https://doi.org/10.46344/JBINO.2020.v09i05.26

CERVICAL CANCER AND GENE EXPRESSION ANALYSIS WITH KEY GENES IDENTIFICATION BY COMPUTATIONAL METHOD

Swaminathan Venkataramnan*, Wan Nurfazreen Binti Zainol Izam Khan

Faculty of Health and Life Sciences, Management and Science University, Seksyen 13, 40100 Shah Alam, Selangor, Malaysia

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

ABSTRACT

The second most unpropitious female illness, Cervical cancer, has been an alarming disease amongst Malaysian women population. This sight has been proven since twenty years ago in the scene with half a million cases worldwide and more than 50% mortality rate in Asia up till present time. Functional annotation analysis was conducted with a total of 247 gene list at Gene Expression Omnibus (GEO) database, then performed in DAVID (Database for Annotation, Visualization and Integrated Discovery) continued in PANTHER (Protein Analysis Through Evolutionary Relationships database) for cervical cancer gene understanding through biological, molecular function to cellular level. Protein-protein interaction network of 61 upregulated and 187 downregulated gene lists were analyzed in STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) server. The network between the gene proteins were represented in nodes and edges with information such as number of networks and PPI enrichment values. As a conclusion, the discovered centre for cervical cancer hub genes were AURKA, UBE2C, TPX2, MCM5 and GINS2 in upregulated gene list and SPRR1B, DSC3, SPRR1A, DSG1 and SPRR2D in downregulated gene list.

Keywords : cervical cancer, cytoscape server, david server, gene profiling, gene expression, geo omnibus, human papillomavirus, microarray, panther server, protein-protein interaction network, string server.

INTRODUCTION

Cervical cancer has been the ultimate agony of the female population in Malaysia since the early days of HPV infection broke out. The disastrous malignancy infection has become this nation's second most horrified female cancer, following after breast cancer [1]. Despite being a developed country with heavily subsidized healthcare benefits [2]. Malaysian females are still heavily blinded by the fact of the discriminated and tabooed topic should never be publicly discussed and combat in public. Precancerous and cancerous lesions are generally detected to be worrisome only at the later stages (Stage 3 and Stage 4). Human papillomavirus (HPV) is a devastating sexually transmitted disease affecting both genders, male and female. The infection is fueled by more than 100 diversed strains, but some being a high-risk strain – HPV 16 and 18. The powerful HPV strain 16 is the main culprit behind almost 70% cervical causing cancers [3], primarily being transmitted through unprotected sexual contact. High risk human papilomavirus or alphapapilomavirus is ignited in the human body when the dangerous strain enters a wounded lesion of basal epithelial cells region. 50 to 100 genomes per cell will be generated there and malignant cells spread like wildfire as they duplicates [4]. Cervical cancer do not show any symptoms that can be experienced by affected individual, but signs arises when the disease is at a daunting stage. Pain during sexual intercourse, abnormal watery and foul-smelling vaginal discharge, unexpected bleeding before and after menstruation cycle are among of the signs one should be aware of. Predominantly, HPV virus are transmitted through sexual contact, but there are studies and evidences of mother- newborn cases that happens during delivery.

Gene expression method have been applied to study comprehensively HPV infection in the human body. The method digs into how it affects human biological system and their metabolic pathway is further elucidated. Malaysian government have made continuous effort in taking preventive measures towards the cancerous HPV infection by introducing vaccination (primary prevention) as early as in 1960's till present era [1]. Screenings and treatments of pre-cancerous lesions falls into the secondary prevention method. In addition to that, therapeutic chemotherapy and radiation treatments are aggressively in action as tertiary treatments for diagnosed, invasive cancer. The Ministry of Health are taking their stand on preventing the spread of this deadly but preventable disease by holding nationwide healthcare campaigns, attitude, prevention and knowledge awareness. In 2010, three years after the first HPV vaccine was approved in Malaysia, nationwide, thrice dosage immunization package was introduced for females in Malaysian high schools and they are free of charge. The effort was indeed a maximized collaboration in aggressive approach for cervical cancer prevention [5].

