ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Review Article*
www.ijptonline.com

IDENTIFICATION OF VITAL GENES AND PATHWAYS IN COLORECTAL CANCER BY BIOINFORMATICS ANALYSIS

S. Venkataramanan*, Nisha Thevandren
Department of Diagnostic and Allied Health Science, Faculty of Health and Life Sciences, Management and Science University, 40100, Shah Alam, Selangor, Malaysia.
Email: s_venkataramanan@msu.edu.my

Received on 02-04-2020 *Accepted on: 29-05-2020*

Abstract:

Colorectal cancer (CRC), also known as bowel cancer, colon cancer, or rectal cancer, is the development of cancer from the colon or rectum. This study aim to explore the differentially expressed genes of colorectal cancer by bioinformatics analysis and identification of the dominant gene that cause colorectal cancer. The gene expression profile GSE 101896 was downloaded from the gene expression omnibus (GEO) database. Then, the genes between male and female were separated into two different groups and analyzed by GEO2R. Functional annotation of genes were performed in DAVID (database for annotation, visualization and integrated database), and pathway enrichment analysis by Gene Ontology database using PANTHER (Protein Analysis Through Evolutionary Relationships) classification system. Then, the analysis was carried out by the STRING server to create protein network interaction and Cystoscape was used to identify the hub gene that cause colorectal cancer. As the result, about 65 up regulated and 174 of down regulated colorectal genes were obtained from the GEO2R. DAVID database, systematically combine functional descriptive of colorectal cancer with intuitive graphical summary. Besides that, Gene Ontology using PANTHER classification system classifies proteins to facilitate high-throughput analysis according to molecular function, biological process and cellular pathway. Further, the STRING server resulted in protein–protein interactions network structure among the genes. Then, from Cytoscape EIF1AX from up regulated and SIRT6 from down regulated was selected as hub gene of the colorectal cancer.

Overally, the hub gene validated as the dominant gene which differentially expressed in colorectal cancer.

Keywords: GSE 101896, colorectal cancer, DAVID database, PANTHER, STRING server and Cytoscape.

1.0 Introduction

Colorectal cancer begins when healthy cells in the lining of the colon or rectum change and grow out of control, forming a mass called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but will not spread. These changes usually take years to develop. Both genetic and environmental factors can cause the changes. However, when a person has an uncommon inherited syndrome changes can occur in months or years [1].

In the present study, the study was carried out by downloading micro array data of GSE and identified the female and male colorectal cancer samples to explore the biological process in this cancer. Experiment type that was conducted to obtain the samples is by expression profiling by array. Where the Messenger RNA were extracted from primary tumor of 60 colorectal cancer patients, hybridized and scanned with Affymetrix Human Genome Gene Chip U133 Plus2.0 array. However, the dataset which was uploaded are not systematically analyzed. Thus, the functional enrichment analysis and functional annotation was carried out by using DAVID database.

Then, the using the genes of up regulated and down regulated the study was carried out Gene Ontology using PATHER to carry on the pathway enrichment analysis that were playing a role in Colorectal Cancer genesis. Later, STRING was performed to create PPI interaction. Finally, the Cytoscape was used to identify the hub gene for both up regulated and down regulated gene. Cancer genesis and then screened out the potential biomarkers for the colorectal cancer. This study is likely play a significant role in IPF genesis and may potentially serve as biomarkers in diagnosis and treatment of colorectal cancer.[2]

1.1 Research Background

This study focuses on identification of dominant gene that cause the colorectal cancer.Bioinformatics analysis by using gene onlogy(PANTHER) and DAVID database was done to identifies the cellular components ,

biological process and molecular pathways to identify the characteristic of the gene involve.The identification was done by step to step to understand the pathway of the colorectal cancer clearly.

1.2 Problem Statement

The number of people affected with colorectal cancer are increasing year by year nowadays.Therefore, the main purpose of this study is to find the hub gene that cause the colorectal cancer.By understanding the cellular component and pathways involve in the hub gene,it will be easier to design a new biomarker to cure this disease.

1.3 Objectives

- To identify the dominant genes that lead to the breast cancer.
- To understand the cellular components and pathways of the colorectal cancer genes.

1.4 Hypothesis

Ho: The genes that identified in the colorectal cancer does not plays the important role in disease associated genes.

H1: The genes that identified in the colorectal cancer plays the important role in disease associated genes.

2.0 Materials and methods

2.1 Acquisition and analysis of datasets using Gene Expression Omnibus database.

The microarray data of GSE101896 were downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/query/) and the platform were Affymetrix Human Genome Gene Chip U133 Plus2.0 array. Samples comprising case study of colorectal tumor response from messenger from primary tumor of 60 colorectal cancer patients. The total of 22 female patients with primary tumor and 38 male patients of primary tumor were selected in GEO2R analysis.[3]

2.2 Inclusion criteria for DEGs

The GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/) is an interactive web tool that allows users to compare two or more groups of sample in a GEO series in order to identify genes that are differently expressed across experimental conditions. GEO2R is a very nice web-based tool to do this graphically and automatically. Enter

the GEO series number in the search. Start by creating groups (e.g. control vs treatment, early vs late time points in a time course, etc), then select samples to add to that group.[4]

2.3 Gene Ontology and pathway enrichment analyses

Gene Ontology (GO) knowledgebase is the largest source of information which describe on the functions of genes. Gene Ontology provide both human readable and machine readable information's. Then for the pathway analysis DAVID database were used to analyze according to the gene functions.[5]

2.4 Protein-protein network construction.

STRING server was used to create a protein-protein network using all the genes that involve in colorectal cancer. STRING server will provide the information. STRING database provides information from numerous sources, including experimental data, computational prediction methods and public text collections. The up regulated and down regulated genes was analyzed separately in this STRING database. The network was then analyzed.[6]

2.5 Identifying hub gene

Cytoscape 3.7.2 was used to determine the hub gene that causes the breast cancer. The hub gene for upregulated and down regulated was identified separately. The hub gene identify based on the highest central betweeness and node degree.[7]

3.0 Results

3.1 Identification of regulated gene

Identification of the colorectal groups in the GSE101896 expression profiling datasets were analyzed using GEO2R tool. The total of 22 female patients with primary tumor and 38 male patients of primary tumor were selected in GEO2R analysis from the total of 60 colorectal cancer of primary tumor patients was analyzed. In this dataset about about of upregulated and of downregulated were identified (Table 1).

Genes	Total
Upregulated	65
Downregulated	174

3.2 Functional Annotation using DAVID database

David database used to identify the functional annotation of the both upregulated and downregulated gene. The highest number gene which is about 11 genes from the total of upregulated genes has identified that they have metal-binding molecular property and has nucleus to carry out basic cellular activities. From the result, 2-oxoglutarate and Fe(II) dioxygenase domain containing protein 1(ofd1) is the functional related gene which has the highest kappa as in Schizosaccharomyces pombe 972h.From the upregulated, identified that obsolete negative regulation of SREBP signaling pathway by positive regulation of transcription factor catabolic process in response to increased oxygen levels.

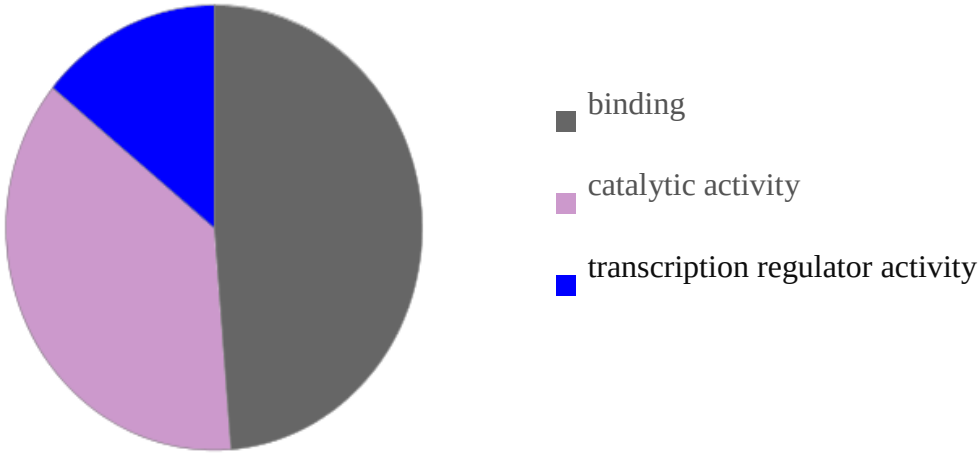
From the downregulated gene, about 63 genes are active in alternative splicing. It shows that protein for which at least two isoforms exist due to distinct pre-mRNA splicing events. Moreover, about 54 genes cointain phosphoprotein. Protein for which at least two isoforms exist due to distinct pre-mRNA splicing events. About 54 genes are involve in protein binding. Protein binding are interacting selectively and non-covalently with any protein or protein complex (a complex of two or more proteins that may include other nonprotein molecules).

3.3 Identification of pathways of GO using PATHER

3.3.1 Molecular function

Total genes: 91

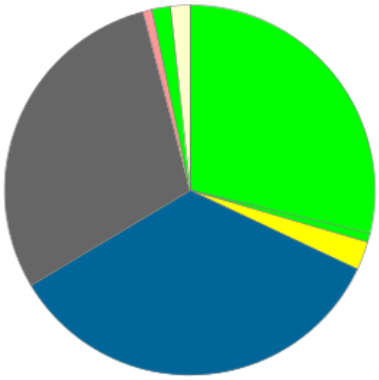
Total function molecular hits: 78



3.3.2 Biological function

Total genes: 98

Total biological hints: 122

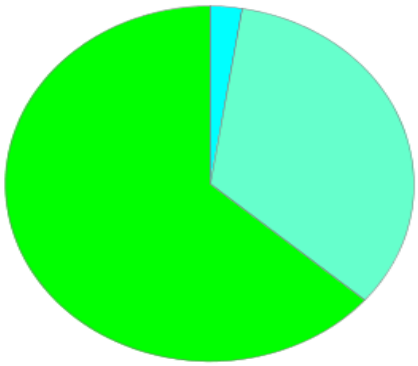


biological regulation
cell proliferation
cellular component organization or biogenesis
cellular process
metabolic process
Multicellular organismal process
reproduction
response to stimulus

3.3.3 Cellular component

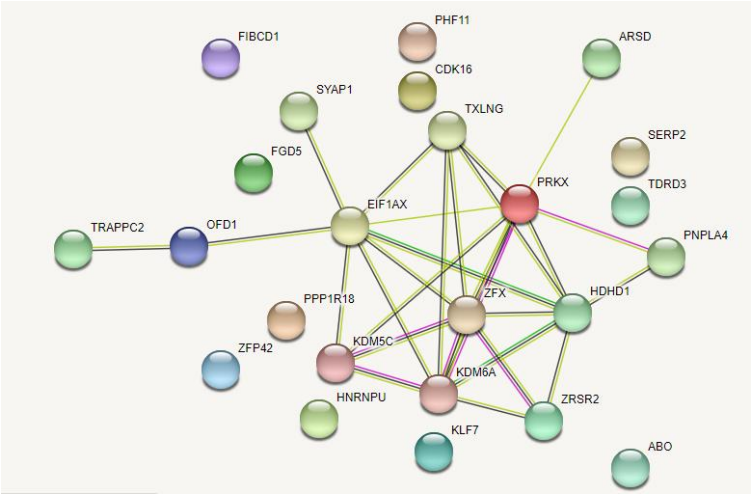
Total cellular: 91

Total component hits: 77



cell
organelle
protein-containing complex

4.0 Protein-protein network construction



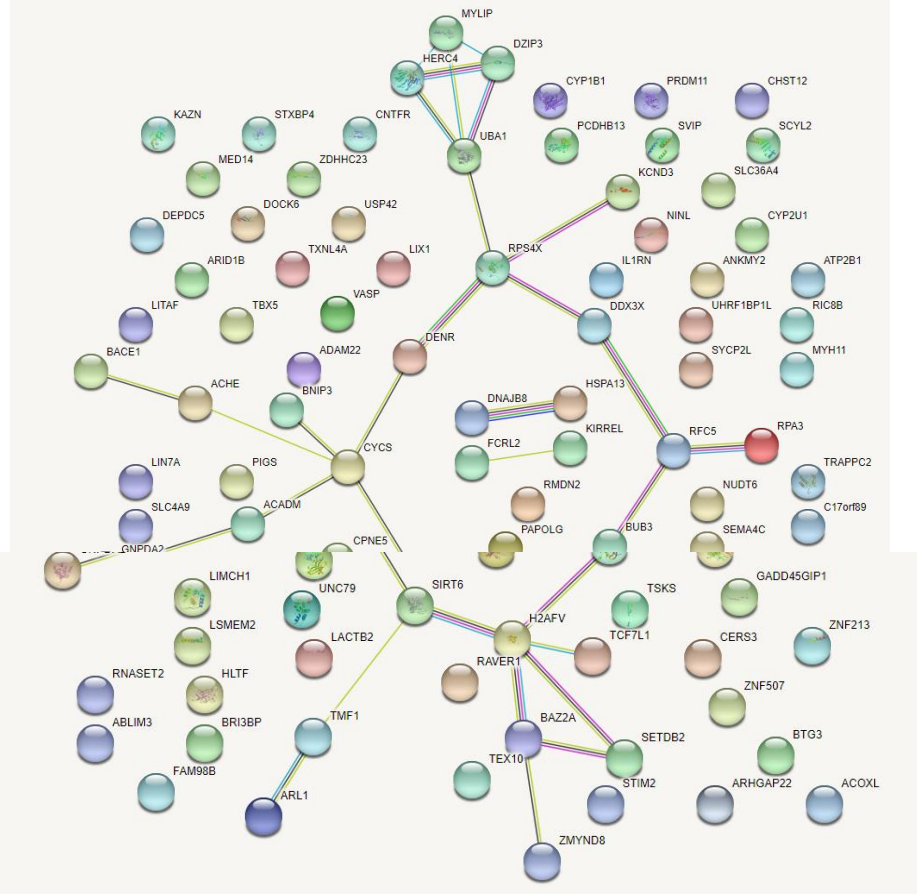
The above protein network shows, the network for the upregulated gene.

Number of nodes: 24

Number of edges: 28

Average node degree: 2.33

Avg. local clustering coefficient: 0.417



The above protein network shows, the network for the down regulated gene.

Number of nodes: 90

Number of edges: 31

Average node degree: 0.689

Avg. local clustering coefficient: 0.188

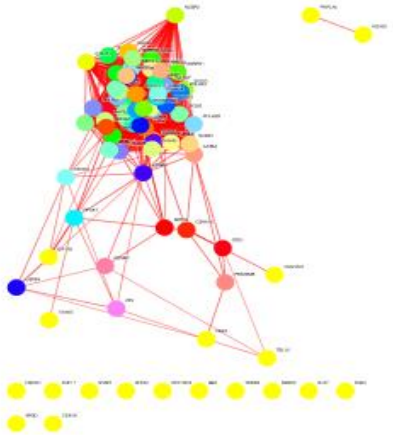
Expected number of edges: 31

PPI enrichment p-value: 0.559

5.0 Identification of hub gene

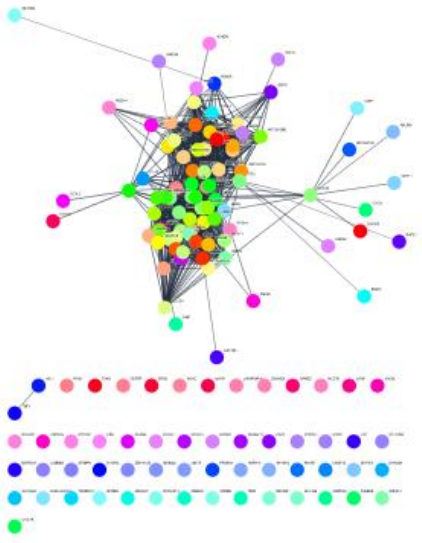
Using Cytoscape the upregulated and downregulated gene was identified. In hub gene identification, the z-score was set at 0.07 and the total amount of interactors set at 55.

For the upregulated genes, about 79 nodes were worked.



Gene	Betweeness	Node Degree
EIF1AX	**22.84**	**5.0**
PRKX	18.23	4.0
TBL1X	9.33	3.0

For down regulated about 145 nodes were worked.



Genes	Betweeness	Node Degree
SIRT6	**13.0**	**15.0**
HLTF	6.81	10 .0
RPS4X	4.0	5.0

6.0. Discussion

Gene ontology (GO) is a common method for annotating genes, gene products and sequences to underlying biological phenomena. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an integrated database resource for biological interpretation of genome sequences and other high-throughput data. Both analyses were available in the DAVID database (https://david.ncifcrf.gov/), which is a bioinformatics data resource composed of an integrated biology knowledge base and analysis tools to extract meaningful biological information from large quantities of genes and protein collections. GO and KEGG analyses were performed using the DAVID database to identify DEGs. A *P* value <0.05 was set as the cut-off criterion.

9 among the 29 nodes from the upregulated gene, it shows by a study that human respiratory conditions are largely influenced by the individual's sex resulting in overall higher risk for males. Sex-based respiratory differences are present at birth suggesting a strong genetic component. Our objective was to characterize early life sex-based genomic signatures determined by variable X-chromosome methylation in the airways. We compared male versus female genome-wide DNA methylation in nasal airway samples from newborns and infants aged 1-6 months (N = 12). We analyzed methylation signals across CpG sites mapped to each X-linked gene using an unsupervised classifier (principal components) followed by an internal evaluation and an exhaustive cross-validation. Results were validated in an independent population of children (N = 72) following the same algorithm. X-linked genes with significant sex-based differential methylation in the nasal airway of infants represented only about 50% of the unique protein coding transcripts. X-linked genes without significant sex-based differential methylation included genes with evidence of escaping X-inactivation and female-biased airway expression. These genes showed similar methylation patterns in males and females suggesting

unbalanced X-chromosome dosage. In conclusion, we identified that the human airways have already sex-based DNA methylation signatures at birth. These early airway epigenomic marks may determine sex-based respiratory phenotypes and overall predisposition to develop respiratory disorders later in life.[8]

Among 88 of 16244 upregulated genes are highly rich in cellular components that are not protein-containing complexes or whole cells. Now the cellular components aspect of the ontology consist of cell, cellular anatomical entity, intracellular ,other organism part, protein-containing complex, virion and virion part. As a part of this reorganization we are obsoleting 'x part' terms such as 'nuclear part'. Instead, parts will be asserted directly in the ontology. For example, a term x that is a nuclear part will have the relationship x part_of nucleus.[9]

From the Cytoscape database, EIF1AX was choosed as the main hub gene from the up regulated gene. Eukaryotic translation initiation factor 1A, X-chromosomal (EIF1A) is a protein that in humans is encoded by the *EIF1AX* gene. This gene encodes an essential eukaryotic translation initiation factor. The protein is a component of the 43S pre-initiation complex (PIC), which mediates the recruitment of the small 40S ribosomal subunit to the 5' cap of messenger RNAs. EIF1A is a small protein (17 kDa in budding yeast) and a component of the 43S preinitiation complexes (PIC). EIF1A binds near the ribosomal A-site, in a manner similar to the functionally related bacterial counterpart IF1[10]. However, SIRT6 was choosed as main hub gene from the downregulated gene.

Sirtuin 6 (SIRT6) is a stress responsive protein deacetylase and mono-ADP ribosyltransferase enzyme encoded by the SIRT6 gene.SIRT6 functions in multiple molecular pathways related to aging, including DNA repair, telomere maintenance, glycolysis and inflammation.Sirt6 has been disrupted, exhibit a severe progeria, or premature aging syndrome, characterized by spinal curvature, greying of the fur, lymphopenia and low levels of blood glucose. The lifespan of Sirt6 knock-out mice is typically one to three months, dependent upon the strain in which the Sirt6 gene has been deleted. SIRT6 and miR-122 negatively regulate each other's expression. SIRT6 was shown to act as a tumor suppressor that blocks the Warburg effect in cancer cells.[11]

7.0 Conclusion

As conclusion, all the results provide detailed information of molecular mechanism of colorectal cancer. All these pathways and analysis will help in identifying potential target biomarkers and therapeutic target. Identification of hub gene has given more ease for these potential target biomarkers. However, molecular biological experiment needed to prove the findings.

References

1. Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., … Watanabe, T. (2015). Colorectal cancer. *Nature Reviews Disease Primers*, *1*(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4874655/.
2. Ren, Y.-M., Duan, Y.-H., Sun, Y.-B., Yang, T., & Tian, M.-Q. (2018). Bioinformatics analysis of differentially expressed genes in rotator cuff tear patients using microarray data. *Journal of Orthopaedic Surgery and Research*, *13*(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6234628/.
3. Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods in Molecular Biology*, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
4. Colquitt, J. L., Mendes, D., Clegg, A. J., Harris, P., Cooper, K., Picot, J., & Bryant, J. (2014). Implantable cardioverter defibrillators for the treatment of arrhythmias and cardiac resynchronisation therapy for the treatment of heart failure: systematic review and economic evaluation. *Health Technology Assessment*, *18*(56), 1–560. https://doi.org/10.3310/hta18560.
5. Manosij Ghosh, Shemim Begum, Ram Sarkar, Debasis Chakraborty and Ujjwal Maulik, Recursive Memetic Algorithm for gene selection in microarray data, *Expert Systems with Applications*, 10.1016/j.eswa.2018.06.057, **116**, (172-185), (2019).
6. Raman, K. (2010). Construction and analysis of protein–protein interaction networks. *Automated Experimentation*, *2*(1), 2. https://doi.org/10.1186/1759-4499-2-2.

7. Le, D.-H. (2017, June 15). HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network. Retrieved November 14, 2019, from https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-017-0437-x.
8. LoMauro, A., & Aliverti, A. (2018). Sex differences in respiratory function. *Breathe, 14*(2), 131–140. https://doi.org/10.1183/20734735.000318.
9. Mule, Dr. S. S., & Waykar, Y. (2015). Role of Use Case Diagram in S/W Development. *International Journal of Management and Economics*. Retrieved from. https://www.researchgate.net/publication/322991847_role_of_use_case_diagram_in_software_development
10. Karunamurthy, A., Panebianco, F., Hsiao, S. J., Vorhauer, J., Nikiforova, M. N., Chiosea, S., & Nikiforov, Y. E. (2016). Prevalence and phenotypic correlations of EIF1AX mutations in thyroid nodules. *Endocrine-Related Cancer, 23*(4), 295–301. https://doi.org/10.1530/erc-16-0043.
11. SIRT6 sirtuin 6 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved December 19, 2019, from https://www.ncbi.nlm.nih.gov/gene/51548.

Corresponding Author:

S. Venkataramanan*,

Department of Diagnostic and Allied Health Science, Faculty of Health and Life Sciences, Management and Science University, 40100, Shah Alam, Selangor, Malaysia.

Email: s_venkataramanan@msu.edu.my