ANALYSIS OF MICROARRAY DATA

FROM OVARIAN CANCER CELLS

Project Report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

In

BIOINFORMATICS

Under the Supervision of

Dr.Pradeep K. Naik

By

Natasha Aggarwal(081502)
Madhurika Sharma(081511)

to

JAYPEE UNIVERSITY OF
INFORMATION TECHNOLOGY

Jaypee University of Information and Technology

Waknaghat, Solan – 173234, Himachal Pradesh

TABLE OF CONTENTS

CERTIFICATE

This is to certify that project report entitled "**Analysis of Microarray Data from Ovarian Cancer Cells** ", submitted by **Natasha Aggarwal (081502), Madhurika Sharma (081511)** in partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: 26/5/2012

Naik

(Pradeep Kumar Naik)

Associate Professor

ACKNOWLEDGEMENT

By the means of this acknowledgement, we would like to extend a word of thanks and appreciation to the people who have contributed to the successful completion of this work.

Firstly, we would like to express my gratitude to Dr. Pradeep Kumar Naik, Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat (H.P), for his valuable support and guidance at all stages of this work. We will not lag behind in expressing our thanks to Mr.Vinay Kumar Mittal and all our friends for their guidance; affectionate attitude and keen interest which made this project a success. Working on this project was a delightful and enriching experience.

Natasha Aggarwal Natasha Aggarwal

Madhurika Sharma Madhurika

Date: 27th May, '12

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The project aims to analyze the cancer data from human ovarian cancer cells for the up regulated genes that play a vital role in regulating the gene expression of cancer cells. The study is based on comparative gene expression profiling analyses of cells of normal and ovarian cancer epithelial cells using microarray technique. The data (.cel files) obtained from this experiment was normalized to obtain the genes which satisfied the criteria of considered threshold. We obtained 100 genes that were highly expressed in cancer cell were analyzed for their role among various pathways involve in cancer.

R- Package was used to preprocess the .cel files so as to remove the noise and set clear grounds for pathway analysis. The statistical tests like mean of expression intensity vales, t-test and average log fold change were used to normalize the data using RMA normalization technique. Pathway analysis was done using online software-WebGestalt. A total of 86 genes were mapped and their functions in different pathways were identified.

The identified genes were found to regulate ovarian cancer and could serve as potent targets for drug development. Further research on these genes can help the scientific community to develop drugs for better therapeutic outcome. This might help to increase the longevity of the patients suffering with cancer.

CHAPTER 1
INTRODUCTION

A DNA microarray also commonly known as gene chip, DNA chip, or biochip is a collection of microscopic DNA spots attached to a solid surface. It is used to measure the expression levels of large numbers of genes simultaneously. Each DNA spot contains a specific DNA sequence, known as probes. A probe is a short section of a gene or other DNA element that are used to hybridize a cDNA sample. It is possible to measure the expression of many thousands of genes bonded to an array simultaneously.

The principle behind microarrays is base pair hybridization between two DNA strands. The base pair hybridization is specific due to complementary nucleic acid sequences which pair with each other by forming hydrogen bonds.

1.1 Problems associated with high throughput gene expression analysis

The DNA microarrays enable the user to study large number of genes at once. Since the right statistical methods and tools to deal with the multiple genes at the same time do not exist, it poses a challenge to analyze the large data sets. Another problem is that the microarray data sets contain a large number of variables (tens of thousands) but only a small number of replicates create unique data analysis sets.

1.2 Normalization

Normalization means to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes. This variation is also known as noise. Imbalances between the labeling dyes may arise from differences between the labeling efficiencies or by the use of different scanner settings. If the imbalance is more complicated than a simple scaling of one channel relative to the other then normalization needs to be intensity dependent. Positions on a slide may differ because of differences between the print-tips on the array printer, variation over the course of the print-run, non-uniformity in the hybridization or from artifacts on the surface of the array which affect one color more than the other.

1.2.1 RMA normalization

RMA stands for Robust Multi-array Average and is used to preprocess expression data. It uses probe-level expression measurements of all arrays in a study to estimate expression values. RMA proceeds in three steps: Background correction, normalization and summarization. RMA successfully reduces the variance of low abundance transcripts using controlled datasets in which known quantities of specific mRNAs are pooled together to distinguish differentially expressed transcripts from the control.

1.2.2 MAS5 Normalization

MAS5 normalizes each array independently and sequentially. It uses data from mismatch probes to calculate a "robust average", based on subtracting mismatch probe value from match probe value.

1.3 Statistical tests

1.3.1 Mean

The mean of a series of variables is the arithmetical average of those numbers. It is determined by summing the numbers, then dividing that sum by the number of variables (count) included in that sum. The formula for calculating mean:

$$\bar{x} = \frac{\sum x}{n}$$

1.3.2 p-value

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. One often "rejects the null hypothesis" when the p-value is less than the significance level α which is often 0.05 or 0.01. When the null hypothesis is rejected, the result is said to be statistically significant.
For our study, a p-value of <=0.001 is considered, it estimates a probability of 0.1 to observe the data, by chance. 561 genes were shortlisted from 54676 genes using this criterion.

1.3.3 t- test

The t-test assesses whether the means of two groups are statistically different from each other or not. The analysis is appropriate whenever we want to compare the means of two groups, and especially appropriate as the analysis for the posttest-only two-group randomized experimental design. The t-test, one-way Analysis of Variance (ANOVA) and a form of regression analysis are mathematically equivalent.

1.3.4 Average log fold

It is calculated for each gene by first finding the average value for each group, and then doing anti-log of the difference between the groups, i.e. 2^(average group A - average group B). Log fold cutoff of <= -0.29 was considered, 100 genes were obtained which highly regulate the ovarian cancer cell related pathways.

1.4 Clustering

Clustering is about assigning a set of objects into groups called clusters so that the objects in the same cluster are more similar in some sense or another to each other than to those in other clusters.

1.4.1 Types of Clustering Algorithms

- Partitioning-based clustering: **are algorithms that determine all the clusters at once in most cases.**
  - K-means clustering
  - K-medoids clustering
  - EM (expectation maximization) clustering
- **Hierarchical clustering:** these algorithms find successive clusters using previously established ones.
  - Divisive clustering is a top down approach.
  - Agglomerative clustering is a bottom up approach.
- **Density-Based Methods**: these clustering algorithms are used to help discover arbitrary-shaped clusters. A cluster is defined as a region in which the density of data objects exceeds some threshold.
  - DBSCAN

- OPTICS

For the clustering of our dataset we used Agglomerative Hierarchical Clustering[5]. It is a method of cluster analysis which seeks to build a hierarchy of clusters based on similarity and their relationships. Agglomerative is a type of hierarchical clustering and is a bottom's up approach because in such a method, each gene is considered as a single entity and based on the $\log_2$ values the clustering is done.

1.5 R Package

R is an open source programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians for developing statistical software and data analysis.

R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. R has stronger object-oriented programming facilities than most statistical computing languages. R's data structures include scalars, vectors, matrices, data frames and lists. The R object system is extensible and includes objects for, among others, regression models, time-series and geo-spatial coordinates. R supports procedural programming with functions and, for some functions, object-oriented programming with generic functions.

1.6 Pathway Analysis

Pathways are collections of genes and proteins that perform a well-defined biological task. For instance, proteins that work to successively synthesize metabolites within a cell are grouped into metabolic pathways. Similarly, proteins that are involved in the transduction of a signal from the cell membrane to the nucleus are grouped into signal transduction pathways. These pathways have been established through decades of molecular biology research and are collected in a variety of public pathway repositories[4].

A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move.

Biological pathways help the body to develop properly and stay healthy; many things must work together at many different levels - from organs to cells to genes. There are many types of biological pathways. Some of the most common are involved in metabolism, the regulation of genes and the transmission of signals.

- Metabolic pathways make possible the chemical reactions that occur in our bodies. An example of a metabolic pathway is the process by which your cells break down food into energy molecules that can be stored for later use. Other metabolic pathways actually help to build molecules.
- Gene regulation pathways turn genes on and off. Such action is vital because genes produce proteins, which are the key components, needed to carry out nearly every task in our bodies. Proteins make up our muscles and organs, help our bodies move and defend us against germs.
- Signal transduction pathways move a signal from a cell's exterior to its interior. Different cells are able to receive specific signals through structures on their surface, called receptors. After interacting with a receptor, the signal travels through the cell where its message is transmitted by specialized proteins that trigger a specific action in the cell. For example, a chemical signal from outside the cell might be turned into a protein signal inside the cell. In turn, that protein signal may be converted into a signal that prompts the cell to move.

When multiple biological pathways interact with each other, it is called a biological network. A lot about human diseases can be learned from studying biological pathways. Identifying what genes, proteins and other molecules are involved in a biological pathway, what goes wrong when a disease strikes can be predicted.

For example, researchers may compare certain biological pathways in a healthy person to the same pathways in a person with a disease to discover the roots of the disorder. Keep in mind that problems in any number of steps along a biological pathway can often lead to the same disease.

1.6.1 Cancer and Biological Pathways

Until recently, most types of cancers were driven by a single genetic error and could be treated by designing drugs to target those specific errors. Much of it was based on the success of a drug that was specifically designed to treat a blood cancer called chronic myeloid leukemia (CML). CML occurs because of a single genetic glitch that leads to the production of a defective protein that spurs uncontrolled cell growth. Gleevec binds to that protein, stopping its activity and producing dramatic results in many CML patients.

Unfortunately, the one-target, one-drug approach has not held up for most other types of cancer. Recently it was deciphered that the genomes of cancer cells have found an array of different genetic mutations that can lead to the same cancer in different patients. Then, based on the genetic profile of their particular tumor, patients could receive the drug or drug combination that is most likely to work for them.

Instead of attempting to discover ways to attack one well-defined genetic enemy, researchers now faced the prospect of fighting lots of little enemies. Fortunately, this complex view can be simplified by looking at which biological pathways are disrupted by the genetic mutations. Rather than designing dozens of drugs to target dozens of mutations, drug developers could focus their attentions on just two or three biological pathways. Patients could then receive the one or two drugs most likely to work for them based on the pathways affected in their particular tumors.[3]

1.6.2 Software Used

WebGestalt: WebGestalt is a WEB-based **GEne SeT AnaLysis Toolkit**. It is designed for functional genomics, proteomic and large-scale genetic studies from which large number of gene lists (e.g. differentially expressed gene sets, co-expressed gene sets etc) are generated. WebGestalt incorporates information from different public resources and provides an easy way for biologists to make sense out of gene lists.[1][2]

We used WebGestalt to get the results and finalize our interpretations to find out the important genes and their vital role in regulating the ovarian cancer. The genes can be easily identified and hence, can be worked upon to protect the mankind from further severity caused by the disease.

CHAPTER 2
MATERIAL AND METHODS

2.1 Software Used

- R- Package
- WebGestalt

2.2 Microarray expression data for ovarian cancer

The data was taken from *http://www.ncbi.nlm.nih.gov/projects/geo/* in the form of .cel zipped files. 6 normal and 6 cancerous .cel files were used for the analysis.

R and Bioconductor (a set of packages that run in R) were used to do most of the mathematical analyses beside Microsoft excel sheet.

2.3 Transforming intensity data into expression data

A bioconductor was installed in order to preprocess affymetrix expression data and then the working directory was changed to the directory where all the raw data files were stored. Then the "library" package containing the Affymetrix microarray code was loaded and .CEL files were read and summarized.

2.4 Normalization of data

Data was normalized with RMA normalization method and the histogram was plotted using the transformed log values of expression data so as to get the normal distribution. To get better reults Quantile normalization was done with the help of limma package . For graphical representation of data a box plot was built before and after quantile normalization in R.

Box plot[12] is a convenient way of graphically depicting groups of numerical data. Box plots display differences between populations without making any assumptions of the underlying statistical distribution: they are non-parametric. The spacing between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and identify outliers.

After obtaining box plot, for better representation MA plot was built.
M in the MA plot[13] stands for the log intensity ratio (or difference between log intensities) and A is the average log intensity for a dot in the plot. MA plots are used to visualize intensity-dependent ratio of raw microarray. Brighter the spot, more observed difference between sample and control. The MA plot uses M as the y-axis and A as the x-axis. The MA plot gives a quick overview of the distribution of the data.

2.5 Statistical analysis of expression data

The differentially expressed genes were identified with the help of average log fold and t-test.
For p-value the cutoff <=0.0011 was taken to determine the significant genes.
The average log fold cutoff was <=0.29.
The genes so obtained after this was visualized with help of heat-map. A heat map[11] is a graphical representation of data where the individual values contained in a matrix are represented as colors. The data is displayed in 2D on a coloured scale. The Heat Map for cancer cells is represented in red colour where as that for the normal cells in green.

