META-ANALYSIS OF PROSTATE CANCER MICROARRAY DATASETS

ENROLLMENT NUMBERS: 131502,131512

STUDENT'S NAME: GUNJAN GUPTA, NIKHITA SOOD

SUPERVISOR NAME: Dr. JAYASHREE RAMANA

Submitted in partial fulfillment of the requirement for the award of the Degree
Of
Bachelor of Technology
In
Bioinformatics

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION AND TECHNOLOGY

WAKNAGHAT, SOLAN – 173234, HIMACHAL PRADESH

ACKNOWLEDGEMENT

It is our pleasure to be indebted to various people, who directly or indirectly contributed in the completion of the work till and influenced our thinking, behaviour, and acts during the course of work. We would like to express our profound gratitude to all of them for their constant guidance and inspiration to me.

We express our sincere gratitude to our guide Dr. Jayashree Ramana, Assistant professor at Biotechnology and Bioinformatics Department, Jaypee University of Information Technology for her exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project.

We would hereby express our sincere thanks to all the faculties of Department of Biotechnology and Bioinformatics, Dr. Chittranjan Rout, Dr. Jayashree Ramana, Dr. Tiratha Raj Singh and Dr. Raghu M. Yennamalli whose teachings gave us conceptual understanding and clarity which ultimately made our work easier. We, greatly, acknowledge Dr. R.S. Chauhan, Head of Department for the valuable assistance.

We would also like to thank the almighty and our parents for their moral support and our friends with whom we shared our day-to-day experience and received lots of suggestions that improved the quality of work.

Signature of Student

Name:

Date:

Signature of Student

Name:

Date:

CERTIFICATE

This is to certify that the work titled "**META-ANALYSIS OF PROSTATE CANCER MICROARRAY DATASETS"** towards partial fulfilment for the award of degree of Bachelor of Technology and submitted to the department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, is an authentic record of work carried out by **Gunjan Gupta** and **Nikhita Sood** during period from July 2016- April 2017 under the supervision of **Dr. Jayashree Ramana**, Assistant Professor, Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology.

Dated:

_____

Dr. Jayashree Ramana

Assistant Professor,

Department of Biotechnology and Bioinformatics,

Jaypee University of Information Technology

Table of contents

CHAPTER-1

INTRODUCTION

1.1) OBJECTIVE:

The objective of the project is to find out the genes which are significantly dysregulated in prostate cancer through meta-analysis of large number of microarray dataset and analyze those genes whose biological functions are clearly associated with cancer using cell signalling pathways and to find out new targets for prostate cancer genes and potential biomarkers for prostate cancer.

1.2) META-ANALYSIS:

Meta analysis is defined as "The statistical analysis of a collection of analysis results from individual studies for integrating the findings". Meta-analysis is a formal, epidemiological, quantitative study design used to assess the results of previous research to derive conclusions about that body of research.

Meta-analysis of microarray datasets aids in finding significantly up-regulated genes in cancer which thereby increases the generalizability and statistical power.

Meta-analysis results can improve precision of estimates of effect, answer questions not posed by the individual studies, settle controversies arising from apparently conflicting studies, and generate new hypotheses [1].

CHAPTER-2

BACKGROUND

2.1) PROSTATE CANCER

Cancer is a class of diseases characterized by uncontrollable cell growth**.** It harms the body when altered cells divide uncontrollably to form lumps or tumours. Tumours can grow and interfere with the digestive, nervous, and circulatory systems and they can release hormones that alter body function. Tumours that stay in one spot and demonstrate limited growth are generally considered to be benign. When a tumour successfully spreads to other parts of the body and grows, invading and destroying other healthy tissues, it is said to have metastasized [2].

Prostate cancer is a disease that begins to grow in the prostate - a gland in the male reproductive system.

In majority of cases the prostate cancer starts in the gland cells, therefore called adenocarcinoma. Prostate cancer starts with tiny alterations in the shape and size of the prostate gland cells - Prostatic intraepithelial neoplasia (PIN).

TYPES:

- Benign prostatic hyperplasia
- Prostatic adenocarcinoma
- Small cell carcinoma
- Squamous cell carcinoma
- Prostatic sarcomas
- Transitional cell carcinomas

CAUSES:

- Age: The older the man is, higher the risk.
- Diet: Prostate cancer can also be developed due to lack of Vitamin D.
- Obesity: Obesity raises the risk of prostate cancer.
- Sexually transmitted diseases: Males with gonorrhoea have a higher chance of having prostate cancer.

TREATMENT:

- Radical prostatectomy
- Branchy therapy
- Conformal radiotherapy
- Intensity modulated radiotherapy
- Radioactive injection in advanced stage.

2.2) MICROARRAY

BACKGROUND:

Microarray is defined as a 2D array on a solid substrate which can be a glass slide or a silicon thin-film cell, that assays large amount of any biological material such as DNA, protein, Antibody, Carbohydrates, tissue etc. using high-throughput screening ,parallel processing and detection methods.

MICROARRAY EXPERIMENT:

Microarray experiment is a large scale experiment that involves monitoring the expression levels of thousands of genes simultaneously under a particular condition. Microarray experiment provides information about the relative gene expression levels. It can be used to compare expression of a set of genes from a cell maintained in a particular condition to the same set of genes from a reference cell maintained under normal conditions.

METHODOLOGY:

The methodology of performing a microarray experiment is as follows:

1. Material processing
    a. Array fabrication
    b. Preparation of biological samples to be tested
    c. Extraction and labeling of the RNA from the samples
2. Hybridization
3. Scanning
4. Information Processing
    a. Image quantization
    b. Data normalization and integration
5. Data management
    a. Gene expression data matrix
    b. Gene annotation

c. Sample annotation

6. Data analysis and modeling
7. Gene expression data analysis and mining
8. Generation of new hypothesis from this analysis

APPLICATIONS:

1. **Gene Discovery**:
Microarray technologies help in identification of new genes, determination of their function and expression levels under different conditions.

2. **Disease Diagnosis**:
By performing microarray experiments, the scientists can learn more about different diseases such as heart diseases, mental illness, infectious disease and cancer.
But now, with the use of microarrays it is feasible for the scientists to further classify the various types of cancers on the basis of the variation in the gene expression levels in the tumor cells as compared to the normal cells. This in turn, will aid the pharmaceutical community to develop more effective drugs because now the treatment strategies will be targeted directly to the specific type of cancer.

3. **Drug Discovery**:
Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. This field extensively employs microarray technologies. On performing comparative analysis of genes from diseased and normal cells, the identification of biochemical constitution of proteins synthesized by diseased genes can be done.

4. **Toxicological Research**:
Microarray technologies serve as a robust platform for research of the impact of toxins on the cells and their passing on to the progeny. The field of toxic genomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants.

CHAPTER -3

MATERIALS AND METHODS

3.1) DATA COLLECTION

3.1.1) Oncomine:

Oncomine is a cancer-profiling database that contains published data that has been collected, annotated and analyzed by Compendia Bioscience. The data in Oncomine is normalized and analyzed using standard protocols, and presented through a web-based interface [3].

Type of analysis available in Oncomine includes:

- Differential gene expression
- Co-Expression
- Outlier Analysis
- Comparing
- Concept List
- Concept Associations

As per the requirements of the project, the data we needed is microarray data and Oncomine is repository of microarray data as we want to study Differential gene expression using microarray data."

Out of all the datasets for "Prostate Carcinoma" we selected the datasets "Normal v\s Prostate Cancer" i.e. datasets which have genes present in normal samples and genes dysregulated in cancer samples.

3.1.2) GEO (Gene Expression Omnibus):

GEO is a public functional genomics data repository accepting Array- and sequence-based data. Various tools are available to help users query and download experiments and curated gene expression profiles [4].

GEO provides a robust, versatile primary data archive database in which we can efficiently store a wide variety of high-throughput functional genomic data sets, offers simple submission procedures and formats that support complete and well-annotated data deposits from the research community.

Data in GEO is organized as:

- **GEO Platform**: A Platform record contains a description of the array or sequencer. A Platform references many Samples that have been deposited by multiple submitters.
- **GEO Sample**: A Sample contains description of the biological material and the experimental protocols to which it was subjected. A Sample entity must reference only one Platform and may be included in multiple Series.
- **GEO Series**: A Series record links together a group of related Samples and provides a focal point and description of the whole study.

To collect the data for Meta –analysis we required "Series" data for "Prostate Cancer". Therefore the following query was used to retrieve the data:

"(prostate cancer OR adenocarcinoma) AND "Homo sapiens"[porgn:__txid9606] NOT MicroRNA NOT shRNA NOT xenograft NOT miR NOT lung NOT renal NOT breast "

3.2) SOFTWARE AND TOOLS USED

3.2.1) MeV (MultiExperiment Viewer):

MeV is a program included in TM4 software. It is an application that allows the viewing of processed microarray slide representations and the identification of genes and expression patterns of interest [5].

MeV can accept data in many formats including Loading TIGR Array Viewer (.tav) File, Loading Tab Delimited, Multiple Sample (.txt) Files (TDMS Format), Loading Affymetrix Data (.txt, .TXT) Files etc.

After loading the data the initial view of the loaded data is "Main Expression Image". Other information which can be obtained from the results includes:

- Result Navigation Tree
- The History Node and Log
- Data Source Selection Information
- Up-regulated genes
- Down-regulated genes
- Non Significant genes
- Total genes

3.2.2) SAM (Significance Analysis of Microarrays):

SAM is a tool included in MeV that is used to pick out significant genes based on differential expression between sets of samples. It is useful when there is a-priori hypothesis that some genes will have significantly different mean expression levels between different sets of samples [6].

For Significant Analysis of Microarrays we need to group our samples in two classes i.e. Normal samples and Cancer samples and then perform SAM analysis.

SAM generates an interactive plot of the observed vs. expected (based on the permuted data) d-values.SAM outputs a SAM graph viewer, a Delta table viewer .SAM also provides information about dysregulated genes, non-significant genes etc.

SAM identifies statistically significant genes by carrying out gene specific t- tests and
Computes a statistic dj for each gene j, which measures the strength of the relationship between gene expression and a response variable.

3.2.3) RStudio-BIOCONDUCTOR-RANKPROD:

R Studio is a set of integrated tools designed to help users to be more productive with R. R Studio includes many packages which are used for biological data analysis. For differential gene expression data the most used package is "BIO-CONDUCTOR" package.

Bio conductor is an open source open development software project to provide tools for the analysis and comprehension of high-throughput genomic data. Bio conductor provides widespread access to a broad range of powerful statistical and graphical methods for the analysis of genomic data. It also facilitates the inclusion of biological metadata in the analysis of genomic data, e.g. literature data from Pub Med, annotation data from Entrez genes.

Bio conductor includes Rank Product method for microarray data analysis and to find significant genes.

