META-ANALYSIS OF BRAIN AND CENTRAL NERVOUS SYSTEM CANCER MICROARRAY DATASETS

ENROLLMENT NUMBERS – 121501 & 121510

STUDENT NAMES – ICXA KHANDELWAL & ADITI SHARMA

SUPERVISOR NAME - Dr. JAYASHREE RAMANA

MAY - 2016

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

TABLE OF CONTENTS

SERIAL NUMBER	TOPIC	PAGE NUMBER
1	Certificate	II
2	Acknowledgement	III
3	Summary	IV
4	List of figures	V
5	List of tables	VII
6	Abbreviation	VIII
7	Chapter1 - Introduction	1
8	Chapter2 - Brain and CNS cancer	3
9	Chapter3 - Datasets and methods	4
10	Chapter4 - Results and Discussion	9
11	Chapter5 - Conclusion and Future Prospects	53
12	Chapter6 - References	54

CERTIFICATE

This is to certify that the work titled "**Meta-analysis of Brain and Central Nervous System Cancer microarray datasets**" submitted by "**Icxa Khandelwal and Aditi Sharma**" in partial fulfilment for the award of degree of **Bachelor of Technology in Bioinformatics** of Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor ……………………..

Name of Supervisor Dr. Jayashree Ramana

Designation Assistant Professor (Senior Grade)

Date ……………………..

ACKNOWLEDGEMENT

Dreams really come true. This thesis represents a part of our dream which would not be possible without the contribution of many others, who directly or indirectly contributed in the development of this work and who influenced our thinking, behaviour and acts during the course of project. 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 and indebtedness to our guide Dr. Jayashree Ramana for her exemplary guidance, valuable feedback and constant encouragement throughout the duration of the project. We are highly grateful to her for providing us the opportunity to do research work in our area of interest. Thank you very much ma'am for being such a wonderful supervisor.

We would hereby express our sincere thanks to all the faculties of BI department, 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.

Credit also goes to all our friends and family as their encouragement kept us in good stead. Their encouragement and continuous support has given us strength and confidence to complete the initial part of the project without any difficulty.

Signature of Student
Name
Date

Signature of Student
Name
Date

SUMMARY

The lack of specific treatment for different types and stages of Brain and Central Nervous System cancer lead to an increase mortality rate. In many common types of Brain and Central Nervous System cancers, various bio-markers have been identified but they are not tested clinically due to their low specificity and/or sensitivity. Through this project we have predicted the potential bio-markers in Brain and CNS cancers using meta-analysis and the functional analysis of these potential genes has been done in terms of metabolic and cell-signalling pathways.

Signature of Student	Signature of Student	Signature of Supervisor
Name	Name	Name
Date	Date	Date

LIST OF FIGURES

FIGURE NUMBER	FIGURE NAME	PAGE NUMBER
1	Pairwise Z score of selected sub-datasets	23
2	Enriched Reactome Pathway obtained through DAVID	25
3	Enriched OMIM Diseases obtained through DAVID	25
4	Enriched KEGG PATHWAY obtained through DAVID	26
5	Enriched GENETIC ASSOCIATION DB DISEASE CLASS obtained through DAVID	26
6	Categories of enriched cellular processes KEGG pathways in co-expression network	27
7	Enriched cellular community KEGG pathways in co-expression network	28
8	Enriched transport and catabolism KEGG pathways in co-expression network	28
9	Enriched signal transduction KEGG pathways in co-expression network	28
10	Enriched signalling molecules and interactions KEGG pathways in co-expression network	29
11	Enriched cancer KEGG pathways in co-expression network	29
12	Enriched neurodegenerative diseases KEGG pathways in co-expression network	29
13	Categories of enriched metabolism KEGG pathways in co-expression network	30
14	Enriched amino acid metabolism KEGG pathways in co-expression network	30
15	Enriched metabolism of cofactors and vitamins KEGG pathways in co-expression network	30
16	Enriched metabolism of other amino acids KEGG pathways in co-expression network	31
17	Enriched endocrine system KEGG pathways in co-expression network	31
18	Enriched immune system KEGG pathways in co-expression network	31
19	Categories of enriched cellular processes KEGG pathways in co-localization network	32
20	Enriched cell growth and death KEGG pathways in co-localization network	32
21	Enriched cellular community KEGG pathways in co-localization network	33
22	Enriched transport and catabolism KEGG pathways in co-localization network	33
23	Enriched signal transduction KEGG pathways in co-localization network	33
24	Enriched signalling molecules interactions and KEGG pathways in co-localization network	34
25	Enriched cancer KEGG pathways in co-localization network	34
26	Categories of enriched metabolism KEGG pathways in co-localization network	34
27	Enriched amino acid metabolism KEGG pathways in co-localization network	35
28	Enriched metabolism of other amino acids KEGG pathways in co-localization network	35
29	Enriched endocrine system KEGG pathways in co-localization network	35
30	Enriched immune system KEGG pathways in co-localization network	36
31	Categories of enriched cellular processes KEGG pathways in genetic interaction network	37
32	Enriched cellular community KEGG pathways in genetic interaction network	37
33	Enriched transport and catabolism KEGG pathways in genetic interaction network	37
34	Enriched signal transduction KEGG pathways in genetic interaction network	38

FIGURE NUMBER	FIGURE NAME	PAGE NUMBER
35	Enriched signalling molecules and interactions KEGG pathways in genetic interaction network	38
36	Enriched cancers KEGG pathways in genetic interaction network	39
37	Enriched neurodegenerative diseases KEGG pathways in genetic interaction network	39
38	Categories of enriched metabolism KEGG pathways in genetic interaction network	39
39	Enriched amino acid metabolism KEGG pathways in genetic interaction network	40
40	Enriched metabolism of cofactors and vitamins KEGG pathways in genetic interaction network	40
41	Enriched metabolism of other amino acids KEGG pathways in genetic interaction network	40
42	Enriched endocrine system KEGG pathways in genetic interaction network	41
43	Enriched immune system KEGG pathways in genetic interaction network	41
44	Categories of enriched cellular processes KEGG pathways in pathway network	42
45	Enriched cellular community KEGG pathways in pathway network	43
46	Enriched transport and catabolism KEGG pathways in pathway network	43
47	Enriched signal transduction KEGG pathways in pathway network	43
48	Enriched signalling molecules and interactions KEGG pathways in pathway network	44
49	Enriched cancers KEGG pathways in pathway network	44
50	Enriched endocrine system KEGG pathways in pathway network	44
51	Enriched immune system KEGG pathways in pathway network	45
52	Categories of enriched cellular processes KEGG pathways in physical interaction network	46
53	Enriched cellular community KEGG pathways in physical interaction network	46
54	Enriched transport and catabolism KEGG pathways in physical interaction network	46
55	Enriched signal transduction KEGG pathways in physical interaction network	47
56	Enriched signalling molecules and interactions KEGG pathways in physical interaction network	47
57	Enriched cancers KEGG pathways in physical interaction network	48
58	Enriched endocrine system KEGG pathways in physical interaction network	48
59	Enriched immune system KEGG pathways in physical interaction network	48
60	Categories of enriched cellular processes KEGG pathways in shared protein domains network	49
61	Enriched cellular community KEGG pathways in shared protein domains network	50
62	Enriched transport and catabolism KEGG pathways in shared protein domains network	50
63	Enriched signal transduction KEGG pathways in shared protein domains network	50
64	Enriched signalling molecules and interactions KEGG pathways in shared protein domains network	51
65	Enriched cancers KEGG pathways in shared protein domains network	51
66	Categories of enriched metabolism KEGG pathways in shared protein domain network	51
67	Enriched immune system KEGG pathways in shared protein domains network	52

LIST OF TABLES

TABLE NUMBER	TABLE NAME	PAGE NUMBER
1	Brain & CNS cancer microarray datasets included in the study	8
2	List of top significantly up-regulated genes in BB2SP1	10
3	List of top significantly up-regulated genes in BB2SP2	10
4	List of top significantly up-regulated genes in BB2SP3	11
5	List of top significantly up-regulated genes in BB2SP4	12
6	List of top significantly up-regulated genes in BB2SP5	12
7	List of top significantly up-regulated genes in BB2SP6	13
8	List of top significantly up-regulated genes in LBSP1	14
9	List of top significantly up-regulated genes in LBSP2	14
10	List of top significantly up-regulated genes in LBSP3	15
11	List of top significantly up-regulated genes in LBSP4	15
12	List of top significantly up-regulated genes in LBSP5	16
13	List of top significantly up-regulated genes in LBSP6	16
14	List of top significantly up-regulated genes in LBSP7	17
15	List of top significantly up-regulated genes in LBSP8	17
16	List of top significantly up-regulated genes in LBSP9	18
17	List of top significantly up-regulated genes in LBSP10	18
18	List of top significantly up-regulated genes in LBSP11	19
19	List of top significantly up-regulated genes in LBSP12	19
20	List of top significantly up-regulated genes in LBSP13	20
21	List of top significantly up-regulated genes in LISP1	20
22	List of top significantly up-regulated genes in LISP2	21
23	List of top significantly up-regulated genes in LISP3	21
24	List of top significantly up-regulated genes in MBSP1	21
25	List of top significantly up-regulated genes in MBSP2	22
26	List of top 26 significantly up-regulated genes	24
27	Network Properties of Co-expression network obtained using Network Analyser tool of Cytoscape	27
28	Network Properties of Co-localization network obtained using Network Analyser tool of Cytoscape	32
29	Network Properties of Genetic Interaction network obtained using Network Analyser tool of Cytoscape	36
30	Network Properties of Pathway1 Interaction network obtained using Network Analyser tool of Cytoscape	42
31	Network Properties of Physical Interaction network obtained using Network Analyser tool of Cytoscape	45
32	Network Properties of Shared Protein Domains network obtained using Network Analyser tool of Cytoscape	49

ABBREVIATION

ABBREVIATED WORD	WORD
CNS	Central Nervous System
RNA	Ribonucleic Acid
mRNA	messenger Ribonucleic Acid
SAM	Significant Analysis of Microarrays
MeV	Multi-experiment viewer
pfp	percentage of false positive predictions
DAVID	Database for Annotation, Visualization and Integrated Discovery
KEGG	Kyoto Encyclopaedia for Genes and Genomes
FDR	False Discovery Rate
GO	Gene Ontology

CHAPTER 1 - INTRODUCTION

1.1. OBJECTIVE

The objective of the project is prediction of potential biomarkers in Brain and CNS cancer and functional analysis of these genes in terms of metabolic and cell signalling pathways.

1.2. META-ANALYSIS

Meta-analysis as the name suggests, refers to analysis of analysis. It can be defined as a statistical technique for combining the findings obtained from independent studies. Meta-analysis of microarray datasets aids in finding significantly up-regulated genes in cancer which thereby increases the generalizability and statistical power [1].
Meta-analysis has been performed for various types of biological data. Since the data obtained from a single experiment is usually biased, therefore this biasness needs to be removed so as to obtain generalized information. For instance, the genes identified from the meta-analysis of multiple datasets belonging to a particular cancer might be enriched in some particular pathway and if the researchers working in these areas focus on these pathways then more concrete results can be generated.

1.3. MICROARRAY EXPERIMENT

1.3.1. OBJECTIVES

The three major objectives of the microarray experiment are to

1. Provide information about the relative gene expression levels.
2. Provide information about overall amount of mRNA
3. Permit the detection of small differences in transcript abundance [2]

1.3.2. METHODOLOGY

The methodology of performing a microarray experiment is as described below:

1. Material processing
    a. Array fabrication
    b. Preparation of biological samples to be tested

   c. Extraction and labelling of the RNA from the samples
2. Hybridization
3. Scanning
4. Information Processing
   a. Image quantitation
   b. Data normalization and integration
5. Data management
   a. Gene expression data matrix
   b. Gene annotation
   c. Sample annotation
6. Data analysis and modelling
7. Gene expression data analysis and mining
8. Generation of new hypothesis from this analysis [2]

1.3.4. APPLICATIONS

Microarray technologies have a wide scale application in various arenas of life sciences. They are:

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 especially cancer. 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 tumour cells as compared to the normal cells.
3. Drug Discovery – Pharmacogenomics 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 2 - BRAIN AND CNS CANCER

2.1. CANCER

Cancer is an abnormal growth of cell likely to affect other parts of the body. It takes place due to factors such as mutations or abnormal transformations in the genes that helps in regulation of growing cells thereby keeping them fit. A tumour can be of two categories: benign and malignant. Benign tumours are not cancerous. These cells are almost normal in appearance, do not attack any nearby tissues nor extend to the entire body. While malignant tumours are considered to be cancerous. If not treated in time can spread in the other parts of the body as well [3] [4].

2.2. BRAIN AND CNS CANCER

Brain and spinal cord tumours may be either benign or malignant. Benign brain and spinal cord tumours are found in the areas near brain and they rarely spread into other parts of the brain. On the other hand, malignant brain and spinal cord tumours are likely to grow quickly and spread into other parts of the brain. This tumour hinders the normal brain functions such as muscle control, sensation, memory and other normal body functions. Cancer cells which have developed from brain tissue are called primary brain tumours while tumours that spread from other body sites to the brain are termed metastatic or secondary brain tumours. Type of the brain cancer indicates what kind of brain cells that gave rise to the tumour. Staging of brain cancer indicates the extent of spread of the cancer. The symptoms of brain and CNS cancer include weakness, difficulty in walking and/or dizziness, seizures, headaches, nausea, vomiting and blurry vision. Also, it has been reported that there is a change in person's alertness, mental capacity, memory, speech or personality is observed. Some patients may even hallucinate. There is no specific age reported when they occur but mostly they are found in young children and adults after 40 years of age. The most common brain tumours found in adults are Glioblastoma and meningioma [5]. According to the statistical report of the UK cancer organization, there were around 10,600 people diagnosed with Brain, other CNS and Intracranial Tumours in UK between years 2011 and 2013. From 2012 to 2014, around 5200 people have been reported dead due to Brain, other CNS and Intracranial Tumours [6].

CHAPTER 3 - DATASETS AND METHODS

3.1. METHODOLOGY

The methodology adopted is described below:

I. Download Brain and CNS cancer datasets present in the Oncomine [7] database which satisfy following criteria:
   a. Include both normal & tumour samples with each type having a count of more than one
   b. Experiment type is mRNA

II. For each downloaded dataset
   a. Perform log_2 normalization if data is not normalized
   b. Create sub-datasets, if required

III. For each sub-dataset/dataset
   a. Determine negatively significant genes using SAM [8] using MeV [9]
   b. If there are 0 negative significant genes, omit the corresponding sub-dataset/dataset from further analysis

IV. Consider all datasets/sub-datasets
   a. For each pair of microarray datasets A and B, calculate Z score using
   b. Z score = $(R_{obs} - n_B P_A)/\text{sqrt}\{ n_B P_A (1 - P_A)\}$ where
      i. R_{obs} is the number of genes up-regulated in both A and B
      ii. n_B is number of genes up-regulated in B
      iii. P_A is probability of gene being up-regulated in A [10]
   c. Reject datasets having Z score < 1.96

V. For each selected dataset,
   a. Rank significantly up-regulated genes using RankProdIt [11]
   b. Select genes having pfp value < 0.15 [11]

VI. Combine all selected genes & remove duplicates

VII. Determine functional annotation for selected genes using DAVID [12][13]

VIII. Create a network using Genemania [14] with official gene symbols as input

IX. 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 [14] 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

3.2. SOFTWARE USED

3.2.1. MeV

The TIGR Multi-experiment viewer can analyse normalized and filtered expression files. It can be defined as a versatile microarray data analysis tool. This is because it includes sophisticated algorithms for clustering, visualization, classification, statistical analysis and biological theme discovery. One of the major advantage of this tool is that it can take several types input file formats. Using this tool, the user can obtain informative as well as interrelated displays of expression and annotation data from single or multiple experiments [9]. Significance Analysis of Microarrays [8] can be performed using this software.

3.2.2. SAM

3.2.2.1. INTRODUCTION

SAM [8] can be defined as a statistical technique for predicting whether the changes appearing in gene expression are statistically significant. Using this approach, significant genes can be picked based on differential expression between sets of samples. It is very useful in situations where there is a-priori hypothesis that some genes will have significantly different mean expression levels between different sets of samples [8].

3.2.2.2. FEATURES

One of the most valuable feature of SAM is that it provides an estimate of FDR. To add on, it is a very interactive algorithm. By tuning the parameter delta the users can set thresholds for significance after looking at the distribution generated based on the test statistic [8].

3.2.2.3. METHODOLOGY

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. This analysis uses non-parametric statistics, since the data may not follow a normal distribution. The response variable describes and groups the data based on experimental conditions. In this method, repeated permutations of the data are used to determine if the expression of any gene is significant related to the response. The use of permutation-based analysis accounts for correlations in genes and avoids parametric assumptions about the distribution of individual genes. This is an advantage over other techniques (e.g., ANOVA and Bonferroni), which assume equal variance and/or independence of genes. The data for each gene are permuted, and a test statistic d is computed for both the original and the permuted data for each gene. In the two-class unpaired design, d is analogous to the t-statistic in a t-test, in that it captures the difference among mean expression levels of experimental conditions, scaled by a measure of variance in the data. Missing values in the input data matrix are imputed by one of two methods: Row average and K nearest neighbours [8].

3.2.3. RANKPROD

For the identification of differentially expressed genes across multiple datasets, 'rank product' method can be used. It is a non-parametric method implemented in the RankProd package [11] [15]. RankProd is a biologically intuitive algorithm and statistically rigorous, which has been shown to be robust against noise in microarray data [16] [17]. This algorithm is shown to have higher specificity and sensitivity as compared to other types of meta-analytic tools for microarrays [15]. Based on the conservative estimation of the percentage of false positive predictions (pfp), a list of up-regulated genes is created [10]. Estimation of the percentage of false positive predictions (pfp), is also known as the false discovery rate. As recommended, a pfp value of, 0.15 [11] can be used to set the threshold for genes that are significantly up-regulated.

3.2.4. DAVID

Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is a tool that provides a comprehensive set of functional annotation tools for analysis of a huge list of genes. Tools provided by DAVID are: Functional Annotation, Gene Functional Classification, Gene ID conversion and Gene Name Batch Viewer [12][13].