2.6 Pathway mapping of highly expressed genes

The 100 genes' that cleared the cut off value of log2 were assembled in a text file named probeset_ids.txt unique Affymetrix Probe-ids served as the input for the WebGestalt user interface, an online software.
Out of 100 ids, 86 ids were mapped in the database out of which only 82 were uniquely identified. The information for the remaining 14 ids is not available at present. From this, we got a comprehensive outlook about the genes these ids correspond to, their role in various metabolic pathways and their role in regulation of ovarian cancer in humans.
WebGestalt enables the user to do the

- Enrichment analysis
- GO slim classification

2.6.1 Enrichment analysis

It enables the user to choose from a range of options to carry out the analysis. Such as

- **Gene Ontology Analysis.** Enrichment analysis for the Gene Ontology categories. The result is visualized in a directed acyclic graph (DAG) in order to maintain the relationship among the enriched GO categories.
- **KEGG Analysis.** Enrichment analysis for the KEGG pathways. Genes can be highlighted in the KEGG pathway maps.
- **Wikipathways Analysis.** Enrichment analysis for the pathways in the Wikipathways database. Genes and corresponding changes can be colored in the Wikipathways maps.
- **Pathway Commons Analysis.** Enrichment analysis for the pathways in the Pathway Commons database.
- **Transcription Factor Target Analysis.** Enrichment analysis for the targets of transcription factors.[7]
- **Micro RNA Target Analysis.** Enrichment analysis for the targets of Micro RNAs[7]
- **Protein Interaction Network Module Analysis.** Enrichment analysis for network modules predicted from Protein Interaction Networks.
- **Cytogenetic Band Analysis.** Enrichment analysis for the cytogenetic bands.

2.6.2 GO slim classification

The user can choose to classify the input genes on basis of

- **Biological processes.** A biological process is a recognized series of events or molecular functions. A process is a collection of molecular events with a defined beginning and end. Mutant phenotypes often reflect disruptions in biological processes
- **Molecular function.** Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level.

- **Cellular components.** Also known as the biological matter refers to the unique, highly organized substances and substances of which cells, and thus living organisms, are composed

The parameters considered were

Reference Set	hsapiens_affy_hg_u133_plus_2
Statistical Method	Hypergeometric
Multiple Test Adjustment	Benjamini and Hochberg (1995) [8]
Significance Level	Top 10
Minimum No. of Genes foe A Category	2

Table1. Parameters for the Enrichment Analysis

Benjamini and Hochberg: It is a procedure that controls the Family Wise Error Rate (FWER) [9] – that is, the probability of falsely rejecting any true null hypothesis. The False Discovery Rate (FDR) [10] – that is, the expected value of the proportion of rejected hypotheses that are falsely rejected can be identified using this procedure when the p-values are independent.

CHAPTER 3
RESULTS AND DISCUSSION

In the study **12** .cel files consisting of 6 normal cells and 6 cancer cells were taken .

A total of **54676** genes were present in these files.

The preprocessing was done using various packages of R language.

These genes were normalized using RMA normalization method. The histogram was then plotted so as to visualize a normal distribution of data.

A second normalization i.e. quantile normalization was done in limma package.

For better visualization in a graphical form, a box plot was built before and after normalization so as to clearly see the difference in processed data.

After obtaining box plot, for better representation MA plot was built.

Differentially expressed genes were obtained after calculating average log fold change and t-test.

A p value cutoff of <=0.001 was taken and a total of 561 significant genes were obtained and for better accuracy a cutoff value of <=0.29 was taken out which 100 significant genes were filtered. Then the pathway analysis was done using WebGestalt

3.1 Histogram

After RMA normalization histogram of intensity values was constructed so as to visualize the normalized data.

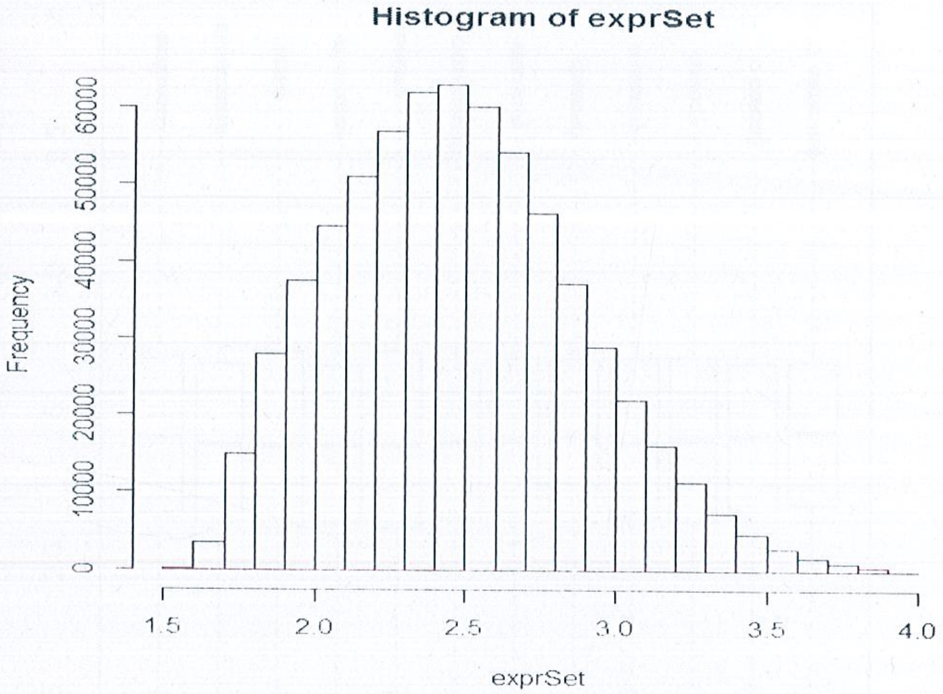


Fig.1 histogram

This histogram depicts that the data is well spread across the range of log intensity values. The variability is approximately constant at all intensities and it appears to be normally distributed.

3.2 Box plot

Before and after quantile normalization box plot was built.

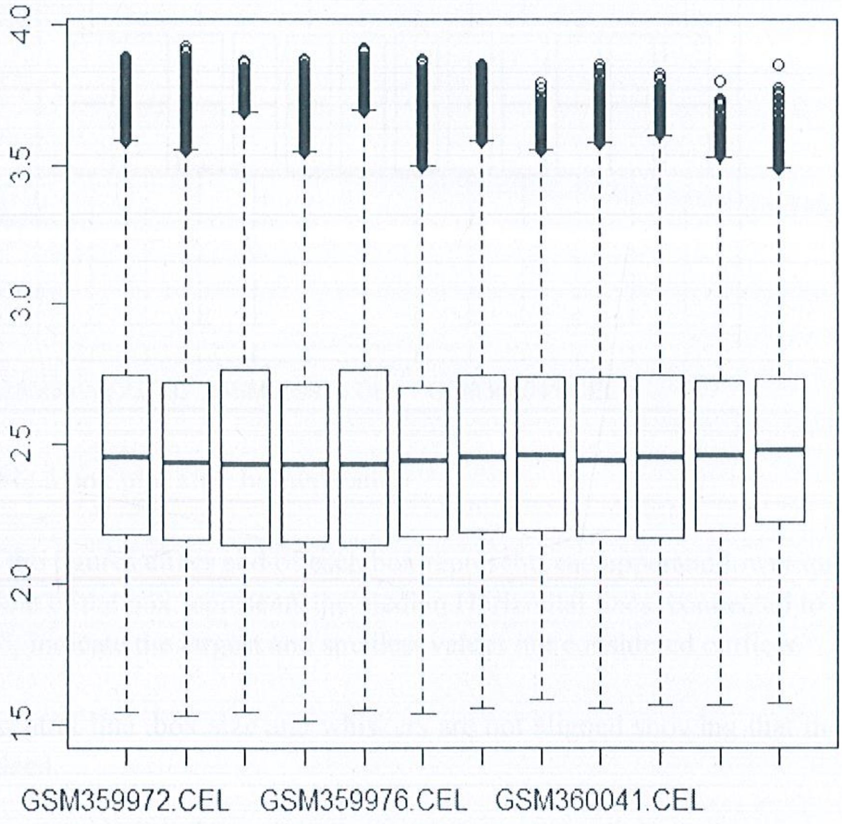


Fig.2 box plot before normalization

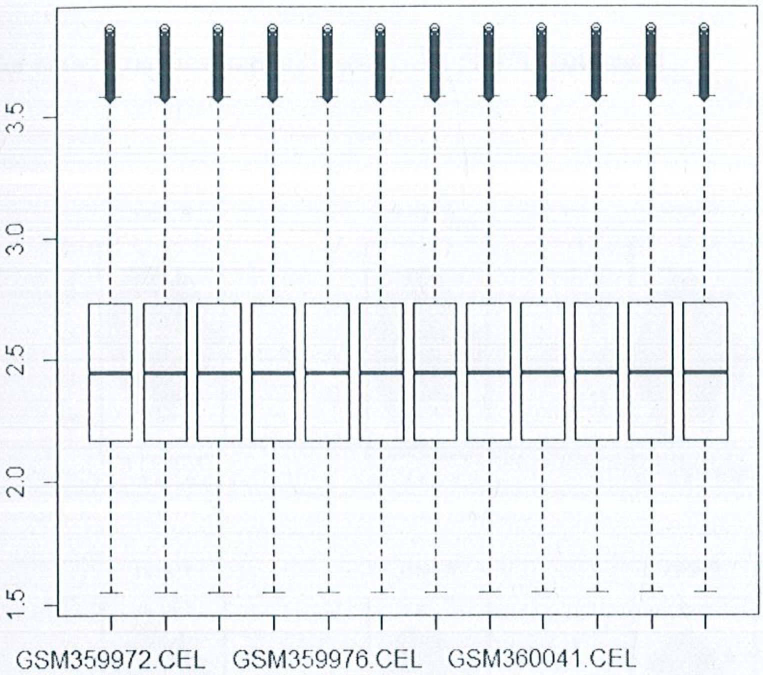


Fig.3 box plot after normalization

In both of the figures either end of each box represents the upperand lower quartile. The line in the middle of the box represents the median.Horizontal lines, connected to the box by "whiskers", indicate the largest and smallest values not considered outliers.

In Fig. 2 central line ,box size and whiskers are not aligned showing that the data needed to be normalized.

In Fig.3 all the parameters of box plot are properly aligned implying that our data is normalized.

3.3MA plot

For better graphical representation MA plot was obtained.

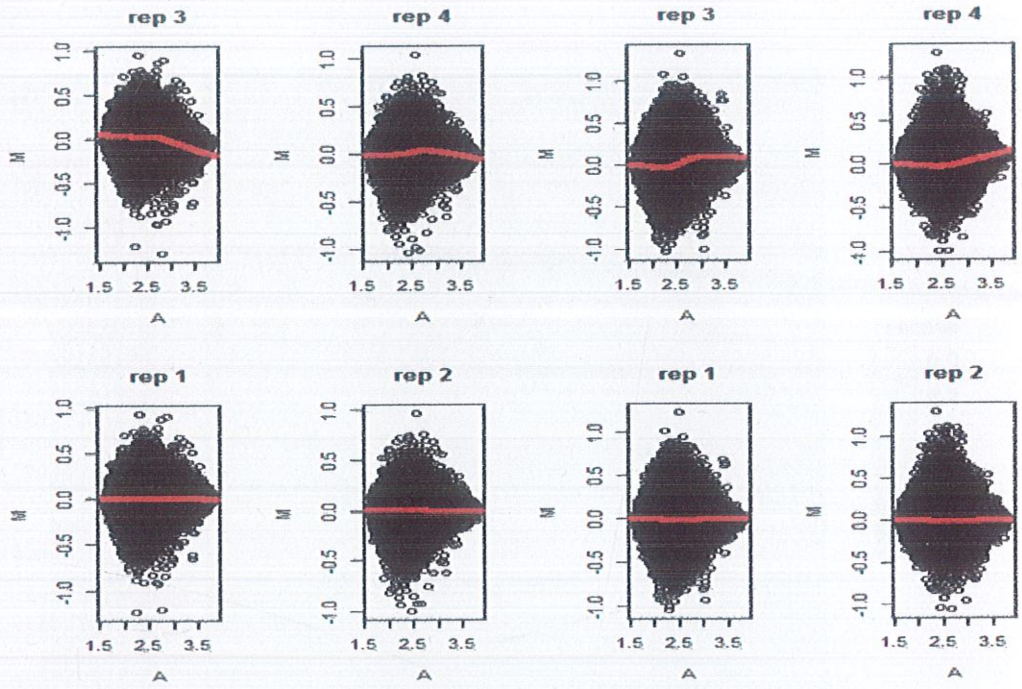


Fig.4 MA plot before and after normalization

The first figure so obtained didn't have a straight line showing that the data is intensity biased and that it has some errors.

We then performed normalization repeatedly until we got straight line in microarray plot resulting in error free data.

3.4 Heat map

The 100 signicant genes obtained by taking the p value cutoff and average log fold cutoff were visualized using the heat map.