RankProd is a non-parametric method for identifying differentially expressed (up- or down-regulated) genes based on the estimated percentage of false predictions (pfp).

The Bio conductor package Rank Prod provides a new and intuitive tool for meta-analysis in detecting differentially expressed genes under two experimental conditions. The package modifies and extends the rank product to integrate multiple microarray studies from different laboratories and platforms .The significance of the detection is assessed by a non-parametric permutation test, and the associated P-value and false discovery rate (FDR) are included in the output alongside the genes that are detected by user-defined criteria. A visualization plot is provided to view actual expression levels for each gene with estimated significance measurements.The function plotRP can be used to plot a graphical display of the estimated pfp

vs. number of identified genes using the output from RankProducts or RP.advance. The function topGene generates a table of the identified genes based on user-specified selection criteria [7].

3.2.4) RANKPRODIT:

Rank Prodit is a web interface developed in haXe. Rank Prodit accepts replicated samples datasets for at least two conditions therefore it can be applied for microarray data analysis. It accepts tab-delimited text file.Once the data is submitted, it is imported into R and performs RankProduct or Rank Sum analysis. [8].

3.2.5) DAVID:

The DAVID Gene Functional Classification Tool uses a novel agglomeration algorithm to condense a list of genes or associated biological terms into organized classes of related genes or biology, called biological modules [9].

Using DAVID, we can do the following annotations:

1. **Individual annotations include:**
    - Diseases
    - Functional Categories
    - Gene Ontology
    - General Annotation
    - Literature
    - Main Accessions
    - Pathways
    - Protein Domains
    - Protein Interactions
    - Tissue Expressions
2. **Combined annotations include:**
    - Functional Annotation Clustering
    - Functional Annotation Chart
    - Functional Annotation Table

3.2.6) GeneMANIA:

Gene MANIA is a user-friendly and flexible web interface for generating hypotheses about gene function, analysing gene lists and prioritizing genes for functional assays [10].
Input to Gene MANIA is the query list and Gene MANIA extends the list with functionally similar genes that it identifies using available genomics and proteomics data. Gene MANIA also reports weights that indicate the predictive value of each selected data set for the query.

Nine types of network can be created in Gene MANIA:

1) **Co-expression**: Two genes are linked in a co-expression network if their expression levels are similar across conditions in a gene expression study.

2) **Physical Interaction**: Two gene products are linked in a physical interaction network if they were found to interact in a protein-protein interaction study.
3) **Genetic interaction**: Two genes are functionally associated in genetic interaction network if the effects of perturbing one gene were found to be modified by perturbations to a second gene.

4) **Shared protein domains**: Two gene products are linked in a shared protein domain network if they have the same protein domain.

5) **Co-localization**: Two genes are linked in a co-localization network if they are both expressed in the same tissue or if their gene products are both identified in the same cellular location.
6) **Pathway**: Two gene products are linked in a pathway if they participate in the same reaction within a pathway
7) **Predicted**: Predicted functional relationships between genes, often protein interactions.
8) **Other**: Networks that do not fit into any of the above categories.
9) **Uploaded**: Networks that we upload.

CHAPTER 4

DATASET AND METHODOLOGY

4.1) METHODOLOGY

The methodology adopted is as follows:

1) Download Prostate cancer datasets from Oncomine database which satisfy following criteria:
   a) Include both normal & tumor samples with each type having a count of more than one.
b) Experiment type is mRNA.

2) For each dataset:
a) Determine significant genes using SAM package in MeV software.
   b) If there are 0 negative significant genes, omit the corresponding dataset from further analysis.

3) Consider all datasets:

a. For each pair of microarray datasets A and B, calculate Z score using

$$Z = \frac{(R_{obs} - n_B P_A)}{\sqrt{n_B P_A (1 - P_A)}}$$

Where
R_{obs} is the number of genes up-regulated in both A and B
n_B is number of genes up-regulated in B
P_A is probability of gene being up-regulated in A

Reject datasets having Z score < 1.96

4) For each selected dataset:
a. Rank significantly up-regulated genes using RankProd
   b. Select genes having pfp value < 0.15
5) Combine all selected genes & remove duplicates.
6) Perform functional annotation of selected genes using DAVID.
7) Create a network using GeneMANIA.
8) For each network,
   a. For each gene

i. Determine the pathways in which the gene is involved using KEGG.
ii. Assign each pathway the GeneMANIA score of its corresponding gene.

b. For each pathway
i. Calculate the combined score and number of times it is present
c. Classify enriched pathways based on the class

4.2) DATASET

In the study, out of 13 datasets (Table 1) we choose 6 datasets for further analysis as they contained a differential analysis of tumour and normal samples experiment type was mRNA and number of samples in both tumor and normal category was more than one. The datasets were taken from Oncomine because the major advantage of using this database is that prior to inclusion in Oncomine database, microarray, datasets obtained from public resources such as Stanford Microarray Database and the NCBI Gene Expression Omnibus [4] or literature sources are reviewed by a panel of experts to ensure that they meet certain quality standards.

SERIES ID	DATAS-ET NAME	CANC-ER TYPE	GENES	PLATFORM	NUMB-ER OF SAMPL ES	DATASET SUMMARY
GSE5594 5	M. Simo Arredoua n	Human clinical prostate cancer	54675	Affymetrix Human Genome U133 Plus 2.0 Array	21	Gene expression of human prostate benign and malignant tissue to identify potential biomarkers and immunotherap y targets.
***GSE622 18**	Carmen Priolo dataset	human prostate cancer		Affymetrix Mapping 250K Sty2 SNP Array	65	Use of genomics to determine how many of the tumors had a genetic aberration

***GSE56288**	Fugen Li Dataset.	human prostate cancer		Illumina HiSeq 2000 (Homo sapiens) Illumina MiSeq (Homo sapiens) Illumina NextSeq 500 (Homo sapiens)	29	The androgen receptor (AR), a nuclear transcription factor (TF), is consistently reprogrammed during prostate tumorgenesis
***GSE39603**	Liang Goh Dataset	Human prostate cancer.	1505	Illumina Golden Gate Methylation Cancer Panel I	111	Genome wide DNA methylation profiling of normal and Prostate cancer samples.
***GSE38240**	Martin Aryee dataset	Human prostate cancer		Illumina HumanMethylation450 Bead Chip (HumanMethylation450_15017482)	12	DNA methylation profiling of normal prostates from organ donors.
***GSE3933**	Lapointe J Prostate cancer	Human prostate cancer	42951	SHBB SHCQ SHBW	112	gene expression in 62 primary prostate tumors, as well as 41 normal prostate specimens and nine lymph node metastases, using cDNA microarrays .

***GSE38073**	Martin Aryee dataset	Human prostate cancer.		Affymetrix Genome-Wide Human SNP 6.0 Array	198	DNA methylation profiling of normal prostates.
***GSE38043**	Alan Lap-Yin Pang dataset	Human prostate cancer	54675	Affymetrix Human Genome U133 Plus 2.0 Array	6	Analyzed the global gene expression pattern of Tregs between healthy donors and prostate cancer patients.
GSE71016	Chad Creighton dataset	Human prostate cancer	62976	Agilent-039494 Sure Print G3 Human GE v2 8x60K Microarray 039381 (Feature Number version)	95	Gene expression profiling of African-American prostate cancer
GSE6919	Federico Alberto Monzon dataset	Human prostate cancer	12553-GPL92 12646-GPL93 12625-GPL8300	Affymetrix Human Genome U95B Array Affymetrix Human Genome U95C Array Affymetrix Human Genome U95 Version 2 Array	504	Expression Data from Normal and Prostate Tumour Tissues
***GSE3325**	Jianjun Yu dataset	Human prostate cancer	None	Affymetrix Human Genome U133 Plus 2.0 Array	19	Integrative Genomic and Proteomic Analysis of Prostate Cancer Reveals Signatures of Metastatic Progression

***GSE45016**	Kenji Tamura dataset	Human prostate cancer	None	Affymetrix Human Genome U133 Plus 2.0 Array	11	Expression data from High-grade prostate cancer cells
***GSE26910**	Paolo Provero dataset	Human prostate cancer	none	Affymetrix Human Genome U133 Plus 2.0 Array	24	Stromal molecular signatures of breast and prostate cancer

Table1: Prostate cancer microarray datasets included in the study

NOTE: The datasets which are marked as bold with asterisk are not being considered for further analysis because their "Series Matrix File" obtained from official website of GEO contained some errors.

CHAPTER -5
RESULTS AND DISCUSSION

5.1) INITIAL DATASETS:

The initial datasets obtained from Oncomine were:

- GSE56288
- GSE39603
- GSE3933-GPL3289
- GSE3933-GPL3044
- GSE3933-GPL2695
- GSE71016
- GSE38043
- GSE55945
- GSE6919-GPL8300
- GSE6919-GPL93
- GSE6919-GPL92

5.2) SIGNIFICANT GENES IDENTIFICATION IN EACH DATASET:

In order to compute Z-Score to filter the significant datasets we need to identify up-regulated genes in each dataset.SAM was used for this purpose using the software MeV.

5.2.1) Lian Goh Dataset:

CLUSTER INFORMATION

NUMBER OF SAMPLES	111
TOTAL NUMBER OF GENES	1505
SIGNIFICANT GENES	89
UPREGULATED GENES	0
DOWNREGULATED GENES	89
NON SIGNIFICANR GENES	1416

Table 2: Lian Goh Dataset

5.2.2) Lapointe J Dataset:

GPL3044 CLUSTER INFORMATION

NUMBER OF SAMPLES	45
TOTAL NUMBER OF GENES	42951
SIGNIFICANT GENES	488
UPREGULATED GENES	462
DOWNREGULATED GENES	26
NON SIGNIFICANR GENES	42463

Table 3: Lapointe J. Dataset (GPL3044)

GPL2695 CLUSTER INFORMATION:

NUMBER OF SAMPLES	26
TOTAL NUMBER OF GENES	44160
SIGNIFICANT GENES	120
UPREGULATED GENES	69
DOWNREGULATED GENES	51
NON SIGNIFICANR GENES	44040

Table 4: Lapointe J Dataset (GPL2695)

5.2.3) Zhang Y Dataset:

CLUSTER INFORMATION

NUMBER OF SAMPLES	95
TOTAL NUMBER OF GENES	62976
SIGNIFICANT GENES	410
UPREGULATED GENES	346
DOWNREGULATED GENES	64
NON SIGNIFICANR GENES	62566

Table 5: Zhang Y Dataset

5.2.4) Lu B Dataset:

CLUSTER INFORMATION:

NUMBER OF SAMPLES	21
TOTAL NUMBER OF GENES	54675
SIGNIFICANT GENES	63
UPREGULATED GENES	0
DOWNREGULATED GENES	63
NON SIGNIFICANR GENES	54612