3.2.5. GENEMANIA

To determine the related genes, this tool searches many large, publicly available biological datasets which include protein-protein, protein-DNA and genetic interactions, pathways, reactions, gene and protein expression data, protein domains and phenotypic screening profiles. Data is regularly updated. The networks names describe the data source and are either generated from the PubMed entry associated with the data source (First author-last author-year), or simply the name of the data source (BioGRID, PathwayCommons-(original data source), Pfam). Two genes are linked in a co-expression network if their expression levels are similar across conditions in a gene expression study. Most of these data are collected from the Gene Expression Omnibus (GEO); we only collect data associated with a publication. Two gene products are linked in a physical interaction network if they were found to interact in a protein-protein interaction study. These data are collected from primary studies found in protein interaction databases, including BioGRID and PathwayCommons. Two genes are functionally associated if the effects of perturbing one gene were found to be modified by perturbations to a second gene. These data are collected from primary studies and BioGRID. Two gene products are linked in shared protein domains network if they have the same protein domain. These data are collected from domain databases, such as InterPro, SMART and Pfam. Two genes are linked in 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. Two gene products are linked in pathway network if they participate in the same reaction within a pathway. These data are collected from various source databases, such as Reactome and BioCyc, via PathwayCommons [14].

3.3. DATASETS

In this study, four brain and CNS cancer datasets (Table1) from Oncomine [7] database were chosen as they contained a differential analysis of tumour and normal samples, experiment type was mRNA and number of samples in both tumour and normal category was more than one. Oncomine [7] currently contain 715 datasets investigating tumour types (Oncomine Research Edition) and it is one of the most comprehensive cancer-specific database. The major advantage of using this database is that prior to inclusion in Oncomine database, the microarray datasets obtained from public resources such as Stanford Microarray Database and the NCBI Gene Expression Omnibus or literature sources are reviewed by a panel of experts to ensure that they meet certain quality standards [7].

Dataset Name*	**Accession ID**	**Genes****	**Platform**
Bredel Brain 2	GSE2223	14386	Platform not pre-defined in Oncomine
Lee Brain	GSE4536	19574	Human Genome U133 Plus 2.0 Array
Liang Brain	GSE4058	9957	Platform not pre-defined in Oncomine
Murat Brain	GSE7696	19574	Human Genome U133 Plus 2.0 Array

Table1: Brain and CNS cancer microarray datasets included in the study

*As identified by the Oncomine database, **Number of genes probed

CHAPTER 4 - RESULTS AND DISCUSSION

4.1. SUB-DATASETS CREATION

Bredel Brain 2 dataset consists specimens belonging to astrocytic, glioblastomas and oligodendroglial types of brain tumours. Hence this dataset was divided on the basis of these tumours and their corresponding subparts and six sub-datasets were created. Lee Brain dataset was divided based on the region from where the tumour was extracted. For instance, a separate sub-dataset was created for each tumour cell line. In the end, thirteen sub-datasets were created. Liang Brain dataset contained data from 5 different platforms. The normal samples belonged to platform GPL2935 and on comparison we found that platform GPL182, GPL2778 and GPL2935 have identical genes whereas GPL2648 and GPL3010 have different genes. So the samples belonging to the latter two platforms were removed from further analysis. The sub-datasets were created on the basis of different types of tumours. In Murat Brain, the dataset was divided to three sub-datasets based on the various treatments provided.

4.2. IDENTIFICATION OF UP-REGULATED GENES

For each sub-dataset, SAM [8] was performed using the software MeV [9]. The results of SAM [8] are mentioned below.

4.2.1. BREDEL BRAIN 2

BB2SP1 - MEV – SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 670

Number of Negative Significant Genes: 20

Total number of Significant Genes: 690

Total number of Non-Significant Genes: 40782

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G40053	40329	0.8480307	-3.5481935	-3.7040496	1.0439255
G27800	27992	0.20470008	-3.3564348	-3.0988002	0.92324156
G8780	8840	-0.3424735	-3.251699	-3.08695	0.9493345
G6106	6148	-0.45969656	-3.2017946	-3.7957	1.1854914
G41019	41301	1.0958507	-3.1818874	-2.5707	0.8079167

Table2: List of top significantly up-regulated genes in BB2SP1

BB2SP2 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 918

Number of Negative Significant Genes: 363

Total Number of Significant Genes: 1281

Total Number of Non-Significant Genes: 40191

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G6106	6148	-0.66767436	-7.1928015	-5.9088335	0.8214927
G13952	14048	-0.25832725	-5.823594	-4.3501577	0.7469885
G5211	5247	-0.7342642	-5.765403	-3.1068242	0.53887373
G27800	27992	0.28035733	-5.6538377	-3.8126483	0.674347
G40665	40947	1.379309	-5.550627	-2.6671202	0.4805079

Table3: List of top significantly up-regulated genes in BB2SP2

BB2SP3 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 5

Number of Negative Significant Genes: 260

Total Number of Significant Genes: 265

Total Number of Non-Significant Genes: 41207

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G6106	6148	-0.39476305	-5.3024573	-6.3835	1.2038758
G34653	34893	0.39388344	-5.0297813	-6.5509996	1.3024422
G6903	6945	-0.36326733	-4.748206	-6.258	1.3179713
G25134	25308	0.1031969	-4.673735	-5.22875	1.1187519
G14765	14867	-0.12811269	-4.6702056	-6.183	1.3239247

Table4: List of top significantly up-regulated genes in BB2SP3

BB2SP4 - MEV- SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 147

Number of Negative Significant Genes: 112

Total Number of Significant Genes: 259

Total Number of Non-Significant Genes: 41213

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G6106	6148	-0.2976352	-4.4902463	-7.1215	1.5859932
G7050	7098	-0.26795313	-3.5124516	-5.5905	1.5916233
G5402	5438	-0.3220763	-3.5098374	-5.7085004	1.6264287
G17270	17384	-0.044481397	-3.2619896	-5.2217503	1.6007869
G26935	27121	0.11490814	-3.110573	-5.19075	1.668744

Table5: List of top significantly up-regulated genes in BB2SP4

BB2SP5 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 56

Number of Negative Significant Genes: 124

Total Number of Significant Genes: 180

Total Number of Non-Significant Genes: 41292

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G12533	12617	-0.2577917	-4.7498717	-3.95425	0.83249617
G13097	13187	-0.23634331	-4.6689863	-3.8734503	0.8296127
G41441	41723	1.9409068	-4.5426674	-3.0896502	0.6801401
G31066	31276	0.37751144	-4.402325	-2.20825	0.5016099
G24285	24453	0.1310783	-4.2328587	-3.7879498	0.8948916

Table6: List of top significantly up-regulated genes in BB2SP5

BB2SP6 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 157

Number of Negative Significant Genes: 35

Total Number of Significant Genes: 192

Total Number of Non-Significant Genes: 41280

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G12461	12545	-0.14428593	-4.426191	-5.5885	1.262598
G6106	6148	-0.3376365	-3.9877808	-6.2510004	1.5675386
G6903	6945	-0.3079923	-3.7153883	-4.8775	1.3127834
G34971	35211	0.36771187	-3.591295	-5.1215	1.4260873
G11229	11301	-0.17633475	-3.5836854	-5.204612	1.4523071

Table7: List of top significantly up-regulated genes in BB2SP6

5.2.2. LEE BRAIN

LBSP1 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 2

Number of Negative Significant Genes: 502

Total Number of Significant Genes: 504

Total Number of Non-Significant Genes: 54109

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	ID_REF	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G10611	201162_at	-0.41715127	-12.722399	-10.279678	0.8079984
G12988	203540_at	-0.3374426	-12.450179	-9.005069	0.7232883
G10612	201163_s_at	-0.4171193	-11.620332	-7.586791	0.6528894
G20350	210982_s_at	-0.13542652	-11.507237	-9.192316	0.7988291
G18310	208894_at	-0.18700576	-10.82327	-8.552448	0.7901908

Table8: List of top significantly up-regulated genes in LBSP1

LBSP2 – MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 11

Number of Negative Significant Genes: 384

Total Number of Significant Genes: 395

Total Number of Non-Significant Genes: 54218

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G26700	217414_x_at	-0.001254091	-13.043336	-10.690708	0.8196299
G27053	217767_at	0.007873991	-12.901984	-8.914035	0.6909042
G20350	210982_s_at	-0.1669093	-12.550909	-9.805683	0.78127277
G10611	201162_at	-0.46261314	-12.191934	-10.868733	0.8914692
G10612	201163_s_at	-0.46257448	-11.941074	-7.8297772	0.6557012

Table9: List of top significantly up-regulated genes in LBSP2

LBSP3 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 11

Number of Negative Significant Genes: 384

Total Number of Significant Genes: 395

Total Number of Non-Significant Genes: 54218

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G12627	203178_at	-0.45897222	-16.733084	-7.9216223	0.4734108
G31764	222484_s_at	0.15820052	-15.21884	-8.321319	0.5467775
G13337	203889_at	-0.43240896	-14.812105	-6.8369436	0.46157813
G38514	229259_at	0.3845055	-14.300953	-7.918677	0.5537167
G21977	212671_s_at	-0.14448987	-13.877378	-5.880843	0.42377192

Table10: List of top significantly up-regulated genes in LBSP3

LBSP4 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 32

Number of Negative Significant Genes: 170

Total Number of Significant Genes: 202

Total Number of Non-Significant Genes: 54411

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G10773	201324_at	-0.6244577	-15.17227	-6.742297	0.4443829
G14887	205439_at	-0.4329719	-14.965078	-6.9934874	0.46732047
G18382	208966_x_at	-0.29140857	-14.94221	-5.9663873	0.3992975
G18774	209360_s_at	-0.27631718	-13.95684	-5.4371285	0.38956732
G15380	205932_s_at	-0.41199967	-13.628781	-4.4402466	0.32579923

Table11: List of top significantly up-regulated genes in LBSP4

LBSP5 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 247

Number of Negative Significant Genes: 165

Total Number of Significant Genes: 412

Total Number of Non-Significant Genes: 54201

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G12635	203186_s_at	-0.26768628	-7.1560173	-9.845611	1.3758506
G33215	223939_at	0.09670599	-6.5314183	-9.507941	1.4557239
G12683	203234_at	-0.2666661	-6.229208	-9.241017	1.483498
G21795	212489_at	-0.09206684	-6.178988	-8.352216	1.3517126
G11853	202404_s_at	-0.28523046	-6.031565	-8.512004	1.411243

Table12: List of top significantly up-regulated genes in LBSP5

LBSP6 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 247

Number of Negative Significant Genes: 165

Total Number of Significant Genes: 412

Total Number of Non-Significant Genes: 54201

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G12629	203180_at	-0.30377248	-9.486974	-10.637856	1.1213118
G20350	210982_s_at	-0.1359427	-9.160765	-9.995325	1.0911016
G10611	201162_at	-0.357209	-8.43605	-10.307954	1.2218933
G10558	201109_s_at	-0.3586634	-7.7537894	-9.420953	1.2150127
G23911	214612_x_at	-0.06799289	-7.478318	-9.3034725	1.2440594

Table13: List of top significantly up-regulated genes in LBSP6

LBSP7 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 247

Number of Negative Significant Genes: 165

Total Number of Significant Genes: 412

Total Number of Non-Significant Genes: 54201

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G11358	201909_at	-0.583945	-18.000202	-5.7069297	0.3170481
G10773	201324_at	-0.61298096	-17.253998	-6.7276807	0.38992012
G44667	235417_at	0.6614682	-17.219368	-6.2822895	0.36483857
G13857	204409_s_at	-0.46937713	-16.250734	-6.1484766	0.3783507
G16712	207266_x_at	-0.35251302	-15.633987	-8.483693	0.54264426

Table14: List of top significantly up-regulated genes in LBSP7

LBSP8 – MEV–SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 193

Number of Negative Significant Genes: 141

Total Number of Significant Genes: 334

Total Number of Non-Significant Genes: 54279

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G10887	201438_at	-0.40441152	-15.106222	-10.233854	0.67745954
G28967	219682_s_at	0.044994902	-11.471538	-8.468	0.73817486
G21795	212489_at	-0.113712296	-10.889008	-7.848991	0.72081786
G41021	231766_s_at	0.33084336	-10.632534	-8.086908	0.7605815
G10955	201506_at	-0.40206048	-10.508613	-9.999392	0.9515425

Table15: List of top significantly up-regulated genes in LBSP8

LBSP9 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 144

Number of Negative Significant Genes: 105

Total Number of Significant Genes: 249

Total Number of Non-Significant Genes: 54364

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G11759	202310_s_at	-0.3856365	-17.63513	-11.875833	0.67341906
G2456	1555623_at	-0.92150146	-13.353509	-10.8735695	0.8142856
G10887	201438_at	-0.4145456	-12.172455	-10.120651	0.8314388
G10611	201162_at	-0.4240482	-11.863988	-9.275802	0.78184515
G11853	202404_s_at	-0.38260493	-11.039992	-8.61555	0.78039455

Table16: List of top significantly up-regulated genes in LBSP9

LBSP10 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 222

Number of Negative Significant Genes: 320

Total Number of Significant Genes: 542

Total Number of Non-Significant Genes: 54071

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G15619	206172_at	-0.20873868	-10.665521	-9.869117	0.9253291
G21795	212489_at	-0.08268734	-10.588763	-8.875637	0.8382128
G23911	214612_x_at	-0.043277383	-10.426174	-9.513362	0.91244996
G29730	220445_s_at	0.06208492	-9.549703	-7.6201315	0.79794437
G16712	207266_x_at	-0.18474282	-9.264542	-8.119384	0.8763935

Table17: List of top significantly up-regulated genes in LBSP10

LBSP11 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 189

Number of Negative Significant Genes: 115

Total Number of Significant Genes: 304

Total Number of Non-Significant Genes: 54309

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G11759	202310_s_at	-0.3511692	-14.644005	-11.457836	0.78242505
G10887	201438_at	-0.3773877	-11.752374	-10.054925	0.8555654
G10955	201506_at	-0.37520188	-9.77415	-9.816697	1.004353
G16712	207266_x_at	-0.22257115	-9.231953	-8.701449	0.9425362
G11853	202404_s_at	-0.3484453	-9.0107975	-8.227518	0.91307324

Table18: List of top significantly up-regulated genes in LBSP11

LBSP12 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 189

Number of Negative Significant Genes: 115

Total Number of Significant Genes: 304

Total Number of Non-Significant Genes: 54309

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G10611	201162_at	-0.305745	-7.899402	-11.181268	1.4154575
G40984	231729_s_at	0.23225701	-7.745438	-9.983899	1.2890038
G10558	201109_s_at	-0.30712357	-6.8821483	-9.636532	1.4002215
G15619	206172_at	-0.19690861	-6.4786205	-9.78559	1.5104434
G11291	201842_s_at	-0.28869736	-6.4515076	-9.183227	1.4234234

Table19: List of top significantly up-regulated genes in LBSP12

LBSP13 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 189

Number of Negative Significant Genes: 115

Total Number of Significant Genes: 304

Total Number of Non-Significant Genes: 54309

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G2456	1555623_at	-0.6829828	-8.211041	-11.294678	1.3755475
G10611	201162_at	-0.31485403	-7.7620497	-11.314462	1.4576641
G40984	231729_s_at	0.23509508	-7.760091	-10.33763	1.3321532
G10558	201109_s_at	-0.3162393	-6.803365	-9.840644	1.4464377
G15619	206172_at	-0.2026293	-6.7408834	-10.45511	1.5509999

Table20: List of top significantly up-regulated genes in LBSP13

4.2.3. LIANG BRAIN

LISP1 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 230

Number of Negative Significant Genes: 17

Total Number of Significant Genes: 247

Total Number of Non-Significant Genes: 23945

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G9131	9131	-0.25554422	-9.032619	-2.722857	0.3014471
G6095	6095	-0.5443013	-8.002071	-2.9275715	0.3658517
G7612	7612	-0.39439735	-7.577522	-2.7129047	0.35802004
G17541	17541	0.47575113	-7.0881057	-2.2450116	0.31672943
G8068	8068	-0.35177836	-6.769095	-1.9995314	0.29539123

Table21: List of top significantly up-regulated genes in LISP1

LISP2 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 230

Number of Negative Significant Genes: 17

Total Number of Significant Genes: 247

Total Number of Non-Significant Genes: 23945

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G8092	8092	-0.38260564	-16.10817	-3.012	0.18698587
G12911	12911	0.0775888	-13.705022	-2.6815834	0.19566429
G18204	18204	0.59442365	-11.722513	-4.1274166	0.35209316
G12148	12148	0.006689409	-10.656303	-3.6226666	0.33995527
G18193	18193	0.59311056	-10.000409	-3.96	0.39598382

Table22: List of top significantly up-regulated genes in LISP2

LISP3 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 230

Number of Negative Significant Genes: 17

Total Number of Significant Genes: 247

Total Number of Non-Significant Genes: 23945

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G7043	7043	-0.44924983	-12.658769	-4.9644995	0.3921787

Table23: List of top significantly up-regulated genes in LISP3

4.2.4. MURAT BRAIN

MBSP1 - MEV-SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 230

Number of Negative Significant Genes: 17

Total Number of Significant Genes: 247

Total Number of Non-Significant Genes: 23945

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G47271	238021_s_at	1.0787748	-16.931166	-1.0434158	0.06162693

Table24: List of top significantly up-regulated genes in MBSP1

MBSP2 –MEV –SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 230

Number of Negative Significant Genes: 17

Total Number of Significant Genes: 247

Total Number of Non-Significant Genes: 23945

LIST OF TOP SIGNIFICANTLY UPREGULATED GENES

GID	GNAME	Expected score (dExp)	Observed score(d)	Numerator(r)	Denominator (s+s0)
G47271	238021_s_at	1.0905435	-9.332379	-1.0732322	0.11500092

Table25: List of top significantly up-regulated genes in MBSP2

MBSP3–MEV –SAM

CLUSTER INFORMATION

Number of Positive Significant Genes: 10

Number of Negative Significant Genes: 0

Total Number of Significant Genes: 10

Total Number of Non-Significant Genes: 54665

4.3. CONGRUENCY BETWEEN MICROARRAY DATASETS

To determine the congruency between microarray datasets, Z score is calculated for all possible pairs of sub-datasets through a Perl script. The formula used is: Z score = (Robs – nBPA)/sqrt{nBPA (1 - PA)} 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 [10]. Since the sub-dataset MBSP3 contains 0 significantly up-regulated genes hence it has not been considered in Z score calculation. The sub-datasets which that had pairwise z score > 1.96 were considered for further analysis. The selected sub-datasets are: BB2SP1, BB2SP2, BB2SP3, BB2SP4, BB2SP5, BB2SP6, LBSP1, LBSP2, LBSP3, LBSP4, LBSP5, LBSP6, LBSP7, LBSP8 , LBSP9, LBSP10, LBSP11, LBSP12, and LBSP13. All the sub-datasets belonging to Liang Brain and Murat Brain were thus excluded from further analysis.