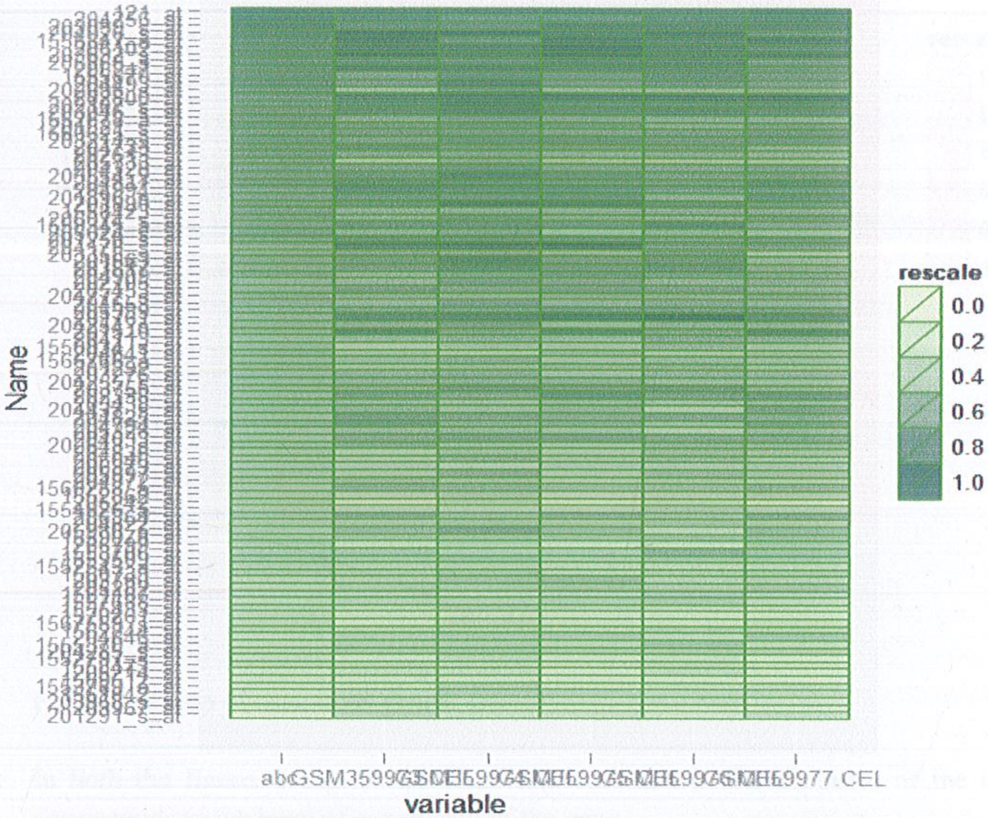


Fig.5 Heat map of normal genes

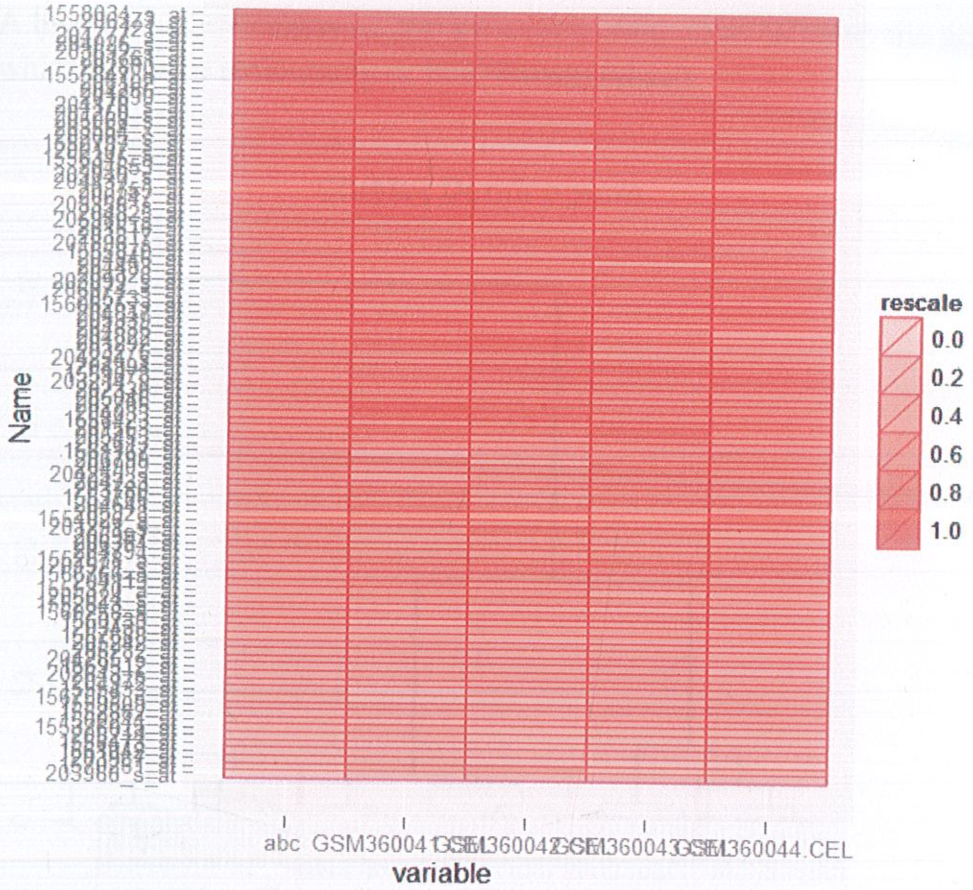


Fig.6 Heat map of cancerous genes

In both the figures ; Fig. 5 and Fig.6 ,the variation in the deepness of the colour corresponds to the level of expression of the gene.
Light colour depicts the under-expression of a gene and deep colour depicts the over-expression .

3.5Clustering results of highly expressed genes

A hierarchical clustering of 100 genes was done so as to group the genes which might be structurally or functionally related.

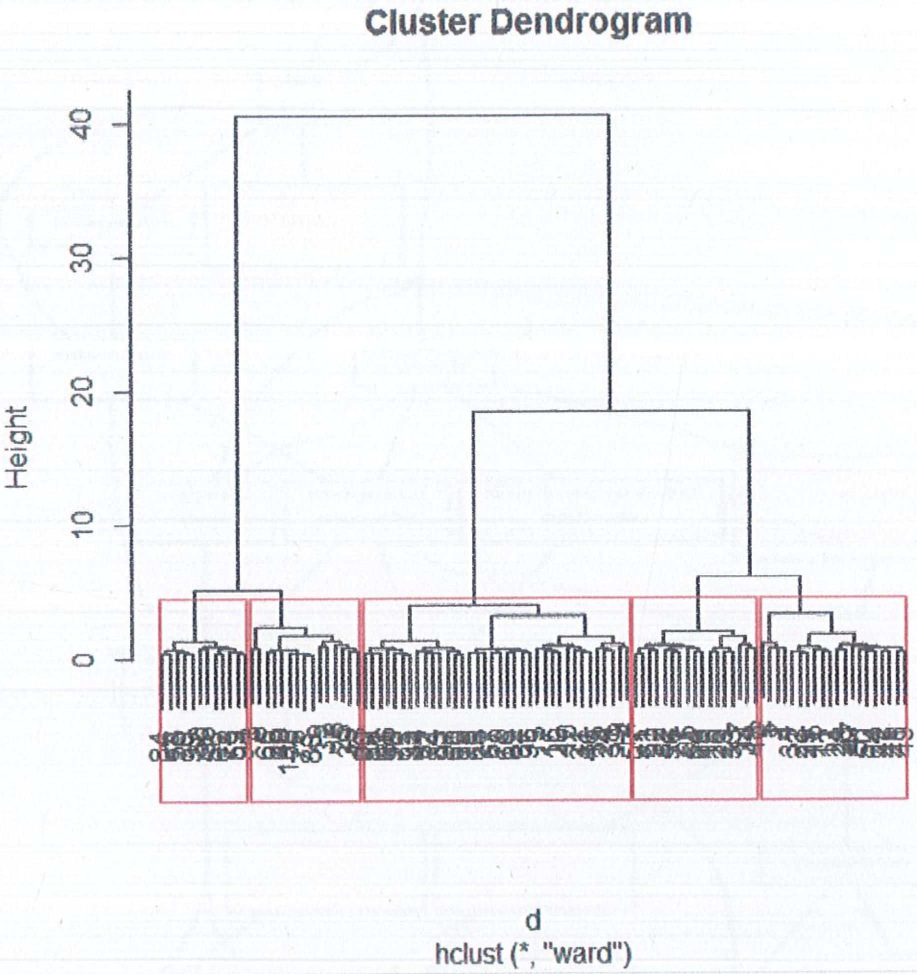


Fig.7 a dendrogram showing results of hierarchical clustering

Cluster dendrogram depicting results of the microarray analysis, generated by R software. The cluster dendrogram was derived from the pooled microarray results of the ovarian cancer genes. Each horizontal line with groupings represent genes that are linked by common structure or a function.

3.6 Pathway Analysis

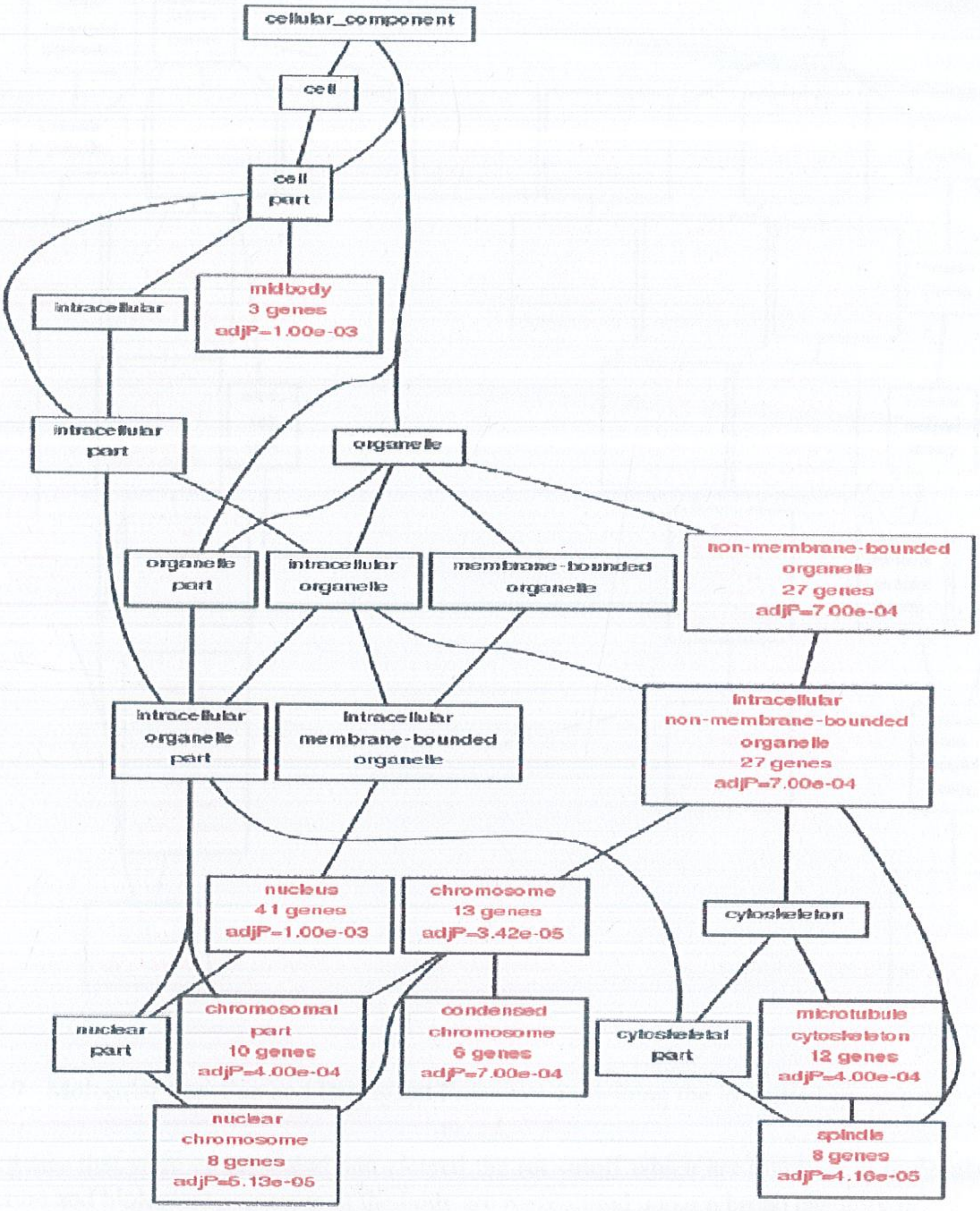

cellular_component
cell
cell part
intracellular
midbody 3 genes adjP=1.00e-03
intracellular part
organelle
organelle part
intracellular organelle
membrane-bounded organelle
non-membrane-bounded organelle 27 genes adjP=7.00e-04
intracellular organelle part
intracellular membrane-bounded organelle
intracellular non-membrane-bounded organelle 27 genes adjP=7.00e-04
nucleus 41 genes adjP=1.00e-03
chromosome 13 genes adjP=3.42e-05
cytoskeleton
nuclear part
chromosomal part 10 genes adjP=4.00e-04
condensed chromosome 6 genes adjP=7.00e-04
cytoskeletal part
microtubule cytoskeleton 12 genes adjP=4.00e-04
nuclear chromosome 8 genes adjP=5.13e-05
spindle 8 genes adjP=4.16e-05

Fig.8 Cellular Components containing the identified genes marked in red

All the genes that were up regulated and cleared the set cutoff which are involved in cellular components in the body are represented under a broad category in red.