Table 6: Lu B Dataset

5.2.5) Yin Pang Dataset:

CLUSTER INFORMATION

NUMBER OF SAMPLES	6
TOTAL NUMBER OF GENES	54675
SIGNIFICANT GENES	11
UPREGULATED GENES	9
DOWNREGULATED GENES	2
NON SIGNIFICANR GENES	54664

Table 7: Yin Pang Dataset

5.2.6) Federico Alberto Dataset:

GPL92 CLUSTER INFORMATION

NUMBER OF SAMPLES	168
TOTAL NUMBER OF GENES	12553
SIGNIFICANT GENES	312
UPREGULATED GENES	199
DOWNREGULATED GENES	113
NON SIGNIFICANR GENES	122241

Table 8: Federico Alberto Dataset (GPL92)

GPL93 CLUSTER INFORMATION:

NUMBER OF SAMPLES	165
TOTAL NUMBER OF GENES	12646
SIGNIFICANT GENES	193
UPREGULATED GENES	170
DOWNREGULATED GENES	23
NON SIGNIFICANR GENES	12453

Table 9: Federico Alberto Dataset (GPL93)

GPL8300 CLUSTER INFORMATION:

NUMBER OF SAMPLES	171
TOTAL NUMBER OF GENES	12625
SIGNIFICANT GENES	758
UPREGULATED GENES	437
DOWNREGULATED GENES	321
NON SIGNIFICANR GENES	11867

Table 10: Federico Alberto Dataset (GPL8300)

MeV RESULTS:

Main Expression Image:

Expression Image viewer gives an image in which each column represents a single sample and each row represents a gene. The names of the samples are displayed vertically above each column. MeV expects that each sample loaded will have the same number of elements, in the same order, and that each gene (spot) is aligned with that element in every other sample loaded.

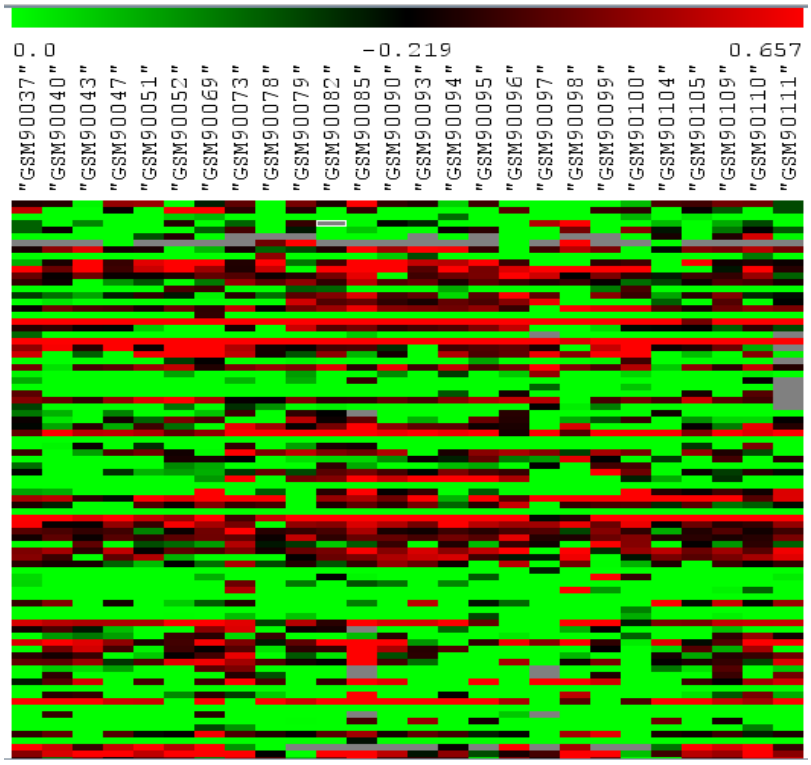

Figure 1: Expression Profile for Lapointe J. dataset

In the above figure red colour represents over-expression of genes; Green colour represents under-expression of genes and black represents normal expression of genes [12].

SAM Graph:

SAM generates an interactive plot of the observed vs. expected (based on the permuted data) d-values. We change the value of the tuning parameter delta using either the slider bar or the text input field below the plot.

Delta is a vertical distance from the solid line of slope 1 (i.e., where observed = expected).

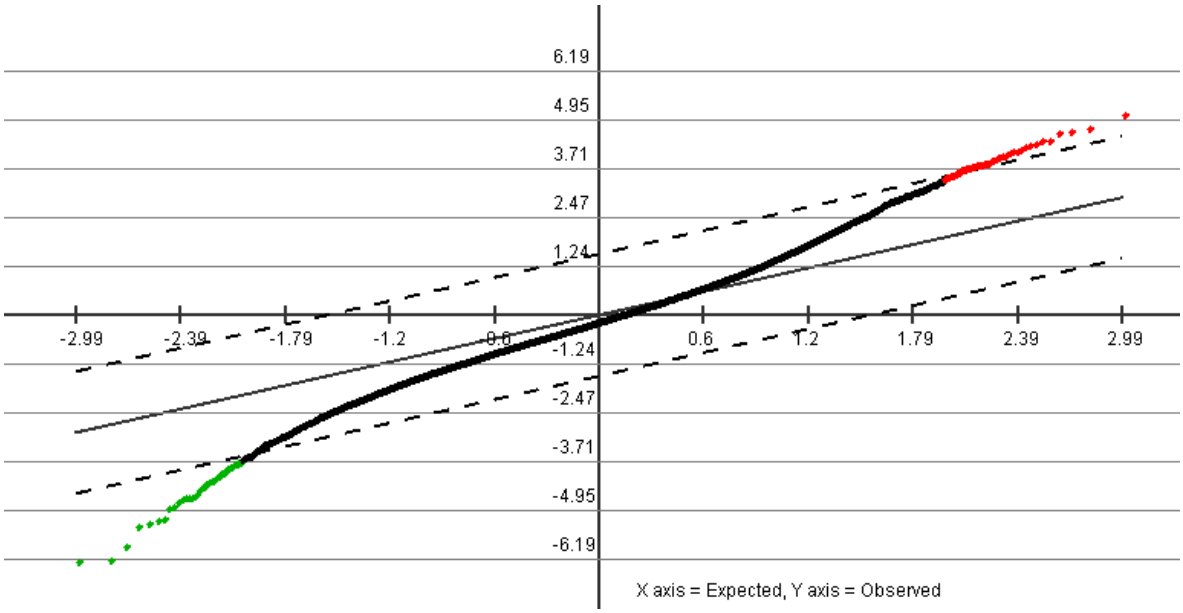

Figure 2: SAM Graph for Lapointe J. dataset (GSE3933)

The two dotted lines represent the region within +/- delta units from the "observed = expected" line. The genes whose plot values are represented by black dots are considered non-significant, those coloured red are positive significant, and the green ones are negative significant.

Expression Profile:

The Expression Profile in SAM gives us the information about

- Positive genes
- Negative genes
- Significant genes
- Non Significant genes
- All genes

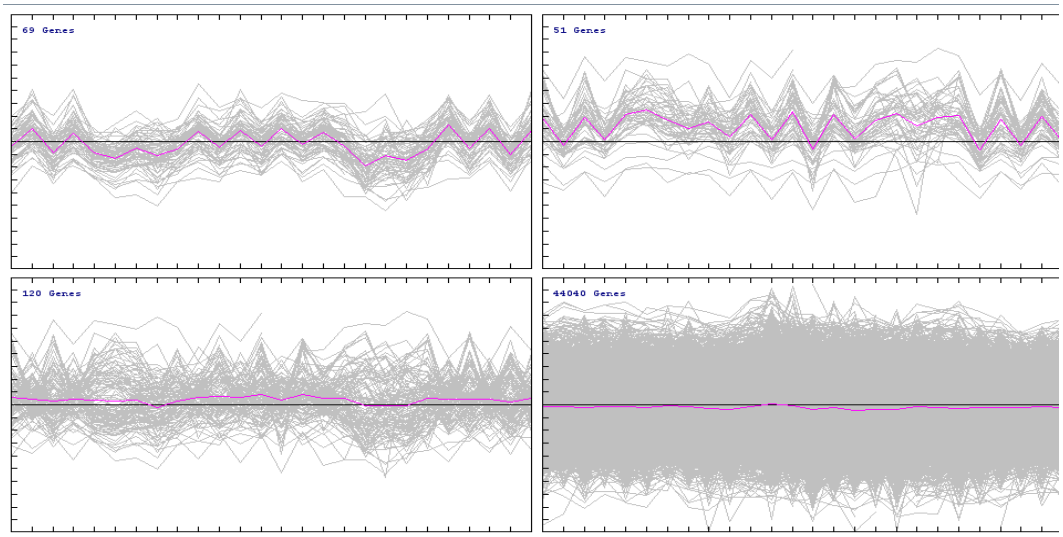

Figure 3: Expression Profile for Lapointe J. dataset (GSE3933)

The above Expression Profile provides the following information:

- 49 up regulated genes
- 51 down regulated genes
- 120 significant genes
- 44046 non-significant genes

LIST OF SIGNIFICANT GENES AND ASSOCIATED VALUES:

SAM lists the "table view" results for all the significant genes as well as non-significant genes. Listed below is the table view of the significant genes which are being up regulated or down regulated in prostate cancer patients.

Original row	ID_REF	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)	Fold change(Unlogged)
257	49091_at	-1.90418	4.436339	486.877	109.7475	NaN
7200	61632_at	0.168516	4.489354	299.1211	66.62899	NaN
9909	64341_f_at	0.746006	4.490792	654.5422	145.7521	NaN
8979	63411_at	0.526701	4.546404	205.8363	45.27453	NaN
9684	64116_at	0.6896	4.546545	105.3338	23.16787	8.19E+17
5385	59809_f_at	-0.17047	4.573236	4961.995	1085.007	NaN
9713	64145_at	0.696581	4.618675	144.2607	31.23421	Infinity
4843	58624_at	-0.27402	4.633412	151.5462	32.70725	Infinity
4149	57261_f_at	-0.41316	4.721728	6668.336	1412.266	NaN
8549	62981_at	0.435492	4.799923	262.1627	54.61809	NaN
4925	58791_f_at	-0.25809	5.091382	5828.236	1144.726	NaN
9855	64287_s_at	0.732202	5.132242	9996.646	1947.813	NaN
8886	63318_at	0.506343	5.216722	1646.179	315.5581	NaN
10017	64449_at	0.773819	-6.17365	-1606.92	260.2875	NaN
4927	58798_at	-0.25773	-6.04575	-367.551	60.79488	NaN
11302	65734_at	1.187869	-5.68802	-722.82	127.0774	NaN
11288	65720_at	1.182016	-5.66464	-1115.55	196.9313	NaN
8048	62480_at	0.333701	-5.65931	-633.39	111.9202	NaN
11473	65905_at	1.262712	-5.63645	-1139.47	202.1608	NaN
9537	63969_f_at	0.654405	-5.61547	-3719.69	662.4	NaN
10319	64751_at	0.856311	-5.53709	-604.415	109.1575	NaN
11431	65863_at	1.244188	-5.4194	-251.003	46.31561	NaN
5626	60058_at	-0.1258	-5.35507	-311.386	58.14789	NaN
9366	63798_at	0.614215	-5.3299	-347.566	65.2105	NaN
1110	50938_at	-1.26068	-5.32156	-509.836	95.80561	NaN
7360	61792_at	0.199309	-5.29498	-78.2843	14.78462	1.52E-09

Table 11: list of significant genes of Lapointe J. Dataset obtained from SAM.

