International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 8, August - 2014

A Novel Approach for Cigarette Smoke-Induced Differential Gene Expression in Blood Cells from Monozygotic Twin Pairs

Akshatha M[1]
Bioinformatics (BBI)
GM Institute of Technology
Davangere, India.

Prof. Manjunath Dammalli[3]
Department of Biotechnology
Siddaganga Institute of Technology
Tumkur, India.

Prof. Dr. H. Gurumurthy[2]
HOD, Dept. of Biotechnology
GM Institute of Technology
Davangere, India.

***Abstract*— Cigarette smoke is a well known source of chemical carcinogens and comprises a complex mixture of over sixty proven, probable and possible carcinogenic agents. Cigarette smoking alone is directly responsible for approximately thirty percent of all cancer deaths. Microarray data analysis provides information on disease and therapeutic approaches for cancer. Samples are downloaded from GEO database and imported to Bioconductor R to do statistical analysis. Expression analysis is performed to Identify and rank common differentially expressed genes in smokers compared to non-smokers, responsible to cause an incredible variety of cancer.**

Keywords— Cigarette smoking, Tobacco, Monozygotic twins, Bioconductor R, gene expression profiles, Cancer.

I. INTRODUCTION

A good health demands a healthy environment. Over the last decades it has been a major interest to understand who and what extent human influence or pollute, their environment and how in turn, the environment influences the human health. The development of civilization and the emergence of industries over the preceding centuries greatly contributed to the increased presence of complex mixtures of hazardous toxic and carcinogenic chemicals in the human environment, many of which are suspected or have been proven to cause cancer. Human exposure to carcinogens may result from many sources, related to the general environment, occupational settings and dietary or life style habits(e.g.: cigarette smoking).Today cancer is one of the leading cause for the death worldwide, with annually increasing number of patients[1]. The word cigare is derived from the Mayan word sikar which means to smoke. The word tobacco is derived from a Spanish word tobaca which is a Y-shaped instrument used by early American Indians to inhale snuff. The term to smoke was introduced during the late sixteenth century by Sir Walter Raleigh to UK [2].

Tobacco smoking has been in vogue for hundreds of years. Chemical carcinogenesis induced by lifestyle factors like cigarette smoking is a major research area in epidemiology. Cigarette smoke is a well known source of chemical carcinogens and comprises a complex mixture of over 60 proven, probable and possible carcinogenic agents.

Monozygotic twin pairs are referred to as identical twins, but they are not perfectly identical, although they can be very similar. Monozygotic twins DNA is very similar, but there are differences, such as copy number variations that indicate a difference in the number of copies of certain parts of the DNA. These differences in DNA can result in slight differences in appearance as well as differences in other physical characteristics, medical conditions or susceptibility to certain diseases. Cigarette smoking alone is directly responsible for approximately 30% of all cancer deaths. Cigarette smoking also contributes to lung disease, heart disease, stroke, and the development of low birth weight babies. Cigarette smoking causes 87% of lung cancer deaths. Lung cancer is the leading cause of cancer death in both men and women. The health risks caused by cigarette smoking are not limited to smokers - exposure to second-hand smoke, or environmental tobacco smoke ETS (environmental tobacco smoke), significantly increases a non-smoker's risk of developing lung cancer[1] [2].

Cancer is a public health problem worldwide. It affects all people from the young to the old, the rich to the poor, men, women and children. Of the several causes investigated for cancer, the use of tobacco has shown strong and consistent associations with cancer at several sites of the body. Presently, more than 10 million people globally are diagnosed with cancer every year due to smoking. It is estimated that by 2020, there will be 15 million new cases every year. Cancer causes 6 million deaths every year, or 12% of deaths worldwide [4]. Microarrays generate large

amounts of numeric data that should be analyzed effectively. Microarrays are used to measure gene expression levels in different ways. An experiment is designed by which the microarray experiment is carried out and data are generated. The analysis of microarray data to produce lists of differentially expressed genes has several steps which can differ based on the type of data being assayed. However, all data follows the same general pipeline which involves reading raw data, quality assessing the data, removing bad spots/arrays from further analysis, preprocessing the data and calculating differential expression by statistical analysis.

II.MATERIALS AND METHODS

The current study planned to take gene expression data from Gene Expression Omnibus (GEO) and aimed to get genes and biomarkers which were not found by the studies conducted previously.