	BB2SP1	BB2SP2	BB2SP3	BB2SP4	BB2SP5	BB2SP6	LBSP1	LBSP2	LBSP3	LBSP4	LBSP5	LBSP6	LBSP7	LBSP8	LBSP9	LBSP10	LBSP11	LBSP12	LBSP13	LIBSP1	LIBSP2	LIBSP3	MBSP1	MBSP2
BB2SP1		35.491151	13.797505	30.169298	65.200388	7.5690304	13.73813	13.51567	14.41603	13.68703	10.35552	9.948795	15.72877	11.24662	13.10995	7.2456635	12.506503	8.7800731	10.162141	10.956394	-0.090566	-0.021966	-0.021966	-0.021966
BB2SP2			88.00194	55.391691	33.710166	21.250753	46.83879	49.1816	46.07779	48.24785	44.82533	47.90067	55.50273	51.39614	57.79013	34.379499	54.13039	42.310952	46.455171	7.4352014	-0.386906	-0.093838	-0.093838	-0.093838
BB2SP3				79.048107	17.356074	36.006643	44.11416	58.68937	46.03869	41.80311	45.42723	45.56254	50.30842	46.09475	58.649	30.513852	51.232367	41.96944	44.586768	5.8304889	-0.326857	-0.079274	-0.079274	-0.079274
BB2SP4					18.631866	19.413465	28.34883	36.69244	30.68205	20.20429	14.47374	31.37249	21.58017	19.03615	18.44576	14.293856	13.951348	16.076843	17.167536	-0.212628	-0.212628	-0.05157	-0.05157	-0.05157
BB2SP5						15.155376	25.74911	27.90118	30.77561	37.21366	36.36875	34.96214	42.74756	39.45308	45.9113	25.640819	43.818702	30.92529	35.699881	4.2163501	4.2163501	-0.054763	-0.054763	-0.054763
BB2SP6							20.86689	22.27619	18.14026	18.10951	26.43594	17.68045	20.81098	19.9558	26.58794	14.880854	25.378557	15.634711	18.054258	-0.11983	-0.11983	-0.029063	-0.029063	-0.029063
LBSP1								116.2882	100.5143	71.07433	39.55049	45.83901	56.8007	49.15613	52.18835	52.167756	48.800745	53.04371	46.018066	-0.397131	-0.397131	-0.096318	-0.096318	-0.096318
LBSP2									121.118	41.12572	47.3671	57.18395	37.87746	48.38648	48.19114	41.981047	44.85376	55.385388	51.981275	-0.346956	-0.346956	-0.084149	-0.084149	-0.084149
LBSP3										58.44789	45.50126	63.74231	42.58437	55.80559	56.26158	43.358579	51.314921	53.904533	55.354014	-0.326341	-0.326341	-0.079149	-0.079149	-0.079149
LBSP4											32.82274	46.34772	103.2766	56.78442	39.72079	62.222086	37.902419	54.213048	41.825611	-0.230398	-0.230398	-0.05588	-0.05588	-0.05588
LBSP5												81.20713	32.93147	68.3967	70.56192	74.388949	69.071999	67.197901	75.915239	-0.226974	-0.226974	-0.055049	-0.055049	-0.055049
LBSP6													42.43372	87.973	66.1895	77.439529	58.284769	89.205242	101.21989	-0.235774	-0.235774	-0.057184	-0.057184	-0.057184
LBSP7														54.72113	53.57028	66.807832	43.488339	58.093524	48.008398	-0.201403	-0.201403	-0.048847	-0.048847	-0.048847
LBSP8															91.78817	75.100439	82.146387	92.512197	94.273183	-0.209772	-0.209772	-0.050877	-0.050877	-0.050877
LBSP9																65.57384	122.99617	84.566852	74.491786	-0.180963	-0.180963	-0.04389	-0.04389	-0.04389
LBSP10																	66.375725	111.52861	100.1942	-0.316539	-0.316539	-0.076772	-0.076772	-0.076772
LBSP11																		79.299582	69.465421	-0.189401	-0.189401	-0.045937	-0.045937	-0.045937
LBSP12																			171.07546	-0.265194	-0.265194	-0.064319	-0.064319	-0.064319
LBSP13																				-0.231076	-0.231076	-0.056044	-0.056044	-0.056044
LIBSP1																					-0.109337	-0.026518	-0.026518	-0.026518
LIBSP2																						-0.026518	-0.026518	-0.026518
LIBSP3																							-0.006429	-0.006429
MBSP1																								-0.004277
MBSP2																								

Figure1: Pairwise Z score of sub-datasets

4.4. IDENTIFICATION OF UPREGULATED GENES

Combining the result of SAM for all the selected sub-datasets, 3861 significant genes were obtained. To improve the result of SAM, the significant genes were further analysed using RankProd [11] with pfp threshold of < 0.15. 271 genes were found to be significantly up-regulated using RankProd [11] program and after removing duplicates, there were 130 significant genes.

GENE SYMBOL	DESCRIPTION
ALDH1A3	aldehyde dehydrogenase 1 family, member A3
ANXA2	annexin A2
C1QC	complement component 1, q subcomponent, C chain
C3	complement component 3
CHI3L2	Chitinase 3-like 2 : chitinase precursor
COL1A1	collagen, type I, alpha 1
COL6A2	collagen, type VI, alpha 2
CTGF	connective tissue growth factor
CXCL14	chemokine (C-X-C motif) ligand 14
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked
GBP1	guanylate binding protein 1, interferon-inducible, 67kDa
GFAP	glial fibrillary acidic protein
HBA1 /// HBA2	hemoglobin, alpha 1 /// hemoglobin, alpha 2
HBB	hemoglobin, beta
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1
HLA-DRA	major histocompatibility complex, class II, DR alpha
IGFBP7	insulin-like growth factor binding protein 7
IL13RA2	interleukin 13 receptor, alpha 2
MAGEA6	melanoma antigen family A, 6
MEOX2	mesenchyme homeobox 2
MGST1	microsomal glutathione S-transferase 1
RBMS1	RNA binding motif, single stranded interacting protein 1
RGS1	regulator of G-protein signaling 1
SERPINA3	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3
SRGN	Serglycin
TGFBI	transforming growth factor, beta-induced, 68kDa

Table26: List of top 26 significantly up-regulated genes

4.5. FUNCTIONAL ANALYSIS OF UP-REGULATED GENES

The significant genes obtained through RankProd were entered as an input to DAVID [12] [13]. 123 genes matched out of 130 genes given as an input.

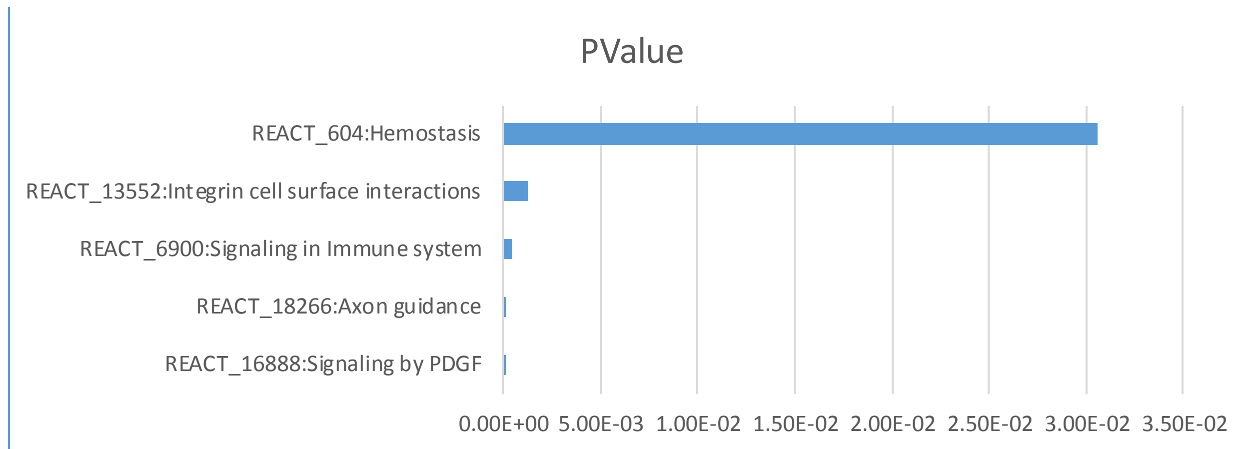

Figure2: Enriched Reactome Pathway obtained through DAVID [12][13].

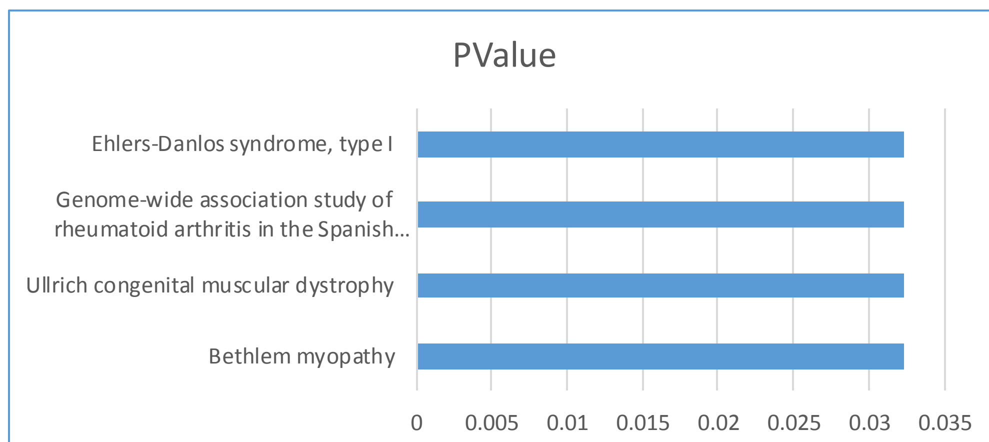

Figure3: Enriched OMIM Diseases obtained through DAVID [12][13]

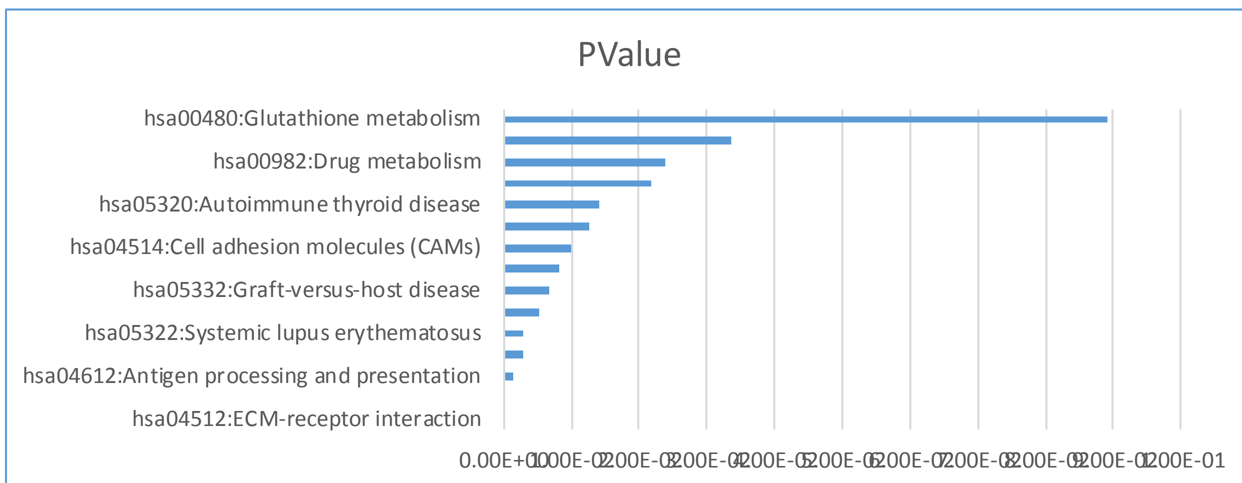

Figure4: Enriched KEGG PATHWAY obtained through DAVID [12][13]

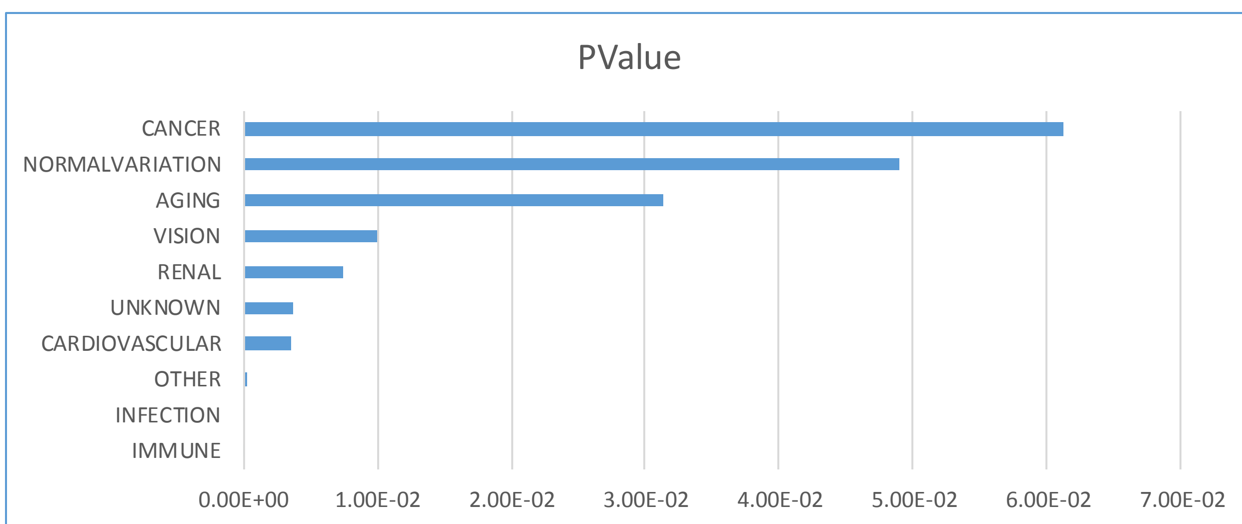

Figure5: Enriched GENETIC ASSOCIATION DB DISEASE CLASS obtained through DAVID [12][13]

4.6. IDENITIFCATION OF SIGNIFICANT PATHWAYS

Using online tool of Genemania [14] a network was created with 127 genes as an input.

The seven types of network which were created are Co-expression, Co-localization, Genetic Interaction, Pathway, Physical Interaction, Predicted and Shared Protein Domains. The predicted pathway contains only one interaction and hence was removed from further analysis.

NETWORK PROPERTIES AND SIGNIFICANT KEGG PATHWAYS FOR CO-EXPRESSION NETWORK

NETWORK PROPERTIES

PROPERTIES	CO-EXPRESSION NETWORK
Clustering Coefficient	0.177
Connected components	2
Network diameter	6
Network radius	1
Shortest paths	4288 (29%)
Characteristic path length	2.091
Average number of neighbours	16.639
Number of nodes	122
Network density	0
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	322
Analysis time (sec)	0.085

Table27: Network Properties of Co-expression network obtained using Network Analyser tool of Cytoscape [18]

SIGNIFICANT PATHWAYS

CELLULAR PROCESSES

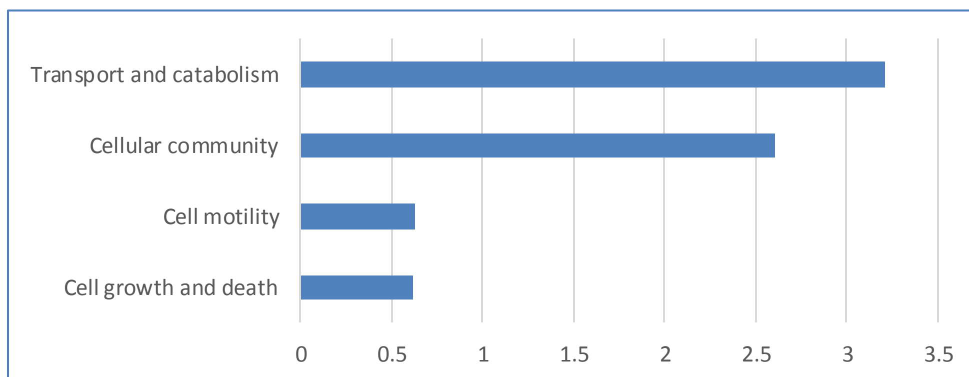

Figure6: Categories of enriched cellular processes KEGG pathways in co-expression network

CELLULAR COMMUNITY

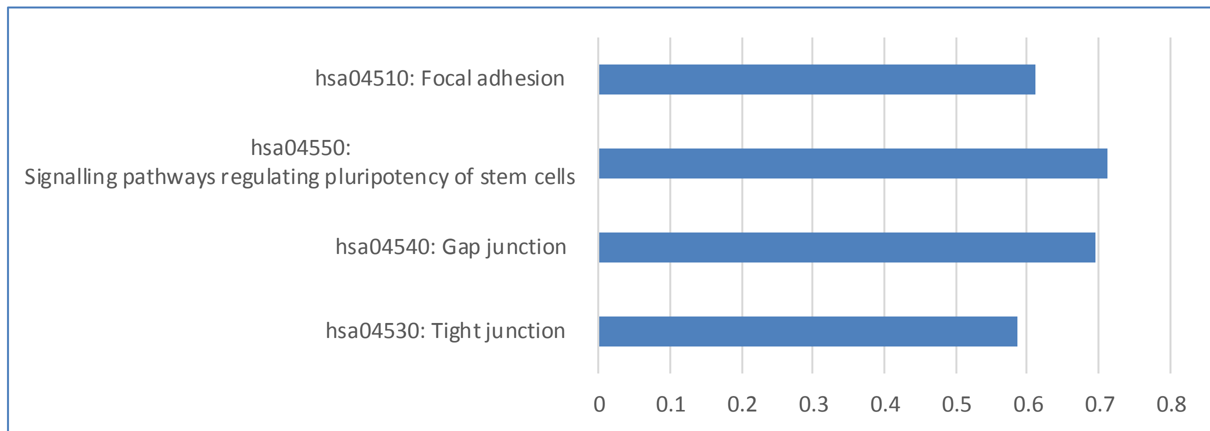

Figure7: Enriched cellular community KEGG pathways in co-expression network

TRANSPORT AND CATABOLISM

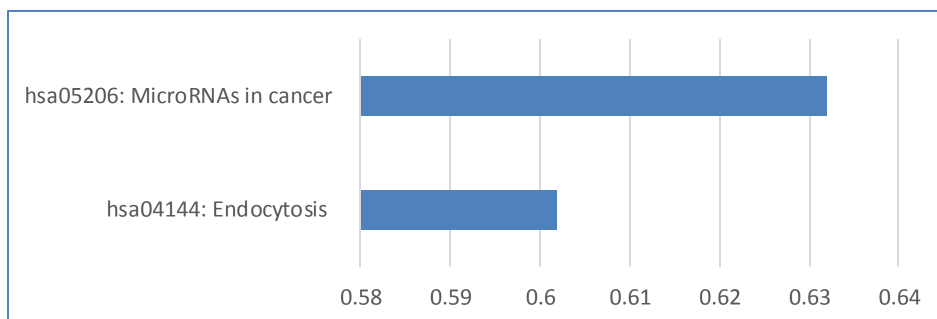

Figure8: Enriched transport and catabolism KEGG pathways in co-expression network