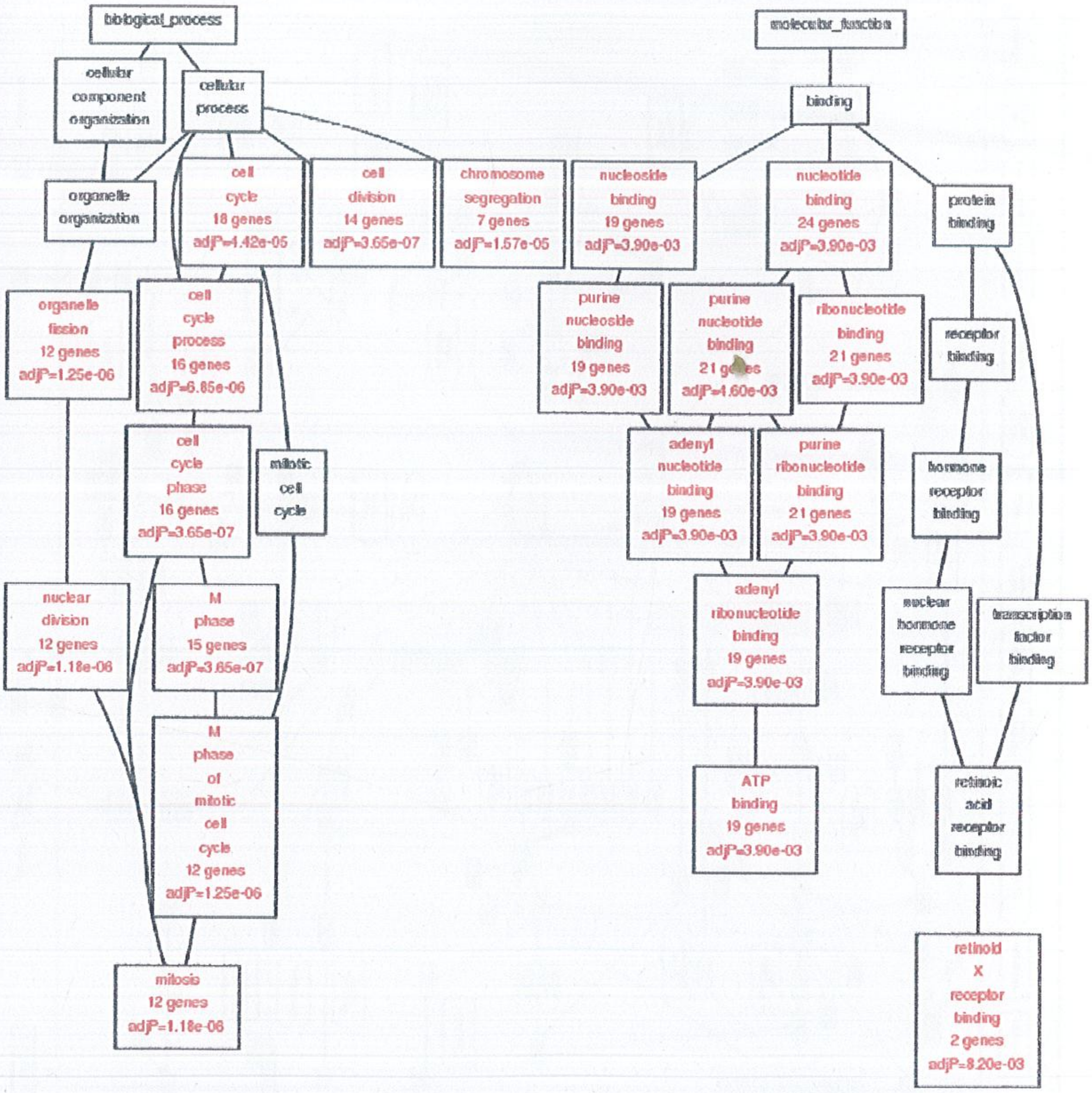

biological_process
cellular component organization
cellular process
organelle organization
cell cycle 18 genes adjP=4.42e-05
cell division 14 genes adjP=3.65e-07
chromosome segregation 7 genes adjP=1.57e-05
organelle fission 12 genes adjP=1.25e-06
cell cycle process 16 genes adjP=6.85e-06
cell cycle phase 16 genes adjP=3.65e-07
mitotic cell cycle
nuclear division 12 genes adjP=1.18e-06
M phase 15 genes adjP=3.65e-07
M phase of mitotic cell cycle 12 genes adjP=1.25e-06
mitosis 12 genes adjP=1.18e-06
molecular_function
binding
nucleoside binding 19 genes adjP=3.90e-03
nucleotide binding 24 genes adjP=3.90e-03
protein binding
purine nucleoside binding 19 genes adjP=3.90e-03
purine nucleotide binding 21 genes adjP=1.60e-03
ribonucleotide binding 21 genes adjP=3.90e-03
receptor binding
adenyl nucleotide binding 19 genes adjP=3.90e-03
purine ribonucleotide binding 21 genes adjP=3.90e-03
hormone receptor binding
adenyl ribonucleotide binding 19 genes adjP=3.90e-03
nuclear hormone receptor binding
transcription factor binding
ATP binding 19 genes adjP=3.90e-03
retinoic acid receptor binding
retinoid X receptor binding 2 genes adjP=8.20e-03

Fig.9 Molecular Function and Biological Processes containing the identified genes marked in red.

The genes that were up regulated and cleared the set cutoff which are involved in molecular function and biological processes in the body are represented under a broad category in red.

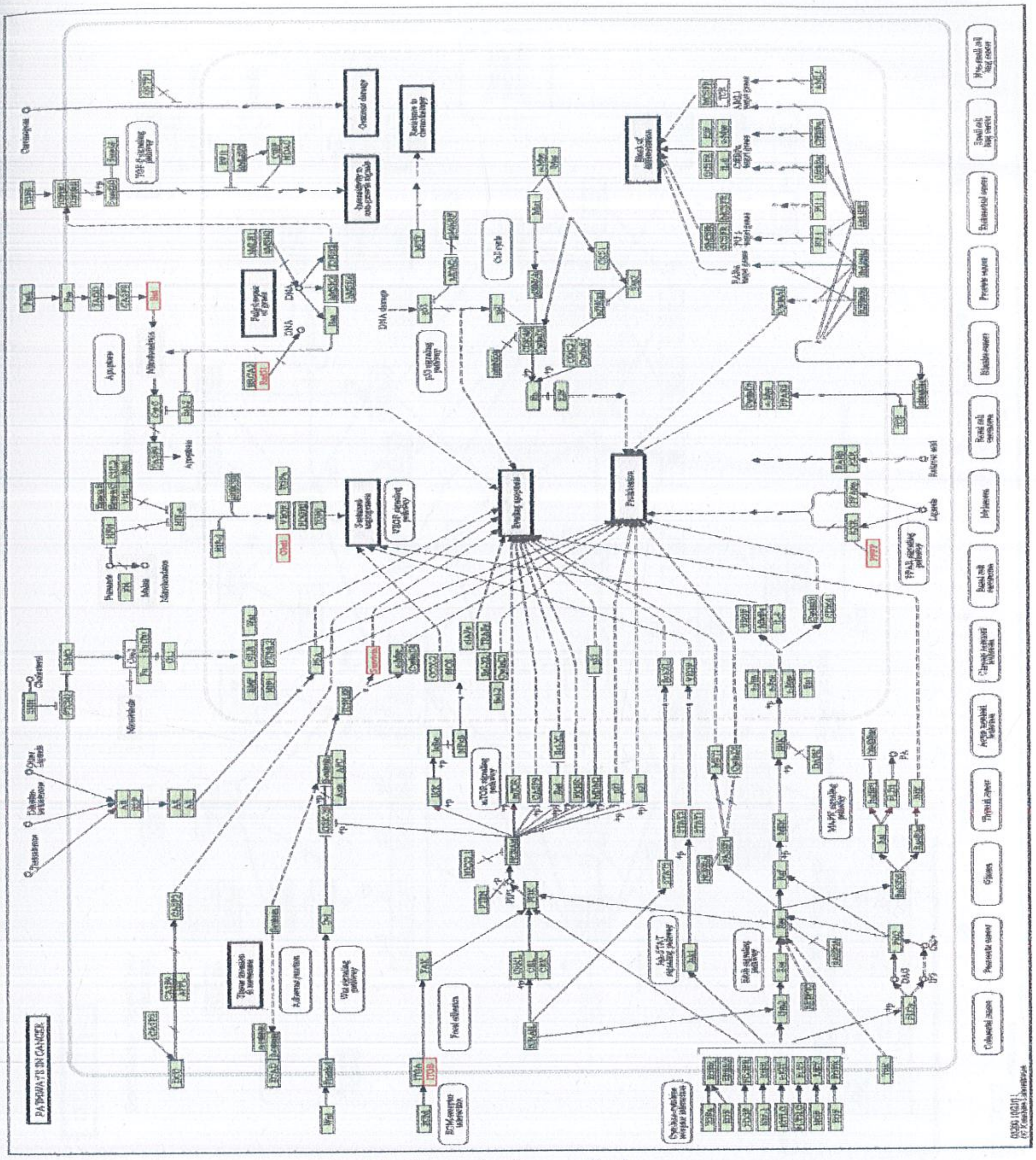


Fig.10 Genes in Cancer Pathways

Identified genes marked in red which were over expressed play role in some cancer pathways as shown

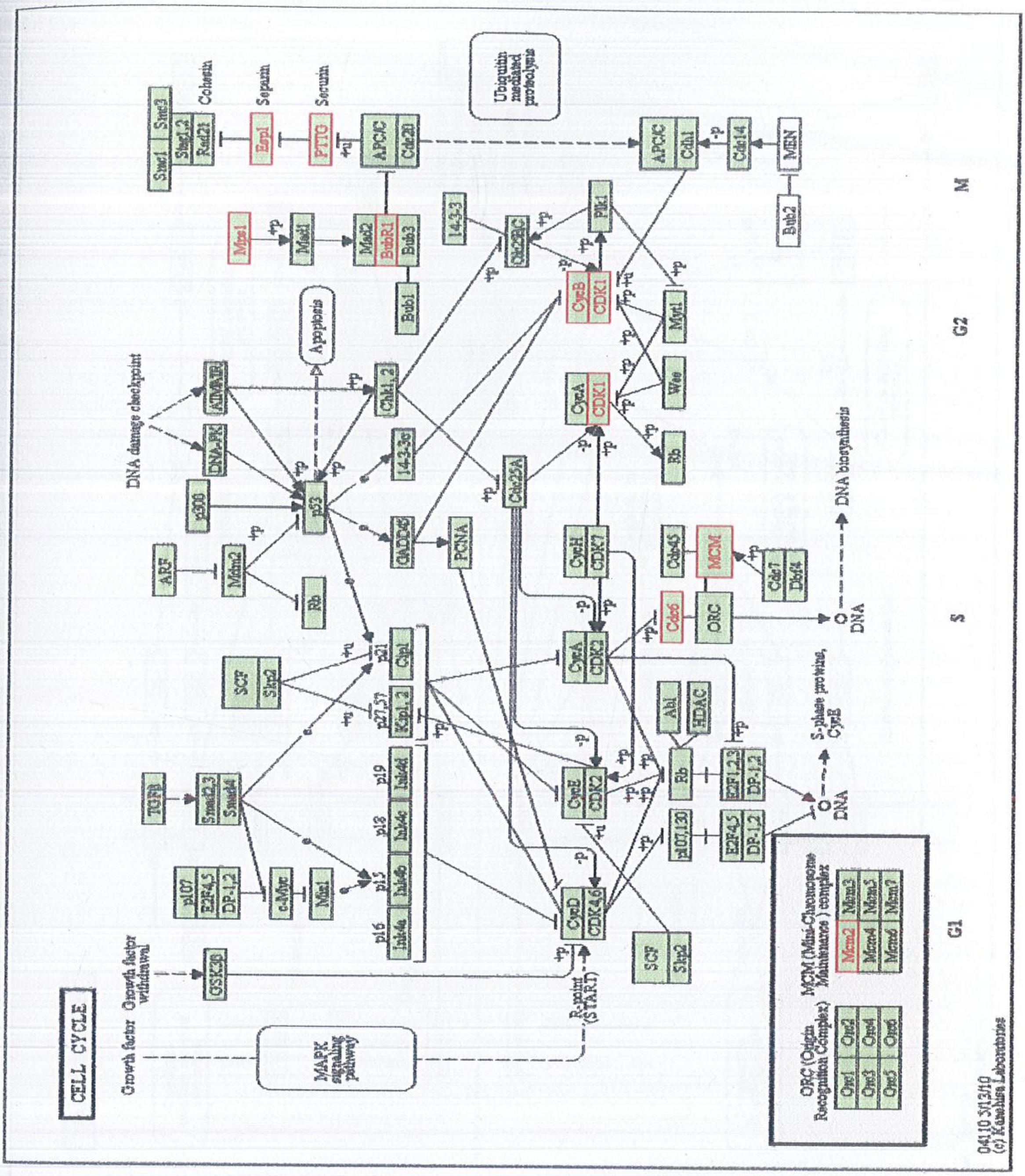


Fig.11 Genes involved in cell cycle; marked in red.

Over expressed genes of ovarian cancer cells are also involved in cell cycle which promotes cancer growth.

