 61,6474-9.

29. Kumar, S., Gadagkar, S.R., 2000. Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. J Mol Evol, 51,544-53.

30. Leung, Y.F.,et. al, 2002. Micro array software review. In a practical approach to micro array data analysis (Berrar, D.P.et al., eds), kluwer academic.

31. Li ,S.,Ross, D.T., Kadin, M.E., et al 2001. Comparative genome scale analysis of gene expression profiles in T-cell lymphoma cells during malignant progression using a complementary DNA microarray. Am J Pathol,158,1231-7.

32. Lockhart, D.J., Dong, H., Byrne, M.C., et al 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 13, 1675-80.

33. Lonnstedt, I., Speed, T., 2002. Replicated microarray data.Stat Sinica Volume 12, 12, 31-46

34. Luo, J.H., Yu Y.P., Cieply, K., et al 2002. Gene expression analysis of prostate cancers, Mol Carcinog, 33, 25-35.

35. Lyons-Weiler ,J., Patel, S., Ngyuen, T., et al. 2004. UPITT Cancer Gene expression Data set Link Database, maintained by the University of Pittsburgh Benedum Oncology Informatics Centre [online]. Accessed 12 Apr 2004. URL: http://bioinformatics.upmc.edu/Help/UPITTGED.html

36. Lyons-Weiler., J.,Patel, S.,Bhattacharya, S., 2003. A classification based machine learning approach for the analysis of genome wide expression data. Genome Res,13,503-12.

37. Nocito, A., Bubendrof, L., Maria Tinner, E., et al 2001. Microarrays of bladder cancer tissue are highly representative of proliferation index and histological grade. J Pathol, 194,349-57.

38. Novak, J.P.,et al., 2002. characterization of variability in large – scale gene expression data: implication for study design .Genomics 79, 104-113,

39. Okutsu, J., Tsunoda, T., Kaneta, Y., et al 2002. Prediction of chemosensitivity for patients with acute myeloid leukaemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis. Mol Cancer Ther, 1,1035-42

40. Pritchard, C.C.,et al., 2002. Project normal: defining normal variance in mouse gene expression. Proc. Natl. Acad.Sci.U.S.A.98, 13266-13271

41. Ramaswamy, S., Ross ,K.N., Lander, E.S., et al. 2003. A molecule signature of metastasis in primary solid tumors. Nat Genet, 33,49-54.

42. Rickman, D.S., Bobek ,M.P., Misek ,D.E., et al 2001. Distinctive Modular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res, 61,6885-91.

43. Saltoh, M.,cherin, chein. M.A., Imara, N., Shimiza, 1994. Modulation of resistance to anticancer drugs by inhibition of metallothionine synthesis cancer .Cancer Res. Oct 15;54(20),5255-7.

44. Saravanakumar, Selvaraj, Jeyakumar Natarajan, 2011. Microarray Data Analysis And Mining Tools : Bioinformation 6(3): 95-99.

45. Sasaki, H., Ide N., Fukai, I., et al 2002. Gene expression analysis of human thymoma correlates with tumor stage. Int J Cancer, 101,342-7

46. Satoh, M., Kloth ,D.M., Kadhim, S.A, Chin ,J.L., Naganuma ,A., Imura,N., Cherian, M.G., 1993. Modulation of both cisplatin nephrotoxicity and drug resistance in murine bladder tumor by controlling metallothionien synthesis.Cancer Res. Apr 15;53(8),1829-32.

47. Scagliotti, G.,Hanna, N., Fosella, F., Sugarman ,K., Blatter, J.,Peterson ,P .,et al. 2009. The differential efficacy of pemetrexed according to NSCLC histology: a review of two phase III studies. Oncologist 14,253-263.

48. Scharpf, R.B., Tjelmeland H., Parmigiani, G., Nobel,A.B., 2009. “A Bayesian model for cross-study differential gene expression. J Am Stat Assoc. 104(488):1295-1310.

49. Schena ,M., Shalon, D., Davis ,R.W., et al. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270,467-70

50. Shaik JS, Yeasin M. 2007. A Unified Framework For Finding Differentially Expressed Genes From Microarray Experiments : BMC Bioinformatics Vol.18;8:347.

51. Shih, J.Y., Yang, S.C., Hong, T.M., et al.2001. Collapsin response mediator protein-I and the invasion and metastasis of cancer cells.J Natl Cancer Inst, 93,1392-400.

52. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res, 15,4876-4882.

53. Tiseo ,M.,Bartolotti, M., Gelsomino, F., Ardizzoni ,A., 2009. First line treatment in advanced non small lung cancer: the emerging role of the histologic subtype. Expert Rev Anticancer Ther 9,425-435.

54. Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R.S., 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol, 8,625-637.