SIGNAL TRANSDUCTION

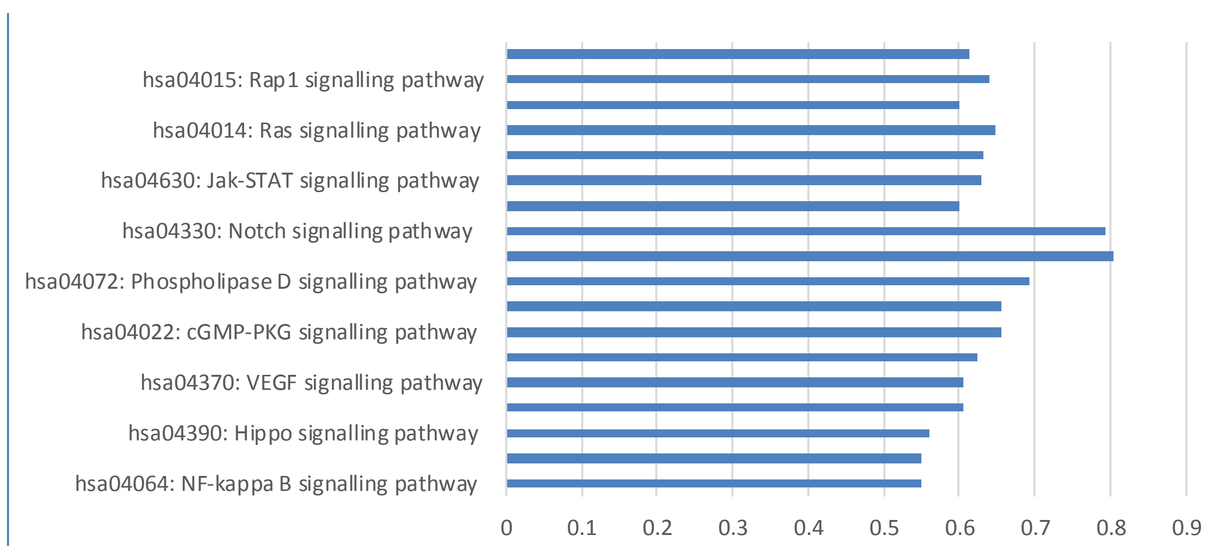

Figure9: Enriched signal transduction KEGG pathways in co-expression network

SIGNALLING MOLECULES AND INTERACTIONS

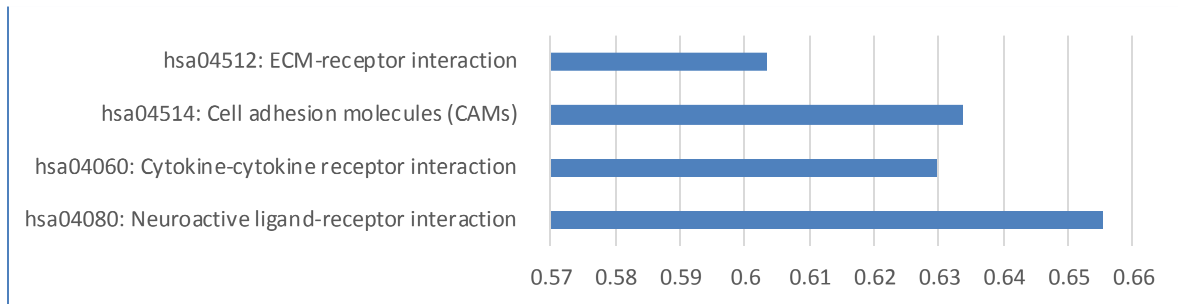

Figure10: Enriched signaling molecules and interactions KEGG pathways in co-expression network

CANCERS

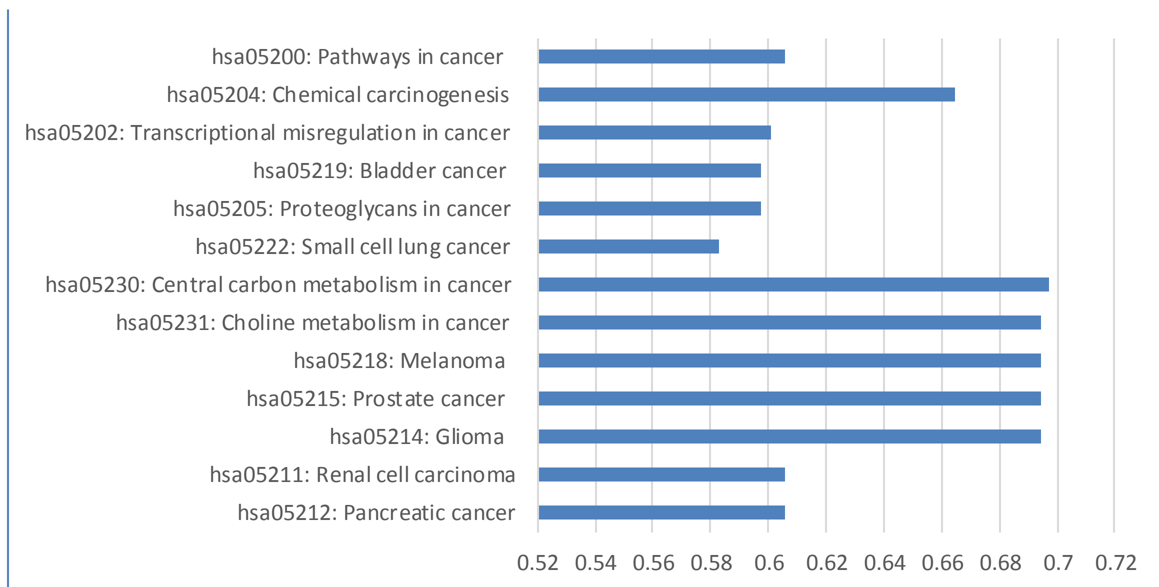

Figure11: Enriched cancer KEGG pathways in co-expression network

NEURODEGENERATIVE DISEASES

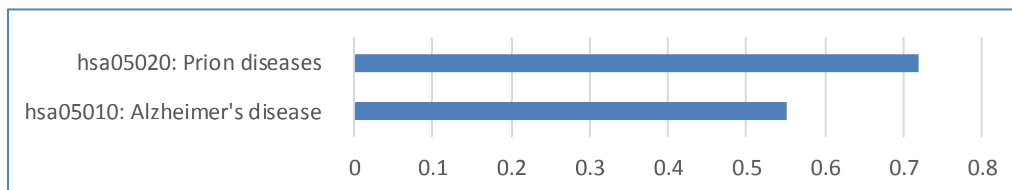

Figure12: Enriched neurodegenerative diseases KEGG pathways in co-expression network

METABOLISM

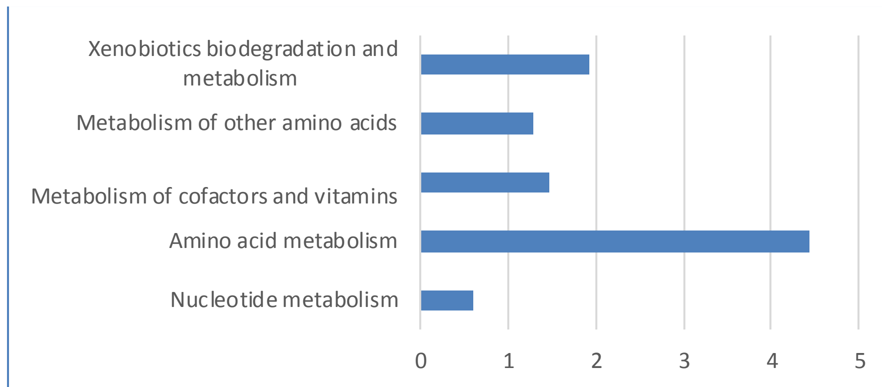

Figure13: Categories of enriched metabolism KEGG pathways in co-expression network

AMINO ACID METABOLISM

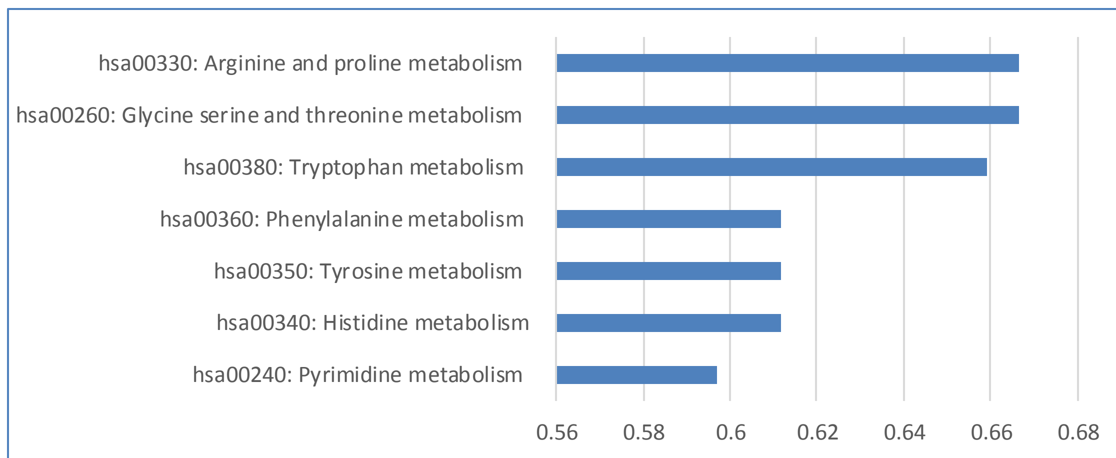

Figure14: Enriched amino acid metabolism KEGG pathways in co-expression network

METABOLISM OF COFACTORS AND VITAMINS

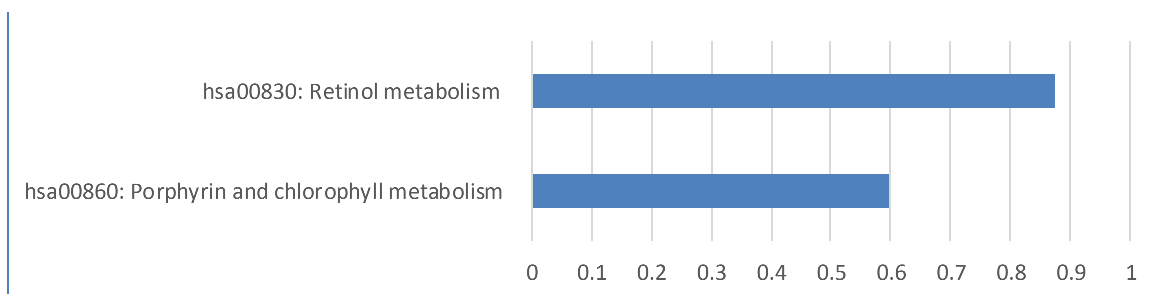

Figure15: Enriched metabolism of cofactors and vitamins KEGG pathways in co-expression network

METABOLISM OF OTHER AMINO ACIDS

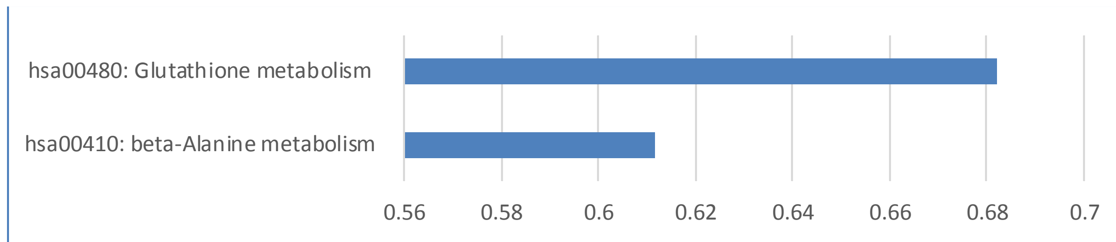

Figure16: Enriched metabolism of other amino acids KEGG pathways in co-expression network

ENDOCRINE SYSTEM

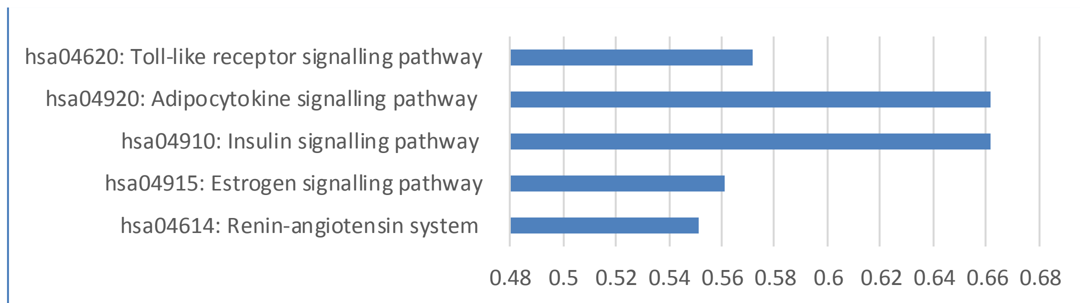

Figure17: Enriched endocrine system KEGG pathways in co-expression network

IMMUNE SYSTEM

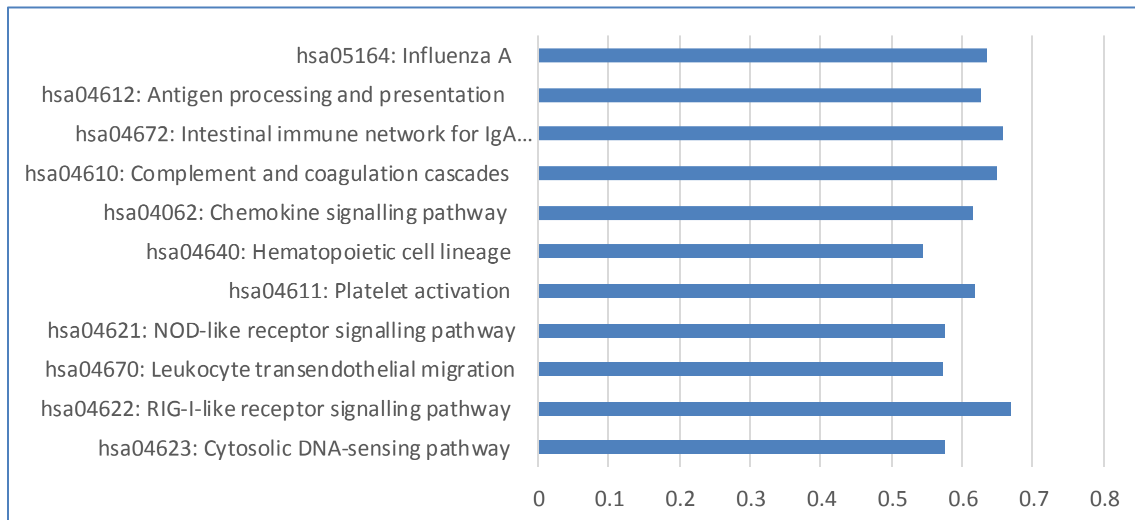

Figure18: Enriched immune system KEGG pathways in co-expression network

NETWORK PROPERTIES AND SIGNIFICANT KEGG PATHWAYS FOR CO-LOCALIZATION NETWORK

NETWORK PROPERTIES

PROPERTIES	CO-LOCALIZATION NETWORK
Clustering Coefficient	0.208
Connected components	3
Network diameter	4
Network radius	1
Shortest paths	319 (5%)
Characteristic path length	1.639
Average number of neighbours	4.184
Number of nodes	76
Network density	0
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	7
Analysis time (sec)	0.303

Table28: Network Properties of Co-localization network obtained using Network Analyser tool of Cytoscape [18]

SIGNIFICANT PATHWAYS

CELLULAR PROCESSES

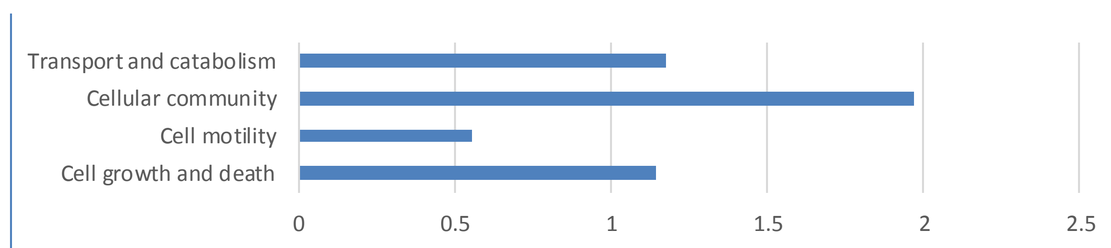

Figure19: Categories of enriched cellular processes KEGG pathways in co-localization network

CELL GROWTH AND DEATH

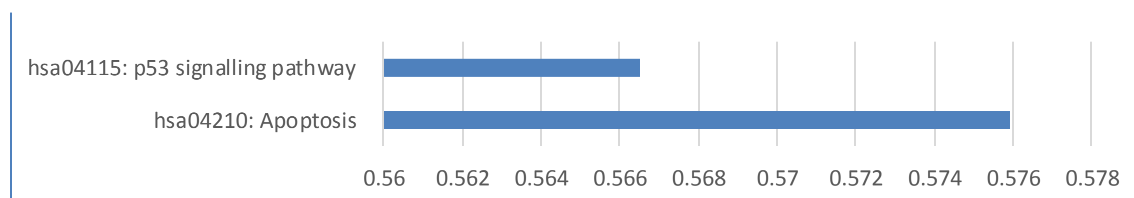

Figure20: Enriched cell growth and death KEGG pathways in co-localization network