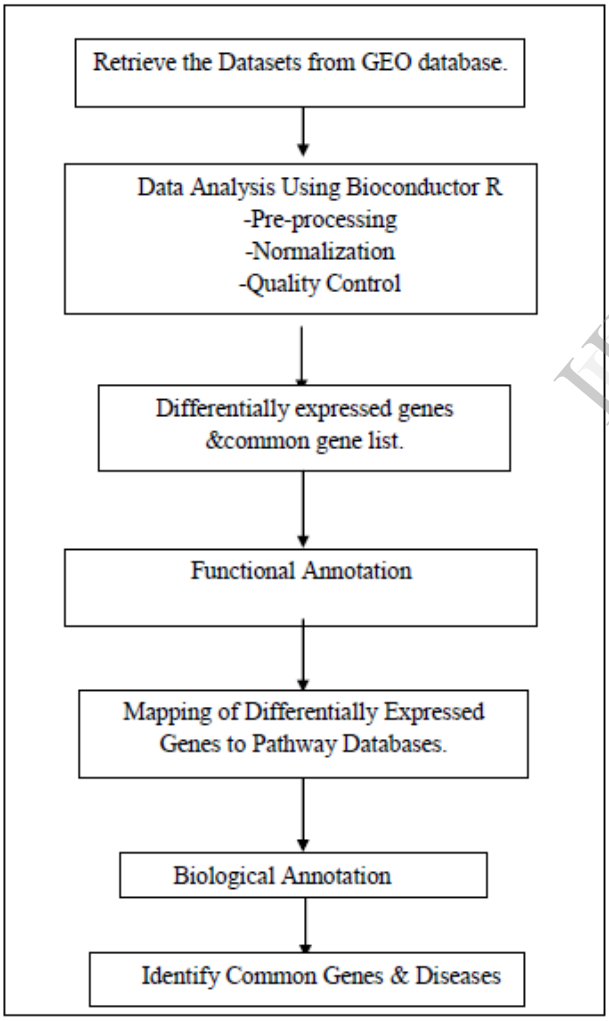

Figure 1: Methodology of Flowchart

The Gene Expression Omnibus is a microarray database that allows users to download experiments and curated gene expression profiles provided by NCBI. The samples for different smokers and non-smokers are downloaded from this database in order to perform expression analysis. The series matrix of the sample is downloaded and they are saved in ZIP/winRAR format. Bioconductor R is used to preprocess and normalize data. jvenn is a an integrative tool for comparing lists with Venn Diagrams. The Database for Annotation, Visualization and Integrated Discovery (DAVID**)** provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. Cytoscape is an open source software platform for visualizing molecular interaction networks and biological pathways. profiler used for functional profiling of gene lists from large-scale experiments. GeneCards is a database of human gene that provides genomic, proteomic, transcriptomics, genetic and functional information on all known and predicted human genes KEGG pathway maps for biological interpretation of higher-level systemic functions.

III. RESULTS AND DISCUSSION

Differential gene expression caused by cigarette smoking in blood cells from monozygotic twins discordant in smoking behaviour. There are some parameters to be considered in the selection of dataset that is the organism of the dataset should be Homo sapiens, experiment type should be expression profiling by array and the data sets should based on Affymetrix human genome platform. The dataset should be in CEL format and there should be more than 2 groups within the dataset.

Different cigarette smoker's samples are downloaded from GEO database in order to perform expression analysis. We only used the data sets based on Affymetrix human genome platform. That's, expression profiles for the dataset GSE30660, GSE7434, GSE3212,GSE12585,GSE4635 were downloaded including all samples.

R is free software which will be downloaded from the link. After a successful R installation, Bioconductor should be installed. After installing the required libraries and packages set the working directory. The datasets which are retrieved in Raw.tar format will be unzipped by using R statistical software into CEL files.

A. Pre-processing

In order to perform meaningful statistical analysis and inferences from the data, we need to ensure that samples are comparable. Systematic differences between the samples that are likely to be noise, biological variability should be removed. To examine and compare the overall distribution of the transformed expression values in the samples Box plot are used.

Affymetrix CEL-files contain slightly processed raw data of these probe intensities. Functions for reading Affymetrix data are available in the package affy. The function ReadAffy() reads in the raw data files, and stores

the data as an AffyBatch object. By default, all CEL-files in the same directory are read. Next it compares the overall distribution of the transformed expression values in the samples.

