Computational Analysis of Gene Expression and Pathway Data for Colorectal Cancer

A

PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

BIOINFORMATICS ENGINEERING

Under the supervision

of

Dr. Tiratha Raj Singh

(Associate Professor)

by

Shorya Singh Thakur (151503)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN – 173234

HIMACHAL PRADESH, INDIA

May – 2019

STUDENT'S DECLARATION

I hereby declare that the work presented in the Project report entitled **"Computational Analysis of Gene Expression and Pathway Data for Colorectal Cancer"** submitted for partial fulfilment of the requirements for the degree of Bachelor of Technology in Bioinformatics Engineering at **Jaypee University of Information Technology, Waknaghat** is an authentic record of my work carried out under the supervision of **Dr. Tiratha Raj Singh.** This work has not been submitted elsewhere for the reward of any other degree/diploma. I am fully responsible for the contents of my project report**.**

Shorya Singh Thakur
151503

Department of Biotechnology and Bioinformatics Engineering
Jaypee University of Information Technology, Waknaghat, India
Date:

CERTIFICATE

This is to certify that the work which is being presented in the project report titled **“Computational Analysis of Gene Expression and Pathway Data for Colorectal Cancer”** in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Bioinformatics Engineering submitted to the Department of Biotechnology and Bioinformatics Engineering, **Jaypee University of Information Technology, Waknaghat** is an authentic record of work carried out by **Shorya Singh Thakur (151503)** during a period from August 2018 to May 2019 under the supervision of **Dr. Tiratha Raj Singh** .

The above statement is correct to the best of my knowledge.

Date: …………………

Signature of Supervisor	Signature of HOD	Signature of External
Dr. Tiratha Raj Singh	Dr. Sudhir Kumar	Examiner
(Associate Professor)	(Professor and Head)	
Department of Biotechnology	Department of	
And Bioinformatics	Biotechnology and	
Engineering	Bioinformatics Engineering	
JUIT, Waknaghat	JUIT, Waknaghat	

ACKNOWLEDGEMENT

I take immense pleasure to express my gratitude to **Dr. Tiratha Raj Singh** (Associate Professor, Department of Biotechnology and Bioinformatics Engineering), my project guide, for the guidance and constructive criticism which was the driving force of the project. He has provided valuable ideas and support during the course of this work. This work would not have been possible without his support. I would also like to thank a few more people who have helped in every likely way for the effective conclusion of the work. I am indebted to Mr. Rohit Shukla for the guidance and moral support and great cooperation during various stages of the work. I would like to thank all other members of the faculty of Department of Biotechnology and Bioinformatics of Jaypee University of Information Technology for providing valuable input through the course of this work. Last but not the least my sincere gratitude to everyone who assisted me during my project.

ABSTRACT

Colorectal disease is one of the most widely known harmful diseases around the globe, yet the included flagging pathways and driven-qualities are to a great extent unknown. Incorporation of 4 associate datasets to clarify the impending key applicant qualities and paths in colorectal cancer was done. Articulation profiles GSE27000, GSE22915, GSE45066, GSE75420, were selected. The number of DEGs shared were 292 (127 down-regulated, 165 up-regulated). They were recognized within the 4 datasets of GSE. Following that, 170 DEGs/hubs were recognized from DEGs protein-protein interaction arrange compound. Finally, 2 utmost important modules were screened from PPI, 28 focal hub genes were recognized and the greater part, where the relating genetic factor are linked with chemokines and cell cycle process. Taken above, utilizing incorporated bioinformatical investigation, DEGs applicant pathways and genetic factor in colon cancer were identified, which might enhance our perspective of the reason and fundamental sub-atomic trials. Thus, these applicant pathways and genes could be beneficial focuses for colorectal cancer.

Keywords: Differentially expressed genes, Colorectal Cancer, Pathways

TABLE OF CONTENTS

ABBREVATIONS

CRC………………..Colorectal cancer

PPI………………….Protein–protein interaction network

DEG………………..Differentially expressed genes

FAP………………...Familial Adenomatous Polyposis

GPCR………………G-protein couple receptor

NCBI……………….National Center for Biotechnology Information

TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Colorectal malignant growth starts in the rectum. Malignancy of the rectum and colon disease are often collected on the reasons that they share various climaxes for all intents and purpose. Usually these cancers initiate with a growth on the inner lining of the rectum. They are called polyps. On the off chance that malignant growth frames in a polyp, it can develop into the mass of the rectum or colon after some time. Mass of the rectum and colon consist of numerous layers. This malignant growth initiates in the cordial level and can mature apparently through the majority or a few of different layers. While divider contains malignancy cells, they would therefore be able to mature into veins. The degree of spread of this malignant growth relies upon how intensely it matures into the separator and the probability of spreading outside the colon is high [1]. All around the world, it is the second most ordinarily analysed disease in females and the second in males. 1.7 million cases and nearly 962,101 deaths in 2017 as per the World Health Organization GLOBOCAN repository. Every year, roughly 135,610 new instances of malignancies are analysed, of which 121,620 are colon and the remaining are rectal diseases. Every year, around 50,630 Americans bite the dust of CRC, representing roughly 8 percent of all disease deaths [2]. Most solemn threat factor for colon malignant growth is the growing age. In 89% of cases individuals aging above 38 and 87% in those aging above 58 [3]. A considerably greater rate of colorectal malignancy within progressively wealthy countries diverged and less developed countries is likewise supposed to be recognised with factors like obesity and deployment of prepared meat.

1.1 OBJECTIVE

Analysing microarray data to explore key genes and their functions in progression of colorectal cancer. To analyse gene expression using various bioinformatics tools like Morpheus, DAVID, Expander, etc. to decipher Differentially Expressed Genes (DEGs) using Microarray data profiling. Identifying microarray datasets from GEO database. The process of enrichment analysis with tools to uncover their biological significance. And when the analysis of Gene Expression Data is done, screening is performed. Thus, helping in understanding the key genes, pathways and function modules involved.

Chapter 2

LITERATURE REVIEW

Colorectal malignant growth emerges from an antecedent injury, the adenomatous polyp, which shapes in a field of epithelial cell hyper proliferation. Movement from this forerunner injury to colorectal malignant growth is a multistep procedure, joined by changes in a few silencer qualities that outcome in anomalies of cell guideline, and has a characteristic history of 10– 15 years. In most cases colorectal diseases emerge from dysplastic adenomatous polyps. It is a multistep procedure includes the inactivation of an assortment of genes that stifle tumours and fix DNA and the concurrent actuation of oncogenes.

2.1 EPIDEMIOLOGY

Comprehensively, the provincial occurrence of CRC shifts more than 10-fold. The most noteworthy frequency rates are in Europe, New Zealand, Australia and North America. South-Asia and Africa have the least rates. Topographical differences like these give off an impression of being owing to contrasts in dietary and ecological experiences that are obligatory beginning with a basis of genetically determined susceptibility. For sporadic CRC, age is a considerable danger cause [4]. Huge bowel cancer is exceptional afore 40; the rate begins to increase fundamentally amid 50 and 40.

Current writing proposes that more than 86 percent of patients analysed younger than 50 are symptomatic at finding, and in spite of this, they have a further developed stage at determination and more unsatisfied results. As of now, most rules don't prescribe screening for asymptomatic people younger than 50 except if they have a constructive family ancestry or an inclining acquired syndrome [5]. Be that as it may, in 2018, the American Cancer Society issued a "qualified" proposal to start screening people at normal hazard for CRC at age 45 years.

2.2 RISK FACTORS

Natural and hereditary elements be able to upsurge the probability of emerging colorectal cancer. However susceptibility that are hereditary results in the increments in threat. These issues be able to isolate in an adequately great hazard to adjust suggestions for colon cancer screening, factors that may change screening proposals, and those that don't modify screening suggestions since they are thought to give a little or unsure extent of risk.

2.2.1 Hereditary CRC syndromes

Many inborn conditions, in which maximum are inherited, are related to a massive risk of mounting colorectal cancer. The most shared of the ancestral colon cancer conditions are the FAP and Lynch syndrome. Nearly 10 percent of unselected patients with CRC convey at least one pathogenic transformation, and that the greater part are not Lynch disorder or FAP [6].