55. Yu ,Wang, Igor,V., Tetko, Mark ,A., Hall, Eibe Frank, Axel Facius , Klaus ,F.X., Mayer, Hans W., Mewes., 2005. Gene selection from microarray data for cancer classification—a machine learning approach. Computational Biology and Chemistry. Volume 29, Issue 1, Pages 37–46

[graphical figures listed below]

Fig. 5 Score histogram of over and under expressed genes

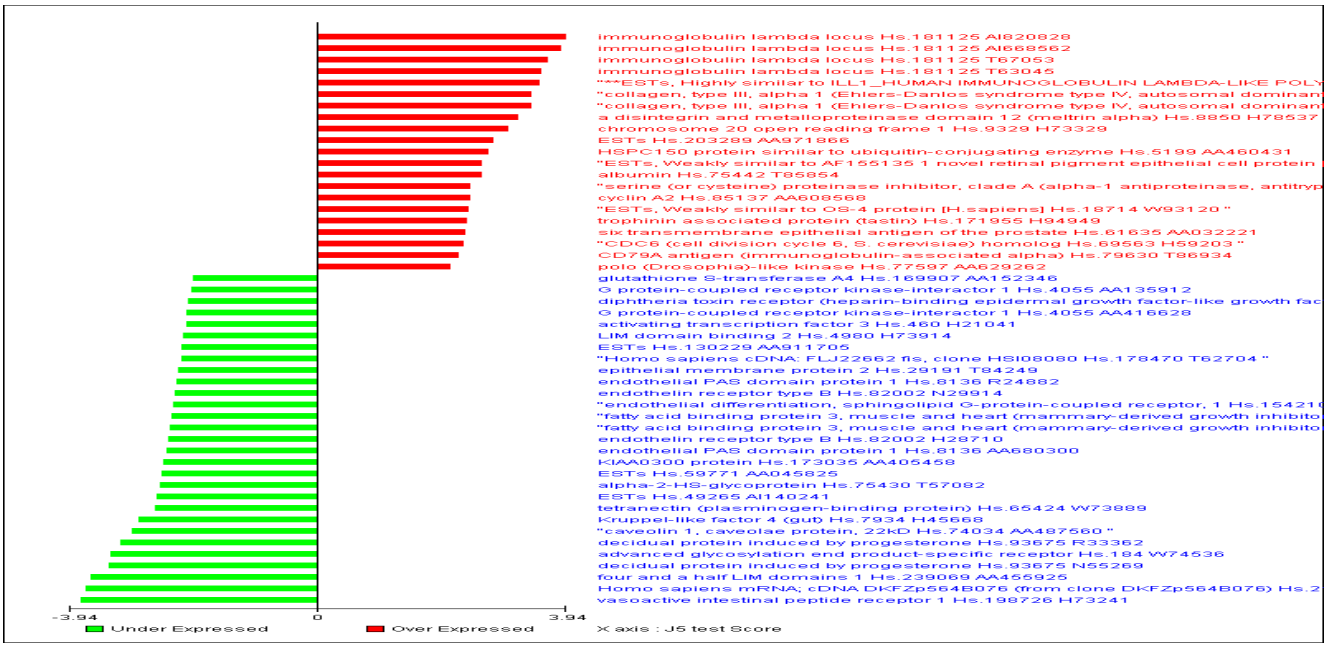

Fig. 6 Mean histogram based on the mean of the data of over expressed and under expressed genes of group 1 & 2.

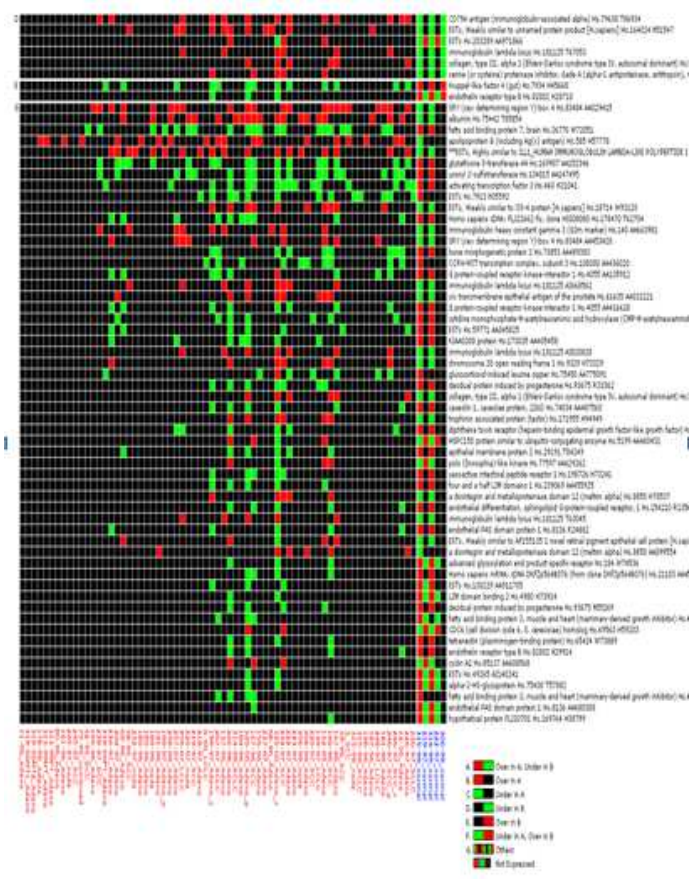

Fig. 7 Expression pattern grid of lung cancer