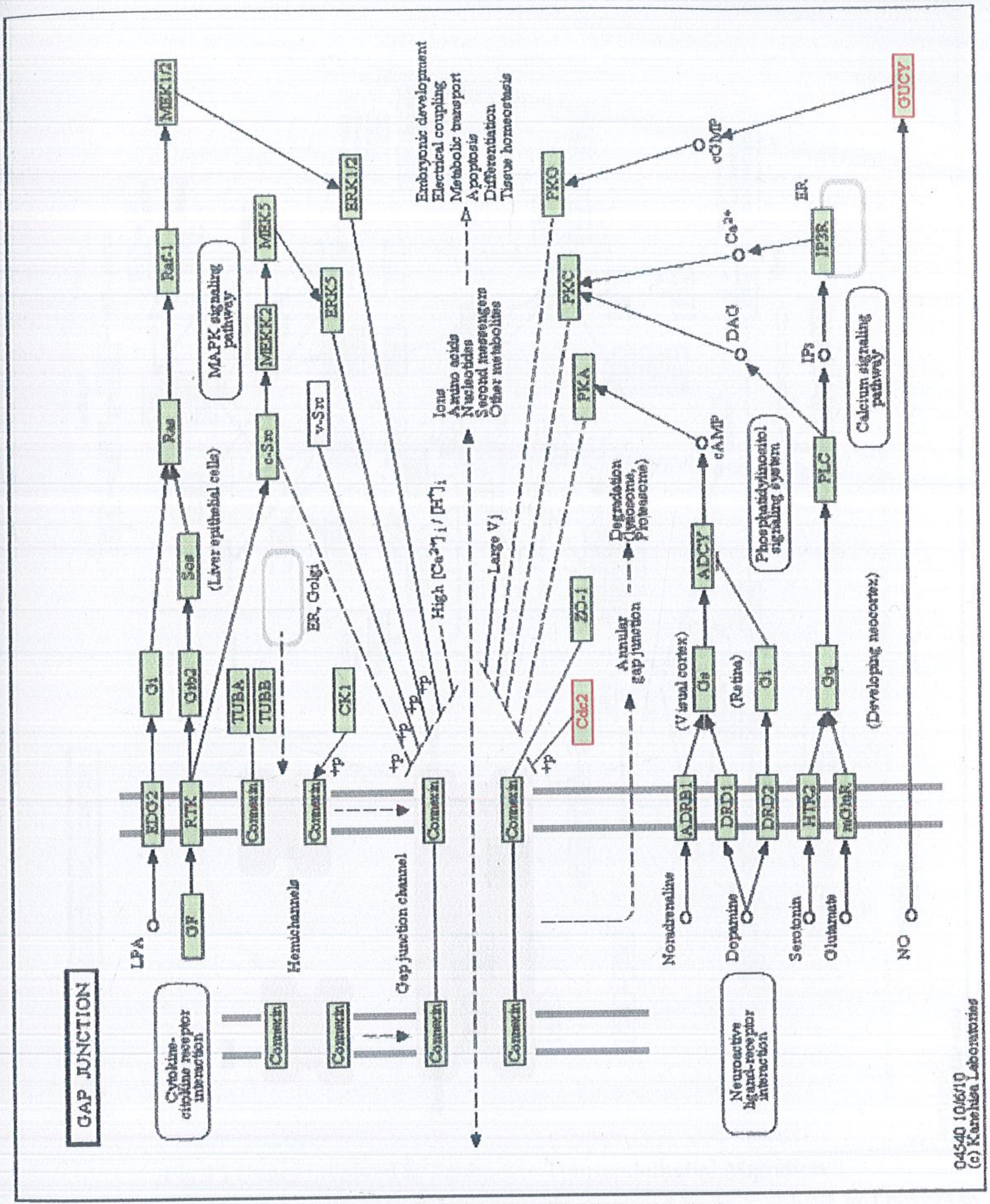


Fig.12 Genes involved in GAP JUNCTION

Listed over expressed genes of ovarian cancer cells were observed to be involved in GAP JUNCTION which regulates the cancer growth.

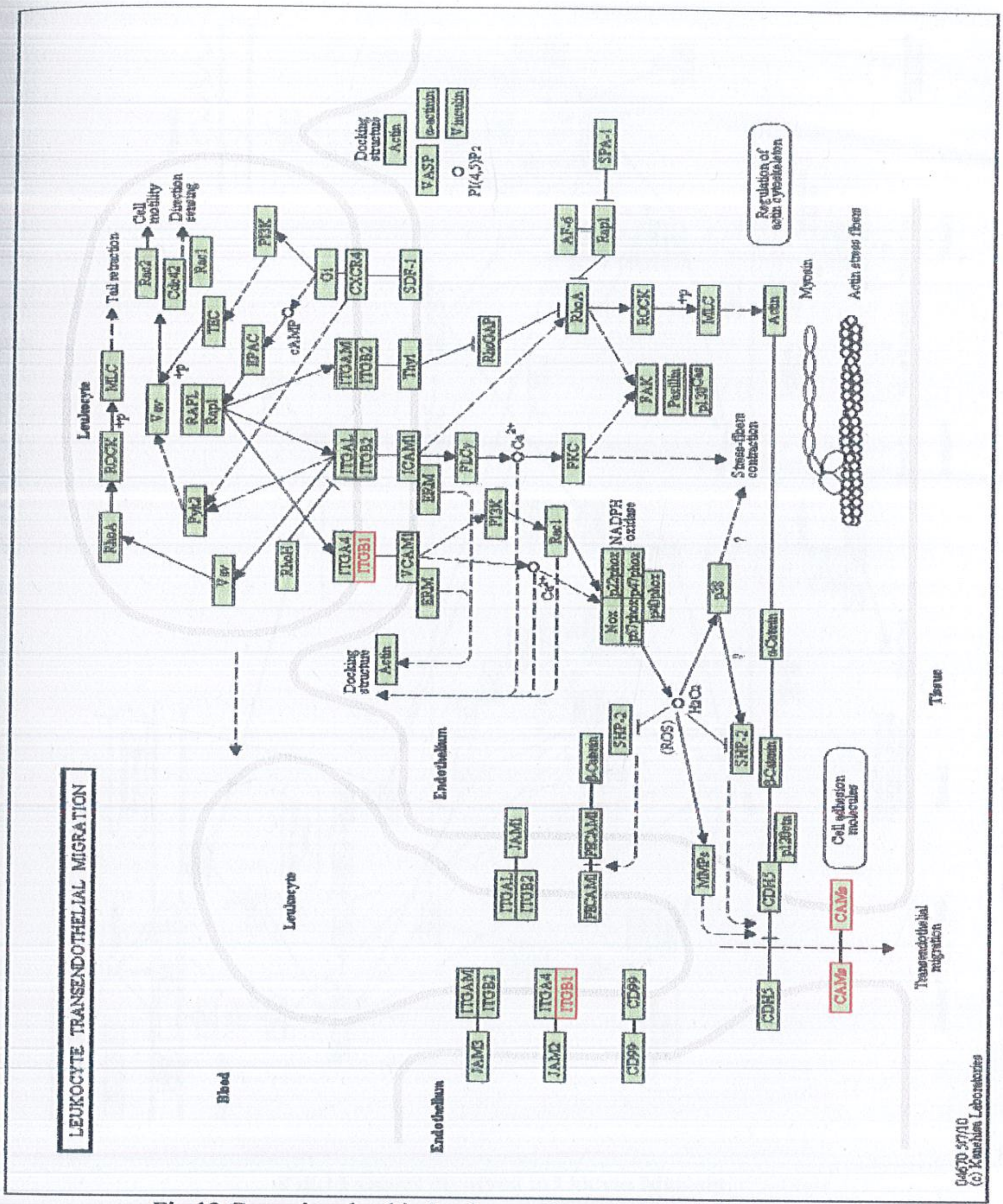


Fig.13 Genes involved in Leukocyte Transendothelial Migration

Genes over expressed in ovarian cell carcinoma were observed to be involved in Leukocyte Transendothelial Migration and increases the cancer growth.

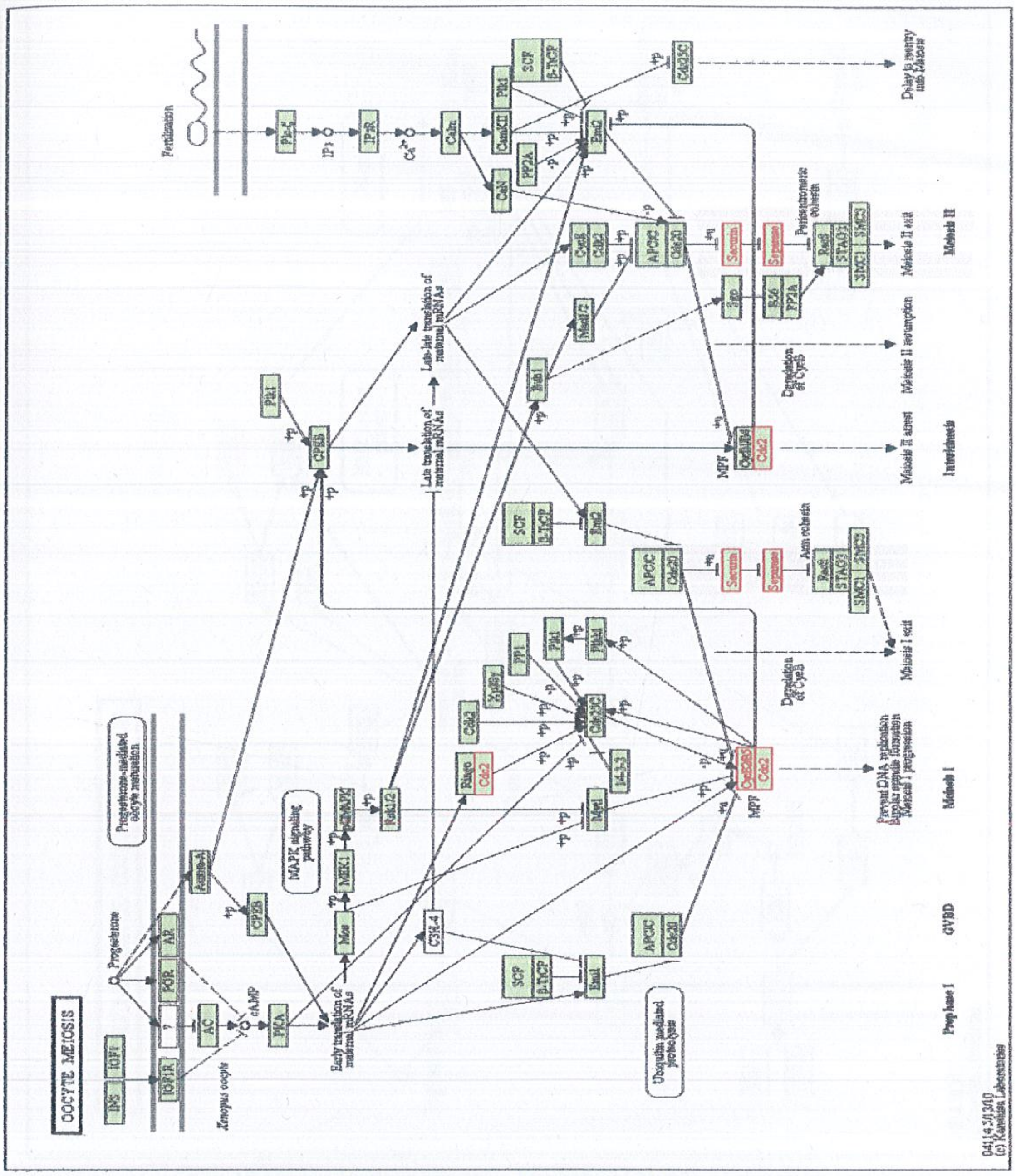


Fig.14 Genes involved in Oocyte Meiosis

The genes marked here are responsible for cell division related anomaly and thus, decreases apoptosis.