CELLULAR COMMUNITY

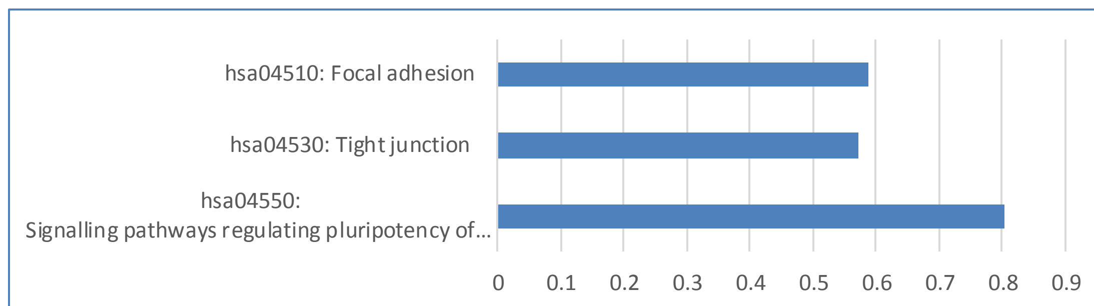

Figure21: Enriched cellular community KEGG pathways in co-localization network

TRANSPORT AND CATABOLISM

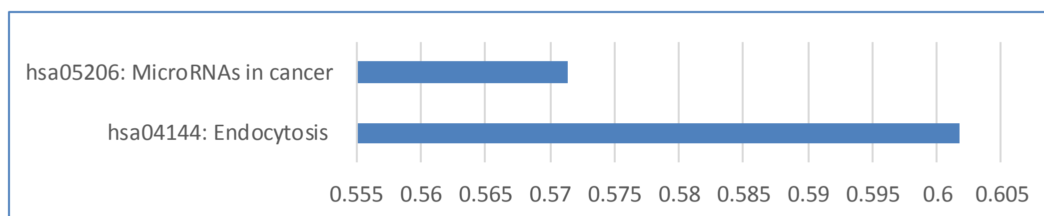

Figure22: Enriched transport and catabolism KEGG pathways in co-localization network

SIGNAL TRANSDUCTION

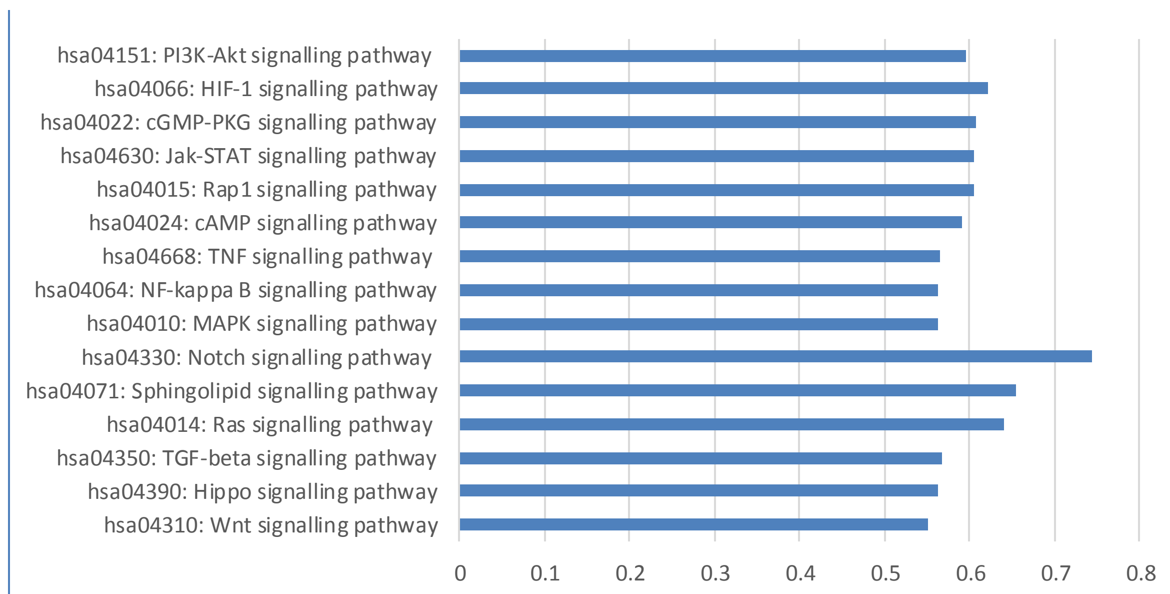

Figure23: Enriched signal transduction KEGG pathways in co-localization network

SIGNALLING MOLECULES AND INTERACTIONS

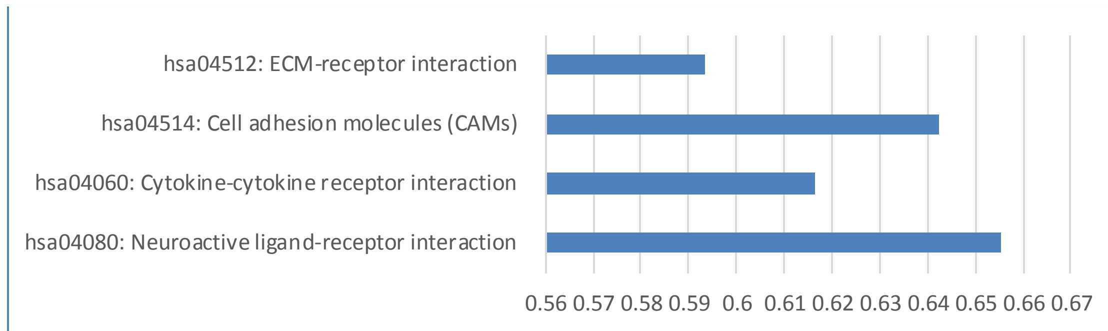

Figure24: Enriched signalling molecules interactions and KEGG pathways in co-localization network

CANCERS

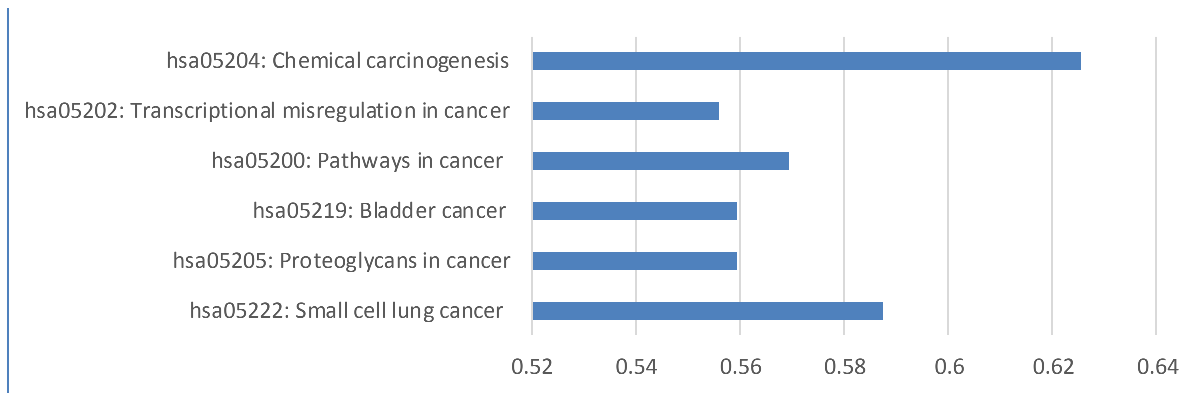

Figure25: Enriched cancer KEGG pathways in co-localization network

METABOLISM

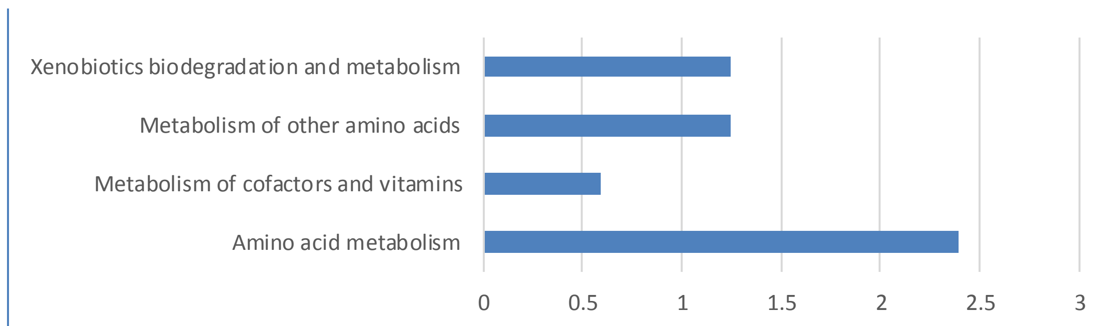

Figure26: Categories of enriched metabolism KEGG pathways in co-localization network

AMINO ACID METABOLISM

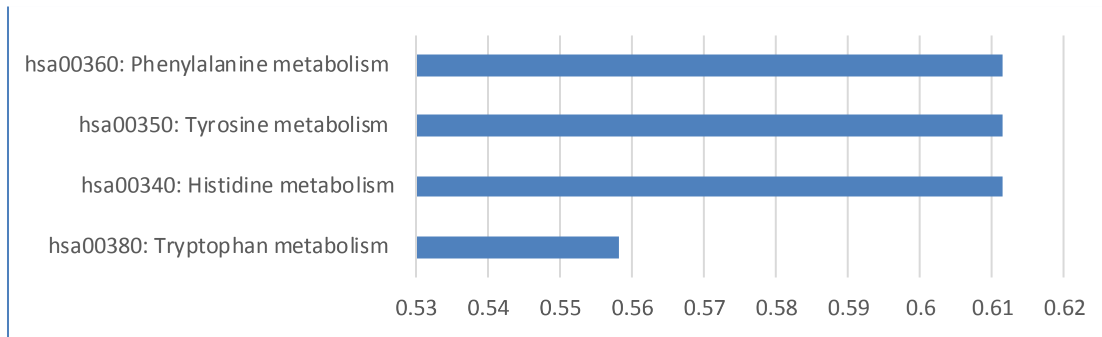

Figure27: Enriched amino acid metabolism KEGG pathways in co-localization network

METABOLISM OF OTHER AMINO ACIDS

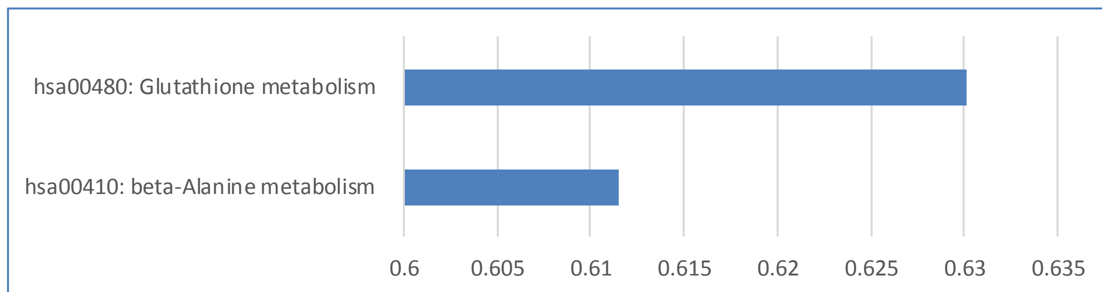

Figure28: Enriched metabolism of other amino acids KEGG pathways in co-localization network

ENDOCRINE SYSTEM

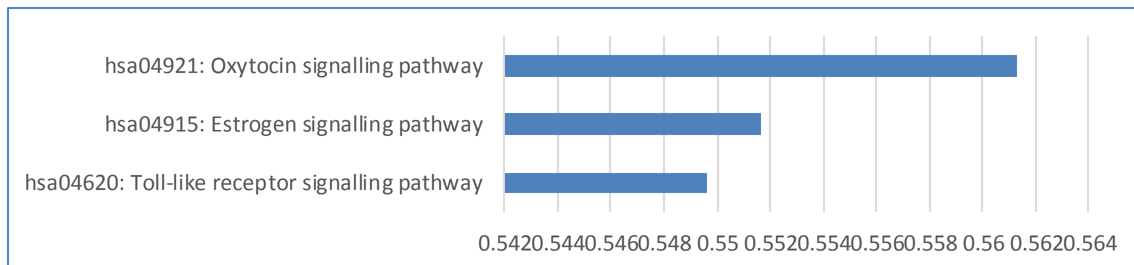

Figure29: Enriched endocrine system KEGG pathways in co-localization network

IMMUNE SYSTEM

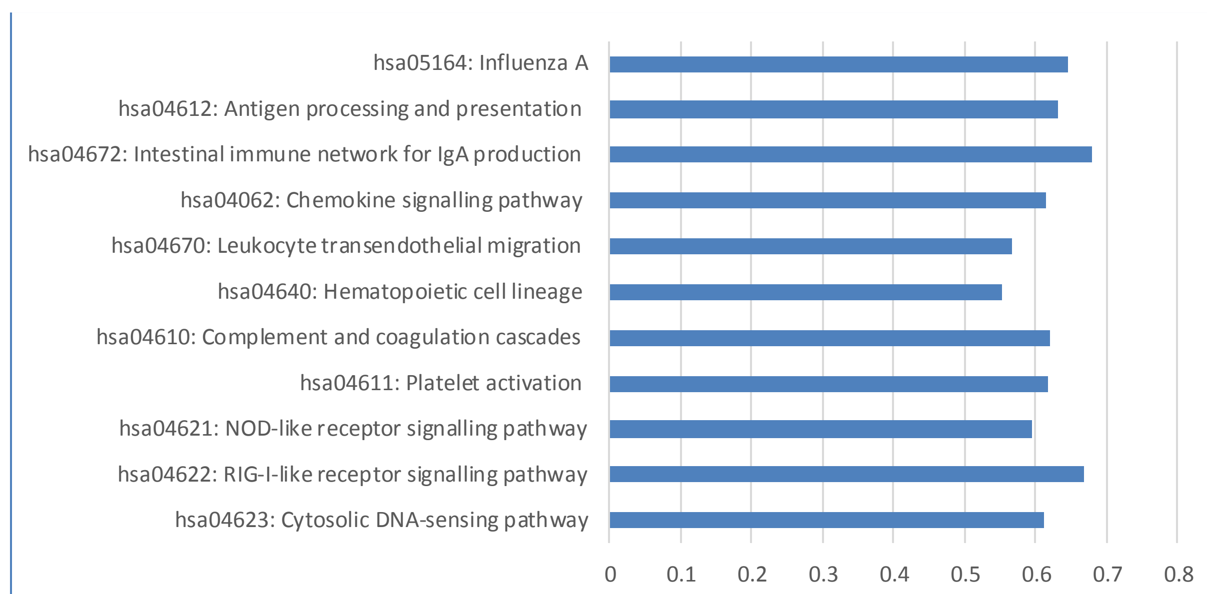

Figure30: Enriched immune system KEGG pathways in co-localization network

NETWORK PROPERTIES AND SIGNIFICANT KEGG PATHWAYS FOR GENETIC INTERACTION NETWORK

NETWORK PROPERTIES

PROPERTIES	GENETIC INTERACTION
Clustering Coefficient	0.08
Connected components	1
Network diameter	6
Network radius	1
Shortest paths	1951 (20%)
Characteristic path length	2.317
Average number of neighbours	7.204
Number of nodes	98
Network density	0
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	0
Analysis time (sec)	0.067

Table29: Network Properties of Genetic Interaction network obtained using Network Analyser tool of Cytoscape[18]

SIGNIFICANT PATHWAYS

CELLULAR PROCESSES

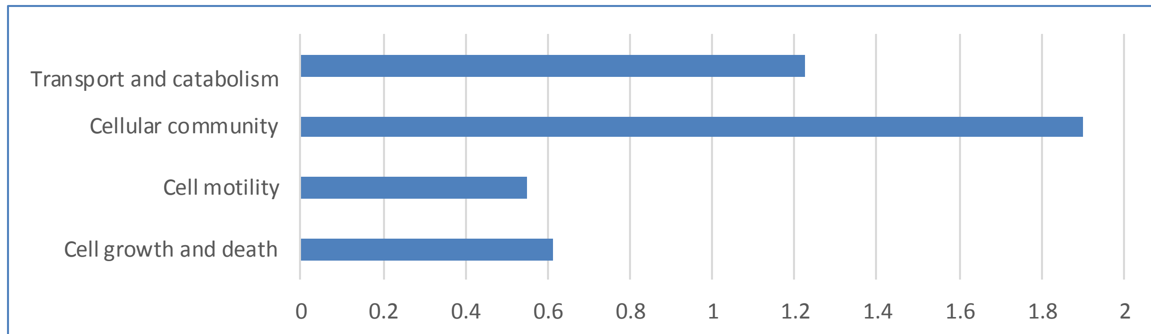

Figure31: Categories of enriched cellular processes KEGG pathways in genetic interaction network

CELLULAR COMMUNITY

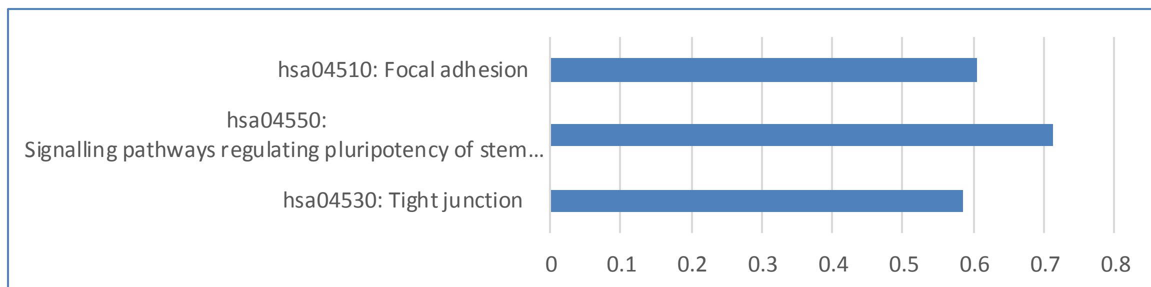

Figure32: Enriched cellular community KEGG pathways in genetic interaction network

TRANSPORT AND CATABOLISM

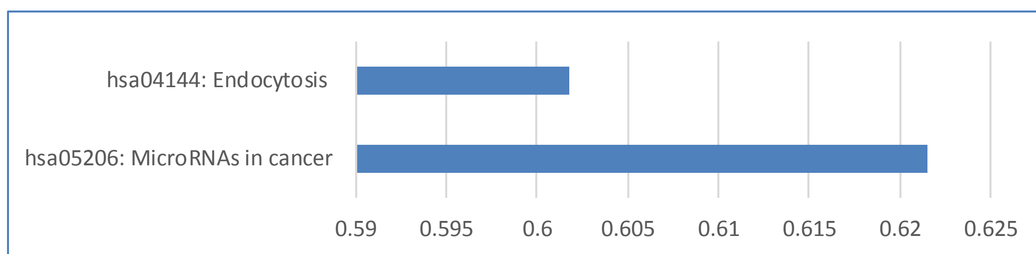

Figure33: Enriched transport and catabolism KEGG pathways in genetic interaction network

SIGNAL TRANSDUCTION

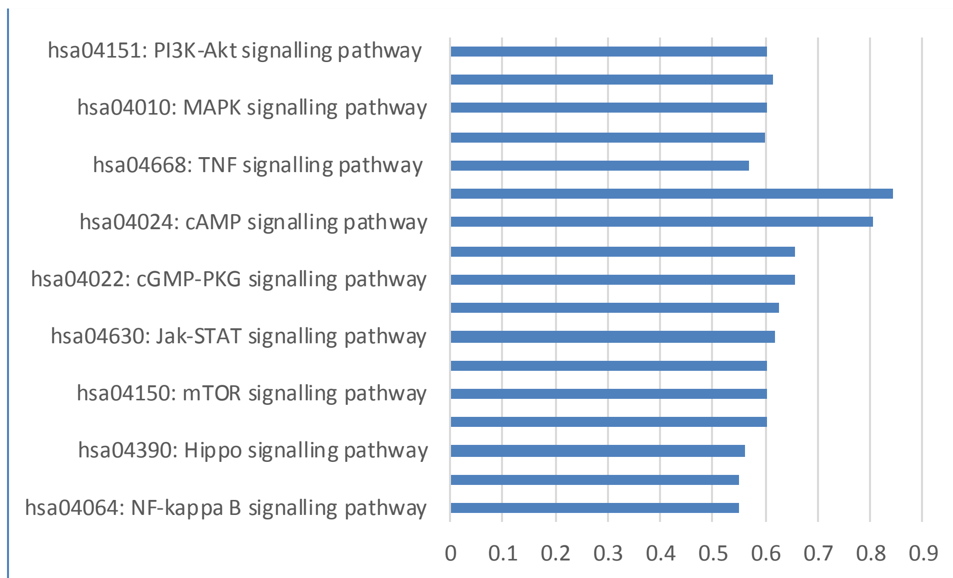

Figure34: Enriched signal transduction KEGG pathways in genetic interaction network

SIGNALLING MOLECULES AND INTERACTIONS

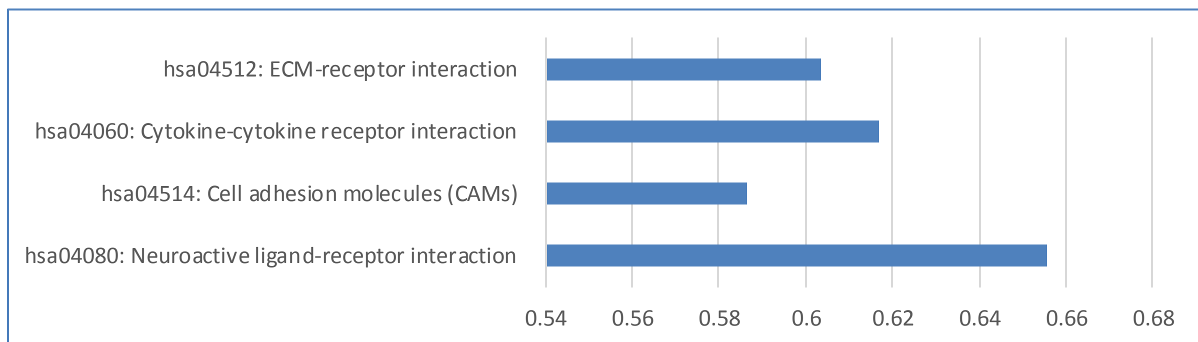

Figure35: Enriched signalling molecules and interactions KEGG pathways in genetic interaction network

CANCERS

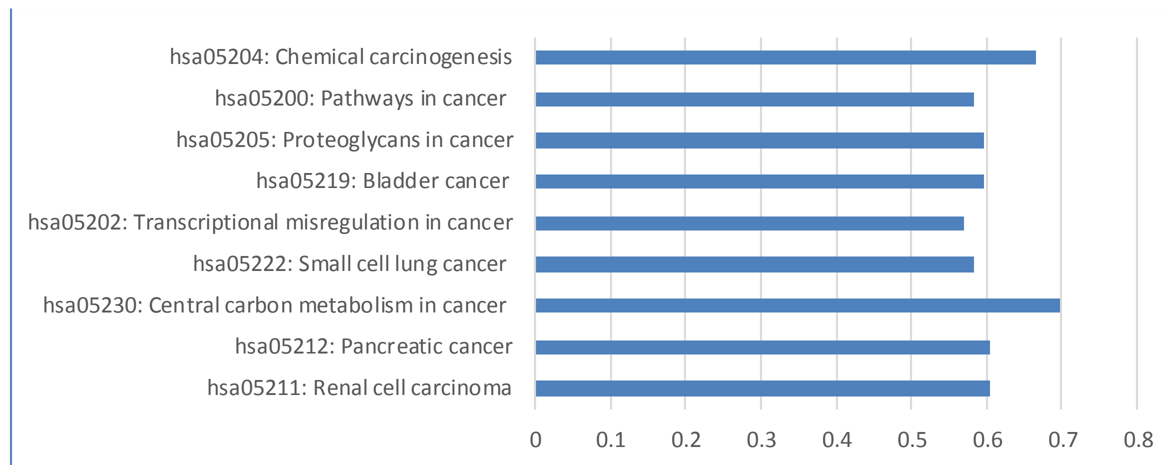

Figure36: Enriched cancers KEGG pathways in genetic interaction network

NEURODEGENERATIVE DISEASES

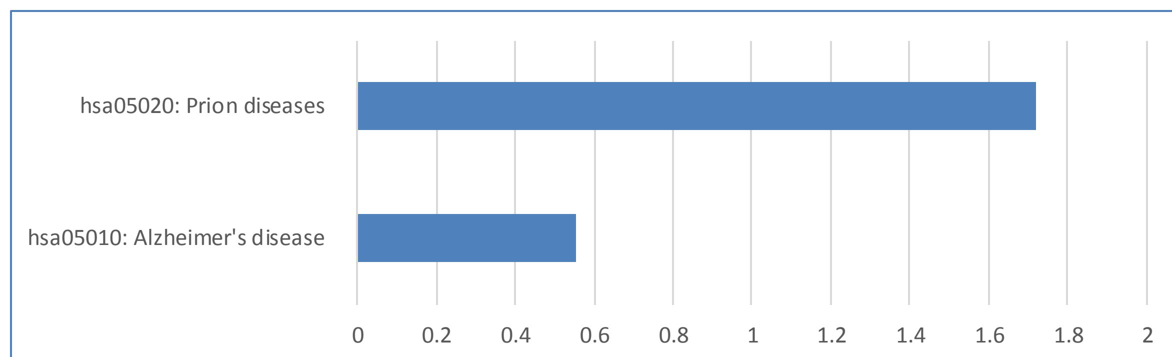

Figure37: Enriched neurodegenerative diseases KEGG pathways in genetic interaction network

METABOLISM

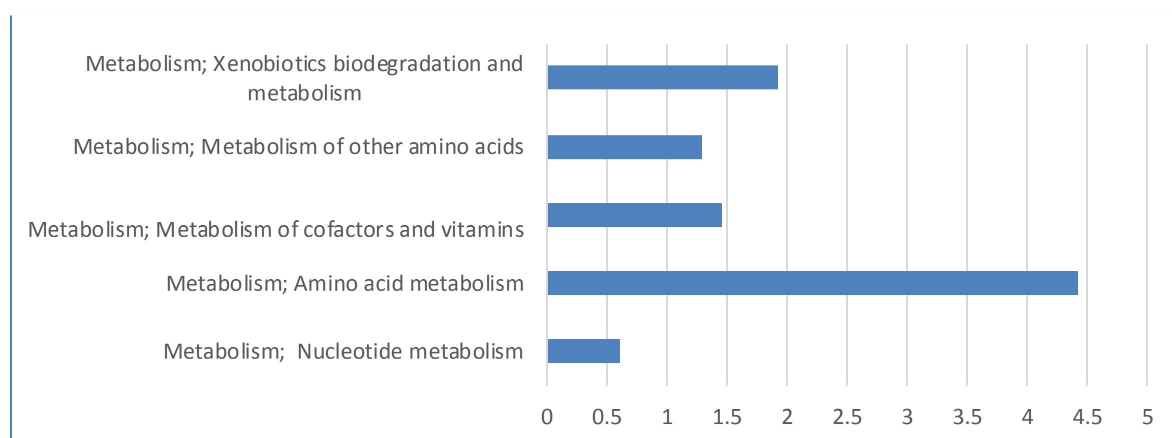

Figure38: Categories of enriched metabolism KEGG pathways in genetic interaction network

AMINO ACID METABOLISM

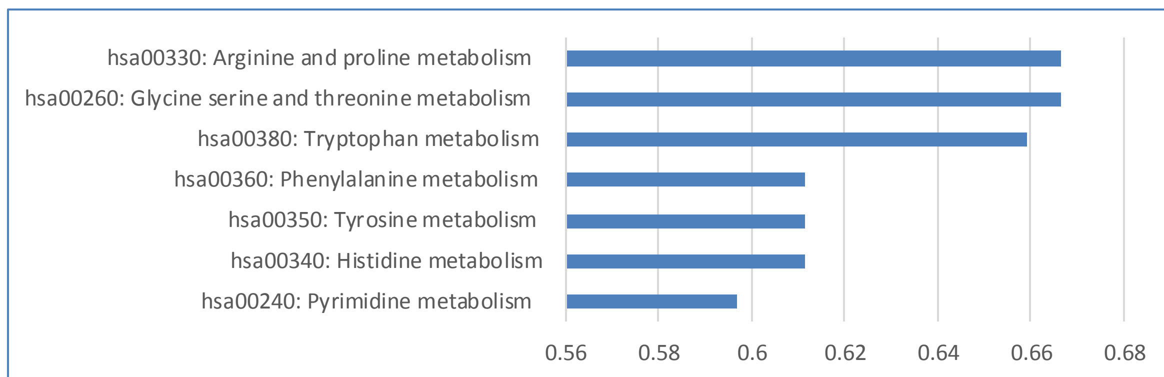

Figure39: Enriched amino acid metabolism KEGG pathways in genetic interaction network

METABOLISM OF COFACTORS AND VITAMINS

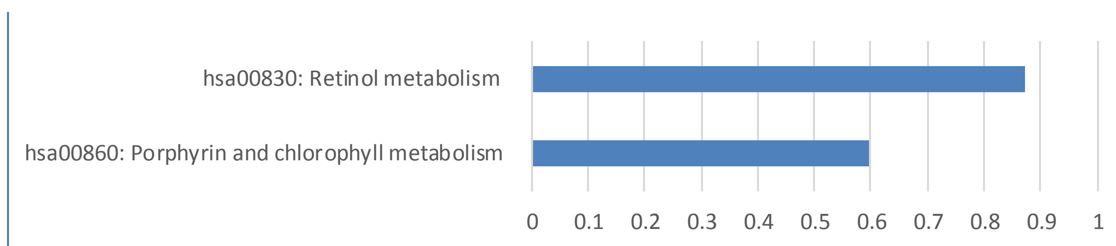

Figure40: Enriched metabolism of cofactors and vitamins KEGG pathways in genetic interaction network

METABOLISM OF OTHER AMINO ACIDS

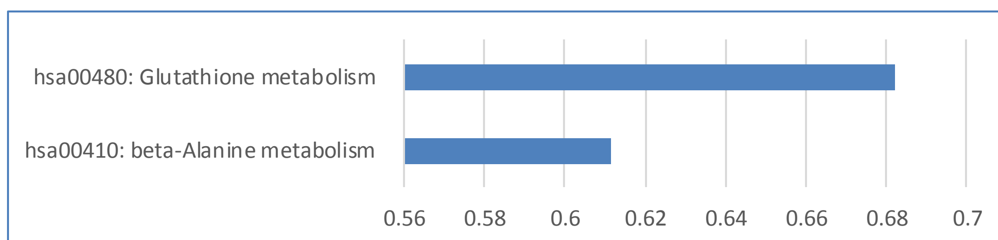

Figure41: Enriched metabolism of other amino acids KEGG pathways in genetic interaction network

ENDOCRINE SYSTEM

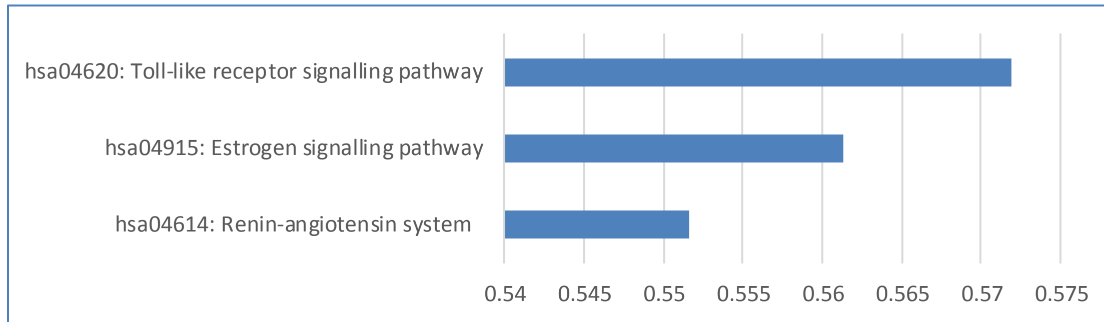

Figure42: Enriched endocrine system KEGG pathways in genetic interaction network

IMMUNE SYSTEM

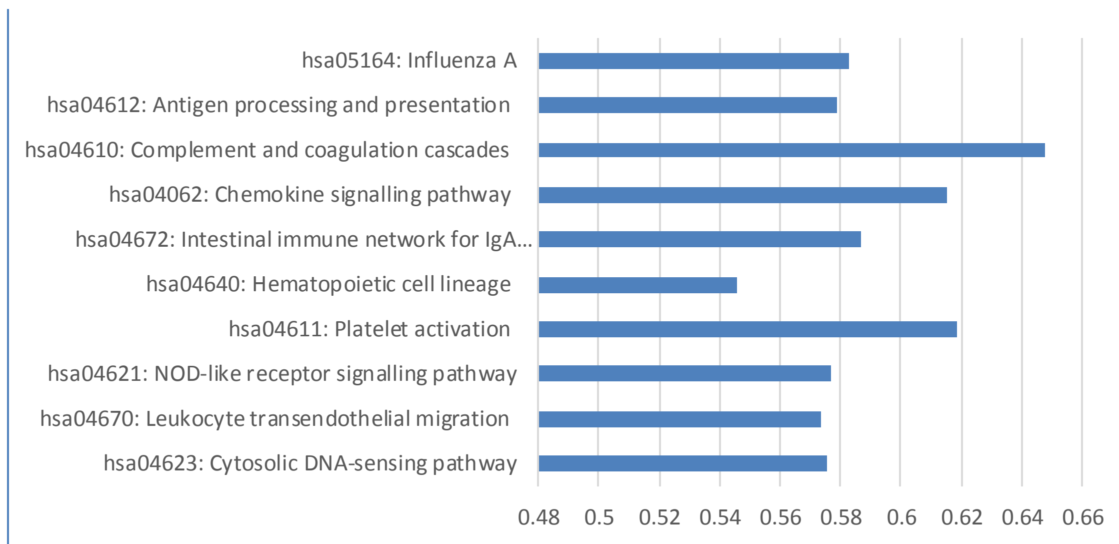

Figure43: Enriched immune system KEGG pathways in genetic interaction network

NETWORK PROPERTIES AND SIGNIFICANT KEGG PATHWAYS FOR PATHWAY NETWORK

NETWORK PROPERTIES

PROPERTIES	PATHWAY
Clustering Coefficient	0.141
Connected components	5
Network diameter	2
Network radius	1
Shortest paths	30 (4%)
Characteristic path length	1.133
Average number of neighbours	2
Number of nodes	26
Network density	0
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	1
Analysis time (sec)	0.035

Table30: Network Properties of Pathway network obtained using Network Analyser tool of Cytoscape [18]

SIGNIFICANT PATHWAYS

CELLULAR PROCESSES

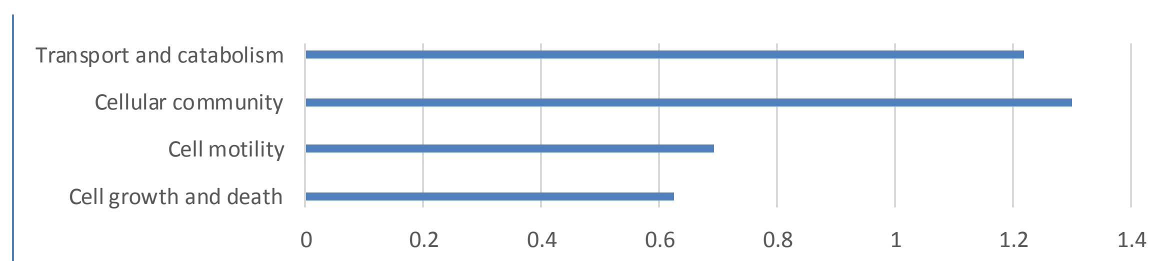

Figure44: Categories of enriched cellular processes KEGG pathways in pathway network