NOTE: The SAM analysis has been shown for Lapointe J. dataset. In a similar way analysis was performed for all the datasets.

5.3) CONGRUENCY BETWEEN MICROARRAY DATASETS

To determine the congruency between microarray datasets, Z score is calculated for all possible pairs of sub-datasets using a Perl script.

The formula used is:

$$Z = \frac{(R_{obs} - n_B P_A)}{\sqrt{n_B P_A (1 - P_A)}}$$

Where Robs is the number of significant genes in both datasets A and B,
nB is the number of genes in dataset B and
PA is the probability of gene being significantly up-regulated in A.
The datasets which that had pair wise absolute z score > 1.96 were considered for further analysis.

	GSE38043	GSE6919	GSE55945	GSE39603	GSE71016
GSE38043	0	673.71	814.21	0	1594.0
GSE6919		0	2477.26	0	4432
GSE55945			0	0	2383
GSE39603				0	0
GSE71016					0

Table 12: Pair wise Z-Score of selected datasets

The above pair wise Zscores were then normalized using "Standard Normalization Method". Normalization was performed so that all the values fits into normal distribution curve and the

standardized values can be compared to 1.96 to choose only the significant datasets for further analysis.

The standard or absolute Z-Score is calculated as:

$$z\ score = \frac{(x - \mu)}{\sigma}$$

Where:

μ is the mean of the population.

σ is the standard deviation of the population.

The absolute value of z represents the distance between the raw score and the population mean in units of the standard deviation. Z is negative when the raw score is below the mean, positive when above.

	GSE38043	**GSE6919**	**GSE55945**	GSE39603	**GSE71016**
GSE38043	0	-0.15	-0.05	0	0.51
GSE6919		0	1.98	0	2.56
GSE55945			0	0	2.68
GSE39603				0	0
GSE71016					0

Table 13: Normalized Pair wise Z-Score values

NOTE: The datasets marked bold are considered for identification of up-regulated genes.
After calculating pair wise Z-scores

- Dataset "GSE39603" was discarded because it had no up-regulated genes therefore numerator in the formula becomes 0 giving the overall value Zero.
- Dataset "GSE38043" was also discarded because absolute Z-Score was below 1.96.
- So, the final datasets for further analysis were: GSE71016, GSE6919, and GSE38043.

5.4) IDENTIFICATION OF UP-REGULATED GENES

In order to identify the up-regulated genes we used "RankProdIt" [8]. We used a tab-delimited text file of Zhan Y. Dataset containing 31689 genes and then we specified two conditions for 95 samples i.e. condition 1 for normal samples and condition 2 for tumor samples.

Following output provided by Rank Product analysis:

Gene_name	Cond1 < Cond2 rank	Cond1 > Cond2 rank	C1 < C2 p. value	C1 > C2 p.value	Average Cond1/Cond2
100_g_at	8732	11684	0.926263736	0.998449594	0.916106789
1000_at	11411	8807	0.998938525	0.923639911	1.033137016
1001_at	5454	2849	0.438647874	0.094788979	1.043387062
1002_f_at	4074	7603	0.203908266	0.829652811	0.965802016
1003_s_at	3276	10516	0.097544195	0.986856984	0.80653978
1004_at	5070	4253	0.367835643	0.299576366	1.022466777
1005_at	4325	2246	0.241457238	0.041379201	0.987506693
1006_at	470	3925	2.39E-05	0.245760471	0.761366403
1007_s_at	7268	11674	0.770052556	0.998409779	0.874629192
1008_f_at	7529	9782	0.80792244	0.96942029	0.943584422
1009_at	6089	8863	0.570846472	0.926812391	0.872161885
101_at	10970	3164	0.99690954	0.134597866	1.132573404
1010_at	1642	8687	0.007179487	0.916340978	0.661793221
1011_s_at	6482	6028	0.641872113	0.612277433	0.962220992
1012_at	7923	398	0.853173276	0	1.181522578
1013_at	84	5821	0	0.580242077	0.780393439
1014_at	9503	11335	0.965710304	0.996571906	0.932995335
1015_s_at	3237	10094	0.093492594	0.978193184	0.834689236
1016_s_at	5658	735	0.48118968	0.000179965	1.11227011
1017_at	1920	9171	0.014534161	0.94395684	0.760562238
1018_at	5215	7601	0.39326724	0.829605829	0.906649804
1019_g_at	6534	9137	0.652167543	0.942397675	0.89193689
102_at	2397	4970	0.032360248	0.434110527	0.867257829
1020_s_at	8654	10460	0.920903806	0.985985029	0.953239841
1021_at	2774	6656	0.055496894	0.709902054	0.807149987

Table 14: RankProduct listed gene results.

The output contains the following results:

- **Gene name**: Gene name contains the gene identifiers as selected by us.

- **Cond1 Rank**: Condition 1 contains normal sample gene's data. Cond1 Rank specifies Rank of the genes in normal sample as calculated by RankProd package.

- **Cond2 Rank**: Condition 2 contains tumor sample gene's data. Cond2 Rank specifies Rank of the genes in tumor sample as calculated by RankProd package.

- **Cond1 p value**: contains the p value of a gene (based on Rank Products of RankProd) when sorting data by condition 1 < condition 2.Genes with the lowest p. value are the most significantly down-regulated genes in condition 1 compared to condition 2.

- **Cond2 p value**: contains the p value of a gene (based on Rank Products of RankProd) when sorting data by condition 2 < condition 1.Genes with the lowest p. value are the most significantly down-regulated genes in condition 2 compared to condition 1.

- **Cond1 pfp value**: contains the probability of false prediction (pfp) of a gene when sorting data by condition 1 < condition 2 (based on Rank Products of RankProd). Genes with the lowest pfp value are the significantly most down-regulated genes in condition 1 compared to condition 2.

- **Cond2 pfp value**: contains the probability of false prediction (pfp) of a gene when sorting data by condition 2 < condition 1 (based on Rank Products of RankProd). Genes with the lowest pfp value are the significantly most down-regulated genes in condition 2 compared to condition 1.

- **Average Cond1/Cond2**: contains the average condition 1/condition 2 fold change on a linear scale. In the data sets containing missing data the average is calculated by removing the missing data.

After getting the above results, the results were sorted on the basis of rank given to genes in tumor samples (decreasing order).

The sorted results are as follows:

Gene_name	Cond1 < Cond2 rank	Cond1 > Cond2 rank	C1 < C2 p. value	C1 > C2 p.value	Average Cond1/Cond2
39278_at	105	1	0	0	1.764344491
40157_s_at	391	2	1.35E-05	0	5.267382976
39063_at	12187	3	0.999988852	0	3.848264705
38615_at	559	4	5.10E-05	0	1.556336567
37582_at	5200	5	0.391912725	0	1.69569389
37125_f_at	10810	6	0.995609173	0	1.917823084
773_at	12522	7	1	0	2.076519896
31737_at	10922	8	0.99653448	0	4.793232909
1042_at	10460	9	0.991755853	0	2.266549661
39095_at	2320	10	0.029093805	0	4.841065938
36113_s_at	5713	11	0.493044275	0	5.337425824
39052_at	9136	12	0.950489728	0	2.094537215
37630_at	12481	13	1	0	1.949294983
37624_at	3666	14	0.14059484	0	4.511508069
32242_at	12548	15	1	0	3.434496985
33505_at	10942	16	0.996724001	0	2.233790399
40795_at	482	17	2.71E-05	0	4.479373087
39544_at	12542	18	1	0	2.009337602
38175_at	11056	19	0.997402453	0	2.941785327
774_g_at	12545	20	1	0	2.053378518
33404_at	12242	21	0.999992833	0	2.078575649
38407_r_at	8861	22	0.935247651	0	1.657654858
32312_at	4310	23	0.239268196	0	4.564922836
37572_at	10403	24	0.991004141	0	1.62141922
40776_at	12497	25	1	0	2.318073102

Table 15: Rank Product top 25 genes.

According to the top results, the top ranked genes were considered for further analysis.
Top 25 genes most significant genes which regulate the gene expression in prostate cancer are listed below:

GENE_ID	GENE_DESCRIPTION
HEBP1	Homo sapiens heme binding protein 1 (HEBP1)
TUSC1	Homo sapiens tumor suppressor candidate 1 (TUSC1)
PAFAH1B2	Homo sapiens platelet-activating factor acetylhydrolase 1b, catalytic subunit 2
STK32C	serine/threonine kinase 32C

CCDC88C	Homo sapiens coiled-coil domain containing 88C
EIF1B	Homo sapiens eukaryotic translation initiation factor 1B (EIF1B)
ARID4A	Homo sapiens AT rich interactive domain 4A (RBP1-like) (ARID4A), transcript variant 1
ZNF501	Homo sapiens zinc finger protein 501 (ZNF501)
HIST2H3D	Homo sapiens histone cluster 2, H3d (HIST2H3D)
PARP1	Homo sapiens poly (ADP-ribose) polymerase 1 (PARP1)
DUSP1	Homo sapiens dual specificity phosphatase 1 (DUSP1)
CKLF	Homo sapiens chemokine-like factor (CKLF), transcript variant 1
CLK1	Homo sapiens CDC-like kinase 1 (CLK1), transcript variant 2
CYB5R4	Homo sapiens cytochrome b5 reductase 4 (CYB5R4)
MAPK8	mitogen-activated protein kinase 8
TMEM14B	Homo sapiens transmembrane protein 14B (TMEM14B), transcript variant 1
XLOC_005179	AY563251 MHC class II antigen beta chain {Aotus nigriceps}
DKFZp547G183	Interleukin-likeUncharacterized protein
UBAP2L	Homo sapiens ubiquitin associated protein 2-like (UBAP2L)
TMEM191B	Homo sapiens transmembrane protein 191B (TMEM191B)
RNF13	Homo sapiens ring finger protein 13 (RNF13), transcript variant 1
TRIM78P	Homo sapiens tripartite motif containing 78, pseudogene (TRIM78P)
PRKD1	Homo sapiens protein kinase D1 (PRKD1)
MAP3K9	Homo sapiens mitogen-activated protein kinase kinase kinase 9 (MAP3K9)
SORD	Homo sapiens sorbitol dehydrogenase (SORD), transcript variant 1

Table 16: Gene names according to the HUGO Gene Nomenclature Committee.