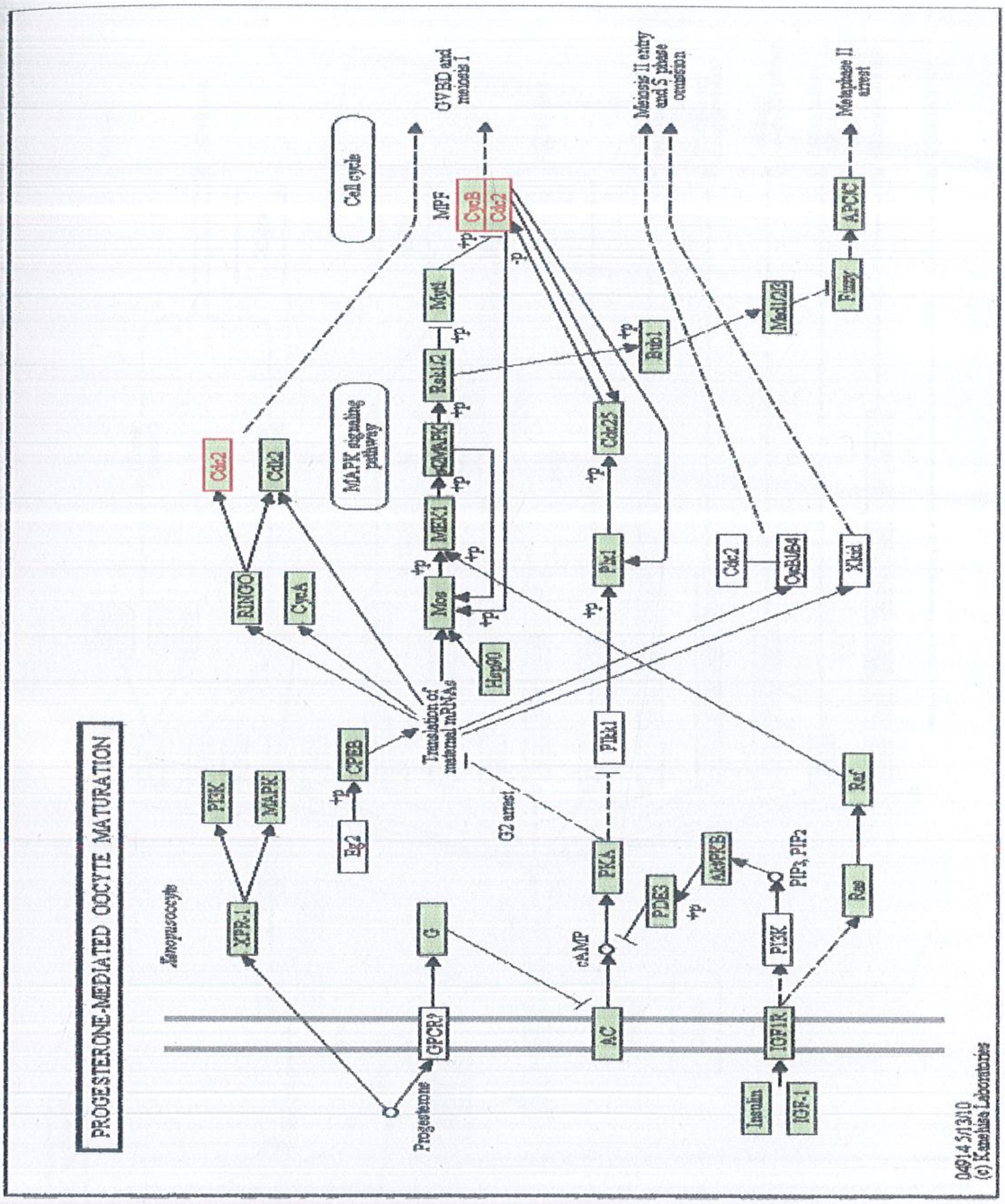


Fig.15 Genes in Progesterone-Mediated Oocyte Maturation

Genes encoding for these cytokines also increased the production of hormones that further sped up the cell division of the ovarian cells.

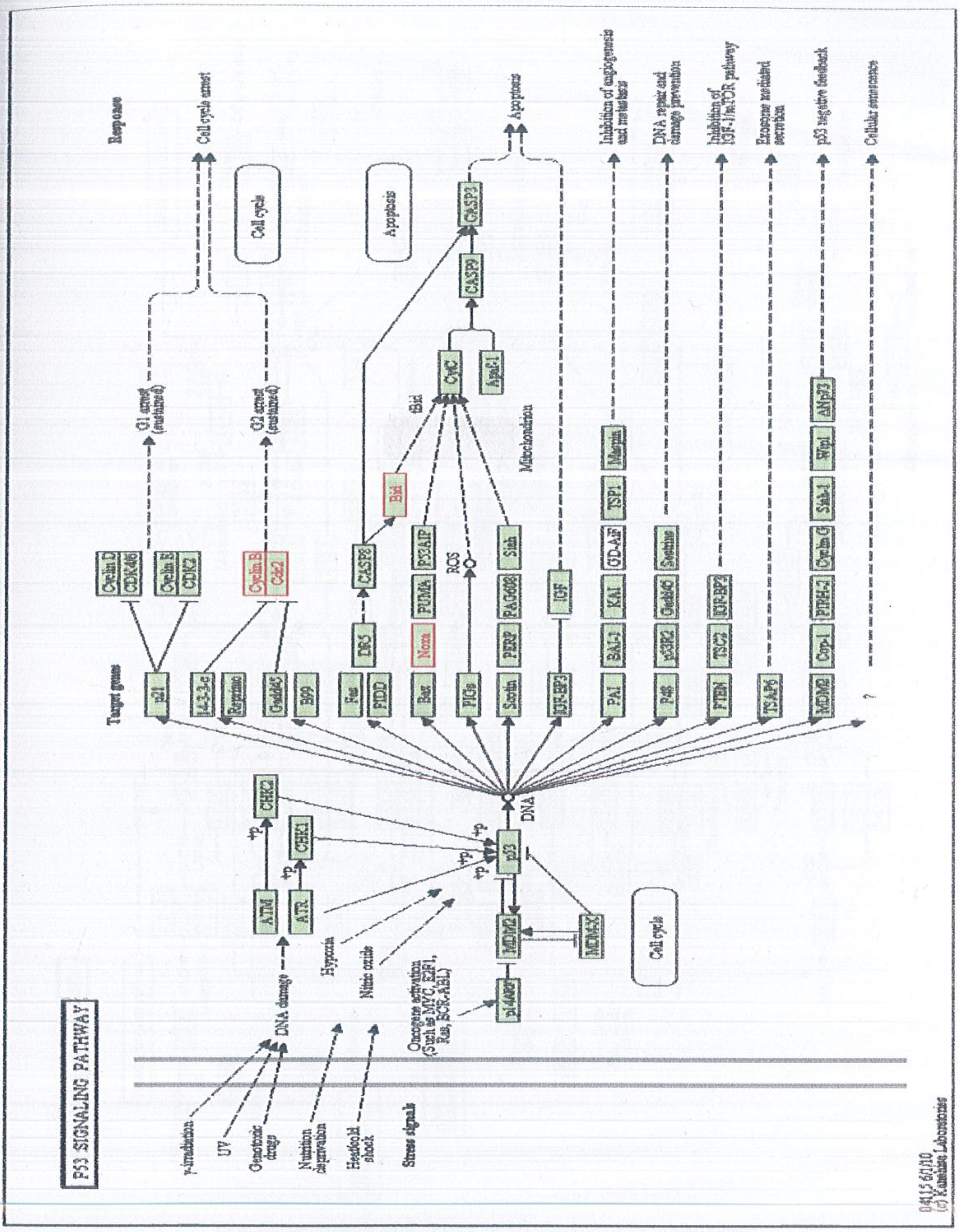


Fig.16 Genes involved in p53 Signaling Pathway

Genes hampered the normal functioning of p53 pathway which delayed apoptosis.

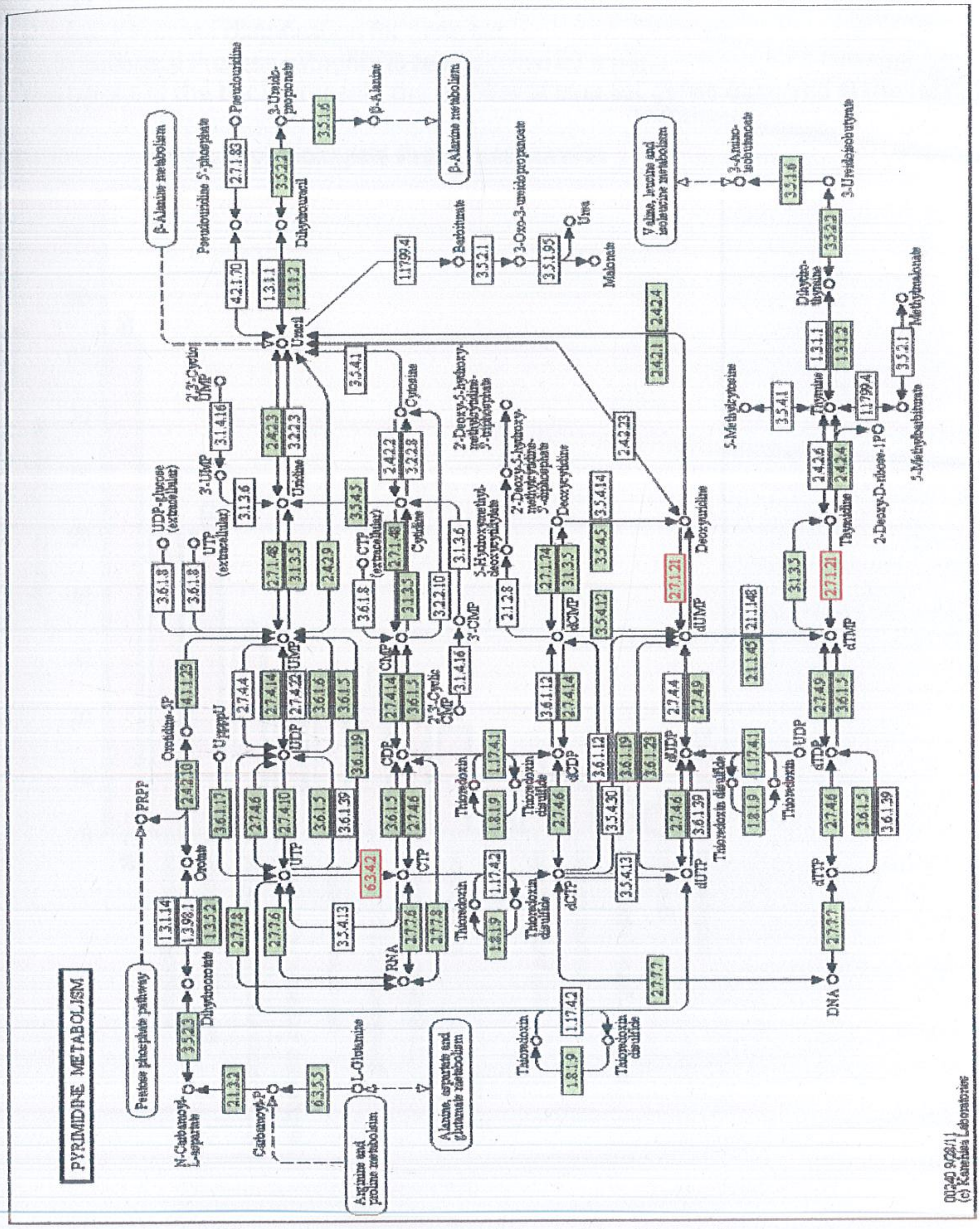


Fig.17 up regulated genes in Pyrimidine Metabolism

Biological Process classification for gene set **probeset_ids.txt**.
Each Biological Process category is represented by a bar.
The height of the bar represents the number of user list genes observed in the category.

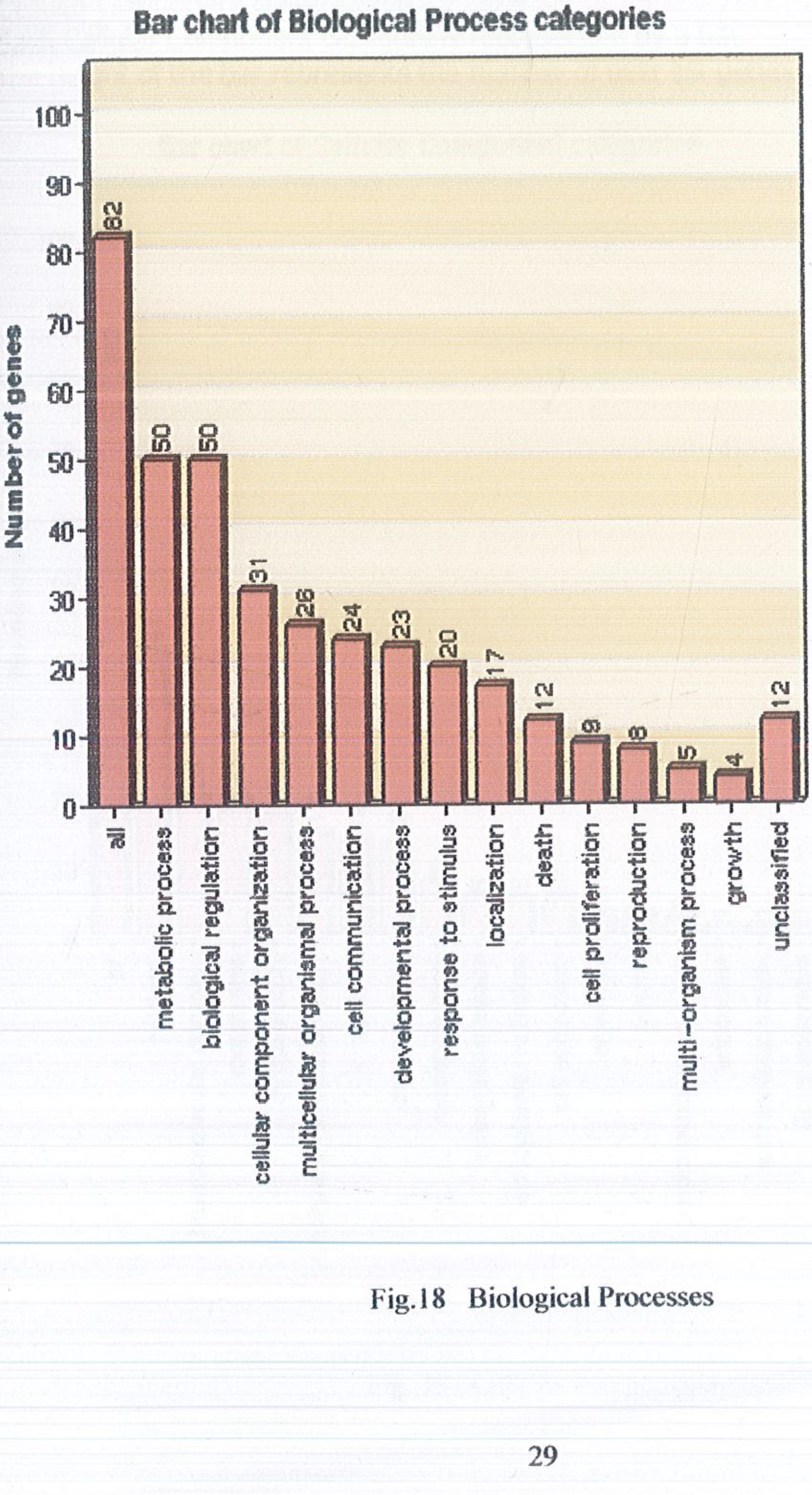


Fig.18 Biological Processes

Out of the total uniquely identified genes, 50 were involved with metabolic process regulation, 50 with biological regulation, 12 that affected apoptosis, 8 triggered the cell division, 4 altered the growth of carcinoma,24 genes hampered the cell signaling pathway.