B. *Normalization*

Normalization is a broad term for methods that are used for removing systematic variation from DNA microarray data. Normalization makes the measurements from different arrays inter-comparable. Normalization is a process in which random elements of the datasets as well as background corrections will be done and it also calculates each of the expression. It is just one part of Affymetrix data processing before estimates of gene expression are ready for further analyses. If we compare the two box plots, differences is seen. If the sample seems quite different from others the dataset like this should be removed considering its bad quality sample.

C. *Quality* Control

The next step in the quality control is to check whether the overall variability of the samples reflect their grouping. This can be done by hierarchical clustering of the samples to see if the samples cluster in the groups we expect. The samples are grouped according to the number of datasets are available.GSE30660 dataset equal number of samples in both groups (4 in each). The groups will be specified with the name as group1 and group2.In order to check the random samples within the two groups, group2 will be subtracted with the group1. Than p-value which is a measure that allows us to control how big a proportion of false positives (genes that we think are differentially expressed but really are not) are willing to accept. Normally the p-value should be less than 0.1 to 0.25. In This significant p-value and adj. p-value has been set to 0.1.Next step is to set the column names for each of the sample after that colour code will be specified for the dataset. In the hierarchical clustering map Red color specifies Highly Expressed Genes and Green Color specifies Low Expressed Genes. Black line shows which of the samples are clustered (Grouped) to one another.

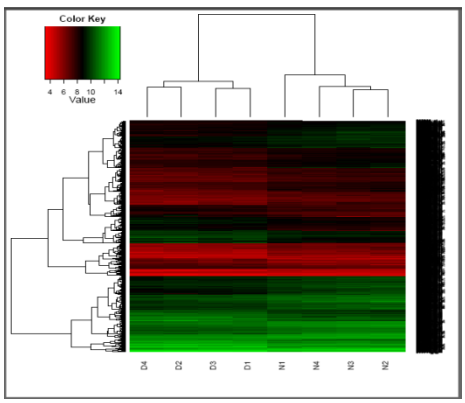

Figure 2: Hierarchical Clustering Map of GSE30660

D. *Differentially Expressed Genes*

Statistical tests are carried out that will be used to identify the genes that are differentially expressed between the two groups. The two significant p-values parameters are selected. For this analysis p-value is used which is a measure that allows us to control how big a proportion of false positives we are willing to accept. To do a more refined selection of the genes that we believe to do list differentially expressed genes. The significant p-values for the GSE30660 dataset has been set to 0.1. Often the results of microarray experiments are verified using other methods, and then we want to filter out genes that exhibit differences in expression that are so small that we will not be able to verify them with another method. This is done by adding one last criterion to the filter. Difference should have an significant value higher than 0 or lower than 0, as we are working with log transformed data, the group mean difference is really the fold change, so this filter means that we require a fold change above 0 and below 0. Note that the significant value > is important because the difference could be negative as well as positive. The result is that we end up with a list of genes that are likely candidates to exhibit differential expression in the two groups [7]. A number of summary statistics are computed for each gene. The log-fold change is the log expression level for that gene. The AveExpr is the average expression level for that gene across all the arrays and channels. The moderated t-statistic (t) is the ratio of the M value to its standard error. Each p-value has an adjacent p-value for each of the gene. The log odd statistics (B) is shown for each gene. Figure 5.6 & 5.7 shows the list of up-regulated and down-regulated gene list.

E. *Common* Gene List

Next step is to paste the probe-id of upregulated and down regulated id in each of the box. Name will be specified for each of the box, In order to identify which of the datasets have the common elements in the five datasets. Jvenn diagram shows which of the genes are overlapped with one another and the graph below the Jvenn shows number of probe-ids taken from each of the dataset.

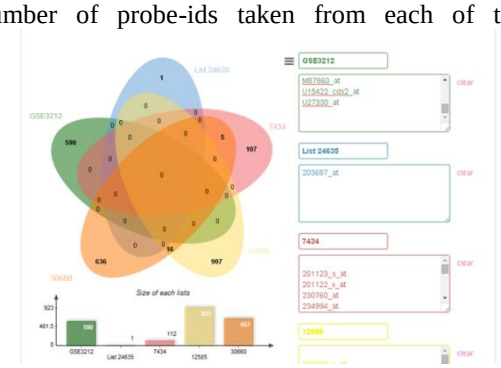

Figure 3: Venn diagram showing overlap among the five datasets upregulated and Downregulated genes in smokers

Than these probe-id s are mapped to DAVID open source database. To determine Gene name of each probe-id

Gene Accession Conversion Tool is used. Same method is followed for all the five datasets. In the DAVID database six random gene names are removed and it gives 101 gene names of differentially expressed genes. Below is the table showing the probe-id of both upregulated and Downregulated common genes along with their gene name, gene-id and from –to location along with the function of each gene.