5.5) FUNCTIONAL ANNOTATION OF GENES

Using RankProdIt we got the top 25 top regulated genes for the prostate cancer. Now we need to do the functional annotation of these genes and for that we used the DAVID tool.
Input to the DAVID [13] was the significant genes obtained through the RankProdIt.

5.5.1) REACTOME PATHWAY:

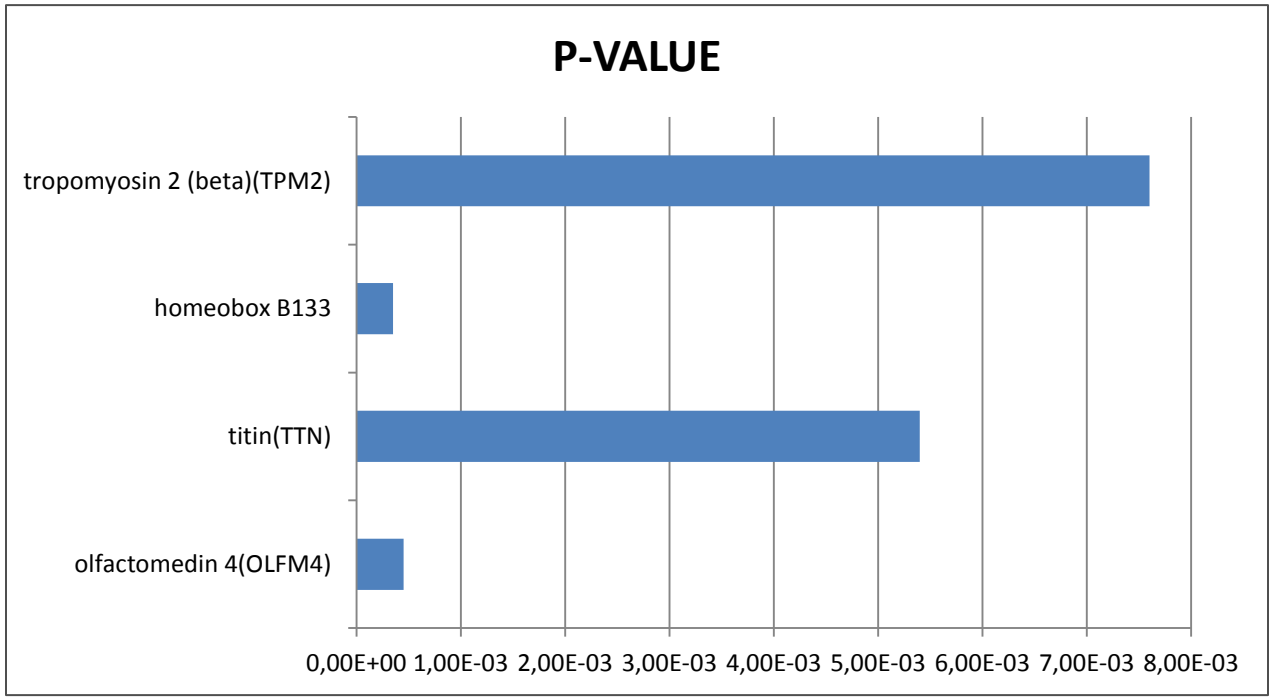

Figure 4: Reactome Pathway for the top regulated genes in prostate cancer.

The above figure lists the genes which have significant P-Value used for the Gene-enrichment analysis. Although we provided list of 25 genes as an input to DAVID but out of 25 only 4 were significantly expressed in Reactome pathway having good P-Value.

P-Value represents the EASE score. The smaller is the P-Value, the more enriched the gene is. Therefore, in the above chart homeobox B133 is the most enriched gene out of all having the least modified Fischer Exact P-Value.

5.5.2) KEGG PATHWAY:

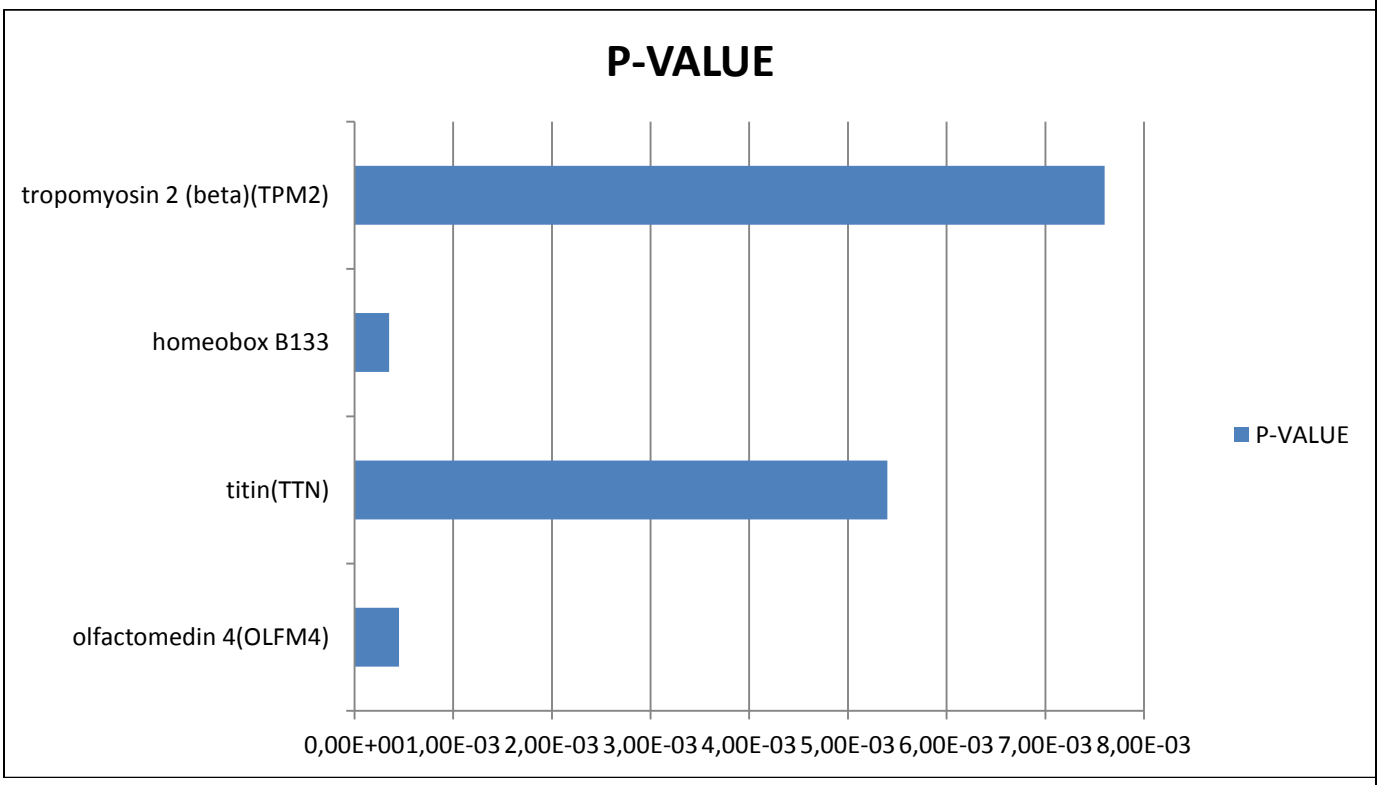

Figure 5: KEGG pathway for the top regulated genes in prostate cancer

The above figure lists the genes which have significant P-Value used for the Gene-enrichment analysis. Out of 25 genes only 4 were significantly expressed in KEGG pathway having good P-Value.
In the above chart homeobox B133 is the most enriched gene out of all having the least modified Fischer Exact P-Value.

Therefore, from the above two pathways we can conclude that **homeobox B133** is the gene which is highly regulated in prostate cancer i.e. the most enriched gene out of 25 top regulated genes.

5.6) IDENTIFICATION OF SIGNIFICANT PATHWAYS

Using online tool of GeneMANIA [14] a network was created with 104 genes as an input.

The six types of network which were created are as follows:

- Co-expression
- Co-localization
- Genetic Interaction

- Pathway
- Physical Interaction
- Shared Protein Domains

Using the CytoScape[15] tool we got the network properties of the network obtained through GeneMANIA.

5.6.1) CO-EXPRESSION

NETWORK IMAGE:

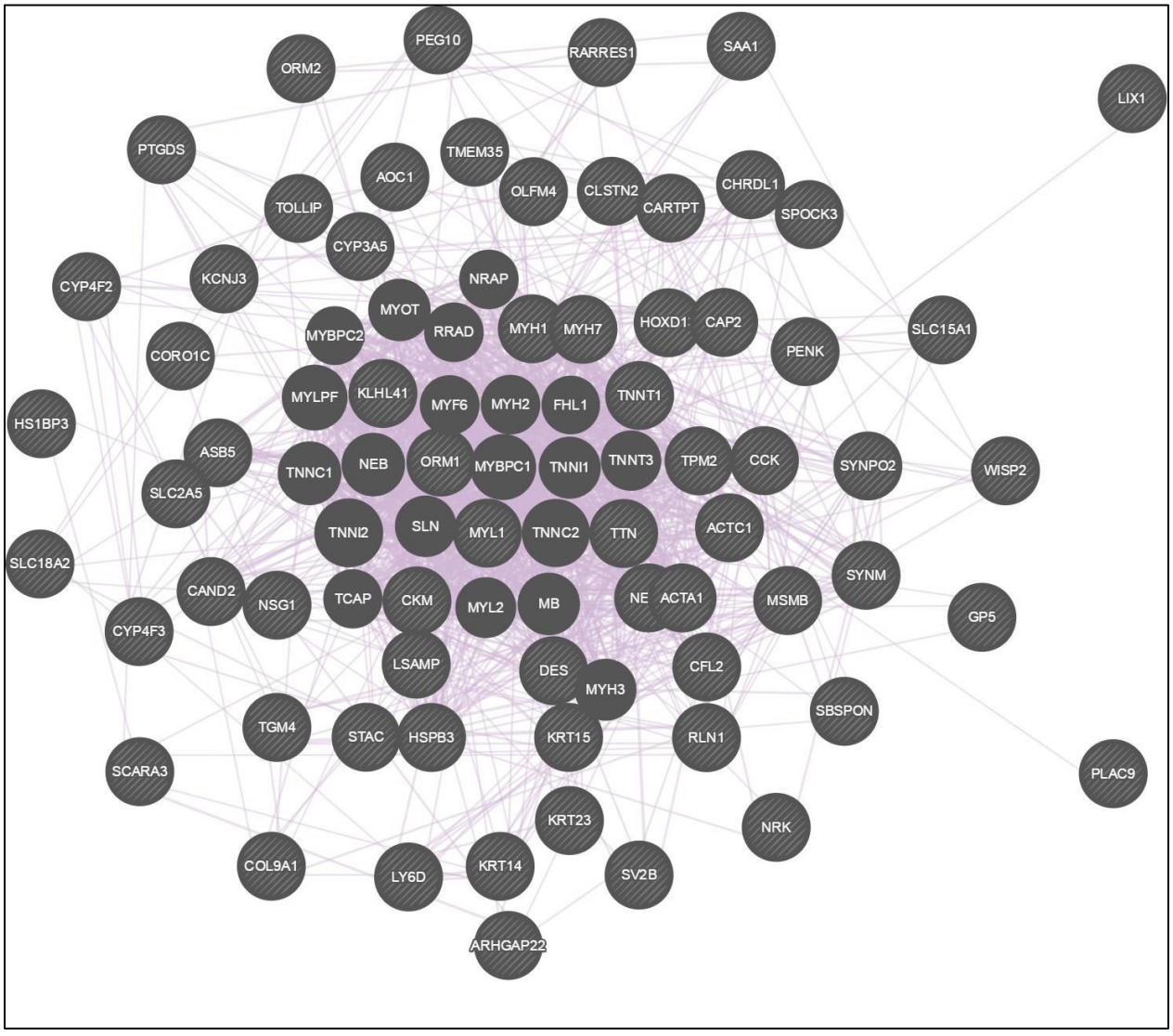

Figure 6: Co-expression network

DISCUSSION: The above network gives us the Co-expression network of the genes we provided as an input and other related genes obtained from different databases.In the Co-expression network the

genes represents the nodes and the edges represents the interactions among the genes. The nodes having maximum number of edges are the genes which are significantly enriched and highly regulated in prostate cancer. For e.g. ORM1 and TNN1 are the genes which are present in the hub and are known as the hub genes. In the above network, all the genes are connected to each other i.e. there is no genes which have different expression level from other genes.

NETWORK PROPERTIES:

PROPERTIES	PHYSICAL INTERACTIONS
CLUSTERING COFFICIENT	0.469
CONNECTED ELEMENTS	1
NETWORK DIAMETER	4
NETWORK RADIUS	3
NETWORK CLUSTER	0.263
SHORTEST PATHS	6972
CHARACTERSTIC PATH LENGTH	1.991
AVG. NUMBER OF NEIGHBORS	18.714
NUMBER OF NODES	84
NETWORK DENSITY	0.225
NETWORK DIVERSITY	0.710
ISOLATED NODES	0
SELF-LOOPS	0
MULTI-EDGE NODE PAIRS	430
ANALYSIS TIME(SEC.)	0.071

Table 17: Network properties of co-expression network

5.6.2) PHYSICAL INTERACTION

NETWORK IMAGE:

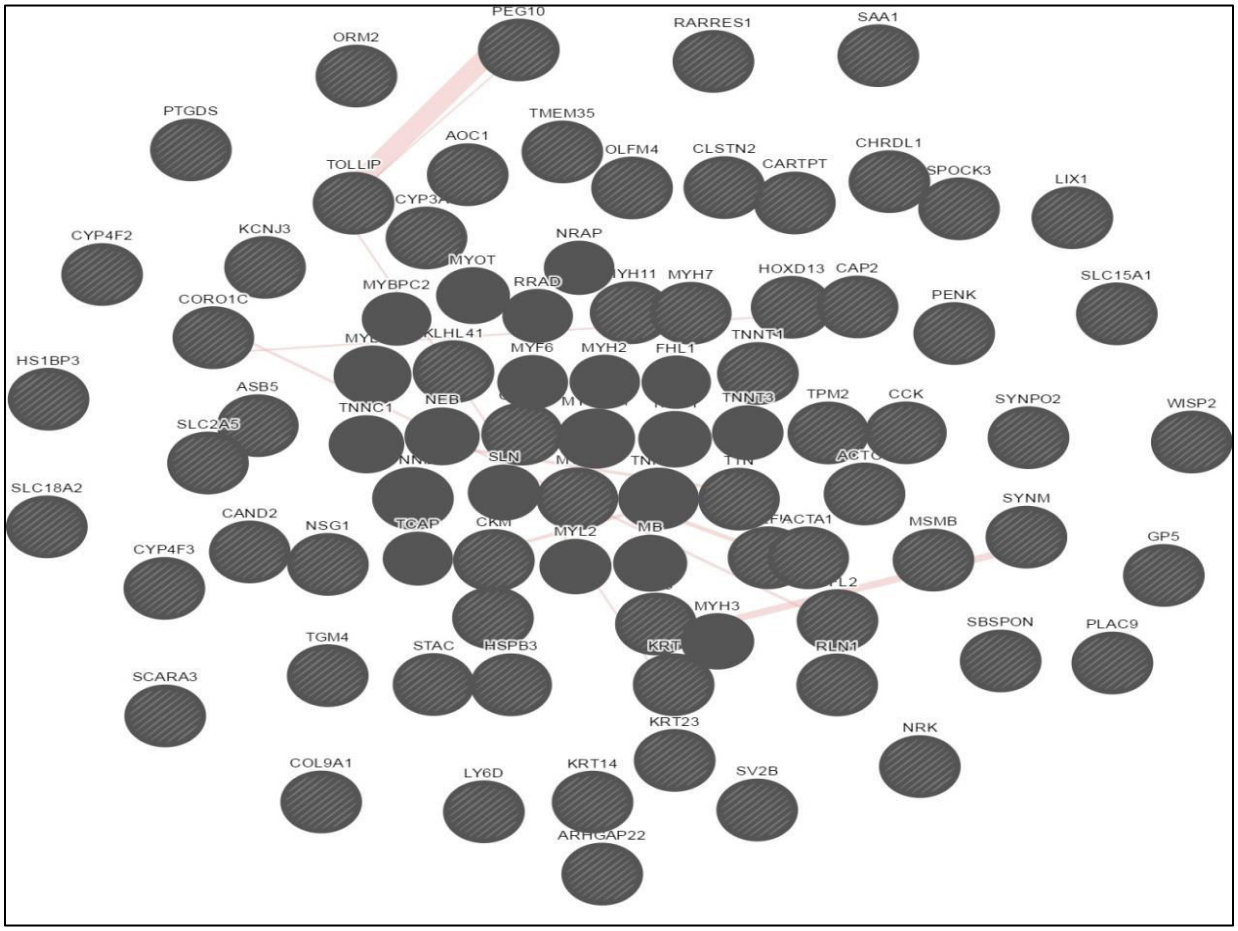

Figure 7: Physical interaction network

DISCUSSION: The above network gives us the Physical interaction network of the genes we provided as an input and other related genes obtained from different databases.In physical interaction network the genes represents the nodes and the edges represents the interactions among the genes. The nodes having maximum number of edges are the genes which are significantly enriched and highly regulated in prostate cancer. For e.g. TOLL1P and PEG10 are the genes which are present in the hub and are known as the hub genes. The nodes which do not have any edges connected to them represents the genes which are not in the Physical interaction network for e.g. SCARA3 and TGM4 etc.

NETWORK PROPERTIES:

PROPERTIES	PHYSICAL INTERACTIONS
CLUSTERING COFFICIENT	0.3
CONNECTED ELEMENTS	5

NETWORK DIAMETER	2
NETWORK RADIUS	1
NETWORK CLUSTER	0.076
SHORTEST PATHS	22
CHARACTERSTIC PATH LENGTH	1.273
AVG. NUMBER OF NEIGHBORS	1.231
NUMBER OF NODES	13
NETWORK DENSITY	0.103
NETWORK DIVERSITY	0.342
ISOLATED NODES	0
SELF-LOOPS	0
MULTI-EDGE NODE PAIRS	1
ANALYSIS TIME(SEC.)	0.032

Table 18: Network properties of Physical Interaction Network

5.6.3) CO-LOCALIZATION

NETWORK IMAGE:

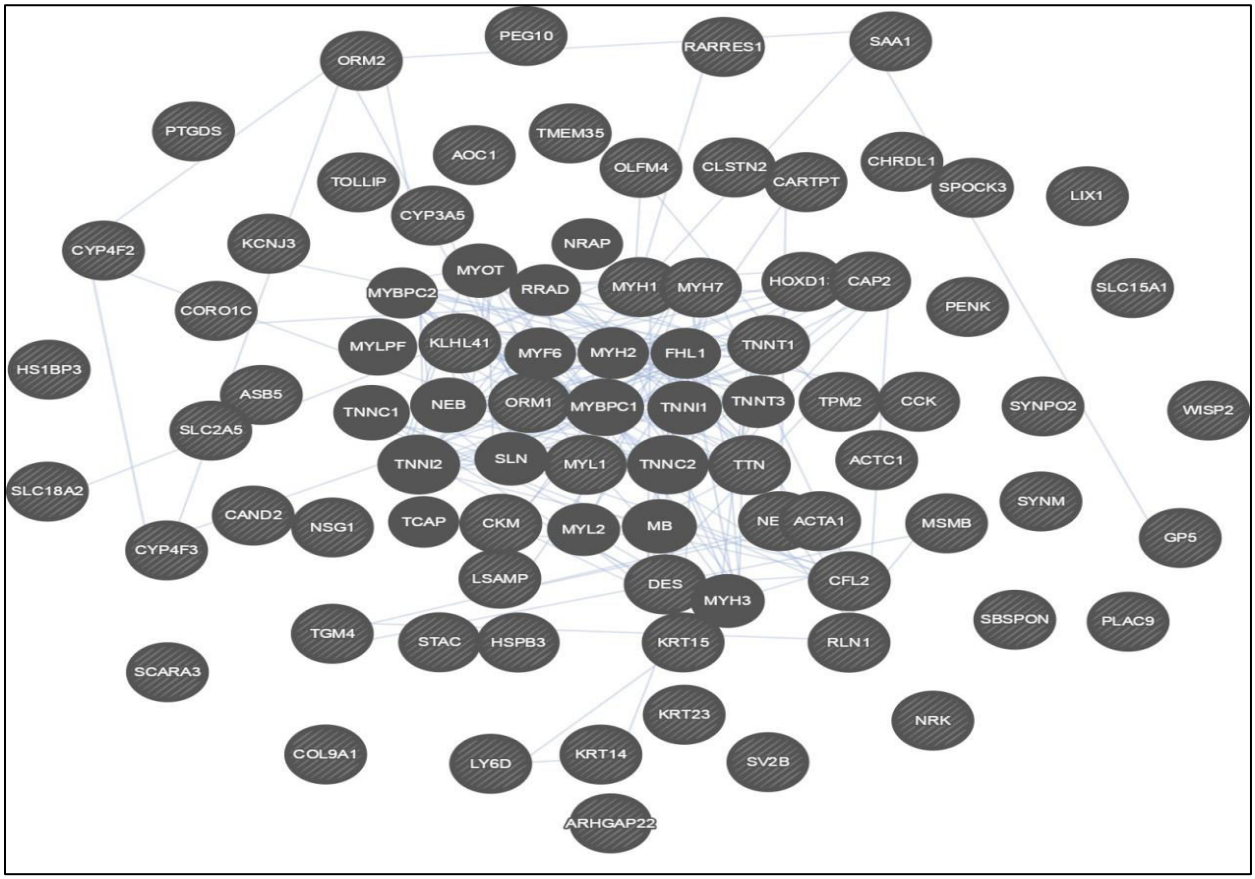

Figure 8: Co-localization network

DISCUSSION: The above network gives us the Co-localization network of the genes we provided as an input and other related genes obtained from different databases.