Cellular Component classification for gene set **probeset_ids.txt**.
Each Cellular Component category is represented by a bar.
The height of the bar represents the number of user list genes observed in the category.

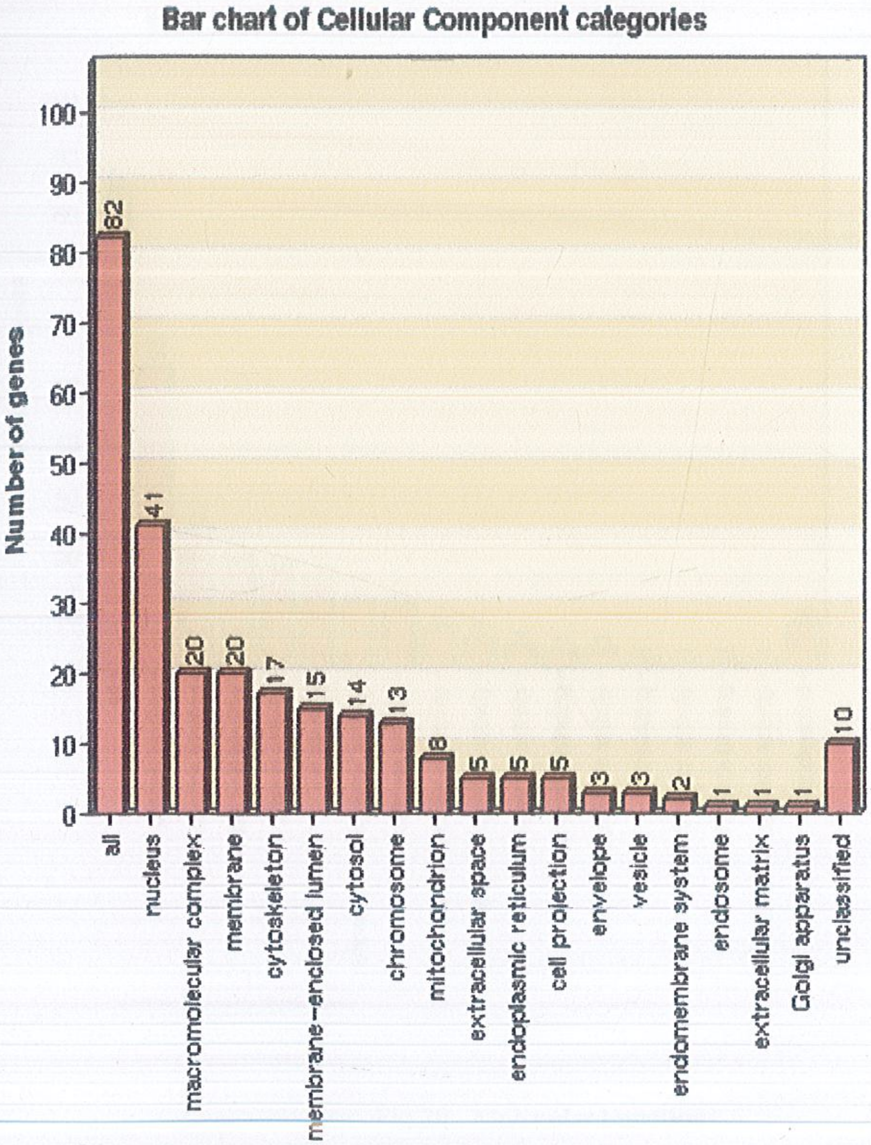
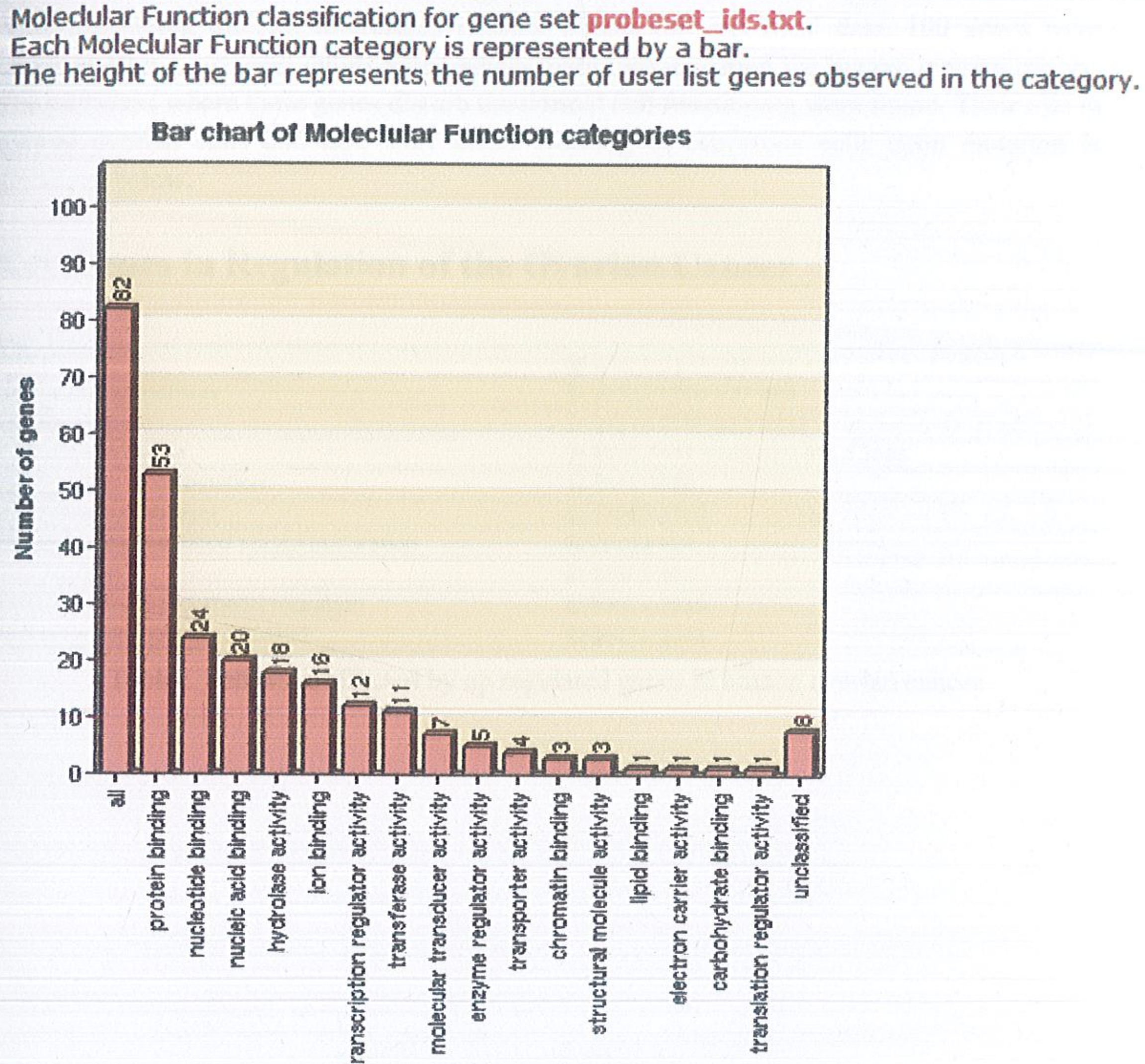


Fig. 19 Cellular components

Of all the 82 genes, 20 genes were found to be present in the membrane and 20 in the macromolecular complex. 14 genes expressed product act in the cytosol. 5 genes were involved in the mechanism of cell protection. 1gene was each mapped to endosome, extracellular matrix and golgi apparatus.



Moleclular Function classification for gene set probeset_ids.txt.
Each Moleclular Function category is represented by a bar.
The height of the bar represents the number of user list genes observed in the category.

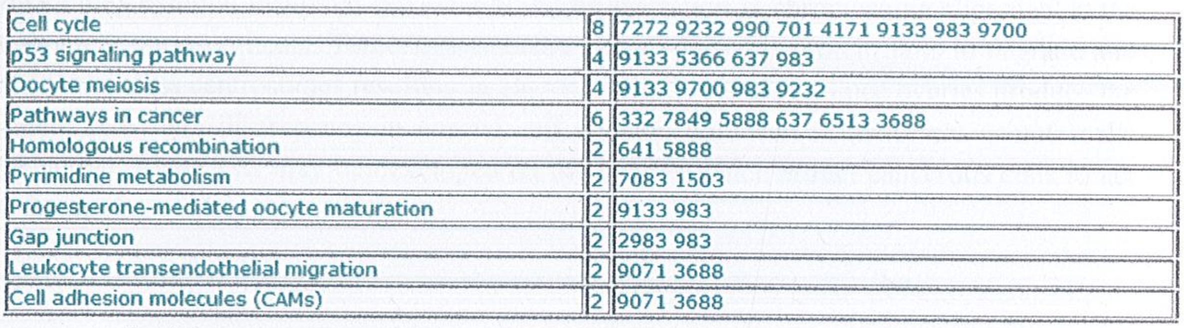
Fig.20 Molecular Function

A total of 53 genes were responsible in faulty protein binding which triggered malfunction of other genes. 4 affected the transporter activity. 8 genes could not be classified as data was not available for them in the database as of now.

CONCLUSION

After processing the raw microarray ovarian cancer cell .cel files' data, 100 genes were obtained after a stringent cutoff value which highly up regulated the human ovarian cancer. The pathways where these genes disturb the normal cell functioning were found. Their role in normal ovarian cells and how they affect working in cancerous cells upon mutation is discussed below.

Key Genes in Regulation of the Ovarian Cancer



Cell cycle	8	7272 9232 990 701 4171 9133 983 9700
p53 signaling pathway	4	9133 5366 637 983
Oocyte meiosis	4	9133 9700 983 9232
Pathways in cancer	6	332 7849 5888 637 6513 3688
Homologous recombination	2	641 5888
Pyrimidine metabolism	2	7083 1503
Progesterone-mediated oocyte maturation	2	9133 983
Gap junction	2	2983 983
Leukocyte transendothelial migration	2	9071 3688
Cell adhesion molecules (CAMs)	2	9071 3688

Table2. Pathways affected by up regulated genes in human ovarian cancer

Cell Signaling Pathway

TTK TTK protein kinase
Gene type: protein coding
Gene ID: 7272
Location: 6q13-q21

Summary

In normal cell functioning, this gene encodes a protein kinase which has an ability to phosphorylate tyrosine, serine and threonine. It is also associated with cell proliferation. Any mutation by external factors in this protein causes aberration in chromosome alignment at the centromere during mitosis. Tumorigenesis may occur when this protein fails to degrade and produces excess centrosomes resulting in aberrant mitotic spindles. Since ovaries produce the gamete cells, any discrepancy in normal cell division will lead to abnormal gamete cells production. Also cysts may be developed on the ovaries which attract cancerous cells to act on them.

PTTG1 pituitary tumor-transforming 1
Gene type: protein coding
Gene ID: 9232
Location: 5q35.1

Summary

In normal cell functioning, the encoded protein is prevents sister chromatid separation. The gene product has transforming activity in vitro and tumorigenic activity in vivo, and the gene is highly expressed in ovarian carcinoma. The gene product contains 2 PXXP motifs, which are required for its transforming and tumorigenic activities, as well as for its stimulation of basic fibroblast growth factor expression. It also contains a destruction box (D box) that is required for its degradation by the APC. So, mutation or even the up regulation of PTTG1 gene inhibits the process of apoptosis.

CDC6 cell division cycle 6 homolog

Gene type: protein coding

Gene ID: 990

Location: 17q21.3

Summary

The protein encoded by this gene is highly essential for the initiation of DNA replication under normal cell working conditions. It is localized in cell nucleus during cell cycle G1 phase, but translocates to the cytoplasm at the start of S phase. When CDC6 is unable to move from nucleus to cytoplasm it creates stress on cell division mechanism which further triggers response from helper genes and an enhanced rate of cell differentiation is promoted. It leads to accumulation of cells with improper disposal of old cells.

p53 Signaling Pathway

CCNB2 cyclin B2

Gene type: protein coding

Gene ID: 9133

Location: 15q22.2

Summary

Cyclin B2 is a member of the cyclin family, specifically the B-type cyclins. The B-type cyclins, B1 and B2, associate with p34cdc2 and are essential components of the cell cycle regulatory machinery. B1 and B2 differ in their subcellular localization. Cyclin B1 co-localizes with microtubules, whereas cyclin B2 is primarily associated with the Golgi region. Cyclin B2 also binds to transforming growth factor beta RII and thus cyclin B2/cdc2 may play a key role in transforming growth factor beta-mediated cell cycle control.