The below table shows the Common Gene List.

Table 1: Common Gene List

205718_at	824764	Homo sapiens	integrin, beta 7	3695	Protein Coding
215125_s_at	780311	Homo sapiens	UDP glucuronosyltransferase 1 family, polypeptide A3	54659	Protein Coding
204798_at	777620	Homo sapiens	v-myb myeloblastosis viral oncogene homolog (avian)	4603	Protein Coding
210051_at	791880	Homo sapiens	Rap guanine nucleotide exchange factor (GEF) 3	10411	Protein Coding
224279_s_at	813712	Homo sapiens	calcium binding tyrosine-(Y)-phosphorylation regulated	26256	Protein Coding
205623_at	801467	Homo sapiens	aldehyde dehydrogenase 3 family, memberA1	218	Protein Coding
209900_s_at	784645	Homo sapiens	solute carrier family 16, member 1 (monocarboxylic acid transporter 1)	6566	Protein Coding
202436_s_at	809769	Homo sapiens	cytochrome P450, family 1, subfamily B, polypeptide 1	1545	Protein Coding
201266_at	816505	Homo sapiens	thioredoxin reductase 1; hypothetical LOC100130902	7296	Protein Coding
220137_at	803349	Homo sapiens	hypothetical protein FLJ20674	652266	Protein Coding
222222_s_at	809384	Homo sapiens	homer homolog 3 (Drosophila)	9454	Protein Coding
203826_s_at	800139	Homo sapiens	phosphatidylinositol transfer protein, membrane-associated 1	9600	Protein Coding
213474_at	820462	Homo sapiens	potassium channel tetramerisation domain containing 7	154881	Protein Coding
209921_at	809324	Homo sapiens	solute carrier family 7, (cationic amino acid transporter, y+ system) member 11	23657	Protein Coding
222304_x_at	773491	Homo sapiens	olfactory receptor, family 7, subfamily E, member 47 pseudogene	403253	Protein Coding
211889_x_at	776067	Homo sapiens	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)	634	Protein Coding
210007_s_at	818567	Homo sapiens	glycerol-3-phosphate dehydrogenase 2 (mitochondrial)	2820	Protein Coding
201194_at	795945	Homo sapiens	selenoprotein W, 1	6415	Protein Coding
212599_at	788447	Homo sapiens	autism susceptibility candidate 2	26053	Protein Coding
204956_at	781573	Homo sapiens	methylthioadenosine phosphorylase	4507	Protein Coding
220500_s_at	815611	Homo sapiens	RAB, member of RAS oncogene family-like 2B	11158	Protein Coding
222039_at	787488	Homo sapiens	kinesin family member 18B	146909	Protein Coding
202275_at	772717	Homo sapiens	glucose-6-phosphate dehydrogenase	2539	Protein Coding
202831_at	806970	Homo sapiens	glutathione peroxidase 2 (gastrointestinal)	2877	Protein Coding

F. Functional Annotation of Differentially Expressed Genes

The differentially expressed genes were mapped to their pathway. This gave the information about the genes and the pathway on which the gene acts. The total differentially expressed genes; Up regulated and Down regulated were mapped to DAVID open source database, this indexing will give curated evidence and confirmation of these genes as differentially expressed.

Following are the screen shots for DAVID database which is used for annotation of common probe id and finding various information. The annotation results show that the probe id list have three functional categories and three protein domains. And all the details of the differentially expressed genes are given in DAVID annotation table.

Following are the screen shots for DAVID database which is used for annotation of common probe id and finding various information. The annotation results show that the probe id list have three functional categories and three protein domains. And all the details of the differentially expressed genes are given in DAVID annotation table.

G. Functional Annotation of Differentially Expressed Genes

The differentially expressed genes were mapped to their pathway. This gave the information about the genes and the pathway on which the gene acts. The total differentially expressed genes; Up regulated and Down regulated were mapped to DAVID open source database, this indexing will give curated evidence and confirmation of these genes as differentially expressed.