In a Co-localization network the genes represents the nodes and the edges represents the interactions among the genes. The nodes having maximum number of edges are the genes which are significantly enriched and highly regulated in prostate cancer. For e.g. TTN and TNNI1 are the genes which are present in the hub and are known as the hub genes. The nodes which do not have any edges connected to them represents the genes which are not in the Co-localization network for e.g. WISP-2 and SCARA3 etc.

NETWORK PROPERTIES:

PROPERTIES	PHYSICAL INTERACTIONS
CLUSTERING COFFICIENT	0.702
CONNECTED ELEMENTS	5
NETWORK DIAMETER	5
NETWORK RADIUS	1
NETWORK CLUSTER	0.273
SHORTEST PATHS	950
CHARACTERSTIC PATH LENGTH	2.044
AVG. NUMBER OF NEIGHBORS	7.708
NUMBER OF NODES	48
NETWORK DENSITY	0.164
NETWORK DIVERSITY	0.790
ISOLATED NODES	0
SELF-LOOPS	0
MULTI-EDGE NODE PAIRS	3
ANALYSIS TIME(SEC.)	0.045

Table 19:Co-localisation network properties

5.6.4) PATHWAY

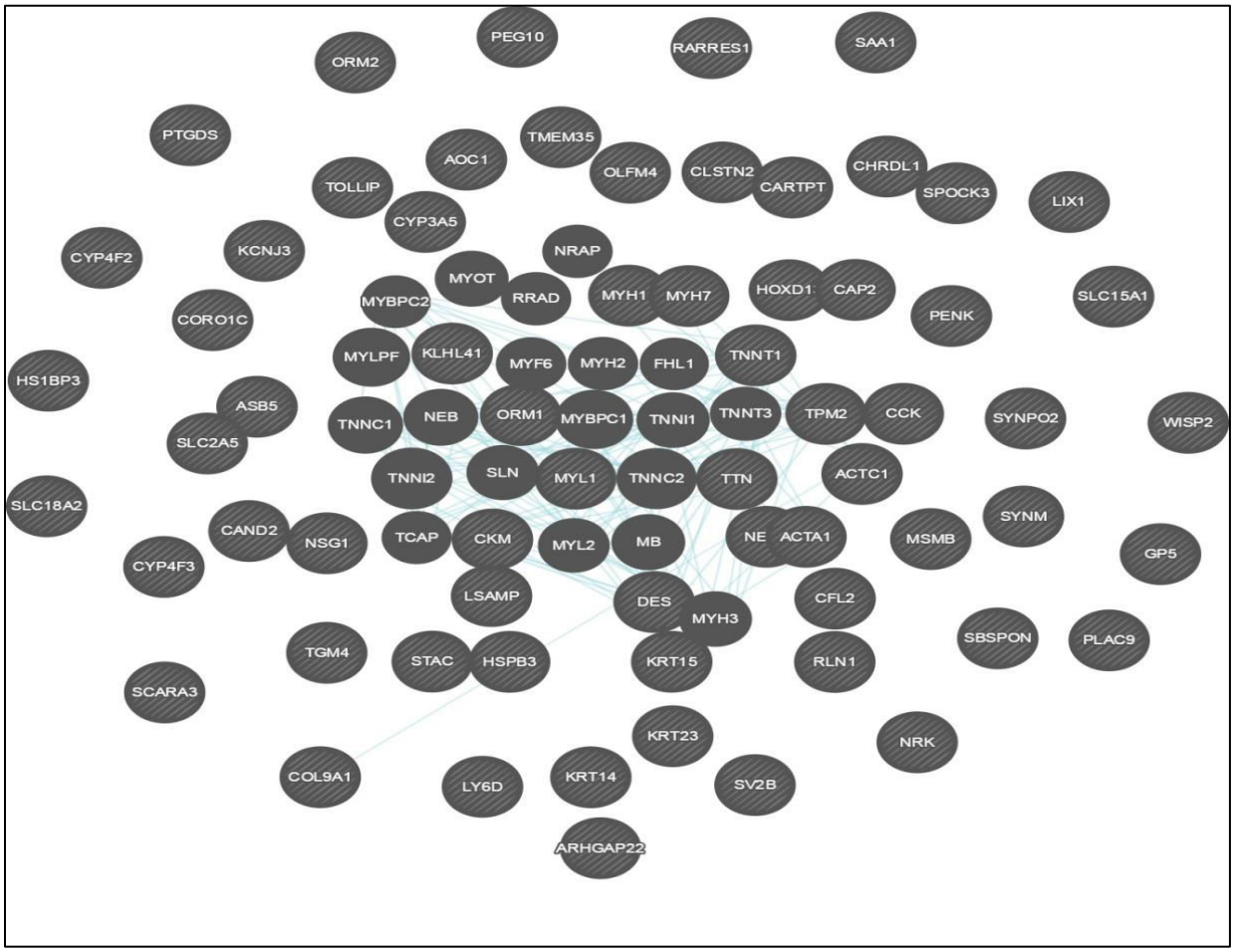

Figure 9: Pathway

DISCUSSION: The above network gives us the Pathway network of the genes we provided as an input and other related genes obtained from different databases.In a Pathway network the genes represents the nodes and the edges represents the interactions among the genes. The nodes having maximum number of edges are the genes which are significantly enriched and highly regulated in prostate cancer. For e.g. DES and MYL2 are the genes which are present in the hub and are known as the hub genes. The nodes which do not have any edges connected to them represents the genes which are not in the Pathway network for e.g. NRK and SCARA3 etc.

NETWORK PROPERTIES:

PROPERTIES	PHYSICAL INTERACTIONS
CLUSTERING COFFICIENT	0.691
CONNECTED ELEMENTS	1
NETWORK DIAMETER	3

NETWORK RADIUS	2
NETWORK CLUSTER	0.295
SHORTEST PATHS	462
CHARACTERSTIC PATH LENGTH	1.723
AVG. NUMBER OF NEIGHBORS	11.364
NUMBER OF NODES	22
NETWORK DENSITY	0.541
NETWORK DIVERSITY	0.477
ISOLATED NODES	0
SELF-LOOPS	0
MULTI-EDGE NODE PAIRS	6
ANALYSIS TIME(SEC.)	0.031

Table 20: Pathway network properties

5.6.5) SHARED PROTEIN DOMAINS

NETWORK IMAGE:

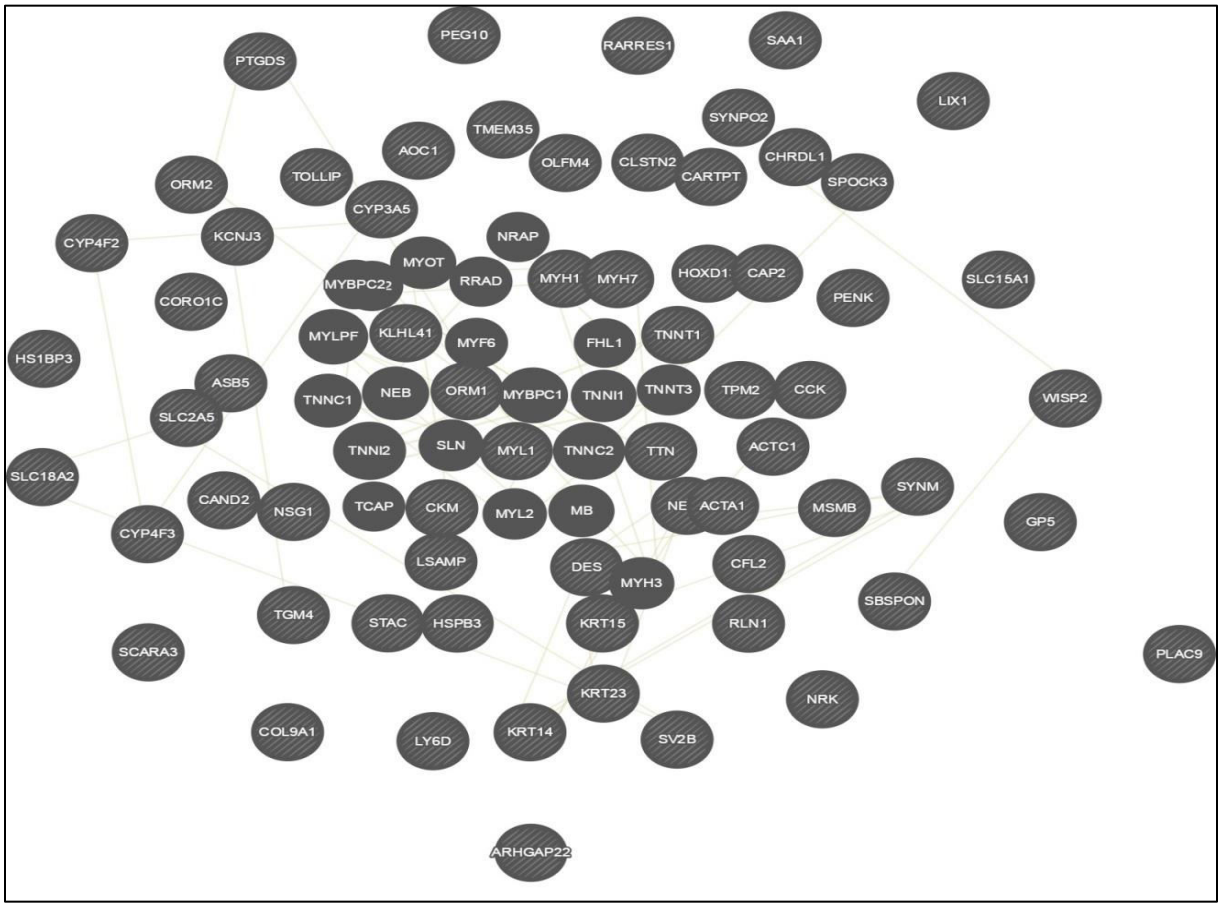

Figure 10: Shared protein domain network

DISCUSSION: The above network gives us the Shared protein domain network of the genes we provided as an input and other related genes obtained from different databases.In a shared protein domain network the genes represents the nodes and the edges represents the interactions among the genes. The nodes having maximum number of edges are the genes which are significantly enriched and highly regulated in prostate cancer. For e.g. DES and ACTA1 are the genes which are present in the hub and are known as the hub genes. The nodes which do not have any edges connected to them represents the genes which are not in the Co-localization network for e.g. ARHGAP22 and PLAC9 etc.