The snowballing public repository demand for microarray experiments that are high-throughput began in 1999 and gene expression omnibus (GEO) was released to the public. This open design server serves as a storage, submission and retrieval of ceaseless amount of data sets that were obtained from antibody array experiments, genomic hybridization and gene expression [6]. At the moment, GEO has a billion of expressed individual genes, that were derived from 1500 laboratories that addresses extensive biological conditions [7]. Differentially expressed genes is an important and crucial step in microarray data analysis as scientific information disseminates across huge datasets and the results are summarized graphically of their functional information for researchers and public [8] certainly the database for annotation, visualization, and integrated discovery (DAVID) gave a wholesome picture of four distinctive categories which are annotation tool, gocharts, kegg charts and domain charts. These categoriess swiftly adjoins descriptive data from multiple open source databases to the gene lists [9].

Proteins interaction in biological pathways, cell to cell communication and the signal regulations displayed between each other or among distinct cells are perfectly potrayed in protein annotation through evolutionary relationship server (PANTHER) is a part of an integral approach in understanding the importance of biological pathway ontologies [10]. The statistical based analysis tool feature in the server gives biologists the opportunity to analyze sequencing, proteomics or gene expression experiments data. The wide range of gene fammilies and subfamilies, including evolutionary relationships are expressed in phylogenetic trees, statistical models (hidden Markov models or HMMs) and multiple sequence alignments [11].

Protein to protein interaction network (PPI) in search tool for the retrieval of interacting genes/proteins server (STRING) provides a system-wide comprehension of the most in depth cellular function details that requires broad functional interactions between the expressed proteins [12]. The connections and associations includes direct (physical) and indirect (indirect) interactions with imported known pathways and complexes of proteins derived from curated databases. The feature of subsets network visualization, enrichment analysis such as gene ontology and metabolic pathways with additional of clustered network hierachy, STRING server maps the protein structures accordingly[13]. Mutually exclusive interactions are also exposed with structural details, nodes (protein) and edges (interactions), after the PPI were constructed [14].

- Author : Wan Nurfazreen Binti Zainol Izam Khan, is currently pursuing Bioinformatics degree in Management and Science University (MSU), Shah Alam, Malaysia. Contact number: 017-4183298. Email : acincyrus@gmail.com
- Co- Author: Venkataramanan Swaminanthan, a Bioinformatics senior lecturer at Management and Science University (MSU), Malaysia. Contact number: 017-3173578. Email : s_venkataramanan@msu.edu.my.

RESEARCH BACKGROUND

This research was conducted in identifying the dominant cervical cancer causing genes and thoroughly expressing them biologically and pathway mechanism wise.

PROBLEM STATEMENT

The HPV gene is not thoroughly comprehended and expressed in cervical cancer disease (Ramakrishnan et al., 2015).

HYPOTHESIS

Null : The identification of HPV gene cannot be expressed through gene expression via Bioinformatics method.

Alternative : HPV gene was expressed through Bioinformatics method.

METHODOLOGY

GENE EXPRESSION ANALYSIS

Omnibus (GEO tool)

The cervical cancer data set with Geo ID : GSE9750 was put into accession box and the results are then seperated into normal and cervical cancer samples, with a total of 66 samples. There were 33 primary tumors, 9 cell lines, and 24 normal cervical epithelium.

GEO2R

GEO2R is a collective web server that distinctively categorizes the cervical cancer samples. The server works on datasets of samples expressed by array expression profiling method. In this study, there were 24 cervical cancer samples and 12 normal sample conditions identified, after the analysis.

DAVID Functional Annotation Clustering

Following after that, DAVID (Database for Annotation, Visualization and Integrated Discovery) server, as a functional annotation clustering tool was used. The main purpose of this tool was to understand the three major aspects of a HPV gene protein which are biological component, cellular component and molecular function. DAVID enhanced the interpretation of a massive gene list in a gene network context. This data mining tool extends module-centric approach where cervical cancer are studied from gene-to-gene or similar relationships applied to it, with information embedded on the server. DAVID has three major stages where firstly, gene pairs were measured in terms of functional relationship, then proceeded to agglomeration procedure to assemble the like-genes into their respective categories. Finally, the results were visualized in two approaches, graphical and text modes.

PANTHER

PANTHER (Protein Annotation Through Evolutionary Relationship) is a classification system server that was used with the aim to maximized HPV gene biologically from molecular function, biological component to cellular level. Protein families and subfamilies were annotated with ontology terms and later curated with PANTHER pathways. The understanding of the processess delivered both graphical and statistical results.

STRING

The gene lists of up regulated and down regulated were analysed on their protein-protein interaction network respectively. The network between the gene proteins were represented in nodes and edges with information such as number of networks and PPI enrichment values were calculated and displayed.

CYTOSCAPE

In addition to that, the upregulated and downregulated gene lists information were thoroughly analysed in this server for the degree of betweenness was set at > 0.4 and centrality at >20.0. The top 5 highest genes from both categories were counted as the hub genes of cervical cancer. The interactions were later visualized as a merged network.

RESULTS AND ANALYSIS

Omnibus GeoTool

The gene list results were obtained from GEO dataset after the Top250 analyzing process were performed. The positive and negative logFC values were differentiated into two separate tables which are upregulated and downregulated gene lists, respectively (Table 1).

CATEGORY	UPREGULATED GENE LISTS	DOWNREGULATED GENE LIST
	61 genes	187 genes

TABLE 1 : Tabulated gene lists results of upregulated and downregulated cervical cancer genes.

DAVID server results

FIGURE 2 : The gene list report was obtained from DAVID server analysis with aurora kinase (AURKA) as the dominant gene.

PANTHER server results

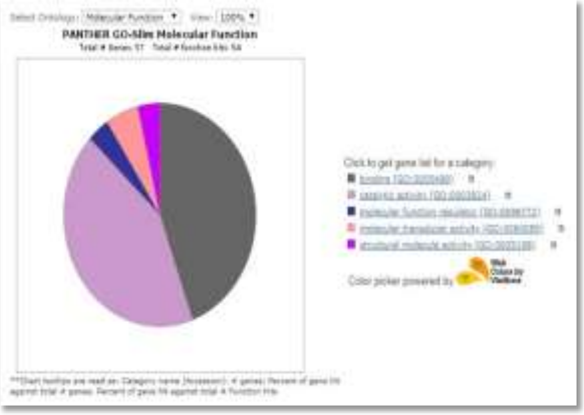


FIGURE 3 : Molecular function analysis of upregulated gene.

Additionally, the gene list was carried forward for further analysis in molecular function, biological function and pathway analysis in PANTHER server. The results were displayed in graphical pie chart for all three analysis. The molecular function results showed 44.4% binding, 42.6% catalytic activity, 3.7% molecular function regulator, 5.6% molecular transducer activity and 3.7% structural molecule activity (Figure 3).

The molecular function analysis of downregulated gene showed 37.5% of binding activity, 33.0% of catalytic activity, 9.8% of transporter activity, 7.1% molecular function regulator, 6.3% molecular transducer activity, 3.6% transcription regulator activity and 2.7% structural molecule activity (Figure 4).

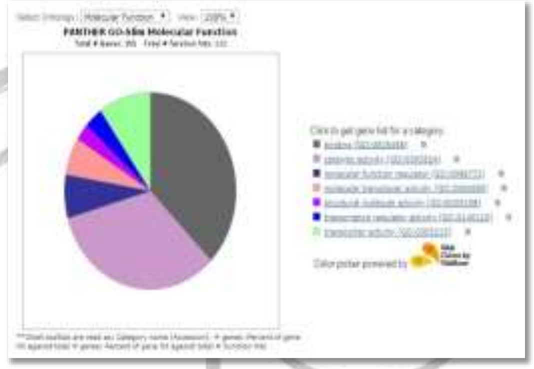
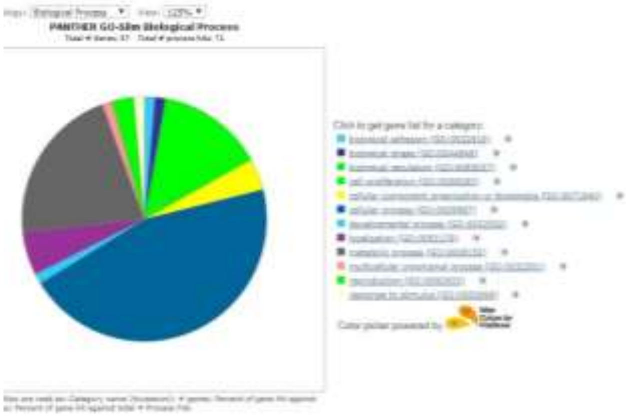


FIGURE 4 : Molecular function analysis of downregulated gene.

The biological process analysis of upregulated gene showed 14% biological adhesion, 1.4% biological phase, 12.7% biological regulation, 1.4% cell proliferation, 4.2% cellular component organization, 45.1% cellular process, 1.4% developmental process, 5.6% localization, 21.1% metabolic process, 1.8% multicellular organismal process, 2.8% reproduction and 1.4% response to stimulus (Figure 5).


PANTHER GO-Slim Biological Process
Click to get gene list for a category:
Color picker powered by

s of upregulated gene.

The biological process analysis of downregulated gene showed 35.7% cellular process, 18.6% metabolic process, 13.2% biological regulation, 9.3% localization, 9.3% response to stimulus, 5.4% immune system process, 5.4% multicellular organismal process, 0.8% biological adhesion and 0.8% pigementation (Figure 6).

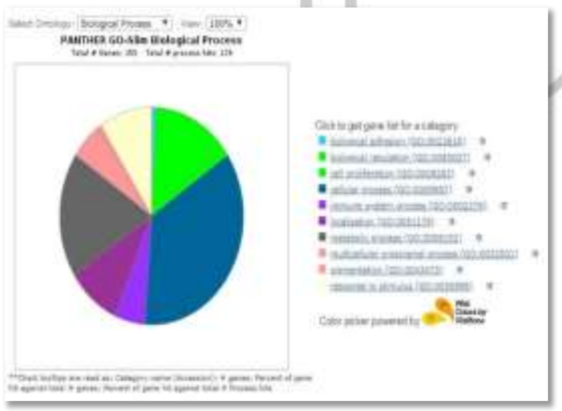

Select Ontology: Biological Process
View: 100%
PANTHER GO-Slim Biological Process
Click to get gene list for a category:
Color picker powered by

FIGURE 6 : Biological process analysis of downregulated gene.

The cellular component analysis of upregulated gene depicted 23.5% cell, 2.0% extracellular region, 52.9% organelle, 19.6% protein-containing complex and 2.0% synapse (Figure 7).

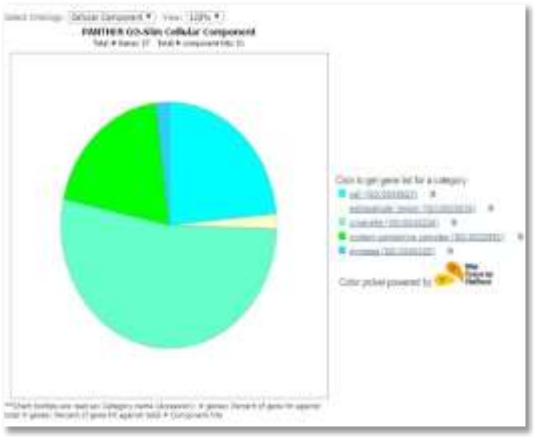

Select Ontology: Cellular Component
PANTHER GO-Slim Cellular Component
Click to get gene list for a category:
Color picker powered by

FIGURE 7 : Cellular component analysis of upregulated gene.

The cellular component analysis of downregulated gene depicted 52.8% cell, 19.1% organelle, 14.6% membrane, 9.0% extracellular region and 4.5% protein containing complex (Figure 8).

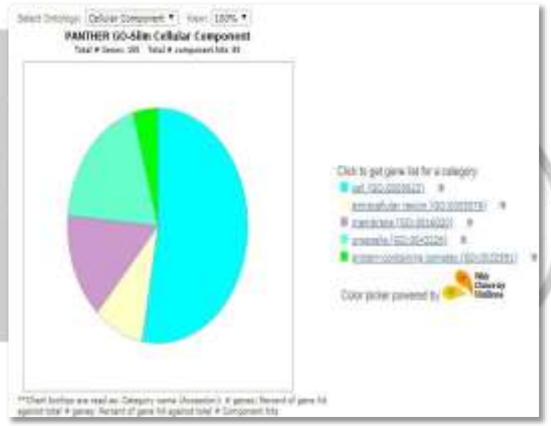

Select Ontology: Cellular Component
View: 100%
PANTHER GO-Slim Cellular Component
Click to get gene list for a category:
Color picker powered by

FIGURE 8 : Cellular component analysis of downregulated gene.

STRING server results

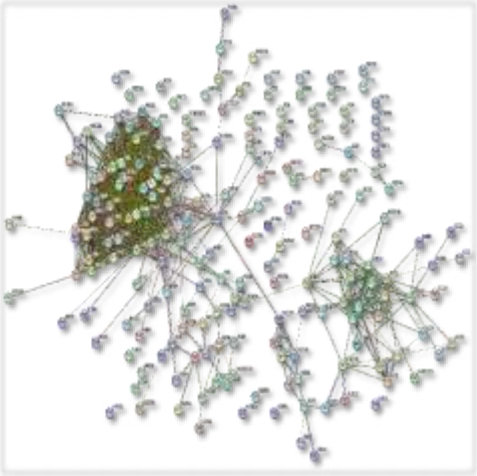


FIGURE 9 : Upregulated gene protein-protein interaction network.

The upregulated gene analysis results presented functional enrichments in the gene network with biological process, molecular function and cellular component as shown below (Figure 9). Biological processes presented the positive regulation of transcription regulatory region DNA, mitotic cell cycle process, sister chromatid cohesion, DNA metabolic process and aortic valve morphogenesis. The KEGG pathways described fat digestion and absorption pathway, pancreatic secretion and cell cycle pathway. In the reactome pathway, five pathways were exposed, the S phase, G1/S transition, activation of the pre-replicative complex, activation of ATR in response to replication stress and E2F mediated regulation of DNA replication.

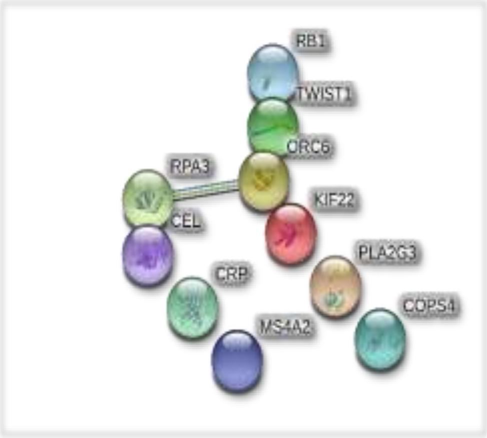
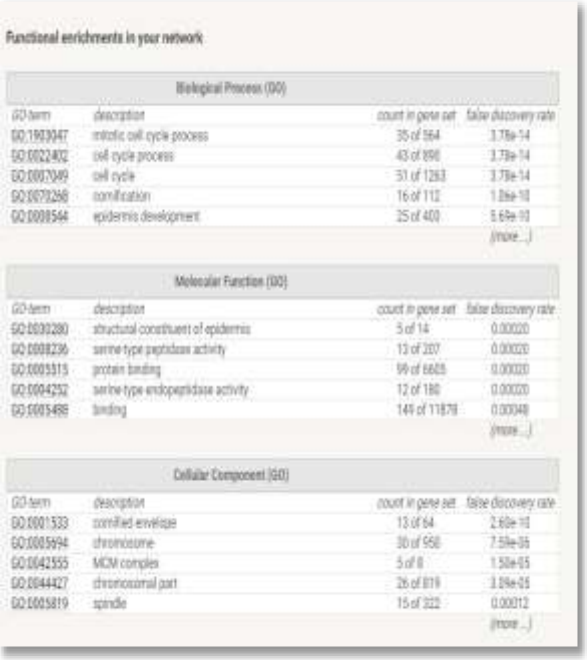


FIGURE 10 : Downregulated gene protein-protein interaction network.



Functional enrichments in your network

Biological Process (GO)

GO-term	description	count in gene set	false discovery rate
GO:1903047	mitotic cell cycle process	35 of 564	3.78e-14
GO:0022402	cell cycle process	43 of 898	3.79e-14
GO:0007049	cell cycle	51 of 1263	3.79e-14
GO:0070268	cornification	16 of 112	1.28e-10
GO:0008544	epidermis development	25 of 400	5.69e-10
			(more...)

Molecular Function (GO)

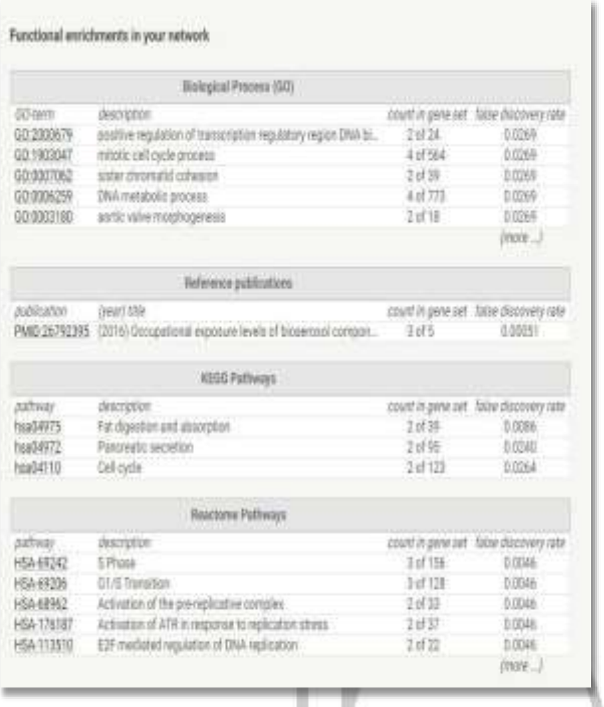
GO-term	description	count in gene set	false discovery rate
GO:0030280	structural constituent of epidermis	5 of 14	0.00020
GO:0008236	serine-type peptidase activity	13 of 207	0.00020
GO:0005515	protein binding	99 of 6605	0.00020
GO:0004252	serine-type endopeptidase activity	12 of 180	0.00020
GO:0005488	binding	149 of 11878	0.00048
			(more...)

Cellular Component (GO)

GO-term	description	count in gene set	false discovery rate
GO:0001533	cornified envelope	13 of 64	2.60e-10
GO:0005694	chromosome	30 of 958	7.59e-05
GO:0042555	MCM complex	5 of 8	1.50e-05
GO:0044427	chromosomal part	26 of 819	3.09e-05
GO:0005819	spindle	15 of 322	0.00012
			(more...)

FIGURE 11: Upregulated gene protein-protein interaction network results.

Moving on, the downregulated gene analysis displayed the functional enrichments in the gene sample network where there were biological process, KEGG Pathways and Reactome Pathways (Figure 4.6.10). The biological process analysis presented mitotic cell cycle process, cell cycle, cell cycle process, cornification and epidermis development. In the molecular function category, there were structural constituent of epidermis, serine type peptidase activity, protein binding, serine type endopeptidase activity and binding. In the cellular component category, cornified envelope, chromosome, MCM complex, chromosomal part and spindle results were presented.



Functional enrichments in your network

Biological Process (GO)

GO-term	description	count in gene set	false discovery rate
GO:2000679	positive regulation of transcription regulatory region DNA bi...	2 of 24	0.0269
GO:1903047	mitotic cell cycle process	4 of 564	0.0269
GO:0007062	sister chromatid cohesion	2 of 39	0.0269
GO:0006259	DNA metabolic process	4 of 773	0.0269
GO:0003180	aortic valve morphogenesis	2 of 18	0.0269
			(more ...)

Reference publications

publication	(year) title	count in gene set	false discovery rate
PMID:26792395	(2016) Occupational exposure levels of bioaerosol compon...	3 of 5	0.00051

KEGG Pathways

pathway	description	count in gene set	false discovery rate
hsa04975	Fat digestion and absorption	2 of 39	0.0086
hsa04972	Pancreatic secretion	2 of 95	0.0240
hsa04110	Cell cycle	2 of 123	0.0264

Reactome Pathways

pathway	description	count in gene set	false discovery rate
HSA-69242	S Phase	3 of 156	0.0046
HSA-69206	G1/S Transition	3 of 128	0.0046
HSA-68962	Activation of the pre-replicative complex	2 of 33	0.0046
HSA-176187	Activation of ATR in response to replication stress	2 of 37	0.0046
HSA-113510	E2F mediated regulation of DNA replication	2 of 22	0.0046
			(more ...)

FIGURE 12 : Downregulated gene STRING analysis results.

The upregulated genes protein-protein interaction network depicted perfectly in the image screen captured below. The discovered centre for hub genes were AURKA, UBE2C, TPX2, MCM5 and GINS2.

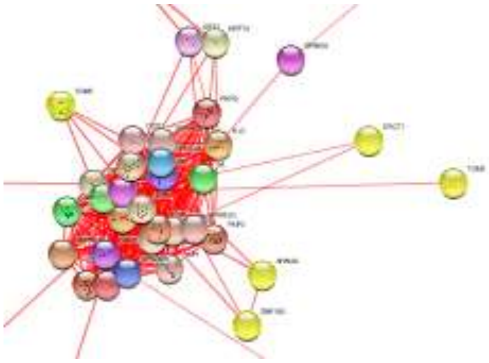


FIGURE 14: Downregulated gene list analysis on molecular interaction network of cervical cancer genes.

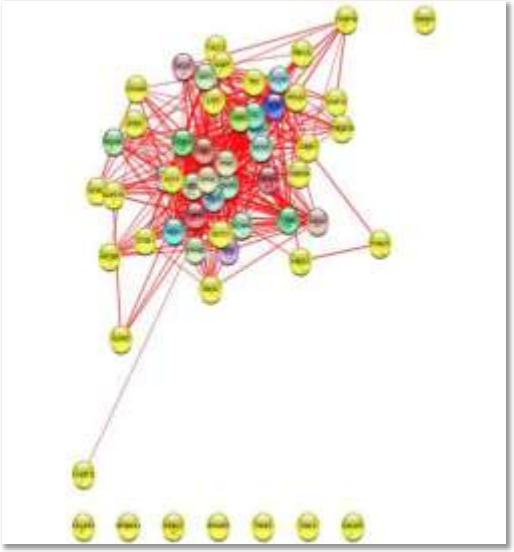


FIGURE 13 : Upregulated gene list analysis on molecular interaction network of cervical cancer genes.

The downregulated genes protein-protein interaction network depicted perfectly in the image screen captured below. The discovered centre for hub genes were SPRR1B, DSC3, SPRR1A, DSG1 and SPRR2D (Figure.14).

Then, the combination of the hub genes were created with a more dominated and transparent view as shown in the image below (Figure 15) of the hub genes involving upregulated and downregulated genes.

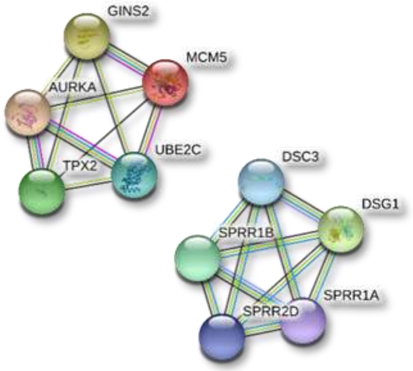


FIGURE 15 : The merged network of strongest 10 cervical cancer hub genes from upregulated and downregulated gene lists, performed in STRING server.

development [15]. The gene list potrayed a whole complete results of this specific gene only and this is supported by the previous studies where it is identified as an oncogene in carcinogenic cell proliferation [16]

The PANTHER analysis results depicted heavily on molecular function, biological processes and cellular component of gene list. The binding receptors were reported to be 44.4% in upregulated and 37.5% in downregulated gene lists. The human papillomavirus gene mechanism supported this result as the pathway of HPV invasion in the keratinocytes were binding receptors-targeted [17]. HPV gene mechanism occured through clathrin- dependent endocytic mechanism and this confirms the cellular processes results where metabolic process and biological adhesion emerged as the top three highest percentages in biological process analysis [17].

The STRING analysis results showed the interactions of the protein to protein interaction network, the KEGG pathways and biological processes that happened.

In Cytoscape analysis, the strongest five hub genes in upregulated gene list analyzed were aurora kinase A (AURKA), E2 ubiquitin-conjugating enzyme C (UBE2C), differentially

DISCUSSIONS

The analysis performed on the gene dataset GSE9750 in DAVID server gave the result of AURKA, a protein coding gene, being highly expressed in it. Aurora kinase A is a cancer related enzyme. This enzyme can be found in the process of microtubule stabilization or formation during chromosome segregation. This gene has the possibility of playing a huge role in tumour progression and expressed gene in cancerous and non-cancerous lung cells 2 (TPX2), minichromosome maintenance complex component 5 (MCM5) and DNA replication complex GINS protein PSF2 (GINS2).

In addition to that, in downregulated gene list analysis, five strongest hub genes were small proline-rich protein 1B (SPRR1B), desmosomal glycoprotein III (DSC3), small proline-rich protein 1A (SPRR1A), pemphigus foliaceus antigen (DSG1) and small proline-rich protein 2D (SPRR2D) were dominating. The protein network genes showed interaction of cancer biomarkers identification in the visualization form.

CONCLUSION

In a nutshell, the hub genes was successfully identified from GSE9750 Geodataset via gene expression analysis for its biological, molecular and cellular component. The pathway mechanism was fully elaborated and depicted in hopes that in the future, an in-depth pathway study on cervical cancer gene biomarker analysis are broaden.

REFERENCES

[1] Z. Shaffie, "A Review of Cervical Cancer Research in Malaysia."

[2] S. N. Buang, S. Ja'afar, I. Pathmanathan, and V. Saint, "Human papillomavirus immunisation of adolescent girls: Improving coverage through multisectoral collaboration in Malaysia," *BMJ (Online)*, vol. 363. BMJ Publishing Group, 2018.

[3] S. Ramakrishnan, S. Partricia, and G. Mathan, "Overview of high-risk HPV's 16 and 18 infected cervical cancer: Pathogenesis to prevention," *Biomedicine and Pharmacotherapy*, vol. 70, no. C. Elsevier Masson SAS, pp. 103–110, 2015.

[4] H. R. Mcmurray, D. Nguyen, T. F. Westbrook, and D. J. Mcance, "Biology of human papillomaviruses," *Int. J. Exp. Pathol.*, vol. 82, no. 1, pp. 15–33, 2001.

[5] S. N. Buang, S. Ja'afar, I. Pathmanathan, and V. Saint, "Human papillomavirus immunisation of adolescent girls: improving coverage through multisectoral collaboration in Malaysia," *BMJ*, vol. 363, 2018.

[6] R. Edgar, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002.

[7] T. Barrett and R. Edgar, "[19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis," *Methods in Enzymology*, vol. 411. pp. 352–369, 2006.

[8] J. Quackenbush, "Computational analysis of microarray data," *Nature Reviews Genetics*, vol. 2, no. 6. pp. 418–427, 2001.

[9] D. W. Huang *et al.*, "The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists," *Genome Biol.*, vol. 8, no. 9, Sep. 2007.

[10] H. Mi and P. Thomas, "PANTHER pathway: an ontology-based pathway database coupled with data analysis tools.," *Methods Mol. Biol.*, vol. 563, pp. 123–140, 2009.

[11] H. Mi *et al.*, "The PANTHER database of protein families, subfamilies, functions and pathways," *Nucleic Acids Res.*, vol. 33, no. DATABASE ISS., Jan. 2005.

[12] D. Szklarczyk *et al.*, "STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.

[13] B. Gemovic, N. Sumonja, R. Davidovic, V. Perovic, and N. Veljkovic, "Mapping of Protein-Protein Interactions: Web-Based Resources for Revealing Interactomes," *Curr. Med. Chem.*, vol. 26, no. 21, pp. 3890–3910, Feb. 2018.

[14] F. Halakou, E. Sen Kilic, E. Cukuroglu, O. Keskin, and A. Gursoy, "Enriching Traditional Protein-protein Interaction Networks with Alternative Conformations of Proteins," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017.

[15] M. T. Kimura *et al.*, “Two functional coding single nucleotide polymorphisms in STK15 (Aurora-A) coordinately increase esophageal cancer risk.,” *Cancer Res.*, vol. 65, no. 9, pp. 3548–54, May 2005.

[16] M. Guo *et al.*, “Increased AURKA promotes cell proliferation and predicts poor prognosis in bladder cancer,” *BMC Syst. Biol.*, vol. 12, Dec. 2018.

[17] C. A. Horvath, G. A. Boulet, V. M. Renoux, P. O. Delvenne, and J. P. J. Bogers, “Mechanisms of cell entry by human papillomaviruses: An overview,” *Virology Journal*, vol. 7. 2010.