2.2.2 Family history of Sporadic CRCs

Family history is likewise a significant hazard factor even outside of the disorders with a characterized hereditary inclination. Having a single influenced first-degree relative (parent or kid) with CRC expands the hazardous risk. Risk is additionally expanded if two first, or one first and at least one first or second-degree relatives on either side of the family have colon disease, or if the case is diagnosed underneath age of 50 years.

2.2.3 Inflammatory Bowel Disease

A sensible gauge of the colon malignant growth frequency is around 0.4 % every year to people with infection length somewhere in the range of 20 to 25 years, at that point 1 % annually from that point [15]. Most reports recommend that the co-event of ulcerative colitis recognizes a subset of patients with a significantly more serious hazard. Others have recognized the nearness of pseudo polyps, especially assuming extensive and complex [7].

2.2.4 Abdominal Radiation

Grown-up overcomers of youth harm who got abdominal radiation are at altogether expanded danger of resulting gastrointestinal neoplasms, the dominant part being CRC. Rules from the Children's Oncology Group suggest colonoscopy at regular intervals for overcomers of youth disease who got 30 Gy or a greater amount of stomach radiation, with screening starting 10 years after radiation or at age 35 years, whichever happens last [14]. A background marked by radiation treatment for prostate disease was related with an expanded danger of rectal malignancy in two substantial databases [8]. The greatness of hazard is around like that seen in patients with a family ancestry of colonic adenomas. Regardless of whether such malignant growths pursue an adenoma to disease grouping, and whether expanded screening in such patients would improve malignancy recognition rates and results, is hazy.

2.3 PROTECTIVE FACTORS

Various factors have been stated by studies that helps with lowering the risk of colorectal cancer. Some of them are everyday exercises, factors including dietary, often use of drugs like aspirin and NSAIDs. And not even single one of them is nowadays used for colorectal cancer screening recommendations.

2.3.1 Physical Activity

Considerable observed data lets us know that daily physical activity, either for work purposes or in the free time, is linked with protection from CRC. In a meta-analysis of 19 studies, there was a considerable 29 percent supressed risk of colorectal cancer as when mirroring the highest versus the lowest active individuals and a 27 percent suppression for distal colon cancer [9]. The mechanism underlying the obvious protective amalgamation of physical activity is unknown and no trails of intervention of physical activity for Colorectal Cancer prevention have been described.

2.3.2 Diet

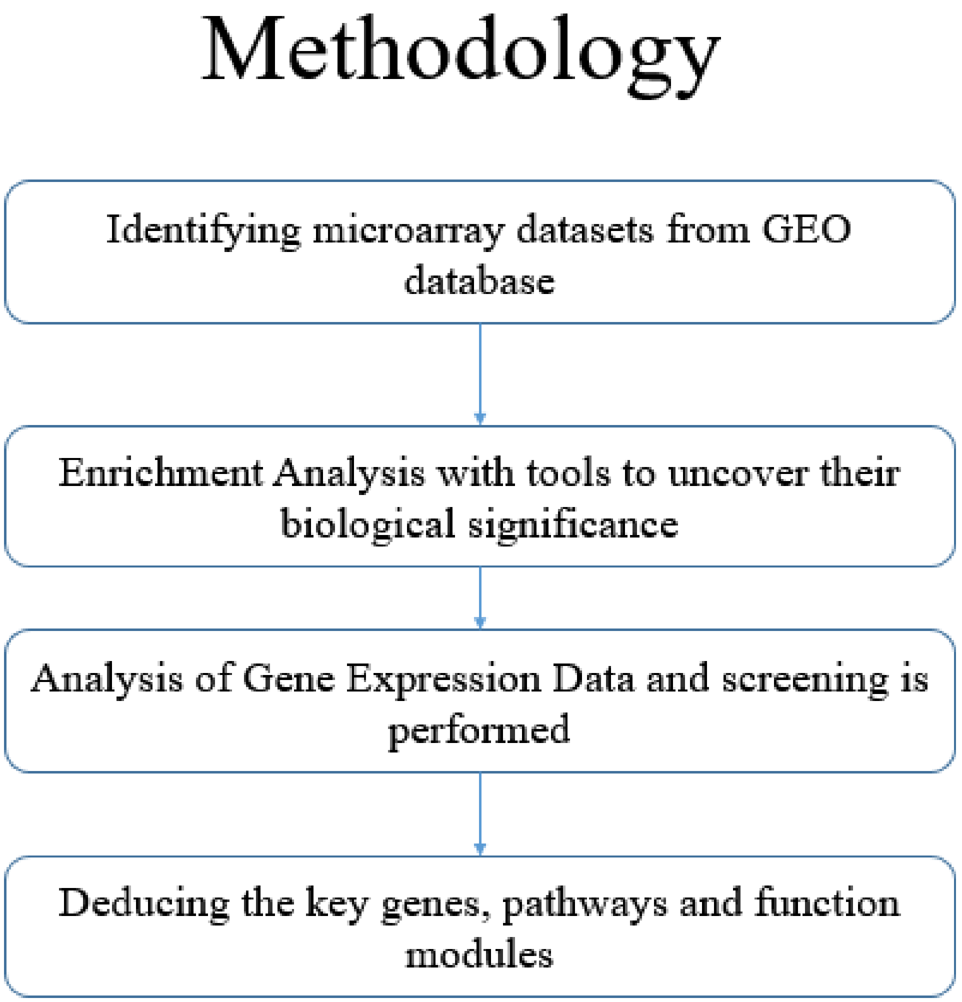
A number of epidemiologic studies have made it clear that an association between the intake of a high fruits and vegetables diet and protection from CRC. The general danger of CRC is roughly 0.4 contrasting gatherings with the most astounding intake with those with the least. In any case, dissonant information have additionally been distributed. The connection between utilization of production and CRC was tested in an imminent partner examine that consolidated subjects from the study of Nurses' Health (77,746 ladies) and Follow-up Study of the Health Professionals' (36,352 men). In that examination, there was no critical relationship between the utilization of organic products, vegetables, or the mix on the occurrence of either colon or rectal disease, autonomous of nutrient enhancement use or smoking propensities [10].

2.3.3 Drugs

A few specialists (most outstandingly nonsteroidal calming operators) have been appeared to have unobtrusive to direct chemo preventive impacts in normal and high-chance populaces [13]. A considerable collection of observational and mediation preliminary proof recommends that headache medicine and other nonsteroidal calming drugs (NSAIDs) ensure against the improvement of colonic adenomas and malignant growth. Customary utilization of headache medicine and different NSAIDs is related with a 20 to 40 percent decrease in the danger of colonic adenomas and CRC in people at normal hazard.

Chapter 3

METHODS AND MATERIALS

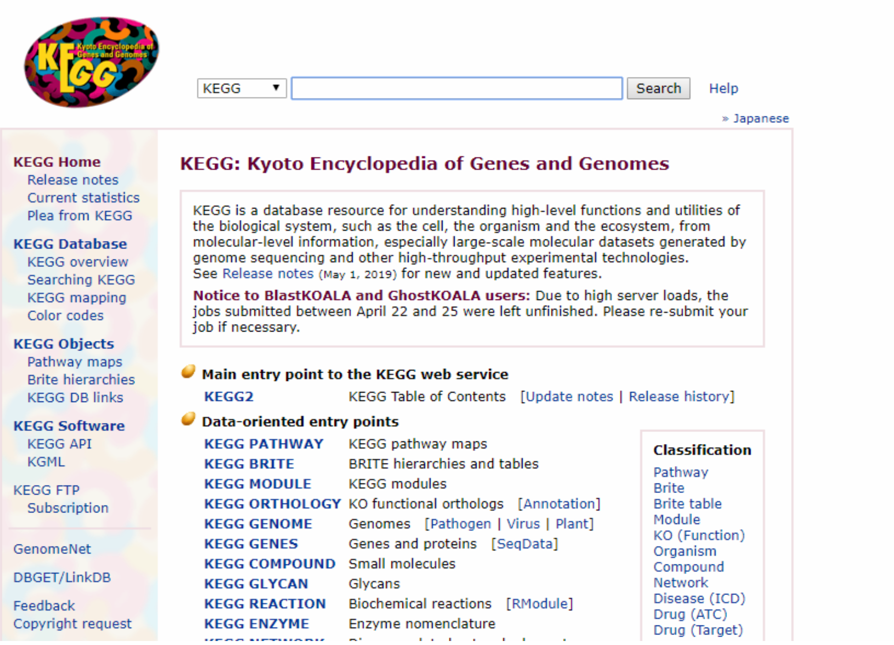

Methodology
Identifying microarray datasets from GEO database
Enrichment Analysis with tools to uncover their biological significance
Analysis of Gene Expression Data and screening is performed
Deducing the key genes, pathways and function modules

3.1 COLLECTION OF DATA

With the help of these repositories, databases and tools process of data collection was performed.