BID BH3 interacting domain death agonist

Gene type: protein coding

Gene ID: 637

Location: 22q11.1

Summary

This gene encodes a death agonist that heterodimerizes with either agonist BAX or antagonist BCL2 in a normal cell. The encoded protein is a member of the BCL-2 family of cell death regulators. It is unable to act as a mediator of mitochondrial damage induced by caspase-8 (CASP8); CASP8 cleaves this encoded protein, and the COOH-terminal part translocates to mitochondria where it triggers cytochrome c release when it is over expressed.

Pathways in Cancer

BIRC5 baculoviral IAP repeat containing 5

Gene type: protein coding

Gene ID: 332

Location: 17q25

Summary

This gene is a member of the inhibitor of apoptosis (IAP) gene family, which encode negative regulatory proteins that prevent apoptotic cell death. IAP family members usually contain multiple baculovirus IAP repeat (BIR) domains, but this gene encodes proteins with only a single BIR domain. The encoded proteins also lack a C-terminus RING finger domain. Gene expression is high during fetal development and in most tumors, yet low in adult tissues. It is expressed in ovaries and any deviation from normal regulation will add another reason of causing ovarian cancer.

PAX8 paired box 8

Gene type: protein coding

Gene ID: 7849

Location: 2q13

Summary

This gene encodes a member of the paired box (PAX) family of transcription factors. Members of this gene family typically encode proteins that contain a paired box domain, an octapeptide, and a paired-type homeodomain. This nuclear protein is involved in thyroid follicular cell development and expression of thyroid-specific genes. Mutations in this gene have been associated with atypical follicular carcinomas. Follicles cover the ovaries and do not allow eggs to form properly. It thus leads to a stressed state where genes try to over exert and sudden rupture may lead to uncontrolled cell growth leading to tumor or cancer formation.

RAD51 RAD51 homolog (S. cerevisiae)

Gene type: protein coding

Gene ID: 5888

Location: 15q15.1

Summary

The protein encoded by this gene is a member of the RAD51 protein family. RAD51 family members are highly similar to bacterial RecA and Saccharomyces cerevisiae Rad51, and are known to be involved in the homologous recombination and repair of DNA. This protein can interact with the ssDNA-binding protein RPA and RAD52, and it is thought to play roles in homologous pairing and strand transfer of DNA. This protein is also found to interact with BRCA1 and BRCA2, which may be important for the cellular response to DNA damage. BRCA2 is shown to regulate both the intracellular localization and DNA-binding ability of this protein. Loss of these controls following BRCA2 inactivation may be a key event leading to genomic instability and tumorigenesis.

BID BH3 interacting domain death agonist

Gene type: protein coding

Gene ID: 637

Location: 2q11.1

Summary

This gene encodes a death agonist that heterodimerizes with either agonist BAX or antagonist BCL2. The encoded protein is a member of the BCL-2 family of cell death regulators. It is a mediator of mitochondrial damage induced by caspase-8 (CASP8); CASP8 cleaves this encoded protein, and the COOH-terminal part translocates to mitochondria where it triggers cytochrome c release in the ovarian cells. Any mal functioning leads to disrupted growth and ovarian cancer in humans.

ITGB1 integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)

Gene type: protein coding

Gene ID: 3688

Location: 10p11.2

Summary

Integrins are heterodimeric proteins made up of alpha and beta subunits. At least 18 alpha and 8 beta subunits have been described in mammals. Integrin family members are membrane receptors involved in cell adhesion and recognition in a variety of processes including embryogenesis, hemostasis, tissue repair, immune response and metastatic diffusion of tumor cells in normal working of cells. This gene encodes a beta subunit and any variation in the expression level leads to human ovarian cancer.

How Genes in Cancer Pathway Regulate Ovarian Cancer

BIRC5 baculoviral IAP repeat containing 5

It is a significant contributor to the development of hormonal therapy resistance in ovarian cancer cells, targeting BIRC5 and blocking it would enhance ovarian cancer cell susceptibility to anti-progesterone therapy. The basis of anti-progesterone therapy involves using drugs that eliminate the presence of progesterone in the cell and cellular environment, since the presence of progesterone are known to enhance tumor immortality in ovarian cancer cells and prostate cancer in men.[14]

PAX8 paired box 8

Pax8 is expressed in embryonal human tissues, in particular in the developing thyroid gland, kidney and nervous system and in the human placenta. Over and under expression of normal PAX8 protein contributes to malignant tumors of the ovary derived from the ovarian surface epithelium.

RAD51 RAD51 homolog (S. cerevisiae)

RAD51D-E233G variant allele has been identified as a potential precursor to ovarian cancers in women with high familial risk but do not possess a BRCA1/BRCA2 mutation[15]. it was determined that the exon 8 mutation led to an increased frequency of ovarian cancer.

BID BH3 interacting domain death agonist

The region around the BID transcriptional start site was methylated in 36% of colorectal and 32% of gastric cancer cell lines and was closely associated with a loss of BID expression in those cell lines[16].

SLC2A1 solute carrier family 2 (facilitated glucose transporter), member 1

This gene is responsible for constitutive glucose uptake. It has a very broad substrate specificity and can transport a wide range of aldoses including both pentoses and hexoses. Any variation in its expression level causes high risk to breast cancer and ovarian cancer.

ITGB1 integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)

This gene is involved in promoting endothelial cell motility and angiogenesis. Also, it is involved in up- regulation of the activity of kinases such as PKC via binding to KRT1. ITGB1 plays a mechanistic adhesive role during telophase, required for the successful completion of cytokinesis. Over expression of this gene causes high risk to breast cancer and ovarian cancer to women.

APPENDIX I

# installing Bioconductor

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

# setting the directory path

```
setwd('/path_to_directory')
```

# loading the "library" that contains the Affymetrix microarray code

```
library(affy)
```

# .CEL files were read, summarized and normalized with RMA normalization method.

```
affy.data <- ReadAffy()
eset.rma <- justRMA()
write.exprs(eset.rma, file="t.csv", sep=",")

exprSet.nologs <- exprs(eset.rma)
```

# List the column (chip) names

```
colnames(exprSet.nologs)
```

# To convert expression values to log form.

```
hist(exprSet.nologs)
exprSet <- log(exprSet.nologs, 2)
hist(exprSet)
```

```
write.table(exprSet, file="Su_mas5_matrix.txt", quote=F, sep="\t")

library(limma)
exprSet.quantile <- normalizeQuantiles(exprSet)

# boxplot before and after normalization

boxplot (exprSet)

# MA plot before and after normalization

par(mfrow=c(2,4))
A<-(exprSet[,1]+exprSet[,2])/2
M<-(exprSet[,1]-exprSet[,2])
plot(A,M,xlab="A",ylab="M",main="rep 3")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
A<-(exprSet[,3]+exprSet[,4])/2
M<-(exprSet[,3]-exprSet[,4])
plot(A,M,xlab="A",ylab="M",main="rep 4")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
A<-(exprSet[,5]+exprSet[,6])/2
M<-(exprSet[,5]-exprSet[,6])
plot(A,M,xlab="A",ylab="M",main="rep 3")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
A<-(exprSet[,7]+exprSet[,8])/2
M<-(exprSet[,7]-exprSet[,8])
plot(A,M,xlab="A",ylab="M",main="rep 4")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
A<-(exprSet.quantile[,1]+exprSet.quantile[,2])/2
M<-(exprSet.quantile[,1]-exprSet.quantile[,2])
```

```
plot(A,M,xlab="A",ylab="M",main="rep 1")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
A<-(exprSet.quantile[,3]+exprSet.quantile[,4])/2
M<-(exprSet.quantile[,3]-exprSet.quantile[,4])
plot(A,M,xlab="A",ylab="M",main="rep 2")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
A<-(exprSet.quantile[,5]+exprSet.quantile[,6])/2
M<-(exprSet.quantile[,5]-exprSet.quantile[,6])
plot(A,M,xlab="A",ylab="M",main="rep 1")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)

A<-(exprSet.quantile[,7]+exprSet.quantile[,8])/2
M<-(exprSet.quantile[,7]-exprSet.quantile[,8])
plot(A,M,xlab="A",ylab="M",main="rep 2")

trend<-lowess(A,M)
lines(trend,col=2,lwd=5)


axprSet <- exprSet.quantile
```

# Calculating log$_2$ ratios

```
normal.mean <- apply(exprSet[, c(1,2,3,4,5,6)], 1, mean)

cancer.mean <- apply(exprSet[, c(7,8,9,10,11,12)], 1, mean)
```

```
normal.cancer.to.human <- normal.mean - cancer.mean

# binding data together in one file

all.data <- cbind(exprSet, normal.mean, cancer.mean, normal.cancer.to.human)

# Checked what data we have here

colnames(all.data)

# Printing data

write.table(all.data,file="Microarray_Analysis_data_1_SOLUTION.txt", quote=F, sep="\t")

# Identifying differentially expressed genes

dataset.1 <- exprSet[1, c(1,2,3,4,5,6,7,8)]

dataset.2 <- exprSet[1, c(7,8,9,10,11,12)]

t.test.gene.1 <- t.test(dataset.1, dataset.2, "two.sided")

brain.p.value.all.genes <- apply(exprSet, 1, function(x)
{ t.test(x[1:2:3:4:5:6], x[7:8:9:10:11:12]) $p.value } )

write.table(brain.p.value.all.genes, file="p_values.txt", quote=F, sep="\t")

# heatmap

# for normal genes

nba <- read.csv("path_to_directory", sep=",")
```

```
nba$Name <- with(nba, reorder(name, abc))
library(ggplot2)

nba.m <- melt(nba)
nba.m <- ddply(nba.m, .(variable), transform,rescale = rescale(value))

(p <- ggplot(nba.m, aes(variable, Gene)) + geom_tile(aes(fill = rescale),colour = "blue") +
scale_fill_gradient2(low = "red",mid="yellow",high = "green", midpoint=0.6))
```

#choosing green colour for normal genes

```
(p <- ggplot(nba.m, aes(variable, Name)) + geom_tile(aes(fill = rescale),colour = "green") +
scale_fill_gradient(low = "#ddfed1",high = "#4A9586"))
```

# For cancer genes

```
nba <- read.csv("path_to_directory", sep=",")
nba$Name <- with(nba, reorder(name, abc))
library(ggplot2)
nba.m <- melt(nba)
nba.m <- ddply(nba.m, .(variable), transform,rescale = rescale(value))
(p <- ggplot(nba.m, aes(variable, Gene)) + geom_tile(aes(fill = rescale),colour = "blue") +
scale_fill_gradient2(low = "red",mid="yellow",high = "green", midpoint=0.6))
```

```
# choosing red colour for cancer genes

(p <- ggplot(nba.m, aes(variable, Name)) + geom_tile(aes(fill = rescale),colour = "red") +
scale_fill_gradient(low = "#ffc8c8",high = "#ff5353"))


# CLUSTERING

nba <- read.csv("path_to_directory", sep=",")

row.names(nba) <- nba$Name
d <- dist(nba, method = "euclidean")

 # distance matrix

fit <- hclust(d, method="ward")
plot(fit) # display dendogram
groups <- cutree(fit, k=5)

 # cut tree into 5 clusters

# draw dendogram with red borders around the 5 clusters

rect.hclust(fit, k=5, border="red")
```

REFERENCES

1. Zhang, B., Kirov, S.A., Snoddy, J.R. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res, 33(Web Server issue),W741-748.
2. Duncan, D.T., Prodduturi, N., Zhang, B. (2010). WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. BMC Bioinformatics, 11(Suppl 4):P10
3. http://www.genome.gov/
4. Kanehisa et al., 2004; Ashburner et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25–29.
5. Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, and Peter Grassberger, "Hierarchical Clustering Based on Mutual Information", (2003) ArXiv q-bio/0311039.
6. http://www.geneontology.org/GO.downloads.ontology.shtml ;Amy Warner, A taxonomy primer.
7. MSigDB. Ashburner, M., *et al.* (2000) Gene ontology: tool for the unification of biology. TheGene Ontology Consortium. *Nat. Genet.*, **25**, 25-29.
8. Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289-300.
9. Dudoit S., Shaffer J.P., and Boldrick J.C., 2003. Multiple hypotheses testing in microarray experiments. Statistical Science 18: 71–103.
10. Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". Journal of the Royal Statistical Society, Series B (Methodological) 57 (1): 289–300.
11. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns". Proc.Natl. Acad. Sci. USA 95 (25):14863-
14868. doi:10.1073/pnas.95.25.14863. PMC 24541. PMID 9843981.
12. Benjamini, Y. (1988). "Opening the Box of a Boxplot". The American Statistician 42 (4): 257–262.
13. DudoitS,Yang,YH,Callow,MJ,Speed,TP(2002).Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.
14. Zhang M, Latham DE, Delaney MA, Chakravarti A (April 2005). "Survivin mediates resistance to antiandrogen therapy in prostate cancer". Oncogene 24 (15):2474–2. doi:10.1038/sj.onc.1208490. PMID 15735703
15. Rodríguez-López et al., 2004; Dowty et al., 2008
16. Obata et al., 2003; Nakamura et al., 2009