243940_at	813490	Homo sapiens	teashirt zinc finger homeobox 2	128553	Protein Coding
219703_at	809716	Homo sapiens	meiosis-specific nuclear structural 1	55329	Protein Coding
220840_s_at	792307	Homo sapiens	chromosome 1 open reading frame 112	55732	Protein Coding
228115_at	800344	Homo sapiens	hypothetical LOC100271832; RNA, Ro-associated Y5 pseudogene 10; RNA, Ro-associated Y1;	6084	Protein Coding
213239_at	796499	Homo sapiens	progesterone immunomodulatory binding factor 1	10464	Protein Coding
1729_at	804117	Homo sapiens	TNFRSF1A-associated via death domain	8717	Protein Coding
203535_at	778820	Homo sapiens	S100 calcium binding protein A9	6280	Protein Coding
201853_s_at	789047	Homo sapiens	cell division cycle 25 homolog B (S. pombe)	994	Protein Coding
204587_at	776931	Homo sapiens	solute carrier family 25 (mitochondrial carrier, brain), member 14	9016	Protein Coding
205569_at	773205	Homo sapiens	lysosomal-associated membrane protein 3	27074	Protein Coding
210367_s_at	797658	Homo sapiens	prostaglandin E synthase	9536	Protein Coding
215920_s_at	790333	Homo sapiens	pyridoxal-dependent decarboxylase domain containing 2	283970	Protein Coding
203573_s_at	788003	Homo sapiens	Rab geranylgeranyltransferase, alpha subunit	5875	Protein Coding
205741_s_at	824370	Homo sapiens	dystrobrevin, alpha	1837	Protein Coding
203896_s_at	814023	Homo sapiens	phospholipase C, beta 4	5332	Protein Coding
212670_at	800879	Homo sapiens	elastin	2006	Protein Coding
220539_at	778705	Homo sapiens	chromosome 10 open reading frame 93; chromosome 10 open reading frame 92	54777	Protein Coding
218812_s_at	781005	Homo sapiens	ORAI calcium release-activated calcium modulator 2	80228	Protein Coding
200425_x_at	818500	Homo sapiens	ropporin, rhophilin associated protein 1B	152015	Protein Coding
206104_at	811857	Homo sapiens	ISL LIM homeobox 1	3670	Protein Coding
207469_s_at	800327	Homo sapiens	pirin (iron-binding nuclear protein)	8544	Protein Coding
201842_s_at	813598	Homo sapiens	EGF-containing fibulin-like extracellular matrix protein 1	2202	Protein Coding
203881_s_at	798170	Homo sapiens	dystrophin	1756	Protein Coding
217610_at	814789	Homo sapiens	speedy homolog E6 (Xenopus laevis)	25939	Protein Coding
215692_s_at	815566	Homo sapiens	metallophosphoesterase domain containing 2	744	Protein Coding
212324_s_at	805236	Homo sapiens	vacuolar protein sorting 13 homolog D (S. cerevisiae)	55187	Protein Coding
1559883_s_at	823405	Homo sapiens	SAM domain and HD domain 1	25939	Protein Coding
221122_at	774416	Homo sapiens	HRAS-like suppressor 2	54979	Protein Coding
204844_at	809624	Homo sapiens	glutamyl aminopeptidase (aminopeptidase A)	2028	Protein Coding
212745_s_at	785713	Homo sapiens	Bardet-Biedl syndrome 4	585	Protein Coding
213075_at	783648	Homo sapiens	olfactomedin-like 2A	169611	Protein Coding
201787_at	809878	Homo sapiens	fibulin 1	2192	Protein Coding
205462_s_at	775231	Homo sapiens	hippocalcin-like 1	3241	Protein Coding
204783_at	799959	Homo sapiens	myeloid leukemia factor 1	4291	Protein Coding
202094_at	798571	Homo sapiens	baculoviral IAP repeat-containing 5	332	Protein Coding
204026_s_at	783593	Homo sapiens	ZW10 interactor	11130	Protein Coding
219202_at	814361	Homo sapiens	rhomboid 5 homolog 2 (Drosophila)	79651,	Protein Coding
201422_at	791547	Homo sapiens	interferon, gamma-inducible protein 30	10437	Protein Coding
201467_s_at	781597	Homo sapiens	NAD(P)H dehydrogenase, quinone 1	1728	Protein Coding
213640_s_at	787120	Homo sapiens	lysyl oxidase	4015	Protein Coding
60815_at	795688	Homo sapiens	polymerase (RNA) II (DNA directed) polypeptide I4, pseudogene	84820	Protein Coding
219000_s_at	784402	Homo sapiens	defective in sister chromatid cohesion 1 homolog (S. cerevisiae)	79075	Protein Coding
218608_at	783390	Homo sapiens	ATPase type 13A2	23400	Protein Coding
213226_at	778615	Homo sapiens	cyclin A2	890	Protein Coding
227236_at	779383	Homo sapiens	tetraspanin 2	10100	Protein Coding
207643_s_at	804156	Homo sapiens	tumor necrosis factor receptor superfamily, member 1A	7132	Protein Coding
209043_at	797969	Homo sapiens	3'-phosphoadenosine 5'-phosphosulfate synthase 1	9061	Protein Coding
221188_s_at	797598	Homo sapiens	cell death-inducing DFFA-like effector b	27141	Protein Coding
221011_s_at	814619	Homo sapiens	limb bud and heart development homolog (mouse)	421301	Protein Coding
217234_s_at	816077	Homo sapiens	hypothetical protein LOC100129652; ezrin	7430	Protein Coding
219517_at	787988	Homo sapiens	elongation factor RNA polymerase II-like 3	80237	Protein Coding
203085_s_at	794775	Homo sapiens	transforming growth factor, beta 1	7040	Protein Coding
214157_at	809539	Homo sapiens	GNAS complex locus	2778	Protein Coding

H. Functional Annotation of Differentially Expressed Genes

The differentially expressed genes were mapped to their pathway. This gave the information about the genes and the pathway on which the gene acts. The total differentially expressed genes; Up regulated and Down

regulated were mapped to DAVID open source database, this indexing will give curated evidence and confirmation of these genes as differentially expressed.

Following are the screen shots for DAVID database which is used for annotation of common probe id and finding various information. The annotation results show that the probe id list have three functional categories and three protein domains. And all the details of the differentially expressed genes are given in DAVID annotation table. Again the file has to be imported to Cytoscape and the columns are set as the source and target. But now the selection of source will be probe-id and target will gene symbol. The network parameters are set to visualize the network. Set the node size and node color along with that edge size and edge color. In this node color is described from low to high i.e. from orange to red. The edges and nodes are shown in the network. The arrows between the nodes can be directed or undirected. The highly expressed genes are shown in red colour and the low expressed genes are shown in orange colour. The analysis of the network statistics are shown in the result.

probe_id	Ensemble id	gene symbol	gene name	platform
215125_S_AT	ENSG00000241119	UGT1A9	UDP glucuronosyltransferase 1 family, polypeptide A9 [Source:HGNC Symbol;Acc:12541]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000226232	RP11-419C5.2	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:B3VL09]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000242366	UGT1A8	UDP glucuronosyltransferase 1 family, polypeptide A8 [Source:HGNC Symbol;Acc:12540]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000244122	UGT1A7	UDP glucuronosyltransferase 1 family, polypeptide A7 [Source:HGNC Symbol;Acc:12539]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000242515	UGT1A10	UDP glucuronosyltransferase 1 family, polypeptide A10 [Source:HGNC Symbol;Acc:12531]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000267143	RP11-677D4.6	N/A	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000236696	PDXDC2P	pyridoxal-dependent decarboxylase domain containing 2, pseudogene [Source:HGNC Symbol;Acc:27559]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000254652	RP11-719N4.2	Protein LOC642778 [Source:UniProtKB/TrEMBL;Acc:E9PF93]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000241535	UGT1A3	UDP glucuronosyltransferase 1 family, polypeptide A3 [Source:HGNC Symbol;Acc:12535]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000240224	UGT1A5	UDP glucuronosyltransferase 1 family, polypeptide A5 [Source:HGNC Symbol;Acc:12537]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000234719	RP11-166K2.1	NIPP-like protein LOC729978 [Source:UniProtKB/Swiss-Prot;Acc:A6NJK4]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000167165	UGT1A6	UDP glucuronosyltransferase 1 family, polypeptide A6 [Source:HGNC Symbol;Acc:12538]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000214967	RP11-467M13.1	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:E9PJI5]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000214940	RP11-1212A22.2	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:E9PJI5]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000244474	UGT1A4	UDP glucuronosyltransferase 1 family, polypeptide A4 [Source:HGNC Symbol;Acc:12536]	AFFY_HG_U133_PLUS_2
215125_S_AT	ENSG00000241635	UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1 [Source:HGNC Symbol;Acc:12530]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000224712	RP11-719N4.1	Protein LOC642778 [Source:UniProtKB/TrEMBL;Acc:E9PN39]	AFFY_HG_U133_PLUS_2
215920_S_AT	ENSG00000183759	PYD1P1	NIPP-like protein 1 [Source:UniProtKB/Swiss-Prot;Acc:Q8WVD5]	AFFY_HG_U133_PLUS_2