- KEGG PATHWAY

A database asset for seeing abnormal state capacities and utilities of the natural framework is KEGG. For example, the cell, the life form and the environment, from atomic dimension data, particularly extensive scale sub-atomic datasets created by genome sequencing and other high-throughput trial advancements.



KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (May 1, 2019) for new and updated features.

Notice to BlastKOALA and GhostKOALA users: Due to high server loads, the jobs submitted between April 22 and 25 were left unfinished. Please re-submit your job if necessary.

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

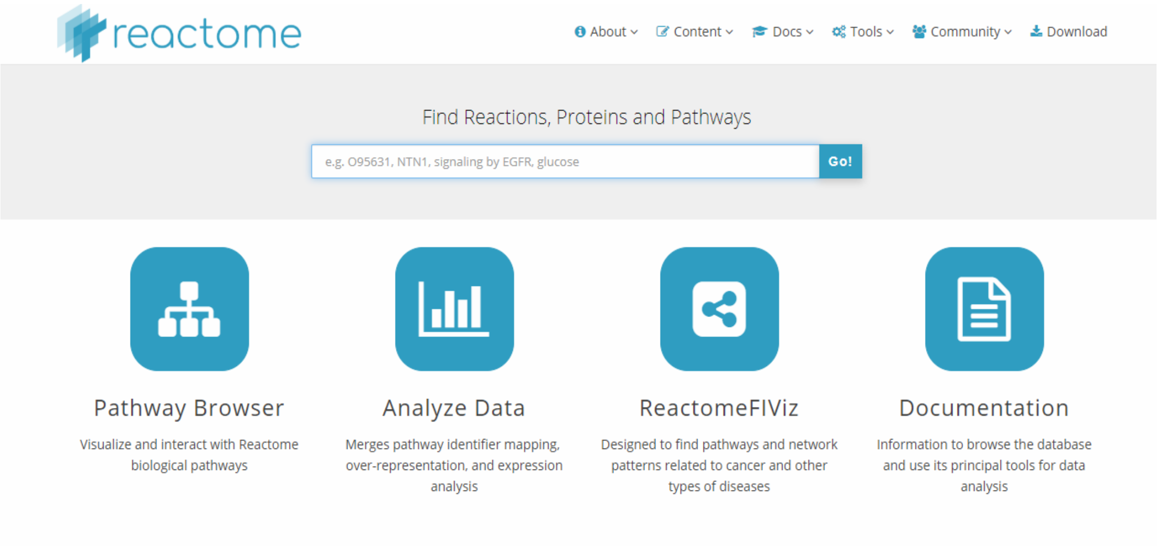
KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs [Annotation]
KEGG GENOME	Genomes [Pathogen | Virus | Plant]
KEGG GENES	Genes and proteins [SeqData]
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature

Classification
Pathway
Brite
Brite table
Module
KO (Function)
Organism
Compound
Network
Disease (ICD)
Drug (ATC)
Drug (Target)

Fig. 1 Kyoto Encylopedia of Genes and Genomes

- Reactome

Reactome is an open-source, free, curated pathway database. The objective is to give instinctive bioinformatics apparatuses to the representation, elucidation and examination of pathway learning to help fundamental research, genome investigation, demonstrating, frameworks science and training.



Find Reactions, Proteins and Pathways

Pathway Browser

Visualize and interact with Reactome biological pathways

Analyze Data

Merges pathway identifier mapping, over-representation, and expression analysis

ReactomeFIViz

Designed to find pathways and network patterns related to cancer and other types of diseases

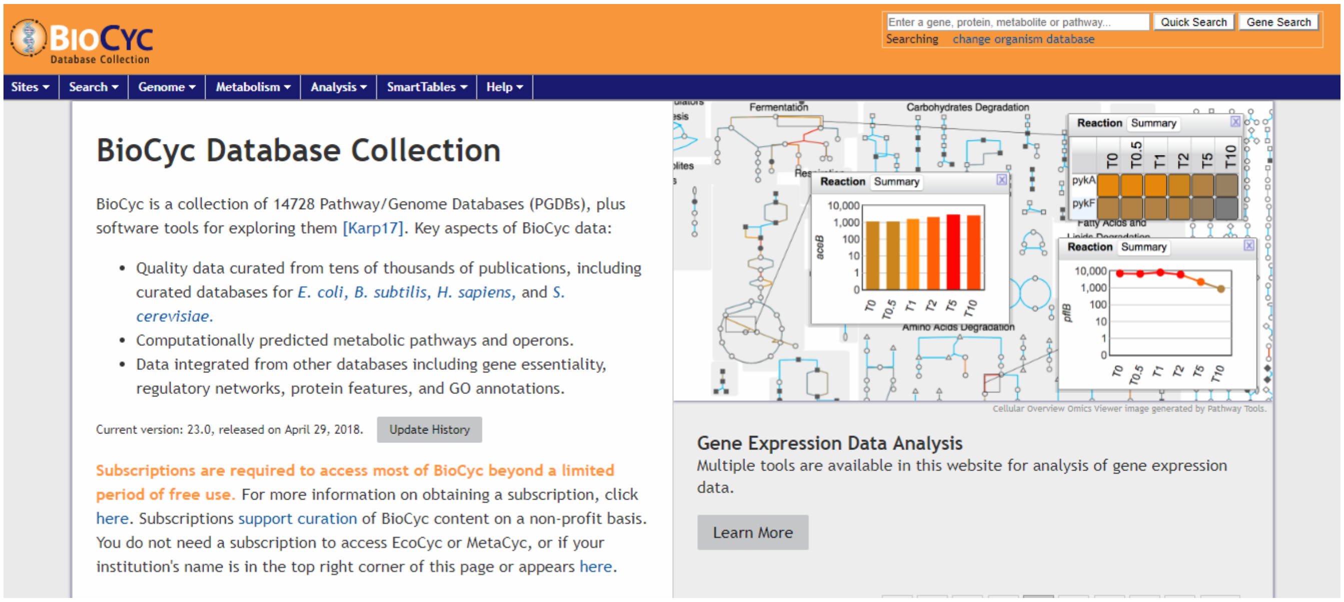
Documentation

Information to browse the database and use its principal tools for data analysis

Fig. 2 Reactome database

- BioCyc

The BioCyc database collection is an assortment of organism specific Pathway/ Genome Databases (PGDBs). They usually gives the evidence to the information in reference to genomic and metabolic pathway information for many organisms.



BioCyc Database Collection

BioCyc is a collection of 14728 Pathway/Genome Databases (PGDBs), plus software tools for exploring them [Karp17]. Key aspects of BioCyc data:

- Quality data curated from tens of thousands of publications, including curated databases for *E. coli*, *B. subtilis*, *H. sapiens*, and *S. cerevisiae*.
- Computationally predicted metabolic pathways and operons.
- Data integrated from other databases including gene essentiality, regulatory networks, protein features, and GO annotations.

Current version: 23.0, released on April 29, 2018.

Subscriptions are required to access most of BioCyc beyond a limited period of free use. For more information on obtaining a subscription, click here. Subscriptions support curation of BioCyc content on a non-profit basis. You do not need a subscription to access EcoCyc or MetaCyc, or if your institution's name is in the top right corner of this page or appears here.

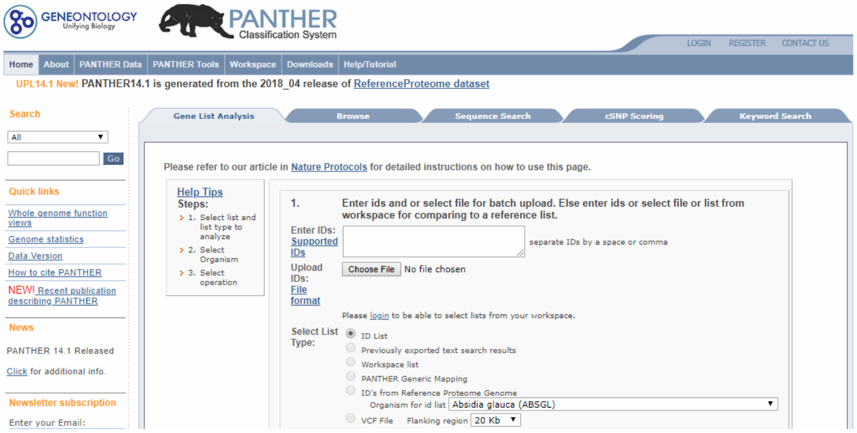
Gene Expression Data Analysis

Multiple tools are available in this website for analysis of gene expression data.

Fig. 3 BioCyc Database Collection

- Panther

PANTHER (Protein Analysis Through Evolutionary Relationships) Grouping System was intended to characterize proteins and their genetic factors so as to encourage high-throughput examinations. Proteins have been arranged by Family and subfamily, pathway, biological procedure and molecular capacity.



UPL14.1 New! PANTHER14.1 is generated from the 2018_04 release of ReferenceProteome dataset

PANTHER 14.1 Released

Click for additional info.

Enter your Email:

Please refer to our article in Nature Protocols for detailed instructions on how to use this page.

Help Tips
Steps:
1. Select list and list type to analyze
2. Select Organism
3. Select operation

1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs: Supported IDs separate IDs by a space or comma

Upload IDs: File format Choose File No file chosen

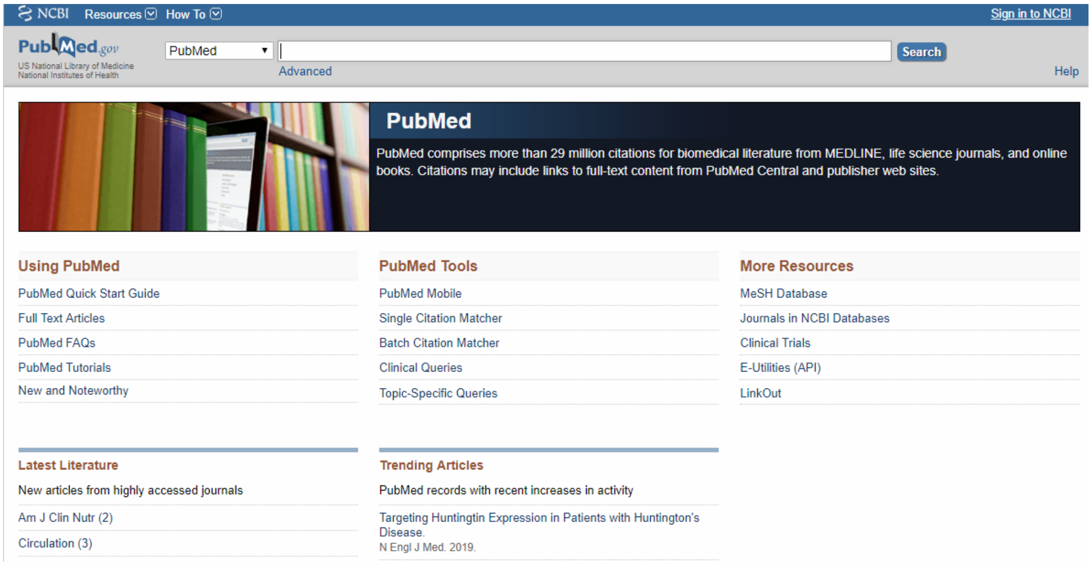
Please login to be able to select lists from your workspace.

Select List Type:
● ID List
○ Previously exported text search results
○ Workspace list
○ PANTHER Generic Mapping
○ ID's from Reference Proteome Genome
Organism for id list: Absidia glauca (ABSGL)
○ VCF File Flanking region: 20 Kb

Fig. 4 PANTHER Classification System

- PUBMED

NCBI (National Center for Biotechnology Information) develops and maintains a free source which is named PubMed. There is also a tool embedded in it which is called MEDLINE with advanced search features and filters which provides an insight to the studies being carried out.



PubMed

PubMed comprises more than 29 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy

PubMed Tools
PubMed Mobile
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries

More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities (API)
LinkOut

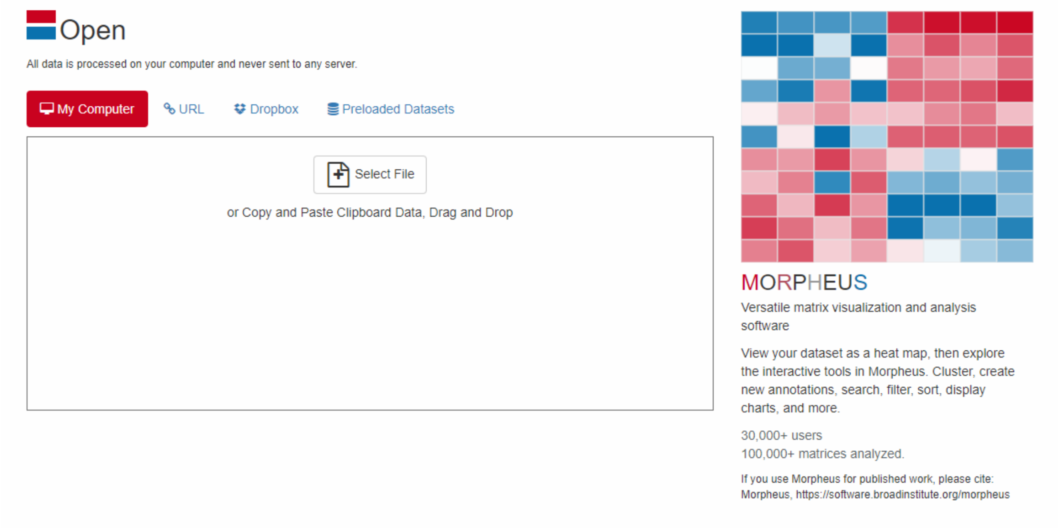
Latest Literature
New articles from highly accessed journals
Am J Clin Nutr (2)
Circulation (3)

Trending Articles
PubMed records with recent increases in activity
Targeting Huntingtin Expression in Patients with Huntington's Disease.
N Engl J Med. 2019.

Fig. 5 PUBMED

- MORPHEUS WEBSITE

It is a versatile matrix visualization and analysis software. It helps view the dataset as a heat map and then explore the interactive tools in Morpheus [12]. Helps to cluster, search, filter, sort and display charts to decipher the meanings behind the same. Also helps to create annotations. Morpheus helps to separate modeling from numerical implementation just by using a domain-specific mark-up language. The model description language allows users to describe their models in mathematical and biological terms.



All data is processed on your computer and never sent to any server.

or Copy and Paste Clipboard Data, Drag and Drop

Versatile matrix visualization and analysis software

View your dataset as a heat map, then explore the interactive tools in Morpheus. Cluster, create new annotations, search, filter, sort, display charts, and more.

30,000+ users
100,000+ matrices analyzed.

If you use Morpheus for published work, please cite:
Morpheus, https://software.broadinstitute.org/morpheus

Fig. 6 MORPHEUS Website

- THE CANCER GENOME ATLAS (TCGA)

The Cancer Genome Atlas (TCGA) is a project that is utilized to list hereditary changes in charge of malignancy, utilizing genome sequencing and bioinformatics [11]. TCGA applies high-throughput genome examination systems to improve our capacity to analyse, treat, and avert malignant growth through a superior comprehension of the hereditary premise of this malady.

- DAVID

The Database for Annotation, Visualization and Integrated Discovery (DAVID) includes a full Knowledgebase update to the unique web-open projects. DAVID presently gives a thorough arrangement of utilitarian explanation instruments for examiners to comprehend natural importance behind large list of genes [16].

3.2 Identification of DEGs and Microarray Data

NCBI-GEO was used for NGS from which colorectal malignant growth and contiguous gene articulation profile of GSE27000, GSE22915, GSE45066 and GSE75420 were collected. GSE27000's microarray information included 41 African Americans and 41 European Africans colorectal tissues and 40 ordinary. GSE22915 included 121 CRC tissue. GSE75420 included 4 sets of colorectal malignant growth tissues.

Morpheus Website was used for the analysis of raw data of high functional genomic expression and was processed in txt format. For the purpose of validation, the colon cancer data from the TCGA was analysed and downloaded.

3.3 Investigation of Pathway Enrichment

Signaling pathway and functional enrichment of DEGs was performed by means of KEGG PATHWAY, Panther and BioCyc. Up-controlled genetic factors that were largely augmented in G α signaling events. The chemokines receptors bind to the chemokines. And also enriched in signaling pathway of cell cycle. The down-controlled genetic factors were primarily augmented in mitotic prometaphase and cycle of the cell. The analysis of signaling pathway conveyed that in the cell cycle DEGs had some common pathways. Various databases were used for the analysis of the candidate DEGs pathways and functions. The most useful one was DAVID because of the visualization, integrated discovery function and gene annotation that it provides and therefore could provide biological meaning of the gene.

3.4 PPI Network Integration

The online database, STRING, was used to engage DEGs-encoded proteins and PPI. Cytoscape was used to produce interaction of protein bond network along with deciphering the interface bond of the applicant DEGs encrypted proteins in colorectal malignancy. Following, Network Analyzer was essentially acquired to decipher degree of the nodes which is basically the quantity of inter-connections to clear out the important genes of PPI. The core proteins and key important genes might be the corresponding proteins in the central nodes that have vital physiological regulatory functions.

Chapter 4

RESULT AND DISCUSSION

4.1 DEGs in Colorectal Cancers

By methods for $p < 0.04$ & logfc > 1 as cut-off measure, 1371, 7831, 2413 and 3949 differentially communicated qualities were deducted from the articulation profile datasets GSE27000, GSE22915, GSE45066 and GSE75420.

After coordinated bioinformatical investigation, aggregate of 282 reliably articulated genes were recognized via 4 profile datasets. 165 up-managed genes. 127 down-directed genes within colon diseased materials, contrasted with ordinary colon materials.

Utilizing Morpheus programming, a heat map was developed in which 165 up-controlled and 127 down-controlled DEGs were found utilizing information profile GSE45066 as a kind of reference, demonstrating the altogether differential dispersion of the 292 DEGs.

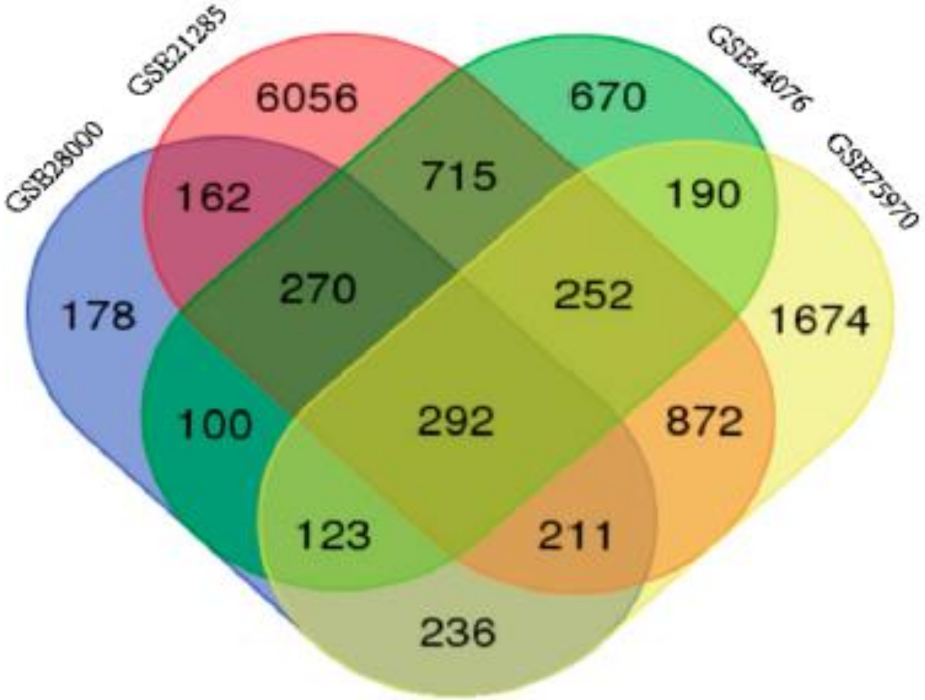
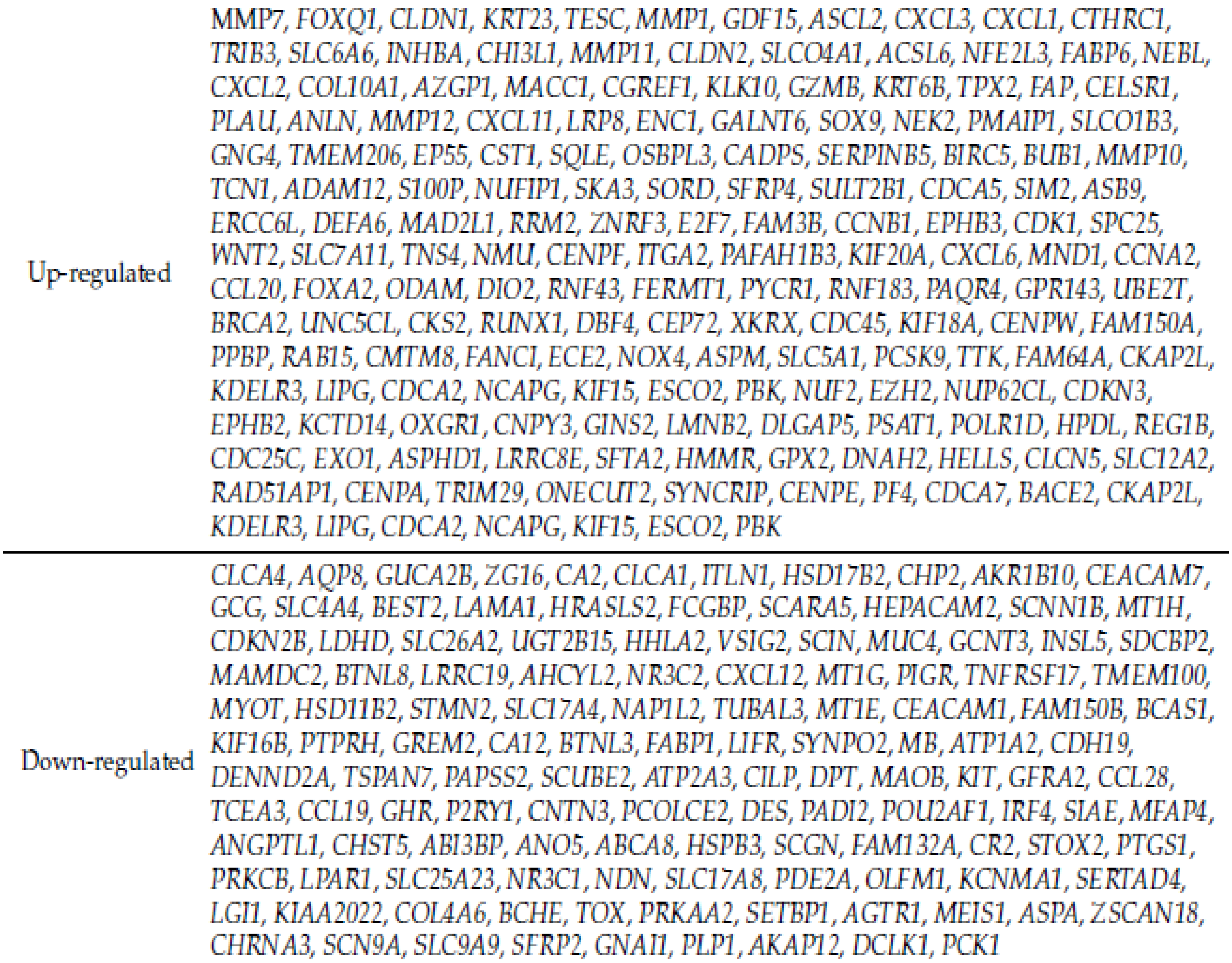


Fig. 7 From 4 profile data sets, 292 commonly alterations DEGs using Morpheus Website. Diverse datasets were represented with diverse colours. Usually changed DEGs were coloured in common and these were identified statistically.

Table 1. Identified DEGs along with the up-controlled and down-controlled genes



DEGs	Genes Name
Up-regulated	MMP7, FOXQ1, CLDN1, KRT23, TESC, MMP1, GDF15, ASCL2, CXCL3, CXCL1, CTHRC1, TRIB3, SLC6A6, INHBA, CHI3L1, MMP11, CLDN2, SLCO4A1, ACSL6, NFE2L3, FABP6, NEBL, CXCL2, COL10A1, AZGP1, MACC1, CGREF1, KLK10, GZMB, KRT6B, TPX2, FAP, CELSR1, PLAU, ANLN, MMP12, CXCL11, LRP8, ENC1, GALNT6, SOX9, NEK2, PMAIP1, SLCO1B3, GNG4, TMEM206, EP55, CST1, SQLE, OSBPL3, CADPS, SERPINB5, BIRC5, BUB1, MMP10, TCN1, ADAM12, S100P, NUFIP1, SKA3, SORD, SFRP4, SULT2B1, CDCA5, SIM2, ASB9, ERCC6L, DEFA6, MAD2L1, RRM2, ZNRF3, E2F7, FAM3B, CCNB1, EPHB3, CDK1, SPC25, WNT2, SLC7A11, TNS4, NMU, CENPF, ITGA2, PAFAH1B3, KIF20A, CXCL6, MND1, CCNA2, CCL20, FOXA2, ODAM, DIO2, RNF43, FERMT1, PYCR1, RNF183, PAQR4, GPR143, UBE2T, BRCA2, UNC5CL, CKS2, RUNX1, DBF4, CEP72, XKRX, CDC45, KIF18A, CENPW, FAM150A, PPBP, RAB15, CMTM8, FANCI, ECE2, NOX4, ASPM, SLC5A1, PCSK9, TTK, FAM64A, CKAP2L, KDELR3, LIPG, CDCA2, NCAPG, KIF15, ESCO2, PBK, NUF2, EZH2, NUP62CL, CDKN3, EPHB2, KCTD14, OXGR1, CNPY3, GINS2, LMNB2, DLGAP5, PSAT1, POLR1D, HPDL, REG1B, CDC25C, EXO1, ASPHD1, LRRC8E, SFTA2, HMMR, GPX2, DNAH2, HELLS, CLCN5, SLC12A2, RAD51AP1, CENPA, TRIM29, ONECUT2, SYNCRIP, CENPE, PF4, CDCA7, BACE2, CKAP2L, KDELR3, LIPG, CDCA2, NCAPG, KIF15, ESCO2, PBK
Down-regulated	CLCA4, AQP8, GUCA2B, ZG16, CA2, CLCA1, ITLN1, HSD17B2, CHP2, AKR1B10, CEACAM7, GCG, SLC4A4, BEST2, LAMA1, HRASLS2, FCGBP, SCARA5, HEPACAM2, SCNN1B, MT1H, CDKN2B, LDHD, SLC26A2, UGT2B15, HHLA2, VSIG2, SCIN, MUC4, GCNT3, INSL5, SDCBP2, MAMDC2, BTNL8, LRRC19, AHCYL2, NR3C2, CXCL12, MT1G, PIGR, TNFRSF17, TMEM100, MYOT, HSD11B2, STMN2, SLC17A4, NAP1L2, TUBAL3, MT1E, CEACAM1, FAM150B, BCAS1, KIF16B, PTPRH, GREM2, CA12, BTNL3, FABP1, LIFR, SYNPO2, MB, ATP1A2, CDH19, DENND2A, TSPAN7, PAPSS2, SCUBE2, ATP2A3, CILP, DPT, MAOB, KIT, GFRA2, CCL28, TCEA3, CCL19, GHR, P2RY1, CNTN3, PCOLCE2, DES, PADI2, POU2AF1, IRF4, SIAE, MFAP4, ANGPTL1, CHST5, ABI3BP, ANO5, ABCA8, HSPB3, SCGN, FAM132A, CR2, STOX2, PTGS1, PRKCB, LPAR1, SLC25A23, NR3C1, NDN, SLC17A8, PDE2A, OLFM1, KCNMA1, SERTAD4, LGI1, KIAA2022, COL4A6, BCHE, TOX, PRKAA2, SETBP1, AGTR1, MEIS1, ASPA, ZSCAN18, CHRNA3, SCN9A, SLC9A9, SFRP2, GNAI1, PLP1, AKAP12, DCLK1, PCK1

4.2 Analysis of Gene Ontology in Colorectal Cancers

With the use of Panther and DAVID, gene ontology analysis was performed on DEGs. The classification was then performed and then deduction was made that there were 3 functional groups.

These were as follows:

- **Biological Process Group**
- **Molecular Functional Group**
- **Cellular Component Group**

(i) In **Biological process group**, the down-regulated genes were mainly augmented in single-organism process and cellular procedure, cycle process of the mitotic cell and localization. The up-controlled genes were largely enriched in the process of single organism and also the single organism cellular processes and the process of cell proliferation.

(ii) In **Molecular functional group**, the down-controlled genes were mainly augmented in binding and binding of the protein. The up-controlled genes were largely augmented in binding of the protein and receptor.

(iii) In **Cellular component group**, the down-regulated genes were mainly augmented in membrane bound organelle and cell part. The up-regulated genes were largely enriched in extracellular region and extracellular space.

With the help of these results, it could be commented that bulk of the DEGs were considerably augmented in mitotic cycle of the cell, cell part, binding & single organism.

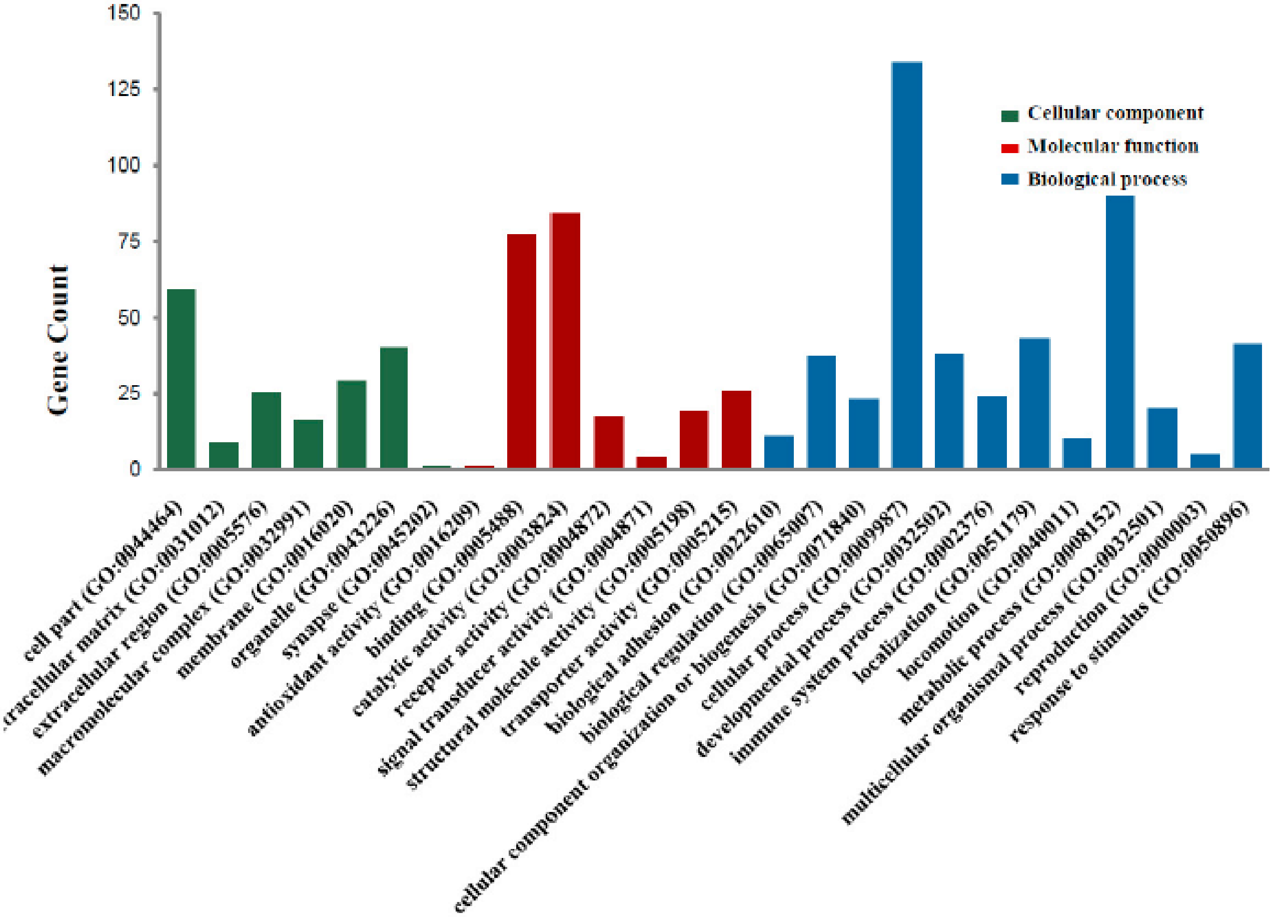


Fig. 8 Three clusters were made of DEGs with analysis of gene ontology

Term	Description	Count
Up-regulated		
GO:0043289	cell-proliferation	120
GO:0008184	extra-cellular	39
GO:0005213	single-organism process	43
GO:0000278	mitotic cell cycle	57
GO:1309074	cellular process	21
GO:0003432	biological regulation	21
GO:0009932	extracellular part	109
Down-regulated		
GO:0043289	single organism process	104
GO:0005232	binding	108
GO:0045387	cell	120
GO:0074833	localization	120
GO:1309074	protein binding	21
GO:0005430	mitotic cell cycle process	21
GO:0007864	organelle	109

Table 2. Analysis of significantly enriched DEGs in colorectal cancer

4.3 Analysis of Signaling Pathway Enrichment

KEGG PATHWAY and Panther was used to conduct signaling pathway enrichment. The down-controlled genes were solely enriched in mitotic prometaphase and binding of the receptors of the chemokine. The up-controlled genes were solely enriched in ligand binding, cycle of the cell. This analysis exhibited that mutual pathways were found in DEGs regarding cycle of the cell and receptors of the chemokine binding to the chemokines.

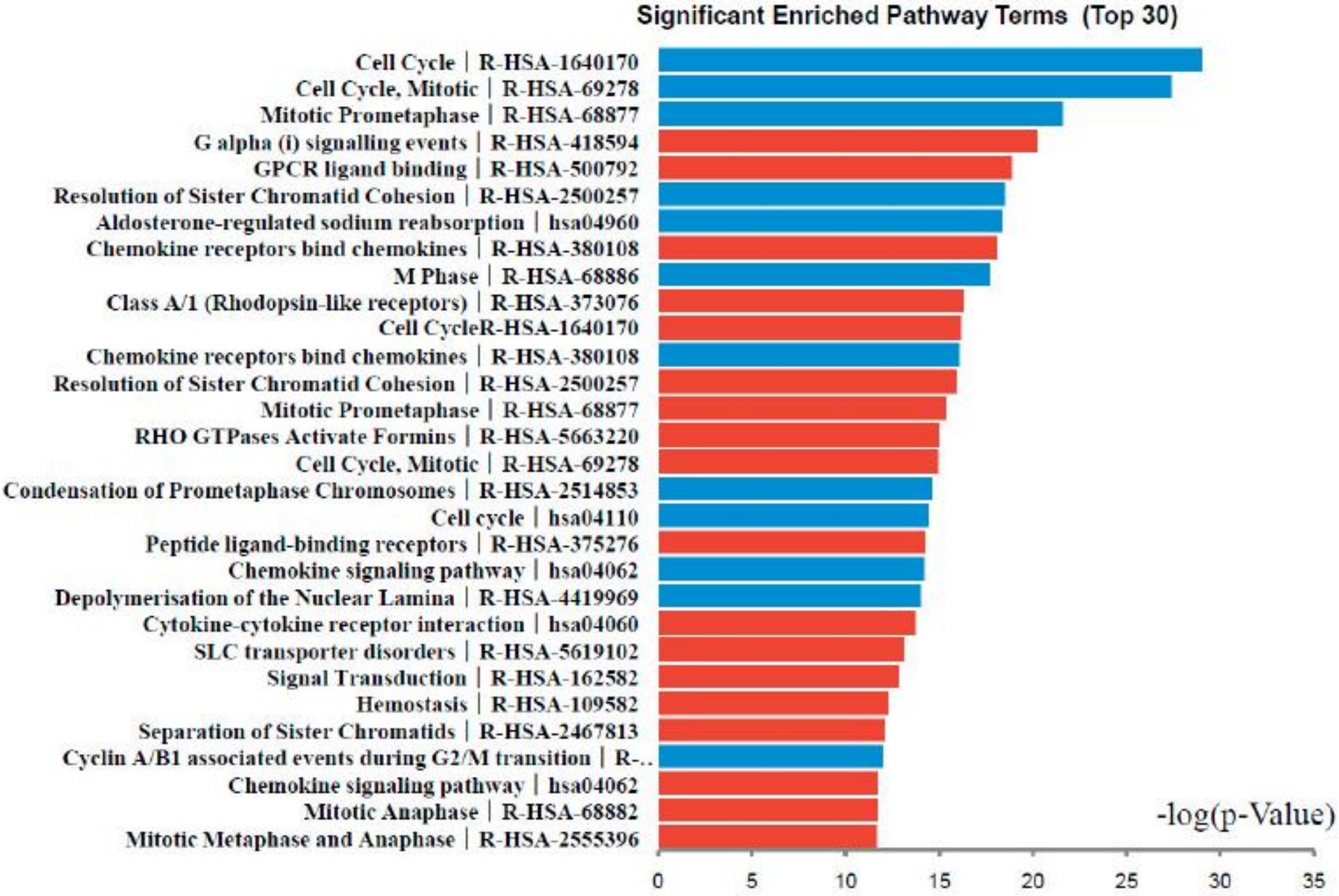


Fig. 9 DEGs in colon cancer with significantly enriched pathway

4.4 Identification of Key Candidate Genes

170 DEGs (111 up-controlled and 65 down-controlled) out of the 282 usually changed DEGs were strained in the complex of the DEGs, having 170 nodes, 510 edges. 110 out of 282 DEGs were falling outside the DEG complex.

Amid 170 nodes, recognition of 28 central node genes was possible with the help of the filters. Out of them the most important ten node degree genes were CKD1, CENB1, CCNP, KOF2A, CXCI2, DLGP, KCNA2, ITCA1, MAD3L2 and NMU1.

Giving the level of importance, 2 most important modules were chosen for additional deciphering.

The investigation of the pathway enrichment gave the following evidence:

- Module 1 had 17 nodes, 106 edges that were majorly connected with processes of cell cycle.

- Module 2 had 14 nodes, 84 edges that were majorly connected with pathway of chemokine signaling.