NETWORK PROPERTIES:

PROPERTIES	PHYSICAL INTERACTIONS
CLUSTERING COFFICIENT	0.673
CONNECTED ELEMENTS	13
NETWORK DIAMETER	3
NETWORK RADIUS	1
NETWORK CLUSTER	0.061
SHORTEST PATHS	126
CHARACTERSTIC PATH LENGTH	1.143
AVG. NUMBER OF NEIGHBORS	2.5
NUMBER OF NODES	44
NETWORK DENSITY	0.058
NETWORK DIVERSITY	0.501
ISOLATED NODES	0
SELF-LOOPS	0
MULTI-EDGE NODE PAIRS	0
ANALYSIS TIME(SEC.)	0.116

Table 21: Shared protein domain network properties

5.6.6) GENETIC INTERACTIONS

NETWORK IMAGE:

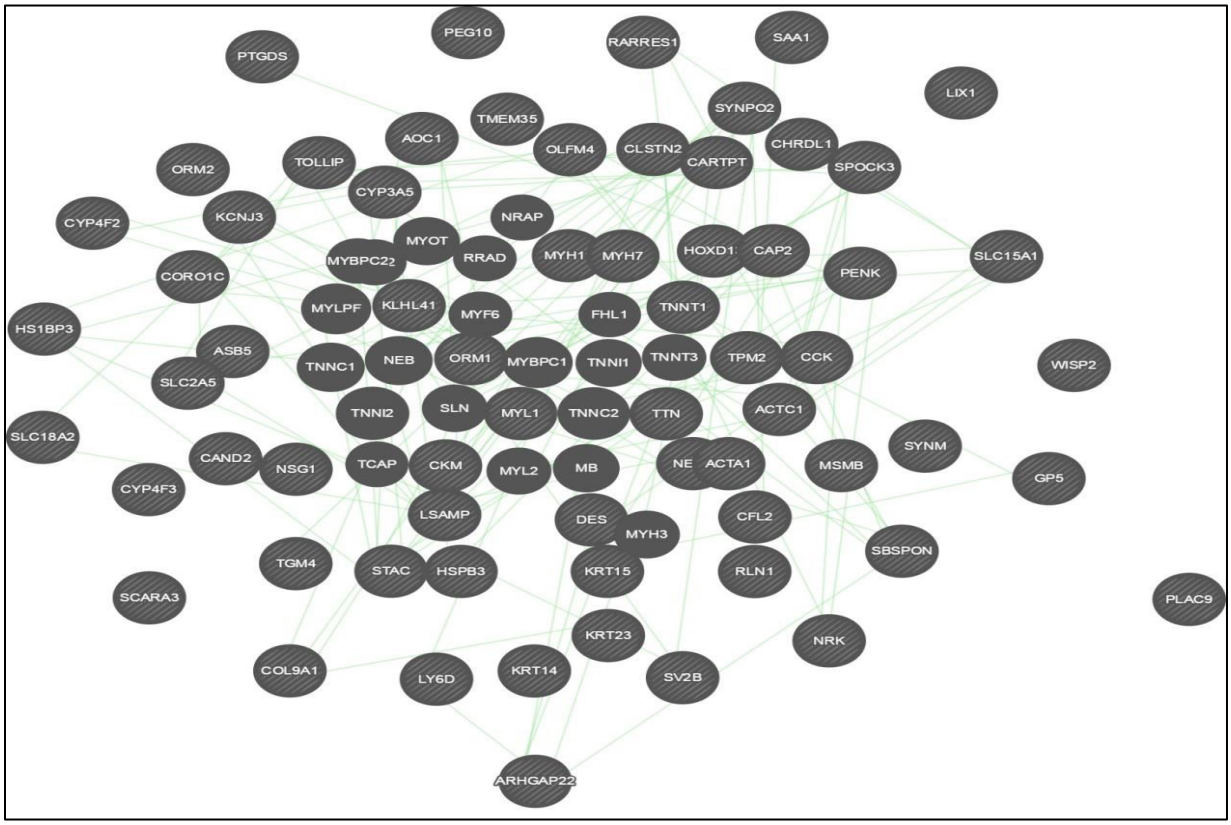

Figure 11: Genetic interaction network

DISCUSSION: The above network gives us the Genetic interaction network of the genes we provided as an input and other related genes obtained from different databases.

In a Genetic interaction network the genes represents the nodes and the edges represents the interactions among the genes. The nodes having maximum number of edges are the genes which are significantly enriched and highly regulated in prostate cancer. For e.g. ACTC1 and TNNIC2 are the genes which are present in the hub and are known as the hub genes. The nodes which do not have any edges connected to them represents the genes which are not in the Genetic interaction network for e.g. WISP-2 and PLAC9 etc.

NETWORK PROPERTIES:

PROPERTIES	PHYSICAL INTERACTIONS
CLUSTERING COFFICIENT	0.100
CONNECTED ELEMENTS	2
NETWORK DIAMETER	6
NETWORK RADIUS	1

NETWORK CLUSTER	0.151
SHORTEST PATHS	3424
CHARACTERSTIC PATH LENGTH	2.827
AVG. NUMBER OF NEIGHBORS	4.230
NUMBER OF NODES	61
NETWORK DENSITY	0.070
NETWORK DIVERSITY	0.692
ISOLATED NODES	0
SELF-LOOPS	0
MULTI-EDGE NODE PAIRS	0
ANALYSIS TIME(SEC.)	0.04

Table 22: Genetic interaction network properties

5.7) HUB GENES- LIST OF GENES IN A NETWORK HAVING THE MAXIMUM WEIGHT AND FORMING THE HUB

GENE 1	GENE 2	WEIGHT
Troponin T3 (TNNT3)	SPARC/osteonectin, cwcv and kazal like domains proteoglycan 3	0.50214535
Desmin(DES)	Synemin(SYNM)	0.50117
toll interacting protein(TOLLIP)	paternally expressed 10(PEG10)	0.47651312
actin, alpha 1, skeletal muscle(ACTA1)	cofilin 2(CFL2)	0.262447
actin, alpha 1, skeletal muscle(ACTA1)	myosin light chain 1(MYL1)	0.22871189
troponin I, skeletal, fast 2 (TNN12)	troponin T1, slow skeletal type(TNNT1)	0.2
troponin I1, slow skeletal type(TNNI1)	troponin T1, slow skeletal type(TNNT1)	0.2
troponin I1, slow skeletal type(TNNI1)	troponin I, skeletal, fast 2 (TNN12)	0.2
Troponin T3 (TNNT3)	troponin T1, slow skeletal type(TNNT1)	0.2
Troponin T3 (TNNT3)	troponin I, skeletal, fast 2 (TNN12)	0.2

Table 23: Hub genes.

CHAPTER- 6
CONCLUSION

The lack of specific treatment for different types of prostate cancer leads to an increase mortality rate. In many common types of prostate cancers, various bio-markers have been identified but they are not tested clinically due to their low specificity and/or sensitivity. Meta analysis of multiple microarray dataset can give us more accurate reliable and comprehensive results rather than single dataset because former has statistical power and generalisability. In our study we performed Meta analysis of 104 significant genes. Through this project we predicted the potential bio-markers in prostate cancer using meta-analysis and the functional analysis of the potential genes has been done in terms of metabolic and cell-signalling pathways.

It can be concluded that meta-analysis of microarray datasets yields more comprehensive and reliable results as compared to a single dataset because the former has generalisability and increased statistical power. On creating different types of networks of significantly unregulated Genes, various pathways that are possibly enriched in prostate cancer can be obtained.

CHAPTER – 7
REFERENCES

1) Campain, Anna, and Yee Yang. "Comparison Study Of Microarray Meta-Analysis Methods". *BMC Bioinformatics* 11.1 (2010): 408. Web.

2) "What Is Cancer?". *National Cancer Institute*. N.p., 2017. Web. 18 Apr. 2017

3) Rhodes, Daniel R. et al. "ONCOMINE: A Cancer Microarray Database And Integrated Data-Mining Platform". *Neoplasia* 6.1 (2004): 1-6. Web.

4) "Home - GEO - NCBI". *Ncbi.nlm.nih.gov*. N.p., 2017. Web. 18 Apr. 2017.

5) Chu, Vu T et al. "Mev+R: Using Mev As A Graphical User Interface For Bioconductor Applications In Microarray Analysis". *Genome Biology* 9.7 (2008): R118. Web.

6) "Mev". *Mev.tm4.org*. N.p., 2017. Web. 18 Apr. 2017.

7) Hong, F. et al. "Rankprod: A Bioconductor Package For Detecting Differentially Expressed Genes In Meta-Analysis". *Bioinformatics* 22.22 (2006): 2825-2827. Web.

8) Laing, Emma, and Colin P Smith. "Rankprodit: A Web-Interactive Rank Products Analysis Tool". *BMC Research Notes* 3.1 (2010): 221. Web.

9) Chu, Vu T et al. "Mev+R: Using Mev As A Graphical User Interface For Bioconductor Applications In Microarray Analysis". *Genome Biology* 9.7 (2008): R118. Web.

10) Warde-Farley, D. et al. "The Genemania Prediction Server: Biological Network Integration For Gene Prioritization And Predicting Gene Function". *Nucleic Acids Research* 38.Web Server (2010): W214-W220. Web.

11) Goonesekere, Nalin C. W. et al. "A Meta Analysis Of Pancreatic Microarray Datasets Yields New Targets As Cancer Genes And Biomarkers". *PLoS ONE* 9.4 (2014): e93046. Web.

12) Huang, Da et al. "The DAVID Gene Functional Classification Tool: A Novel Biological Module-Centric Algorithm To Functionally Analyze Large Gene Lists". *Genome Biology* 8.9 (2007): R183. Web.

13)"DAVID Functional Annotation Bioinformatics Microarray Analysis". *David-d.ncifcrf.gov*. N.p., 2017. Web. 18 Apr. 2017.

14)"Genemania". *Genemania.org*. N.p., 2017. Web. 18 Apr. 2017.

15)Shannon, P. "Cytoscape: A Software Environment For Integrated Models Of Biomolecular Interaction Networks". *Genome Research* 13.11 (2003): 2498-2504. Web.

APPENDIX

LIST OF TABLES:

LIST OF FIGURES: