Common up Regulated and down regulated Genes for Multiple Cancers using Microarray Gene Expression Analysis

Apoorva.D[*], Dr.Gurumurthy.H[**]

[*] Department of Information science and Engg, Dr.TTIT, India
[**] Department of Biotechnology, G M. Institute of Technology, India

Abstract- Cells are building blocks of living things. Normal cells multiply when body needs them and die when body doesn't need them. Cancer appears to occur when the growth of the cell in the body is out of control and cells divide too quickly. In cancer cells display uncontrolled multiplication, invasion and metastasis and caused by abnormalities in genetic material of transformed cells. Samples of microarray experiments performed on homo sapiens of normal and cancerous cells were downloaded from GEO database and these samples were imported to CLC Main Workbench software and expression analysis is performed to identify and rank common differentially expressed Genes in multiple type of cancer, by using technique called DNA Microarray data analysis which is used to find out expression of large number of genes simultaneously and it provides invaluable information on disease pathology, progression, resistance to treatment and therapeutic approaches for cancer. These genes found are useful for drug design as they act as biomarkers and also can be used in further analysis of fundamental signal transduction pathways that lead to carcinomas, since most genes causes cancer which are responsible for causing other cancer e.g. gene causing breast cancer have chances of casing ovarian cancer hence by finding such genes prevention of getting multiple cancer can be done.

Index Terms- GEO database, CLC Main Workbench, Differentially Expressed Genes, Microarray

I. INTRODUCTION

Cancer is a category of disease in which rapid creation of abnormal cells grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organ. Cancers are caused by abnormalities in cells which may be due to affects of carcinogens, such as tobacco smoke, radiation, chemicals, or infectious agents. Other cancer promoting genetic abnormalities may randomly occur through errors in DNA replication, or inherited. The heritability of cancers is usually affected by complex interactions between carcinogens and the host's genome. Since cancers can occur due to gene mutation in the cells finding those genes can be helpful. There are large no of genes present in the cell so there is one popular technique called microarray technology to find out expression of large number of genes simultaneously.

DNA Microarray is a collection of DNA spots attached on solid surface this is known as Affymetrix chip. Each DNA spot has specific sequence and is called as probe. Microarray methods were initially developed to study differential gene expression using complex populations of RNA. Refinement of these methods now permits the analysis of copy number imbalances and gene amplification of DNA.

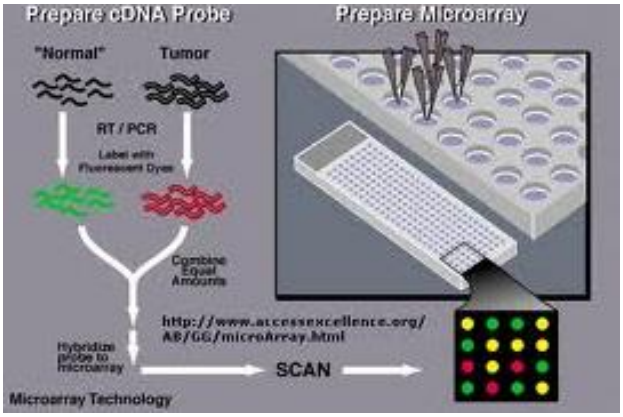

Figure 1: Microarray technology

For the research ten type of cancers were selected they are Colon, Breast, Ovarian, Lung, Pancreatic, Gastric, Liver, Thyroid, Salivary Gland, Pituitary cancer. These various types of cancers were selected based on the statistical data obtained from authentic sources like GEO and the supplementary information of published manuscripts. Since, previously no common up regulated genes and down regulated genes were identified for multiple types of cancers this research concentrates on identifying the common up regulated and down regulated genes by performing expression analysis using CLC Main Workbench. Performing statistical analysis for large no of genes is very hard and causes error but by using CLC Main Workbench software it is very easy to perform statistical analysis.

II. MATERIALS AND METHOD

Some of the databases and tools used are:

A. Gene Expression omnibus: it is a public repository that archives and distributes Microarray, next generation sequencing, and other forms of functional genomic data which is submitted by scientific community. It is a microarray database that allows users to download experiments and curated gene expression profiles provided by NCBI (http://www.ncbi.nlm.nih.gov). The datasets for different cancers are downloaded from this database

in order to perform expression analysis. The downloaded datasets are stored in ZIP/winRAR format the link for GEO database is http://www.ncbi.nlm.nih.gov/geo.

The data in GEO database is organized into platform, samples, series and datasets.

Platform is composed of summary description of array and sequencer and, for array based platform, a data table defining array template. Each platform record is assigned a unique and stable accession number (GPLxxx). A platform may reference samples submitted by submitters.

Sample records describe conditions under which the sample was handled, a manipulation is undergone, and abundance measurement of each element derived from it. Each sample is assigned a unique and stable accession number (GSMxxx) sample entity must reference only one platform and may include multiple series.

Series record links to group of related samples and provides focal point and discussion on whole studies. Each series data has unique and stable accession number (GSExxx).

Dataset is a curate collection of biologically and statistically comparable GEO samples and forms the basis of GEO's suite of data display and analysis tool. Samples within the dataset refer to same platform i.e. they share a common set of array elements.

B.CLC Main Workbench: It is the graphical user interface and the functions of CLC main Workbench are used by thousands of researches for DNA, RNA and protein sequence analyzing. Such as gene expression analysis, primer design, molecular cloning, phylogenetic analyses, and sequence data management. It is available on windows, MAC OS X, and Linux

CLC Main Workbench has Navigation area, View area, Menu Bar, Toolbar, Status Bar and Toolbox.

Navigation area is located in the left side of the screen, under the toolbar. It is used for organizing and navigating data. Its behavior is similar to the way files and folders are usually displayed on computer. The data in navigation area is organized into number of locations. When it is started for first time, there is one location called CLC_Data. Data can be added to navigation area in a number of ways. Files can be imported from the file system or by dragging it into the navigation area.

View area is the right hand part of the screen, displaying current work. The View area may consists of one or more views, represented by tabs at the top of the view area.

Tool box and Status bar: the toolbox is placed in the left side of the user interface of CLC Main Workbench below the navigation area. The toolbox shows a processes tab and a toolbox tab. By clicking the processes tab the toolbox displays the previous and running processes. The tools in the toolbox can be accessed by double clicking or by dragging elements from the navigation area to an item in the toolbox. The status bar is located at the bottom of the window. In the left side of the bar is an indication of whether the computer is making calculations or whether it is idle. The right side of the status bar indicates the range of selection of a sequence.

Workspace: if we are working on a project and have arranged the views for the project, we can save this arrangement using workspaces. The workspace remembers the way we have arranged the views, and can switch between different workspaces.

The method begins with retrieving the datasets from the GEO database after retrieving of datasets expression analysis is carried out using CLC main Workbench. First, the datasets are imported into CLC main Workbench by clicking import in the toolbar and file is selected. Samples are stored in navigation area. The next step is to tell the CLC Main Workbench how the samples are related this is done by setting up an experiment. An experiment is the central data type when analyzing expression data in the CLC Main Workbench. It includes a set of samples and information about how the samples are related. The experiment is also used to accumulate calculations like t-tests and clustering.

After setting up an experiment an experiment table will be opened. The table includes the expression values for each sample and in addition a few extra values such as the range, interquartile range, fold change and difference values. The experiment is saved and can proceed to expression analysis.

Next Quality control is performed. First MA plot is created since MA plot compares two samples, select two of the arrays and create a plot. Next select same two arrays used for the plot, choose log 2 transformation and create a MA plot again. This will result in a quite different plot.

The next step is to transform the expression values within the experiment, since this is the data we are going to use in further analysis. If the table is opened all the samples have an extra column with transformed expression values there is also an extra column for group mean and transformed IQR.

The next step is to examine and compare the overall distribution of the transformed expression values in the sample so Box plot is created. The next step in quality control is to check whether the overall variability of the samples reflect the grouping so principle component analysis is performed. In order to complement the principle component analysis hierarchical clustering of samples is done to see if the samples cluster in the groups we expect.

Next step is to identify and investigate the genes that are differentially expressed. Some statistical test is carried out that will be used to identify the genes that are differentially expressed between the two groups. The transformed values of FDR p-value correction is used to refine the genes that only have value below 0.0005.next one last criterion is added to the filter that is difference should have absolute value higher than 2.

Next step is to perform annotation test in which the gene list is annotated and use the annotation to see if there is a pattern in the biological annotations of genes in the list of candidate differentially expressed genes. Two types of annotation methods are used: Hyper geometric Tests on annotations and Gene Set Enrichment Analysis (GSEA).

First step is to import an annotation file used to annotate the arrays. The annotation file can be downloaded from the website http://www.affymetrix.com/support/technical/annotationfilesmain.affx.signing

The annotation file is imported to CLC Main Workbench. The add annotation is selected in annotation test present in expression analysis of toolbox the experiment and the annotation file is selected and next and finish is clicked. Next in toolbox hyper geometric test on annotation is selected present in annotation test within expression analysis. The two experiments are selected and next is clicked. Go biological process and

transformed expression values are selected and next, finish is clicked. And the test is performed.

Next select Gene Set Enrichment analysis present in annotation test within expression analysis of toolbox. The original full experiment is selected and next is clicked. Transformed expression value is selected and finish is clicked.

III. RESULTS

A.Differentially Expressed Genes.

Statistical analysis will be done to identify genes that are differentially expressed between the two groups. The two corrected p-values, bonferroni corrected and FDR corrected parameters are selected. For the analysis FDR p-value is used 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) we are willing to accept.

To do more refined selection of the genes that we believe to be differentially expressed, advanced filtering is used which is located at the top of the experiment table. Transformed-FDR p-value correction is selected in the first drop-down box, select < in the next and enter 0.0005(or 0.005 depending on locale settings). Declaring 0.0005 means we are setting specificity to be 95. In diagonistic testing when the disease prevalence is small, we need a test with very high specificity, as otherwise there are too many false positive results.

B.Annotation test.

The gene list will be annotated and used to see if there is a pattern in the biological annotations of the genes in the list of candidate differentially expressed genes.

Table 1: The number of up regulated and down regulated genes obtained for the cancers

Sl. no	Type of cancers	up regulated	down regulated
1	Colon cancer	5	0
2	Breast cancer	204	24
3	Gastric cancer	560	148
4	Liver cancer	174	13
5	Lung cancer	120	22
6	Ovarian cancer	949	405
7	Pancreatic cancer	116	13
8	Pituitary cancer	137	42
9	Salivary gland cancer	739	313
10	Thyroid cancer	900	407

Table 2: Shows up regulated common genes.

Sl. no	GO id	Types of Cancers	Description	Gene name
1	19915	Colon, breast, salivary.	Lipid storage.	PLIN2 perilipin 2 PLIN2 perilipin 2 PLIN2 perilipin 2
2	14070	Colon, gastric, liver, pancreatic	response to organic cyclic substance	
3	42493	Colon, gastric, liver, ovarian, pancreatic, salivary, thyroid	response to drug	
4	6955	Colon, gastric, salivary	immune response	TRBC 1 T cell receptor beta constant 1
5	60748	Breast, ovarian, salivary, thyroid.	tertiary branching involved in mammary gland duct morphogenesis	PGR progesterone receptor
6	8064	Breast, liver.	regulation of actin polymerization or depolymerization	CXCL 12 chemokine (C-X-C motif) ligand 12
7	2070	Breast, ovarian, salivary, thyroid.	epithelial cell maturation	PGR progesterone recepto

8	50847	Breast, ovarian, salivary.	progesterone receptor signaling pathway	r PGR progest erone recepto r
9	33603	Breast, ovarian, thyroid.	positive regulation of dopamine secretion	CXCL 12 chemo kine (C-X-C motif) ligand 12 CXCL 12 chemo kine (C-X-C motif) ligand 12
10	1667	Breast, gastric, pancreatic,	ameboidal cell migration	

Table 3: Shows down regulated common genes.

Sl. no	GO id	Types of cancers	Description	Gene name
1	6915	Breast, liver, ovarian, thyroid.	apoptosis	IL19 interleukin 19
2	6810	Breast, gastric, ovarian, pancreatic, pituitary, salivary, thyroid.	transport	PDZK1 PDZ domain containing 1
3	30154	Breast, lung, pancreatic.	cell differentiation	Ifrd1 interferon-related developmental regulator 1
4	15031	Breast, gastric, liver, ovarian, pancreatic, pituitary, salivary, thyroid.	protein transport	Napb N-ethylmaleimide sensitive fusion protein attachment protein beta
5	122	Breast, gastric, pituitary, salivary.	negative regulation of transcription from RNA polymerase II.	Ctnnbip1 catenin beta interacting protein 1
6	6281	Breast, gastric, salivary.	DNA repair	TOP2A topoisomerase (DNA) II alpha 170kDa
7	6412	Breast, gastric, liver, lung, ovarian, pituitary, salivary, thyroid.	translation	Akt1 thymoma viral proto-oncogene 1
8	16192	Breast, lung, ovarian, pituitary, salivary, thyroid.	vesicle-mediated transport	Gsn gelsolin
9	6397	Breast, gastric, liver, ovarian, pituitary,salivary, thyroid.	mRNA processing	Srsf2 serine/arginine-rich splicing factor2
10	6468	Breast, gastric, lung, ovarian, pituitary, thyroid.	protein phosphorylation	Akt1 thymoma viral proto-oncogene 1

IV. CONCLUSION

The work comprises of investigating the common genes responsible to cause a staggering variety of cancer through their differential expression patterns using a technique called DNA Microarray. As an overview of entire process, relevant data from GEO is obtained, tabulated them and subjected them to analysis and found common differentially expressed genes.

Future scope is that the identified set of genes which are common and differentially expressed in multiple cancers might be useful in the further analysis of fundamental signal transduction pathways that lead to carcinomas, so that these genes can act as biomarkers for drug design. Since most of the cancers are caused by genes which are already responsible to cause other cancer e.g. women having breast cancer have chances of getting ovarian cancer, so it can be prevented by doing gene therapy on common genes responsible to cause cancer.

REFERENCES

[1] http://www.ncbi.nlm.nih.gov/pubmedhealth.

[2] Jain N, Thatte J, Braciale T, Ley K, O'Connell M and Lee JK. Local-Pooled-error test for Identifying Differentially Expressed genes with a small number of replicated microarrays. Oxford Journals Bioinformatics 2003; 19: 1945-1951.

[3] http://www.ncbi.nlm.nih.gov/geo.

[4] http://www.clcbio.com/products/clc-main-workbench.

[5] http://www.clcbio.com.

[6] Yudi Pawitan and Stefan Michiels. False Discovery rate, sensitivity and sample size for microarray studies 2005; 1: 1-2.

AUTHORS

First Author – Apoorva.D, B.E, M.TECH, DR.TTIT, India, apoorva.bhavikatte@gmail.com.
Second Author – Dr.Gurumurthy.H , M.Sc, PGD (Bioinformatics),PhD., MISTE., G M. Institute of Technology, India.