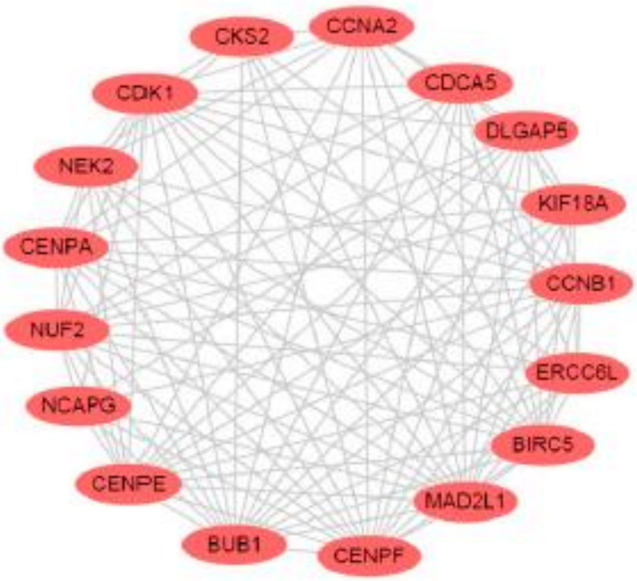


Fig. 10 17 nodes and 106 edges of the first module

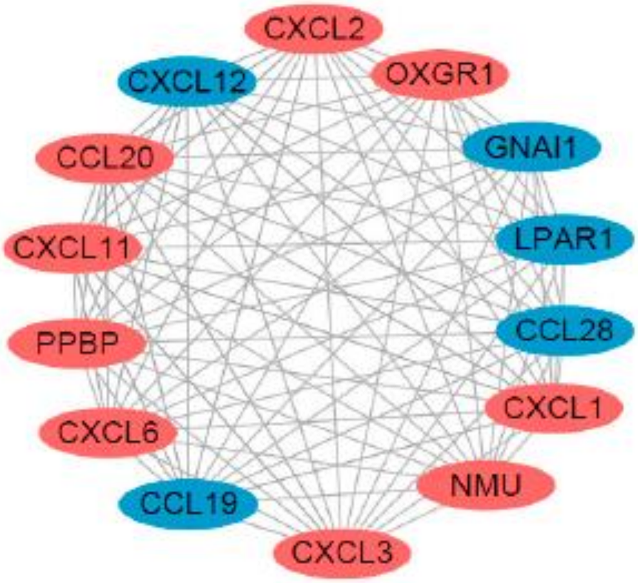


Fig. 11 14 nodes and 84 edges of the second module

4.5 DEGs Validation in TCGA Dataset

In order to approve the dependability of the recognized DEGs, TCGA CRC dataset was downloaded. Using similar policy the data was analysed. It was found that altogether 165 up-controlled genes that were identified were highly articulated too in the TCGA colorectal malignancies significantly. 114 down-controlled genes that were identified were also lowly articulated in the TCGA colorectal malignancies significantly. Some of the down-controlled genes couldn’t be on the gradient. Consistency of the up-controlled and the down-controlled was 89.7%, signifying that the consequences of the recognized nominee genes are dependable.

Chapter 5

CONCLUSIONS

With the help of various cohorts profile datasets and analysis using integrated bioinformatics, identification of commonly changed 292 DEGs was successful and out of them 28 usually changed hub genes was identified and were significantly enriched in various pathways. They're mainly linked with chemokines, processes of cell cycle and G-protein receptor signaling pathways. Verdicts like these considerably enhance the thoughtfulness of the root and the essential subatomic procedures in colon cancer. Thus, the pathways and the candidate genes involved can be taken forward as beneficial markers.

Chapter 6

DISCUSSION

Colorectal malignant growth is basically a gathering of molecularly and histologically heterogeneous infections described with varying arrangements of hereditary and epigenetic adjustments engaged with numerous useful flagging pathways, and the latter is tweaked by hereditary and epigenetic occasions, prompting the adjustments of quality articulation at transcriptional as well as translational dimensions. Subsequently, the attributes of CRC can't be clarified just by profiles of the gene expression, however genomic articulation profile can speak to adjustments in attributes of tumours. Numerous elements ought to be kept in mind, comprising genomic changes, non-coding RNAs and microRNAs, etc., that can take an interest in colorectal carcinogenesis and are related with existence to some extent.

REFERENCES

1. Guo, Y., Bao, Y., Ma, M. and Yang, W., 2017. Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis. *International journal of molecular sciences*, *18*(4), p.722.
2. Liang, B., Li, C. and Zhao, J., 2016. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. *Medical Oncology*, *33*(10), p.111.
3. Lynch, H.T. and De la Chapelle, A., 2003. Hereditary colorectal cancer. *New England Journal of Medicine*, *348*(10), pp.919-932.
4. Markowitz, S.D. and Bertagnolli, M.M., 2009. Molecular basis of colorectal cancer. *New England journal of medicine*, *361*(25), pp.2449-2460.
5. Siegel, R.L., Miller, K.D., Fedewa, S.A., Ahnen, D.J., Meester, R.G., Barzi, A. and Jemal, A., 2017. Colorectal cancer statistics, 2017. *CA: a cancer journal for clinicians*, *67*(3), pp.177-193.
6. Sano, H., Kawahito, Y., Wilder, R.L., Hashiramoto, A., Mukai, S., Asai, K., Kimura, S., Kato, H., Kondo, M. and Hla, T., 1995. Expression of cyclooxygenase-1 and-2 in human colorectal cancer. *Cancer research*, *55*(17), pp.3785-3789.
7. Center, M.M., Jemal, A., Smith, R.A. and Ward, E., 2009. Worldwide variations in colorectal cancer. *CA: a cancer journal for clinicians*, *59*(6), pp.366-378.
8. Haggar, F.A. and Boushey, R.P., 2009. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, *22*(04), pp.191-197.
9. Boyle, P. and Langman, J.S., 2000. ABC of colorectal cancer: Epidemiology. *BMJ: British Medical Journal*, *321*(7264), p.805.
10. Ng, E.K., Chong, W.W., Jin, H., Lam, E.K., Shin, V.Y., Yu, J., Poon, T.C., Ng, S.S. and Sung, J.J., 2009. Differential expression of microRNAs in plasma of patients with colorectal cancer: a potential marker for colorectal cancer screening. *Gut*, *58*(10), pp.1375-1381.
11. Peltomaki, P., Aaltonen, L.A., Sistonen, P., Pylkkanen, L., Mecklin, J.P., Jarvinen, H., Green, J.S., Weber, J.L. and Leach, F.S., 1993. Genetic mapping of a locus predisposing to human colorectal cancer. *Science*, *260*(5109), pp.810-812.

12. Saha, S., Bardelli, A., Buckhaults, P., Velculescu, V.E., Rago, C., Croix, B.S., Romans, K.E., Choti, M.A., Lengauer, C., Kinzler, K.W. and Vogelstein, B., 2001. A phosphatase associated with metastasis of colorectal cancer. *Science*, *294*(5545), pp.1343-1346.
13. Parsons, D.W., Wang, T.L., Samuels, Y., Bardelli, A., Cummins, J.M., DeLong, L., Silliman, N., Ptak, J., Szabo, S., Willson, J.K. and Markowitz, S., 2005. Colorectal cancer: mutations in a signalling pathway. *Nature*, *436*(7052), p.792.
14. Leary, R.J., Lin, J.C., Cummins, J., Boca, S., Wood, L.D., Parsons, D.W., Jones, S., Sjöblom, T., Park, B.H., Parsons, R. and Willis, J., 2008. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proceedings of the National Academy of Sciences*, *105*(42), pp.16224-16229.
15. Tol, J., Dijkstra, J.R., Klomp, M., Teerenstra, S., Dommerholt, M., Vink-Börger, M.E., van Cleef, P.H., van Krieken, J.H., Punt, C.J. and Nagtegaal, I.D., 2010. Markers for EGFR pathway activation as predictor of outcome in metastatic colorectal cancer patients treated with or without cetuximab. *European journal of cancer*, *46*(11), pp.1997-2009.
16. Hong, Y., Ho, K.S., Eu, K.W. and Cheah, P.Y., 2007. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clinical Cancer Research*, *13*(4), pp.1107-1114.
17. Whiffin, N., Hosking, F.J., Farrington, S.M., Palles, C., Dobbins, S.E., Zgaga, L., Lloyd, A., Kinnersley, B., Gorman, M., Tenesa, A. and Broderick, P., 2014. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Human molecular genetics*, *23*(17), pp.4729-4737.
18. Gulmann, C., Sheehan, K.M., Conroy, R.M., Wulfkuhle, J.D., Espina, V., Mullarkey, M.J., Kay, E.W., Liotta, L.A. and Petricoin Iii, E.F., 2009. Quantitative cell signalling analysis reveals down-regulation of MAPK pathway activation in colorectal cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 218*(4), pp.514-519.
19. Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P. and Romero, R., 2008. A novel signaling pathway impact analysis. Bioinformatics, 25(1), pp.75-82.

20. Longley, D.B., Allen, W.L. and Johnston, P.G., 2006. Drug resistance, predictive markers and pharmacogenomics in colorectal cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, *1766*(2), pp.184-196.

21. Zanke, B.W., Greenwood, C.M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S.M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E. and Ferretti, V., 2007. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics*, *39*(8), p.989.