CELLULAR COMMUNITY

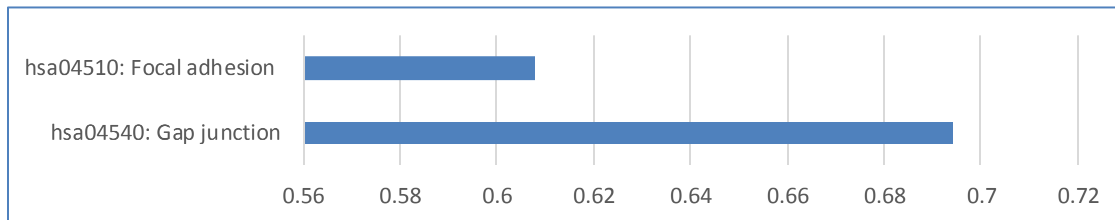

Figure45: Enriched cellular community KEGG pathways in pathway network

TRANSPORT AND CATABOLISM

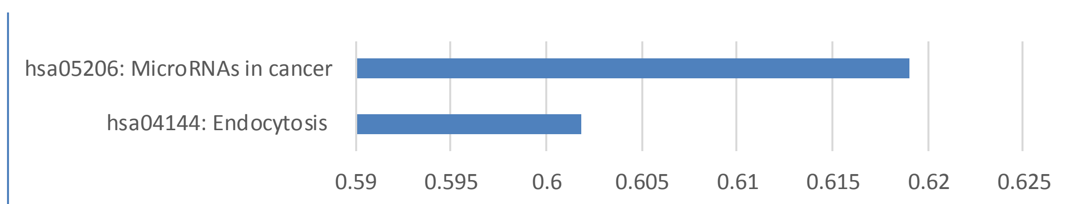

Figure46: Enriched transport and catabolism KEGG pathways in pathway network

SIGNAL TRANSDUCTION

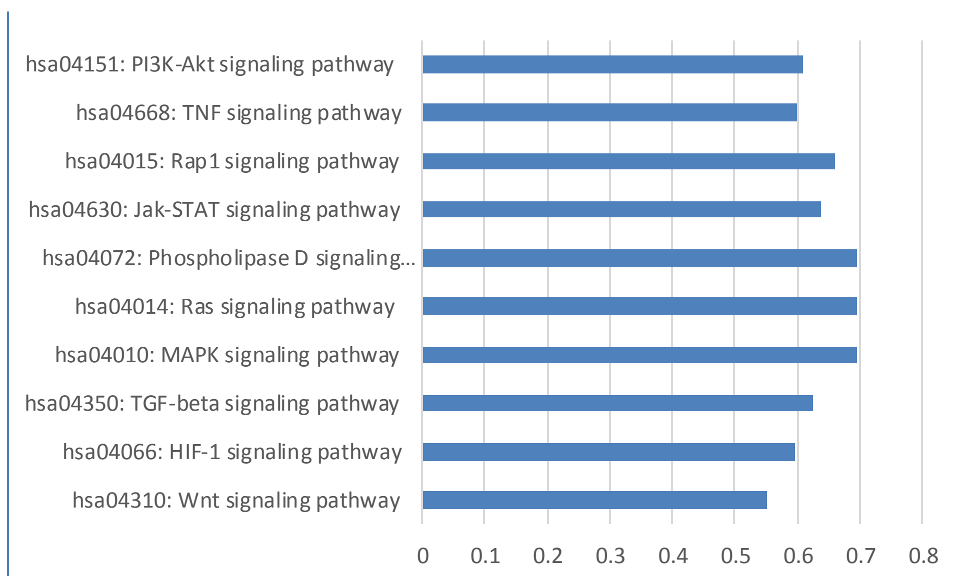

Figure47: Enriched signal transduction KEGG pathways in pathway network

SIGNALLING MOLECULES AND INTERACTIONS

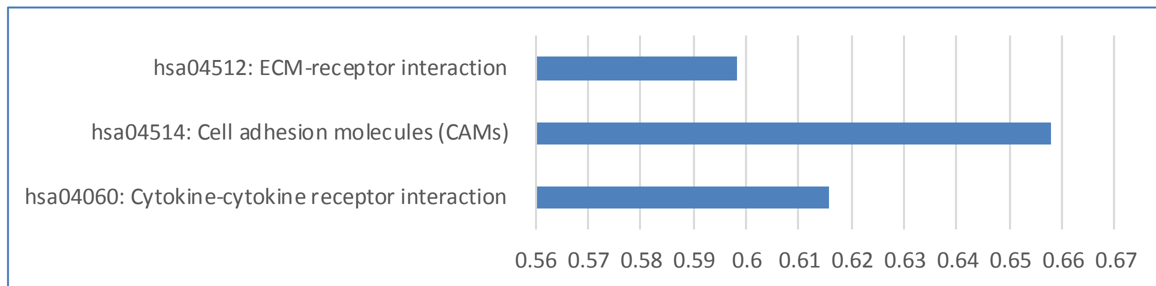

Figure48: Enriched signalling molecules and interactions KEGG pathways in pathway network

CANCERS

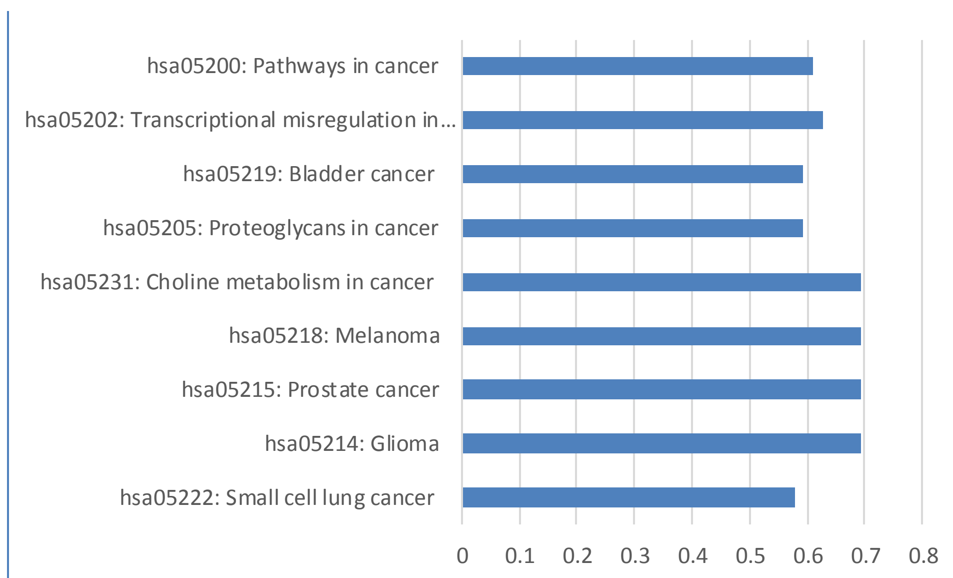

Figure49: Enriched cancers KEGG pathways in pathway network

ENDOCRINE SYSTEM

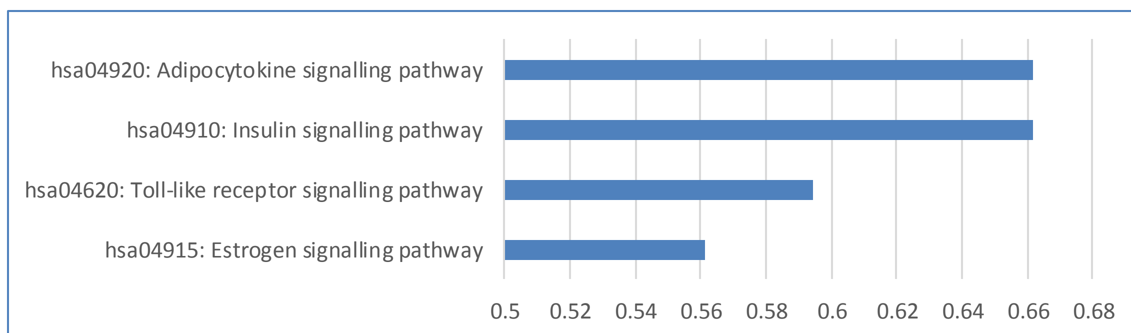

Figure50: Enriched endocrine system KEGG pathways in pathway network

IMMUNE SYSTEM

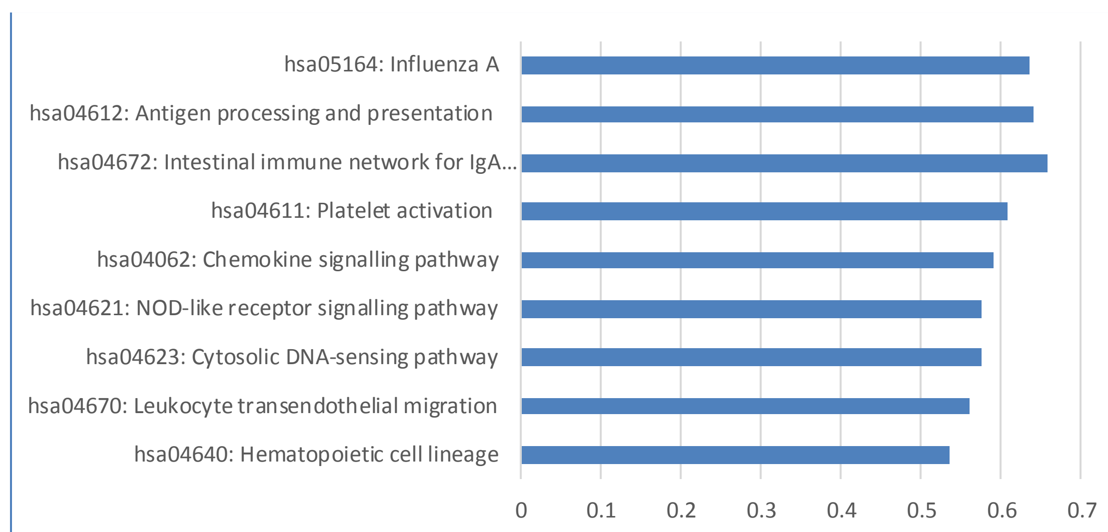

Figure51: Enriched immune system KEGG pathways in pathway network

NETWORK PROPERTIES AND SIGNIFICANT KEGG PATHWAYS FOR PHYSICAL INTERACTION NETWORK

NETWORK PROPERTIES

PROPERTIES	PHYSICAL INTERACTION
Clustering Coefficient	0.024
Connected components	8
Network diameter	4
Network radius	1
Shortest paths	87 (6%)
Characteristic path length	1.77
Average number of neighbours	2.222
Number of nodes	36
Network density	0
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	35
Analysis time (sec)	0.301

Table31: Network Properties of Physical Interaction network obtained using Network Analyser tool of Cytoscape [18]

SIGNIFICANT PATHWAYS

CELLULAR PROCESSES

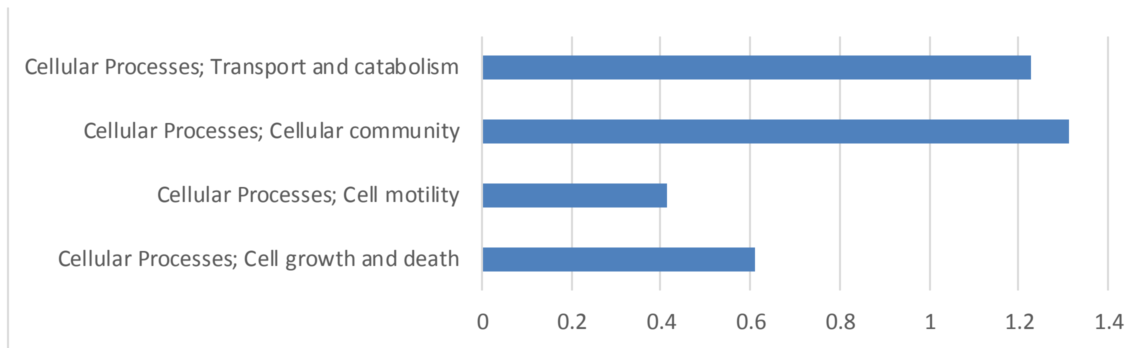

Figure52: Categories of enriched cellular processes KEGG pathways in physical interaction network

CELLULAR COMMUNITY

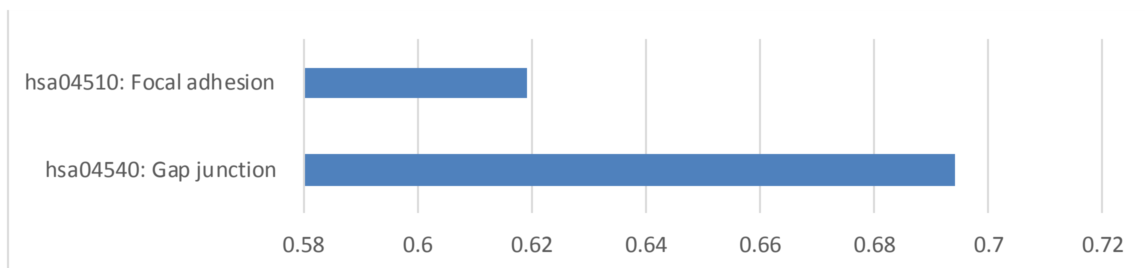

Figure53: Enriched cellular community KEGG pathways in physical interaction network

TRANSPORT AND CATABOLISM

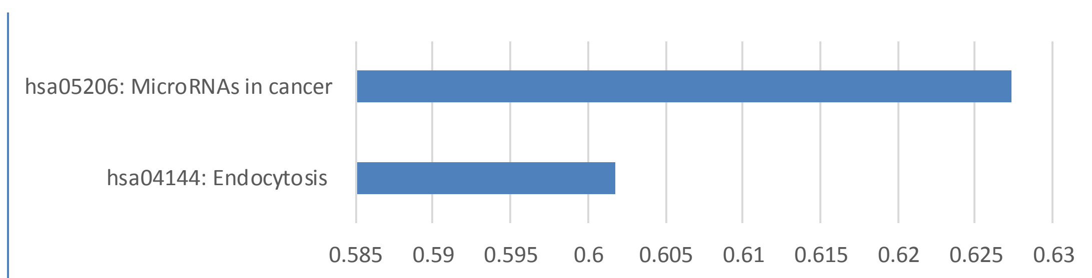

Figure54: Enriched transport and catabolism KEGG pathways in physical interaction network

SIGNAL TRANSDUCTION

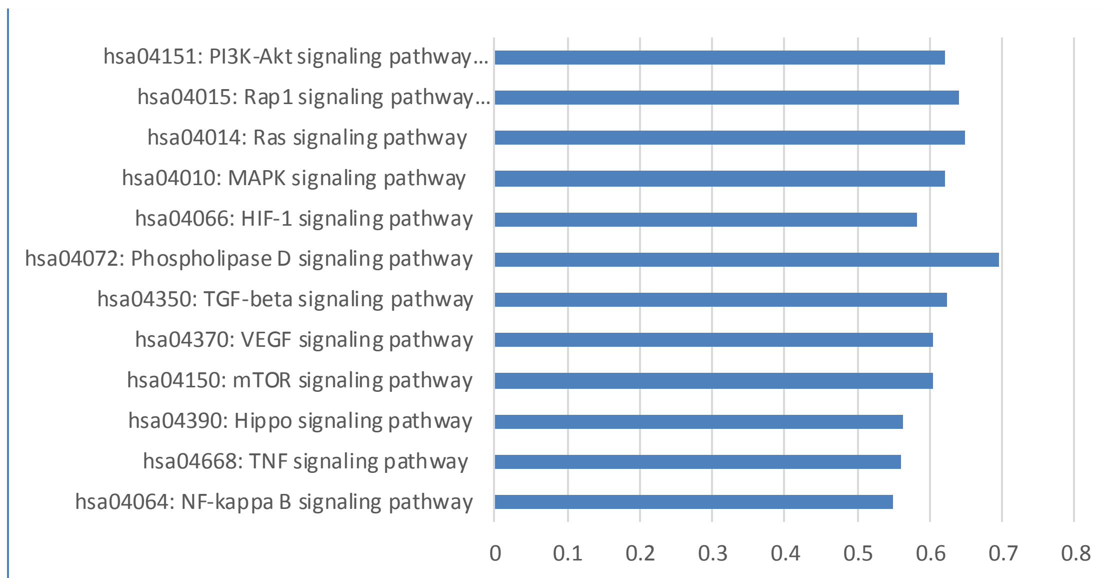

Figure55: Enriched signal transduction KEGG pathways in physical interaction network

SIGNALLING MOLECULES AND INTERACTIONS

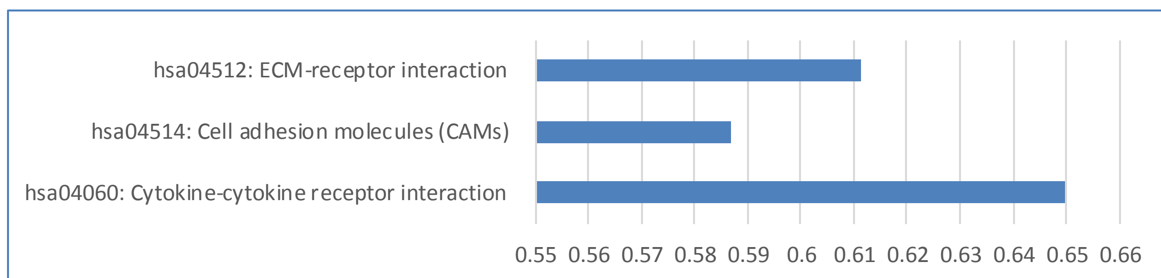

Figure56: Enriched signalling molecules and interactions KEGG pathways in physical interaction network

CANCERS

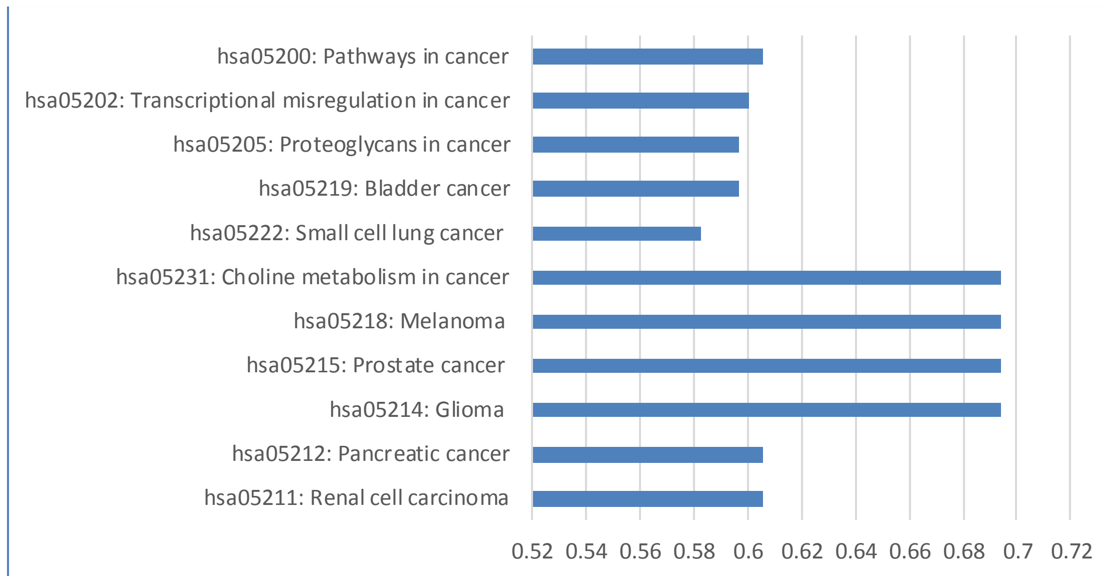

Figure57: Enriched cancers KEGG pathways in physical interaction network

ENDOCRINE SYSTEM

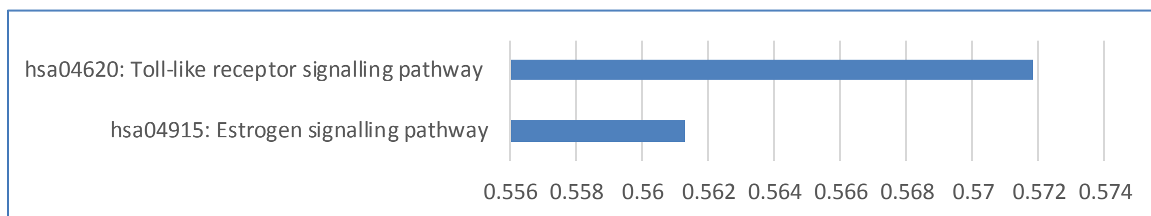

Figure58: Enriched endocrine system KEGG pathways in physical interaction network

IMMUNE SYSTEM

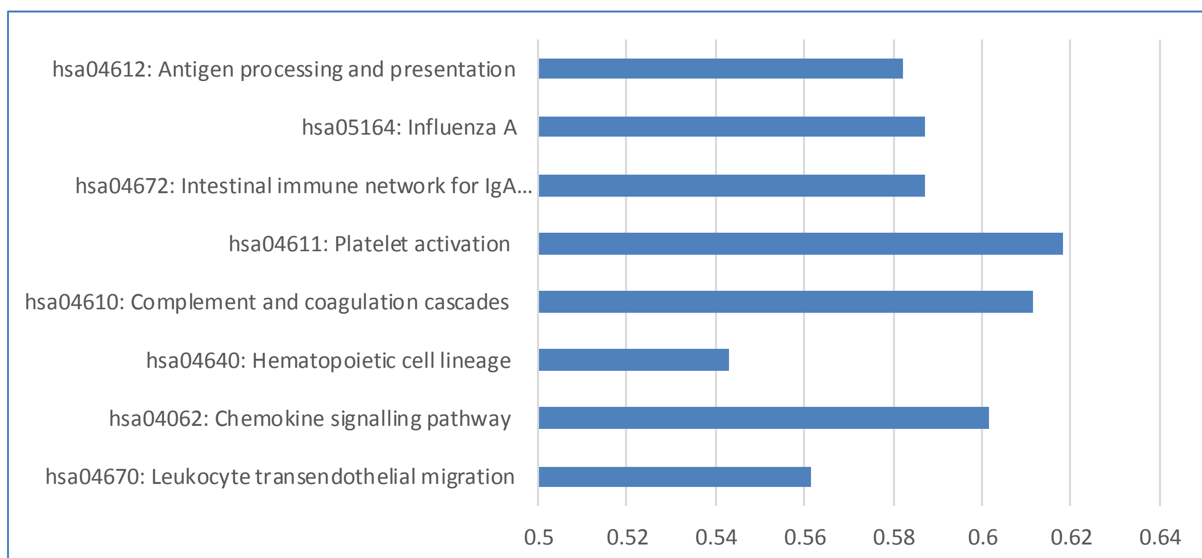

Figure59: Enriched immune system KEGG pathways in physical interaction network

NETWORK PROPERTIES AND SIGNIFICANT KEGG PATHWAYS FOR SHARED PROTEIN DOMAINS

NETWORK PROPERTIES

PROPERTIES	SHARED PROTEIN DOMAINS
Clustering Coefficient	0.236
Connected components	18
Network diameter	3
Network radius	1
Shortest paths	81 (2%)
Characteristic path length	1.16
Average number of neighbours	2.1875
Number of nodes	64
Network density	0
Isolated nodes	0
Number of self-loops	0
Multi-edge node pairs	0
Analysis time (sec)	0.584

Table32: Network Properties of Shared Protein Domains network obtained using Network Analyser tool of Cytoscape [18]

SIGNIFICANT PATHWAYS

CELLULAR PROCESSES

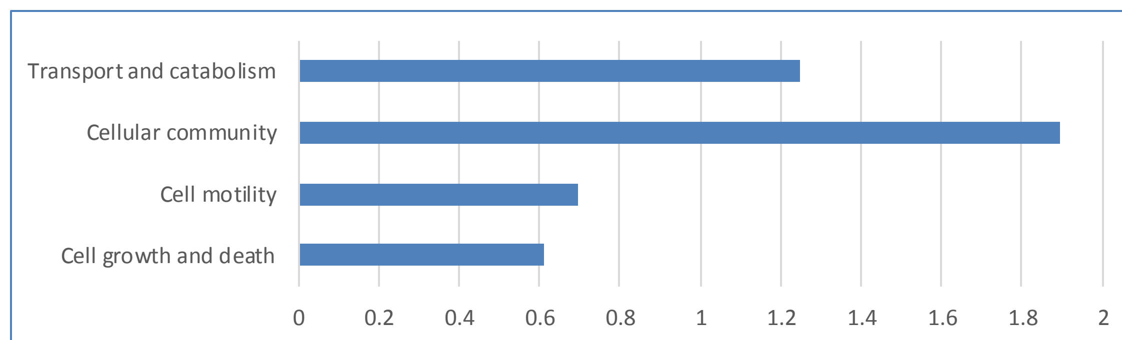

Figure60: Categories of enriched cellular processes KEGG pathways in shared protein domains network

CELLULAR COMMUNITY

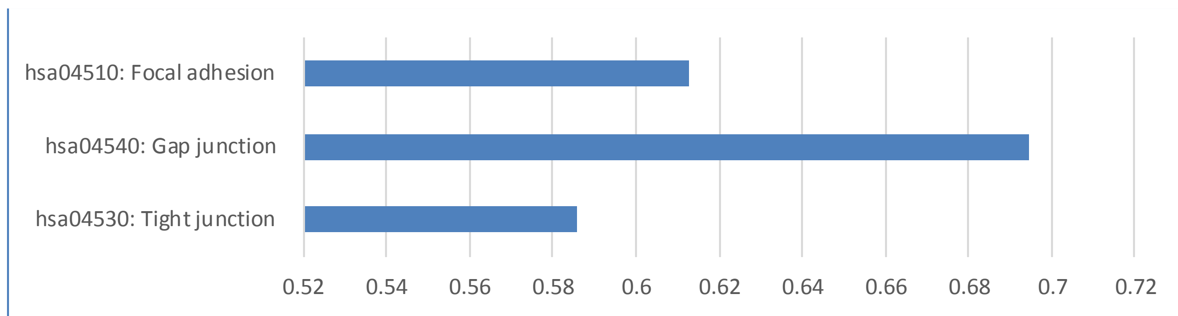

Figure61: Enriched cellular community KEGG pathways in shared protein domains network

TRANSPORT AND CATABOLISM

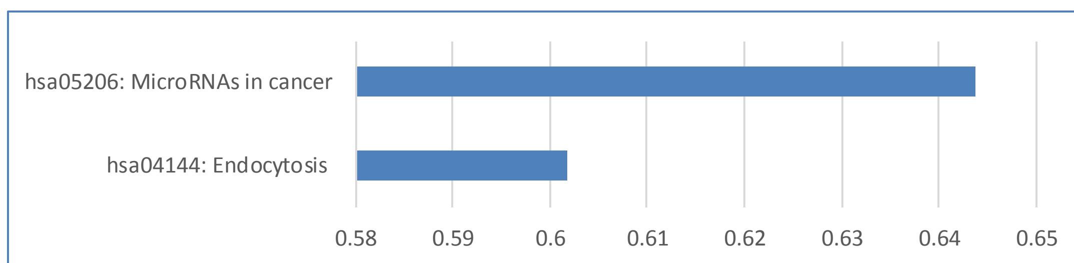

Figure62: Enriched transport and catabolism KEGG pathways in shared protein domains network

SIGNAL TRANSDUCTION

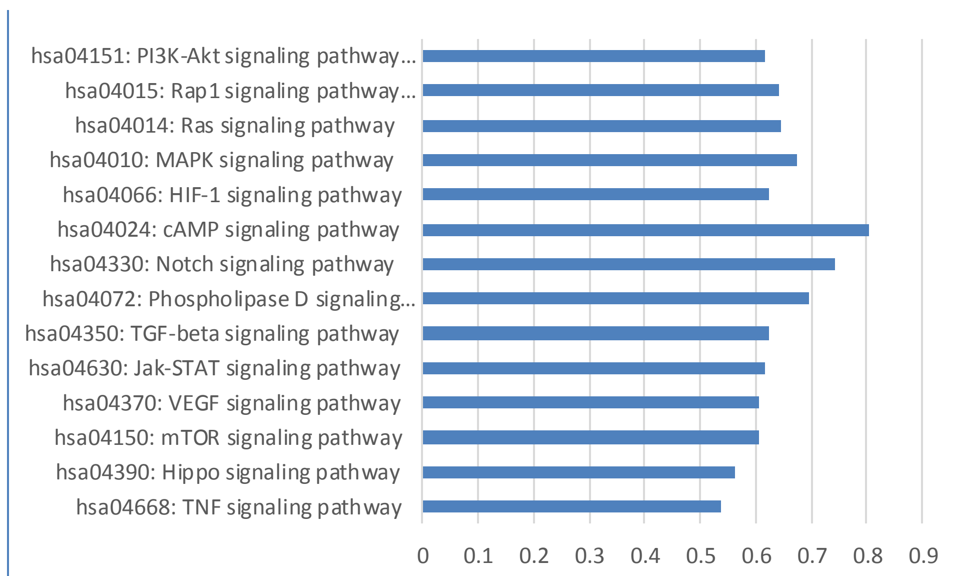

Figure63: Enriched signal transduction KEGG pathways in shared protein domains network

SIGNALLING MOLECULES AND INTERACTIONS

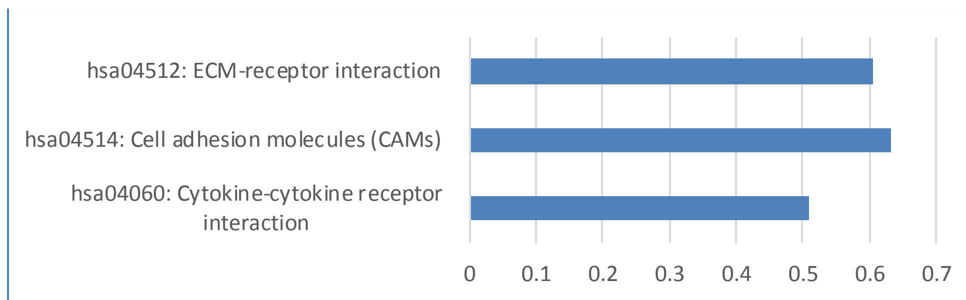

Figure64: Enriched signalling molecules and interactions KEGG pathways in shared protein domains network

CANCERS

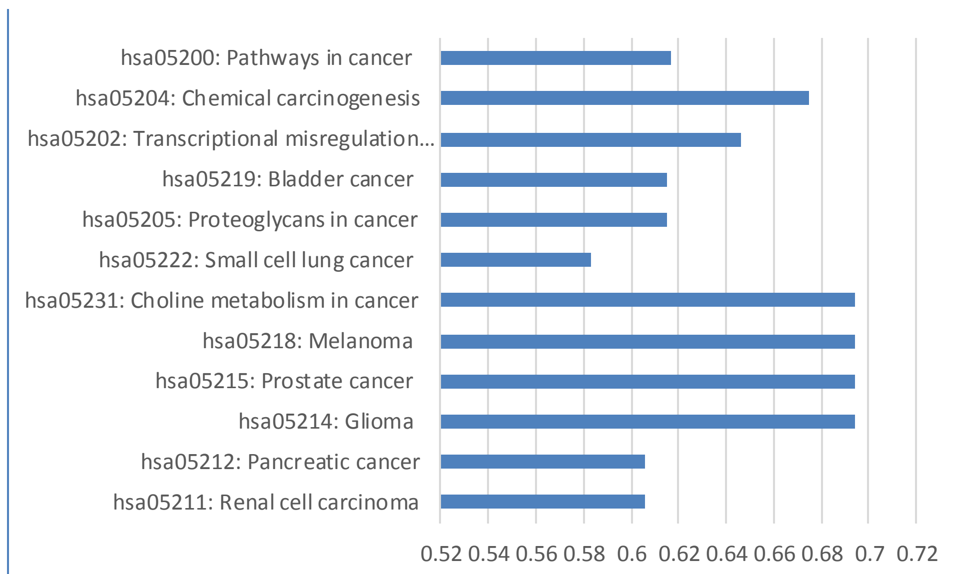

Figure65: Enriched cancers KEGG pathways in shared protein domains network

METABOLISM

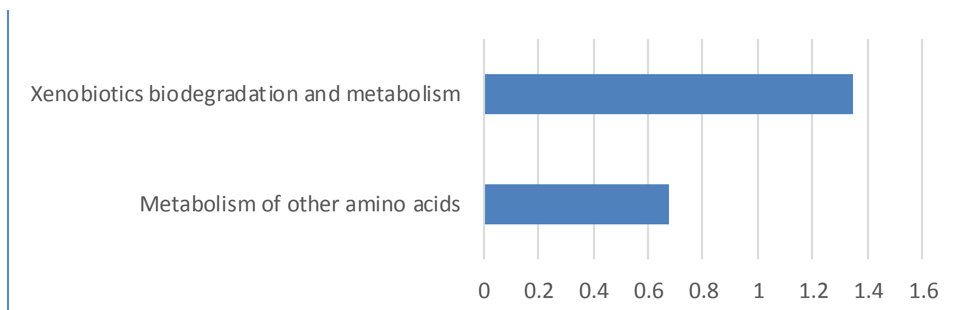

Figure66: Categories of enriched metabolism KEGG pathways in shared protein domain network

IMMUNE SYSTEM

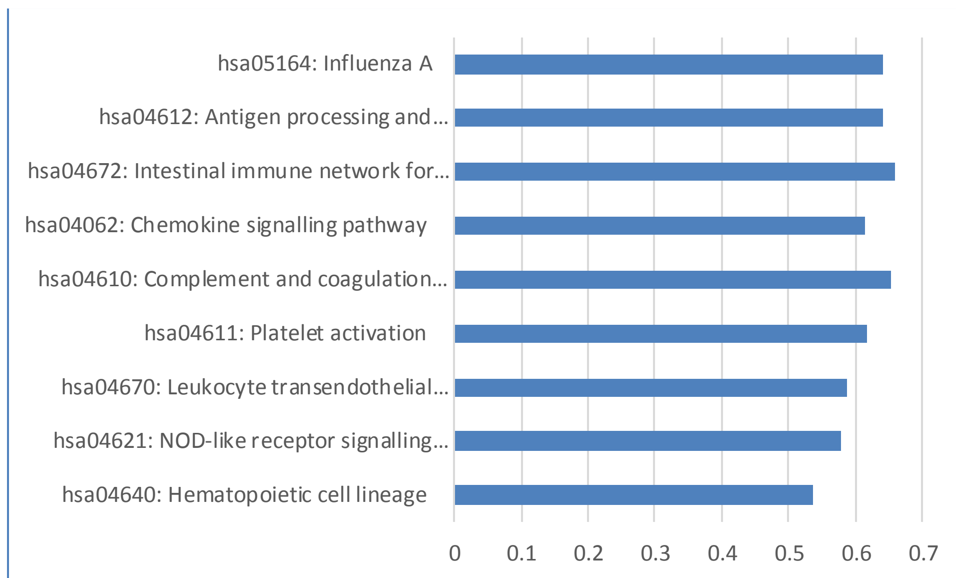

Figure67: Enriched immune system KEGG pathways in shared protein domains network

CHAPTER 5 – CONCLUSION AND FUTURE PROSPECTS

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 generalizability and increased statistical power. On creating different types of networks of significantly up-regulated genes, various pathways that are possibly enriched in Brain and CNS cancer have be obtained. The significantly up-regulated genes and pathways can be verified experimentally and then effective treatment against this ailment can be designed accordingly.

CHAPTER-6 REFERENCES

1. Ramasamy A, Mondry A, Holmes CC, Altman DG. "Key issues in conducting a meta-analysis of gene expression microarray datasets." PLoS Med 5: e184,2008.
2. Alvis Brazma, Helen Causton, John Quackenbush. "Microarray gene expression data analysis"
3. http://www.cancer.gov/about-cancer/what-is-cancer
4. http://www.who.int/mediacentre/factsheets/fs297/en/
5. http://www.cancer.gov/types/brain
6. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours
7. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al."ONCOMINE: a cancer microarray database and integrated data-mining platform." Neoplasia 6: 1–6,2004.
8. Tusher, V. G., Tibshirani, R., & Chu, G. "Significance analysis of microarrays applied to the ionizing radiation response. "Proceedings of the National Academy of Sciences, 98(9), 5116-5121,2001.
9. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. "TM4: a free, open-source system for microarray data management and analysis. Biotechniques.";34(2):374-8,Feb 2003.
10. Goonesekere, N. C., Wang, X., Ludwig, L., & Guda, C. "A meta analysis of pancreatic microarray datasets yields new targets as cancer genes and biomarkers." ,PloS one, 9(4), e93046,2014.
11. Laing E, Smith CP "RankProdIt: A web-interactive Rank Products analysis tool. "BMC Res Notes 3: 221,2010.
12. Huang, D. W., Sherman, B. T., & Lempicki, R. A." Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists."Nucleic acids research, 37(1), 1-13,2009.

13. Huang, D. W., Sherman, B. T., & Lempicki, R. A. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.” Nature protocols, 4(1), 44-57,2009.
14. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P.& Maitland, A.”The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. “Nucleic acids research, 38(suppl 2), W214-W220,2010.
15. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, et al. “RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. “Bioinformatics 22: 2825–2827,2006.
16. Breitling R, Armengaud P, Amtmann A, Herzyk P “Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.” FEBS Lett 573: 83–92,2004.
17. Dudley JT, Tibshirani R, Deshpande T, Butte AJ “Disease signatures are robust across tissues and experiments.” Mol Syst Biol 5: 307,2009.
18. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T.“Cytoscape: a software environment for integrated models of biomolecular interaction networks.”Genome research, 13(11), 2498-2504,2003.