I. List of Common Genes

For each of the gene Gene Ontology-id can be known in the Cytoscape. Select the highly expressed gene. When we select the gene UGT1A1 node color changes to yellow. Then select external links in that choose ontology click on Gene Ontology (Quick by name).it shows the list in which it gives information about a gene that in which of the species these genes are present.

J. Identified Common Genes and Diseases

In five dataset of monozygotic twin pairs discordant for smoking and non-smoking, we investigated whether cigarette smoking causes differential gene expression of toxicologically relevant genes in peripheral blood cells. By analyzing cigarette smoke-induced differential gene expression in monozygotic twins, we reduced the impact of interindividual variability due to variation in genetic background. It also provided the opportunity to perform pair-wise analyses, adding statistical power to the study. The analyses revealed several genes to be reproducibly differentially expressed due to cigarette smoking. Genes which were differentially expressed in smokers compared to non-smokers were identified by a combination Jvenn and Cytoscape software. The genes that were found by all approaches are considered the most discriminatively relevant genes. Resulting genes are shown in the Table 2 Identified Common Genes and Diseases.

Table 2: Identified common genes and diseases.

Gene Symbol	Gene Name	Diseases
UGT1A1	UDP glucuronosyltransferase 1 family,	Colorectal cancer, Stomach cancer, Liver cancer, colon cancer, Lung cancer, Thyroid cancer, Esophageal cancer, Breast cancer, Gastric cancer, Pancreatic cancer
UGT1A3	UDP glucuronosyltransferase 1 family, polypeptide A3	Lung cancer
UGT1A4	UDP glucuronosyltransferase 1 family, polypeptide A4	Lung cancer
UGT1A5	UDP glucuronosyltransferase 1 family, polypeptide A5	Colon cancer, Bone cancer
UGT1A6	UDP glucuronosyltransferase 1 family, polypeptide A6	Colorectal cancer, Ovarian cancer, Colon cancer, Gastric cancer, Breast cancer
UGT1A7	UDP glucuronosyltransferase 1 family, polypeptide A7	Pancreatic cancer, Colorectal cancer, Lung cancer, Colon cancer, Gastric cancer, Breast cancer
UGT1A8	UDP glucuronosyltransferase 1 family, polypeptide A8	Esophageal cancer, Colorectal cancer
UGT1A9	UDP glucuronosyltransferase 1 family, polypeptide A9	Colorectal cancer, Pancreatic cancer, Lung cancer, Colon cancer, Breast cancer
UGT1A10	UDP glucuronosyltransferase 1 family, polypeptide A10	Colon cancer, Lung cancer

CONCLUSION

Tobacco use causes a wide range of major diseases which impact nearly every organ of the body. The work investigated the genes which are differentially expressed in monozygotic twin pairs (smokers and non smokers). By analyzing both the up &down –regulated results, ninety six highly expressed genes and eighteen common genes are investigated.

ACKNOWLEDGMENT

The authors would like to say thank Siddaganga Institute of Technology, Tumkur, GM Institute of Technology, for their Technical support and, Davangere for their support and guidance.

REFERENCES

[1]. Pindborg, J.J., Mehta, F.S., Gupta, P.C., Daftary, D.K., Smith,C.J., Reverse smoking in Andhra Pradesh, India: A study of palatal lesions among 10,169 villagers, British Journal of Cancer., 1971.
[2]. http://www.medivisionindia.com/addiction/tobacco
[3]. http://www.pmusa.com/health_issues/
[4]. http://www.who.int/cancer/en/
[5]. Shah and Jatin,., Atlas of Clinical Oncology, Cancer of the Head and Neck, BC Decker Inc, 24 Edition, 2001.
[6]. Colin White., Research on Smoking and Lung Cancer: A Landmark in the History of Chronic Disease Epidemiology, The Yale Journal Of Biology And Medicine, 1989.
[7]. http://www.cancer.org/cancer/news/expertvoices/post/2013/01/02/light-smoking-as-risky-as-a-pack-a-day.aspx

[8]. Kolonen, S., Tuomisto, J., Puustinen, P., Airaksinen, M.M., Effects of smoking abstinence and chain-smoking on puffing topography and diurnal nicotine exposure, Pharmacol Biochem Behav., Vol **42** , 1992,pp 327–32.
[9]. International Journal of Environmental Research and Public Health.
[10]. Smith, C.J., Perfetti, T.A., Garg, R., and Hansch, C., IARC carcinogens reported in cigarette mainstream smoke and their calculated log P values, Food Chem. Toxicol., Vol 41, 2003, pp 807- 817.
[11]. Van Schooten, F.J., Hirvonen, A., Maas, L.M., De Mol, B.A., Kleinjans, J.C., Bell, D.A., and Durrer, J.D., Putative susceptibility markers of coronary artery disease: association between VDR genotype, smoking, and aromatic DNA adduct levels in human right atrial tissue, FASEB J., Vol 12, 1998, pp 1409–1417.
[12]. DeFlora, S., D'Agostini, F., Balansky, R., Modulation of cigarette smoke-related end-points in mutagenesis and carcinogenesis, Mutat. Res., 2003, pp 523–524.
[13]. http://www.biomedcentral.com/1755-8794/3/24
[14]. Hamadeh, H.K., Amin, R.P., Paules, R.S., And Afshari, C.A., An overview of toxicogenomics. Curr, Issues Mol. Biol., Vol 4, 2002, pp 45–56.
[15]. Toraason, M., Albertini, R., Bayard, S., Applying new biotechnologies to the study of occupational cancer—a workshop summary, Environ. Health Perspect, Vol 112, 2004, pp 413–416.
[16]. Wu, M.M., Chiou, H.Y., Ho, I.C., Chen, C.J. and Lee, T.C., Gene expression of inflammatory molecules in circula ting lymphocytes from arsenic-exposed human subjects, Environ. Health Perspect. Vol 111, 2003, pp 1429–1438.
[17]. Lampe, J.W., Stepaniants, S.B, S.B., Mao, M., Radich, J.P., Dai, H., Linsley, P.S., Friend, S.H. and Potter, J.D., Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol, Biomarkers Prev., Vol 13, 2004, pp 445–453.
[18]. Dougall, I., The Effect of Repeated Whole Cigarette Smoke Challenge on Human Air-Liquid Interface Lung Epithelial Culture,, Jul 2011.
[19]. Huuskonen, P.,[1], Storvik, M., Reinisalo, M., Honkakoski, P., Rysä, J., Hakkola, J., Pasanen, M., Microarray analysis of the global alterations in the gene expression in the placentas from cigarette-smoking mothers, Clin Pharmacol Ther. 2008, pp 542-50
[20]. Heguy , A., O'Connor, T.P., Luettich, K., Worgall, S., Cieciuch ,A., Harvey, B.G., Hackett, N.R., Crystal, R.G., Gene expression profiling of human alveolar macrophages of phenotypically normal smokers and nonsmokers reveals a previously unrecognized subset of genes modulated by cigarette smoking,. J Mol Med (Berl). 2006 , pp 318-28.
[21]. Chen , J., Shields, P. G., Expression data from healthy smokers, Aug 2008.
[22]. Steiling, K., Kadar, A.Y., Bergerat, A., Flanigon, J., Sridhar, S., Shah, V., Ahmad, Q.R., Brody, J.S., Lenburg, M.E., Steffen, M., Spira , A., Comparison of proteomic and transcriptomic profiles in the bronchial airway epithelium of current and never smokers, PLoS One. Apr 2009.

Websites

http://www.ncbi.nlm.nih.gov/geo/
http://www.r-project.org
bioinfo.genotoul.fr/jvenn/index.html
http://david.abcc.ncifcrf.gov/
http://www.cytoscape.org/
biit.cs.ut.ee/gprofiler/
www.genecards.org/
www.genome.jp/kegg/pathway.html